Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Expected Conditional Characteristic Function-based Measures for Testing Independence

Chenlu Ke & Xiangrong Yin

To cite this article: Chenlu Ke & Xiangrong Yin (2020) Expected Conditional Characteristic Function-based Measures for Testing Independence, Journal of the American Statistical Association, 115:530, 985-996, DOI: 10.1080/01621459.2019.1604364

To link to this article: https://doi.org/10.1080/01621459.2019.1604364

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 115, NO. 530, 985–996: Theory and Methods
https://doi.org/10.1080/01621459.2019.1604364

Taylor & Francis
Taylor & Francis Group

Expected Conditional Characteristic Function-based Measures for Testing Independence

Chenlu Ke and Xiangrong Yin

Department of Statistics, University of Kentucky, Lexington, KY

ABSTRACT

We propose a novel class of independence measures for testing independence between two random vectors based on the discrepancy between the conditional and the marginal characteristic functions. The relation between our index and other similar measures is studied, which indicates that they all belong to a large framework of reproducing kernel Hilbert space. If one of the variables is categorical, our asymmetric index extends the typical ANOVA to a kernel ANOVA that can test a more general hypothesis of equal distributions among groups. In addition, our index is also applicable when both variables are continuous. We develop two empirical estimates and obtain their respective asymptotic distributions. We illustrate the advantages of our approach by numerical studies across a variety of settings including a real data example. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2018
Revised March 2019

KEYWORDS
Categorical variables; Characteristic function; Independence; Reproducing kernel Hilbert space.

1. Introduction

Statistical independence/dependence tests have been proposed with a broad variety of measures. However, most classical methods can only detect certain types of dependence or have assumptions that are difficult to assess and meet. For example, the well-known Spearman's correlation can only capture monotonic relationships between the two variables. Likelihood-based methods such as Wilk's Lambda are not applicable if the dimension exceeds the sample size, or when distribution assumptions do not hold. Therefore, testing independence is a challenging task, especially in high-dimensional spaces with complicated dependence structures.

More flexible measures have been developed to overcome these difficulties in statistical literature. Wang, Jiang and Liu (2017) proposed a new measure, G^2, to test whether two univariate continuous random variables are dependent and measure the strength of their relationship. The G^2 can be considered as the piecewise R^2 between the sliced variables. And $G^2 = 0$ if and only if $E(X|Y)$, $E(Y|X)$, $\text{var}(X|Y)$ and $\text{var}(Y|X)$ are all constant, which in fact, is not equivalent to the independence of X and Y. The measure can handle nonlinearity and heteroscedastic errors compared to R^2. Its generalization to continuous multivariate variables is intuitive, but it may become difficult and complicated, due to its slicing scheme. Székely, Rizzo, and Bakirov (2007) proposed a novel measure for multivariate variables termed distance covariance (DCOV) and related distance correlation (DCOR). Unlike the classical correlation or the G^2, DCOR is zero if and only if the random variables are independent. This measure has led to applications in variable selection (Li, Zhong, and Zhu 2012) and dimension reduction (Sheng and Yin 2013, 2016). In addition, developing conditional independence tests has also been attractive since they are essential to statistical inference such as graphical models, Bayesian network analysis and dimension reduction. Su and White (2003, 2007, 2008) proposed a series of difference measures between conditional densities based on smoothing empirical likelihood, conditional characteristic function, and weighted Hellinger distance, respectively. Wang et al. (2015) extended the work of Székely, Rizzo, and Bakirov (2007) and developed a conditional independence measure.

Related research exists in machine learning literature as well. Kernel-based methods have been developed and successfully used for detecting dependence of variables (Bach and Jordan 2002a; Gretton et al. 2005a; Gretton et al. 2005b; Sun et al. 2007). Applications of kernel-based approaches can be found in areas including gene selection (Yamanishi, Vert and Kanehisa 2004), fitting graphical models (Bach and Jordan 2002b), dependence detection in fMRI signals (Gretton et al. 2005), and variable selection (Fukumizu, Bach and Jordan 2004). Hilbert Schmidt independence criterion (HSIC) is one of the kernel-based measures for independence that has been proposed (Gretton et al. 2005a, 2005b, 2008). HSIC is computed as the Hilbert-Schmidt norm of a cross-covariance operator on mappings of variables into reproducing kernel Hilbert spaces (RKHS). Those mappings inherit properties of interest such as independence and homogeneity. Other than covariance, an RKHS dependence statistic can also rely on distance (Smola et al. 2007) or correlation (Dauxois and Nkiet 1998; Bach and Jordan 2002a; Fukumizu, Bach and Gretton 2007) between the feature mappings. Extensions of HSIC include an associated measure for conditional independence developed by Fukumizu et al. (2008). Sejdinovic et al. (2013) proposed a framework that nicely links HSIC with DCOV; that is, DCOV is precisely an example of HSIC.

CONTACT Xiangrong Yin yinxiangrong@uky.edu Department of Statistics, University of Kentucky, 725 Rose St. Lexington, KY 40536-0082.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

In this article, we develop a new class of measures for testing independence of two random vectors based on the discrepancy between the conditional and the marginal characteristic functions. Whereas most of the independence measures treat two variables symmetrically, we consider one of the variables conditioning on the other, which is a common idea in regression, classification, and discriminant analysis. More importantly, when one of the variables is nominal, independence tests based on symmetric measures like HSIC and DCOV still rely on the values of the nominal variable, which is problematic. Hence, there is a lack of appropriate and powerful tests other than the classical (M)ANOVA or nonparametric methods like Kruskal–Wallis. Our work fills this gap. Intuitively, the relation between HSIC/DCOV and our measure is analogous to the relation between a linear regression/correlation and an ANOVA. In fact, our method extends the classical ANOVA to a kernel ANOVA that can test a more general hypothesis of equal distributions among groups. Note that Rizzo and Székely (2010) proposed a measure called distance components (DISCO) that also focuses on multi-sample hypothesis for equal distributions, but our approach generates a much broader class of measures. Essentially, we develop a parallel RKHS framework to HSIC/DCOV that unifies our index and DISCO. Although we are motivated by aforementioned setting, our index is also applicable when both variables are continuous. In addition, if necessary, we can simply obtain a symmetric index by adding a term with switched roles of the two variables.

The rest of the article is organized as follows. Section 2 introduces the new measure, a development in a RKHS framework and affiliated properties. Section 3 constructs two empirical estimates and obtains their respective asymptotic distributions. Section 4 provides an algorithm to carry out independence tests using permutations. Section 5 briefly extends the marginal independence measure to a conditional independence measure. Section 6 numerically demonstrates the advantages of our method compared to some existing approaches. Section 7 concludes the article with a short discussion. All proofs are contained in a supplementary file.

2. Expected Conditional Characteristic Function-based Independence Criterion (ECCFIC)

In this section, we introduce a new class of independence measures through two different approaches and discuss related properties.

2.1. ECCFIC via Bochner's Theorem

Suppose random vectors $\mathbf{X} \in \mathbb{R}^p$ and $\mathbf{Y} \in \mathbb{R}^q$. Let $(\mathbf{X}', \mathbf{Y}')$ be an i.i.d. copy of $(\mathbf{X}, \mathbf{Y})$, $\varphi_{\mathbf{X}}$ denote the characteristic function of $\mathbf{X}$ and $\varphi_{\mathbf{X}|\mathbf{Y}}$ denote the conditional characteristic function of $\mathbf{X}$ given $\mathbf{Y}$. We use $E_{\mathbf{X}_\mathbf{y}}(\cdot)$ to represent conditional expectation $E(\cdot|\mathbf{Y} = \mathbf{y})$ and $E_{\mathbf{X}_\mathbf{y},\mathbf{X}'_\mathbf{y}}(\cdot)$ to denote $E(\cdot|\mathbf{Y} = \mathbf{y}, \mathbf{Y}' = \mathbf{y})$. A hypothesis test of independence between $\mathbf{X}$ and $\mathbf{Y}$ is given by $H_0 : \varphi_{\mathbf{X}|\mathbf{Y}} = \varphi_{\mathbf{X}}$ vs. $H_1 : \varphi_{\mathbf{X}|\mathbf{Y}} \neq \varphi_{\mathbf{X}}$. Thus, it is natural to define a measure based on the discrepancy between $\varphi_{\mathbf{X}|\mathbf{Y}}$ and $\varphi_{\mathbf{X}}$. We consider the following distance-like quantity between the two characteristic functions $\varphi_{\mathbf{X}|\mathbf{Y}}$ and $\varphi_{\mathbf{X}}$:

$$\psi^2_\omega(\mathbf{Y}) \equiv \int |\varphi_{\mathbf{X}|\mathbf{Y}}(\mathbf{u}) - \varphi_{\mathbf{X}}(\mathbf{u})|^2 d\omega(\mathbf{u}),$$

where ω is a finite nonnegative Borel measure on $\mathbb{R}^p$. Although $\psi^2_\omega(\mathbf{Y})$ is an intuitive measure of the discrepancy between $\varphi_{\mathbf{X}|\mathbf{Y}}$ and $\varphi_{\mathbf{X}}$, its calculation may be computationally demanding in practice. However, $\psi^2_\omega(\mathbf{Y})$ can also be generated by a positive semi-definite kernel that is induced by ω, which results in an equivalent but simpler representation. The following theorem of Bochner (Wendland 2004, Theorem 6.6) is the key.

Theorem 1 (Bochner). A continuous function $K : \mathbb{R}^p \to \mathbb{C}$ is positive semi-definite if and only if it is the Fourier transform of a finite nonnegative Borel measure ω on $\mathbb{R}^p$, that is,

$$K(\mathbf{x}) = \int_{\mathbb{R}^p} \mathrm{e}^{-i\mathbf{x}^{\mathrm{T}}\mathbf{u}} d\omega(\mathbf{u}), \tag{1}$$

for any $\mathbf{x} \in \mathbb{R}^p$.

In the remaining text, we state that a positive semi-definite kernel K is induced by a finite nonnegative Borel measure ω if K is defined by ω according to (1). We then obtain an alternative formula for $\psi^2_\omega(\mathbf{Y})$ by applying Bochner's Theorem.

Theorem 2. For a given event $\mathbf{Y} = \mathbf{Y}' = \mathbf{y}$,

$$\psi^2_\omega(\mathbf{y}) \equiv E_{\mathbf{X}_\mathbf{y},\mathbf{X}'_\mathbf{y}} K(\mathbf{X}-\mathbf{X}') - 2E_{\mathbf{X}_\mathbf{y},\mathbf{X}'} K(\mathbf{X}-\mathbf{X}') + E_{\mathbf{X},\mathbf{X}'} K(\mathbf{X}-\mathbf{X}'), \tag{2}$$

where ω is a finite nonnegative Borel measure on $\mathbb{R}^p$ and K is a positive semi-definite kernel induced by ω.

