On a Nonparametric Notion of Residual and its Applications

Rohit Kumar Patra[a,*], Bodhisattva Sen[a,**], Gábor Székely[b]

[a] *Department of Statistics, Columbia University, New York, NY 10027*
[b] *National Science Foundation, Arlington, VA 22230*

Abstract

Let $(X, \mathbf{Z})$ be a continuous random vector in $\mathbb{R} \times \mathbb{R}^d$, $d \geq 1$. In this paper, we define the notion of a nonparametric residual of X on $\mathbf{Z}$ that is always independent of the predictor $\mathbf{Z}$. We study its properties and show that the proposed notion of residual matches with the usual residual (error) in a multivariate normal regression model. Given a random vector $(X, Y, \mathbf{Z})$ in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$, we use this notion of residual to show that the conditional independence between X and Y, given $\mathbf{Z}$, is equivalent to the mutual independence of the residuals (of X on $\mathbf{Z}$ and Y on $\mathbf{Z}$) and $\mathbf{Z}$. This result is used to develop a test for conditional independence. We propose a bootstrap scheme to approximate the critical value of this test. We compare the proposed test, which is easily implementable, with some of the existing procedures through a simulation study.

Keywords: Bootstrap; conditional distribution function; energy statistic; one sample multivariate goodness-of-fit test; partial copula; testing conditional independence.

1. Introduction

Let $(X, \mathbf{Z})$ be a random vector in $\mathbb{R} \times \mathbb{R}^d = \mathbb{R}^{d+1}$, $d \geq 1$. We assume that $(X, \mathbf{Z})$ has a joint density on $\mathbb{R}^{d+1}$. If we want to predict X using $\mathbf{Z}$ we usually formulate the following regression problem:

$$X = m(\mathbf{Z}) + \epsilon, \tag{1.1}$$

where $m(\mathbf{z}) = \mathbb{E}(X|\mathbf{Z} = \mathbf{z})$ is the conditional mean of X given $\mathbf{Z} = \mathbf{z}$ and $\epsilon := X - m(\mathbf{Z})$ is the *residual* (although ϵ is usually called the error, and its estimate the residual, for this paper we feel that the term residual is more appropriate). Typically we further assume that the residual ϵ is *independent* of

*Corresponding author
**Supported by NSF CAREER Grant DMS-1150435.
Email addresses: rohit@stat.columbia.edu (Rohit Kumar Patra), bodhi@stat.columbia.edu (Bodhisattva Sen), gszekely@nsf.gov (Gábor Székely)

$\mathbf{Z}$. However, intuitively, we are just trying to break the information in $(X, \mathbf{Z})$ into two parts: a part that contains all relevant information about X, and the "residual" (the left over) which does not have anything to do with the relationship between X and $\mathbf{Z}$.

In this paper we address the following question: given any random vector $(X, \mathbf{Z})$ how do we define the notion of a "residual" of X on $\mathbf{Z}$ that matches with the above intuition? Thus, formally, we want to find a function $\varphi : \mathbb{R}^{d+1} \to \mathbb{R}$ such that the residual $\varphi(X, \mathbf{Z})$ satisfies the following two conditions:

(C.1) the residual $\varphi(X, \mathbf{Z})$ is independent of the predictor $\mathbf{Z}$, i.e.,

$$\varphi(X, \mathbf{Z}) \perp\!\!\!\perp \mathbf{Z}, \qquad \text{and}$$

(C.2) the information content of $(X, \mathbf{Z})$ is the same as that of $(\varphi(X, \mathbf{Z}), \mathbf{Z})$, i.e.,

$$\sigma(X, \mathbf{Z}) = \sigma(\varphi(X, \mathbf{Z}), \mathbf{Z}), \tag{1.2}$$

where $\sigma(X, \mathbf{Z})$ denotes the σ-field generated by X and $\mathbf{Z}$. We can also express (1.2) as: there exists a measurable function $h : \mathbb{R}^{d+1} \to \mathbb{R}$ such that

$$X = h(\mathbf{Z}, \varphi(X, \mathbf{Z})); \tag{1.3}$$

see e.g., Theorem 20.1 of Billingsley (1995).

In this paper we propose a notion of a residual that satisfies the above two conditions, under any joint distribution of X and $\mathbf{Z}$. We investigate the properties of this notion of residual in Section 2. We show that this notion indeed reduces to the usual residual (error) in the multivariate normal regression model. Further, we use this notion of residual to develop a test for conditional independence.

Suppose now that $(X, Y, \mathbf{Z})$ has a joint density on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d = \mathbb{R}^{d+2}$. The assumption of conditional independence means that X is independent of Y given $\mathbf{Z}$, i.e., $X \perp\!\!\!\perp Y|\mathbf{Z}$. Conditional independence is an important concept in modeling causal relations (Dawid (1979), Pearl (2000)), in graphical models (Lauritzen (1996); Koller and Friedman (2009)), in economic theory (see Chiappori and Salanié (2000)), and in the literature of program evaluations (see Heckman et al. (1997)) among other fields. Traditional methods for testing conditional independence are either restricted to the discrete case (Lauritzen (1996); Agresti (2013)) or impose simplifying assumption when the random variables are continuous (Lawrance (1976)). However, recently there has been a few nonparametric testing procedures proposed for testing conditional independence without assuming a functional form between the distributions of X, Y, and $\mathbf{Z}$. Su and White (2007) consider testing conditional independence based on the difference between the conditional characteristic functions, while Su and White (2008) use the Hellinger distance between conditional densities of X given Y and $\mathbf{Z}$, and X given Y to test for conditional independence. A test based on estimation of the maximal nonlinear conditional correlation is proposed in Huang (2010). Bergsma (2011) develops a test based on partial copula. Fukumizu et al. (2007)

propose a measure of conditional dependence of random variables, based on normalized cross-covariance operators on reproducing kernel Hilbert spaces; Zhang et al. (2012) propose another kernel-based conditional independence test. Poczos and Schneider (2012) extend the concept of distance correlation (developed by Székely et al. (2007) to measure dependence between two random variables or vectors) to characterize conditional dependence. Székely and Rizzo (2014) investigate a method that is easy to compute and can capture non-linear dependencies but does not completely characterize conditional independence; also see Györfi and Walk (2012) and the references therein.

In Section 3 we use the notion of residual defined in Section 2 to show that the conditional independence between X and Y given $\mathbf{Z}$ is equivalent to the mutual independence of three random vectors: the residuals of X on $\mathbf{Z}$ and Y on $\mathbf{Z}$, and $\mathbf{Z}$. We reduce this testing of mutual independence to a one sample multivariate goodness-of-fit test. We further propose a modification of the easy-to-implement *energy* statistic based method (Székely and Rizzo (2005); also see Székely and Rizzo (2013)) to test the goodness-of-fit; see Section 3.1. In Section 3.2 we use our notion of nonparametric residual and the proposed goodness-of-fit test to check the null hypothesis of conditional independence. Moreover, we describe a bootstrap scheme to approximate the critical value of this test. In Section 4 we compare the finite sample performance of the procedure proposed in this paper with other available methods in the literature through a finite sample simulation study. We end with a brief discussion, Section 5, where we point to some open research problems and outline an idea, using the proposed residuals, to define (and test) a nonparametric notion of partial correlation.

