

A class of consistent tests of independence between covariate and error in nonparametric regression with missing at random responses

Sthitadhi Das¹ and Saran Ishika Maiti²

¹Department of Mathematics
Brainware University, Barasat
West Bengal, India
dsd.mt@brainwareuniversity.ac.in

²Department of Statistics
Visva-Bharati
Santiniketan, India
saranishika.maiti@visva-bharati.ac.in

ABSTRACT

In nonparametric regression statistical relationship between covariate and random error is a matter of interest. For a traditional nonparametric regression model, $Y = m(X) + \epsilon$ with Y the response, X the covariate, ϵ the random error and $m(\cdot)$ a suitably chosen smooth function the null hypothesis of interest is X and ϵ are independent against all possible alternatives citing dependence between X and ϵ . A valid extension is if for an incomplete data set with several missing responses, similar testing of independence is applicable. In the present article, some observations on Y are missing at random whereas X possesses complete set of observations. Thereafter, in this missing at random (MAR) type data, test of independence between X and ϵ is investigated. Test statistics, based on some available or modified measures of association are constructed to develop consistent test procedures against a sequence of contiguous alternatives. The asymptotic powers of the test statistics are further determined through a finite sample simulation study. Finally, real data analysis is executed to examine the usefulness of the proposed test statistics.

Keywords: Kendall's τ , Measures of Association, Asymptotic Power, Contiguous Alternative, Nonparametric Regression Model, Missing at Random.

2000 Mathematics Subject Classification: 62G08, 62G30.

1 Introduction

We consider the simple nonparametric regression model $Y = m(X) + \epsilon$ with Y being response, X the covariate and ϵ the unexplained part of Y with the assumptions (i) $E[\epsilon|X = x] = 0 \forall x$ and (ii) $E[\epsilon^2|X = x] = \sigma^2(x)(> 0)$. ϵ is measured as $(Y - m(X))$ where $m(\cdot)$ is a smooth function. Usually, most statistical processes are designed for complete data. The elaborate testimony on nonparametric regression analysis technique for complete data has marked a

deep root in statistical literature while nonparametric regression analysis in presence of missing values is yet to garner substantial footing. In regression, missing observations may occur either on X or on Y or on both. The investigator should detect the nature of missingness of observations. In general, some observations on a random variable can be either *missing at random* (MAR) or *missing completely at random* (MCAR) or *not missing at random* (NMAR) (see Little and Rubin (1987)(Little and Rubin, 1987) for precise definition). In this study, we impose *missing at random* (MAR) assumption on Y . Pivotal studies in this direction date back in the context of estimation of the mean of response variable Y by Cheng (1994). Cheng provided various schemes of estimation of $m(x) = E(Y|X = x)$ under nonparametric setup through estimating $m(\cdot)$ using Nadaraya-Watson kernel density estimation technique. Later, Chu *et al.* (1995) estimated the regression function $m(X)$ in the model $Y = m(X) + \epsilon$ using local linear smoothing approach where Y assumes some missing values. In that paper, the estimation procedure was performed in two stages — (i) by ignoring the paired observations having missing responses and estimating $m(\cdot)$ as $\hat{m}_s(\cdot)$ using local linear regression equations, then (ii) imputing the missing Y -values by their corresponding functional values of $\hat{m}_s(\cdot)$, thereby estimating the regression function in the transformed model as $\hat{m}_l(\cdot)$. The asymptotic properties of both $\hat{m}_s(\cdot)$ and $\hat{m}_l(\cdot)$ have been then discussed through bias, variance and mean asymptotic squared error (MASE)s. Later on, Nitter(Nittner, 2003) considered an additive nonparametric regression model $Y = f(X) + \epsilon$ under the assumption that X contains missing observations whereas Y is completely observed. Along with *complete case analysis* by deleting all the incomplete bivariate observations on (X, Y) , methods like *single imputation*, *stochastic mean imputation* and *nearest neighbour imputation* for estimating the regression function $f(\cdot)$ are used here. Efromovich (Efromovich, 2011) studied the impact of MAR assumption on the estimated regression function $\hat{m}(x)$ through mean integrated squared error (MISE) and minimax rate in case of both small sample and large sample data. The present article delineates about a class of rank based and distance based test to verify the independence of ϵ and X in nonparametric regression model when missing at random (MAR) observations occur in Y . The test statistics involved herewith are shown as the functions of U-statistics with certain orders of degeneracy. The same issue of independence under nonparametric regression model with Missing completely at random (MCAR) response is already studied by Das *et al.*(Das and Maiti, 2024). In fact the present article may be considered as a valid extension of this work. The organization of this paper is as follows. In Section 2 the regression model is proposed under MAR for the response variable. The estimation of regression function is done using Nadaraya-Watson method as well as local linear smoothing technique. In Section 3 the hypotheses are formalized while Section 4 unfurls some pertinent nonparametric measures of association which are commonly used as the measures of dependence. In Section 5 the theory of contiguity is brought in for constructing a sequence of alternative hypotheses against the null hypothesis of interest suggesting independence between covariate and error. The limiting distributions of the test statistics under both the null hypothesis as well as contiguous alternatives are determined using the theory of U-statistics. Then the asymptotic powers of each of the test statistics are evaluated. Finite sample simulation study is carried out for each test statistic in Section 6. Finally, real data study is performed in Section 7 to check the efficacy of

each of the considered test statistics by computing simulated p-value.

2 Nonparametric Regression Setup under MAR

A random sample of size n is drawn from the joint cumulative distribution function of (X, Y) , say $(X_1, Y_1), \dots, (X_n, Y_n)$. Let us assume that among these data some Y_i 's are unobserved and remaining are observed. The corresponding X_i 's to missing Y 's are hence said to be *unresponsive*. Let δ denotes an indicator variable defined as

$$\delta_i = \begin{cases} 1, & \text{if } Y_i \text{ is observed} \\ 0, & \text{if } Y_i \text{ is missing, } i = 1, \dots, n. \end{cases}$$

For better understanding, i -th observation can be represented by the triplet (X_i, Y_i, δ_i) , $i = 1, \dots, n$. The probability expression $P(\delta = 1|X, Y)$ is defined as the probability of missingness of Y -value for given value of X . In MAR assumption $P(\delta = 1|X, Y) = P(\delta = 1|X) = p(X)$ where $p(\cdot)$ is a function of X . The missing Y -value is dependent on the corresponding value of X only. As a preliminary tool, the listwise deletion may be employed to estimate the regression function. But that would provide less efficient estimate. So here we adopt imputation technique.

Let there be k bivariate observations with reported responses and for the remaining $(n - k)$ pairs, Y values are not reported. Without loss of generality let for first (X_i, Y_i) 's, $i = 1, \dots, k$, Y_i 's are observed whereas for remaining (X_i, Y_i) 's for $i = (k + 1), \dots, n$, $Y_{k+1}, \dots, Y_n$ are missing. Then, a nonparametric model based only on the complete observations can be renamed as

$$Y = m_1(X) + \epsilon \quad (2.1)$$

with similar assumptions on ϵ as in the original model. The regression function $m_1(X)$ is called the *first stage regression function*. Upon deriving the estimator of $m_1(X)$ as $\hat{m}_1(X)$ by any traditional nonparametric estimation technique(s), missing Y -values are filled with $\hat{m}_1(X)$ and hence known as *imputed responses*. Thus, a set of n values on Y is achieved. Denoting the new response variable as $\tilde{Y}$, the following regression model is considered based on complete bivariate data on $(X, \tilde{Y})$ as

$$\tilde{Y} = m_2(X) + \tilde{\epsilon} \quad (2.2)$$

where the regression function $m_2(\cdot)$ is called the *second stage regression function*. Next, we have estimate $m_2(X)$ using any suitable nonparametric kernel density estimation method or smoothing technique as mentioned earlier.

2.1 Estimation of regression functions

2.1.1 Estimation using Nadaraya-Watson (NW) method

We begin with $m_1(x) = E(Y|X = x)$ in (2.1), which is further simplified as

$$\begin{aligned} m_1(x) &= E(Y|X = x) \\ &= \int_y y h_{Y|X}(y|x) dy = \int_y y \cdot \frac{f_{X,Y}(x, y)}{g_X(x)} dy = \frac{\int_y y f_{X,Y}(x, y) dy}{\int_y f_{X,Y}(x, y) dy}. \end{aligned}$$

where $f_{X,Y}(\cdot, \cdot)$ is the joint p.d.f. of (X, Y) , $g_X(\cdot)$ is the marginal p.d.f. of X and $h_{Y|X}(\cdot|x)$ is the conditional p.d.f. of Y at given value of X as x . In nonparametric situation, all the three densities are to be estimated using kernel smoothing approach (Härdle (Härdle, 1990), p. 32), originally provided by Nadaraya-Watson (1964) known as *Nadaraya-Watson estimation of kernel density method* (NW method).

Based on complete bivariate dataset (x_i, y_i) for $i = 1, \dots, k$, the first stage regression function $m_1(X)$ in (2.1) is estimated by Nadaraya-Watson method as follows.

$$\hat{m}_1(x) = \frac{\sum_{i=1}^k V\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^k V\left(\frac{X_i - x}{h}\right)} \quad (2.3)$$

where $V(\cdot)$ is the kernel density function defined on $\mathbb{R}$. h is the bandwidth for estimation and $h \rightarrow 0$ with $nh \rightarrow \infty$ as $n \rightarrow \infty$. Here, $V(u)$ is chosen as the widely used *Epanechnikov kernel*, i.e., $V(u) = \frac{3}{4}(1 - u^2)$ for $|u| \leq 1$. h is fixed by *Silverman's thumbrule* of bandwidth selection (Silverman (Silverman, 2018)). After estimating $\hat{m}_1(\cdot)$ at the values of X for which responses are missing, Y figures are imputed by,

$$y_j^* = \hat{m}_1(x_j), \quad j = (k+1), \dots, n.$$

and therefore a new set of responses is obtained as $\{\tilde{y}_i\}$, $i = 1, \dots, n$ where

$$\tilde{y}_i = \begin{cases} y_i, & \text{for } i = 1, \dots, k \\ y_i^*, & \text{for } i = (k+1), \dots, n. \end{cases}$$

Finally, we have n paired observations $(x_i, \tilde{y}_i)$'s ($i = 1, \dots, n$) on $(X, \tilde{Y})$ with $\tilde{Y}$ denoting imputed responses. The second stage regression function $m_2(\cdot)$ (2.2) is estimated by NW method of estimation on the basis of bivariate observations on $(X, \tilde{Y})$ as

$$\hat{m}_2(x) = \frac{\sum_{i=1}^n V\left(\frac{X_i - x}{h}\right) \tilde{Y}_i}{\sum_{i=1}^n V\left(\frac{X_i - x}{h}\right)} \quad (2.4)$$

where $\hat{m}_2(\cdot)$ is known as *final estimator* of $m(\cdot)$.

2.1.2 Estimation using imputed local linear smoothing (ILLS) method

Apart from kernel density estimation method, *local linear smoothing* technique is another bonafide technique to estimate $m(X)$ (Chu et al.(Chu and Cheng, 1995)). Here regression function is approximated by a polynomial of lowest degree, say, 1 in the neighbourhood of any point x , say $(x - h, x + h)$ with $h > 0$. Hence, a local linear approximation is given as

$$m(x_i) \approx c_1 + c_2(x_i - x), \quad i = 1, \dots, n \quad (2.5)$$

where c_1 and c_2 are *intercept* and *slope* of the approximated local linear smoother.

The parameters c_1 and c_2 are estimated by minimization of the following local weighted function

$$\sum_{i=1}^k (y_i - c_1 - c_2(x - x_i))^2 V\left(\frac{x - x_i}{h}\right) (= S(c_1, c_2), \text{ say}) \quad (2.6)$$

in the first stage where k responses are observed, by solving the following normal equations.

$$\frac{\delta}{\delta c_1} S(c_1, c_2) = 0 \quad (2.7)$$

$$\frac{\delta}{\delta c_2} S(c_1, c_2) = 0. \quad (2.8)$$

Equations (2.7) and (2.8) are further simplified as

$$\sum_{i=1}^k y_i V\left(\frac{x - x_i}{h}\right) = c_1 \sum_{i=1}^k V\left(\frac{x - x_i}{h}\right) + c_2(x - x_i) \sum_{i=1}^k V\left(\frac{x - x_i}{h}\right) \quad (2.9)$$

$$\sum_{i=1}^k y_i(x - x_i) V\left(\frac{x - x_i}{h}\right) = c_1 \sum_{i=1}^k (x - x_i) V\left(\frac{x - x_i}{h}\right) + c_2 \sum_{i=1}^k (x - x_i)^2 V\left(\frac{x - x_i}{h}\right) \quad (2.10)$$

By solving (2.9) and (2.10), we obtain

$$\hat{c}_1 = \frac{A_0 E_2 - A_1 E_1}{E_0 E_2 - E_1^2} \quad (2.11)$$

$$\hat{c}_2 = \frac{A_0 E_1 - A_1 E_0}{E_1^2 - E_0 E_2} \quad (2.12)$$

where $A_v = \sum_{i=1}^k y_i(x - x_i)^v V\left(\frac{x - x_i}{h}\right)$ for $v = 0, 1$ and $E_j = \sum_{i=1}^k (x - x_i)^j V\left(\frac{x - x_i}{h}\right)$ for $j = 0, 1, 2$.