Here we restrict ourselves to a translation-invariant kernel $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ that can be written in terms of the difference of their arguments. We henceforth use notation $\psi^2_K(\mathbf{Y})$ instead of $\psi^2_\omega(\mathbf{Y})$ without ambiguity. Note that $\psi^2_K(\mathbf{Y})$ is a $\mathbf{Y}$-measurable random variable. Then a measure of the independence between $\mathbf{X}$ and $\mathbf{Y}$, $\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y})$, is naturally defined by taking the expectation over $\mathbf{Y}$, that is, $\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y}) \equiv E_\mathbf{Y}\psi^2_K(\mathbf{Y})$. To this end, we formally define our new index.

Definition 1. Let $\mathbf{X}$ and $\mathbf{Y}$ be two random variables on $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. For a given kernel K induced by a finite nonnegative Borel measure ω on $\mathbb{R}^p$, the expected conditional characteristic function-based independence criterion (ECCFIC) is defined as

$$\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y}) \equiv E_\mathbf{Y}\left[\int \left|\varphi_{\mathbf{X}|\mathbf{Y}}(\mathbf{u}) - \varphi_{\mathbf{X}}(\mathbf{u})\right|^2 d\omega(\mathbf{u})\right]. \tag{3}$$

Or equivalently,

$$\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y}) \equiv E_\mathbf{Y} E_{\mathbf{X}_\mathbf{Y},\mathbf{X}'_\mathbf{Y}} K(\mathbf{X}-\mathbf{X}') - E_{\mathbf{X},\mathbf{X}'} K(\mathbf{X}-\mathbf{X}'). \tag{4}$$

Note that the weight function used in DCOV is not integrable and hence, Bochner's theorem does not apply for DCOV. The counterpart of our index using the weight function of DCOV is developed by Yin and Yuan (2019), named as ECD, showing that ECD statistic is actually equivalent to DISCO. In the next two subsections, we introduce a general framework of RKHS that unifies ECD and ECCFIC.

2.2. Generalized ECCFIC via MMD

Previously, we proposed ECCFIC in an intuitive way based on the characteristic functions. However, this new class of measures can also be developed via maximum mean discrepancy (MMD, Gretton et al. 2012a) without the requirement for kernels to be translation-invariant. MMD is the largest difference in expectations over functions in the unit ball of a RKHS and can be used to determine if two samples are drawn from different distributions (Gretton et al. 2012a). MMD can be employed to develop HSIC and hence, to measure statistical independence. Now we follow the framework of Gretton et al. (2012a) and Sejdinovic et al. (2013) to generalize ECCFIC via MMD on real spaces. We begin with an introduction to RKHS.

Definition 2 (RKHS, Sejdinovic et al. 2013, Definition 8). Let $\mathbb{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathbb{H}$ if:

1. $\forall \mathbf{x} \in \mathcal{X}, K(\cdot, \mathbf{x}) \in \mathbb{H}$;
2. $\forall \mathbf{x} \in \mathcal{X}, \forall f \in \mathbb{H}, < f, K(\cdot, \mathbf{x}) >_{\mathbb{H}} = f(\mathbf{x})$.

If $\mathbb{H}$ has a reproducing kernel K, it is said to be a reproducing kernel Hilbert space (RKHS) and denoted by $\mathbb{H}_k$.

A reproducing kernel is positive definite. Conversely, Moore-Aronszajn Theorem (Berlinet and Thomas-Agnan 2011, Theorem 3) states that, for every positive-definite function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, there is an associated RKHS $\mathbb{H}_K$ of real-valued functions on $\mathcal{X}$ with the reproducing kernel k. The map $\phi : \mathcal{X} \to \mathbb{H}_K, \phi : \mathbf{x} \to K(\cdot, \mathbf{x})$ is called the canonical feature map of K. We say that K is a nondegenerate kernel if its feature map is injective. The concept of feature map can be extended to kernel embeddings of finite signed Borel measures on $\mathcal{X}$. Let $\mathcal{M}(\mathcal{X})$ be the set of all finite signed Borel measures on $\mathcal{X}$ and $\mathcal{M}_+^1(\mathcal{X})$ be the set of all Borel probability measures on $\mathcal{X}$.

Definition 3 (Kernel embedding, Sejdinovic et al. 2013, Definition 9). Let K be a measurable kernel on $\mathcal{X}$, and $\nu \in \mathcal{M}(\mathcal{X})$. The kernel embedding of ν into the RKHS, $\mathbb{H}_K$, is $\mu_K(\nu) \in \mathbb{H}_K$ such that $\langle f, \mu_K(\nu) \rangle_{\mathbb{H}_K} = \int f(\mathbf{x}) d\nu(\mathbf{x})$ for all $f \in \mathbb{H}_K$.

The kernel embedding can alternatively be defined by the Bochner integral: $\mu_K(\nu) = \int K(\cdot, \mathbf{x}') d\nu(\mathbf{x}')$. To ensure that the kernel embeddings exist, we need to restrict ourselves to a particular class of measures. For a measurable kernel K on $\mathcal{X}$ and $\theta > 0$, denote,

$$\mathcal{M}_K^\theta(\mathcal{X}) \equiv \left\{ \nu \in \mathcal{M}(\mathcal{X}) : \int K^\theta(\mathbf{x}, \mathbf{x}) d|\nu|(\mathbf{x}) < \infty \right\}.$$

The kernel embedding $\mu_K(\nu)$ is well defined for $\nu \in \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$ by Riesz representation theorem (Sejdinovic et al. 2013). Therefore, kernel embeddings of Borel probability measures in $\mathcal{M}_+^1(\mathcal{X}) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$ exist, and we can introduce a discrepancy between two Borel probability measures in terms of the Hilbert space distance between their embeddings.

Definition 4 (MMD, Sejdinovic et al. 2013, Definition 10). Let K be a measurable kernel on $\mathcal{X}$, and let probability measures $\mathbf{P}$, $\mathbf{Q} \in \mathcal{M}_+^1(\mathcal{X}) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$. The MMD γ_K between $\mathbf{P}$ and $\mathbf{Q}$ is given by

$$\gamma_K(\mathbf{P}, \mathbf{Q}) \equiv \|\mu_K(\mathbf{P}) - \mu_K(\mathbf{Q})\|_{\mathbb{H}_K}.$$

Let $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$ be the probability distribution of $\mathbf{X}$ given $\mathbf{Y}$, $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}}$ be the probability distribution of $\mathbf{X}$ and K be a measurable kernel on $\mathbb{R}^p$. Assuming $\mathbf{P}, \mathbf{Q} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathbb{R}^p)$, we have

$$\begin{aligned}
\gamma_K^2(\mathbf{Y}) &\equiv \gamma_K^2(\mathbf{P}, \mathbf{Q}) \\
&= \langle \mu_K(\mathbf{P}), \mu_K(\mathbf{P}) \rangle_{\mathbb{H}_K} + \langle \mu_K(\mathbf{Q}), \mu_K(\mathbf{Q}) \rangle_{\mathbb{H}_K} \\
&\quad -2 \langle \mu_K(\mathbf{P}), \mu_K(\mathbf{Q}) \rangle_{\mathbb{H}_K} \\
&= E_{\mathbf{X}_\mathbf{Y}} E_{\mathbf{X}'_\mathbf{Y}} K(\mathbf{X}, \mathbf{X}') + E_{\mathbf{X}} E_{\mathbf{X}'} K(\mathbf{X}, \mathbf{X}') \\
&\quad -2 E_{\mathbf{X}_\mathbf{Y}} E_{\mathbf{X}'} K(\mathbf{X}, \mathbf{X}'). \qquad (5)
\end{aligned}$$

Definition 5 (Generalized ECCFIC). Let $\mathbf{X}$ and $\mathbf{Y}$ be two random variables on $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. Assume $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}, \mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathbb{R}^p)$. For a given measurable kernel K on $\mathbb{R}^p$, the generalized ECCFIC, is defined as

$$\begin{aligned}
\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y}) &\equiv E_\mathbf{Y} \gamma_K^2(\mathbf{Y}) \\
&= E_\mathbf{Y} E_{\mathbf{X}_\mathbf{Y}, \mathbf{X}'_\mathbf{Y}} K(\mathbf{X}, \mathbf{X}') - E_{\mathbf{X}, \mathbf{X}'} K(\mathbf{X}, \mathbf{X}'). \qquad (6)
\end{aligned}$$

In particular, $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y}) = \mathcal{I}_K^2(\mathbf{X}|\mathbf{Y})$ if $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x} - \mathbf{x}')$.

2.3. A Unified Framework

We now show that ECD (Yin and Yuan 2019) belongs to the family of generalized ECCFIC measures. The connection between negative type semi-metrics and distance-induced kernels, which are translation-variant positive-definite kernels, is the key to our main result. We begin with an introduction to negative type semi-metrics as to define the generalized ECD as well as distance-induced kernels.

Definition 6 (Sejdinovic et al. 2013, Definitions 1 and 2). Let $\rho : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ be a function such that $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$,

1. $\rho(\mathbf{x}, \mathbf{x}') = 0$ if and only if $\mathbf{x} = \mathbf{x}'$;
2. $\rho(\mathbf{x}, \mathbf{x}') = \rho(\mathbf{x}', \mathbf{x})$;
3. $\forall n \geq 2, \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^p$, and $\alpha_1, \alpha_2, \ldots \alpha_n \in \mathbb{R}$, with $\sum_{i=1}^n \alpha_i = 0$,

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \rho(\mathbf{x}_i, \mathbf{x}_j) \leq 0.$$

Then $(\mathcal{X}, \rho)$ is said to be a negative type semi-metric and ρ is called a semi-metric on $\mathcal{X}$.

Before we proceed to formally define generalized ECD, we need to introduce a new class of Borel measures

$$\widetilde{\mathcal{M}}_\rho^\theta(\mathcal{X}) \equiv \{ \nu \in \mathcal{M}(\mathcal{X}) : \exists \mathbf{x}_0 \in \mathcal{X} \text{ s.t.} \int \rho^\theta(\mathbf{x}, \mathbf{x}_0) d|\nu|(\mathbf{x}) < \infty \}.$$

We say that $\nu \in \widetilde{\mathcal{M}}_\rho^\theta(\mathcal{X})$ has a finite θ-moment ($\theta > 0$) with respect to a semi-metric ρ of negative type.

Definition 7 (Generalized ECD). Suppose $(\mathbb{R}^p, \rho)$ is a semi-metric space of negative type. Assume $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$, $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \widetilde{\mathcal{M}}_\rho^1(\mathbb{R}^p)$, the generalized ECD is defined as

$$\mathcal{V}_\rho^2(\mathbf{X}|\mathbf{Y}) \equiv E_{\mathbf{Y}} D_\rho(\mathbf{Y}),$$

where $D_\rho(\mathbf{Y}) \equiv -\int \rho d([\mathbf{P}-\mathbf{Q}] \times [\mathbf{P}-\mathbf{Q}])$.

Note that if $\rho(\mathbf{x}, \mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|_p$, then $\mathcal{V}_\rho^2(\mathbf{X}|\mathbf{Y})$ is precisely the ECD of Yin and Yuan (2019). Similar to DCOV, we can further extend ECD to a class of α-distance measures by choosing $\rho(\mathbf{x}, \mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\|_p^\alpha$, where $0 < \alpha \leq 2$.

