

A distribution-free test of independence based on a modified mean variance index

Weidong Ma, Fei Ye, Jingsong Xiao & Ying Yang

To cite this article: Weidong Ma, Fei Ye, Jingsong Xiao & Ying Yang (2023) A distribution-free test of independence based on a modified mean variance index, *Statistical Theory and Related Fields*, 7:3, 235-259, DOI: [10.1080/24754269.2023.2201101](https://doi.org/10.1080/24754269.2023.2201101)

To link to this article: <https://doi.org/10.1080/24754269.2023.2201101>



© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 28 Apr 2023.



Submit your article to this journal [↗](#)



Article views: 396



View related articles [↗](#)



View Crossmark data [↗](#)

A distribution-free test of independence based on a modified mean variance index

Weidong Ma^a, Fei Ye^b, Jingsong Xiao^a and Ying Yang^a

^aDepartment of Mathematical Sciences, Tsinghua University, Beijing, People's Republic of China; ^bSchool of Statistics, Capital University of Economics and Business, Beijing, People's Republic of China

ABSTRACT

Cui and Zhong (2019), (*Computational Statistics & Data Analysis*, 139, 117–133) proposed a test based on the mean variance (MV) index to test independence between a categorical random variable Y with R categories and a continuous random variable X . They ingeniously proved the asymptotic normality of the MV test statistic when R diverges to infinity, which brings many merits to the MV test, including making it more convenient for independence testing when R is large. This paper considers a new test called the integral Pearson chi-square (IPC) test, whose test statistic can be viewed as a modified MV test statistic. A central limit theorem of the martingale difference is used to show that the asymptotic null distribution of the standardized IPC test statistic when R is diverging is also a normal distribution, rendering the IPC test sharing many merits with the MV test. As an application of such a theoretical finding, the IPC test is extended to test independence between continuous random variables. The finite sample performance of the proposed test is assessed by Monte Carlo simulations, and a real data example is presented for illustration.

ARTICLE HISTORY

Received 23 February 2022
Revised 14 March 2023
Accepted 3 April 2023

KEYWORDS

Test of independence;
asymptotic null distribution;
mean variance index;
 k -sample Anderson Darling
test statistic; concentration
type inequality

1. Introduction

As a fundamental task in statistical inference and data analysis, testing independence of random variables has been explored for decades in the literature. Based on different types of random variables, many approaches to test independence have been proposed. For instance, if one wants to test independence between two categorical random variables, then the contingency table analysis and the Pearson chi-square test can be used. If both variables are continuous, there are also many important tests, such as, Hoeffding (1948), Rosenblatt (1975), Csörgö (1985) and Zhou and Zhu (2018), among others. Testing independence between random vectors has also received much attention in recent years, for instance, Székely et al. (2007), Székely and Rizzo (2009), Heller et al. (2012), Zhu et al. (2017), Pfister et al. (2018) and Xu et al. (2020).

It is also important to test independence between a continuous variable and a categorical variable. Suppose X is a continuous variable with support \mathbb{R}_X and $Y \in \{1, \dots, R\}$ is a categorical variable with R categories. We are interested in the following test of hypothesis:

$$H_0: X \text{ and } Y \text{ are independent, versus } H_1: X \text{ and } Y \text{ are not independent.}$$



Or, equivalently,

$$\begin{aligned} H_0: F(x) = F_r(x), \text{ for any } x \in \mathbb{R}_X \text{ and } r = 1, \dots, R, \\ \text{versus } H_1: F(x) \neq F_r(x), \text{ for some } x \in \mathbb{R}_X \text{ and } r = 1, \dots, R, \end{aligned} \quad (1)$$

where $F(x) = P(X \leq x)$, $p_r = P(Y = r)$, and $F_r(x) = P(X \leq x | Y = r)$, $r = 1, \dots, R$. Thus, testing independence between X and Y is equivalent to testing the equality of conditional distributions, which is known as the k -sample problem in the literature (see e.g., Jiang et al., 2015).

Recently, Cui and Zhong (2019) proposed the mean variance (MV) test based on a new measure of dependence between X and Y , the MV index (Cui et al., 2015), to test hypothesis (1). The MV index is defined as

$$MV(X | Y) = E_X [\text{Var}_Y (F(X | Y))] = \sum_{r=1}^R p_r \int [F(x) - F_r(x)]^2 dF(x),$$

CONTACT Ying Yang  yangying@tsinghua.edu.cn  Department of Mathematical Sciences, Tsinghua University, Beijing 100084, People's Republic of China

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

where $F(x | Y) = P(X \leq x | Y)$. Given $\{(X_i, Y_i), i = 1, \dots, n\}$ with sample size n , the MV test statistic is proposed:

$$n\widehat{MV}_n(X | Y) = n \int \sum_{r=1}^R \hat{p}_r [F_n(x) - F_{rn}(x)]^2 dF_n(x),$$

where $F_n(x)$, \hat{p}_r and $F_{rn}(x)$ are the empirical counterparts of $F(x)$, p_r and $F_r(x)$, respectively. An important theoretical finding of Cui and Zhong (2019) is that when the number of categories of Y is allowed to diverge with the sample size, the standardized MV test statistic is a standard normal distribution. Cui and Zhong (2019) has argued many appealing merits of this finding. For instance, this makes it convenient for obtaining any critical value of the MV test by using an approximated normal distribution when R is large.

For any fixed $x \in \mathbb{R}_X$, dividing MV test statistic's integrand by $F_n(x)(1 - F_n(x))$ leads to the Pearson chi-square test statistic

$$\chi_n^2(x) = n \sum_{r=1}^R \hat{p}_r \frac{[F_n(x) - F_{rn}(x)]^2}{F_n(x)(1 - F_n(x))} \tag{2}$$

$$= \sum_{r=1}^R \sum_{l=1}^2 \frac{(n_{lr}(x)n - n_{l+}(x)n_{+r})^2}{n_{l+}(x)n_{+r}n}, \tag{3}$$

which is widely used in practice to test independence between the indicator function $I(X \leq x)$ and Y . Here $n_{lr}(x)$ ($l = 1, 2, r = 1, \dots, R$) are the counts in a $2 \times R$ contingency table (Table 1) determined in the following way

$$n_{1r}(x) = |\{(X_i, Y_i) : X_i \leq x \text{ and } Y_i = r\}|, \quad \text{for } r = 1, \dots, R,$$

$$n_{2r}(x) = |\{(X_i, Y_i) : X_i > x \text{ and } Y_i = r\}|, \quad \text{for } r = 1, \dots, R,$$

where $|A|$ denotes the cardinality of a set A , and $n_{l+}(x) = \sum_{r=1}^R n_{lr}(x)$, $n_{+r} = \sum_{l=1}^2 n_{lr}(x)$, for $l = 1, 2, r = 1, \dots, R$. As the Pearson chi-square test is more widely used in testing independence, we can imitate the MV test statistic to take the integral of $\chi_n^2(x)$ with respect to $F_n(x)$, and propose the following test statistic:

$$n\widehat{IPC}_n(X, Y) = \sum_{i=1}^n \sum_{r=1}^R \sum_{l=1}^2 \frac{(n_{lr}(X_i)n - n_{l+}(X_i)n_{+r})^2}{n_{l+}(X_i)n_{+r}n}$$

$$= n \sum_{r=1}^R \hat{p}_r \int \frac{[F_n(x) - F_{rn}(x)]^2}{F_n(x)(1 - F_n(x))} dF_n(x). \tag{4}$$

We call $\widehat{IPC}_n(X, Y)$ as the integral Pearson chi-squared (IPC) statistic, and $n\widehat{IPC}_n(X, Y)$ as the IPC test statistic.

It is not difficult to see that the IPC test statistic is essentially a reestablishment of the k -sample Anderson Darling test statistic proposed by Scholz and Stephens (1987). The reader is referred to He et al. (2019) and Ma et al. (2022) for some recent work on this statistics. The asymptotic null distribution of the IPC test statistic when R is fixed was established in Scholz and Stephens (1987). The promising performance of the k -sample Anderson Darling statistic (IPC test statistic) has been verified by many subsequent works in the literature and a variety of applications in practice. However, to our best knowledge, its theoretical property when the number of categories of Y is diverging remains unknown. The main goal of this paper is to fill in gaps in this area. In analogy to the MV test, we find that the IPC test also enjoys an appealing property, that is, the asymptotic null distribution of the standardized IPC test statistic when R is diverging is a standard normal distribution. This important theoretical finding allows the IPC test to share many distinguished merits with the MV test. Our work, together with Cui and Zhong (2019), establishes a solid theoretical foundation and empirical evidence for independence testing between a continuous variable and a categorical variable with a diverging number of categories. As an application of such a theoretical finding, we also extend the IPC test to test independence between two continuous random variables. The approach is carried out by slicing one of the variables on its support to get a categorical variable, and then the IPC test can be applied. We

Table 1. Empirical bivariate distribution for a fixed x .

Events	$Y = 1$	$Y = 2$...	$Y = R$	Total
$X \leq x$	$n_{11}(x)$	$n_{12}(x)$...	$n_{1R}(x)$	$n_{1+}(x)$
$X > x$	$n_{21}(x)$	$n_{22}(x)$...	$n_{2R}(x)$	$n_{2+}(x)$
Total	n_{+1}	n_{+2}	...	n_{+R}	n

allow the slicing scheme to be finer as the sample size increases, which ensures us to obtain a satisfactory test power. Slicing technique is widely used across many statistical fields, such as feature screening (Mai & Zou, 2015b; Yan et al., 2018; Zhong et al., 2021) and k -sample test (Jiang et al., 2015). It has also been used for testing independence. For instance, it is commonly seen in practice to slice two univariate variables into categorical variables and apply Pearson chi-squared test to test their independence. Please refer to Zhang et al. (2022) for more recent development of sliced independence test. Our research enriches the application of the slicing skill in the field of independence testing. The proposed approach also provides a computationally tractable way to compute the p -value efficiently. Simulation studies show that the proposed test has satisfactory test power in many scenarios.

The rest of the paper is organized as follows. Section 2 introduces some preliminaries of the IPC test. Section 3 presents the main results, including the asymptotic null distribution of the test statistic when R is diverging with the sample size. Simulation studies of the proposed test and a real data application are included in Section 4. Section 5 concludes the paper. Due to the limited space, all the technical proofs of theorems are given in Appendix.

2. Preliminaries

Let X be a continuous random variable with support \mathbb{R}_X , $Y \in \{1, \dots, R\}$ be a categorical variable with R categories. Motivated by the IPC statistic in (4), we define the following IPC index between X and Y .

$$\text{IPC}(X, Y) = \sum_{r=1}^R p_r \int_{\mathbb{R}_X} \frac{[F(x) - F_r(x)]^2}{F(x)(1 - F(x))} dF(x). \quad (5)$$

The IPC statistic is a natural estimator of the IPC index. Note that the $n_{l+}(X_i)$ in the denominator of the right-hand side of the first equality of (4) will take zero when X_i is the largest or smallest one among all $\{X_i\}_{i=1}^n$. A solution is to follow Mai and Zou (2015a) and consider the Winsorized empirical CDF

$$\tilde{F}_n(x) = \begin{cases} b, & \text{if } F_n(x) \geq b; \\ F_n(x), & \text{if } a < F_n(x) < b; \\ a, & \text{if } F_n(x) \leq a \end{cases}$$

at a predefined pair of number (a, b) . The Winsorization will cause bias in estimating the IPC index. Though such bias can automatically vanish if we let $a \rightarrow 0$ and $b \rightarrow 1$ as $n \rightarrow \infty$. However, how to properly choose a and b is beyond the scope of this paper. At the same time we notice that, if X_i is the largest or smallest one, the numerator of the first equality of (4) will also take zero. Therefore, we hereafter denote $0/0 = 0$ following the common practice in the literature (see for example, He et al., 2019; Ma et al., 2022) to avoid confusion. Then we have the following lemmas.

Lemma 2.1: Let $Y \in \{1, \dots, R\}$ be a categorical variable with R categories and X a continuous variable with support \mathbb{R}_X ,

$$\widehat{\text{IPC}}_n(X, Y) \xrightarrow{P} \text{IPC}(X, Y), \quad (6)$$

as $n \rightarrow \infty$.

Lemma 2.1 shows that $\widehat{\text{IPC}}_n(X, Y)$ is a consistent estimate of the IPC index.

Lemma 2.2: $0 \leq \text{IPC}(X, Y) < 1$ and $\text{IPC}(X, Y) = 0$ if and only if X and Y are independent.