Next, by the Taylor's series expansion of $m(x_i)$ upto first order in the neighbourhood of x ,

$$m(x_i) \approx m(x) - (x - x_i)m'(x) \quad (2.13)$$

for all $i = 1, \dots, k$ where $m'(\cdot)$ is the first order derivative of $m(\cdot)$. Then, $y_i \approx m(x) - (x - x_i)m'(x) + \epsilon_i$, $i = 1, \dots, k$ ($\approx$ means 'approximately equals to').

Then (2.3) can be simplified further as

$$\begin{aligned}\hat{m}_1(x) &= \frac{\sum_{i=1}^k V\left(\frac{x_i - x}{h}\right) \{\hat{c}_1 + \hat{c}_2(x - x_i)\}}{\sum_{i=1}^k V\left(\frac{x_i - x}{h}\right)} \\ &= \hat{c}_1 + h\hat{c}_2 \frac{\sum_{i=1}^k \left(\frac{x_i - x}{h}\right) V\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^k V\left(\frac{x_i - x}{h}\right)}\end{aligned}$$

which approaches to $\hat{c}_1$ for very small bandwidth h such that $h \rightarrow 0$.

It is to be noted that by deleting the incomplete pairs to estimate the regression function $m(X)$ as $\hat{m}_1(X)$ the loss of information, hence occurred, could be significant. Hence, imputed values of $\hat{m}_1(\cdot)$ at the covariate of missing responses could be a reasonable substitute. After imputation, the complete set of paired observations on (X, Y) is obtained. Then the regression function $m(\cdot)$ needs to be estimated again on the basis of newly obtained observations. Therefore, the new set of responses is expressed as

$$\tilde{y}_i = \delta_i y_i + (1 - \delta_i) \hat{m}_1(x_i), \quad i = 1, \dots, n.$$

Proceeding in similar fashion, we have to minimize

$$\sum_{i=1}^n (\tilde{y}_i - d_1 - d_2(x - x_i))^2 V\left(\frac{x - x_i}{h}\right) (= H(d_1, d_2), \text{ say}) \quad (2.14)$$

and the estimators of d_1 and d_2 are obtained as

$$\hat{d}_1 = \frac{\tilde{A}_0 \tilde{E}_2 - \tilde{A}_1 \tilde{E}_1}{\tilde{E}_0 \tilde{E}_2 - \tilde{E}_1^2} \quad (2.15)$$

$$\hat{d}_2 = \frac{\tilde{A}_0 \tilde{E}_1 - \tilde{A}_1 \tilde{E}_0}{\tilde{E}_1^2 - \tilde{E}_0 \tilde{E}_2} \quad (2.16)$$

where $\tilde{A}_v = \sum_{i=1}^k \tilde{y}_i (x - x_i)^v V\left(\frac{x - x_i}{h}\right)$ for $v = 0, 1$ and $\tilde{E}_j = \sum_{i=1}^k (x - x_i)^j V\left(\frac{x - x_i}{h}\right)$ for $j = 0, 1, 2$. The final estimator of $m(\cdot)$ is derived by simplifying (2.3) as

$$\hat{m}_2(x) = \frac{\sum_{i=1}^n V\left(\frac{x_i - x}{h}\right) \{\hat{d}_1 + \hat{d}_2(x - x_i)\}}{\sum_{i=1}^n V\left(\frac{x_i - x}{h}\right)} = \hat{d}_1 - h\hat{d}_2 \frac{\sum_{i=1}^n \left(\frac{x_i - x}{h}\right) V\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^n V\left(\frac{x_i - x}{h}\right)}$$

which tends to $\hat{d}_1$ as $h \rightarrow \infty$.

Therefore, $\hat{m}_2(x) \approx \hat{d}_1$, the final estimator of $m(\cdot)$ is called *imputed local linear smoother* (ILLS) of $m(X)$. In the subsequent sections, we proceed with the bivariate observations on $(X, \tilde{Y})$ to derive the setup of testing of independence between X and ϵ .

3 Hypotheses of interest

This article revolves with the objective of testing $H_0: X$ and ϵ independent ($X \perp\!\!\!\perp \epsilon$) against all possible alternative hypotheses citing no independence between them. So, a common way-out is to consider the joint probability distribution function of (X, ϵ) and verify whether it can be expressed in terms of product of marginal CDF of X and marginal CDF of ϵ . Intuitively, ϵ can not be observed as it may vary for particular $(X_i, \tilde{Y}_i)$ ($1 \leq i \leq n$). To check independence between X and ϵ , one may re-frame H_0 as X is independent to any properly defined function of ϵ , say $s(\epsilon)$. Since ϵ is not available practically, $s(\epsilon)$ needs to be approximated by a suitably defined function of observed values of Y -observations. As per the assumption of smoothness of $m(\cdot)$, we equate that function of Y that approximates the corresponding function of ϵ , i.e., $s(\tilde{Y}) \approx s(\epsilon)$. A common choice of $s(\epsilon)$ would be a linear contrast $\sum_{i=1}^r l_i \epsilon_i$ with coefficients $l_1, \dots, l_r$. By choosing $l_1 = 1 = l_3$ and $l_2 = -2$, Dhar et al. (Dhar, Bergsma and Dassios, 2018) considered $s(\epsilon_{i-1}, \epsilon_i, \epsilon_{i+1}) = \epsilon_{i-1} - 2\epsilon_i + \epsilon_{i+1}$ which has maximum absolute moment of order 3 among all possible linear functions based on $\{\epsilon_{i-1}, \epsilon_i, \epsilon_{i+1}\}$ ($i = 1, \dots, n$) subject to the constraints $l_1 + l_2 + l_3 = 0$ and $l_1^2 + l_2^2 + l_3^2$ is constant. Extending the idea, Das et al. ((Das and Maiti, 2022), (Das and Maiti, 2024)) defined $s(\epsilon_{i-2}, \epsilon_{i-1}, \epsilon_i, \epsilon_{i+1}) = \epsilon_{i-2} - 3\epsilon_{i-1} + 3\epsilon_i - \epsilon_{i+1}$ and verified that it has maximal fourth order absolute moment among all possible linear combinations of four *i.i.d.* error observations with the constraints $\sum_{i=1}^4 l_i = 0$ and $\sum_{i=1}^4 l_i^2$ being constant. In this article, we proceed with the third order difference of errors (third order linear contrast) under MAR condition in the response Y . Due to smoothness of $m(\cdot)$, instead of $s(\epsilon_{i-2}, \epsilon_{i-1}, \epsilon_i, \epsilon_{i+1})$, $i = 1, \dots, n$ therefore it is further approximated $s(\tilde{Y}_{i-2}, \tilde{Y}_{i-1}, \tilde{Y}_i, \tilde{Y}_{i+1}) (=Y_i^*$, say, for $i = 1, \dots, n$), the third order difference of Y -observations. Then the hypotheses of interest can be reshaped as $H_0: X \perp\!\!\!\perp s(\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3, \tilde{Y}_4)$ against $H_1: X \not\perp\!\!\!\perp s(\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3, \tilde{Y}_4)$, as $s(Y_1, Y_2, Y_3, Y_4) \approx s(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)$ with four *i.i.d.* errors $\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4$ without loss of generality.

4 Measures of association & test statistics

In current investigation, the concept of association is less implicating compared to the concept of dependence of two variables. For example, one may develop a measure of association, say $\psi(X, Y)$, indicating $X \perp\!\!\!\perp Y \iff \psi = 0$. To study the relation between X and ϵ , the test statistics are constructed in line with three such already existing nonparametric measures of associations, viz. (i) Kendall's τ , (ii) Szekely et al.'s $dCov$, (iii) Bergsma et al.'s τ^* .

4.1 Kendall's τ

If Π_c and Π_d are the probabilities of concordance and discordance of the bivariate observations obtained on (X, Y) , then most popular Kendall's $\tau = \Pi_c - \Pi_d$. In general, $-1 \leq \tau \leq 1$. But τ zero does not ensure the independence of X and Y .

4.2 Székely's $dCov$

Szekely et al.(Székely, Rizzo and Bakirov, 2007) invented a distance based covariance measure $dCov$ based on Euclidean distances between sample observations on jointly distributed random variables (X, Y) . To detect if two random variables are independent or not, but the distance covariance works in such situation. In general, $0 \leq dCor(X, Y) < \infty$ and $dCov(X, Y) = 0 \iff X \perp\!\!\!\perp Y$, otherwise X is dependent to Y .

4.3 Bergsma's τ^*

Another approach for testing independence between two random variables in case of an ordinal dataset is provided by Bergsma et al.(Bergsma and Dassios, 2014)'s τ^* . It is an improvement of Kendall's τ , where $\tau^* = 0$ implies and is implied by the independence of X and Y . Based on four *i.i.d.* replications (X_i, Y_i) , $i = 1, \dots, 4$ of (X, Y) , τ^* is defined as

$$\tau^* = E[a(X_1, X_2, X_3, X_4)a(Y_1, Y_2, Y_3, Y_4)] \quad (4.1)$$

where $a(s_1, s_2, s_3, s_4) = \text{sign}\{|s_1 - s_2| + |s_3 - s_4| - |s_1 - s_3| - |s_2 - s_4|\}$, $z_i \in \mathbb{R} \forall i = 1, \dots, 4$. $0 \leq \tau^*(X, Y) \leq 1$.

We would like to consider the above three measures of association as well as dependence to test the independence between X and ϵ .

4.4 Proposition of test statistics

As we discussed in **Section 3** on bivariate observations (X, Y^*) with Y^* being third order difference of Y , the bivariate observations (x_i, y_i^*) 's ($i = 1, \dots, n$) are generated first. Then X observations are ordered and the corresponding Y^* observations are taken. Let us name those Y^* observations as "induced" ordered Y^* . Therefore, the test statistics would be constructed on $\{(x_{(1)}, y_{(1)}^*), \dots, (x_{(n)}, y_{(n)}^*)\}$. In addition to the proposed test statistics, three other popular goodness-of-fit statistics are reported for experimental comparison.

$$T'_{n,KS} = \sqrt{n} \sup_{x,y \in \mathbb{R}} |F_n(x, y) - \hat{G}_X(x) \hat{H}(y)| \quad (4.2)$$

$$T'_{n,CM} = n \int \int \left(F_n(x, y) - \hat{G}_X(x) \hat{H}(y) \right)^2 d\hat{G}_X(x) d\hat{H}(y) \quad (4.3)$$

$$T'_{n,AD} = n \int \int \frac{\left(F_n(x, y) - \hat{G}_X(x) \hat{H}(y) \right)^2}{\hat{G}_X(x) \hat{H}(y) (1 - \hat{G}_X(x)) (1 - \hat{H}(y))} d\hat{G}_X(x) d\hat{H}(y) \quad (4.4)$$

where F_- is the left continuous portion of F , $F_n(\cdot, \cdot)$ is the empirical joint probability distribution function of (X, Y^*) and $\hat{G}(\cdot)$, $\hat{H}(\cdot)$ are the marginal empirical joint probability distribution functions of X and Y^* respectively.

$$T'_{n,1} = \frac{1}{\binom{n}{2}} \sum_{1 \leq i < j \leq n} \text{sign}\{(x_{(i)} - x_{(j)})(y_{(i)}^* - y_{(j)}^*)\} \quad (4.5)$$

$$T'_{n,2} = \frac{1}{\binom{n}{4}} \sum_{1 \leq i < j < k < l \leq n} a(x_{(i)}, x_{(j)}, x_{(k)}, x_{(l)}) a(y_{(i)}^*, y_{(j)}^*, y_{(k)}^*, y_{(l)}^*) \quad (4.6)$$

$$T'_{n,3} = \frac{1}{\binom{n}{4}} \sum_{1 \leq i < j < k < l \leq n} \frac{1}{4} h(x_{(i)}, x_{(j)}, x_{(k)}, x_{(l)}) h(y_{(i)}^*, y_{(j)}^*, y_{(k)}^*, y_{(l)}^*) \quad (4.7)$$

where $sign(t) = \frac{t}{|t|}$ if $t \neq 0$ or 0 otherwise, $a(p, q, r, s) = sign\{|p - q| + |r - s| - |p - r| - |q - s|\}$ and $h(p, q, r, s) = \{|p - q| + |r - s| - |p - r| - |q - s|\}$; $p, q, r, s \in \mathbb{R}$. $T'_{n,1}$ is the sample *Kendall's tau* between X and Y^* whereas $T'_{n,2}$ is the sample version of Bergsma et al. (Bergsma and Dassios, 2014)'s τ^* . $T'_{n,3}$ is the sample counterpart of $dCov$ inspired by Székely et al. (Székely et al., 2007). The set of bivariate observations on (X, Y^*) are $\{(x_{(1)}, y_{(1)}^*), \dots, (x_{(n)}, y_{(n)}^*)\}$ where $y_{(i)}^* := \tilde{y}_{(i+1)} - 3\tilde{y}_{(i)} + 3\tilde{y}_{(i-1)} - \tilde{y}_{(i-2)}$ with $\tilde{y}_{(-1)} = \tilde{y}_{(0)} = \tilde{y}_{(1)}$, $\tilde{y}_{(n+1)} = \tilde{y}_{(n)}$ (Y^* is the third order difference of the imputed set of Y , viz. $\tilde{Y}$). The corresponding errors are $\epsilon_{(1)}^*, \dots, \epsilon_{(n)}^*$ which are also the induced ordered values on $\epsilon_1^*, \dots, \epsilon_n^*$.