We now introduce distance-induced kernels and illustrate the relation with the semi-metrics of negative type.

Definition 8 (Distance-induced kernel, Sejdinovic et al. 2013, Definition 13). Let ρ be a semi-metric of negative type on $\mathcal{X}$ and let $\mathbf{x}_0 \in \mathcal{X}$. The kernel

$$K(\mathbf{x}, \mathbf{x}') = \frac{1}{2}[\rho(\mathbf{x}, \mathbf{x}_0) + \rho(\mathbf{x}', \mathbf{x}_0) - \rho(\mathbf{x}, \mathbf{x}')]$$

is said to be the distance-induced kernel induced by ρ and centered at $\mathbf{x}_0$.

By varying $\mathbf{x}_0$, ρ induces a family of distance-induced kernels:

$$\mathcal{F}_\rho \equiv \left\{ \frac{1}{2}[\rho(\mathbf{x}, \mathbf{x}_0) + \rho(\mathbf{x}', \mathbf{x}_0) - \rho(\mathbf{x}, \mathbf{x}')] : \mathbf{x}_0 \in \mathcal{X} \right\}.$$

Every $K \in \mathcal{F}_\rho$ is positive definite and nondegenerate, that is, $\mathbf{x} \mapsto K(\cdot, \mathbf{x})$ is injective. In the opposite, any nondegenerate kernel K on $\mathcal{X}$ can generate a valid semi-metric ρ of negative type on $\mathcal{X}$ by defining

$$\rho(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}, \mathbf{x}) + K(\mathbf{x}', \mathbf{x}') - 2K(\mathbf{x}, \mathbf{x}').$$

It is clear that every distance-induced kernel $K \in \mathcal{F}_\rho$ induced by ρ, also generates ρ. Furthermore, if K generates ρ, then $\mathcal{M}_K^{\frac{n}{2}}(\mathcal{X}) = \widetilde{\mathcal{M}}_\rho^{\frac{n}{2}}(\mathcal{X})$ for all $n \in \mathbb{N}$ (Sejdinovic et al. 2013, Proposition 20). Note that $\mathcal{M}_K^1(\mathcal{X}) \subset \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$ and hence, $\mathcal{M}_\rho^1(\mathcal{X}) \subset \mathcal{M}_K^{\frac{1}{2}}(\mathcal{X})$. We are set to build up the connection between $\mathcal{V}_\rho^2(\mathbf{X}|\mathbf{Y})$ and $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})$.

Theorem 3. Let $(\mathbb{R}^p, \rho)$ be a semi-metric space of negative type and let K be any kernel that generates ρ. Suppose $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$ and $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}}$ satisfy $\mathbf{P}, \mathbf{Q} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \widetilde{\mathcal{M}}_\rho^1(\mathbb{R}^p)$. Then $\mathcal{V}_\rho^2(\mathbf{X}|\mathbf{Y}) = 2\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})$.

2.4. Properties of Generalized ECCFIC

To make generalized ECCFIC a legitimate measure of independence, the kernel is required to be characteristic, meaning that the feature mapping from the space of probability measures to the RKHS is injective. Conditions under which kernels are characteristic can be found in Fukumizu et al. (2009) and Sriperumbudur et al. (2008, 2010). Examples of characteristic kernels include Gaussian, Laplacian, and inverse multiquadratics. When K is characteristic, $\gamma_K(\mathbf{P}, \mathbf{Q}) = 0$ iff $\mathbf{P} = \mathbf{Q}$, $\forall \mathbf{P}, \mathbf{Q} \in \mathcal{M}_+^1(\mathbb{R}^p)$ (Sejdinovic et al. 2013). As a result, the following theorem is trivial but important.

Theorem 4. Let K be a characteristic kernel on $\mathbb{R}^p$, and $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$, $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}} \in \mathcal{M}_+^1(\mathbb{R}^p) \cap \mathcal{M}_K^{\frac{1}{2}}(\mathbb{R}^p)$. Then $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent.

Note that $\mathcal{H}_K^2(\mathbf{X}|\mathbf{X}) = E_{\mathbf{X}} K(\mathbf{X}, \mathbf{X}) - E_{\mathbf{X},\mathbf{X}'} K(\mathbf{X}, \mathbf{X}')$. A statistic similar to correlation can be then defined as $\rho_K(\mathbf{X}|\mathbf{Y}) \equiv \dfrac{\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})}{\mathcal{H}_K^2(\mathbf{X}|\mathbf{X})}$.

Theorem 5. Suppose $E_{\mathbf{X}} k(\mathbf{X}, \mathbf{X}) < \infty$. The following properties hold:

1. $0 \leq \mathcal{H}_k^2(\mathbf{X}|\mathbf{Y}) \leq \mathcal{H}_k^2(\mathbf{X}|\mathbf{X}) < \infty$, and thus $0 \leq \rho_k(\mathbf{X}|\mathbf{Y}) \leq 1$.
2. $\rho_k(\mathbf{X}|\mathbf{Y}) = 1$ if and only if $\mathbf{X}$ is a function of $\mathbf{Y}$.

Another critical property of $\mathcal{H}_k^2(\mathbf{X}|\mathbf{Y})$ involves a decomposition analogous to ANOVA. Recall the feature map of a reproducing kernel K, $\phi(\mathbf{x}) = K(\cdot, \mathbf{x})$, and note that $< \phi(\mathbf{x}), \phi(\mathbf{x}') >_{\mathbb{H}_K} = K(\mathbf{x}, \mathbf{x}')$ by Definition 2. If we treat $(\phi(\mathbf{X}), \mathbf{Y})$ rather than $(\mathbf{X}, \mathbf{Y})$ as an individual, then the kernel embedding of $\mathbf{Q} \equiv \mathbf{P}_{\mathbf{X}}$ into the RKHS $\mathbb{H}_K$, $\mu_K(\mathbf{Q})$, can be viewed as the overall mean of $\phi(\mathbf{X})$, while $\mu_K(\mathbf{P})$ for $\mathbf{P} \equiv \mathbf{P}_{\mathbf{X}|\mathbf{Y}}$ can be viewed as the mean of $\phi(\mathbf{X})$ given $\mathbf{Y}$. Let $\mathcal{W}_K^2(\mathbf{X}|\mathbf{Y}) \equiv E_{\mathbf{X},\mathbf{Y}} \|\phi(\mathbf{X}) - \mu_K(\mathbf{P})\|_{\mathbb{H}_K}$, then we have the following decomposition:

$$\begin{aligned} \mathcal{H}_k^2(\mathbf{X}|\mathbf{X}) &= E_{\mathbf{X}} \|\phi(\mathbf{X}) - \mu_K(\mathbf{Q})\|_{\mathbb{H}_K} \\ &= E_{\mathbf{Y}} \|\mu_K(\mathbf{P}) - \mu_K(\mathbf{Q})\|_{\mathbb{H}_K} \\ &\quad + E_{\mathbf{X},\mathbf{Y}} \|\phi(\mathbf{X}) - \mu_K(\mathbf{P})\|_{\mathbb{H}_K} \\ &= \mathcal{H}_k^2(\mathbf{X}|\mathbf{Y}) + \mathcal{W}_K^2(\mathbf{X}|\mathbf{Y}). \end{aligned} \tag{7}$$

If Y is categorical, Equation (7) can be regarded as the population decomposition of total variability into between group dispersion and within group dispersion, which is henceforth referred as the kernel ANOVA population decomposition.

One may consider a more general setting as follows:

$$\mathcal{H}_{K,a}^2(\mathbf{X}|\mathbf{Y}) \equiv E_{\mathbf{Y}}\left[a(\mathbf{Y})\gamma_K^2(\mathbf{Y})\right], \tag{8}$$

where $a(\cdot)$ is a given nonnegative weight function. Note that under the same conditions of Theorem 4, if $a(\cdot)$ has the same support as the probability density function of $\mathbf{Y}$, then $\mathcal{H}_{K,a}^2(\mathbf{X}|\mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent.

3. Empirical Estimators and Asymptotic Properties

In this section, we provide two different approaches to estimate $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})$: a slicing method that can be applied to both a categorical or a continuous $\mathbf{Y}$, and a kernel regression estimation that is intended for a continuous $\mathbf{Y}$ only. When $\mathbf{Y}$ is categorical or sliced, our method extends the typical ANOVA to a kernel ANOVA that is able to test a more general hypothesis of equal distributions among groups. While if $\mathbf{Y}$ is continuous, our test, based on the kernel regression estimator of generalized ECCFIC, provides an alternative to HSIC.

3.1. Slicing on Y: A Kernel ANOVA

Estimating $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y})$ is straightforward when $\mathbf{Y}$ is categorical. If $\mathbf{Y}$ is continuous, we can make it discrete by slicing.

Let $(\mathbf{X}_t, \mathbf{Y}_t)$, $t = 1, \ldots, n$ be a random sample of $(\mathbf{X}, \mathbf{Y})$. Assume that $\mathbf{Y}$ has L levels $\{\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(L)}\}$ with probability $\{p_1, p_2, \ldots, p_L\}$ and each level has n_l observations $(\mathbf{X}_t^{(l)}, \mathbf{y}^{(l)})$, $t = 1, \ldots, n_l$, $l = 1, \ldots, L$. The empirical generalized ECCFIC is defined as

$$\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n}\sum_{l=1}^{L}\frac{1}{n_l}\sum_{i,j=1}^{n_l} K(\mathbf{X}_i^{(l)}, \mathbf{X}_j^{(l)}) - \frac{1}{n^2}\sum_{i,j=1}^{n} K(\mathbf{X}_i, \mathbf{X}_j) \tag{9}$$

$$= \frac{1}{n^2}\text{trace}(\mathbf{KL}), \tag{10}$$

where $\mathbf{K}$ is the $n \times n$ Gram matrix with entries $K_{ij} \equiv K(\mathbf{X}_i, \mathbf{X}_j)$, $\mathbf{L}$ is an $n \times n$ matrix with entries $L_{ij} \equiv \sum_{l=1}^{L} \frac{n}{n_l} I\{\mathbf{Y}_i = \mathbf{y}^{(l)}\} I\{\mathbf{Y}_j = \mathbf{y}^{(l)}\} - 1$ and $I\{\cdot\}$ is the indicator function.

Theorem 6 (Consistency). Assuming that $E_{\mathbf{X}} K(\mathbf{X}, \mathbf{X}) < \infty$, we have

$$\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} \mathcal{H}^2_K(\mathbf{X}|\mathbf{Y}).$$

Theorem 7 (Null Distribution). Under H_0, if $E_{\mathbf{X}} K(\mathbf{X}, \mathbf{X}) < \infty$, then

$$n\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} (L-1)\mathcal{H}^2_K(\mathbf{X}|\mathbf{X}) \sum_{i=1}^{\infty} \lambda_i Z_i^2,$$

where $Z_i \overset{\text{iid}}{\sim} N(0,1)$ and λ_i are positive constants with $\sum_{i=1}^{\infty} \lambda_i = 1$.