2. A nonparametric notion of residual

Conditions (C.1)–(C.2) do not necessarily lead to a unique choice for φ. To find a meaningful and unique function φ that satisfies conditions (C.1)–(C.2) we impose the following natural restrictions on φ. We assume that

(C.3) $\quad x \mapsto \varphi(x, \mathbf{z})$ is strictly increasing in its support, for every fixed $\mathbf{z} \in \mathbb{R}^d$.

Note that condition (C.3) is a slight strengthening of condition (C.2). Suppose that a function φ satisfies conditions (C.1) and (C.3). Then any strictly monotone transformation of $\varphi(\cdot, \mathbf{z})$ would again satisfy (C.1) and (C.3). Thus, conditions (C.1) and (C.3) do not uniquely specify φ. To handle this identifiability issue, we replace condition (C.1) with (C.4), described below.

First observe that, by condition (C.1), the conditional distribution of the random variable $\varphi(X, \mathbf{Z})$ given $\mathbf{Z} = \mathbf{z}$ does not depend on $\mathbf{z}$. We assume that

(C.4) $\quad \varphi(X, \mathbf{Z})|\mathbf{Z} = \mathbf{z}$ is uniformly distributed, for all $\mathbf{z} \in \mathbb{R}^d$.

Condition (C.4) is again quite natural – we usually assume that the residual has a fixed distribution, e.g., in regression we assume that the (standardized) residual in normally distributed with zero mean and unit variance. Note that condition (C.4) is slightly stronger than (C.1) and will help us uniquely identify

φ. The following result shows that, indeed, under conditions (C.3)–(C.4), a unique φ exists and gives its form.

Lemma 2.1. *Let $F_{X|\mathbf{Z}}(\cdot|\mathbf{z})$ denote the conditional distribution function of $X|\mathbf{Z} = \mathbf{z}$. Under conditions (C.3) and (C.4), we have a unique choice of $\varphi(x, \mathbf{z})$, given by*

$$\varphi(x, \mathbf{z}) = F_{X|\mathbf{Z}}(x|\mathbf{z}).$$

Also, $h(\mathbf{z}, u)$ can be taken as

$$h(\mathbf{z}, u) = F^{-1}_{X|\mathbf{Z}}(u|\mathbf{z}). \tag{2.1}$$

Proof. Fix $\mathbf{z}$ in the support of $\mathbf{Z}$. Let $u \in (0, 1)$. Let us write $\varphi_{\mathbf{z}}(x) = \varphi(x, \mathbf{z})$. By condition (C.4), we have $\mathbb{P}[\varphi(X, \mathbf{Z}) \le u|\mathbf{Z} = \mathbf{z}] = u$. On the other hand, by (C.3),

$$\mathbb{P}[\varphi(X, \mathbf{Z}) \le u|\mathbf{Z} = \mathbf{z}] = \mathbb{P}[X \le \varphi^{-1}_{\mathbf{z}}(u)|\mathbf{Z} = \mathbf{z}] = F_{X|\mathbf{Z}}(\varphi^{-1}_{\mathbf{z}}(u)|\mathbf{z}).$$

Thus, we have

$$F_{X|\mathbf{Z}}(\varphi^{-1}_{\mathbf{z}}(u)|\mathbf{z}) = u, \quad \text{for all } u \in (0, 1),$$

which is equivalent to $\varphi_{\mathbf{z}}(x) = F_{X|\mathbf{Z}}(x|\mathbf{z})$.

Let h be as defined in (2.1). Then,

$$h(\mathbf{z}, \varphi(x, \mathbf{z})) = F^{-1}_{X|\mathbf{Z}}(\varphi(x, \mathbf{z})|\mathbf{z}) = F^{-1}_{X|\mathbf{Z}}(F_{X|\mathbf{Z}}(x|\mathbf{z})|\mathbf{z}) = x,$$

as required. □

Thus from the above lemma, we conclude that in the nonparametric setup, if we want to have a notion of a residual satisfying conditions (C.3)–(C.4) then the residual has to be $F_{X|\mathbf{Z}}(X|\mathbf{Z})$. The following remarks are in order now.

Remark 2.2. Let us first consider the example when $(X, \mathbf{Z})$ follows a multivariate Gaussian distribution, i.e.,

$$\begin{pmatrix} X \\ \mathbf{Z} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \Sigma := \begin{pmatrix} \sigma_{11} & \boldsymbol{\sigma}_{12}^\top \\ \boldsymbol{\sigma}_{12} & \Sigma_{22} \end{pmatrix} \right),$$

where $\mu_1 \in \mathbb{R}$, $\mu_2 \in \mathbb{R}^d$, Σ is a $(d+1) \times (d+1)$ positive definite matrix with $\sigma_{11} > 0$, $\sigma_{12} \in \mathbb{R}^{d\times 1}$ and $\Sigma_{22} \in \mathbb{R}^{d\times d}$.

Then the conditional distribution of X given $\mathbf{Z} = \mathbf{z}$ is $N(\mu_1 + \boldsymbol{\sigma}_{12}^\top \Sigma_{22}^{-1}(\mathbf{z} - \boldsymbol{\mu}_2), \sigma_{11} - \boldsymbol{\sigma}_{12}^\top \Sigma_{22}^{-1} \boldsymbol{\sigma}_{12})$. Therefore, we have the following representation in the form of (1.1):

$$X = \mu_1 + \boldsymbol{\sigma}_{12}^\top \Sigma_{22}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_2) + \left(X - \mu_1 - \boldsymbol{\sigma}_{12}^\top \Sigma_{22}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_2) \right)$$

where the usual residual is $X - \mu_1 - \boldsymbol{\sigma}_{12}^\top \Sigma_{22}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_2)$, which is known to be independent of $\mathbf{Z}$. In this case, using Lemma 2.1, we get

$$\varphi(X, \mathbf{Z}) = \Phi\left(\frac{X - \mu_1 - \boldsymbol{\sigma}_{12}^\top \Sigma_{22}^{-1}(\mathbf{Z} - \boldsymbol{\mu}_2)}{\sqrt{\sigma_{11} - \boldsymbol{\sigma}_{12}^\top \Sigma_{22}^{-1} \boldsymbol{\sigma}_{12}}} \right),$$

where $\Phi(\cdot)$ is the distribution function of the standard normal distribution. Thus $\varphi(X, \mathbf{Z})$ is just a fixed strictly increasing transformation of the usual residual, and the two notions of residual essentially coincide.

Remark 2.3. The above notion of residual does not extend so easily to the case of discrete random variables. Conditions (C.1) and (C.2) are equivalent to the fact that $\sigma(X, \mathbf{Z})$ factorizes into two sub σ-fields as $\sigma(X, \mathbf{Z}) = \sigma(\varphi(X, \mathbf{Z})) \otimes \sigma(\mathbf{Z})$. This may not be always possible as can be seen from the following simple example.