According to Lemma 2.2, the IPC index is an effective measure of dependence between a continuous variable and a categorical variable. Thus we can construct test of independence via the IPC statistic.

Let $T_n = n\widehat{\text{IPC}}_n(X, Y)$. Note that T_n is essentially the k -sample Anderson Darling test statistic proposed by Scholz and Stephens (1987), and then we can directly derive the asymptotic null distribution of T_n .

Theorem 2.3: Suppose X is a continuous random variable and Y is a categorical random variable with a fixed class number R . Under H_0 ,

$$T_n = n\widehat{\text{IPC}}_n(X, Y) \xrightarrow{d} \sum_{j=1}^{\infty} \frac{1}{j(j+1)} \chi_j^2(R-1), \quad (7)$$

where $\chi_j^2(R-1)$'s, $j = 1, 2, \dots$, are identically and independent distributed (i.i.d.) χ^2 random variables with $R-1$ degree of freedom, and \xrightarrow{d} denotes the convergence in distribution.

Though Theorem 2.3 gives an explicit form of the asymptotic null distribution, the exact distribution of $\sum_{j=1}^{\infty} [j(j+1)]^{-1} \chi_j^2(R-1)$ is not accessible since it is a summation of infinitely many chi-square random variables. To address this issue, a widely adopted approach is to approximate $\sum_{j=1}^{\infty} \frac{\chi_j^2(R-1)}{j(j+1)}$ by $D_N + (R-1)/(N+1)$ for a sufficiently large N , where $D_N = \sum_{j=1}^N \frac{\chi_j^2(R-1)}{j(j+1)}$, and $\frac{R-1}{N+1}$ is the expectation of $\sum_{j=N+1}^{\infty} \frac{1}{j(j+1)} \chi_j^2(R-1)$. However, as a chi-square type mixture, D_N 's cumulative distribution function does not have a known closed form. In practice, we usually generate many samples from D_N and then use the empirical distribution as a surrogate of the true distribution. We can also use permutation test or bootstrap to compute the p -value for the IPC test. However, though these numerical methods are valid, they do make the IPC test less convenient for independence testing.

Lemma 2.1 declares that $\widehat{\text{IPC}}_n(X, Y)$ converges in probability to $\text{IPC}(X, Y)$, which is a new result not discussed in Scholz and Stephens (1987). Furthermore, we have a better result about the convergence rate.

Theorem 2.4: *Under the conditions of Lemma 2.1, for any $\varepsilon > 0$,*

$$P\left(\left|\widehat{\text{IPC}}_n(X, Y) - \text{IPC}(X, Y)\right| > \varepsilon\right) \leq C_1 n R \exp\left(-C_2 n \varepsilon^2 / R^2\right) \rightarrow 0, \quad (8)$$

as $n \rightarrow 0$. Here C_1 is a positive constant, and $C_2 > 0$ depends only on $\min_{1 \leq r \leq R} p_r$.

Theorem 2.4 follows directly from Theorem 3.2 in Section 3.1. The probability inequality in (8) allows us to give a lower bound of the power of the test with finite sample size. In specific, according to Theorem 2.3, we compute the critical value C_α for a given significance level $\alpha > 0$. Then under H_1 , the power is

$$\begin{aligned} P(T_n \geq C_\alpha | H_1) &= 1 - P\left(\widehat{\text{IPC}}_n(X, Y) < \frac{C_\alpha}{n} \middle| H_1\right) \\ &= 1 - P\left(\text{IPC}(X, Y) - \widehat{\text{IPC}}_n(X, Y) > \text{IPC}(X, Y) - \frac{C_\alpha}{n} \middle| H_1\right) \\ &\geq 1 - P\left(\left|\text{IPC}(X, Y) - \widehat{\text{IPC}}_n(X, Y)\right| > \text{IPC}(X, Y) - \frac{C_\alpha}{n} \middle| H_1\right) \\ &\geq 1 - C_1 n R \exp\left\{-C_2 n \left(\text{IPC}(X, Y) - \frac{C_\alpha}{n}\right)^2 / R^2\right\}. \end{aligned}$$

According to Lemma 2.2, we have $\text{IPC}(X, Y) > 0$ under H_1 . Therefore, the power of the test converges to 1 as the sample size increases to infinity. In other words, this ensures that the IPC test of independence is a consistent test.

We would like to conclude this section by introducing two relevant recent work in the literature on IPC index. The application of the dependence measure in marginal feature screening has received increasing attention. Recently, He et al. (2019) proposed a novel feature screening procedure based on the IPC index (which they referred to as the AD index) for ultrahigh-dimensional discriminant analysis where the response is a categorical variable with a fixed number of classes. The theoretical guarantee of the IPC statistic in He et al. (2019) has focused primarily on concentration inequality, rather than the asymptotic distribution. They showed that the proposed screening method is more competitive than many other existing methods. The promising numerical performance of He et al. (2019)'s method soon inspired subsequent work. Later, Ma et al. (2022) extended He et al. (2019)'s work with the help of slicing technique, and proposed an IPC index-based screening procedure which can handle many types of response variable, including continuous variable, categorical variable and discrete variable taking finite or infinite values. Especially, the slicing technique used in Ma et al. (2022) is further considered in this article to develop method for testing independence between two continuous random variables. The details are postponed in Section 3.2.

3. Main results

In this section, we allow the number of categories of Y to approach infinity with the sample size n , and consider the properties of the IPC test. Research on the categorical variable with a diverging number of categories has received increasing attention in the literature. For instance, Cui et al. (2015) established the sure screening property of the MV index for discriminant analysis with a diverging number of response classes. In their setting, they allow the number of categories R to approach infinity at a slow rate of n . And Ni and Fang (2016) also proposed an entropy-based feature screening for ultrahigh dimensional multiclass classification allowing the number of response classes to diverge. Readers are also referred to Ni et al. (2017), Yan et al. (2018), Ni et al. (2020) and Ma et al. (2022), among others, for more examples.

Here, we emphasize that it is also important to study test of independence between a continuous variable and a categorical variable with a diverging number of categories. One of its applications is to provide a feasible approach for testing independence between a continuous variable and a categorical variable taking infinite values. To be specific, suppose Y is a categorical variable taking infinite values (e.g., Poisson variable) and X is a continuous variable. To test independence between X and Y , we can define a new variable $Y' = Y \wedge R$ for some R , where $a \wedge b = \min(a, b)$. The IPC test is then applied to test independence between X and Y' , which gives us important information about whether X and Y are independent. Then a natural question is how to choose an appropriate R . A reasonable approach is to allow R to go to infinity with the sample size n so as to obtain satisfactory test power. This is one of the reasons that motivates us to study the asymptotic properties of the IPC statistic when R is diverging.

3.1. Asymptotic properties when R is diverging

In the following, we establish the large sample properties of the IPC statistic when R is diverging with the sample size n . To avoid any ambiguity, in Section 3.1, we actually consider a sequence of problems indexed by k , $k = 1, 2, \dots$. For each k , $Y_k \in \{1, \dots, R_k\}$ denotes the categorical variable with R_k categories, $p_{r,k} = P(Y_k = r)$, for $r = 1, \dots, R_k$, X_k denotes the continuous variable, and $\{(X_{ki}, Y_{ki}) : i = 1, 2, \dots, n_k\}$ is a random sample with sample size n_k from (X_k, Y_k) . The following theorem shows the asymptotic normality of the standardized test statistic if X_k and Y_k are independent for any $k = 1, 2, \dots$.

Theorem 3.1: Assume that $n_k \rightarrow \infty$ as $k \rightarrow \infty$. Let $T_{n_k} = n_k \widehat{\text{IPC}}_{n_k}(X_k, Y_k)$. If $\sqrt{R_k} / \min_{1 \leq r \leq R_k} p_{r,k} = o(n_k^{3/8})$ and $R_k \rightarrow \infty$ as $n_k \rightarrow \infty$, and X_k and Y_k are independent for $k = 1, 2, \dots$, we have

$$\frac{T_{n_k} - (R_k - 1)}{\sqrt{2 \left(\frac{\pi^2}{3} - 3 \right) (R_k - 1)}} \xrightarrow{d} N(0, 1), \quad (9)$$

as $k \rightarrow \infty$.

If $\min_{1 \leq r \leq R_k} p_{r,k} = O(n_k^{-\gamma})$ where $0 < \gamma < 3/8$, then we derive that $R_k = O(n_k^\eta)$ for some $0 < \eta < 3/4 - 2\gamma$, namely, we allow the number of categories to go to infinity with the sample size n at the relatively slow rate. Cui and Zhong (2019) also gave a similar result for the MV test with R diverging.

Let $V(R) = \sum_{j=1}^{\infty} \chi_j^2(R-1)/[j(j+1)]$ be the asymptotic null distribution in Theorem 2.3 where R is fixed. A direct application of Theorem 3.1 is that we can use a normal distribution with mean $R-1$ and variance $2(\pi^2/3 - 3)(R-1)$ to approximate the asymptotic null distribution of the IPC test (i.e., $V(R)$) when R is large. Denote $W(R) = N(R-1, 2(\pi^2/3 - 3)(R-1))$. To gain more insight into the connection between the normal distribution $W(R)$ and $V(R)$, one can notice that the mean and the variance of $V(R)$ are also $R-1$ and $2(\pi^2/3 - 3)(R-1)$, respectively. This result is a distinguished merit of the IPC test. It enables us to reduce the computational cost since it is more easy to calculate the critical value of $W(R)$ than of $V(R)$.

To further check the validity of using $W(R)$ as a surrogate for $V(R)$ to compute the critical value of the IPC test when R is large, we compare the empirical quantiles of the IPC test statistic with the theoretical quantiles of the normal distribution $W(R)$ in (9) and the asymptotic null distribution $V(R)$ in (7). We generate $Y \in \{1, \dots, R\}$ with equal probabilities and X independently from $U(0, 1)$. We consider $R = 10, 15, \dots, 35$. For each R , let $n = 40 \times R$, and we repeat the simulation 1000 times to obtain 1000 values of the IPC test statistic T_n . We report the 90% and 95% quantiles of 1000 T_n 's (denoted by empirical quantile in Table 2), as these two quantiles are most widely used in hypothesis testing. The 90% and 95% quantiles of $V(R)$ (denoted by theoretical quantile 1) and $W(R)$ (denoted by theoretical quantile 2) are also computed. The results are gathered in Table 2. The empirical quantiles are close to the theoretical quantiles of $W(R)$ even when $R = 10$, which further supports our proposed method of using the approximated normal distribution to calculate the critical value of the IPC test when R is relatively large. Looking further into the results in Table 2, we can see that T_n 's empirical quantiles seem to be almost systematically smaller than the quantiles of $V(R)$ (with the exception of the 95% quantile when $R = 35$), while

Table 2. Comparison of empirical quantiles with two theoretical quantiles.

R	90%						95%					
	10	15	20	25	30	35	10	15	20	25	30	35
Empirical quantile	11.930	17.726	23.448	28.915	34.379	39.655	13.099	19.114	24.943	30.460	36.149	41.802
Theoretical quantile 1	12.027	17.806	23.401	28.923	34.425	39.785	13.206	19.178	24.995	30.636	36.298	41.592
Theoretical quantile 2	11.927	17.651	23.253	28.780	34.255	39.690	12.757	18.686	24.459	30.135	35.744	41.303

larger than the quantiles of $W(R)$ (both by a very small amount). Note that the asymptotic distribution $V(R)$ can be viewed as a chi-square-type mixture. Such chi-square-type mixture follows an asymmetrical, positively skewed (or right-skewed) distribution, in which the left tail is shorter while the right tail is longer. To be specific, the skewness of $V(R)$ is $E(V(R) - EV(R))^3 / \text{Var}(V(R))^{3/2} = (80 - 8\pi^2) / \{(2\pi^2/3 - 6)^{3/2}(R - 1)^{1/2}\} > 0$, which will tend to zero as R goes to infinity. While the normal distribution $W(R)$ is symmetric, its skewness is 0. Since $V(R)$ is a better approximation of the exact distribution of T_n , it makes sense that the 90% and 95% quantiles of both the T_n 's empirical distribution and $V(R)$ will be slightly larger than that of $W(R)$. It is also interesting that the T_n 's empirical quantiles fall between the quantiles of $V(R)$ and the quantiles of $W(R)$. This may implicate that the skewness of the exact distribution of T_n seems to be smaller than that of $V(R)$.

