On a Nonparametric Notion of Residual and its Applications

Rohit Kumar Patra^{a,*}, Bodhisattva Sen^{a,**}, Gábor Székely^b

^aDepartment of Statistics, Columbia University, New York, NY 10027

^bNational Science Foundation, Arlington, VA 22230

Abstract

Given a random vector $(X, \mathbf{Z})$, we define a notion of nonparametric residual of X on $\mathbf{Z}$ that is always independent of $\mathbf{Z}$. Given $(X, Y, \mathbf{Z})$, we use this notion of residual to develop a test for the conditional independence between X and Y , given $\mathbf{Z}$.

Keywords: Conditional distribution function; 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:

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

5 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 $\mathbf{Z}$. However, intuitively, we are just trying to break the information in $(X, \mathbf{Z})$
10 into two parts: a part that contains all relevant information about X , and

*Corresponding author

**Supported by NSF CAREER Grant DMS-1150435.

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
15 the above intuition? Thus, formally, we want to find a function $\varphi : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ 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., $\varphi(X, \mathbf{Z}) \perp\!\!\!\perp \mathbf{Z}$.

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

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

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

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

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
25 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 $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d = \mathbb{R}^{d+2}$.
30 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 \(1996\)](#)), and in economic theory (see [Chiappori and Salanié \(2000\)](#)) among other fields. Traditional methods for testing conditional independence
35 are either restricted to the discrete case or impose simplifying assumption when the random variables are continuous. 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 \(2008\)](#) 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. [Póczos and Schneider \(2012\)](#) extend the concept of distance correlation (developed by [Székely et al. \(2007\)](#) to measure dependence between two random vectors) to characterize conditional dependence; also see [Székely and Rizzo \(2014\)](#) and [Györfi and Walk \(2012\)](#) and the references therein.

In Section 3 we use the notion of residual defined in Section 2 to develop a test for the conditional independence between X and Y given $\mathbf{Z}$. We first 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 the 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. We end with a brief discussion in Section 4 where we point to some open research problems. In the accompanying online supplementary material, we give the proofs of the results stated in the paper and compare the finite sample performance of the proposed procedure with other available methods in the literature.

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

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

Note that condition (C.3) is a 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

$$(C.4) \quad \varphi(X, \mathbf{Z}) | \mathbf{Z} = \mathbf{z} \sim \mathcal{U}(0, 1) \text{ for all } \mathbf{z} \in \mathbb{R}^d,$$

where $\mathcal{U}(0, 1)$ denotes the uniform distribution on $(0, 1)$. 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 is 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 (proved in Section 9 of the online supplementary material) 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, $h(\mathbf{z}, u)$ (see (C.2)) can be taken as*

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

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. From the proof of Lemma 2.1 it can be seen that for continuous random variables there always exists a notion of residual $\varphi(x, \mathbf{z}) = F_{X|\mathbf{Z}}(x|\mathbf{z})$ which satisfies conditions (C.1) and (C.2). However without conditions (C.3) and (C.4) we cannot guarantee its uniqueness.

Remark 2.3. Let us first consider the example when $(X, \mathbf{Z})$ follows a multivariate Gaussian distribution, i.e.,

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

where $\mu_1 \in \mathbb{R}$, $\boldsymbol{\mu}_2 \in \mathbb{R}^d$, $\boldsymbol{\Sigma}$ is a $(d+1) \times (d+1)$ positive definite matrix with $\sigma_{11} > 0$, $\boldsymbol{\sigma}_{12} \in \mathbb{R}^{d \times 1}$ and $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{d \times d}$.

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

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

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

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

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 2.4. 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 $\varphi(X, Z) = X$.

110 *Remark 2.5.* Lemma 2.1 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 $U \sim \mathcal{U}(0, 1)$ which will act as the residual. Then (1.3), where h is defined in (2.1), shows that we can generate $X = F_{X|\mathbf{Z}}^{-1}(U|\mathbf{Z})$.