Let (X, Z) take values in $\{0, 1\}^2$ such that $\mathbb{P}[X = i, Z = j] > 0$ for all $i, j \in \{0, 1\}$. Then it can be shown that such a factorization exists if and only if X and Z are independent, in which case $\varphi(X, Z) = X$.

Remark 2.4. Lemma 2.1 also gives an way to generate X, using $\mathbf{Z}$ and the residual. We can first generate $\mathbf{Z}$, following its marginal distribution, and an independent random variable $U \sim \mathcal{U}(0, 1)$ (here $\mathcal{U}(0, 1)$ denotes the Uniform distribution on $(0, 1)$) which will act as the residual. Then (1.3), where h is defined in (2.1), shows that we can generate $X = F^{-1}_{X|\mathbf{Z}}(U|\mathbf{Z})$.

In practice, we need to estimate the residual $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ from observed data, which can be done both parametrically and non-parametrically. If we have a parametric model for $F_{X|\mathbf{Z}}(\cdot|\cdot)$, we can estimate the parameters, using e.g., maximum likelihood, etc. If we do not want to assume any structure on $F_{X|\mathbf{Z}}(\cdot|\cdot)$, we can use any nonparametric smoothing method, e.g., standard kernel methods, for estimation; see Bergsma (2011) for such an implementation. We will discuss the estimation of the residuals in more detail in Section 3.3.

3. Conditional independence

Suppose now that $(X, Y, \mathbf{Z})$ has a joint density on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d = \mathbb{R}^{d+2}$. In this section we state a simple result that reduces testing for the conditional independence hypothesis $H_0 : X \perp\!\!\!\perp Y|\mathbf{Z}$ to a problem of testing mutual independence between three random variables/vectors that involve our notion of residual. We also briefly describe a procedure to test the mutual independence of the three random variables/vectors (see Section 3.1). We start with the statement of the crucial lemma.

Lemma 3.1. *Suppose that $(X, Y, \mathbf{Z})$ has a continuous joint density on $\mathbb{R}^{d+2}$. Then, $X \perp\!\!\!\perp Y|\mathbf{Z}$ if and only if $F_{X|\mathbf{Z}}(X|\mathbf{Z}), F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ and $\mathbf{Z}$ are mutually independent.*

Proof. Let us make the following change of variable

$$(X, Y, \mathbf{Z}) \mapsto (U, V, \mathbf{Z}) := (F_{X|\mathbf{Z}}(X), F_{Y|\mathbf{Z}}(Y), \mathbf{Z}).$$

The joint density of $(U, V, \mathbf{Z})$ can be expressed as

$$f_{(U,V,\mathbf{Z})}(u, v, \mathbf{z}) = \frac{f(x, y, \mathbf{z})}{f_{X|\mathbf{Z}=\mathbf{z}}(x) f_{Y|\mathbf{Z}=\mathbf{z}}(y)} = \frac{f_{(X,Y)|\mathbf{Z}=\mathbf{z}}(x, y) f_{\mathbf{Z}}(\mathbf{z})}{f_{X|\mathbf{Z}=\mathbf{z}}(x) f_{Y|\mathbf{Z}=\mathbf{z}}(y)}, \tag{3.1}$$

where $x = F^{-1}_{X|\mathbf{Z}=\mathbf{z}}(u)$, and $y = F^{-1}_{Y|\mathbf{Z}=\mathbf{z}}(v)$. Note that as the Jacobian matrix is upper-triangular, the determinant is the product of the diagonal entries of the matrix, namely, $f_{X|\mathbf{Z}=\mathbf{z}}(x)$, $f_{Y|\mathbf{Z}=\mathbf{z}}(y)$ and 1.

If $X \perp\!\!\!\perp Y|\mathbf{Z}$ then $f_{(U,V,\mathbf{Z})}(u, v, \mathbf{z})$ reduces to just $f_{\mathbf{Z}}(\mathbf{z})$, for $u, v \in (0, 1)$, from the definition of conditional independence, which shows that $U, V, \mathbf{Z}$ are independent (note that it is easy to show that U, V are marginally $\mathcal{U}(0, 1)$, the Uniform distribution on $(0, 1)$). Now, given that $U, V, \mathbf{Z}$ are independent, we know that $f_{(U,V,\mathbf{Z})}(u, v, \mathbf{z}) = f_{\mathbf{Z}}(\mathbf{z})$ for $u, v \in (0, 1)$, which from (3.1) easily shows that $X \perp\!\!\!\perp Y|\mathbf{Z}$. □

Remark 3.2. Note that the joint distribution of $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ is known as the *partial copula*; see e.g., Bergsma (2011). Bergsma (2011) developed a test for conditional independence by testing mutual independence between $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$. However, as the following example illustrates, the independence of $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ is not enough to guarantee that $X \perp\!\!\!\perp Y|\mathbf{Z}$. Let W_1, W_2, W_3 be i.i.d. $\mathcal{U}(0, 1)$ random variables. Let $X = W_1 + W_3$, $Y = W_2$ and $Z = \text{mod}(W_1 + W_2, 1)$, where 'mod' stands for the modulo (sometimes called modulus) operation that finds the remainder of the division $W_1 + W_2$ by 1. Clearly, the random vector (X, Y, Z) has a smooth continuous density on $[0, 1]^3$. Note that Z is independent of W_i, for $i = 1, 2$. Hence, X, Y and Z are pairwise independent. Thus, $F_{X|\mathbf{Z}}(X) = F_X(X)$ and $F_{Y|\mathbf{Z}}(X) = F_Y(Y)$, where F_X and F_Y are the marginal distribution functions of X and Y, respectively. From the independence of X and Y, $F_X(X)$ and $F_Y(Y)$ are independent. On the other hand, the value of W_1 is clearly determined by Y and Z, i.e., $W_1 = Z - Y$ if $Y \leq Z$ and $W_1 = Z - Y + 1$ if $Y > Z$. Consequently, X and Y are not conditionally independent given Z. To see this, note that for every $z \in (0, 1)$,

$$\mathbb{E}[X|Y, Z = z] = \begin{cases} z - Y + 0.5 & \text{if } Y \leq z \\ z - Y + 1 + 0.5 & \text{if } Y > z, \end{cases}$$

which obviously depends on Y. In Remark 3.6 we illustrate this behavior with a finite sample simulation study.

Remark 3.3. We can extend the above result to the case when X and Y are random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. In that case we define the conditional multivariate distribution transform $F_{X|\mathbf{Z}}$ by successively conditioning on the co-ordinate random variables, i.e., if $X = (X_1, X_2)$ then we can define $F_{X|\mathbf{Z}}$ as $(F_{X_2|X_1,\mathbf{Z}}, F_{X_1|\mathbf{Z}})$. With this definition, Lemma 3.1 still holds.

To use Lemma 3.1 to test the conditional independence between X and Y given $\mathbf{Z}$, we need to first estimate the residuals $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ from observed data, which can be done by any nonparametric smoothing procedure, e.g., standard kernel methods (see Section 3.3). Then, any procedure for testing the mutual independence of $F_{X|\mathbf{Z}}(X|\mathbf{Z}), F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ and $\mathbf{Z}$ can be used. In this paper we advocate the use of the *energy* statistic (see Rizzo and Székely (2010)), described briefly in the next subsection, to test the mutual independence of three or more random variables/vectors.