We further compare the empirical null distribution with $W(R)$. Still generate $Y \in \{1, \dots, R\}$ with equal probabilities and X independently from $U(0, 1)$. Consider four scenarios: (a) $R = 5, n = 100 \times R = 500$; (b) $R = 10, n = 80 \times R = 800$; (c) $R = 20, n = 40 \times R = 800$; (d) $R = 50, n = 30 \times R = 1500$. We run the simulation 100000 times for each scenario to obtain 100000 values of the IPC test statistic T_n . Then we compare the empirical distribution of the standardized IPC test statistic $[T_n - (R - 1)] / \sqrt{2(\pi^2/3 - 3)(R - 1)}$ with the standard normal distribution $N(0, 1)$ in Figure 1. In scenario (a) when $R = 5$ is too small, the empirical density curve of the standardized IPC test statistic deviates to some extent from the normal density function, even though the sample size $n = 500$ is large. Also, when $R = 5$, the empirical density is positively skewed, with more values clustered around the left tail while the right tail is slightly longer. The empirical density curve, however, is very well matched to the

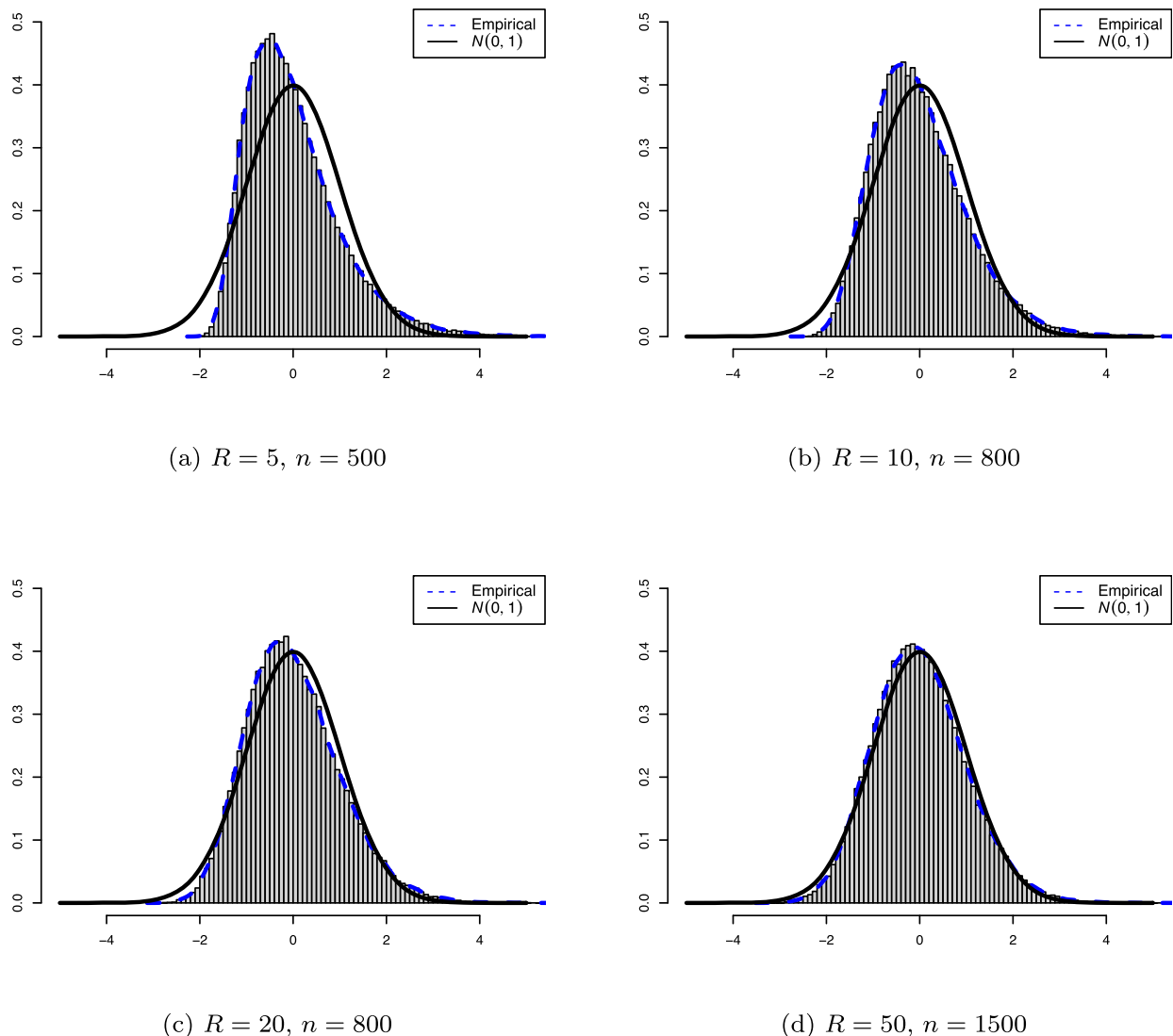


Figure 1. Comparing the empirical distribution of the standardized IPC test statistic with the standard normal distribution. The blue broken line represents the empirical density and the black solid line represents the standard normal density. The empirical density is a kernel density estimate using Gaussian kernels based on 100000 values of T_n . In each panel, the histogram of the standardized IPC test statistic is also displayed. (a) $R = 5, n = 500$. (b) $R = 10, n = 800$. (c) $R = 20, n = 800$ and (d) $R = 50, n = 1500$.

standard normal density curve when R increases, such as in scenario (c) when $R = 20$. This further emphasizes that R should be large enough (say, larger than 10) to ensure the normal approximation in Theorem 3.1 to hold.

The following theorem allows us to bound the deviation of the IPC statistic when R is diverging, which is parallel to Theorem 3.1 in Ma et al. (2022).

Theorem 3.2: Suppose $R_k = O(n_k^\eta)$ for some $0 \leq \eta < 1/2$ and there exists a positive constant c_1 such that $c_1/R_k \leq p_{r,k}$ for $r = 1, \dots, R_k, k = 1, 2, \dots$. Then for any $\varepsilon \in (0, 1)$,

$$P\left(\left|\widehat{\text{IPC}}_{n_k}(X_k, Y_k) - \text{IPC}(X_k, Y_k)\right| > \varepsilon\right) \leq C_1 n_k R_k \exp\left(\frac{-C_2 n_k \varepsilon^2}{R_k^2}\right), \quad (10)$$

where C_1 is a positive constant and $C_2 > 0$ depends only on c_1 .

Remark 3.1: He et al. (2019) has also established a concentration inequality for the IPC statistic. However, their theoretical guarantee relies on a fixed number of categories (i.e., $\eta = 0$). Thus, Theorem 3.2 is different to Lemma 4 in He et al. (2019).

The condition $c_1/R_k \leq p_{r,k}$ for $r = 1, \dots, R_k$, which is also used in Cui et al. (2015) and Cui and Zhong (2019), requires that the proportion of each category of Y_k can not be too small. Indeed, the condition can be relaxed in a way that c_1 is allowed to tend to 0 at a slow rate. Specifically, if we assume $c_1 = o(n_k^{-\tau})$ for some $0 < \tau < 1/2 - \eta$, then the probability in (10) will still converge to zero, but the convergence rate will be relatively slower. Note that Theorem 2.4 is a special case of Theorem 3.2 when $\eta = 0$, i.e., R_k is fixed, and the condition on $p_{r,k}$ is automatically satisfied.

3.2. Extension of the IPC test

A natural application of Theorem 3.1 is to extend the IPC test to test independence between two continuous variables via the slicing technique. Consider two continuous random variables X and Z . Without loss of generality, we assume that the supports of X and Z are \mathbb{R} . We define a partition of the support of Z with a given positive integer R :

$$\mathbb{S} = \{[q_{r-1}, q_r) : q_{r-1} < q_r, r = 1, \dots, R\}, \quad (11)$$

where $q_0 = -\infty, q_R = \infty$. Each interval $[q_{r-1}, q_r)$ is called a slice in the literature (Mai & Zou, 2015b; Yan et al., 2018). And a new random variable can be accordingly defined as $Y^{\mathbb{S}} = r$ if and only if $q_{r-1} \leq Z < q_r$ for $r = 1, \dots, R$. The IPC test can be applied to test independence between X and $Y^{\mathbb{S}}$. If the distribution of Z is known, we suggest a uniform slicing to partition Z such that $q_r = F_Z^{-1}(r/R)$ for $r = 1, \dots, R$, where $F_Z(z)$ is the cumulative distribution function of Z . However, in practice, $F_Z(z)$ is usually unknown. But given observations $\{(X_i, Z_i), i = 1, \dots, n\}$ with sample size n , we can use $\hat{q}_r = \widehat{F}_Z^{-1}(r/R)$ to estimate q_r for $r = 1, \dots, R$, where $\widehat{F}_Z(z)$ is the empirical distribution of Z . And $\widehat{\mathbb{S}} = \{[\hat{q}_{r-1}, \hat{q}_r), r = 1, \dots, R\}$ is regarded as an intuitive uniform slicing scheme (Yan et al., 2018). We also define $Y_i^{\widehat{\mathbb{S}}} = r$ if and only if $Z_i \in [\hat{q}_{r-1}, \hat{q}_r)$ for $r = 1, \dots, R, i = 1, \dots, n$. Now, we compute $\widehat{\text{IPC}}_n(X, Y^{\widehat{\mathbb{S}}})$ as

$$\widehat{\text{IPC}}_n(X, Y^{\widehat{\mathbb{S}}}) := \sum_{r=1}^R \tilde{p}_r \int \frac{[F_n(x) - \tilde{F}_{rn}(x)]^2}{F_n(x)(1 - F_n(x))} dF_n(x),$$

where $\tilde{p}_r = \frac{1}{n} \sum_{i=1}^n I(Y_i^{\widehat{\mathbb{S}}} = r) = 1/R$, and $\tilde{F}_{rn}(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x, Y_i^{\widehat{\mathbb{S}}} = r) / \tilde{p}_r$ is the empirical conditional distribution of X based on the subjects for which $\hat{q}_{r-1} \leq Z_i < \hat{q}_r$. We reject hypothesis $H_0: X$ and Z are independent, if $(n\widehat{\text{IPC}}_n(X, Y^{\widehat{\mathbb{S}}}) - R + 1) / \sqrt{2(\pi^2/3 - 3)(R - 1)} \geq \Phi^{-1}(1 - \alpha)$ for some given significance value $\alpha \in (0, 1)$, where $\Phi(x)$ is the standard normal distribution function.

Obviously, it is important to choose an appropriate R for testing independence. If R is too large, then the sample size in each slice is too small, making the estimate of the IPC index inaccurate. And if R is too small, then much information of Z may be lost, making the test power poor. In the slicing literature (Mai & Zou, 2015b; Yan et al., 2018; Zhong et al., 2021), a common choice is to set $R = \lfloor \log n \rfloor$, where $\lfloor x \rfloor$ is the integer part of x . And according to Theorem 3.1, we can also choose $R < \lfloor n^{1/4} \rfloor$. In practice, we recommend choosing $R = \lfloor n/k \rfloor$ for some $20 \leq k \leq 50$, so that the sample size in each slice is about 20 to 50.

3.3. Comparison with the MV test

In this subsection, we would like to discuss the advantages of the IPC test compared to the MV test. As explained in Cui and Zhong (2019), the MV index can be considered as the weighted average of Cramér-von Mises distances between $F_r(x)$, the conditional distribution of X given $Y = r$, and $F(x)$, the unconditional distribution function of X . Note that the IPC index can be viewed as a modification of the MV index by adding a weight function $\{F(x)(1 - F(x))^{-1}\}$. Such weight function is large for $F(x)$ near 0 and 1, and smaller near $F(x) = 1/2$. Hence, the IPC test emphasizes more on the difference between $F_r(x)$ and $F(x)$ near the tail of $F(x)$. As it is known, $F_r(x) - F(x) = \sum_{j=1}^R p_j(F_r(x) - F_j(x))$. Accordingly, the IPC test is more sensitive to tail differences among the conditional distributions. In the following, we consider the test of independence between a continuous random variable and a categorical variable with a relatively large number of classes (i.e., R is large) and the test of independence for two continuous random variables, and further illustrate the IPC test’s sensitivity to differences in the tails of the conditional distributions through numerical simulations.

1. *When R is large or is allowed to diverge.* In this case, we recommend using a normal distribution to approximate the IPC test’s null distribution due to Theorem 3.1. It is not surprising that given a large R , IPC test still retains sensitivity to tail differences when using a normal distribution instead of $V(R)$ to calculate p -value. The following example is used to illustrate this issue.