115 *Remark 2.6.* The above notion of nonparametric residual can be extended to the case of a continuous random vector $\mathbf{X} \in \mathbb{R}^p$. Without loss of generality assume let $p = 2$ and $\mathbf{X} = (X_1, X_2)$. Using arguments similar to those in the proof of Lemma 2.1 it can be shown that $\varphi(\mathbf{X}, \mathbf{Z}) := (F_{X_1|\mathbf{Z}}(X_1|\mathbf{Z}), F_{X_2|X_1, \mathbf{Z}}(X_2|X_1, \mathbf{Z}))$ satisfies conditions (C.1) and (C.2). In fact, now $\varphi(\mathbf{X}, \mathbf{Z})|\mathbf{Z} = \mathbf{z} \sim \mathcal{U}([0, 1]^2)$ for
120 all $\mathbf{z} \in \mathbb{R}^d$.

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
125 can use any nonparametric smoothing method, e.g., standard kernel methods, for estimation; see Bergsma (2011) for such an implementation. We discuss the estimation of the residuals in more detail in Section 5 of the online supplementary material.

3. Conditional independence

130 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 the defined notion of residual. We also briefly describe a procedure to test the mutual independence of the
135 three random variables/vectors (see Section 3.1). We start with the statement of the crucial lemma (proved in Section 9 of the online supplementary material).

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.*

140 *Remark 3.2.* As in Remark 2.6, we can extend the above result to the case when $\mathbf{X}$ and $\mathbf{Y}$ are random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively. In that case we define the conditional multivariate distribution transform $\mathbf{F}_{X|\mathbf{Z}}$ by successively conditioning on the co-ordinate random variables, i.e., if $p = 2$ and $\mathbf{X} = (X_1, X_2)$ then we can define $\mathbf{F}_{X|\mathbf{Z}}$ as $(F_{X_2|X_1, \mathbf{Z}}, F_{X_1|\mathbf{Z}})$. With this definition, Lemma 3.1
145 still holds.

Bergsma (2011) developed a test for conditional independence by testing the mutual independence of $F_{X|\mathbf{Z}}(X|\mathbf{Z})$ and $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$. In Section 7 of the online supplementary material we show that the test proposed in Bergsma (2011) may fail.

150 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 5 of the online supplementary material). Then, any procedure for testing the mutual independence of
155 $F_{X|\mathbf{Z}}(X|\mathbf{Z})$, $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$ and $\mathbf{Z}$ can be used. In this paper we advocate the use of the *energy* statistic (see Székely and Rizzo (2005)), 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 160 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
165 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) $V_1, \dots, V_r$ and we want to test their mutual independence. We assume that we know the marginal distributions of $V_1, \dots, V_r$; without loss of generality, we can assume that V_i 's are standard Gaussian random variables (vectors). We write $\mathbf{T} := (V_1, V_2, \dots, V_r) \in \mathbb{R}^k$ and introduce $\mathbf{T}_{\text{ind}} := (V_1^*, V_2^*, \dots, V_r^*)$ where V_j^* is an i.i.d. copy of V_j , $j = 1, 2, \dots, r$, but in $\mathbf{T}_{\text{ind}}$ the marginals, $V_1^*, V_2^*, \dots, V_r^*$, are independent. To test the mutual independence of $V_1, V_2, \dots, V_r$ all we need to test now is whether $\mathbf{T}$ and $\mathbf{T}_{\text{ind}}$ are identically distributed. If we observed a sample from $\mathbf{T}$, we can test for the equality of distributions of $\mathbf{T}$ and $\mathbf{T}_{\text{ind}}$ through a one-sample goodness-of-fit test for the standard multivariate normal distribution, i.e.,

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

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

In this paper we consider the following *energy* statistic

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

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

Suppose now that we observe n i.i.d. samples $\mathbf{T}_1, \dots, \mathbf{T}_n$ of $\mathbf{T}$. The (scaled) sample version of the energy statistic for testing the goodness-of-fit hypothesis is

$$\mathcal{E}_n(\mathbf{T}_1, \dots, \mathbf{T}_n) := 2 \sum_{i=1}^n \mathbb{E}\|\mathbf{T}_i - \mathbf{T}_{\text{ind}}\| - \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{T}_i - \mathbf{T}_j\| - n\mathbb{E}\|\mathbf{T}_{\text{ind}} - \mathbf{T}'_{\text{ind}}\|. \quad (3.1)$$