3.1. Testing mutual independence of three or more random vectors with known marginals

Testing independence of two random variables (or vectors) has received much recent attention in the statistical literature; see e.g., Székely et al. (2007), Gretton et al. (2005), and the references therein. However, testing the mutual independence of three or more random variables is more complicated and we could not find any easily implementable method in the statistical literature.

In this sub-section, we test the mutual independence of three or more random variables (vectors) with known marginals by converting the problem to a one-sample goodness-of-fit test for multivariate normality. In the following we briefly describe our procedure in the general setup.

Suppose that we have $r \geq 3$ continuous random variables (or vectors) $V_1, \ldots, V_r$ and we want to test their mutual independence. We assume that we know the marginal distributions of $V_1, \ldots, V_r$; without loss of generality, we can assume that V_i's are standard Gaussian random variables (vectors). We write $T := (V_1, V_2, \ldots, V_r) \in \mathbb{R}^k$ and introduce $T_{\text{ind}} := (V_1^*, V_2^*, \ldots, V_r^*)$ where V_j^* is an i.i.d. copy of V_j, $j = 1, 2, \ldots, r$, but in T_{ind} the coordinates, $V_1^*, V_2^*, \ldots, V_r^*$, are independent. To test the mutual independence of $V_1, V_2, \ldots, V_r$ all we need to test now is whether T and T_{ind} are identically distributed. If we observed a sample from T, we can test for the equality of distributions of T and T_{ind} through a one-sample goodness-of-fit test for the standard multivariate normal distribution, i.e.,

$$H_0 : T \sim N(\mathbf{0}, \mathbf{I}_{k\times k}),$$

as $T_{\text{ind}} \sim N(\mathbf{0}, \mathbf{I}_{k\times k})$, where $\mathbf{I}_{k\times k}$ is the identity matrix of order k and $\mathbf{0} := (0, \ldots, 0) \in \mathbb{R}^k$.

In this paper we consider the following *energy* statistic (see Székely and Rizzo (2005) and Rizzo and Székely (2010))

$$\Lambda(T) = 2\mathbb{E}\|T - T_{\text{ind}}\| - \mathbb{E}\|T - T'\| - \mathbb{E}\|T_{\text{ind}} - T'_{\text{ind}}\|, \tag{3.2}$$

where T' and T'_{ind} are i.i.d. copies of T and T_{ind}, respectively ($\|\cdot\|$ denotes the Euclidean norm). Note that $\Lambda(T)$ is always nonnegative, and equals 0, if and only if T and T_{ind} are identically distributed, i.e., if and only if $V_1, V_2, \ldots, V_r$ are mutually independent (see Corollary 1 of Székely and Rizzo (2005)).

Suppose now that we observe n i.i.d. samples $T_1, \ldots, T_n$ of T. The (scaled) sample version of the energy statistic for testing the goodness-of-fit hypothesis

is

$$\mathcal{E}_n(T_1,\ldots,T_n) := 2\sum_{i=1}^{n}\mathbb{E}\|T_i - T_{\text{ind}}\| - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\|T_i - T_j\| - n\mathbb{E}\|T_{\text{ind}} - T'_{\text{ind}}\|. \tag{3.3}$$

Note that the first expectation in the above display is with respect to T_{ind}. Under the null hypothesis of mutual independence, the test statistic $\mathcal{E}_n(T_1,\ldots,T_n)$ has a limiting distribution, as $n\to\infty$, while under the alternative hypothesis $\mathcal{E}_n(T_1,\ldots,T_n)$ tends to infinity; see Section 4 of Székely and Rizzo (2005) and Section 8 of Székely and Rizzo (2013) for detailed discussions. Thus any test that rejects the null for large values of $\mathcal{E}_n(T_1,\ldots,T_n)$ is consistent against general alternatives.

As T_{ind} and T'_{ind} are i.i.d. $N(\mathbf{0},\mathbf{I}_{k\times k})$ random variables. The statistic $\mathcal{E}_n(T_1,\ldots,T_n)$ is easy to compute:

$$\mathbb{E}\|T_{\text{ind}} - T'_{\text{ind}}\| = \sqrt{2}\mathbb{E}\|T_{ind}\| = 2\frac{\Gamma\left(\frac{d+3}{2}\right)}{\Gamma\left(\frac{d+2}{2}\right)}$$

and for any $a\in\mathbb{R}^{d+2}$, we have

$$\mathbb{E}\|a - T_{\text{ind}}\| = \frac{\sqrt{2}\Gamma\left(\frac{d+3}{2}\right)}{\Gamma\left(\frac{d+2}{2}\right)} + \sqrt{\frac{2}{\pi}}\sum_{k=0}^{\infty}\frac{(-1)^k}{k!\,2^k}\frac{|a|^{2k+2}}{(2k+1)(2k+2)}\frac{\Gamma\left(\frac{d+3}{2}\right)\Gamma\left(k+\frac{3}{2}\right)}{\Gamma\left(k+\frac{d}{2}+2\right)}.$$

The expression for $\mathbb{E}\|a - T_{\text{ind}}\|$ follows from the discussion in Zacks (1981) (see page 55). See the source code "energy.c" in the *energy* package of R language (R Development Core Team (2008)) for a fast implementation of this; also see Székely and Rizzo (2013).

3.2. Testing conditional independence

In this sub-section we use Lemma 3.1 and the test for mutual independence proposed in the previous sub-section (Section 3.1) to test for the conditional independence of X and Y given $\mathbf{Z}$. We start with a simple lemma.

Lemma 3.4. *Suppose that $(X,Y,\mathbf{Z})$ has a continuous joint density on $\mathbb{R}^{d+2}$. Then $X \perp\!\!\!\perp Y|\mathbf{Z}$ if and only if*

$$W := (F_{X|\mathbf{Z}}(X|\mathbf{Z}), F_{Y|\mathbf{Z}}(Y|\mathbf{Z}), F_{\mathbf{Z}}(\mathbf{Z})) \sim \mathcal{U}([0,1]^{d+2}),$$

where $F_{\mathbf{Z}}(\mathbf{z}) = \left(F_{Z_d|Z_{d-1},\ldots,Z_1}(z_d|z_{d-1},\ldots,z_1),\ldots,F_{Z_2|Z_1}(z_2|z_1),F_{Z_1}(z_1)\right)$, $\mathbf{Z} = (Z_1,\ldots,Z_d)$, $\boldsymbol{z} = (z_1,\ldots,z_d)$, and $\mathcal{U}([0,1]^{d+2})$ denote the Uniform distribution on $[0,1]^{d+2}$.