Let $Y \in \{1, \dots, 20\}$ with $P(Y = r) = 1/20$, for $r = 1, \dots, 20$. When $Y = r$, generate $X \sim BW + (1 - B)V_r$, where $B \sim \text{Binomial}(1, p)$, W and V_r are independent, $W = N(0, 1)$ and $V_r = N(10 + r, 1)$. To intuitively gain some understanding of our simulation setting, set $p = 0.8$. We draw the conditional distributions of X given $Y = 1$ and $Y = 5$, respectively in Figure 2. It is easy to see that the conditional distributions differ from each other only at their right tails. We choose the sample size $n = 400$, and $p = 0.7, 0.75, 0.8, 0.85, 0.9$. We apply the IPC test and the MV test, and compute the p -values for these two tests by using their approximated normal distributions. The empirical powers of these two tests based on 500 replicates at the significance level $\alpha = 0.05$ are presented in Table 3. To further validate the robustness of the IPC test against heavy-tails, we further consider $W \sim t(1)$ in the above setting. The empirical powers are also shown in Table 3. A larger p indicates that the differences among the conditional distributions occur in a more extreme right tail end, and thus are more difficult to detect the dependence between X and Y . We can see from Table 3 that the IPC test is significantly more powerful than the MV test when $p < 0.9$. When $p = 0.9$, neither the IPC nor the MV has sufficient statistical power to detect the dependence between X and Y . The simulation validates that the IPC test has a better power to tail differences among the conditional distributions. In Example 4.1 we will compare with other existing methods to further validate the IPC test’s sensitivity towards tail differences.

2. *Testing independence between continuous random variables.* We follow the notation in Section 3.2. Let X and Z be two continuous random variables. It is natural to expect that the IPC test will be more powerful than the MV test to detect the tail differences among the conditional distribution of X given Z . Consider a straightforward extension of the IPC index in (5) and define the following index between X and Z :

$$\text{IPC}(X, Z) = \int \int \frac{[F(x | Z = z) - F(x)]^2}{F(x)(1 - F(x))} dF(x) dF_Z(z), \tag{12}$$

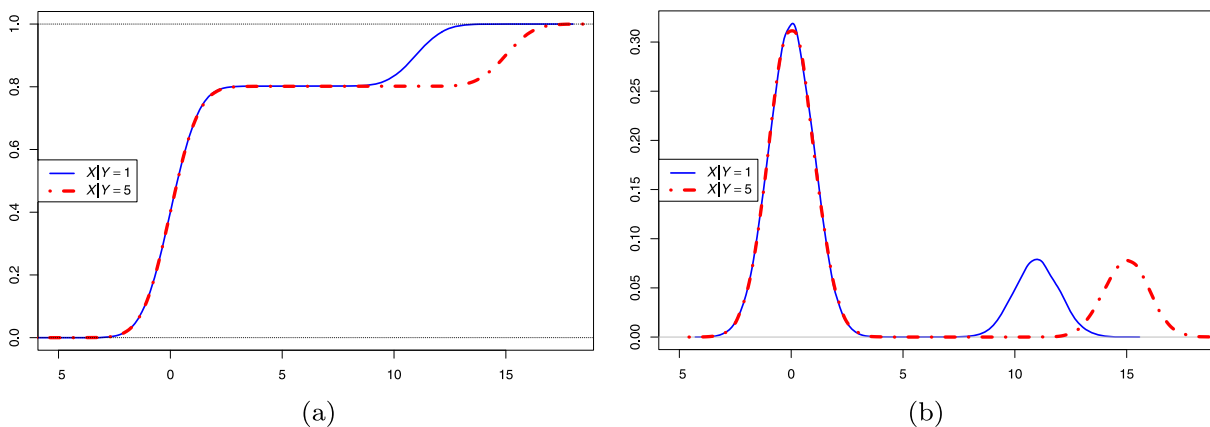


Figure 2. Panel (a) shows the pair of conditional distributions. The blue solid line represents the conditional distribution of X given $Y = 1$, that is, $BN(0, 1) + (1 - B)N(11, 1)$ where $B \sim \text{Binomial}(1, 0.8)$; and the red dot-dash line represents the conditional distribution of X given $Y = 5$, that is, $BN(0, 1) + (1 - B)N(15, 1)$ where $B \sim \text{Binomial}(1, 0.8)$. Panel (b) shows the corresponding conditional density functions.

Table 3. Test of independence between a continuous variable and a categorical variable with $R = 20$ classes.

p	$W \sim N(0, 1)$					$W \sim t(1)$				
	0.70	0.75	0.80	0.85	0.90	0.70	0.75	0.80	0.85	0.90
IPC	0.984	0.838	0.498	0.226	0.096	0.972	0.816	0.440	0.230	0.102
MV	0.654	0.338	0.166	0.074	0.062	0.692	0.370	0.176	0.118	0.064

Notes: The empirical powers of the IPC and MV tests at significance level $\alpha = 0.05$ against different p are computed based on 500 replications. A larger p indicates that the differences among the conditional distributions occur in a more extreme tail end.

Table 4. Test of independence between two continuous random variables.

p	$W \sim N(0, 1)$					$W \sim t(1)$				
	0.70	0.75	0.80	0.85	0.90	0.70	0.75	0.80	0.85	0.90
IPC	0.996	0.912	0.624	0.266	0.126	0.992	0.828	0.560	0.238	0.106
MV	0.762	0.430	0.202	0.110	0.060	0.794	0.400	0.190	0.082	0.074

Notes: The empirical powers of the IPC and MV tests at significance level $\alpha = 0.05$ against different p are computed based on 500 replications. A larger p indicates that the differences among the conditional distributions occur in a more extreme tail end.

where $F(\cdot | Z = z)$ is the conditional distribution of X given $Z = z$, and $F(x)$ and $F_Z(z)$ are the distributions of X and Z , respectively. Given a positive integer R and a corresponding uniform slicing scheme \mathbb{S} defined as in (11) with $q_r = F_Z^{-1}(r/R)$ for $r = 1, \dots, R$, recall that $Y^{\mathbb{S}} = r$ if and only if $q_{r-1} \leq Z < q_r$. Under certain mild conditions, Ma et al. (2022) has shown that $\text{IPC}(X, Y^{\mathbb{S}}) \rightarrow \text{IPC}(X, Z)$, as $R \rightarrow \infty$.

From (12), again, we have some insights that the IPC test of independence emphasizes more on the difference between $F(x | Z = z)$ and $F(x)$ near the tail of $F(x)$. We use a toy sample to further illustrate this issue. Generate $Z \sim \text{Unif}(4, 6)$, and generate $X = BW + 5(1 - B)Z$, where $B \sim \text{Binomial}(1, p)$. We still consider two settings of W : (i) $W \sim N(0, 1)$ and (ii) $W \sim t(1)$. Choose the sample size $n = 400$, and $p = 0.7, 0.75, 0.8, 0.85, 0.9$. We follow the step in Section 3.2 and choose $R = 20$ to conduct the test of independence. Table 4 presents the empirical powers of IPC and MV tests based on 500 replicates at the significance level $\alpha = 0.05$. IPC test outperforms the MV test in these settings. Note that when $p = 0.8$, the MV test is almost invalid. However, the IPC test still has a reasonably acceptable power.

4. Numerical studies and data application

4.1. Numerical studies

In this section, we assess the finite-sample performance of the IPC test by comparing with some powerful methods proposed in recent years: the MV test (Cui & Zhong, 2019), the distance correlation (DC) test (Székely et al., 2007), the HHG test (Heller et al., 2012, 2016) and the Hilbert-Schmidt independence criterion (HSIC) test (Gretton et al., 2005, 2007; Pfister et al., 2018). The R packages *energy*, *HHG*, and *dHSIC* are used to implement the DC test, the HHG test and the HSIC test, respectively. Note that the DC test can not be directly applied to a categorical variable, so in our simulations we will transfer a categorical variable with R categories into a random vector with $R-1$ binary dummy variables and apply *dcov.test* to this dummy vector instead of the original data. For the DC, HHG, and HSIC tests, the permutation test with $K = 200$ is used to calculate the p -value.

Example 4.1: In this example, we evaluate the performance of IPC test for the large- R case. Let $R = 15$, and we consider the following two cases.

Model 1.1. Generate $Y \in \{1, \dots, 15\}$ with equal probabilities. And let $\mu = (\mu_1, \dots, \mu_{15})$, where $\mu_{5j+l} = l + 1$ for $1 \leq l \leq 3$, and $\mu_{5j+l} = l + 2$ for $l = 4, 5, j = 0, 1, 2$. For $Y = r$, generate $X = BU + (1 - B)(V_{\mu_r} + 20)$, where $B \sim \text{Binomial}(1, p)$, $U \sim \text{Unif}(-20, 20)$, $V_{\mu_r} \sim \text{Beta}(3, \mu_r)$.

Model 1.2. Generate $Y \sim \text{Unif}(0, 4)$. And let $X \sim BU + (1 - B)W$, where $W \sim \text{Unif}(\cos(Y\pi) + 21, \cos(Y\pi) + 24)$. B, U are the same as in Model 1.1.

Let $n = 400$. In Model 1.2, we uniformly slice Y into a categorical variable with $R = 15$ classes in order to apply the IPC and MV tests. Let p vary from 0 to 1 in both two models. We compute the p -value for the IPC test by using the asymptotic distribution in Theorem 3.1. The empirical power of each test based on 500 simulations at the significance level $\alpha = 0.05$ is shown in Figure 3. Note that, when $p = 1$, X is independent with Y in both models. We deliberately report the results, i.e., the type I error rates of each test, in Table 5. The type I error rates of the IPC test (and other tests) are close to the nominal significance level $\alpha = 0.05$, which further supports Theorem 3.1. Figure 3 clearly shows that the IPC test outperforms other competitors. And the power differences between IPC test and MV test exceed 0.25 when $p = 0.6$ for both models.

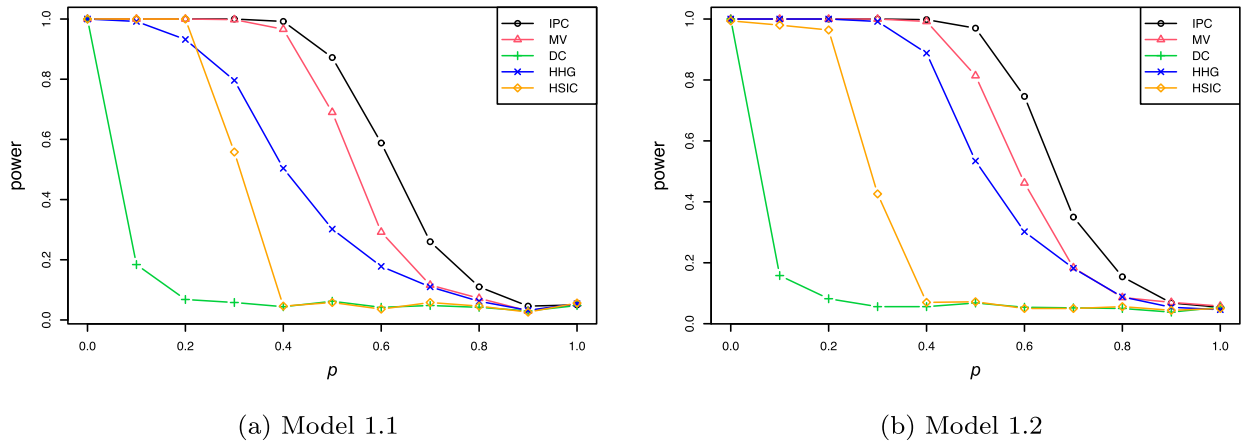


Figure 3. Comparison of powers of several tests of independence against different p in Example 4.1. In each case, 500 simulations are used to estimate the power. (a) Model 1.1 and (b) Model 1.2.

Table 5. Empirical type I error rates at the significance level $\alpha = 0.05$ in Example 4.1.

n	Model 1.1					Model 1.2				
	IPC	MV	DC	HHG	HSIC	IPC	MV	DC	HHG	HSIC
400	0.050	0.042	0.048	0.056	0.044	0.054	0.058	0.054	0.046	0.05

Looking further into the models considered in this example. In both Model 1.1 and Model 1.2, the conditional distributions of X given Y differ from each other only in their right tails when $p > 0.5$. A larger p indicates that the conditional distribution functions differ from each other in a more extreme tail end. And when $p = 1$, X and Y are independent. Thus it could be more difficult to detect the dependence between X and Y for a larger $p < 1$. As a result, we can see from Figure 3 that the power of each test decreases with the growth of p . Among the tests considered, the DC test and the HSIC test perform the worst in both models. Their powers rapidly decrease to near 0 when p increases to 0.4. It can be seen that the IPC test and the MV test have a better performance compared to other tests. Furthermore, the IPC test has a significant higher power than the MV test when p is between 0.6 and 0.8 in both models. This further supports our observation in Section 3.3 that the IPC test is more sensitive to tail differences.