A natural consistent estimator of $\mathcal{H}^2_K(\mathbf{X}|\mathbf{X})$ is given by

$$\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X}) \equiv \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{X}_i, \mathbf{X}_i) - \frac{1}{n^2}\sum_{i,j=1}^{n} K(\mathbf{X}_i, \mathbf{X}_j).$$

Define $\rho_{K,n}(\mathbf{X}|\mathbf{Y}) \equiv \dfrac{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})}{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X})}$, then we have the following results.

Corollary 1. Assuming that $E_{\mathbf{X}} K(\mathbf{X}, \mathbf{X}) < \infty$,

1. under H_0, $n\rho_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} (L-1)\sum_{i=1}^{\infty} \lambda_i Z_i^2$, where $Z_i \overset{\text{iid}}{\sim} N(0,1)$ and λ_i are positive constants with $\sum_{i=1}^{\infty} \lambda_i = 1$;
2. under H_1, then $n\rho_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} \infty$.

Recall that we introduced a kernel ANOVA population decomposition in the previous section. We now formulate an empirical kernel ANOVA test for equal distributions among groups, where $n\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})$ has a nice interpretation as the treatment sum of squares. Considering a random sample $(\phi(\mathbf{X}_t), \mathbf{Y}_t)$, $t = 1, \ldots, n$ in RKHS $\mathbb{H}_K$, where ϕ denotes the feature map of kernel K, we have the total sum of squares (SST) as

$$\text{SST} \equiv \sum_{i=1}^{n} \left\| \phi(\mathbf{X}_i) - \frac{1}{n}\sum_{j=1}^{n} \phi(\mathbf{X}_j) \right\|^2_{\mathbb{H}_k}.$$

Then SST can be decomposed into treatment sum of squares (SSTr) and error sum of squares (SSE), that is, SST = SSTr+SSE, where

$$\text{SSTr} \equiv \sum_{l=1}^{L} n_l \left\| \frac{1}{n_l}\sum_{i=1}^{n_l} \phi(\mathbf{X}_i^{(l)}) - \frac{1}{n}\sum_{j=1}^{n} \phi(\mathbf{X}_j) \right\|^2_{\mathbb{H}_k}$$

and

$$\text{SSE} \equiv \sum_{l=1}^{L}\sum_{i=1}^{n_l} \left\| \phi(\mathbf{X}_i^{(l)}) - \frac{1}{n_l}\sum_{j=1}^{n_l} \phi(\mathbf{X}_j^{(l)}) \right\|^2_{\mathbb{H}_k}.$$

After some algebra, we can show that, in fact, SST $= n\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X})$ and SSTr $= n\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})$. As a consequence, we can propose a test statistic analogous to the F-statistic in ANOVA, namely

$$\mathcal{F}_{K,n}(\mathbf{X}|\mathbf{Y}) \equiv \frac{\text{SSTr}/(L-1)}{\text{SSE}/(n-L)} = \frac{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})/(L-1)}{(\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{X}) - \mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}))/(n-L)}.$$

Under H_0, $\mathcal{F}_{K,n}(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} Q$, where $Q = \sum_{i=1}^{\infty} \lambda_i Z_i^2$, $Z_i \overset{\text{iid}}{\sim} N(0,1)$ and $E(Q) = 1$. Székely and Bakirov (2003) proved that

$$P\left(Q \geq \left(\Phi^{-1}(1-\alpha_0/2)\right)^2\right) \leq \alpha_0,$$

for $\alpha_0 \leq 0.215$, where $\Phi(\cdot)$ is the standard normal c.d.f.

3.2. Kernel Regression Estimation

In this section, we construct a more sophisticated estimator via kernel estimation for a continuous $\mathbf{Y}$. We start from an alternative formula of $\gamma^2_K(\mathbf{y})$. By formula (5),

$$\begin{aligned}\gamma^2_K(\mathbf{y}) &= E\left[K(\mathbf{X}_1, \mathbf{X}_2)|\mathbf{Y}_1 = \mathbf{y}, \mathbf{Y}_2 = \mathbf{y}\right] \\ &\quad - E\left[K(\mathbf{X}_1, \mathbf{X}_3)|\mathbf{Y}_1 = \mathbf{y}\right] \\ &\quad - E\left[K(\mathbf{X}_2, \mathbf{X}_4)|\mathbf{Y}_2 = \mathbf{y}\right] + EK(\mathbf{X}_3, \mathbf{X}_4) \\ &= E\left[d_{1234}|\mathbf{Y}_1 = \mathbf{y}, \mathbf{Y}_2 = \mathbf{y}\right],\end{aligned}$$

where $d_{1234} \equiv K_{12} - K_{13} - K_{24} + K_{34}$ and $K_{ts} \equiv K(\mathbf{X}_t, \mathbf{X}_s)$.

For a kernel function $G : \mathbb{R}^q \to \mathbb{R}$ and a bandwidth $h \equiv h(n)$, define $G_h(\mathbf{y}) \equiv h^{-q} G(\mathbf{y}/h)$. The Nadaraya–Watson kernel estimator of the conditional expectation $\gamma^2_K(\mathbf{Y}_{t_1})$ is given by

$$\gamma^2_{K,n}(\mathbf{Y}_{t_1}) = \frac{\frac{1}{n^4}\sum_{t_2t_3t_4t_5} G_{t_1t_2} G_{t_1t_3} d_{t_2t_3t_4t_5}}{\widehat{f}^2_h(\mathbf{Y}_{t_1})},$$

where $G_{ts} \equiv G_h(\mathbf{Y}_t - \mathbf{Y}_s)$ and $\widehat{f}_h(\mathbf{Y}_{t_1}) \equiv \frac{1}{n}\sum_{s=1}^{n} G_{t_1 s}$. A natural estimator immediately follows as

$$\Gamma_{K,G,n}(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n^5}\sum_{t_1t_2t_3t_4t_5} \frac{G_{t_1t_2} G_{t_1t_3} d_{t_2t_3t_4t_5}}{\widehat{f}^2_h(\mathbf{Y}_{t_1})}. \tag{11}$$

Note that the smoothing kernel G applied on $\mathbf{Y}$ is different from the reproducing kernel K applied on $\mathbf{X}$. We have different requirements on choosing these two kernels, although we use Gaussian for both in our simulations later.

It is known that kernel estimate suffers from random denominator issues, but it can be alleviated by different strategies. Intuitively, we may either assume that the density of $\mathbf{Y}$, $f(\mathbf{y})$, is bounded below by some positive number or impose a trimming function $a_\epsilon(\mathbf{y}) = I\{f(\mathbf{y}) > \epsilon\}$, where $\epsilon > 0$. Another option is to apply a proper weight function $a(\cdot)$ on $\mathbf{Y}$ for the same purpose of dealing with the possible large bias near 0. We adopt the latter approach as in Su and White (2007) and Wang et al. (2015) to deal with the weighted measure $\mathcal{H}^2_{K,a}(\mathbf{X}|\mathbf{Y})$. Consider the following estimator:

$$\Gamma_{K,G,a,n}(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n^5} \sum_{t_1 t_2 t_3 t_4 t_5} \frac{G_{t_1 t_2} G_{t_1 t_3} d_{t_2 t_3 t_4 t_5} a(\mathbf{Y}_{t_1})}{\widehat{f}^2_h(\mathbf{Y}_{t_1})}. \quad (12)$$

We choose $a(\mathbf{Y}_t) = f^2_h(\mathbf{Y}_t)$ ($\hat{a}(\mathbf{Y}_t) = \widehat{f}^2_h(\mathbf{Y}_t)$) so that the denominator in (12) vanishes and there is no need for any additional assumption or trimming function. Eventually, we obtain the following statistic (some subscripts are omitted without ambiguity for simplicity):

$$\Gamma^V_n(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n^5} \sum_{t_1 t_2 t_3 t_4 t_5} G_{t_1 t_2} G_{t_1 t_3} d_{t_2 t_3 t_4 t_5} \quad (13)$$

$$= \frac{1}{n^3} \text{trace}(\mathbf{KHGGH}), \quad (14)$$

where $\mathbf{G}$ is a $n \times n$ matrix with entries G_{ij}, $\mathbf{H} \equiv \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is an n vector of ones. A corresponding bias-adjusted statistic is given by

$$\Gamma^U_n(\mathbf{X}|\mathbf{Y}) \equiv \frac{1}{n_5} \sum_{t_1 \neq t_2 \neq t_3 \neq t_4 \neq t_5} G_{t_1 t_2} G_{t_1 t_3} d_{t_2 t_3 t_4 t_5}, \quad (15)$$

where $n_5 = n(n-1)(n-2)(n-3)(n-4)$.

Developing the asymptotic properties of $\Gamma^U_n(\mathbf{X}|\mathbf{Y})$ requires some regularity conditions, which are common in the literature for kernel estimation (Su and White 2007; Wang et al. 2015). Let $\mathcal{F}^\alpha_\mu$ ($\mu > 0$ and $\alpha > 0$) be the class of functions $f : \mathbb{R}^q \to \mathbb{R}$ satisfying: f is $(m-1)$-times partially differentiable, for $m-1 < \mu \leq m$; for some $\rho > 0$, $\sup_{\mathbf{y}' \in \phi_{\mathbf{y}\rho}} |f(\mathbf{y}') - f(\mathbf{y}) - Q_f(\mathbf{y}, \mathbf{y}')|/||\mathbf{y}' - \mathbf{y}||^\mu \leq R_f(\mathbf{y})$ for all $\mathbf{y}$, where $\phi_{\mathbf{y}\rho} \equiv \{\mathbf{y}' : ||\mathbf{y}' - \mathbf{y}|| < \rho\}$; $Q_f = 0$ when $m = 1$; Q_f is a $(m-1)$th degree homogeneous polynomial in $\mathbf{y}' - \mathbf{y}$ with coefficients the partial derivatives of f at $\mathbf{y}$ of orders 1 through $m-1$ when $m > 1$; and f, its partial derivatives of order $m-1$ and less, and R_f, have finite αth moments (Robinson 1988). Our conditions are as follows.

(C1) The kernel G is a product of univariate kernel $g : \mathbb{R} \to \mathbb{R}$ such that $\int_{\mathbb{R}} u^i g(u) du = \delta_{i0}$ ($i = 0, 1, \ldots, \nu - 1$), and $g(u) = O((1 + |u|^{\nu+1+\epsilon})^{-1})$ for some $\epsilon > 0$, where $\nu \geq 2$ is an integer and δ_{ij} is Kronecker's delta.
(C2) $h^q \to 0$ and $nh^q \to \infty$ as $n \to \infty$.
(C3) The marginal density of $\mathbf{Y}$, $f(\mathbf{y}), \in \mathcal{F}^\infty_\nu$.
(C4) Let $m(\mathbf{y}) \equiv E_{\mathbf{X}_\mathbf{y}} K(\mathbf{X}, \mathbf{X})$, then $m(\mathbf{y}) \in \mathcal{F}^\infty_\nu$.