Proof. Note that by Lemma 3.1, $X \perp\!\!\!\perp Y|\mathbf{Z}$ if and only if $F_{X|\mathbf{Z}}(X|\mathbf{Z})$, $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ and $\mathbf{Z}$ are mutually independent. Furthermore, note that $F_{X|\mathbf{Z}}(X|\mathbf{Z})$, $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ are i.i.d. $\mathcal{U}(0,1)$ random variables. Thus the proof of the lemma will be complete if we show that $F_{\mathbf{Z}}(\mathbf{Z}) \sim \mathcal{U}([0,1]^d)$.

As each of $F_{Z_d|Z_{d-1},\ldots,Z_1}(Z_d|Z_{d-1},\ldots,Z_1),\ldots,F_{Z_2|Z_1}(Z_2|Z_1)$, and $F_{Z_1}(Z_1)$ are $\mathcal{U}(0,1)$ random variables, it is enough to show that they are mutually independent. For simplicity of notation, we will only prove the independence of $F_{Z_2|Z_1}(Z_2|Z_1)$ and $F_{Z_1}(Z_1)$, independence of other terms can be proved similarly. Note that

$$\begin{aligned}\mathbb{P}(F_{Z_2|Z_1}(Z_2|Z_1)\leq z_2|F_{Z_1}(Z_1)=z_1) &= \mathbb{P}(F_{Z_2|Z_1}(Z_2|Z_1)\leq z_2|Z_1=F_{Z_1}^{-1}(z_1))\\ &= \mathbb{P}\Big(Z_2\leq F_{Z_2|Z_1}^{-1}\big(z_2|F_{Z_1}^{-1}(z_1)\big)\Big|Z_1=F_{Z_1}^{-1}(z_1)\Big)\\ &= F_{Z_2|Z_1}\Big(F_{Z_2|Z_1}^{-1}\big(z_2|F_{Z_1}^{-1}(z_1)\big)|F_{Z_1}^{-1}(z_1)\Big)\\ &= z_2.\end{aligned}$$

As the conditional distribution of $F_{Z_2|Z_1}(Z_2|Z_1)$ given $F_{Z_1}(Z_1)=z_1$ does not depend on z_1, we have that $F_{Z_2|Z_1}(Z_2|Z_1)$ and $F_{Z_1}(Z_1)$ are independent. □

Let us now assume $X\perp\!\!\!\perp Y|\mathbf{Z}$ and define

$$W:=\big(F_{X|\mathbf{Z}}(X|\mathbf{Z}),F_{Y|\mathbf{Z}}(Y|\mathbf{Z}),F_{Z_d|\mathbf{Z}_{-d}}(Z_d|\mathbf{Z}_{-d}),\ldots,F_{Z_2|Z_1}(Z_2|Z_1),F_{Z_1}(Z_1)\big).$$

By Lemma 3.4, we have

$$W\overset{\mathcal{D}}{=}(U_1,\ldots,U_{d+2}),$$

where $U_1,U_2,\ldots,U_{d+2}$ are i.i.d. $\mathcal{U}(0,1)$ random variables. An equivalent formulation is

$$H_0: T:=\Phi^{-1}(W)\overset{\mathcal{D}}{=}N(\mathbf{0},\mathbf{I}_{(d+2)\times(d+2)}), \tag{3.4}$$

where Φ is the distribution function corresponding to the standard Gaussian random variable, and for any $\mathbf{a}\in\mathbb{R}^{d+2}$, $\Phi^{-1}(\mathbf{a}):=(\Phi^{-1}(a_1),\ldots,\Phi^{-1}(a_{d+2}))$.

We observe i.i.d. data $\{(X_i,Y_i,\mathbf{Z}_i):i=1,\ldots,n\}$ from the joint distribution of $(X,Y,\mathbf{Z})$ and we are interested in testing $X\perp\!\!\!\perp Y|\mathbf{Z}$. Suppose first that the distribution functions $F_{X|\mathbf{Z}}(\cdot|\cdot),F_{Y|\mathbf{Z}}(\cdot|\cdot)$, and $F_{\mathbf{Z}}(\cdot)$ are known. Then we have an i.i.d. sample $T_1,\ldots,T_n$ from T, where

$$T_i:=\Phi^{-1}(F_{X|\mathbf{Z}}(X_i|\mathbf{Z}_i),F_{Y|\mathbf{Z}}(Y_i|\mathbf{Z}_i),F_{\mathbf{Z}}(\mathbf{Z}_i)). \tag{3.5}$$

Now we can use the the test statistic (3.3) to test the hypothesis of conditional independence.

As the true conditional distribution functions $F_{X|\mathbf{Z}},F_{Y|\mathbf{Z}}$, and $F_{\mathbf{Z}}$ are unknown, we can replace them by their estimates $\widehat{F}_{X|\mathbf{Z}},\widehat{F}_{Y|\mathbf{Z}}$, and $\widehat{F}_{\mathbf{Z}}$, respectively, where $\widehat{F}_{\mathbf{Z}}(\mathbf{z})=\Big(\widehat{F}_{Z_d|Z_{d-1},\ldots,Z_1}(z_d|z_{d-1},\ldots,z_1),\ldots,\widehat{F}_{Z_2|Z_1}(z_2|z_1),\widehat{F}_{Z_1}(z_1)\Big)$; see Section 3.3 for more details on how to compute these estimates. Let us now define

$$\widehat{T}_i:=\Phi^{-1}(\widehat{F}_{X|\mathbf{Z}}(X_i|\mathbf{Z}_i),\widehat{F}_{Y|\mathbf{Z}}(Y_i|\mathbf{Z}_i),\widehat{F}_{\mathbf{Z}}(\mathbf{Z}_i)), \tag{3.6}$$

for $i=1,2,\ldots,n$. We will use

$$\widehat{\mathcal{E}_n}:=\mathcal{E}_n(\hat{T}_1,\ldots\hat{T}_n) \tag{3.7}$$

to test the hypothesis of conditional independence.

3.2.1. Approximating the asymptotic distribution through bootstrap

The limiting behavior of $\mathcal{E}_n$ is not very useful in computing the critical value of the test statistic $\widehat{\mathcal{E}_n}$ proposed in the the previous sub-section. In a related but slightly different problem studied in Sen and Sen (2014), it was shown that, the analogous versions of $\mathcal{E}_n$ and $\widehat{\mathcal{E}_n}$ have very different limiting distributions.

In independence testing problems it is quite standard and natural to approximate the critical value of the test, under H_0, by using a permutation test; see e.g., Székely and Rizzo (2009), Gretton et al. (2007). However, in our problem as we use $\hat{T}_i$ instead of T_i, the permutation test is not valid; see Sen and Sen (2014).

In this sub-section, we propose a bootstrap procedure to approximate the distribution of $\widehat{\mathcal{E}_n}$, under the null hypothesis of conditional independence. We now describe the bootstrap procedure. Let $\mathbb{P}_{n,\mathbf{Z}}$ be the empirical distribution of $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$.

Step 1: Generate an i.i.d. sample $\{U^*_{i,1}, U^*_{i,2}, \mathbf{Z}^*_{n,i}\}_{1\le i\le n}$ of size n from the measure $\mathcal{U}(0,1)\times\mathcal{U}(0,1)\times\mathbb{P}_{n,\mathbf{Z}}$; recall that $\mathcal{U}(0,1)$ denotes the Uniform distribution on $(0,1)$.