Example 4.2: This example considers a Poisson regression model. Let $Z \sim \text{Poisson}(u)$, where $u = \exp(0.8X_1 - 0.8X_2 + \log 4)$, $(X_1, X_2) \sim N((0, 1)^T, \Sigma)$, $\Sigma = (0.5^{|i-j|})_{1 \leq i, j \leq 2}$. Let $Y = Z$ if $Y \leq 8$; otherwise $Y = 9$. As a consequence, Y is a 10-categories variable. Consider $n = 100, 150, \dots, 300$. We apply the testing methods to test independence between Y and X_1 , Y and X_2 , respectively. And the asymptotic normal distribution in Theorem 3.1 is used to compute p -value for the IPC test. The empirical powers of each test based on 500 replications are summarized in Table 6. The IPC test has most excellent power performances in all settings. The HHG test and the HSIC test perform poorly when the sample size $n \leq 150$.

The power of the IPC test is only slightly higher than that of the MV test. However, it is significantly higher than that of HHG and HSIC. The DC test has moderate performance, inferior to the MV test, but better than HSIC.

Example 4.3: In this example, we evaluate the power of the IPC test in testing independence between continuous variables. Simulations are carried out with sample size $n = 400$. We choose $R = 15$ to implement the IPC test. Generating $Z \sim \text{Unif}(-2, 2)$, the following alternatives are considered.

Table 6. Empirical powers of each test at the significance level $\alpha = 0.05$ against the sample sizes in Example 4.2.

n	X_1					X_2				
	IPC	MV	DC	HHG	HSIC	IPC	MV	DC	HHG	HSIC
100	0.708	0.686	0.626	0.342	0.422	0.724	0.714	0.648	0.350	0.452
150	0.918	0.910	0.856	0.536	0.652	0.918	0.908	0.860	0.556	0.664
200	0.990	0.988	0.968	0.746	0.840	0.986	0.978	0.964	0.718	0.830
250	0.998	0.992	0.992	0.872	0.932	0.996	0.990	0.990	0.880	0.928
300	1.000	0.998	0.992	0.908	0.954	0.998	0.998	0.994	0.918	0.968

- (a) Linear: $X = Z/2 + 12\gamma\varepsilon$, where γ is a noise parameter ranging from 0 to 1, and $\varepsilon \sim \text{Unif}(-2, 2)$ is independent of Z .
- (b) Quadratic: $X = (\frac{1}{2}Z)^2 + 4.5\gamma\varepsilon$.
- (c) Step function: $X = f(Z) + 25\gamma\varepsilon$, where f takes value 2 in interval $[-2, -1) \cup [0, 1)$ and value -2 in $[-1, 0) \cup [1, 2]$.
- (d) W -shaped: $X = |Z + 1|I(Z < 0) + |Z - 1|I(Z \geq 0) + 4\gamma\varepsilon$.
- (e) Sinusoid: $X = \cos(4\pi Z) + 5\gamma\varepsilon$.
- (f) Ellipse: $X = \sqrt{1 - (Z/2)^2} + 1.5\gamma\varepsilon$.

To conduct the IPC test and the MV test, we uniformly slice Z into a categorical variable Y with $R = 15$ classes. The choices of the coefficients in all of the above are to make sure that a full range of powers can be observed when γ varies from 0 to 1. In addition to the test methods mentioned before, in this example, we further consider a comparison with a new test, the modified Blum-Kiefer-Rosenblatt (MBKR) test (Zhou & Zhu, 2018) which is applied for testing independence between continuous variables. Figure 4 presents the empirical power of each test based on 500 simulations at the significance level $\alpha = 0.05$. We see from the figure that the IPC test performs quite excellent when the relationship has an oscillatory nature (the W -shaped and the sinusoid). It is also better than other competitors for the step function, and comparably well to the MBKR test for the quadratic function. However, the IPC test has poor performance compared to other tests for some smooth alternatives: the linear and the ellipse. For the linear function, the MBKR test has the highest performance. IPC test has comparable performance to HSIC. For the ellipse function, HHG test has the highest power and DC test performs the poorest. The performance of the IPC test, on the other hand, is moderate.

We give an intuitive explanation here for the excellent performance of the IPC test in detecting oscillatory relationships. Denote $X | Y = r$ as the random variable which follows the conditional distribution of X given $Y = r$. By simple calculation, we find that if X and Z have an oscillatory relationship, then the variances of $X | Y = r$ differ from each other more significantly. As a comparison, if X and Z have a linear relationship, then $\text{Var}\{X | Y = 1\} = \dots = \text{Var}\{X | Y = 15\}$. Consequently, the IPC test has a higher test power when there is an oscillatory relationship between X and Z .

4.2. Real data application

Example 4.4: We consider a data set from AIDS Clinical Trials Group Protocol 175 (ACTG175), which is available from the R package *speff2trial*. Many researchers have studied this data set, such as Tsiatis et al. (2008), Zhang et al. (2008), Lu et al. (2013) and Zhou et al. (2020). The data set contains 2139 HIV-infected subjects. And all the subjects were randomized to four different treatment groups with equal probability: zidovudine (ZDV) monotherapy, ZDV+didanosine (ddI), ZDV+zalcitabine, and ddI monotherapy. In addition to the treatment indicators indicating which group each subject was assigned to, the data contains many other important variables, such as the CD4 count at 20 ± 5 weeks post-baseline (CD420), the CD4 count at baseline (CD40), the history of intravenous drug use, et al.

In this study, in order to get more elaborated results, we only consider the subjects under ZDV+zalcitabine groups (524 subjects) in the following analysis. The goal of our study is to check whether the treatment effect under ZDV + zalcitabine groups is dependent on some other covariates. Following Hammer et al. (1996) and Tsiatis et al. (2008), we use the change from baseline to 20 ± 5 weeks in CD4 cell count, i.e., $CD420 - CD40$, to measure the treatment effect. And the covariates of interest are listed below: history of intravenous drug use (0 =no, 1 =yes), gender (0 =female, 1 =male), antiretroviral history (0 =naive, 1 =experienced), age, and CD8 count at baseline (CD80). Thus the first three covariates are categorical, and the last two are continuous covariates. Let $X = CD420 - CD40$, and then there are 5 candidates Y . The null hypotheses are listed as follows.

- H_0^1 : X is independent of Y with $Y =$ history of intravenous drug use;
- H_0^2 : X is independent of Y with $Y =$ gender;
- H_0^3 : X is independent of Y with $Y =$ antiretroviral history;
- H_0^4 : X is independent of Y with $Y =$ age;
- H_0^5 : X is independent of Y with $Y =$ CD8 count at baseline.

We apply the IPC, MV, DC, HHG and HSIC tests to these five hypotheses. The permutation test with $K = 1000$ permuted times is used for DC, HHG and HSIC tests to compute the p -values. And for H_0^4 and H_0^5 , we follow the approach in Section 3.2 to slice Y into a categorical variable with 15 classes to implement the IPC test and MV test. Table 7 summarizes the p -values of each test. If we only consider the significance level $\alpha = 0.05$, then we observe

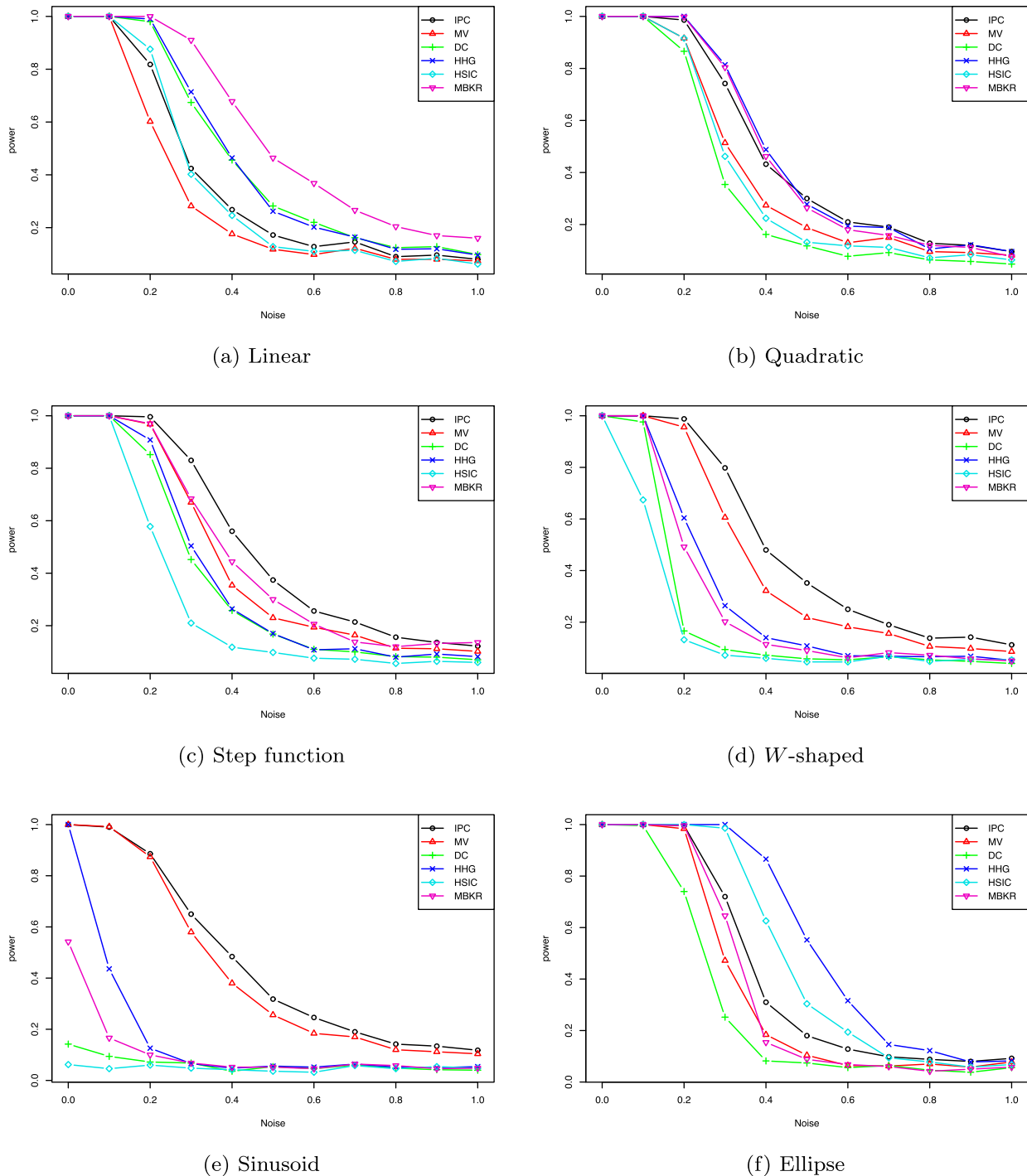
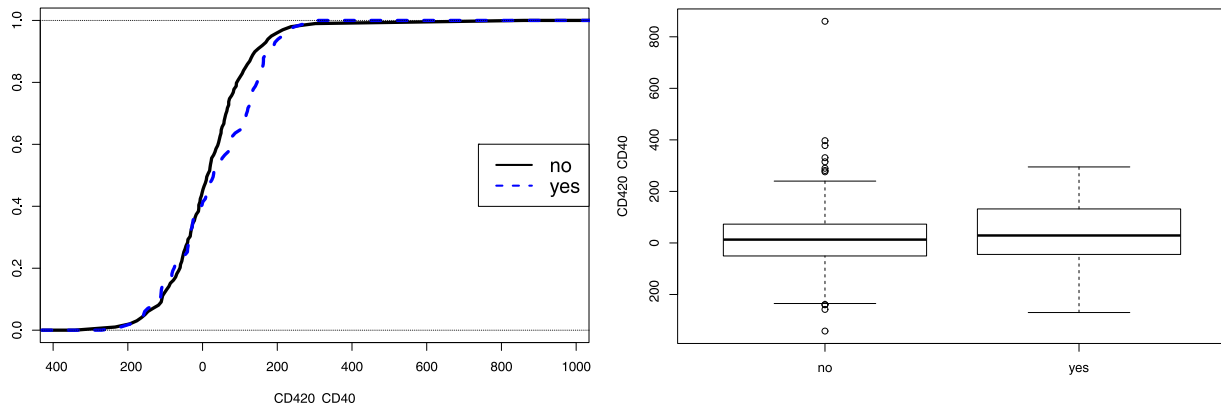


Figure 4. Comparison of powers of several tests of independence in Example 4.3. The noise level increases from left to right. In each case, 500 simulations are used to estimate the power of each test. (a) Linear. (b) Quadratic. (c) Step function. (d) *W*-shaped. (e) Sinusoid and (f) Ellipse.

that all the tests reject H_0^3 , H_0^4 and H_0^5 , and accept H_0^2 . That is, the treatment effect under the ZDV+zalcitabine group depends on antiretroviral history, age and CD80, but not on gender. Regarding the history of intravenous drug use, the IPC, DC, HHG and HSIC tests declare statistical dependence between this and the treatment effect. However, the MV test has a p -value larger than 0.05, and thus it can not reject H_0^1 . We draw the empirical conditional distributions of X given $Y = 0$ and 1 as well as the side-by-side boxplots in Figure 5, where $Y =$ history of intravenous drug use. We see that the conditional distributions of X are different across different Y . However, the difference is relatively small and mainly occurs in the right tails. According to the discussion in Section 3.3, IPC test will be more powerful in such case. Also, the categories of Y are very unbalanced with $\#\{Y = 0\} = 448$ and $\#\{Y = 1\} = 76$, making the MV test more difficult to detect the dependence between X and Y .