Under H_0 the critical regions for $T'_{n,i}$'s are derived as $\omega_{n,i} : T'_{n,i} > t_{i,\alpha}$, $i = KS, CM, AD, 1, 2, 3$ for $0 < \alpha < 1$ being the level of significance such that $P_{H_0}[T'_{n,i} > t_{i,\alpha}] = \alpha$ and $t_{i,\alpha}$ is the α -th critical point of the asymptotic null distribution of $T'_{n,i}$.

5 Asymptotic power analysis of test statistics

With the help of the test statistics defined in (4.5), (4.6) and (4.7) we establish some relevant test procedures. The null hypothesis H_0 indicates independence of X and ϵ . The alternative hypothesis is a general two-sided one which can be re-expressed into a special set of local hypotheses, technically known as *contiguous alternatives*.

The concept of *contiguity* is popular in the theory of asymptotic distributions. Le Cam (1960) introduced the theory of contiguity on two arbitrary sequences of probability measures, say, P_n and Q_n , on a specified sequence of measurable spaces $(\mathcal{X}_n, \mathcal{A}_n)$.

Definition 5.1. For an arbitrary sequence of events $A_n \in \mathcal{A}_n$, if $P_n(A_n) \rightarrow 0 \implies Q_n(A_n) \rightarrow 0$ for sufficiently large sample size n , then Q_n is concluded as contiguous with respect to P_n . It is symbolically expressed as $P_n \triangleleft Q_n$.

To check whether $P_n \triangleleft Q_n$, first lemma of contiguity, proposed by Le Cam, is worth mentioning.

Lemma 5.1. If $\log \frac{dQ_n}{dP_n} \sim N(-\frac{\theta}{2}, \theta)$ where $(\theta > 0)$, Q_n is contiguous with respect to P_n , i.e. $P_n \triangleleft Q_n$.

For detailed proof of the above, Hajek et al. (Šidák, Sen and Hájek, 1999), Van Der Vaart (van der Vaart, 2002) are recommended.

A sequence of alternative hypotheses, say H_n , that converges to H_0 as $n \rightarrow \infty$, is defined as a sequence of contiguous alternatives. Here, H_n is defined as

$$H_n : F_{n;X,\epsilon^*}(x, y) = \left(1 - \frac{\gamma}{\sqrt{n}}\right) G_X(x) H_{\epsilon^*}(y) + \frac{\gamma}{\sqrt{n}} \xi_{X,\epsilon^*}(x, y) \quad (5.1)$$

where $F_{n;X,\epsilon^*}(\cdot, \cdot)$ is the joint CDF of (X, ϵ^*) under local alternative sequence H_n . $H_{\epsilon^*}(\cdot)$ and $G_X(\cdot)$ are respectively the marginal CDFs of ϵ^* and X . $\xi_{X,\epsilon^*}(\cdot, \cdot)$ is the joint distribution function of (X, ϵ^*) under an alternative hypothesis (not local), known as proper CDF of (X, ϵ^*) . γ is the mixing quantity and $\gamma > 0$.

In subsection 4.4 we already discussed about generation of bivariate samples from the joint

CDF of (X, Y^*) and construction of three test statistics. Additionally, the proposed test statistics $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$ are degenerate U-statistics with orders of degeneracy 0, 1 and 1 respectively. Using the theory of limiting distribution of an U-statistic of order of degeneracy $v(\geq 0)$, the asymptotic distributions of $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$ are to be derived. To address this, Le Cam's lemma is useful. Let us assume

Assumption 1. $f_{X,\epsilon^*}(x, e) > 0$ for all x, e , where $f_{X,\epsilon^*}(\cdot, \cdot)$ is the joint PDF of (X, ϵ^*) under H_0 .

Assumption 2. $E_f \left(\frac{\xi'_{X,\epsilon^*}(x, e)}{f_{X,\epsilon^*}(x, e)} - 1 \right)^2 < \infty$ where $\xi'_{X,\epsilon^*}(\cdot, \cdot)$ is the joint proper PDF of (X, ϵ^*) , defined as $\xi'_{X,\epsilon^*}(u, v) = \frac{\delta^2}{\delta u \delta v} \xi_{X,\epsilon^*}(u, v)$.

Theorem 5.2. Under Assumption 1 and Assumption 2, H_n is a sequence of contiguous alternatives.

An in-depth proof of the above is available in Das et al. (Das and Maiti, 2022). Now, assembling the concepts including definition, order of degeneracy, variance etc. related to a degenerate U statistic (see Serfling(Serfling, 2009)), the next objective is to determine the asymptotic distributions of proposed test statistics.

5.1 Asymptotic distributions of $T'_{n,KS}$, $T'_{n,CM}$ and $T'_{n,AD}$

The asymptotic power analysis of the traditional test statistics $T'_{n,KS}$, $T'_{n,CM}$ and $T'_{n,AD}$, based on the Kolmogorov-Smirnov (KS), Cramer von Mises(CM) and Anderson-Darling(AD) distances, between the joint CDF of (X, Y^*) and the product of marginal CDFs of X and Y^* is performed taking Y^* as second/third order difference of Y . An empirical Gaussian process $\frac{\sqrt{n}(F_n(x, y) - \hat{G}_X(x)\hat{H}_\epsilon(y))}{\{P(x, y)\}^\lambda}$, is the pivot for the construction of the test statistics provided by (4.2), (4.3) and (4.4), where $P(x, y) = G_X(x)H_\epsilon(y)(1 - G_X(x))(1 - H_\epsilon(y))$ with $0 < \lambda \leq \frac{1}{2}$.

Result 1. (Einmahl et al.(Einmahl and V.Keilegom, 2008))

Under assumptions (i) $X \sim f_X(\cdot)$ defined on a bounded interval $I(X)(\subseteq \mathbb{R})$, (ii) $\epsilon \sim g_\epsilon(\cdot)$ such that $\sup_{u \in \mathbb{R}} g_\epsilon(u) := M(< \infty)$ and (iii) $m(\cdot)$ are smooth in $\mathbb{R}$, the empirical Gaussian process

$$\frac{\sqrt{n}(F_n(x, y) - \hat{G}_X(x)\hat{H}_\epsilon(y))}{\{P(x, y)\}^\lambda}$$

for $x \in I(X)$, $y \in \mathbb{R}$ weakly approaches a Gaussian process $Z_0(G_X(x), H_\epsilon(y))$ defined on $[0, 1]^2$.

Here, $Z_0(u, v) = Z(u, v) - uZ(1, v) - vZ(u, 1) + uvZ(1, 1)$ where Z is a standard Gaussian process defined on $[0, 1]^2$. Moreover, Z_0 is a 4-sided tied down Wiener process with $Z_0(0, 0) = 0 = Z_0(1, 1)$. Then the centered, bivariate Gaussian process Z_0 satisfies

1. $E(Z_0(u, v)) = 0$ for all $0 \leq u, v \leq 1$.
2. $Cov(Z_0(u_1, v_1), Z_0(u_2, v_2)) = (\min(u_1, u_2) - u_1u_2)(\min(v_1, v_2) - v_1v_2)$ for all $0 \leq u_1, v_1, u_2, v_2 \leq 1$.

Proof is available in Einmahl et al. (Einmahl and V.Keilegom, 2008). Instead of considering ϵ (or, $\hat{\epsilon}$), the test of independence between X and second/third difference of Y can be performed with the aid of Result 1 in similar way. In that case, the joint CDF of (X, Y^*) is taken instead of that of (X, ϵ) . The result can be rewritten using the joint CDF as well as marginal CDFs of (X, Y^*) .

Using the above concept, now, we discuss on the limiting distributions of $T'_{n,KS}$, $T'_{n,CM}$ and $T'_{n,AD}$ along with $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$ under H_0 and H_n both. Let us assume

Assumption 3. $X_1, \dots, X_n$ are *i.i.d.* random variables with distribution function $G_X(\cdot)$.

Assumption 4. $\tilde{Y}_1, \dots, \tilde{Y}_n$ are obtained from the model $\tilde{Y} = m(X) + \epsilon$ with ϵ having bounded probability density function and $E(\epsilon|X) = 0$.

Taking all the Assumptions 1-4 the limiting distributions of all the test statistics are given by the following theorems.

Theorem 5.3. Under H_0 ,

1.

$$T'_{n,KS} \xrightarrow{L} \sup_{0 \leq u, v \leq 1} |Z_0(u, v)|$$

as discussed by Einmahl et al. (Einmahl and V.Keilegom, 2008).

2.

$$T'_{n,CM} \xrightarrow{L} n \int_0^1 \int_0^1 Z_0^2(u, v) du dv \stackrel{D}{=} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mu_{ij} T_{ij}^2$$

where T_{ij} 's are the *i.i.d.* standard normal variables, μ_{ij} 's are the eigenvalues with corresponding eigenfunctions ξ_i, ξ_j 's ($i, j = 1, \dots, n$) as obtained from the following integral equation

$$\int_0^1 \int_0^1 K((x, y), (u, v)) \xi_i(u) \xi_j(v) du dv = \mu_{ij} \xi_i(x) \xi_j(y)$$

where K is the kernel defined as

$$K((u_1, v_1), (u_2, v_2)) = (\min(u_1, u_2) - u_1 u_2)(\min(v_1, v_2) - v_1 v_2).$$

Here, the eigenvalues μ_{ij} 's are solved as $\frac{1}{\pi^4 i^2 j^2}$ and the corresponding eigenfunctions are determined as

$$\psi_{ij}(u, v) = 2 \sin(\pi i u) \sin(\pi j v), \quad i, j = 1, 2, \dots$$

The asymptotic distribution of $T'_{n,CM}$ was derived by Blum et al. (Blum, Kiefer and Rosenblatt, 1961).

3.

$$T'_{n,AD} \xrightarrow{L} n \int_0^1 \int_0^1 \frac{Z_0^2(u, v)}{uv(1-u)(1-v)} du dv \stackrel{D}{=} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mu_{ij}^* T_{ij}^{*2}.$$

where T_{ij}^* 's are the *i.i.d.* standard normal variables, μ_{ij}^* 's are the eigenvalues with corresponding eigenfunctions ξ_i^*, ξ_j^* 's ($i, j = 1, \dots, n$) as obtained from the following integral equation

$$\int_0^1 \int_0^1 K^*((x, y), (u, v)) \xi_i^*(u) \xi_j^*(v) du dv = \mu_{ij}^* \xi_i^*(x) \xi_j^*(y)$$

where K^* is the kernel given by

$$K^*((u_1, v_1), (u_2, v_2)) = \frac{(\min(u_1, u_2) - u_1 u_2)(\min(v_1, v_2) - v_1 v_2)}{\sqrt{u_1 u_2(1 - u_1)(1 - u_2)} \sqrt{v_1 v_2(1 - v_1)(1 - v_2)}}.$$

The eigenvalues and the corresponding eigenfunctions associated with K^* were determined by Koning et al. (Koning and Protasov, 2003) as

$$\mu_{ij}^* = \frac{1}{ij(i+1)(j+1)}$$

and

$$\psi_{ij}^*(u, v) = \sqrt{2 \cdot \frac{2i+1}{i(i+1)}} \sqrt{2 \cdot \frac{2j+1}{j(j+1)}} u(u-1)v(v-1)L_{i-1}(2u-1)L_{j-1}(2v-1)$$

for $i, j = 1, \dots, n$, where $L_s(\cdot)$ is the Legendre's polynomial of degree s .

Theorem 5.4. Under H_n ,

1.

$$T'_{n,KS} \xrightarrow{L} \sup_{u,v} |Z_1(u, v)|.$$

where $Z_1(x, y)$ is a bivariate Gaussian process with

$$\begin{aligned} E[Z_1(x_i, y_i)] &= Cov_{H_0} \left(\log \frac{dF_n(x, y)}{dF_0(x, y)}, \sqrt{n}(F_n(x_i, y_i) - \hat{F}_X(x_i)\hat{G}(y_i)) \right) \\ &= E_{H_0} \left[\gamma \left(\frac{dK(x, y)}{dF_0(x, y)} - 1 \right) (F_n(x_i, y_i) - \hat{F}_X(x_i)\hat{G}(y_i)) \right], \end{aligned}$$

for $i = 1, \dots, n$, where $k(\cdot, \cdot)$ denotes the joint pdf of (X, Y^*) under H_1 , $f(\cdot, \cdot)$ denotes the joint pdf of (X, Y^*) under H_0 . $F_0(\cdot, \cdot)$ and $K(\cdot, \cdot)$ are the joint CDFs of (X, Y^*) under H_0 and H_1 respectively. Moreover, $Cov(Z_1(u_1, v_1), Z_1(u_2, v_2)) = Cov(Z_0(u_1, v_1), Z_0(u_2, v_2))$ for all $0 \leq u_1, v_1, u_2, v_2 \leq 1$.