Note that the first expectation in the above display is with respect to $\mathbf{T}_{\text{ind}}$. Under the null hypothesis of mutual independence, the test statistic $\mathcal{E}_n(\mathbf{T}_1, \dots, \mathbf{T}_n)$

has a limiting distribution, as $n \rightarrow \infty$, while under the alternative hypothesis $\mathcal{E}_n(\mathbf{T}_1, \dots, \mathbf{T}_n)$ tends to infinity with probability one; 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 $\mathcal{E}_n(\mathbf{T}_1, \dots, \mathbf{T}_n)$ is consistent against general alternatives. The energy statistic $\mathcal{E}_n$ is easy to compute. As $\mathbf{T}_{\text{ind}}$ and $\mathbf{T}'_{\text{ind}}$ are i.i.d. $N(\mathbf{0}, \mathbf{I}_{k \times k})$ random variables:

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

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

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

The expression for $\mathbb{E}\|\mathbf{a} - \mathbf{T}_{\text{ind}}\|$ 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 \(2013\)](#). The appropriately modified code can also be found at <http://stat.columbia.edu/~rohit/research.html>.

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 (proved in Section 9 of the online supplementary material).

Lemma 3.3. *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*

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

where $\mathbf{F}_{\mathbf{Z}}(\mathbf{z}) = (F_{Z_d|Z_{d-1}, \dots, Z_1}(z_d|z_{d-1}, \dots, z_1), \dots, F_{Z_2|Z_1}(z_2|z_1), F_{Z_1}(z_1))$, $\mathbf{Z} = (Z_1, \dots, Z_d)$, and $\mathcal{U}([0, 1]^{d+2})$ denotes the uniform distribution on $[0, 1]^{d+2}$.

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

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

By Lemma 3.3, we have $\mathbf{W} \stackrel{\mathcal{D}}{=} (U_1, \dots, U_{d+2})$, where $U_1, U_2, \dots, U_{d+2}$ are i.i.d. $\mathcal{U}(0, 1)$ random variables. An equivalent formulation is

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

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

195 We observe i.i.d. data $\{(X_i, Y_i, \mathbf{Z}_i) : i = 1, \dots, n\}$ 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 $\mathbf{F}_{\mathbf{Z}}(\cdot)$ are known. Then we have an i.i.d. sample $\mathbf{T}_1, \dots, \mathbf{T}_n$ from $\mathbf{T}$, where

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

Now we can use the test statistic (3.1) to test the hypothesis of conditional independence. As the true conditional distribution functions $F_{X|\mathbf{Z}}$, $F_{Y|\mathbf{Z}}$, and $\mathbf{F}_{\mathbf{Z}}$ are 200 unknown, we can replace them by their estimates $\hat{F}_{X|\mathbf{Z}}$, $\hat{F}_{Y|\mathbf{Z}}$, and $\hat{\mathbf{F}}_{\mathbf{Z}}$, respectively, where $\hat{\mathbf{F}}_{\mathbf{Z}}(\mathbf{z}) = (\hat{F}_{Z_d|Z_{d-1}, \dots, Z_1}(z_d|z_{d-1}, \dots, z_1), \dots, \hat{F}_{Z_2|Z_1}(z_2|z_1), \hat{F}_{Z_1}(z_1))$; see Section 5 of the online supplemental material for more details on how to compute these estimates. Let us now define

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

205 for $i = 1, 2, \dots, n$. We will use $\widehat{\mathcal{E}}_n := \mathcal{E}_n(\hat{\mathbf{T}}_1, \dots, \hat{\mathbf{T}}_n)$ to test the hypothesis of conditional independence.

Remark 3.4. Note that the statistic $\widehat{\mathcal{E}}_n$ is not invariant under reordering of the coordinates of $\mathbf{Z}$ as $\hat{\mathbf{F}}_{\mathbf{Z}}(\mathbf{Z})$ depends on the order. But as pointed out in Sen and Chaudhuri (2012) and Serfling (2004), when the coordinates have a special 210 physical ordering we suggest conditioning successively on the less important co-ordinate random variables. In the absence of such an ordering, we suggest fixing any ordering before doing the experiment; asymptotically the effect of the particular choice will be negligible.

3.2.1. Approximating the asymptotic distribution through bootstrap

215 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 to approximate the 220 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 $\widehat{\mathbf{T}}_i$ instead of $\mathbf{T}_i$, 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 225 now describe the bootstrap procedure.

Let $\mathbb{P}_{n,\mathbf{Z}}$ be the empirical distribution of $\mathbf{Z}_1, \dots, \mathbf{Z}_n$.