Condition (C1) characterizes the class of νth order kernel function and it implies $\int u^\nu g(u) du < \infty$. Condition (C2) requires the bandwidth to be chosen appropriately according to n. Conditions (C3) and (C4) impose smoothness and moment conditions on the marginal and conditional distribution. Similar constraints were used in Su and White (2007).

Theorem 8 (Consistency). Assume that conditions (C1)–(C4) hold and $E_\mathbf{X} K^2(\mathbf{X}, \mathbf{X}) < \infty$, then

$$\Gamma^U_n(\mathbf{X}|\mathbf{Y}) \xrightarrow{P} E_\mathbf{Y}\left[\gamma^2_K(\mathbf{Y}) f^2(\mathbf{Y})\right].$$

Theorem 9 (Asymptotic Null Distribution). Assume that conditions (C1)–(C3) hold and $E_\mathbf{X} K^2(\mathbf{X}, \mathbf{X}) < \infty$. Under H_0, we have

$$nh^{q/2} \Gamma^U_n(\mathbf{X}|\mathbf{Y}) \xrightarrow{d} N(0, 2\sigma^2),$$

where $\sigma^2 = C^q \left[E K^2_{12} - 2E K_{12} K_{13} + E^2 K_{12}\right] E f^3(\mathbf{Y})$ and $C = \int_{\mathbb{R}} [\int_{\mathbb{R}} g(\mu + \nu) g(\mu) d\mu]^2 d\nu$.

4. An Algorithm Via Permutation Procedure

Nonparametric tests that rely on the asymptotic results may have poor power in finite samples (Su and White 2007). An alternative is to use permutation approach (Efron and Tibshirani 1994; Davison and Hinkley 1997) and it has been proved to be successful in area that is related to our measure, such as DCOV tests (Székely, Rizzo and Bakirov 2007; Székely and Rizzo 2009) and DISCO tests (Rizzo and Székely 2010). Davison and Hinkley (1997) suggested that at least 99 and at most 999 random permutations should be sufficient practically. We illustrate the permutation procedure using test statistic $\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})$ as follows:

1. Compute $\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})$ using formula (9) or (10) for the data;
2. For each replicate $b = 1, \ldots, B$, generate a random permutation π^b and compute the statistic of permuted sample $(\mathbf{X}_k, \mathbf{Y}_{\pi^b(k)})$, $k = 1, \ldots, n$, denoted by $\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}_b)$;
3. Compute the empirical p-value by

$$\widehat{p} = \frac{1 + \sum_{b=1}^B I\left\{\mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y}_b) \geq \mathcal{H}^2_{K,n}(\mathbf{X}|\mathbf{Y})\right\}}{B+1},$$

where $I\{\cdot\}$ is the indicator function.

5. An Extension to Conditional version

Although we used the conditional characteristic function to develop ECCFIC at the beginning, the measure is still a marginal test statistic. However, we can simply extend ECCFIC to a conditional independence measure for testing $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}|\mathbf{Z}$ based on the same idea.

Let $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ be three random vectors in $\mathbb{R}^{p_1}$, $\mathbb{R}^{p_2}$ and $\mathbb{R}^q$, respectively. For a given translation-invariant RKHS kernel K, we consider the following discrepancy between two characteristic functions $\varphi_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$ and $\varphi_{\mathbf{X}|\mathbf{Z}}$:

$$\widetilde{\psi}^2_K(\mathbf{Y}, \mathbf{Z}) \equiv \int |\varphi_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}(\mathbf{u}) - \varphi_{\mathbf{X}|\mathbf{Z}}(\mathbf{u})|^2 \omega(\mathbf{u}) d\mathbf{u},$$

where ω is a finite nonnegative Borel measure on $\mathbb{R}^{p_1}$ that induces K. Then we define the expected conditional characteristic function-based conditional independence criterion (ECCFCIC) as $\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv E_{(\mathbf{Y},\mathbf{Z})} \widetilde{\psi}^2_K(\mathbf{Y}, \mathbf{Z})$. After some algebra, we can show that

$$\mathcal{I}^2_K(\mathbf{X}|\mathbf{Y};\mathbf{Z}) = E_{(\mathbf{Y},\mathbf{Z})} E_{\mathbf{X}_{(\mathbf{Y},\mathbf{Z})}, \mathbf{X}'_{(\mathbf{Y},\mathbf{Z})}} K(\mathbf{X} - \mathbf{X}')$$
$$- E_\mathbf{Z} E_{\mathbf{X}_\mathbf{Z}, \mathbf{X}'_\mathbf{Z}} K(\mathbf{X} - \mathbf{X}')$$

$$= \mathcal{I}_K^2(\mathbf{X}|(\mathbf{Y},\mathbf{Z})) - \mathcal{I}_K^2(\mathbf{X}|\mathbf{Z}).$$

That is, ECCFCIC can be written as the difference between two marginal indices. We can also develop a generalized ECCFCIC via MMD, which will result in a more general index as follows:

$$\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv \mathcal{H}_K^2(\mathbf{X}|(\mathbf{Y},\mathbf{Z})) - \mathcal{H}_K^2(\mathbf{X}|\mathbf{Z}).$$

Then we can easily estimate $\mathcal{H}_K^2(\mathbf{X}|\mathbf{Y};\mathbf{Z})$ by

$$\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y};\mathbf{Z}) \equiv \mathcal{H}_{K,n}^2(\mathbf{X}|(\mathbf{Y},\mathbf{Z})) - \mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Z}).$$

We omit the asymptotic properties of $\mathcal{H}_{K,n}^2(\mathbf{X}|\mathbf{Y};\mathbf{Z})$ as well as the algorithm here as they follow straightforward from Sections 3 and 4.

Note that Su and White (2007) proposed a conditional independence measure, denoted as Γ, that is also based on the difference between $\varphi_{\mathbf{X}|\mathbf{Y},\mathbf{Z}}$ and $\varphi_{\mathbf{X}|\mathbf{Z}}$. However, their index and ours are distinct in terms of how to quantify the discrepancy. Su and White (2007) first take a Fourier transformation of the difference $\varphi_{\mathbf{X}|\mathbf{Y},\mathbf{Z}} - \varphi_{\mathbf{X}|\mathbf{Z}}$ (adding an extra parameter), then compute the square norm, while we directly measure the distance between the two characteristic functions, resulting in a simpler formula. In addition, while our index has a clear and intuitive interpretation of kernel ANOVA and RKHS, the counterpart explanation for Γ is not clear. Along the line of ECCFIC, we can estimate ECCFCIC through two paths. One is the slicing method intending for categorical $\mathbf{Y}$ and $\mathbf{Z}$, which was not considered in Su and White (2007). For continuous $\mathbf{Y}$ and $\mathbf{Z}$, we apply the kernel regression estimation approach as in Su and White (2007) as well as Wang et al. (2015) mainly for estimating the conditional expectation terms. Again, the smoothing kernel we apply on $(\mathbf{Y},\mathbf{Z})$ or $\mathbf{Z}$ is different from the reproducing kernel we apply on $\mathbf{X}$, while Su and White (2007) only used the smoothing kernel.

6. Numerical Studies

In this section we provide empirical examples of independence tests using ECCFIC and power comparisons with other existing tests, in particular, the HSIC, I^{NOCCO} (Fukumizu et al. 2008), DCOV and DISCO (or equivalently, ECD). All the tests are implemented using the permutation procedure presented in Section 4. R code for ECCFIC is included in supplements but is also available upon request from the first author.

Basic settings are as follows unless otherwise specified. For ECCFIC, HSIC and I^{NOCCO}, Gaussian kernel, which is a translation-invariant characteristic kernel, is applied with parameter σ setting to the heuristic median pairwise distances of the data (Gretton et al. 2008), although more sophisticated methods are available (Fukumizu et al. 2009; Gretton et al. 2012b). For the kernel regression estimation of a continuous $\mathbf{Y}$, we assume $a(\cdot) \equiv 1$ and use Gaussian kernel for G in formula (11). The bandwidth h is set to $1.06\tilde{\sigma}n^{-\frac{1}{5}}$ as Silverman (1986) suggested, where n is the sample size and $\tilde{\sigma}$ is estimated by sample standard deviation. One may also use cross-validation, test graph method and other techniques to choose the smoothing parameter. $B = [200 + 5000/n]$ permutation replicates are carried out in each test (Székely, Rizzo, and Bakirov 2007). ϵ_n of I^{NOCCO} is set to be 10^{-6}. Empirical power or Type I error rate is computed as the proportion of significant tests on 10,000 random samples at significance level of 0.1.

Example 1. This example is to examine the Type I error rates, similarly to Example 1 in Székely, Rizzo and Bakirov (2007). Set $\mathbf{X} \in \mathbb{R}^5$ and $\mathbf{Y} \in \mathbb{R}$ to be independent. In model (a), $\mathbf{X} \sim N(\mathbf{0}, I_5)$, $\mathbf{Y} \sim N(0,1)$. In model (b)–(d), we repeat the same scheme except that the marginals of $\mathbf{X}$ and $\mathbf{Y}$ are $t(\nu)$, $\nu = 1,2,3$. For our slicing method and DISCO, the number of slices is set to 5. Results in Table 1 indicate that the empirical Type I error rates of all the methods are under reasonable control.

Example 2. This example is to examine the performance of ECCFIC when one of the variables is categorical. The setting imitates Example 3 in Rizzo and Székely (2010), a four group balanced design with common sample size 30. Data are generated from distributions with identical independent marginals. Groups 1–3 all have central t(4) distributions as marginals. Group 4 has a mixture distribution of two equal-weighted noncentral t(4) with noncentrality parameter δ and $-\delta$, respectively. We treat group indicator as response naturally. Monte Carlo power comparison of our method with others is summarized in Figure 1. We use 199 permutations in each test.

In Figure 1(a), power curves with respect to noncentrality parameter δ are presented with dimension fixed at $p = 10$. Each method roughly achieves the nominal significance level 0.1 under the null hypothesis (δ=0). ECCFIC test is generally more powerful than I^{NOCCO}, DCOV and DISCO but slightly less preferred than HSIC. In Figure 1(b), noncentrality parameter is fixed at $\delta = 1$ and power varies with dimension p. We notice that I^{NOCCO} is less capable of detecting dependence than ECCFIC and HSIC when p is small, although it outperforms all the other methods when p gets to 20.

We then replace group indicator 1–4 by 1, 8, 0.2 and 2.5. Figure 1(c,d) shows that the power of HSIC and DCOV decreases dramatically while the performance of our method and DISCO remain the same. I^{NOCCO} is also robust to this variation. When $\mathbf{Y}$ is nominal and its values are not meaningful, ECCFIC tests fix the issue of symmetric measures like HSIC and DCOV because they are only subject to the cohorts of the data but not the values of $\mathbf{Y}$, as seen in (9).

Table 1. Example 1: Empirical Type I error rates of ECCFIC, HSIC, I^{NOCCO}, DCOV, and DISCO tests.