Step 2: The bootstrap sample is then $\{X^*_{n,1}, Y^*_{n,1}, \mathbf{Z}^*_{n,1}\}_{1\le i\le n}$, where

$$X^*_{n,i} := \widehat{F}^{-1}_{X|Z}(U^*_{i,1}|\mathbf{Z}^*_{n,1}) \qquad \text{and} \qquad Y^*_{n,i} := \widehat{F}^{-1}_{Y|Z}(U^*_{i,2}|\mathbf{Z}^*_{n,1}). \tag{3.8}$$

Step 3: Use the bootstrap sample $\{X^*_{n,i}, Y^*_{n,i}, \mathbf{Z}^*_{n,i}\}_{1\le i\le n}$ to get smooth estimators $\widehat{F}^*_{X|\mathbf{Z}}, \widehat{F}^*_{Y|\mathbf{Z}}$, and $\widehat{F}^*_{\mathbf{Z}}$ of $F_{X|\mathbf{Z}}, F_{Y|\mathbf{Z}}$, and $F_{\mathbf{Z}}$; see Section 3.3 for a discussion on smooth estimation of the conditional distribution functions.

Step 4: Compute the bootstrap test statistic $\mathcal{E}^*_n := \mathcal{E}_n(\widehat{T}^*_1, \ldots, \widehat{T}^*_n)$ where

$$\widehat{T}^*_i = \Phi^{-1}\big(\widehat{F}^*_{X|\mathbf{Z}}(X^*_{n,i}|\mathbf{Z}^*_{n,i}), \widehat{F}^*_{Y|\mathbf{Z}}(Y^*_{n,i}|\mathbf{Z}^*_{n,i}), \widehat{F}^*_{\mathbf{Z}}(\mathbf{Z}^*_{n,i})\big). \tag{3.9}$$

We can now approximate the distribution of $\widehat{\mathcal{E}_n}$ by the conditional distribution of $\mathcal{E}^*_n$ given the data $\{X_i, Y_i, \mathbf{Z}_i\}_{1\le i\le n}$. In Section 4 we study the finite sample performance of the above procedure through a simulation study and illustrate that our procedure indeed yields a valid test for conditional independence.

Remark 3.5. In steps 1 and 2 above, we generate the bootstrap sample from the approximated joint distribution of $(X, Y, \mathbf{Z})$ under the null hypothesis of conditional independence. In steps 3 and 4 we mimic the evaluation of the test statistic $\widehat{\mathcal{E}_n}$ using the bootstrap sample. This is an example of a model based bootstrap procedure. Sen and Sen (2014) prove the consistency of a similar bootstrap procedure in a related problem. As the sample size increases the approximated joint distribution of $(X, Y, \mathbf{Z})$ (under H_0) would converge to the truth and the bootstrap distribution would replicate the distribution of $\widehat{\mathcal{E}_n}$.

3.3. Nonparametric estimation of the residuals

In this sub-section we discuss procedures to nonparametrically estimate $F_{X|\mathbf{Z}}, F_{Y|\mathbf{Z}}$, and $F_{\mathbf{Z}}$ given data $\{X_i, Y_i, \mathbf{Z}_i\}_{1\le i\le n}$. The nonparametric estimation of the conditional distribution functions would involve smoothing. In the following we briefly describe the standard approach to estimating the conditional distribution functions using kernel smoothing techniques (also see Lee et al. (2006), Yu and Jones (1998), and Hall et al. (1999)). For notational simplicity, we restrict to the case $d = 1$, i.e., $\mathbf{Z}$ is a real-valued random variable. Given an i.i.d. sample of $\{(X_i, Z_i) : i = 1, \ldots, n\}$ from $f_{X,Z}$, the joint density of (X, Z), we can use the following kernel density estimator of $f_{X,Z}$:

$$\widehat{f}_n(x,z) = \frac{1}{nh_{1,n}h_{2,n}} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h_{1,n}}\right) k\left(\frac{z - Z_i}{h_{2,n}}\right)$$

where k is a symmetric probability density on $\mathbb{R}$ (e.g., the standard normal density function), and $h_{i,n}, i = 1, 2$, are the smoothing bandwidths. It can be shown that if $nh_{1,n}h_{2,n} \to \infty$ and $\max\{h_{1,n}, h_{2,n}\} \to 0$, as $n \to \infty$, then $\widehat{f}_n(x,z) \xrightarrow{P} f_{X,Z}(x,z)$. In fact, the theoretical properties of the above kernel density estimator are very well studied; see e.g., Fan and Gijbels (1996) and Einmahl and Mason (2005) and the references therein. For the convenience of notation, we will write $h_{i,n}$ as h_i, $i = 1, 2$.

The conditional density of X given Z can then be estimated by

$$\widehat{f}_{X|Z}(x|z) = \frac{\widehat{f}_n(x,z)}{\widehat{f}_Z(z)} = \frac{\frac{1}{nh_1h_2}\sum_{i=1}^{n} k\left(\frac{x-X_i}{h_1}\right) k\left(\frac{z-Z_i}{h_2}\right)}{\frac{1}{nh_2}\sum_{i=1}^{n} k\left(\frac{z-Z_i}{h_2}\right)}.$$

Thus the conditional distribution function of X given Z can be estimated as

$$\widehat{F}_{X|Z}(x|z) = \frac{\int_{-\infty}^{x} \widehat{f}_n(t,z)\,dt}{\widehat{f}_Z(z)} = \frac{\frac{1}{nh_2}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_1}\right) k\left(\frac{z-Z_i}{h_2}\right)}{\frac{1}{nh_2}\sum_{i=1}^{n} k\left(\frac{z-Z_i}{h_2}\right)} = \sum_{i=1}^{n} w_i(z) K\left(\frac{x - X_i}{h_1}\right)$$

where K is the distribution function corresponding to k (i.e., $K(u) = \int_{-\infty}^{u} k(v)\,dv$) and $w_i(z) = \frac{\frac{1}{nh_2}k\left(\frac{z-Z_i}{h_2}\right)}{\frac{1}{nh_2}\sum_{j=1}^{n} k\left(\frac{z-Z_j}{h_2}\right)}$ are weights that sum to one for every z. Least square cross-validation method proposed in Hall et al. (2004) can be used to find the optimal choices for h_1 and h_2. For general d, the optimal parameters must satisfy $h_1 \sim n^{-2/(d+4)}$ and $h_2 \sim n^{-1/(d+4)}$; see Section 6.2 of Li and Racine (2007) and Li et al. (2013) for a thorough discussion.

Remark 3.6. Now we provide empirical evidence for the failure of the test proposed in Bergsma (2011) in the example discussed in Remark 3.2. We plot (see Figure 1) the histogram of p-values obtained from the proposed test (see Section 3.2) and that of the p-values obtained from testing the independence of $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ (using their estimates $\widehat{F}_{X|\mathbf{Z}}(\cdot|\cdot)$ and $\widehat{F}_{Y|\mathbf{Z}}(\cdot|\cdot)$). We

use the distance covariance test statistic (see Székely et al. (2007)) to test for the independence of $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$. Figure 1 demonstrates that a test for mutual independence of $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ can fail to capture the conditional dependence between X and Y given $\mathbf{Z}$.