Step 1: Generate an i.i.d. sample $\{U_{i,1}^*, U_{i,2}^*, \mathbf{Z}_{n,i}^*\}_{1 \leq i \leq n}$ of size n from the measure $\mathcal{U}(0, 1) \times \mathcal{U}(0, 1) \times \mathbb{P}_{n,\mathbf{Z}}$.

Step 2: The bootstrap sample is then $\{X_{n,1}^*, Y_{n,1}^*, \mathbf{Z}_{n,1}^*\}_{1 \leq i \leq n}$, where

$$X_{n,i}^* := \widehat{F}_{X|Z}^{-1}(U_{i,1}^* | \mathbf{Z}_{n,1}^*) \quad \text{and} \quad Y_{n,i}^* := \widehat{F}_{Y|Z}^{-1}(U_{i,2}^* | \mathbf{Z}_{n,1}^*).$$

226 **Step 3:** Use the bootstrap sample $\{X_{n,i}^*, Y_{n,i}^*, \mathbf{Z}_{n,i}^*\}_{1 \leq i \leq n}$ 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 5](#) for a discussion on smooth estimation of the conditional distribution functions.

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

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

We can now approximate the distribution of $\widehat{\mathcal{E}}_n$ by the conditional distribution 235 of $\mathcal{E}_n^*$ given the data $\{X_i, Y_i, \mathbf{Z}_i\}_{1 \leq i \leq n}$.

In [Section 6](#) of the online supplementary material 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. We

also compare the performance of the test proposed here with those proposed in
240 [Fukumizu et al. \(2007\)](#) and [Zhang et al. \(2012\)](#).

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
245 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 $\widehat{\mathcal{E}}_n$.

4. Discussion

250 In this paper we use nonparametric methods to estimate $F_{X|\mathbf{Z}}(\cdot|\cdot)$. A natural way of mitigating the 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
255 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.

We can also use these residuals to come up with a nonparametric notion of
260 partial correlation. The partial correlation of X and Y measures the degree of association between X and Y , removing the effect of $\mathbf{Z}$; see [Székely and Rizzo \(2014\)](#) and the references therein. 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
265 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 $F_{Y|\mathbf{Z}}(Y|\mathbf{Z})$.

Acknowledgments: 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
 270 also like to thank Probal Chaudhuri for motivating the problem. The research of second and third authors is supported by National Science Foundation.

References

- Bergsma, W. P., 2011. Nonparametric testing of conditional independence by means of the partial copula.
- 275 Billingsley, P., 1995. Probability and measure, 3rd Edition. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York.
- 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.
 280
- 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
 285 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.
- 290 Györfi, L., Walk, H., 2012. Strongly consistent nonparametric tests of conditional independence. *Statist. Probab. Lett.* 82 (6), 1145–1150.
- Huang, T.-M., 2010. Testing conditional independence using maximal nonlinear conditional correlation. *Ann. Statist.* 38 (4), 2047–2091.

- Lauritzen, S. L., 1996. Graphical models. Vol. 17 of Oxford Statistical Science
 295 Series. The Clarendon Press Oxford University Press, oxford Science Publications.
- Pearl, J., 2000. Causality. Cambridge University Press, Cambridge, models, reasoning, and inference.
- Poczos, B., Schneider, J., 2012. Conditional distance variance and correlation.
- 300 R Development Core Team, 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Sen, A., Sen, B., 2014. Testing independence and goodness-of-fit in linear models. *Biometrika* 101 (4), 927–942.
- Sen, B., Chaudhuri, P., 2012. On fractile transformation of covariates in regression.
 305 *Journal of the American Statistical Association* 107 (497), 349–361.
- Serfling, R., 2004. Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of statistical Planning and Inference* 123 (2), 259–278.
- Su, L., White, H., 2008. A nonparametric hellinger metric test for conditional
 310 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.
- 315 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. *Ann. Statist.* 42 (6), 2382–2412.

- 320 Székely, G. J., Rizzo, M. L., Bakirov, N. K., 2007. Measuring and testing dependence by correlation of distances. *Ann. Statist.* 35 (6), 2769–2794.
- 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.
- 325 Zhang, K., Peters, J., Janzing, D., Schölkopf, B., 2012. Kernel-based conditional independence test and application in causal discovery. arXiv preprint.