Table 7. The p -values of each test in Example 4.4.

Y	IPC	MV	DC	HHG	HSIC
history of intravenous drug use	0.0387	0.0580	0.0330	0.0060	0.0070
gender	0.8914	0.8832	0.8751	0.8232	0.7223
antiretroviral history	0.0007	0.0007	0.0020	0.0020	0.0060
age	0.0420	0.0210	0.0030	0.0120	0.0140
CD80	0.0040	0.0030	0.0060	0.0240	0.0100

**Figure 5.** The left panel shows the empirical conditional distributions of CD420 – CD40 given $Y = 0$ and $Y = 1$. And the right panel shows the side-by-side boxplots of CD420 – CD40 against $Y = 0$ and $Y = 1$. Here $Y =$ history of intravenous drug use.

5. Discussion

In this paper, we studied the IPC test of independence between a continuous variable X and a categorical variable Y . When the number of categories of Y is fixed, the IPC test statistic is in essence the k -sample Anderson Darling test statistic, and its theoretical properties were studied in Scholz and Stephens (1987). Our work mainly focused on two aspects. First, we derived the convergence rate of the IPC statistic to the IPC index and thus a lower bound of the power of the test at a given significance level with a finite sample size could be derived. Second, we showed that the standardized test statistic has an asymptotic normal distribution when the number of categories R diverges to infinity with the sample size. A distinguished merit is thereby shared by the IPC test, that is, its critical values can be easily obtained by using an approximated normal distribution when R is relatively large. As an application, we extended the IPC test to test independence between two continuous random variables. We uniformly slice a continuous variable into a discrete variable in order to apply the IPC test. And by allowing more slices as the sample size increases, the IPC test is allowed to gain more test power. The proposed test was compared to the DC test, HHG test, HSIC test and MV test on many simulation experiments. The results showed that the IPC test has a better performance in many scenarios. It is also possible to consider more different slicing schemes for independence testing of continuous variables. We left it for further research.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by National Natural Science Foundation of China [Grant numbers 12271286, 11931001 and 11771241].

References

- Csörgő, S. (1985). Testing for independence by the empirical characteristic function. *Journal of Multivariate Analysis*, 16(3), 290–299. [https://doi.org/10.1016/0047-259X\(85\)90022-3](https://doi.org/10.1016/0047-259X(85)90022-3)
- Cui, H., Li, R., & Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510), 630–641. <https://doi.org/10.1080/01621459.2014.920256>
- Cui, H., & Zhong, W. (2018). A distribution-free test of independence and its application to variable selection. Available at arXiv:1801.10559.
- Cui, H., & Zhong, W. (2019). A distribution-free test of independence based on mean variance index. *Computational Statistics & Data Analysis*, 139, 117–133. <https://doi.org/10.1016/j.csda.2019.05.004>
- Dvoretzky, A., Kiefer, J., & Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3), 642–669. <https://doi.org/10.1214/aoms/1177728174>

- Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In S. Jain, H. U. Simon, & E. Tomita (Eds.), *Algorithmic learning theory* (pp. 63–77). Springer Berlin Heidelberg.
- Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., & Smola, A. J. (2007). A kernel statistical test of independence. In *Proceedings of the 20th International Conference on Neural Information Processing Systems* (pp 585–592). Curran Associates Inc. NIPS'07.
- Hall, P., & Heyde, C. C (1980). *Martingale limit theory and its application*, Probability and mathematical statistics, Inc, Academic Press [Harcourt Brace Jovanovich, Publishers].
- Hammer, S. M., Katzenstein, D. A., Hughes, M. D., Gundacker, H., Schooley, R. T., Haubrich, R. H., Henry, W. K., Lederman, M. M., Phair, J. P., Niu, M., Hirsch, M. S., & Merigan, T. C. (1996). A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15), 1081–1090. <https://doi.org/10.1056/NEJM199610103351501>
- He, S., Ma, S., & Xu, W. (2019). A modified mean-variance feature-screening procedure for ultrahigh-dimensional discriminant analysis. *Computational Statistics & Data Analysis*, 137, 155–169. <https://doi.org/10.1016/j.csda.2019.02.003>
- Heller, R., Heller, Y., & Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2), 503–510. <https://doi.org/10.1093/biomet/ass070>
- Heller, R., Heller, Y., Kaufman, S., Brill, B., & Gorfine, M. (2016). Consistent distribution-free k -sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(29), 1–54.
- Hoeffding, W. (1948). A non-parametric test of independence. *The Annals of Mathematical Statistics*, 19(4), 546–557. <https://doi.org/10.1214/aoms/1177730150>
- Jiang, B., Ye, C., & Liu, J. S. (2015). Nonparametric k -sample tests via dynamic slicing. *Journal of the American Statistical Association*, 110(510), 642–653. <https://doi.org/10.1080/01621459.2014.920257>
- Lu, W., Zhang, H. H., & Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research*, 22(5), 493–504. <https://doi.org/10.1177/0962280211428383>
- Ma, W., Xiao, J., Yang, Y., & Ye, F. (2022). Model-free feature screening for ultrahigh dimensional data via a Pearson chi-square based index. *Journal of Statistical Computation and Simulation*, 92(15), 3222–3248. <https://doi.org/10.1080/00949655.2022.2062358>
- Mai, Q., & Zou, H. (2015a). Sparse semiparametric discriminant analysis. *Journal of Multivariate Analysis*, 135, 175–188. <https://doi.org/10.1016/j.jmva.2014.12.009>
- Mai, Q., & Zou, H. (2015b). The fused Kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics*, 43(4), 1471–1497. <https://doi.org/10.1214/14-AOS1303>
- Ni, L., & Fang, F. (2016). Entropy-based model-free feature screening for ultrahigh-dimensional multiclass classification. *Journal of Nonparametric Statistics*, 28(3), 515–530. <https://doi.org/10.1080/10485252.2016.1167206>
- Ni, L., Fang, F., & Shao, J. (2020). Feature screening for ultrahigh dimensional categorical data with covariates missing at random. *Computational Statistics & Data Analysis*, 142, Article 106824. <https://doi.org/10.1016/j.csda.2019.106824>
- Ni, L., Fang, F., & Wan, F. (2017). Adjusted Pearson chi-square feature screening for multi-classification with ultrahigh dimensional data. *Metrika*, 80(6–8), 805–828. <https://doi.org/10.1007/s00184-017-0629-9>
- Pfister, N., Bühlmann, P., Schölkopf, B., & Peters, J. (2018). Kernel-based tests for joint independence. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, 80(1), 5–31. <https://doi.org/10.1111/rssb.12235>
- Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *The Annals of Statistics*, 3(1), 1–14. <https://doi.org/10.1214/aos/1176342996>
- Scholz, F.-W., & Stephens, M. A. (1987). k -sample Anderson–Darling tests. *Journal of the American Statistical Association*, 82(399), 918–924. <https://doi.org/10.2307/2288805>
- Székel, G. J., & Rizzo, M. L. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4), 1236–1265. <https://doi.org/10.1214/09-AOAS312>
- Székel, G. J., Rizzo, M. L., & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6), 2769–2794. <https://doi.org/10.1214/009053607000000505>
- Tsiatis, A. A., Davidian, M., Zhang, M., & Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27(23), 4658–4677. <https://doi.org/10.1002/sim.3113>
- Xu, K., Shen, Z., Huang, X., & Cheng, Q. (2020). Projection correlation between scalar and vector variables and its use in feature screening with multi-response data. *Journal of Statistical Computation and Simulation*, 90(11), 1923–1942. <https://doi.org/10.1080/00949655.2020.1753057>
- Yan, X., Tang, N., Xie, J., Ding, X., & Wang, Z. (2018). Fused mean-variance filter for feature screening. *Computational Statistics & Data Analysis*, 122, 18–32. <https://doi.org/10.1016/j.csda.2017.10.008>
- Zhang, M., Tsiatis, A. A., & Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3), 707–715. <https://doi.org/10.1111/j.1541-0420.2007.00976.x>
- Zhang, Y., Chen, C., & Zhu, L. (2022). Sliced independence test. *Statistica Sinica*, 32(Special online issue), 2477–2496. <https://doi.org/10.5705/ss.202021.0203>
- Zhong, W., Wang, J., & Chen, X. (2021). Censored mean variance sure independence screening for ultrahigh dimensional survival data. *Computational Statistics & Data Analysis*, 159, Article 107206. <https://doi.org/10.1016/j.csda.2021.107206>
- Zhou, N., Guo, X., & Zhu, L. (2020). A projection-based model checking for heterogeneous treatment effect. Available at arXiv:2009.10900.
- Zhou, Y., & Zhu, L. (2018). Model-free feature screening for ultrahigh dimensional data through a modified Blum-Kiefer-Rosenblatt correlation. *Statistica Sinica*, 28(3), 1351–1370. <https://doi.org/10.5705/ss.202016.0264>
- Zhu, L., Xu, K., Li, R., & Zhong, W. (2017). Projection correlation between two random vectors. *Biometrika*, 104(4), 829–843. <https://doi.org/10.1093/biomet/asx043>

Appendix. Proof of theorems

This appendix contains the technical proofs of Lemma 2.2 and Theorem 3.1. Lemma 2.1 and Theorem 2.4 are direct corollaries of Theorem 3.2, and the proof of Theorem 3.2 follows from Lemma 4 in Ma et al. (2022), and thus their proofs are omitted.

A.1 Notations and preliminaries

Recall that the IPC index of (X, Y) , where X is a continuous random variable with support \mathbb{R}_X and $Y \in \{1, \dots, R\}$ is a categorical variable with R categories is defined as

$$\begin{aligned} \text{IPC}(X, Y) &= \sum_{r=1}^R p_r \int \frac{[F(x) - F_r(x)]^2}{F(x)(1 - F(x))} dF(x) \\ &= \sum_{r=1}^R \int \frac{[p_r F(x) - F(x, r)]^2}{F(x) \bar{F}(x) p_r} dF(x), \end{aligned}$$

where $F(x)$ is the distribution function of X , $F_r(x) = P(X \leq x | Y = r)$, $\bar{F}(x) = 1 - F(x)$, $p_r = P(Y = r)$ and $F(x, r) = P(X \leq x, Y = r)$, $r = 1, \dots, R$. And given i.i.d. samples $Z_i = (X_i, Y_i)$ for $i = 1, \dots, n$, the IPC statistic is defined as

$$\begin{aligned} \widehat{\text{IPC}}_n(X, Y) &= \sum_{r=1}^R \hat{p}_r \int \frac{(F_n(x) - F_{rn}(x))^2}{F_n(x) \bar{F}_n(x)} dF_n(x) \\ &= \sum_{r=1}^R \int \frac{(\hat{p}_r F_n(x) - F_n(x, r))^2}{F_n(x) \bar{F}_n(x) \hat{p}_r} dF_n(x) \\ &= \frac{1}{n} \sum_{r=1}^R \sum_{i=1}^n \frac{(\hat{p}_r F_n(X_i) - F_n(X_i, r))^2}{F_n(X_i) \bar{F}_n(X_i) \hat{p}_r}, \end{aligned}$$

where $F_n(x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x)$, $\bar{F}_n(x) = 1 - F_n(x)$, $\hat{p}_r = \frac{1}{n} \sum_{i=1}^n I(Y_i = r)$, $F_n(x, r) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x, Y_i = r)$, and $F_{rn}(x) = F_n(x, r)/\hat{p}_r$ for $r = 1, \dots, R$.