2.

$$T'_{n,CM} \xrightarrow{L} n \int_0^1 \int_0^1 Z_1^2(u, v) du dv \stackrel{D}{=} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mu_{ij} (T_{ij} + a_{ij})^2$$

where T_{ij} 's are the i.i.d. standard normal variables, μ_{ij} 's are the eigenvalues associated with kernel K , ψ_{ij} 's are the corresponding eigenfunctions as given by Theorem 5.4(II) and

$$a_{ij} = \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\frac{k_{X,\epsilon}(x, y)}{f_{X,\epsilon}(x, y)} - 1 \right) \psi_{ij}(x, y) f_{X,\epsilon}(x, y) dx dy$$

for $i, j = 1, 2, \dots$

3.

$$T'_{n,AD} \xrightarrow{L} n \int_0^1 \int_0^1 \frac{Z_1^2(u, v)}{uv(1-u)(1-v)} du dv \stackrel{D}{=} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mu_{ij}^* (T_{ij}^* + a_{ij}^*)^2,$$

where T_{ij}^* 's are the i.i.d. standard normal variables, μ_{ij}^* 's are the eigenvalues associated with kernel K^* , ψ_{ij}^* 's are the corresponding eigenfunctions as given by Theorem 5.4(III) and

$$a_{ij}^* = \gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\frac{k_{X,\epsilon}(x, y)}{f_{X,\epsilon}(x, y)} - 1 \right) \psi_{ij}^*(x, y) f_{X,\epsilon}(x, y) dx dy$$

for $i, j = 1, 2, \dots$

5.2 Asymptotic distributions of $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$

Theorem 5.5. 1. $T'_{n,1}$ has kernel of order of degeneracy 0.

2. $T'_{n,1} \xrightarrow{P} 0$ under H_0 .

3. Under H_0 , $\sqrt{n}(T'_{n,1} - E(T'_{n,1})) \xrightarrow{L} N(0, 4\zeta_1)$.

4. Under H_n , $\sqrt{n}(T'_{n,1} - E(T'_{n,1})) \xrightarrow{L} N(\mu_1, 4\zeta_1)$, where

$$\mu_1 = 2\gamma \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[2 \int_{-\infty}^x \int_{-\infty}^y dG_X(u) dH_{\epsilon^*}(v) + 2 \int_x^{\infty} \int_y^{\infty} dG_X(u) dH_{\epsilon^*}(v) \right] d\xi_{X,\epsilon^*}(x, y) \quad (5.2)$$

and,

$$\zeta_1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[2 \int_{-\infty}^x \int_{-\infty}^y dG_X(u) dH_{\epsilon^*}(v) + 2 \int_x^{\infty} \int_y^{\infty} dG_X(u) dH_{\epsilon^*}(v) \right]^2 dG_X(x) dH_{\epsilon^*}(y). \quad (5.3)$$

Next, the asymptotic distributions are proposed, borrowing the ideas from Lee (Lee, 1990).

Theorem 5.6. 1. $T'_{n,2}$ has kernel of order of degeneracy 1.

2. $T'_{n,2} \xrightarrow{P} 0$ under H_0 .

3. The asymptotic distribution for $T'_{n,2}$ under H_0 is given by

$$n(T'_{n,2} - E(T'_{n,2})) \xrightarrow{L} \sum_{k=1}^{\infty} \gamma_k \{\psi_k^2 - 1\}$$

where $\psi_1, \psi_2, \dots$ are i.i.d. $N(0, 1)$ random variables, γ_k 's are the eigenvalues associated with

$$\begin{aligned} r(x, y) &= E[\text{sign}\{|X_{(1)} - X_{(2)}| + |X_{(3)} - X_{(4)}| - |X_{(1)} - X_{(3)}| - |X_{(2)} - X_{(4)}|\}] \\ &\quad \times \text{sign}\{|Y_{(1)}^* - Y_{(2)}^*| + |Y_{(3)}^* - Y_{(4)}^*| - |Y_{(1)}^* - Y_{(3)}^*| - |Y_{(2)}^* - Y_{(4)}^*|\}] \\ &\quad [X_{(1)} = x, Y_{(1)}^* = y]. \end{aligned}$$

4. The asymptotic distribution for $T'_{n,2}$ under H_n is given by

$$n(T'_{n,2} - E(T'_{n,2})) \xrightarrow{L} \sum_{k=1}^{\infty} \gamma_k \{(\psi_k + b_k)^2 - 1\} \quad (5.4)$$

where $\psi_1, \psi_2, \dots$ are i.i.d. $N(0, 1)$ random variables, γ_k 's are the eigenvalues associated with $l(x, y)$ given in iii. The quantities b_k 's are defined as

$$b_k = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h \phi_k(x) \phi_k(y) dG_X(x) dH_{\epsilon^*}(y). \quad (5.5)$$

where ϕ_k 's are the eigen functions corresponding to γ_k 's, $k = 1, 2, \dots$ satisfying the following eigenvalue-eigenfunction equation

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} r(x, y) \prod_{m=2}^4 \phi_k(X_i) \phi_k(Y_i) dF_{X_i, Y_i}(X_i, Y_i) = \gamma_k \phi_k(x) \phi_k(y) \quad (5.6)$$

and h is defined as $\lim_{n \rightarrow \infty} h_n$ where h_n is determined from the relation

$$\frac{dQ_n}{dP_0} = 1 + \frac{1}{\sqrt{n}} h_n$$

with P_0 and Q_n being the probability measures associated with H_0 and H_n respectively.

Theorem 5.7. 1. $T'_{n,3}$ has kernel of order of degeneracy 1.

2. $T'_{n,3} \xrightarrow{P} 0$ under H_0 .

3. The asymptotic distribution for $T'_{n,3}$ under H_0 is given by

$$n(T'_{n,3} - E(T'_{n,3})) \xrightarrow{L} \sum_{k=1}^{\infty} \tilde{\gamma}_k \{\tilde{\psi}_k^2 - 1\}$$

where $\tilde{\psi}_1, \tilde{\psi}_2, \dots$ are i.i.d. $N(0, 1)$ random variables, $\tilde{\gamma}$'s are the eigenvalues associated with

$$\begin{aligned} \tilde{r}(x, y) &= E[\{|X_{(1)} - X_{(2)}| + |X_{(3)} - X_{(4)}| - |X_{(1)} - X_{(3)}| - |X_{(2)} - X_{(4)}|\} \\ &\quad \times \{|Y_{(1)}^* - Y_{(2)}^*| + |Y_{(3)}^* - Y_{(4)}^*| - |Y_{(1)}^* - Y_{(3)}^*| - |Y_{(2)}^* - Y_{(4)}^*|\} \\ &\quad |X_{(1)} = x, Y_{(1)}^* = y]. \end{aligned}$$

4. The asymptotic distribution for $T'_{n,3}$ under H_n is given by

$$n(T'_{n,3} - E(T'_{n,3})) \xrightarrow{L} \sum_{k=1}^{\infty} \tilde{\gamma}_k \{(\tilde{\psi}_k + \tilde{b}_k)^2 - 1\} \quad (5.7)$$

where $\tilde{\psi}_1, \tilde{\psi}_2, \dots$ are i.i.d. $N(0, 1)$ random variables, $\tilde{\gamma}_k$'s are the eigenvalues associated with $\tilde{r}(x, y)$ given in (iii). The quantities $\tilde{b}_k$'s are defined as

$$\tilde{b}_k = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h \tilde{\phi}_k(x) \tilde{\phi}_k(y) dG_X(x) dH_{\epsilon^*}(y). \quad (5.8)$$

where $\tilde{\phi}_k$'s are the eigenfunctions corresponding to $\tilde{\gamma}_k$'s, $k = 1, 2, \dots$
 h is defined in Theorem 5.7.

Both Theorem 5.6 and Theorem 5.7 describe the limiting distributions of $T'_{n,2}$ and $T'_{n,3}$, the U -statistics with order of degeneracy 1 each (Gregory(Gregory, 1977)).

5.3 Examples

The asymptotic power analysis of $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$ are to be conducted for different values of γ in (5.1). Under the assumption of MAR of Y -observations, the values of powers of each of the test statistics at various values of γ are determined using Theorem 5.3, Theorem 5.4, Theorem 5.5, Theorem 5.6 and Theorem 5.7. We consider three examples under H_1 as follows. Note that for each of example H_0 assumes that ϵ and X are independent standard normal, i.e., $f(\epsilon, X) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2 + \epsilon^2)}$ while under H_1 $X \sim N(0, 1)$ with the conditional distribution of ϵ given X as mentioned below.

Example 5.3.1. $\epsilon|X = x \sim N\left(0, \frac{1+5x}{100}\right)$ with $X \sim N(0, 1)$.

Example 5.3.2. $\epsilon|X = x \stackrel{d}{=} \frac{1}{10} \sqrt{1 - (0.2x)^{\frac{1}{4}}} T_x$ where $T_x \sim t_{2/(0.2x)^{\frac{1}{4}}}$ with $X \sim N(0, 1)$.

Example 5.3.3. $\epsilon|X = x \stackrel{d}{=} \frac{V_x - d_x}{10\sqrt{2d_x}}$ where $V_x \sim \chi_{d_x}^2$, $d_x = \frac{1}{3x}$, with $X \sim N(0, 1)$.

In Example 5.3.1 the conditional error distribution is normal with zero mean but variance depends on x . A t -type conditional error distribution is provided by 5.3.2. Example 5.3.3 states the conditional distribution of ϵ given $X = x$ as a standardized version of a Chi-square distribution with degrees of freedom $s_x = \frac{1}{3x}$. In both cases, the null distribution of (X, ϵ) is taken as a bivariate normal distribution with parameters $(0, 0, 1, 1, 0)$. In all examples, the proportions of missingness of Y -observations are fixed to 5%, 10% and 20%. The sample size is taken as 100 and the number of replications is fixed to 500.

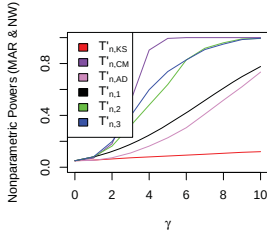
First, we generate *i.i.d.* observations $(x_1, \epsilon_1), \dots, (x_{100}, \epsilon_{100})$ from $F_{X, \epsilon}(\cdot, \cdot)$, the CDF a standard bivariate normal distribution. Taking the regression function as $m(x) = x^2 - 2x^3$, the responses y_i 's are computed as $(m(x_i) + \epsilon_i)$'s for $i = 1, \dots, 100$. Suppose $100p$ ($0 < p < 1$) observations on Y are missing at random.

In this exercise, probability of being missing is modelled by nonparametric additive logistic function in $m(X)$, i.e., $p(X) = \frac{e^{m(X)}}{1 + e^{m(X)}}$. The estimation of $m(\cdot)$ is performed by **Nadaraya-Watson (NW)** method of estimation of kernel density and by **imputed local linear smoothing (ILLS)** using the only available bivariate observations on (X, Y) and the estimated regression function is $\hat{m}_1(\cdot)$. Next, the values of $\hat{m}_1(\cdot)$ are obtained at the X -values with missing responses. The finally obtained paired observations are $(x_1, \tilde{y}_1), \dots, (x_{100}, \tilde{y}_{100})$ and the regression function is hence estimated as $\hat{m}_2(\cdot)$. Then the random errors $\epsilon_1, \dots, \epsilon_n$ are estimated as $(\tilde{y}_i - \hat{m}_2(x_i))$'s, denoted by $\hat{\epsilon}_i$'s ($i = 1, \dots, 100$). Next, the empirical CDF of ϵ is formed on the basis of the estimated errors $\hat{\epsilon}_1, \dots, \hat{\epsilon}_{100}$, say $\hat{F}_{\hat{\epsilon}}(\cdot)$. A bootstrap sample of size n from $\hat{F}_{\hat{\epsilon}}(\cdot)$ is obtained as $\hat{\epsilon}_{1,b}, \dots, \hat{\epsilon}_{100,b}$ for $b = 1, \dots, 500$. Then, the centered errors are obtained as $\hat{\epsilon}_{i,cent}^b = \hat{\epsilon}_{i,b} - \frac{1}{100} \sum_{k=1}^n \hat{\epsilon}_{k,b}$, $i = 1, \dots, 100$, $b = 1, \dots, 500$ and the bootstrap responses are determined as $\hat{y}_i^b = \hat{m}_2(x_i) + \hat{\epsilon}_{i,cent}^b$, $i = 1, \dots, 100$, $b = 1, \dots, 500$. Now, ordering the X -observations $x_{(1)}, \dots, x_{(100)}$ the corresponding induced ordered responses are arranged as $\tilde{y}_{(1)}, \dots, \tilde{y}_{(100)}$. The third order differences based on induced order responses are calculated as $\tilde{y}_{(i-2)} - 3\tilde{y}_{(i-1)} + 3\tilde{y}_{(i)} - \tilde{y}_{(i+1)}$, denoted by $\hat{y}_{(i)}^*$ for $i = 1, \dots, 100$ with $\tilde{y}_{(-1)} = 0$, $\tilde{y}_{(0)} = \tilde{y}_{(1)}$ and $\tilde{y}_{(101)} = \tilde{y}_{(100)}$. Based upon the bivariate observations $\{(x_{(i)}, \hat{y}_{(i)}^*)\}$ for $i = 1, \dots, n$, the test statistics $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$ are evaluated under H_0 . The bootstrap values of $T_{n,g}^b$'s ($g = KS, CM, AD, 1, 2, 3$) are the b estimates of $T_{n,g}^b$ computed for the b -th bootstrap sample $\{(x_{(1)}^b, \hat{y}_{(1)}^{*b}), \dots, (x_{(100)}^b, \hat{y}_{(100)}^{*b})\}$ on $(X, \tilde{Y}^*)$, $g = 1, \dots, 3$, $b = 1, \dots, 500$. The α -th critical value of $T_{n,g}^b$ is estimated as the point containing $500(1-\alpha)$ number of $T_{n,g}^b$ ($g = KS, CM, AD, 1, 2, 3$), here $\alpha = 0.05$.