		Method					
Model	n	ECCFIC (kernel)	HSIC	I^{NOCCO}	DCOV	ECCFIC (slicing)	DISCO
(a)	25	0.1017	0.1018	0.0999	0.1005	0.1009	0.1012
	50	0.1007	0.1047	0.0966	0.1021	0.1065	0.1045
	100	0.0984	0.0993	0.1006	0.0963	0.0959	0.0944
(b)	25	0.0988	0.0989	0.0948	0.0976	0.1000	0.0995
	50	0.0942	0.0967	0.1014	0.1000	0.0978	0.0961
	100	0.0953	0.0991	0.0981	0.1022	0.1001	0.1020
(c)	25	0.1027	0.1018	0.0999	0.0968	0.1017	0.0996
	50	0.1000	0.1005	0.0922	0.0977	0.0974	0.0973
	100	0.1036	0.1019	0.0959	0.1026	0.1025	0.1014
(d)	25	0.1008	0.0952	0.0962	0.1031	0.1018	0.0988
	50	0.0977	0.1004	0.0990	0.0961	0.1019	0.1042
	100	0.1019	0.0998	0.1000	0.1064	0.1024	0.0978

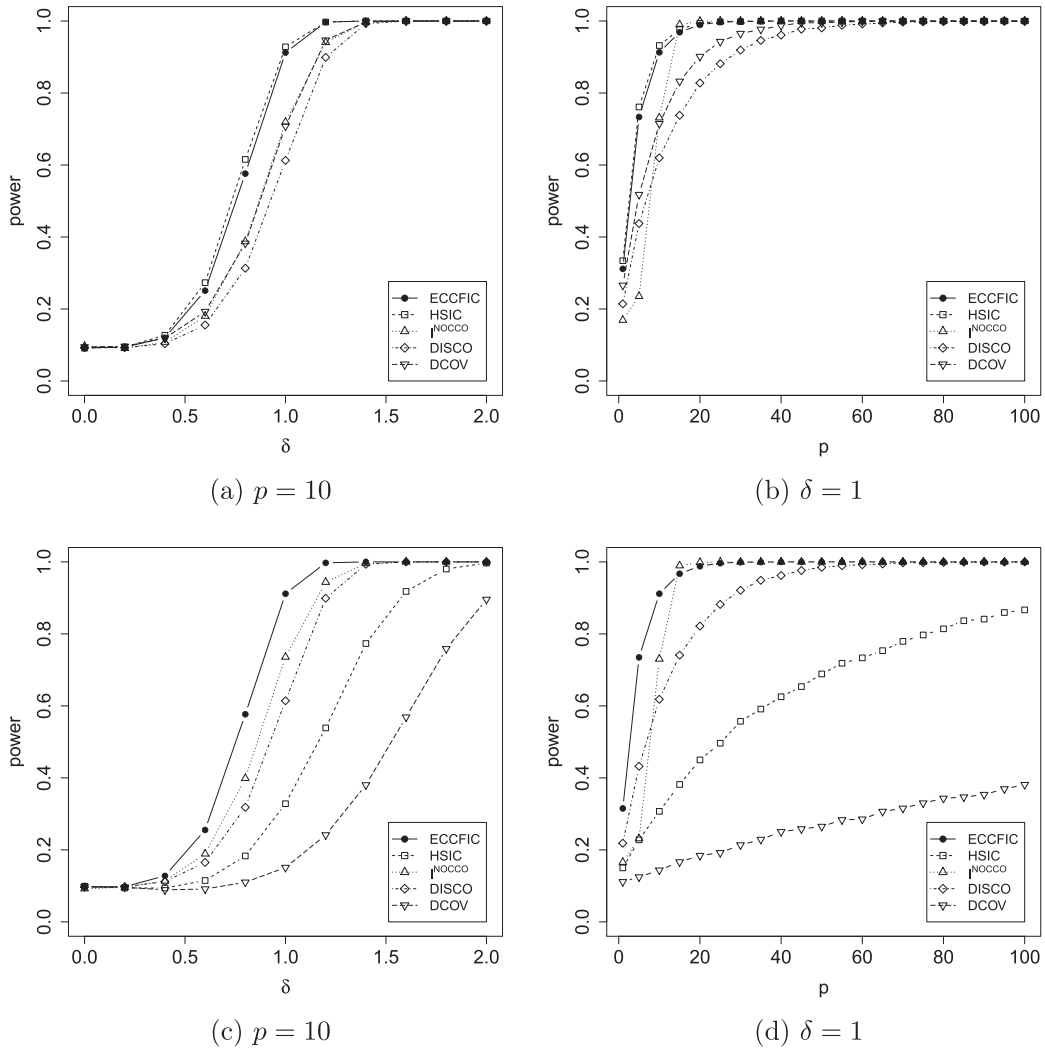

Figure 1. Example 2: Empirical power of the ECCFIC, HSIC, I^{NOCCO}, DISCO, and DCOV tests.

Example 3. This example aims to show the effect of different kernels on the performance of ECCFIC and HSIC. We imitate the Genome-Wide Association Studies (GWAS) example in Cui, Li, and Zhong (2015). In GWAS, we typically have genetic data containing a large number of single-nucleotide polymorphisms (SNPs). The SNPs are categorical predictors with three classes, denoted by {AA, Aa, aa}. We adopt a simple model with only two SNPs and denote Z_{ij} as the indicators of the dominant effect of the jth SNP for ith subject. Z_{ij} is generated in the following way $Z_{ij} = \begin{cases} 1, & \text{if } X_{ij} < q_1 \\ 0, & \text{if } q_1 \le X_{ij} < q_3 \\ -1, & \text{if } X_{ij} \ge q_3 \end{cases}$, where $\mathbf{X}_i = (X_{i1}, X_{i2}) \sim N(0, \Sigma)$, $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$, and q_1 and q_3 are the first and third quartiles of a standard normal distribution, respectively. Then we generate the response by $Y = \beta_1 Z_1 + \beta_2 Z_2 + \epsilon$, where $\beta_j = (-1)^U (a + |Z|)$ for $j = 1, 2$, $a = 2 \log n / \sqrt{n}$, $U \sim$Bernoulli(0.4) and $Z \sim N(0, 1)$. The error term $\epsilon \sim t(1)$, which is largely heavy-tailed. Monte Carlo power comparison of ECCFIC and HSIC with different kernels (Gaussian kernel and kernels induced by semi-metric $\rho(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_p^\alpha$ with $\alpha = 1/2, 1, 3/2$) are summarized in Figure 2, assuming a significance level 0.05. In general, ECCFIC is more powerful and less sensitive to the choice of kernel than HSIC.

Example 4. This example is to investigate the effect of number of slices on the performance of the ECCFIC slicing method and DISCO. The Saviotti aircraft data contain six variables of aircraft designs in the twentieth century (Bowman and Azzalini 1997). Two variables, wing span (m) and speed (km/h) for the 230 designs of the third (of three) periods are considered here. As discussed in Example 3 of Székely and Rizzo (2009), the nonlinear relation between speed and wing span is quite evident from the scatterplot and contour plot. Our goal is to test the independence of log(speed) and log(span).

We slice on log(span) to apply ECCFIC slicing method and DISCO. Results are listed in Table 2 with respect to different number of slices. Although our method is not very sensitive to the number of slices, we suggest that each slice should have at least 5 and at most $n/2$ data points.

Example 5. This example is to examine the performance of ECCFIC when both variables are univariate continuous.

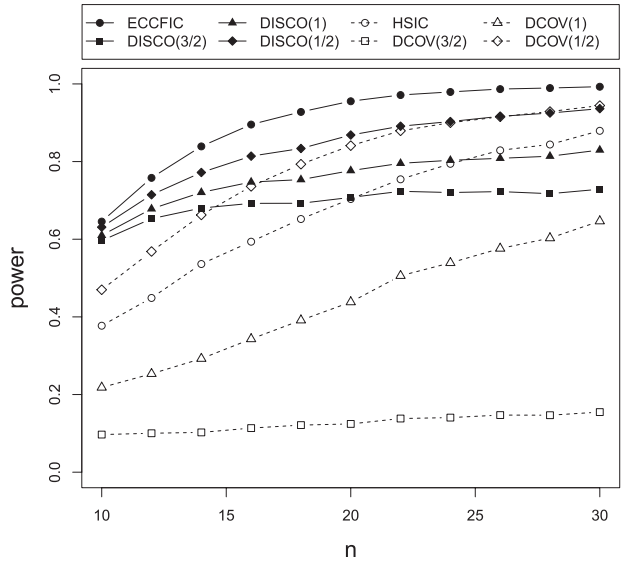

Figure 2. Example 3: Empirical power of ECCFIC, HSIC and DCOV tests.

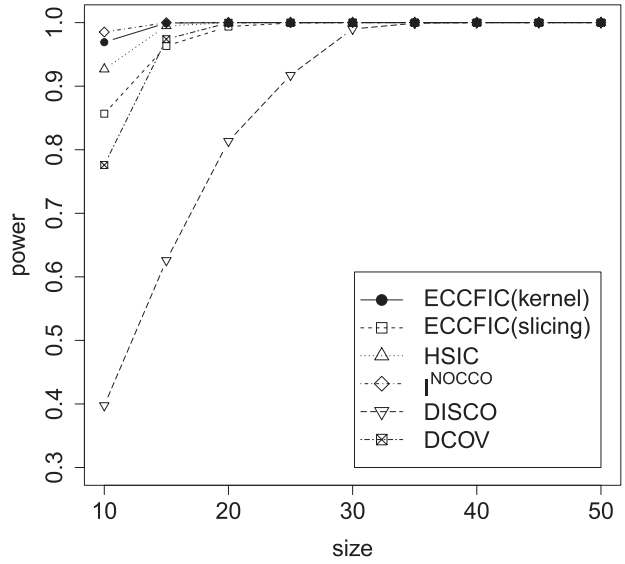

Figure 3. Example 5: Empirical power of ECCFIC, HSIC, I^{NOCCO}, DISCO, and DCOV tests.

Table 2. Example 4: *p*-values of ECCFIC and DISCO tests.

# of slices	2	5	10	23	46	115
ECCFIC	0.001	0.001	0.001	0.001	0.001	0.001
DISCO	0.005	0.006	0.001	0.002	0.002	0.004

Consider Example 2 in Székely and Rizzo (2009): $(X, Y) = (X, \varphi(X))$, where X is standard normal and $\varphi(\cdot)$ is the probability density function of standard normal. The relation between X and Y is deterministic but not monotone. Monte Carlo power comparisons are shown in Figure 3 for varied sample size n. To apply the slicing method and DISCO, we use 2, 3, and 4 slices when $n = 10, 15$ and 20, respectively. While for n greater than 20, we use 5 slices. Figure 3 reveals that the ECCFIC test with the kernel regression estimator has decent performance against the alternative, even with very small sample size. In addition, although slicing on $\mathbf{Y}$ is less preferred in continuous case, we note that the ECCFIC slicing method is still better than DISCO.