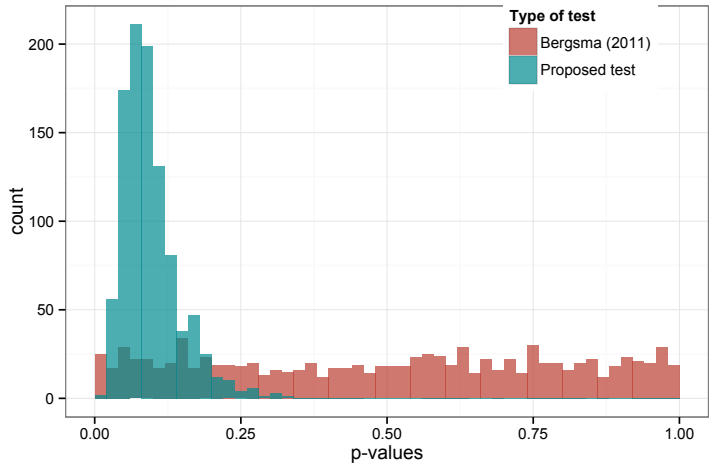

Figure 1: Histograms of p-values (estimated using 1000 bootstrap samples) over 1000 independent replications. Here, for $i = 1, \ldots, 200$, $\{X_i, Y_i, Z_i\}$ are i.i.d. samples from the example discussed in Remark 3.2.

4. Simulation

We now investigate the finite sample performance of the testing procedure developed in this paper through a simulation study. We also compare the performance of the our testing procedure to those proposed in Fukumizu et al. (2007) and Zhang et al. (2012). We denote the the testing procedure proposed in Fukumizu et al. (2007) by CI_{perm} and use KCI to denote the kernel based conditional independence test proposed in Zhang et al. (2012).

To illustrate and compare the performance of different testing procedures, we consider the following sampling scenario borrowed from Zhang et al. (2012). Let us assume that X and Y are only dependent on Z_1 (the first coordinate of $\mathbf{Z}$) and that all other conditioning variables are independent of X, Y, and Z_1. We assume that $\mathbf{Z} \sim N_d(\mathbf{0}, \sigma_z^2 \mathbf{I}_{d\times d})$, $X := W + Z_1 + \epsilon$, and $Y := W + Z_1 + \epsilon'$, where ϵ, ϵ', and W are three independent mean zero Gaussian random variables. Moreover, we assume that ϵ, ϵ', and W are independent of $\mathbf{Z}$, $var(\epsilon) = var(\epsilon') = \sigma_E^2$, and $var(W) = \sigma_W^2$, where for any real random variable V, $var(V)$ denotes its variance. Note that $X \perp\!\!\!\perp Y|\mathbf{Z}$ if and only if $\sigma_W = 0$.

In our finite sample simulations we fixed $\sigma_E = 0.3$ and $\sigma_z = 0.2$. We generate 500 i.i.d. samples $\{X_i, Y_i, \mathbf{Z}_i\}_{1\le i\le 500}$ for each of $d = 1, 3$, and 5 and for different values of σ_W. For each such sample, we use 1000 bootstrap replicates

to estimate the p-value of the proposed test procedure. We have used the "np" (see Hayfield and Racine (2008)) package in R (R Core Team (2015)) to estimate the conditional distribution functions with the tuning parameters chosen using least-squares cross validation (see Section 3.3). In Figure 2 we plot the power (estimated using 500 independent experiments) of the testing procedure proposed in Section 3.2 along with those of CI_{perm} and KCI as σ_W increases from 0 to 0.25, for dimensions $1, 3$, and 5. We fix the significance level at 0.05.

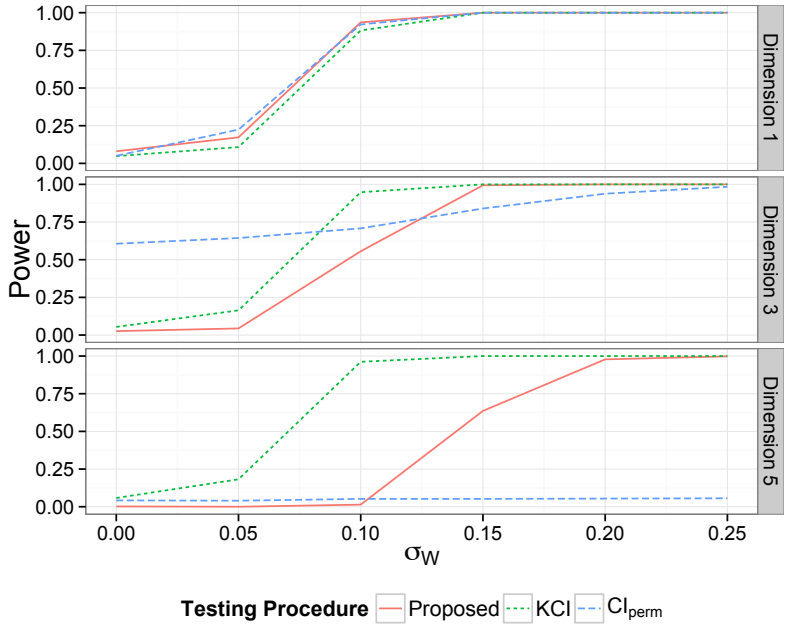

Figure 2: The power (at significance level 0.05) of the three testing procedures for sample size $n = 500$ as the dimension d and σ_W increase.

The distribution of the KCI test statistic under the null hypothesis of conditional independence is estimated with a Monte Carlo procedure suggested in Zhang et al. (2012). To implement the CI_{perm} and the KCI testing procedures, we have used the MATLAB source codes provided in Zhang et al. (2012); the source code can be found at `http://people.tuebingen.mpg.de/kzhang/KCI-test.zip`. The R language codes used to implement our procedure are available at `http://stat.columbia.edu/~rohit/research.html`.

Observe that for CI_{perm}, the probability of type I error is much greater than the significance level for $d = 3$. Furthermore, for $d = 5$, it fails to detect the alternative for all values of σ_W. The performance of CI_{perm} is sensitive to the dimension of the conditioning variable. The probability of type I error for both the proposed and the KCI testing procedures are around the specified

significance level. Moreover, the powers of KCI and the proposed test increase to 1 as σ_W increases. Overall, we think that for this simulation scenario the KCI method has the best performance.

5. Discussion

Given a random vector $(X, \mathbf{Z})$ in $\mathbb{R} \times \mathbb{R}^d = \mathbb{R}^{d+1}$ we have defined the notion of a nonparametric residual of X on $\mathbf{Z}$ as $F_{X|\mathbf{Z}}(X|\mathbf{Z})$, which is always independent of the response $\mathbf{Z}$. We have studied the properties of the nonparametric residual and showed that it indeed reduces to the usual residual in a multivariate normal regression model. However, nonparametric estimation of $F_{X|\mathbf{Z}}(\cdot|\cdot)$ requires smoothing techniques, and hence suffers from the curse of dimensionality. A natural way of mitigating this curse of dimensionality could be to use dimension reduction techniques in estimating the residual $F_{X|\mathbf{Z}}(X|\mathbf{Z})$. Another alternative would be to use a parametric model for the conditional distribution function.