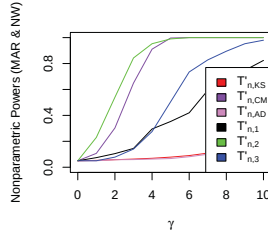
Under the alternative hypothesis, the above mentioned procedure is followed with the aid of specified conditional error distribution. The proportion of $T_{n,g}^b$'s obtained hence that is greater than the critical value obtained previously is the asymptotic power of $T_{n,g}^b$, $g = KS, CM, AD, 1, 2, 3$, $b = 1, \dots, 500$. Sections 5.4.1 and 5.4.2 display the comparative power curves of all six test

statistics.

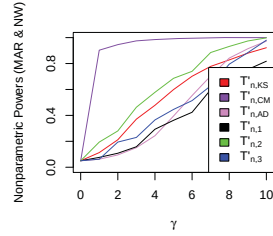
5.3.1 Under MAR when $m(X)$ is estimated by NW method



(a) Example 5.3.1

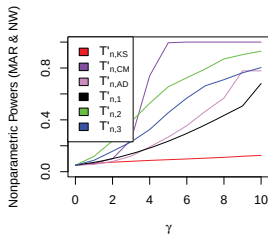


(b) Example 5.3.2

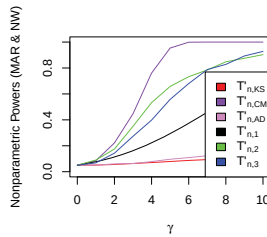


(c) Example 5.3.3

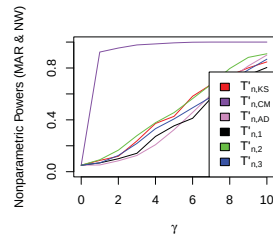
Figure 1: Powers of test statistics under 5% MAR context when $m(X)$ is estimated by NW method



(a) Example 5.3.1

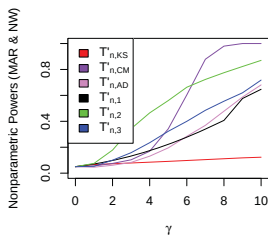


(b) Example 5.3.2

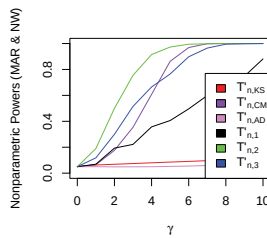


(c) Example 5.3.3

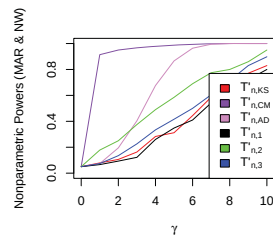
Figure 2: Powers of test statistics under 10% MAR context when $m(X)$ is estimated by NW method



(a) Example 5.3.1



(b) Example 5.3.2



(c) Example 5.3.3

Figure 3: Powers of test statistics under 20% MAR context when $m(X)$ is estimated by NW method

5.3.2 Under MAR when $m(X)$ is estimated by ILLS method

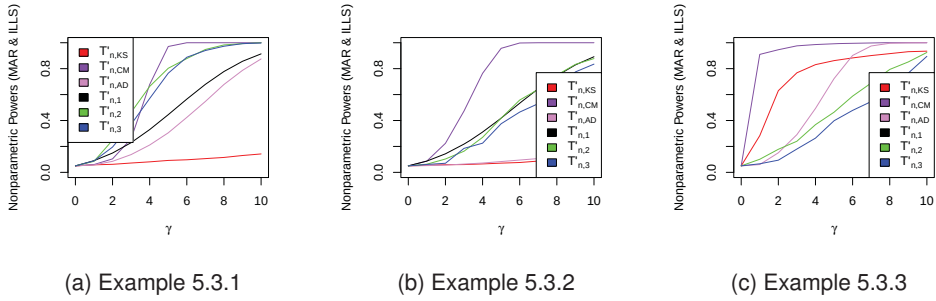


Figure 4: Powers of test statistics under 5% MAR context when $m(X)$ is estimated by ILLS method

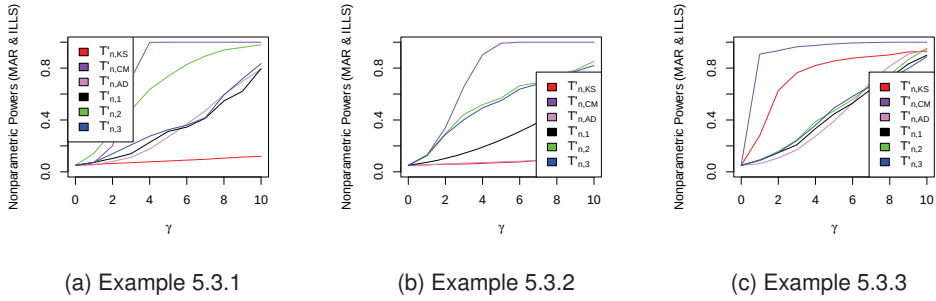


Figure 5: Powers of test statistics under 10% MAR context when $m(X)$ is estimated by ILLS method

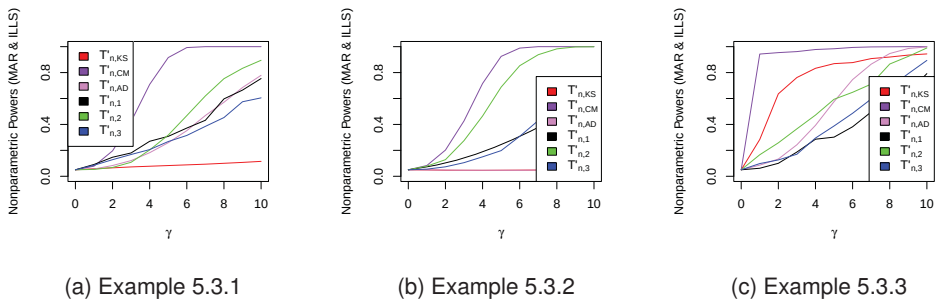


Figure 6: Powers of test statistics under 20% MAR context when $m(X)$ is estimated by ILLS method

From the power curves of the six test statistics under nonparametric regression with missing data setup, the dominance of the distance based Cramer von Mises test statistic $T'_{n,CM}$ is recognised undoubtedly. The Anderson-Darling test statistic $T'_{n,AD}$ performs reasonably good

for the conditional error distributions given by Example 5.3.1 and Example 5.3.3, but for Example 5.3.2 the performance of $T'_{n,AD}$ is highly degrading, under both MAR and MCAR type data and NW and ILLS estimation of $m(X)$.

Among the proposed three test statistics based on modified measures of association, viz. $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$, the rank based test statistic $T'_{n,2}$ emerges as the best one. But its asymptotic power performance is not as impressive as $T'_{n,CM}$. Another rank based test statistic, the sample counterpart of the traditional Kendall's τ , is consistent too, but its performance is lesser compared to the distance based test statistic $T'_{n,3}$ although with increase of missingness it gains its stability.

The proposed test statistic $T'_{n,2}$ is still the best choice among the three proposed test statistics as $T'_{n,2}$ is robust to outliers because of its structure.

The tables of powers of the six test statistics are presented in **Appendix-I**.

6 Finite sample simulation study

The theme of finite sample simulation study is to investigate the performance of the test statistics under various conditional error distributions with varying parametric values for a specific number of sample observations. Based on the conditional distribution of $\epsilon|X$, the random samples (X_i, Y_i^*) , $i = 1, \dots, n$, are drawn from the joint distribution of (X, Y^*) . The empirical distribution function of (X, Y^*) is needed instead of the actual cumulative distribution function for the generation of random samples. Here, finite sample study is performed with $n = 150$ which is sufficiently large to explain the power under H_1 , based on the asymptotic distributions of the test statistics. In our study, we consider the regression function as $m(X) = x + 3x^2 - 2x^3$. For a comparative study, we study the powers of the test statistics $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$ along with the power by the other traditional goodness of fit statistics namely, Kolmogorov -Smirnov ($T'_{n,KS}$), Cramer-von-Mises ($T'_{n,CM}$) and Anderson-Darling ($T'_{n,AD}$). Each of these test statistics are constructed from the bivariate samples on (X, Y^*) , Y^* being the third order difference of Y . We consider a pair of examples as discussed in Einmahl et al. ((Einmahl and Keilegom, 2008)). Under H_0 , the covariate X is uniformly distributed on $[0, 1]$, whereas the error ϵ follows a normal distribution with mean 0 and standard deviation 0.1. The alternative hypotheses are considered as follows.

Example 1. $H_1 : \epsilon|X = x \sim N\left(0, \frac{1+ax}{100}\right)$ with $a > 0$ being the parameter controlling variance.

Example 2. $H_1 : \epsilon|X = x \stackrel{L}{=} \frac{W_x - \omega_x}{10\sqrt{2\omega_x}}$ where $W_x \sim \chi^2_{\omega_x}$ with $\omega_x = \frac{1}{bx}$. Here, $b > 0$ controls the skewness of the distribution of $\epsilon|X$.

The asymptotic powers are evaluated in a similar manner as described in **Section 5**. Furthermore, one may see Neumeyer (Neumeyer, 2009) to determine the critical values of $T'_{n,KS}$, $T'_{n,CM}$, $T'_{n,AD}$, $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$ at level of significance α and their limiting powers for various parameter values. All necessary computations are carried out in **R 4.0.5** and the bandwidth is selected by Silverman(Silverman, 2018)'s approach as $h = 0.9 \times IQR(X) \times n^{-\frac{1}{5}}$ where IQR denotes the interquartile range. For different values of the parameters a (in Example 1) and b (in Example 2), the calculated powers are presented in **Appendix-I**. The power curves ($m(X)$

estimated by **NW** and **ILLS** methods) of the six test statistics in the 5%, 10% and 20% MAR situations are displayed below.

6.1 Power curves of test statistics under MAR for estimation of $m(X)$ by NW method

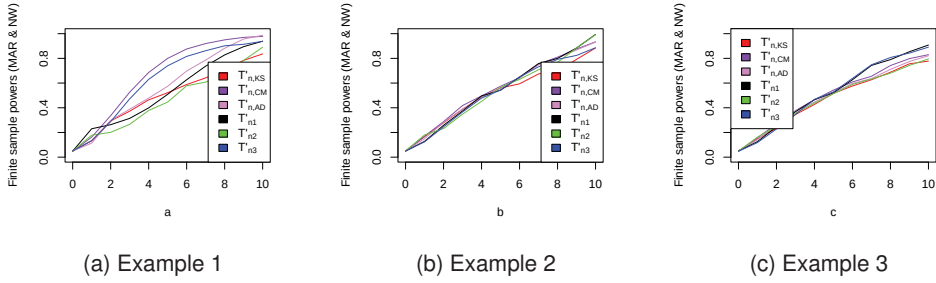


Figure 7: Finite sample power curves of test statistics under 5% MAR context when $m(X)$ is estimated by NW method

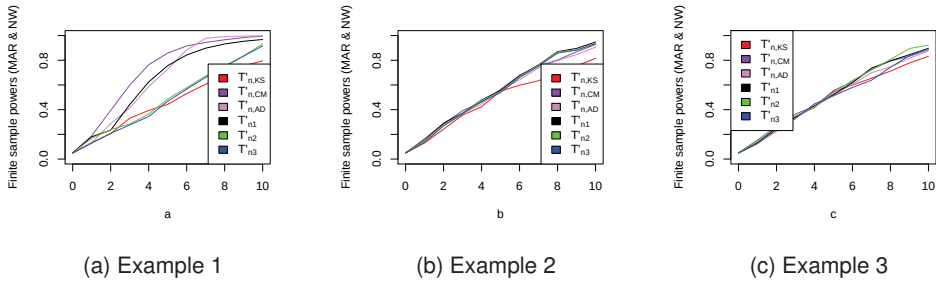


Figure 8: Powers of test statistics under 10% MAR context when $m(X)$ is estimated by NW method