Example 6. This example (Székely, Rizzo, and Bakirov 2007) is to examine the power of ECCFIC when both variables are multivariate continuous. Suppose that the distribution of $\mathbf{X}$ is standard multivariate normal with dimension $p = 5$, and $Y_{kj} = X_{kj}\epsilon_{kj}, j = 1, \ldots, p$, where ϵ_{kj} are independent standard normal variables and independent of X. For multivariate continuous $\mathbf{Y}$, existing slicing techniques in other areas, for example in sufficient dimension reduction such as Zhu, Zhu and Feng (2010), Li, Wen and Zhu (2008), and Cook and Zhang (2014) can be very helpful. However, the kernel regression estimator is still more applicable and accurate. Thus, we only compare with HSIC and DCOV. Figure 4 indicates that ECCFIC with the kernel regression estimator works the best.

Example 7. This example evaluates the performance of ECCFIC in regression setups. Two models are generated: (A) $Y = (\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^2 + \epsilon$; (B) $Y = 0.2(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})^2\epsilon$. Let $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)^{\mathrm{T}}$ and $\epsilon \sim N(0, 1)$. Predictors are generated based on the following schemes: part (1), standard normal predictors $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I}_{10})$; part (2), nonnormal predictors; part (3), discrete predictors. We report the power for sample size $n = 10, 20$ and 50, respectively. To apply the slicing method and DISCO, we slice Y into 2, 4, and 5 levels for n=10, 20, and 50, respectively.

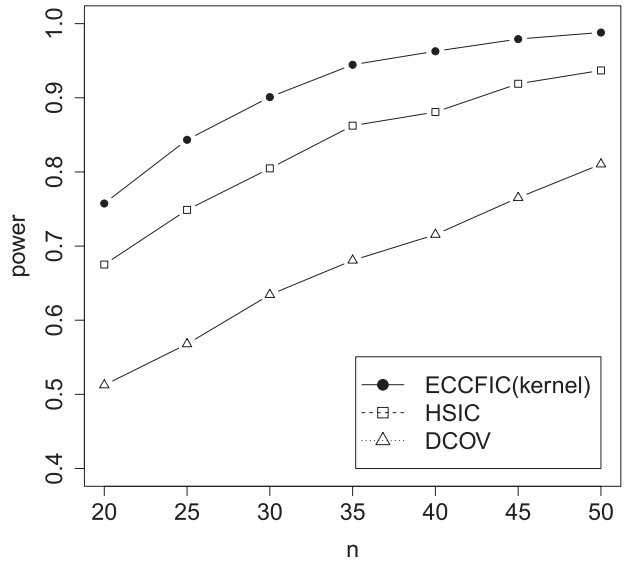

Figure 4. Example 6: Empirical power of ECCFIC, HSIC, and DCOV tests.

Model A. This is the first model in Sheng and Yin (2013), which has a nonlinear structure in the regression mean function. Predictors for part (2) and part (3) are generated as follows: part (2), $X_1 \sim N(-8, 4), X_2 \sim F(4, 10), X_3 \sim \chi^2(5), X_4 \sim t(15), X_5 \sim t(3), X_i \sim N(0, 1), i = 6, \ldots, 10$; part (3), $X_i \sim$ Poisson$(1), i = 1, \ldots, 5, X_i \sim N(0, 1), i = 6, \ldots, 10$.

Model B. This is the third model in Sheng and Yin (2013), which has a nonlinear structure in the regression variance function. Predictors for part (2) and part (3) are generated as follows: part (2), $X_1 \sim N(-8, 4), X_2 \sim t(5), X_3 \sim$ Gamma$(9, 0.5), X_4 \sim F(5, 12), X_5 \sim \chi^2(13), X_i \sim N(0, 1), i = 6, \ldots, 10$; part (3), $X_i \sim$ Poisson$(1), i = 1, \ldots, 5, X_i \sim N(0, 1), i = 6, \ldots, 10$.

Table 3. Example 7: Empirical power of ECCFIC, HSIC, I^{NOCCO}, DCOV, and DISCO tests.

Model	n	Method					
		ECCFIC (kernel)	HSIC	I^{NOCCO}	DCOV	ECCFIC (slicing)	DISCO
A(1)	10	0.2433	0.1775	0.1218	0.2013	0.1220	0.1169
	20	0.3852	0.2290	0.1149	0.2738	0.1462	0.1221
	50	0.6595	0.3480	0.1069	0.4006	0.2115	0.1599
A(2)	10	0.5690	0.4935	0.1760	0.5001	0.3310	0.2401
	20	0.8857	0.8008	0.1268	0.7768	0.5810	0.4019
	50	0.9989	0.9962	0.1037	0.9935	0.9621	0.7810
A(3)	10	0.6528	0.5877	0.1422	0.6945	0.4778	0.4355
	20	0.9433	0.9104	0.1203	0.9783	0.7363	0.7148
	50	1.0000	1.0000	0.1042	1.0000	0.9986	0.9984
B(1)	10	0.2193	0.1786	0.1224	0.2001	0.1062	0.1052
	20	0.3357	0.2173	0.1221	0.2578	0.1168	0.1110
	50	0.5790	0.3054	0.1171	0.3459	0.1459	0.1235
B(2)	10	0.3504	0.3087	0.1275	0.3059	0.1293	0.1376
	20	0.5870	0.5174	0.1171	0.5131	0.3036	0.3415
	50	0.9084	0.8740	0.1002	0.8816	0.6784	0.7645
B(3)	10	0.3548	0.3286	0.1431	0.3042	0.1260	0.1254
	20	0.5723	0.5304	0.1280	0.4739	0.2869	0.2757
	50	0.9131	0.9124	0.1014	0.8362	0.6859	0.6815

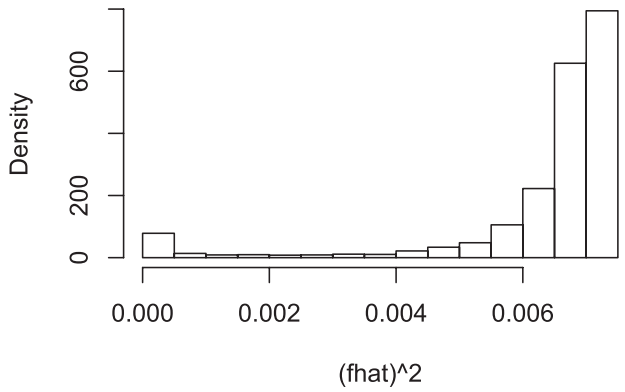

Figure 5. Example 8: histogram of $\widehat{f}_h^2(Y_t)$.

Results in Table 3 indicate that ECCFIC with the kernel regression estimator has the best power in all the models except A(3), in which case ECCFIC is second but very close to the best, DCOV.

Example 8. This example is to elaborate the use of the weight function $a(\cdot)$ in the kernel regression estimation of ECCFIC. Although we use $a(\cdot) \equiv 1$ in all the previous examples, our method is actually sensitive to the choice of the weight function, especially when extreme values exist. This is logically, similar to Ordinary Least Squares (OLS) vs. Weight Least Squares (WLS). Below we provide an example where the use of a weight function is appropriate. Suppose $Y = \frac{1}{|X|} + \epsilon$, where $X \sim Unif(-3, 3)$ and $\epsilon \sim N(0, 1)$. A histogram of $\widehat{f}_h^2(Y_t)$ values is provided in Figure 5, which shows a heavy tail near 0. Then we compare the power of four methods—our method with $a(\mathbf{Y}_t) \equiv f_h^2(\mathbf{Y}_t)$, our method with $a(\cdot) \equiv 1$, HSIC and DCOV based on 1,000 Monte Carlo simulations.

Table 4. Example 8: Empirical power of ECCFIC, HSIC, and DCOV tests.

n	ECCFIC $a(\mathbf{Y}_t) \equiv f_h^2(\mathbf{Y}_t)$	ECCFIC $a(\cdot) \equiv 1$	HSIC	DCOV
20	0.430	0.225	0.414	0.190
35	0.690	0.357	0.675	0.408
50	0.845	0.443	0.827	0.631

Table 5. Example 9: Empirical Type I error rate and power of ECCFCIC and Γ.

n	α	ECCFCIC	Γ
100	0.05	0.055	0.050
	0.10	0.100	0.095
200	0.05	0.035	0.070
	0.10	0.060	0.115

Model A, Type I error rate

n	α	ECCFCIC	Γ
100	0.05	0.240	0.160
	0.10	0.365	0.240
200	0.05	0.490	0.210
	0.10	0.595	0.385

Model B, power

As we can see from Table 4, the use of the weight function can improve the performance of our method when there are extremely small $\widehat{f}_h^2(\mathbf{Y}_t)$ values. We suggest to check the distribution of $\widehat{f}_h^2(\mathbf{Y}_t)$ before choosing the weight function, which however, could be somewhat subjective.

Example 9. This example is to compare ECCFCIC with the measure Γ of Su and White (2007), in terms of the Type I error rate and the power. Two time series models (Su and White 2007) are generated: (A) $Y_t = 0.3Y_{t-1} + \epsilon_t$, where $\epsilon_t \overset{\text{iid}}{\sim} N(0, 1)$; (B) $Y_t = e^{-Y_{t-1}^2} + |0.1Y_{t-2}(16 - Y_{t-2})|\epsilon_t$, where $\{\epsilon_t\}$ are iid sum of 10 uniformly independently distributed random variables each over the range $[-1/7, 1/7]$. We test the null hypothesis H_0 : $f(Y_t|Y_{t-1}, Y_{t-2}) = f(Y_t|Y_{t-1})$, that is, Y_{t-2} has no explanatory power for Y_t, which is true for Model A but not for Model B. The results based on 200 Monte Carlo simulations are listed in Table 5, which shows that our measure has a reasonable Type I error rate and is more powerful than Γ.

Example 10. These data focus on comparing our kernel ANOVA with the typical ANOVA and DCOV. The leaf dataset contains a collection of shape and texture features extracted from digital images of leaf specimens originating from a total of 30 different plant species (*http://archive.ics.uci.edu/ml/datasets/Leaf*, Silva, Marçal and da Silva 2013; Silva 2013). Specifically, the relation between elongation and species is studied. The typical ANOVA and the kernel ANOVA decompositions are listed in Table 6. Both tests indicate that elongation is a significant aspect to distinct different leaf species. However, a residual plot of the fitted ANOVA model in Figure 6 reveals nonconstant variance of the elongation measurements across the species.

Therefore, we further examine the assumption of ANOVA by testing the dependence between ANOVA residuals and species. Our method is compared with DCOV. Our kernel ANOVA test in Table 7 also suggests a violation of constant variance. From the above two kernel ANOVA tests, we can conclude that distributions of elongation are different between species in the second moment or higher. As for the DCOV test on ANOVA residuals, since species is a categorical variable, we consider both the original coding as well as a dummy coding

Table 6. Example 10: ANOVA and Kernel ANOVA.

