Statistics and Probability Letters 109 (2016) 208–213

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

On a nonparametric notion of residual and its applications

Rohit K. Patra [a,*], Bodhisattva Sen [a], Gábor J. Székely [b]

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

ARTICLE INFO

Article history:
Received 30 September 2015
Received in revised form 20 October 2015
Accepted 20 October 2015
Available online 28 November 2015

Keywords:
Conditional distribution function
Testing conditional independence

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 $\mathbf{Z}$.

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 $\mathbf{Z}$. 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. , $\varphi(X, \mathbf{Z}) \perp\!\!\!\perp \mathbf{Z}$.
(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).

* Corresponding author.
E-mail address: rohit@stat.columbia.edu (R.K. Patra).

http://dx.doi.org/10.1016/j.spl.2015.10.011
0167-7152/

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, 2009), in graphical models (Lauritzen, 1996), and in economic theory (see Chiappori and Salanié, 2000) among other fields. Traditional methods for testing conditional independence 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. Poczos 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 (see Appendix A), 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) $x \mapsto \varphi(x, \mathbf{z})$ is strictly increasing in its support, for every fixed $\mathbf{z} \in \mathbb{R}^d$.

Note that condition (C.3) is a 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) $\varphi(X, \mathbf{Z})|\mathbf{Z} = \mathbf{z} \sim \mathcal{U}(0, 1)$ 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 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 (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*

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

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.