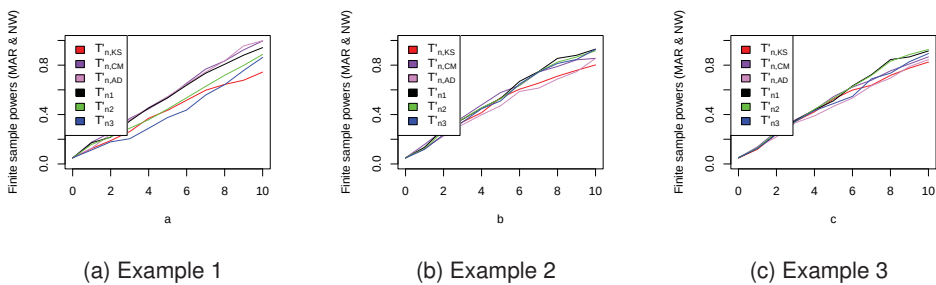


Figure 9: Powers of test statistics under 20% MAR context when $m(X)$ is estimated by NW method

6.2 Power curves of test statistics under MAR for estimation of $m(X)$ by ILLS method

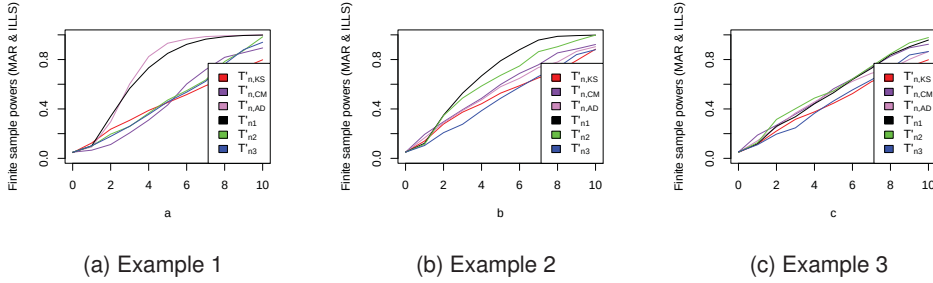


Figure 10: Finite sample power curves of test statistics under 5% MAR context when $m(X)$ is estimated by ILLS method

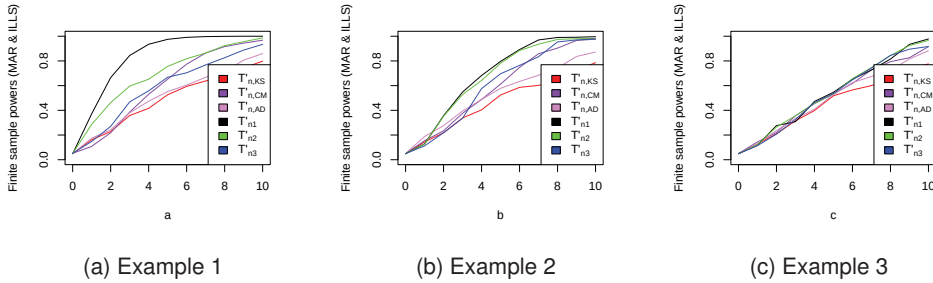


Figure 11: Powers of test statistics under 10% MAR context when $m(X)$ is estimated by ILLS method

From the above power diagrams, all test statistics perform reasonably well with the changing of missingness. The rank based test statistics $T'_{n,1}$ and $T'_{n,2}$ produce significant power performances for Example 2 and Example 3. The traditional test statistics $T'_{n,KS}$ and $T'_{n,CM}$ are more or less better performing test statistics for Example 1. In all cases, $T'_{n,KS}$ performs worse than any other test statistic.

7 Real data analysis

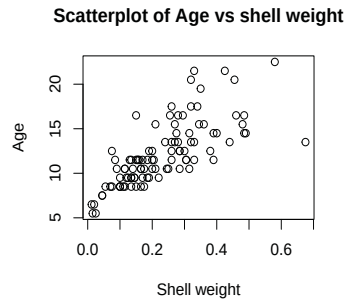
The effectiveness of the proposed statistics $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$ is judged with respect to real data with the proportion of missingness 5% and 10% respectively. A test statistic with higher p-value is considered to be more inclined to null hypothesis with comparison to the other statistics. We carry out this exercise on a Zoological data; *viz.* The Abalone data, which is freely available in the UCI Machine Learning Repository Data Set

(<https://archive.ics.uci.edu/ml/datasets/Abalone>). This data is collected by the Department of Primary Industry and Fisheries, Tasmania. It has 4177 observations each

with 10 variables extracted on different physical measurements of abalone (a common species of mollusc mainly found in warm seas). The age of abalone from 9 independent variables is sex (nominal), length (in mm) for longest shell measurement, diameter (in mm) perpendicular to length, height (in mm) with meat in shell, whole weight (in grams) of abalone, shucked weight (in grams), i.e. weight of meat, viscera weight (in grams) i.e. gut weight (after bleeding), shell weight (in grams) after being dried, rings (integer) and the single dependent variable is age (in years). We take a single regressor *i.e.* shell weight (X) and age as regressand (Y). The scatter plot of shell weight vs. age displayed below confirms that the data are almost free of outliers and a more or less positive linear trend exists between the variables.

MAR	Estimation of $m(X)$	p-values		
		$T'_{n,1}$	$T'_{n,2}$	$T'_{n,3}$
5%	NW	0.36	0.29	0.14
10%	NW	0.33	0.26	0.11
5%	ILLS	0.45	0.32	0.18
10%	ILLS	0.44	0.28	0.15

Table 1: p-values of test statistics in Abalone data in MAR context



To test the independence of X and ϵ we carry out bootstrap tests using $T'_{n,k}$'s ($k = 1, 2, 3$). First, the observed values of the test statistics under the null hypothesis of independence are obtained for the fixed sample size of 100. Then 200 bootstrap resamples are generated from the null distribution of each of $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$. The proportion of resamples that exceed the respective observed value of $T'_{n,k}$ provides the bootstrap p-value of $T'_{n,k}$, $k = 1, 2, 3$. Suppose the b^{th} resample of $T'_{n,k}$ be $T'^b_{n,k}$, $b = 1, \dots, 200$, $k = 1, 2, 3$. The estimated p-value of $T'_{n,k}$ is $\frac{\#\{T'^b_{n,k} > T^*_{n,k}\}}{200}$, $b = 1, \dots, 200$, $k = 1, 2, 3$ where $T^*_{n,k}$ is the observed value of $T'_{n,k}$ under H_0 . Here, the p-values of the tests based on $T'_{n,1}$, $T'_{n,2}$ and $T'_{n,3}$ under 5% and 10% observations missing at random are provided in Table 1.

It is noted that $T'_{n,3}$ performs with lowest p-values in all four situations where $m(X)$ is estimated either by NW or by ILLS methods. On the other hand, $T'_{n,1}$ provides best support to independence of X and ϵ with highest p-value in NW and ILLS methods.

8 Conclusion

In this article, we have made an attempt to formulate some consistent test procedures with a motive of investigating the nature of dependence between X and ϵ under missing at random response situation. The power analysis of the test statistics for a sequence of contiguous alternatives H_n proves the superiority of $T'_{n,1}$ under missingness over $T'_{n,2}$ and $T'_{n,3}$ for increasing values of γ . $T'_{n,1}$ and $T'_{n,2}$ are both rank based test statistics but missingness leaves heavier toll in $T'_{n,2}$ while for complete case $T'_{n,2}$ stands supreme. In comparison with other statistics ($T'_{n,KS}$, $T'_{n,AD}$ and $T'_{n,CM}$), $T'_{n,1}$ performs the best. Tacitly speaking, the kernels of $T'_{n,2}$ and $T'_{n,3}$ involve

four *i.i.d.* pairs (X_i, Y_i^*) , $i = 1, \dots, 4$, whereas $T'_{n,1}$ involves only two bivariate observations on (X, Y^*) . Hence, the effect of missingness is more impulsive in case of $T'_{n,2}$ and $T'_{n,3}$ with comparison to $T'_{n,1}$. Moreover, $T'_{n,2}$ is the sole competitor of $T'_{n,1}$ in the context of asymptotic power analysis for increasing value of γ . Altogether, both $T'_{n,1}$ and $T'_{n,2}$ are significant test statistics in detecting dependence between X and ϵ under MAR assumption. Hence, rank based test statistics contribute a firm foothold for testing independence (or association) between two random variables in missing data approach. The proposed test statistics are sensitive to all types of deflections of independence encompassing nonlinear dependence, even in the presence of missing responses.

Since the third order difference of Y is taken to develop the hypothesis testing procedure, one may also investigate whether a higher order difference of Y , say ' r ', is more suitable in addressing the same issue. By checking the maximization of absolute moments of various ordered differences of Y , this optimal order can be deduced for which the best power curves of the test statistics can be achieved. In addition, instead of a single nonparametric covariate, the issue of independence may be a point of concern in multiple nonparametric model with missing response by using center outward ranking induced by data depth theory. The relevant R codes are available in github.com/sthdas999/MAR-nonparametric-R-codes.

Acknowledgment

The authors are indebted to the Reviewer and the Editor for their valuable suggestions.

References

- Bergsma, W. and Dassios, A. 2014. A consistent test of independence based on a sign covariance related to kendall's tau, *Bernoulli* **20**(2): 1006 – 1028.
URL: <https://doi.org/10.3150/13-BEJ514>
- Blum, J. R., Kiefer, J. and Rosenblatt, M. 1961. Distribution free tests of independence based on the sample distribution function, *The Annals of Mathematical Statistics* **32**(2): 485 – 498.
URL: <https://doi.org/10.1214/aoms/1177705055>
- Chu, C. and Cheng, P. 1995. Nonparametric estimation of mean functionals with data missing at random, *Journal of Statistical planning and Inference* **89**: 85–99.
- Das, S. and Maiti, S. 2022. On the test of association between nonparametric covariate and error in semiparametric regression model, *Journal of the Indian Society for Probability and Statistics* **23**(2): 541–564.
- Das, S. and Maiti, S. 2024. A rank based test of independence of covariate and error in non-parametric regression with missing completely at random response situation,, *Statistics and Applications* **22**(1): 195–219.

- Dhar, S. S., Bergsma, W. and Dassios, A. 2018. Testing independence of covariates and errors in non-parametric regression, *Scandinavian Journal of Statistics* **45**(3): 421–443.
- Efromovich, S. 2011. Nonparametric regression with responses missing at random, *Journal of Statistical Planning and Inference* **141**(12): 3744–3752.
- Einmahl, J. and Keilegom, I. 2008. Tests for independence in nonparametric regression, *Statistica Sinica* **18**: 601–615.
- Einmahl, J. and V.Keilegom, I. 2008. Specification tests in nonparametric regression, *Journal of Econometrics* **143**(1): 88–102.
- Gregory, G. G. 1977. Large sample theory for u-statistics and tests of fit, *The annals of statistics* pp. 110–123.
- Härdle, W. 1990. *Applied nonparametric regression*, number 19, Cambridge university press.
- Hoeffding, W. 1994. A non-parametric test of independence, *The Collected Works of Wassily Hoeffding* pp. 214–226.
- Koning, A. J. and Protasov, V. 2003. Tail behaviour of gaussian processes with applications to the brownian pillow, *Journal of multivariate analysis* **87**(2): 370–397.
- Lee, A. 1990. *U-Statistics. Theory and practice* Marcel Dekker, Inc., New York.
- Little, T. and Rubin, D. 1987. *Analysis with missing data*, JohnWiley & Sons.
- Neumeyer, N. 2009. Testing independence in nonparametric regression, *Journal of Multivariate Analysis* **100**(7): 1551–1566.
- Nittner, T. 2003. Missing at random (mar) in nonparametric regression-a simulation experiment, *Statistical Methods and Applications* **12**(2): 195–210.
- Serfling, R. J. 2009. *Approximation theorems of mathematical statistics*, John Wiley & Sons.
- Šidák, Z., Sen, P. K. and Hájek, J. 1999. *Theory of rank tests*, Academic Press.
- Silverman, B. W. 2018. *Density estimation for statistics and data analysis*, Routledge.
- Székely, G. J., Rizzo, M. L. and Bakirov, N. K. 2007. Measuring and testing dependence by correlation of distances, *The Annals of Statistics* **35**(6): 2769–2794.
- van der Vaart, A. 2002. The statistical work of lucien le cam, *The Annals of Statistics* **30**(3): 631–682.