We first provide a proof of Lemma 2.2.

Proof of Lemma 2.2.: It is obvious that $\text{IPC}(X, Y) = 0$ if and only if X and Y are independent. By noticing that $\sum_{r=1}^R p_r = 1$ and $\sum_{r=1}^R F(x, r) = F(x)$, we have

$$\begin{aligned} \frac{1}{F(x) \bar{F}(x)} \sum_{r=1}^R \frac{(p_r F(x) - F(x, r))^2}{p_r} &= \frac{1}{F(x) \bar{F}(x)} \left(\sum_{r=1}^R \frac{F^2(x, r)}{p_r} - F^2(x) \right) \\ &< \frac{1}{F(x) \bar{F}(x)} \left(\sum_{r=1}^R \frac{F(x, r) p_r}{p_r} - F^2(x) \right) \\ &= \frac{1}{F(x) \bar{F}(x)} \left(\sum_{r=1}^R F(x, r) - F^2(x) \right) \\ &= 1. \end{aligned}$$

Hence we have $\text{IPC}(X, Y) < 1$. ■

Next, we give some preparations for the proof of Theorem 3.1. For given constant $C > 0$, let $F^{n,C}(x) = F(x) \vee n^{-\frac{1}{2+C}}$, $\bar{F}^{n,C}(x) = \bar{F}(x) \vee n^{-\frac{1}{2+C}}$, $F_n^C(x) = F_n(x) \vee n^{-\frac{1}{2+C}}$ and $\bar{F}_n^C(x) = \bar{F}_n(x) \vee n^{-\frac{1}{2+C}}$. Then we have the following lemmas.

Lemma A.1: Let $\Delta_1 F(x) = F^{n,C}(x) - F_n^C(x)$ and $\Delta_2 F(x) = \bar{F}^{n,C}(x) - \bar{F}_n^C(x)$. Then

$$\sup_{x \in \mathbb{R}} |\Delta_1 F(x)| = O_p(n^{-1/2}), \quad \text{and} \quad \sup_{x \in \mathbb{R}} |\Delta_2 F(x)| = O_p(n^{-1/2}).$$

Proof: It is easy to show that

$$\left| F^{n,C}(x) - F_n^C(x) \right| \leq |F(x) - F_n(x)|.$$

Hence by Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Dvoretzky et al., 1956),

$$\sup_x |\Delta_1 F(x)| \leq \sup_x |F(x) - F_n(x)| = O_p(n^{-1/2}).$$

Similarly, we have $\sup_x |\Delta_2 F(x)| = O_p(n^{-1/2})$. ■

Lemma A.2: $\sup_x \left| \frac{F_n^{n,C}(x) \bar{F}_n^C(x) - F_n^C(x) \bar{F}_n^C(x)}{F_n^C(x) \bar{F}_n^C(x)} \right| = O_p(n^{-\frac{C}{4+2C}}) = o_p(1).$

Proof: Note that

$$\begin{aligned} F_n^{n,C}(x) \bar{F}_n^C(x) &= \left(F_n^C(x) + \Delta_1 F(x) \right) \left(\bar{F}_n^C(x) + \Delta_2 F(x) \right) \\ &= F_n^C(x) \bar{F}_n^C(x) + \bar{F}_n^C(x) \Delta_1 F(x) + F_n^C(x) \Delta_2 F(x) + \Delta_1 F(x) \Delta_2 F(x). \end{aligned}$$

Then,

$$\begin{aligned} \sup_x \left| \frac{F_n^C(x) \bar{F}_n^C(x) - F_n^{n,C}(x) \bar{F}_n^C(x)}{F_n^C(x) \bar{F}_n^C(x)} \right| &\leq \sup_x \left| \frac{\Delta_1 F(x)}{F_n^C(x)} \right| + \sup_x \left| \frac{\Delta_2 F(x)}{\bar{F}_n^C(x)} \right| + \sup_x \left| \frac{\Delta_1 F(x) \Delta_2 F(x)}{F_n^C(x) \bar{F}_n^C(x)} \right| \\ &= O_p\left(n^{-1/2 + \frac{1}{2+C}}\right) + O_p\left(n^{-1/2 + \frac{1}{2+C}}\right) + O_p\left(n^{-1 + \frac{1}{2+C}}\right) \\ &= O_p\left(n^{-\frac{C}{4+2C}}\right). \end{aligned}$$

A.2 Proof of Theorem 3.1

To avoid any ambiguity, Theorem 3.1 considers a sequence of problems indexed by $(n_k, R_k, p_{1,k}, \dots, p_{R_k,k})$, $k = 1, 2, \dots$, where the sample size $n_k \rightarrow \infty$, the number of categories $R_k \rightarrow \infty$, and let $Y_k = Y(R_k)$ denote the categorical variable with R_k categories and $p_{r,k} = P(Y(R_k) = r)$, $r = 1, \dots, R_k$. From now on, we shall omit the subscript unless specifically mentioned. Moreover, in Section A.2, we should keep in mind that X and Y are independent.

A.2.1 Architecture of the proof

Our aim here is to provide a general overview of the proof of Theorem 3.1. At a high level, the general structure is fairly simple. And to make the structure clear, we divide the proof into three parts.

- (1) First, given a positive constant C , we substitute $F_n^{n,C}(x)$, \bar{F}_n^C and p_r for $F_n(x)$, $\bar{F}_n(x)$ and \hat{p}_r in the denominator of the IPC statistic, and thereby obtain

$$\widehat{\text{IPC}}_{n,C}(X, Y) := \sum_{r=1}^R \int \frac{1}{p_r} \frac{[\hat{p}_r F_n(x) - F_n(x, r)]^2}{F_n^{n,C}(x) \bar{F}_n^C(x)} dF_n(x).$$

And then prove that the difference between $n\widehat{\text{IPC}}_n(X, Y)/\sqrt{R}$ and $n\widehat{\text{IPC}}_{n,C}(X, Y)/\sqrt{R}$ is bounded by $n\widehat{\text{IPC}}_{n,C}(X, Y)/\sqrt{R} \times O_p(n^{-\frac{C}{4+2C}} + \frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} n^{-1/2}) + O_p(n^{-\frac{1}{2+C}} \sqrt{R})$, provided that $\frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} = o(n^{1/2})$.

- (2) Fixing $C = 6$, let

$$f_i(x, r) = [I(X_i \leq x) - F(x)] [I(Y_i = r) - p_r],$$

and

$$f_{i,n}(x, r) = \frac{f_i(x, r)}{\sqrt{F_n^{n,6}(x) \bar{F}_n^{n,6}(x)}},$$

and define

$$\widetilde{\text{IPC}}_n(X, Y) = \sum_{r=1}^R \frac{1}{p_r} \int \left[\frac{1}{n} \sum_{i=1}^n f_{i,n}(x, r) \right]^2 dF(x).$$

Under the condition $\frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} = o(n^{3/8})$, showing that $n\widehat{\text{IPC}}_{n,6}(X, Y)/\sqrt{R}$ is close to $n\widetilde{\text{IPC}}_n(X, Y)/\sqrt{R}$ and combined with the first part of the proof, we can derive that

$$n\widehat{\text{IPC}}_n(X, Y) - n\widetilde{\text{IPC}}_n(X, Y) = o_p(\sqrt{R}).$$

- (3) Finally, consider

$$n\widetilde{\text{IPC}}_n(X, Y) = J_{1n} + J_{2n},$$

where

$$J_{1n} = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^R \frac{1}{p_r} \int f_{i,n}^2(x, r) dF(x),$$

and

$$J_{2n} = \frac{1}{n} \sum_{i \neq j} \sum_{r=1}^R \frac{1}{p_r} \int f_{i,n}(x, r) f_{j,n}(x, r) dF(x).$$

We show that

$$\frac{J_{1n} - (R-1)}{\sqrt{2(\frac{\pi^2}{3} - 3)(R-1)}} \xrightarrow{P} 0, \quad \text{and} \quad \frac{J_{2n}}{\sqrt{2(\frac{\pi^2}{3} - 3)(R-1)}}$$

can be viewed as a martingale difference sequence. Then by the well-developed theory of central limit theorem of the martingale difference (Hall & Heyde, 1980), we can complete the proof.

Combined with Lemmas A.1 and A.2, the proof in part 1 is not difficult. And the proofs in part 2 and part 3 follow from Cui and Zhong (2018) and Cui and Zhong (2019) with a small modification.

A.2.2 Part 1

We summarize the conclusion we want to prove in part 1 into the following lemma.

Lemma A.3: For a fixed constant C , let

$$\widehat{\text{IPC}}_{n,C}(X, Y) = \sum_{r=1}^R \int \frac{1}{p_r} \frac{[\hat{p}_r F_n(x) - F_n(x, r)]^2}{F_n(x) \bar{F}_n(x)} dF_n(x).$$

For simplicity, write $\widehat{\text{IPC}}_n = \widehat{\text{IPC}}_n(X, Y)$, and $\widehat{\text{IPC}}_{n,C} = \widehat{\text{IPC}}_{n,C}(X, Y)$. Then if $\frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} = o(n^{1/2})$, and under condition that X and Y are independent, we have

$$|\widehat{\text{IPC}}_n - \widehat{\text{IPC}}_{n,C}| = O_p\left(n^{-\frac{3+C}{2+C}R}\right) + \widehat{\text{IPC}}_{n,C} \left(O_p\left(n^{-\frac{C}{4+2C}}\right) + O_p\left(\frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} n^{-1/2}\right) \right).$$

Proof: Let

$$\widehat{\text{IPC}}'_n = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^R \frac{[F_n(X_i, r) - \hat{p}_r F_n(X_i)]^2}{F_n(X_i) \bar{F}_n(X_i) p_r}.$$

Then

$$\begin{aligned} |\widehat{\text{IPC}}_n - \widehat{\text{IPC}}'_n| &\leq \max_{1 \leq r \leq R} \left| 1 - \frac{p_r}{\hat{p}_r} \right| \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^R \frac{[\hat{p}_r F_n(X_i) - F_n(X_i, r)]^2}{F_n(X_i) \bar{F}_n(X_i) p_r} \\ &= \widehat{\text{IPC}}'_n \max_{1 \leq r \leq R} \left| 1 - \frac{\hat{p}_r}{p_r} \right|. \end{aligned}$$

Since

$$E(\sqrt{n}(\hat{p}_r - p_r))^2 = p_r(1 - p_r),$$

we have

$$\begin{aligned} E\left(\max_{1 \leq r \leq R} |\hat{p}_r - p_r|\right)^2 &\leq E\left(\sum_{r=1}^R |\hat{p}_r - p_r|\right)^2 \\ &\leq R \sum_{r=1}^R E(\hat{p}_r - p_r)^2 \\ &= R \sum_{r=1}^R \frac{p_r(1 - p_r)}{n} \leq \frac{R}{n}. \end{aligned}$$

So, $\max_{1 \leq r \leq R} |\hat{p}_r - p_r| = O_p(\sqrt{R/n})$. Then

$$\max_{1 \leq r \leq R} \left| \frac{\hat{p}_r - p_r}{\hat{p}_r} \right| = \max_{1 \leq r \leq R} \left| \frac{\hat{p}_r - p_r}{p_r + \hat{p}_r - p_r} \right| \leq \max_{1 \leq r \leq R} |\hat{p}_r - p_r| \max_{1 \leq r \leq R} \frac{1}{p_r + \hat{p}_r - p_r}.$$

Since $\hat{p}_r - p_r = O_p(\sqrt{\frac{R}{n}}) = o_p(\min_{1 \leq r \leq R} p_r)$, we have

$$\max_{1 \leq r \leq R} \left| \frac{\hat{p}_r - p_r}{\hat{p}_r} \right| = O_p\left(\frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} n^{-1/2}\right) = o_p(1).$$