		ANOVA				Kernel ANOVA			
Source	Df	Sum	Mean	F	p-value	Sum	Mean	$\mathcal{F}$	p-value
Species	29	11.9107	0.4107	120.46	<0.001	445.1185	15.3489	42.1277	0.001
Error	310	1.0569	0.0034			112.9461	0.3643		
Total	339	12.9676				558.0646			

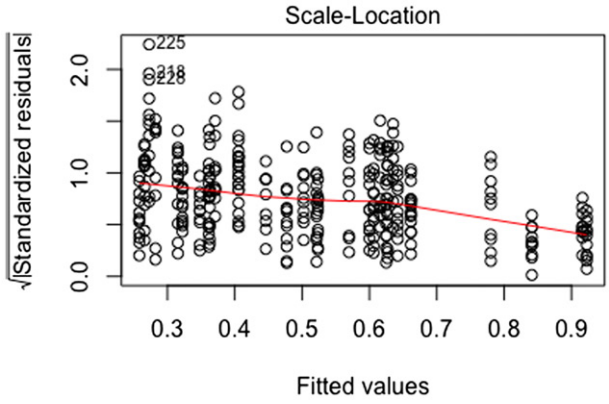

Figure 6. Example 10: Analysis of ANOVA residuals.

Table 7. Example 10: Kernel ANOVA and DCOV test on analysis of ANOVA residuals.

Kernel ANOVA					
Source	Df	Sum	Mean	$\mathcal{F}$	p-value
Species	29	315.8016	10.8897	1.6937	0.001
Error	310	1993.1540	6.4295		
Total	339	2308.9550			
DCOV					
$nV^2 = 0.0489, p\text{-value} = 0.326$					
DCOV (dummy coding)					
$nV^2 = 0.0932, p\text{-value} = 0.12$					

(see Cui, Li, and Zhong 2015), but neither of them detects the heteroscedasticity. Thus, our kernel ANOVA method is more powerful than DCOV.

7. Discussion

In this article, we proposed ECCFIC as a flexible and powerful measure for testing independence between two random vectors, which is especially useful when one of them is categorical. We provided two empirical estimators for the new measure and their associated asymptotic properties. Similar asymptotic results on non-iid samples may also be obtained by using U-statistic (Lee 1990) and those of Su and White (2007). Another direction of investigating asymptotic distributions is to let the dimension tend to infinity. Székely and Rizzo (2013) indicated that the sample DCOR tends to 1 as the dimension goes to infinity even when $\mathbf{X}$ and $\mathbf{Y}$ are independent. Therefore, they propose a modified DCOR statistic and under independence the distribution of a transformation of the statistic converges to a t-distribution as the dimension tends to infinity. Dueck et al. (2014) studies the limiting theorems of an affinely invariant version of DCOR assuming normal distributions. We can certainly follow these work to develop asymptotic results for our measure when the dimension tends to infinity. We can also further study the optimization over kernels and parameters. However, this is a very challenging problem, although discussion can be found in the literature (Fukumizu et al. 2009, Gretton et al. 2012b).

Acknowledgments

The authors thank the editor, the associate editor, and four referees for their insightful and constructive comments, which lead to a greatly improved version. This work is supported in part by an NSF grant CIF-1813330.

References

Bach, F. R., and Jordan, M. I. (2002a), "Kernel Independent Component Analysis," *Journal of Machine Learning Research*, 3, 1–48. [985]

Bach, F. R., and Jordan, M. I. (2002b), "Tree-dependent Component Analysis," in *Proceedings of the Eighteenth conference on Uncertainty in Artificial Intelligence*, eds. A. Darwiche, and N. Friedman, San Francisco, CA: Morgan Kaufmann Publishers Inc., pp. 36–44. [985]

Bowman, A. W., and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations* (Vol. 18), Oxford: OUP. [992]

Cook, R. D., and Zhang, X. (2014), "Fused Estimators of the Central Subspace in Sufficient Dimension Reduction," *Journal of the American Statistical Association*, 109, 815–827. [993]

Cui, H., Li, R., and Zhong, W. (2015), "Model-free Feature Screening for Ultrahigh Dimensional Discriminant Analysis," *Journal of the American Statistical Association*, 110, 630–641. [992,995]

Dauxois, J., and Nkiet, G. M. (1998), "Nonlinear Canonical Analysis and Independence Tests," *The Annals of Statistics*, 26, 1254–1278. [985]

Davison, A. C., and Hinkley, D. V. (1997), *Bootstrap Methods and Their Application* (Vol. 1), Cambridge: Cambridge University Press. [990]

Dueck, J., Edelmann, D., Gneiting, T., Richards, D. (2014), "The Affinely Invariant Distance Correlation," *Bernoulli*, 20(4), 2305–2330. [995]

Efron, B., and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Boca Raton, FL: CRC Press. [990]

Fukumizu, K., Bach, F. R., and Gretton, A. (2007), "Statistical Consistency of Kernel Canonical Correlation Analysis," *Journal of Machine Learning Research*, 8, 361–383. [985]

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004), "Dimensionality Reduction for Supervised Learning With Reproducing Kernel Hilbert Spaces," *Journal of Machine Learning Research*, 5, 73–99. [985]

Fukumizu, K., Gretton, A., Lanckriet, G. R., Schölkopf, B., and Sriperumbudur, B. K. (2009), "Kernel Choice and Classifiability for RKHS Embeddings of Probability Distributions," in *Advances in Neural Information Processing Systems 22*, eds. Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Red Hook, NY: Curran Associates, Inc., pp. 1750–1758. [988,991,995]

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008), "Kernel Measures of Conditional Dependence," in *Advances in Neural Information Processing Systems 20*, eds. J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Red Hook, NY: Curran Associates, Inc., pp. 489–496. [985,991]

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a), "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, 13, 723–773. [987]

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005b), "Measuring Statistical Dependence with Hilbert–Schmidt Norms," in *Algorithmic Learning Theory*, eds. S. Jain, H. U. Simon, and E. Tomita, Berlin, Heidelberg: Springer, pp. 63–77. [985]

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2008), "A Kernel Statistical Test of Independence," in *Advances in*

Neural Information Processing Systems, eds. J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Red Hook, NY: Curran Associates, Inc., pp. 585–592. [985,991]

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005a), "Kernel Methods for Measuring Independence," *Journal of Machine Learning Research*, 6, 2075–2129. [985]

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b), Optimal Kernel Choice for Large-scale Two-sample Tests," in *Advances in Neural Information Processing Systems 25*, eds. F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Red Hook, NY: Curran Associates, Inc., pp. 1205–1213. [991,995]

Gretton, A., Smola, A. J., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. K. (2005), Kernel Constrained Covariance for Dependence Measurement," in *AISTATS* (Vol. 10), pp. 112–119. [985]

Li, B., Wen, S., and Zhu, L. (2008), "On a Projective Resampling Method for Dimension Reduction with Multivariate Responses," *Journal of the American Statistical Association*, 103, 1177–1186. [993]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening Via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [985]

Rizzo, M. L., and Székely, G. J. (2010), "Disco Analysis: A Nonparametric Extension of Analysis of Variance," *The Annals of Applied Statistics*, 4, 1034–1055. [986,990,991]

Robinson, P. M. (1988), "Root-N-consistent Semiparametric Regression," *Econometrica*, 56(4), 931–954. [990]

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013), "Equivalence of Distance-based and RKHS-based Statistics in Hypothesis Testing," *The Annals of Statistics*, 41, 2263–2291. [985,987,988]

Sheng, W., and Yin, X. (2013), "Direction Estimation in Single-index Models Via Distance Covariance," *Journal of Multivariate Analysis*, 122, 148–161. [985,993]

Sheng, W., and Yin, X. (2016), "Sufficient Dimension Reduction Via Distance Covariance," *Journal of Computational and Graphical Statistics*, 25, 91–104. [985]

Silva, P. F. B. (2013), "Development of a System for Automatic Plant Species Recognition," Master Thesis, University of Porto, Mathematics Department. [994]

Silva, P. F. B., Marçal, A. R. S., and da Silva, R. M. A. (2013), "Evaluation of Features for Leaf Discrimination", in *Image Analysis and Recognition*, eds. M. Kamel, and A. Campilho, Berlin, Heidelberg: Springer, pp. 197–204. [994]

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis* (Vol. 26), Boca Raton: CRC press. [991]

Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007), "A Hilbert Space Embedding for Distributions," in *Algorithmic Learning Theory*, eds. M. Hutter, R. A. Servedio, and E. Takimoto, Berlin, Heidelberg: Springer, pp. 13–31. [985]

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. (2008), "Injective Hilbert Space Embeddings of Probability Measures," in *Proceedings of the 21st Annual Conference on Learning Theory*, eds. R. Servedio, and T. Zhang, Madison, WI: Omnipress, pp. 111–122. [988]

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010), "Hilbert Space Embeddings and Metrics on Probability Measures," *Journal of Machine Learning Research*, 11, 1517–1561. [988]

Su, L., and White, H. (2003), "Testing Conditional Independence Via Empirical Likelihood," *Department of Economics, UCSD*. [985]

——— (2007), "A Consistent Characteristic Function-based Test for Conditional Independence," *Journal of Econometrics*, 141, 807–834. [985,990,991,994,995]

——— (2008), "A Nonparametric Hellinger Metric Test for Conditional Independence," *Econometric Theory*, 24, 829–864. [985]

Sun, X., Janzing, D., Schölkopf, B., and Fukumizu, K. (2007), "A Kernel-based Causal Learning Algorithm," in *Proceedings of the 24th International Conference on Machine Learning*, ed. Z. Ghahramani, New York, NY: ACM, pp. 855–862. [985]

Székely, G. J., and Bakirov, N. K. (2003), "Extremal Probabilities for Gaussian Quadratic Forms," *Probability Theory and Related Fields*, 126, 184–202. [989]

Székely, G. J., and Rizzo, M. L. (2009), "Brownian Distance Covariance," *The Annals of Applied Statistics*, 3, 1236–1265. [990,992,993]

——— (2013), "The Distance Correlation T-test of Independence in High Dimension," *Journal of Multivariate Analysis*, 117, 193–213. [995]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [985,990,991,993]

Wang, X., Jiang, B., and Liu, J. S. (2017), "Generalized R-squared for Detecting Dependence," *Biometrika*, 104, 129–139. [985]

Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015), "Conditional Distance Correlation," *Journal of the American Statistical Association*, 110, 1726–1734. [985,990,991]

Wendland, H. (2004), *Scattered Data Approximation* (Vol. 17), Cambridge: Cambridge University Press. [986]

Yamanishi, Y., Vert, J. -P., and Kanehisa, M. (2004), "Heterogeneous Data Comparison and Gene Selection with Kernel Canonical Correlation Analysis," in *Kernel Methods in Computational Biology*, eds. B. Schölkopf, K. Tsuda and J. -P. Vert, Cambridge, MA: MIT Press, pp. 209–229. [985]

Yin, X., and Yuan, Q. (2019), "A New Class of Measures for Testing Independence," *Statistica Sinica*, [online], DOI: 10.5705/ss.202017.0538. Available at *http://www3.stat.sinica.edu.tw/ss_newpaper/SS-2017-0538_na.pdf*. [986,987,988]

Zhu, L. -P., Zhu, L. -X., and Feng, Z. -H. (2010), "Dimension Reduction in Regressions Through Cumulative Slicing Estimation," *Journal of the American Statistical Association*, 105, 1455–1466. [993]