Appendix-I

8.1 Powers of test statistics in Example 5.3.1

λ	Powers in 5% MAR			Powers in 10% MAR			Powers in 20% MAR		
	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
1	0.098	0.086	0.100	0.082	0.074	0.086	0.092	0.080	0.088
2	0.171	0.123	0.184	0.130	0.127	0.164	0.124	0.118	0.156
3	0.275	0.165	0.276	0.204	0.166	0.224	0.200	0.170	0.218
4	0.393	0.244	0.368	0.277	0.245	0.290	0.258	0.255	0.278
5	0.528	0.315	0.458	0.368	0.324	0.392	0.316	0.354	0.376
6	0.635	0.416	0.537	0.470	0.438	0.488	0.450	0.442	0.464
7	0.748	0.478	0.662	0.574	0.526	0.583	0.530	0.538	0.560
8	0.852	0.511	0.781	0.656	0.596	0.654	0.622	0.614	0.648
9	0.887	0.652	0.854	0.712	0.620	0.746	0.700	0.648	0.736
10	0.940	0.706	0.928	0.810	0.684	0.808	0.794	0.692	0.792

Table 2: Table of powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ in 5%, 10% and 20% MAR contexts where $m(X)$ is estimated using Nadaraya-Watson method of estimation of kernel density in Example 5.3.1

λ	Powers in 5% MAR			Powers in 10% MAR			Powers in 20% MAR		
	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
1	0.078	0.078	0.094	0.077	0.074	0.086	0.092	0.076	0.080
2	0.138	0.150	0.166	0.126	0.148	0.144	0.156	0.126	0.134
3	0.236	0.234	0.252	0.224	0.210	0.234	0.242	0.204	0.228
4	0.354	0.352	0.358	0.348	0.348	0.352	0.352	0.326	0.350
5	0.471	0.468	0.484	0.460	0.458	0.470	0.474	0.426	0.464
6	0.596	0.588	0.616	0.564	0.574	0.556	0.592	0.552	0.542
7	0.648	0.644	0.658	0.636	0.632	0.620	0.658	0.616	0.616
8	0.778	0.768	0.784	0.760	0.748	0.748	0.774	0.718	0.722
9	0.858	0.846	0.860	0.844	0.836	0.830	0.826	0.790	0.826
10	0.934	0.924	0.948	0.918	0.916	0.914	0.900	0.894	0.912

Table 3: Table of powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ in 5%, 10% and 20% MAR contexts where $m(X)$ is estimated using local linear smoothing method of estimation in Example 5.3.1

8.2 Powers of test statistics in Example 5.3.2

λ	Powers in 5% MAR			Powers in 10% MAR			Powers in 20% MAR		
	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
1	0.154	0.148	0.124	0.126	0.116	0.112	0.078	0.077	0.074
2	0.295	0.260	0.238	0.266	0.250	0.242	0.178	0.111	0.108
3	0.431	0.397	0.354	0.374	0.348	0.318	0.280	0.236	0.228
4	0.626	0.498	0.442	0.562	0.472	0.448	0.398	0.336	0.286
5	0.788	0.668	0.518	0.740	0.558	0.520	0.510	0.474	0.472
6	0.864	0.772	0.641	0.864	0.638	0.604	0.674	0.510	0.508
7	0.942	0.865	0.746	0.952	0.792	0.744	0.756	0.606	0.614
8	0.985	0.926	0.872	0.982	0.884	0.824	0.862	0.794	0.782
9	0.998	0.978	0.914	0.996	0.906	0.888	0.974	0.894	0.885
10	1	0.994	0.962	0.999	0.993	0.960	0.988	0.978	0.974

Table 4: Table of powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ in 5%, 10% and 20% MAR contexts where $m(X)$ is estimated using Nadaraya-Watson method of estimation of kernel density in Example 5.3.2

λ	Powers in 5% MAR			Powers in 10% MAR			Powers in 20% MAR		
	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
1	0.092	0.087	0.082	0.099	0.092	0.088	0.086	0.082	0.078
2	0.152	0.143	0.132	0.168	0.158	0.146	0.166	0.130	0.128
3	0.246	0.226	0.196	0.236	0.230	0.212	0.210	0.206	0.200
4	0.348	0.306	0.296	0.336	0.320	0.318	0.328	0.286	0.296
5	0.458	0.416	0.414	0.426	0.404	0.374	0.412	0.346	0.365
6	0.585	0.568	0.558	0.562	0.540	0.476	0.558	0.496	0.482
7	0.700	0.682	0.670	0.680	0.676	0.584	0.658	0.582	0.552
8	0.796	0.774	0.738	0.784	0.762	0.682	0.740	0.690	0.648
9	0.878	0.864	0.842	0.864	0.848	0.738	0.868	0.748	0.730
10	0.950	0.948	0.926	0.942	0.934	0.884	0.936	0.888	0.860

Table 5: Table of powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ in 5%, 10% and 20% MAR contexts where $m(X)$ is estimated using local linear smoothing method of estimation in Example 5.3.2

8.3 Powers of test statistics in Example 1

p	a	Powers in MAR setup (NW)			Powers in MAR setup (ILLS)		
		$T_{n,KS}$	$T_{n,CM}$	$T_{n,AD}$	$T_{n,KS}$	$T_{n,CM}$	$T_{n,AD}$
5%	0	0.048	0.049	0.048	0.047	0.049	0.051
	1	0.110	0.154	0.068	0.126	0.074	0.094
	2	0.126	0.334	0.114	0.240	0.112	0.298
	3	0.232	0.538	0.262	0.310	0.210	0.602
	4	0.374	0.728	0.450	0.388	0.304	0.822
	5	0.478	0.846	0.620	0.452	0.436	0.930
	6	0.552	0.924	0.782	0.516	0.600	0.962
	7	0.618	0.966	0.868	0.584	0.722	0.984
	8	0.740	0.986	0.924	0.652	0.820	0.992
	9	0.820	0.992	0.944	0.736	0.858	0.996
	10	0.894	0.994	0.964	0.792	0.890	0.998
	a	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$
	0	0.049	0.049	0.048	0.048	0.05	0.049
	1	0.286	0.086	0.110	0.104	0.098	0.102
	2	0.442	0.222	0.224	0.342	0.198	0.182
	3	0.564	0.418	0.362	0.570	0.262	0.272
	4	0.616	0.588	0.454	0.664	0.374	0.358
	5	0.674	0.652	0.546	0.740	0.478	0.462
	6	0.714	0.720	0.644	0.848	0.554	0.540
	7	0.784	0.758	0.722	0.896	0.640	0.626
	8	0.826	0.812	0.768	0.942	0.788	0.766
	9	0.874	0.862	0.810	0.976	0.876	0.882
	10	0.916	0.900	0.852	0.998	0.988	0.942

Table 6: Powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ in 5% missingness in Example 1 where $m(X)$ is estimated by both NW and ILLS methods

8.4 Powers of test statistics in Example 2

p	a	Powers in MAR setup (NW)			Powers in MAR setup (ILLS)		
		$T_{n,KS}$	$T_{n,CM}$	$T_{n,AD}$	$T_{n,KS}$	$T_{n,CM}$	$T_{n,AD}$
10%	0	0.049	0.05	0.048	0.049	0.05	0.05
	1	0.138	0.088	0.136	0.166	0.112	0.176
	2	0.192	0.132	0.254	0.240	0.222	0.244
	3	0.228	0.242	0.532	0.362	0.388	0.384
	4	0.372	0.408	0.684	0.422	0.532	0.474
	5	0.406	0.552	0.818	0.528	0.654	0.552
	6	0.562	0.688	0.896	0.598	0.776	0.606
	7	0.624	0.770	0.950	0.642	0.866	0.670
	8	0.690	0.852	0.966	0.684	0.920	0.724
	9	0.748	0.894	0.982	0.746	0.946	0.810
	10	0.786	0.914	0.998	0.800	0.968	0.862
	a	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$
	0	0.048	0.05	0.05	0.05	0.05	0.05
	1	0.318	0.144	0.120	0.360	0.290	0.152
	2	0.594	0.214	0.248	0.634	0.458	0.274
	3	0.724	0.282	0.298	0.800	0.584	0.474
	4	0.840	0.370	0.382	0.912	0.640	0.542
	5	0.876	0.454	0.464	0.938	0.746	0.660
	6	0.900	0.512	0.596	0.966	0.822	0.710
	7	0.944	0.592	0.684	0.986	0.870	0.784
	8	0.978	0.648	0.770	0.994	0.926	0.830
	9	0.994	0.776	0.836	0.999	0.962	0.892
	10	1	0.888	0.896	1	0.990	0.948

Table 7: Powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ in 10% missingness in Example 1 where $m(X)$ is estimated by both NW and ILLS methods

p	b	Powers in MAR setup (NW)			Powers in MAR setup (ILLS)		
		$T_{n,KS}$	$T_{n,CM}$	$T_{n,AD}$	$T_{n,KS}$	$T_{n,CM}$	$T_{n,AD}$
5%	0	0.048	0.05	0.052	0.051	0.05	0.05
	1	0.070	0.246	0.156	0.140	0.194	0.166
	2	0.120	0.330	0.254	0.278	0.296	0.288
	3	0.286	0.428	0.398	0.374	0.394	0.388
	4	0.442	0.550	0.444	0.442	0.486	0.474
	5	0.584	0.688	0.554	0.528	0.598	0.580
	6	0.618	0.748	0.630	0.588	0.688	0.648
	7	0.668	0.792	0.676	0.652	0.758	0.736
	8	0.728	0.834	0.762	0.704	0.854	0.780
	9	0.776	0.852	0.798	0.792	0.886	0.866
	10	0.800	0.886	0.852	0.884	0.922	0.904
	b	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$
	0	0.05	0.05	0.05	0.05	0.05	0.05
	1	0.144	0.076	0.114	0.122	0.114	0.104
	2	0.210	0.170	0.324	0.350	0.346	0.208
	3	0.472	0.328	0.532	0.528	0.488	0.276
	4	0.682	0.510	0.700	0.668	0.586	0.386
	5	0.810	0.700	0.826	0.790	0.672	0.486
	6	0.881	0.788	0.880	0.880	0.748	0.576
	7	0.920	0.870	0.916	0.958	0.864	0.664
	8	0.943	0.924	0.940	0.988	0.904	0.734
	9	0.957	0.956	0.964	0.994	0.954	0.840
	10	0.965	0.974	0.978	0.998	0.998	0.880

Table 8: Powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ in 5% missingness in Example 2 where $m(X)$ is estimated by both NW and ILLS methods

p	b	Powers in MAR setup (NW)			Powers in MAR setup (ILLS)		
		$T_{n,KS}$	$T_{n,CM}$	$T_{n,AD}$	$T_{n,KS}$	$T_{n,CM}$	$T_{n,AD}$
10%	0	0.048	0.05	0.05	0.048	0.049	0.05
	1	0.145	0.106	0.082	0.144	0.154	0.188
	2	0.208	0.254	0.104	0.218	0.234	0.276
	3	0.318	0.446	0.164	0.338	0.376	0.394
	4	0.368	0.608	0.278	0.404	0.488	0.488
	5	0.422	0.734	0.394	0.524	0.612	0.578
	6	0.480	0.826	0.412	0.586	0.748	0.634
	7	0.536	0.868	0.540	0.602	0.858	0.684
	8	0.614	0.916	0.682	0.642	0.904	0.744
	9	0.724	0.934	0.736	0.724	0.966	0.836
	10	0.810	0.952	0.812	0.786	0.976	0.872
	b	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$	$T_{n,1}$	$T_{n,2}$	$T_{n,3}$
	0	0.05	0.05	0.05	0.05	0.05	0.05
	1	0.135	0.146	0.100	0.128	0.126	0.112
	2	0.190	0.332	0.218	0.356	0.352	0.218
	3	0.310	0.448	0.300	0.548	0.528	0.336
	4	0.491	0.682	0.440	0.682	0.642	0.576
	5	0.688	0.768	0.534	0.796	0.782	0.696
	6	0.845	0.874	0.612	0.892	0.884	0.764
	7	0.938	0.938	0.666	0.970	0.936	0.834
	8	0.980	0.958	0.708	0.990	0.974	0.954
	9	0.995	0.984	0.740	0.992	0.980	0.970
	10	0.999	0.992	0.780	0.996	0.982	0.978

Table 9: Powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ in 10% missingness in Example 2 where $m(X)$ is estimated by both NW and ILLS methods

9 Appendix-II

9.1 Proof of Theorem 5.4

$\log L_n$ can be derived as

$$\begin{aligned} \log L_n &= \log \prod_{i=1}^n \frac{f_{n;X,\epsilon^*}(x_i, e_i)}{f_{X,\epsilon^*}(x_i, e_i)} \\ &= \sum_{i=1}^n \log \left\{ \left(1 - \frac{\lambda}{\sqrt{n}}\right) f_{X,\epsilon^*}(x_i, e_i) + \frac{\lambda}{\sqrt{n}} \xi'_{X,\epsilon^*}(x_i, e_i) \right\} - \sum_{i=1}^n \log f_{X,\epsilon^*}(x_i, e_i). \end{aligned}$$

By Taylor's expansion of $\log(1+r)$ for $r > -1$ and WLLN, $\log L_n$ becomes

$$\frac{\lambda}{\sqrt{n}} \sum_{i=1}^n \left(\frac{\xi'_{X,\epsilon^*}(x_i, e_i)}{f_{X,\epsilon^*}(x_i, e_i)} - 1 \right) - \frac{\lambda^2}{2n} \sum_{i=1}^n \left(\frac{\xi'_{X,\epsilon^*}(x_i, e_i)}{f_{X,\epsilon^*}(x_i, e_i)} - 1 \right)^2 + O_P\left(\frac{1}{\sqrt{n}}\right). \quad (9.1)$$