Suppose now that $(X, Y, \mathbf{Z})$ has a joint density on $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d = \mathbb{R}^{d+2}$. We have used this notion of residual to show that the conditional independence between X and Y, given $\mathbf{Z}$, is equivalent to the mutual independence of the residuals $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ and the predictor $\mathbf{Z}$. We have used this result to propose a test for conditional independence, based on the energy statistic.

We can also use these residuals to come up with a nonparametric notion of partial correlation. The partial correlation of X and Y measures the degree of association between X and Y, removing the effect of $\mathbf{Z}$. In the nonparametric setting, this reduces to measuring the dependence between the residuals $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$. We can use distance covariance (Székely et al. (2007)), or any other measure of dependence, for this purpose. We can also test for zero partial correlation by testing for the independence of the residuals $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$.

Acknowledgements: The second author would like to thank Arnab Sen for many helpful discussions, and for his help in writing parts of the paper. He would also like to thank Probal Chaudhuri for motivating the problem. The research of second and third authors is supported by National Science Foundation.

References

Agresti, A., 2013. Categorical data analysis, 3rd Edition. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Bergsma, W. P., 2011. Nonparametric testing of conditional independence by means of the partial copula.

Billingsley, P., 1995. Probability and measure, 3rd Edition. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, a Wiley-Interscience Publication.

Chiappori, P., Salanié, B., 2000. Testing for asymmetric information in insurance markets. Journal of Political Economy 108 (1), 56–78.

Dawid, A. P., 1979. Conditional independence in statistical theory. J. Roy. Statist. Soc. Ser. B 41 (1), 1–31.
URL http://links.jstor.org.ezproxy.cul.columbia.edu/sici?sici=0035-9246(1979)41:1<1:CIIST>2.0.CO;2-T&origin=MSN

Einmahl, U., Mason, D. M., 2005. Uniform in bandwidth consistency of kernel-type function estimators. Ann. Statist. 33 (3), 1380–1403.
URL http://dx.doi.org/10.1214/009053605000000129

Fan, J., Gijbels, I., 1996. Local polynomial modelling and its applications. Vol. 66 of Monographs on Statistics and Applied Probability. Chapman & Hall, London.

Fukumizu, K., Gretton, A., Sun, X., Schölkopf, B., 2007. Kernel measures of conditional dependence. In: Advances in Neural Information Processing Systems. pp. 489–496.

Gretton, A., Bousquet, O., Smola, A., Schölkopf, B., 2005. Measuring statistical dependence with hilbert-schmidt norms. Proceedings of the Conference on Algorithmic Learning Theory (ALT), 63–77.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., Smola, A. J., 2007. A kernel statistical test of independence. In: Advances in Neural Information Processing Systems. pp. 585–592.

Györfi, L., Walk, H., 2012. Strongly consistent nonparametric tests of conditional independence. Statist. Probab. Lett. 82 (6), 1145–1150.
URL http://dx.doi.org/10.1016/j.spl.2012.02.023

Hall, P., Racine, J., Li, Q., 2004. Cross-validation and the estimation of conditional probability densities. Journal of the American Statistical Association 99 (468).

Hall, P., Wolff, R. C. L., Yao, Q., 1999. Methods for estimating a conditional distribution function. J. Amer. Statist. Assoc. 94 (445), 154–163.

Hayfield, T., Racine, J. S., 2008. Nonparametric econometrics: The np package. Journal of Statistical Software 27 (5).
URL http://www.jstatsoft.org/v27/i05/

Heckman, J., Ichimura, H., Todd, P., 1997. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. The Review of Economic Studies 64 (4), 605.

Huang, T.-M., 2010. Testing conditional independence using maximal nonlinear conditional correlation. Ann. Statist. 38 (4), 2047–2091.

Koller, D., Friedman, N., 2009. Probabilistic graphical models. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, principles and techniques.

Lauritzen, S. L., 1996. Graphical models. Vol. 17 of Oxford Statistical Science Series. The Clarendon Press Oxford University Press, oxford Science Publications.

Lawrance, A., 1976. On conditional and partial correlation. The American Statistician 30 (3), 146–149.

Lee, Y. K., Lee, E. R., Park, B. U., 2006. Conditional quantile estimation by local logistic regression. J. Nonparametr. Stat. 18 (4-6), 357–373.

Li, Q., Lin, J., Racine, J. S., 2013. Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. J. Bus. Econom. Statist. 31 (1), 57–65.
URL http://dx.doi.org/10.1080/07350015.2012.738955

Li, Q., Racine, J. S., 2007. Nonparametric econometrics. Princeton University Press, Princeton, NJ, theory and practice.

Pearl, J., 2000. Causality. Cambridge University Press, Cambridge, models, reasoning, and inference.

Poczos, B., Schneider, J., 2012. Conditional distance variance and correlation.

R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org

Rizzo, M. L., Székely, G. J., 2010. DISCO analysis: a nonparametric extension of analysis of variance. Ann. Appl. Stat. 4 (2), 1034–1055.

Sen, A., Sen, B., 2014. Testing independence and goodness-of-fit in linear models. Biometrika 101 (4), 927–942.

Su, L., White, H., 2007. A consistent characteristic function-based test for conditional independence. J. Econometrics 141 (2), 807–834.

Su, L., White, H., 2008. A nonparametric hellinger metric test for conditional independence. Econometric Theory 24 (04), 829–864.

Székely, G. J., Rizzo, M. L., 2005. A new test for multivariate normality. J. Mult. Anal. 93 (1), 58–80.

Székely, G. J., Rizzo, M. L., 2009. Brownian distance covariance. Ann. Appl. Stat. 3 (4), 1236–1265.
URL http://dx.doi.org/10.1214/09-AOAS312

Székely, G. J., Rizzo, M. L., 2013. Energy statistics: A class of statistics based on distances. Journal of Statistical Planning and Inference 143 (8), 1249–1272.

Székely, G. J., Rizzo, M. L., 2014. Partial distance correlation with methods for dissimilarities, to appear in *Ann. Statist.*

Székely, G. J., Rizzo, M. L., Bakirov, N. K., 2007. Measuring and testing dependence by correlation of distances. Ann. Statist. 35 (6), 2769–2794.

Yu, K., Jones, M. C., 1998. Local linear quantile regression. J. Amer. Statist. Assoc. 93 (441), 228–237.

Zacks, S., 1981. Parametric statistical inference. Vol. 4 of International Series in Nonlinear Mathematics: Theory, Methods and Applications. Pergamon Press, Oxford-New York, basic theory and modern approaches.

Zhang, K., Peters, J., Janzing, D., Schölkopf, B., 2012. Kernel-based conditional independence test and application in causal discovery. arXiv preprint arXiv:1202.3775.