$\therefore$ It is evident that

$$\log L_n - \frac{\lambda}{\sqrt{n}} \sum_{i=1}^n \left(\frac{\xi'_{X,\epsilon^*}(x_i, e_i)}{f_{X,\epsilon^*}(x_i, e_i)} - 1 \right) + \frac{\lambda^2}{2n} \sum_{i=1}^n \left(\frac{\xi'_{X,\epsilon^*}(x_i, e_i)}{f_{X,\epsilon^*}(x_i, e_i)} - 1 \right)^2 = O_P\left(\frac{1}{\sqrt{n}}\right) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

Define a sequence of random variables V_n as $\frac{\lambda}{\sqrt{n}} \sum_{i=1}^n \left(\frac{\xi'_{X,\epsilon^*}(x_i, e_i)}{f_{X,\epsilon^*}(x_i, e_i)} - 1 \right)$. Then using Lindeberg's condition, the asymptotic distribution of V_n is obtained as

$$\frac{V_n - E(V_n)}{\sqrt{\text{Var}(V_n)}} \xrightarrow{L} N(0, 1)$$

under H_0 , where

$$\frac{\lambda}{\sqrt{n}} \sum_{i=1}^n E \left(\frac{\xi'_{X,\epsilon^*}(x_i, e_i)}{f_{X,\epsilon^*}(x_i, e_i)} - 1 \right) = 0 \text{ and } \text{Var}(V_n) = \frac{\lambda^2}{n} \sum_{i=1}^n E \left(\frac{\xi'_{X,\epsilon^*}(x_i, e_i)}{f_{X,\epsilon^*}(x_i, e_i)} - 1 \right)^2 = \lambda^2 E \left(\frac{\xi'_{X,\epsilon^*}(x_1, e_1)}{f_{X,\epsilon^*}(x_1, e_1)} - 1 \right)^2.$$

So under H_0 ,

$$V_n \xrightarrow{L} N \left(0, \lambda^2 E \left(\frac{\xi'_{X,\epsilon^*}(x_1, e_1)}{f_{X,\epsilon^*}(x_1, e_1)} - 1 \right)^2 \right).$$

Define another sequence of random variables $\tilde{V}_n = \frac{\lambda^2}{2n} \sum_{i=1}^n \left(\frac{\xi'_{X,\epsilon^*}(x_i, e_i)}{f_{X,\epsilon^*}(x_i, e_i)} - 1 \right)^2$. It is to be noted that $\tilde{V}_n$ converges weakly to $\frac{\lambda^2}{2} E \left(\frac{\xi'_{X,\epsilon^*}(x_1, e_1)}{f_{X,\epsilon^*}(x_1, e_1)} - 1 \right)^2$. Then, $\log L_n - V_n + \tilde{V}_n = O_P(\frac{1}{\sqrt{n}}) = o_p(1)$ and by Slutsky's theorem the asymptotic distribution of the sequence of random variables $G_n = V_n - \tilde{V}_n$ converges to a random variable G that follows a normal distribution with mean $-\frac{1}{2} \lambda^2 E \left(\frac{\xi'_{X,\epsilon^*}(x_1, e_1)}{f_{X,\epsilon^*}(x_1, e_1)} - 1 \right)^2$ and variance $\lambda^2 E \left(\frac{\xi'_{X,\epsilon^*}(x_1, e_1)}{f_{X,\epsilon^*}(x_1, e_1)} - 1 \right)^2$. Therefore it is clear that $\log L_n - G_n = o_p(1)$, i.e. $\log L_n$ has the distribution same with the distribution of G_n in asymptotic sense. Notationally, $\log L_n \sim N(-\frac{1}{2}\theta, \theta)$ where $\theta = \lambda^2 E \left(\frac{\xi'_{X,\epsilon^*}(x_1, e_1)}{f_{X,\epsilon^*}(x_1, e_1)} - 1 \right)^2$. Hence, H_n is established as a sequence of contiguous alternatives due to asymptotic normality of $\log L_n$.

9.2 Proof of Theorem 5.5

1. The kernel of $T_{n,1}$ is $\text{sign}\{(x_{(i)} - x_{(j)})(y_{(i)}^* - y_{(j)}^*)\} = h((x_{(i)}, y_{(i)}^*), (x_{(j)}, y_{(j)}^*))$, $1 \leq i < j \leq n$. Define

$$\begin{aligned} h_1(u, v) &= E[h((x_{(i)}, y_{(i)}^*), (x_{(j)}, y_{(j)}^*)) | x_{(i)} = u, y_{(i)}^* = v] \\ &= E[\text{sign}\{(x_{(i)} - x_{(j)})(y_{(i)}^* - y_{(j)}^*)\} | x_{(i)} = u, y_{(i)}^* = v] \\ &= 2P[(x_{(i)} - x_{(j)})(y_{(i)}^* - y_{(j)}^*) > 0 | x_{(i)} = u, y_{(i)}^* = v] - 1. \end{aligned}$$

Note that, $E[h_1(u, v)] = E[h((x_{(i)}, y_{(i)}^*), (x_{(j)}, y_{(j)}^*))] = 0$ for $1 \leq i < j \leq n$ under H_0 . Then, $\zeta_1 = \text{Var}[h_1(u, v)] = E[h_1^2(u, v)] > 0$. Therefore, the order of degeneracy of h is 0.

2. Due to Serfling(Serfling, 2009), the variance of $T_{n,1}$ is approximated as $\frac{4\zeta_1}{n}$ for large n . Then,

$$P(|T_{n,1}| > \epsilon) < \frac{\text{Var}(T_{n,1})}{\epsilon^2}$$

by Markov's inequality as $E(T_{n,1}) = 0$ under H_0 . Now, $\text{Var}(T_{n,1}) \rightarrow 0$ as $n \rightarrow \infty$ from which it can be concluded that $T_{n,1} \xrightarrow{P} 0$ under H_0 .

3. The asymptotic normality of $\sqrt{n}(T_{n,1} - E(T_{n,1}))$ is established using the wayout furnished by Hoeffding(Hoeffding, 1994) under H_0 .

4. By applying Le Cam's third lemma it can be shown that the asymptotic distribution of $(\sqrt{n}(T_{n,1} - E(T_{n,1})), \log L_n)$ weakly converges in law to another random variable following a $N_2 \left(\begin{pmatrix} 0 \\ -\frac{\theta}{2} \end{pmatrix}, \begin{pmatrix} 4\zeta_1 & \Delta \\ \Delta & \theta \end{pmatrix} \right)$ ($\theta > 0$) distribution under H_0 . Then the asymptotic law of $\sqrt{n}(T_{n,1} - E(T_{n,1}))$ under H_n is derived as $N(\Delta, 4\zeta_1)$ with $\Delta = \lim_{n \rightarrow \infty} \text{Cov}_{H_0}(\sqrt{n}(T_{n,1} - E(T_{n,1})), \log L_n)$. The expression of Δ is determined to be same as (5.2).

9.3 Proof of Theorem 5.6

1. Let the kernel of $T_{n,2}$ is denoted by $A((X_{(1)}, Y_{(1)}^*), (X_{(2)}, Y_{(2)}^*), (X_{(3)}, Y_{(3)}^*), (X_{(4)}, Y_{(4)}^*))$ which is deduced as

$$\begin{aligned} & A((X_{(1)}, Y_{(1)}^*), (X_{(2)}, Y_{(2)}^*), (X_{(3)}, Y_{(3)}^*), (X_{(4)}, Y_{(4)}^*)) \\ &= a(X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}) \times a(Y_{(1)}^*, Y_{(2)}^*, Y_{(3)}^*, Y_{(4)}^*) \\ &= \text{sign}(|X_{(1)} - X_{(2)}| + |X_{(3)} - X_{(4)}| - |X_{(1)} - X_{(3)}| - |X_{(2)} - X_{(4)}|) \\ &\quad \times \text{sign}(|Y_{(1)}^* - Y_{(2)}^*| + |Y_{(3)}^* - Y_{(4)}^*| - |Y_{(1)}^* - Y_{(3)}^*| - |Y_{(2)}^* - Y_{(4)}^*|), \end{aligned}$$

where $I(\cdot)$ is an indicator function on $\mathbb{R}$ satisfying $I(S) = 1$ if S is true or 0 otherwise.

Define, $A_\eta((x_{(1)}, y_{(1)}^*), \dots, (x_{(\eta)}, y_{(\eta)}^*)) = E[A((x_{(1)}, y_{(1)}^*), \dots, (x_{(\eta)}, y_{(\eta)}^*), (X_{(\eta+1)}, Y_{(\eta+1)}^*), \dots, (X_{(4)}, Y_{(4)}^*))]$ and $\zeta_\eta = \text{Var}[A_\eta((X_{(1)}, Y_{(1)}^*), \dots, (X_{(\eta)}, Y_{(\eta)}^*))]$ for $\eta = 0, \dots, 4$.

Using two inequalities $|Y_{(1)}^* - Y_{(3)}^*| \leq |Y_{(1)}^* - Y_{(2)}^*| + |Y_{(2)}^* - Y_{(3)}^*|$ and $|Y_{(2)}^* - Y_{(4)}^*| \leq |Y_{(2)}^* - Y_{(3)}^*| + |Y_{(3)}^* - Y_{(4)}^*|$, one can evaluate $P(Y_{(2)}^* > Y_{(3)}^*, Y_{(1)}^* > Y_{(4)}^*)$ as

$$\begin{aligned} & P(Y_{(2)}^* > Y_{(3)}^*, Y_{(1)}^* > Y_{(4)}^*) \\ &= P(Y_{(2)}^* > Y_{(3)}^*, Y_{(1)}^* > Y_{(4)}^*, Y_{(3)}^* > Y_{(1)}^*) + P(Y_{(2)}^* > Y_{(3)}^*, Y_{(1)}^* > Y_{(4)}^*, Y_{(3)}^* \leq Y_{(1)}^*) \\ &= \frac{1}{4!} \times 6 = \frac{1}{4}. \end{aligned}$$

Similarly, $P(Y_{(2)}^* > Y_{(3)}^*, Y_{(1)}^* \leq Y_{(4)}^*) = \frac{1}{4}$. Then, $P(Y_{(2)}^* < Y_{(3)}^*) = \frac{1}{2} = P(Y_{(2)}^* > Y_{(3)}^*)$ and finally $P(|Y_{(1)}^* - Y_{(2)}^*| + |Y_{(3)}^* - Y_{(4)}^*| - |Y_{(1)}^* - Y_{(3)}^*| - |Y_{(2)}^* - Y_{(4)}^*| < 0) = \frac{1}{2}$.

Therefore $E[K((X_{(1)}, Y_{(1)}^*), (X_{(2)}, Y_{(2)}^*), (X_{(3)}, Y_{(3)}^*), (X_{(4)}, Y_{(4)}^*))]$ is obtained as 0. Again, ζ_1 is determined as $\{1 + 4P[Y_{(2)}^* > Y_{(3)}^*, Y_{(5)}^* > Y_{(6)}^*] - 2P[Y_{(2)}^* > Y_{(3)}^*] - 2P[Y_{(5)}^* > Y_{(6)}^*]\}$. For four distinct integers (i_1, i_2, i_3, i_4) with $1 \leq i_1 \neq i_2 \neq i_3 \neq i_4 \neq 7$ one can find out $P[Y_{(i_1)}^* > Y_{(i_2)}^* > Y_{(i_3)}^* > Y_{(i_4)}^*] = \frac{6}{4!} = \frac{1}{4}$ and then $P[Y_{(i_1)}^* > Y_{(i_2)}^*] = \frac{1}{2}$. Therefore, $\zeta_1 = 1 + 4 \times \frac{1}{4} - 2 \times \frac{1}{2} - 2 \times \frac{1}{2} = 0$.

Furthermore, ζ_2 is computed as

$$\begin{aligned} & 1 + 4P[Y_{(2)}^* > Y_{(3)}^*, Y_{(2)}^* > Y_{(5)}^*] - 2P[Y_{(2)}^* > Y_{(3)}^*] - 2P[Y_{(2)}^* > Y_{(5)}^*] \\ &= P[Y_{(2)}^* > Y_{(3)}^* > Y_{(5)}^* > Y_{(6)}^* > Y_{(4)}^* > Y_{(1)}^*] + 4 \times 5 \times 6 \times P[Y_{(2)}^* > Y_{(5)}^* > Y_{(3)}^* > Y_{(6)}^* > Y_{(4)}^* > Y_{(1)}^*] \\ &= 2 \times \frac{4 \times 5 \times 6}{6!} = \frac{1}{3} > 0, \end{aligned}$$

that implies that $T_{n,2}$ has order of degeneracy 1.

2. Analogous proof of **Proposition 1** in Das et al.(Das and Maiti, 2022).
3. The similar way of proof is in Serfling(Serfling, 2009)'s theorem on the limiting distribution of a degenerate U-statistic of order of degeneracy 1.
4. Proof regarding the asymptotic distribution of $n(T_{n,2} - E(T_{n,2}))$ under H_n is available in Gregory(Gregory, 1977).

9.4 Proof of Theorem 5.7

Theorem 5.7 can be proved in the same line as the proof of Theorem 5.6 done earlier.