Hence, $\widehat{\text{IPC}}_n = (1 + O_p(\frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} n^{-1/2})) \widehat{\text{IPC}}'_n$. Next, let

$$\widehat{\text{IPC}}_n^* = \frac{1}{n} \sum_{r=1}^R \sum_{i=1}^n \frac{[\hat{p}_r F_n(X_i) - F_n(X_i, r)]^2}{F_n^C(X_i) \bar{F}_n^C(X_i) p_r}.$$

Let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be the ordered statistics of X_1, \dots, X_n . Since X is continuous, there are no ties among X_1, \dots, X_n . We can assume that $X_{(1)} < \dots < X_{(n)}$. Let $A_n = \lfloor n^{1-\frac{1}{2+C}} \rfloor$, and define

$$S_{n1} = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^R \frac{[\hat{p}_r F_n(X_i) - F_n(X_i, r)]^2}{F_n(X_i) \bar{F}_n(X_i) p_r} I(X_i \leq X_{(A_n)}),$$

$$S_{n2} = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^R \frac{[\hat{p}_r F_n(X_i) - F_n(X_i, r)]^2}{F_n(X_i) \bar{F}_n(X_i) p_r} I(X_i \geq X_{(n-A_n)}).$$

Indeed, we have $0 \leq \widehat{\text{IPC}}'_n - \widehat{\text{IPC}}^*_n \leq S_{n1} + S_{n2}$. And

$$\begin{aligned} ES_{n1} &= \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^R \sum_{j=1}^n E \left\{ \frac{[\hat{p}_r F_n(X_i) - F_n(X_i, r)]^2}{F_n(X_i) \bar{F}_n(X_i) p_r} I(X_i \leq X_{(A_n)}) I(X_i = X_{(j)}) \right\} \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^R \sum_{j=1}^{A_n} E \left\{ \frac{[\hat{p}_r F_n(X_i) - F_n(X_i, r)]^2}{F_n(X_i) \bar{F}_n(X_i) p_r} I(X_i = X_{(j)}) \right\} \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{A_n} \sum_{r=1}^R \frac{1}{j} \left\{ \frac{(n-j)[(n-1)p_r + 1]}{n^2} - \frac{2(n-j)}{n^2} [(n-1)p_r + 1] + \frac{(n-j-1)p_r + 1}{n} \right\} \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{A_n} \frac{R-1}{n^2} \\ &= \frac{A_n(R-1)}{n^2}. \end{aligned}$$

Similarly, we also have $ES_{n2} = \frac{A_n}{n^2}(R-1)$. Therefore,

$$\widehat{\text{IPC}}'_n - \widehat{\text{IPC}}^*_n = O_p\left(n^{-\frac{3+C}{2+C}}R\right).$$

Finally, according to Lemma A.2,

$$\begin{aligned} \left| \widehat{\text{IPC}}^*_n - \widehat{\text{IPC}}_{n,C} \right| &= \frac{1}{n} \left| \sum_{i=1}^n \sum_{r=1}^R \frac{[\hat{p}_r F_n(X_i) - F_n(X_i, r)]^2}{F^{n,C}(X_i) \bar{F}^{n,C}(X_i) p_r} - \sum_{i=1}^n \sum_{r=1}^R \frac{[\hat{p}_r F_n(X_i) - F_n(X_i, r)]^2}{F_n^C(X_i) \bar{F}_n^C(X_i) p_r} \right| \\ &= \frac{1}{n} \left| \sum_{i=1}^n \sum_{r=1}^R \frac{[\hat{p}_r F_n(X_i) - F_n(X_i, r)]^2}{F^{n,C}(X_i) \bar{F}^{n,C}(X_i) p_r} \times \left(\frac{F^{n,C}(X_i) \bar{F}^{n,C}(X_i)}{F_n^C(X_i) \bar{F}_n^C(X_i)} - 1 \right) \right| \\ &\leq \widehat{\text{IPC}}_{n,C} \times \sup_x \left| \frac{F^{n,C}(x) \bar{F}^{n,C}(x)}{F_n^C(x) \bar{F}_n^C(x)} - 1 \right| \\ &= \widehat{\text{IPC}}_{n,C} O_p\left(n^{-\frac{C}{4+2C}}\right). \end{aligned}$$

Hence

$$\left| \widehat{\text{IPC}}_n - \widehat{\text{IPC}}_{n,C} \right| = O_p\left(n^{-\frac{3+C}{2+C}}R\right) + \widehat{\text{IPC}}_{n,C} \left(O_p\left(n^{-\frac{C}{4+2C}}\right) + O_p\left(\frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} n^{-1/2}\right) \right).$$

A.2.3 Part 2

Recall that

$$f_i(x, r) = [I(X_i \leq x) - F(x)][I(Y_i = r) - p_r], \quad f_{i,n}(x, r) = \frac{f_i(x, r)}{\sqrt{F^{n,6}(x) \bar{F}^{n,6}(x) p_r}}$$

and

$$\widetilde{\text{IPC}}_n(X, Y) = \sum_{r=1}^R \frac{1}{p_r} \int \left[\frac{1}{n} \sum_{i=1}^n f_{i,n}(x, r) \right]^2 dF(x).$$

The following lemma is what we want to prove in part 2.

Lemma A.4: If $\frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} = o(n^{3/8})$, and Under H_0 : X and Y are independent, then

$$\widehat{\text{IPC}}_n(X, Y) - \widetilde{\text{IPC}}_n(X, Y) = O_p(Rn^{-9/8}) + O_p\left(\frac{Rn^{-5/4}}{\min_{1 \leq r \leq R} p_r}\right) + \widetilde{\text{IPC}}_n(X, Y) o_p(n^{-1/8}).$$

Proof: For simplicity, write $\widehat{\text{IPC}}_n = \widehat{\text{IPC}}_n(X, Y)$. Given $C = 6$, according to Lemma A.3, and under the condition that $\frac{\sqrt{R}}{\min_{1 \leq r \leq R} p_r} = o(n^{3/8})$, we have

$$\begin{aligned} \widehat{\text{IPC}}_n - \widehat{\text{IPC}}_{n,6} &= O_p(n^{-9/8}R) + \widehat{\text{IPC}}_{n,6} [O_p(n^{-3/8}) + o_p(n^{-1/8})] \\ &= O_p(n^{-9/8}R) + \widehat{\text{IPC}}_{n,6} o_p(n^{-1/8}). \end{aligned} \quad (\text{A1})$$

Let

$$\widehat{\text{IPC}}_{1n} = \sum_{r=1}^R \frac{1}{p_r} \int \left[\frac{1}{n} \sum_{i=1}^n f_{i,n}(x, r) \right]^2 dF_n(x).$$

Next, we follow the proof of Lemma A.1 in Cui and Zhong (2019), and show that

$$\begin{aligned} \widehat{\text{IPC}}_{n,6} - \widehat{\text{IPC}}_{1n} &= \sum_{r=1}^R \frac{1}{p_r} \int \frac{1}{F^{n,6}(x) \bar{F}^{n,6}(x)} \left\{ [\hat{p}_r F_n(x) - F_n(x, r)]^2 - \left[\frac{1}{n} \sum_{i=1}^n f_i(x, r) \right]^2 \right\} dF_n(x) \\ &= O(n^{1/8}) \sum_{r=1}^R \frac{1}{p_r} \int \left\{ [\hat{p}_r F_n(x) - F_n(x, r)]^2 - \left[\frac{1}{n} \sum_{i=1}^n f_i(x, r) \right]^2 \right\} dF_n(x). \end{aligned}$$

Let $\bar{f}_n(x, r) = \frac{1}{n} \sum_{i=1}^n f_i(x, r)$. By the DKW inequality, we have

$$\begin{aligned} \sup_x \left| [\hat{p}_r F_n(x) - F_n(x, r)]^2 - \left[\frac{1}{n} \sum_{i=1}^n f_i(x, r) \right]^2 \right| &= \sup_x |\hat{p}_r F_n(x) - F_n(x, r) - \bar{f}_n(x, r)| |\hat{p}_r F_n(x) - F_n(x, r) + \bar{f}_n(x, r)| \\ &= \sup_x |F_n(x) - F(x)| |\hat{p}_r - p_r| \left\{ \sup_x |\hat{p}_r F_n(x) - F_n(x, r)| + \sup_x |\bar{f}_n(x, r)| \right\} \\ &= O_p(n^{-1/2}) O_p(n^{-1/2}) O_p(n^{-1/2}) = O_p(n^{-3/2}). \end{aligned}$$

Here, the second equality follows by

$$\begin{aligned} \hat{p}_r F_n(x) - F_n(x, r) - \bar{f}_n(x, r) &= \left\{ \frac{1}{n} \sum_{i=1}^n I(X_i \leq x, Y_i = r) - F_n(x) \hat{p}_r \right\} \\ &\quad - \left\{ \frac{1}{n} \sum_{i=1}^n I(X_i \leq x, Y_i = r) - F(x) \hat{p}_r - p_r F_n(x) + p_r F(x) \right\} \\ &= -[F_n(x) - F(x)] [\hat{p}_r - p_r], \end{aligned}$$

and the last equality follows by

$$\begin{aligned} \sup_x |\hat{p}_r F_n(x) - F_n(x, r)| &= O_p(n^{-1/2}), \\ \sup_x |\bar{f}_n(x, r)| &= O_p(n^{-1/2}). \end{aligned}$$

Indeed,

$$\begin{aligned} \sup_x |\hat{p}_r F_n(x) - F_n(x, r)| &\leq \sup_x \left| \frac{1}{n} \sum_{i=1}^n [I(X_i \leq x) - F(x)] I(Y_i = r) \right| \\ &\quad + \sup_x \left| F(x) \frac{1}{n} \sum_{i=1}^n [I(Y_i = r) - p_r] \right| + |\hat{p}_r - p_r| + \sup_x |F_n(x) - F(x)| \\ &= \sup_x \left| \frac{1}{n} \sum_{i=1}^n [I(X_i \leq x) - F(x)] I(Y_i = r) \right| + \sup_x F(x) |\hat{p}_r - p_r| + O_p(n^{-1/2}) \\ &= \sup_x \left| \frac{1}{n} \sum_{i=1}^n [I(X_i \leq x) - F(x)] I(Y_i = r) \right| + O_p(n^{-1/2}), \end{aligned}$$

and

$$\begin{aligned} E \left[\sup_x \left| \frac{1}{n} \sum_{i=1}^n [I(X_i \leq x) - F(x)] I(Y_i = r) \right| \right] &= \sum_{m=1}^n E \left[\sup_x \left| \frac{1}{n} \sum_{i=1}^n [I(X_i \leq x) - F(x)] I(Y_i = r) \right|, m Y_i' = r \right] \\ &= \sum_{m=1}^n \binom{n}{m} p_r^m (1-p_r)^{n-m} \frac{\sqrt{m}}{n} E \left[\sup_x \left| \frac{1}{\sqrt{m}} \sum_{i=1}^m [I(X_i \leq x) - F(x)] \right| \right] \end{aligned}$$

$$\begin{aligned} &\leq 4 \sum_{m=1}^n \binom{n}{m} p_r^m (1-p_r)^{n-m} \frac{\sqrt{m}}{n} \\ &\leq 4n^{-1/2}, \end{aligned}$$

where the first inequality follows by the DKW inequality. Hence, $\sup_x |\hat{p}_r F_n(x) - F_n(x, r)| = O_p(n^{-1/2})$ and similarly $\sup_x |\bar{f}_n(x, r)| = O_p(n^{-1/2})$. Therefore, we have

$$\widehat{\text{IPC}}_{n,6} - \widetilde{\text{IPC}}_{1n} = \frac{R}{\min_{1 \leq r \leq R} p_r} O_p(n^{-11/8}). \tag{A2}$$

Combining (A1) and (A2), we have

$$\widehat{\text{IPC}}_n - \widetilde{\text{IPC}}_{1n} = O_p(Rn^{-9/8}) + \frac{R}{\min_{1 \leq r \leq R} p_r} O_p(n^{-11/8}) + \widetilde{\text{IPC}}_{1n} O_p(n^{-1/8}).$$

To complete the proof, we only need to show that

$$\begin{aligned} \widetilde{\text{IPC}}_{1n} - \widetilde{\text{IPC}}_n &= \sum_{r=1}^R \frac{1}{p_r} \int \left[\frac{1}{n} \sum_{i=1}^n f_{i,n}(x, r) \right]^2 d[F_n(x) - F(x)] \\ &= \frac{R}{\min_{1 \leq r \leq R} p_r} O_p(n^{-11/8}). \end{aligned}$$

It is enough to show that

$$I_n(r) := \int \left[\frac{1}{n} \sum_{i=1}^n f_{i,n}(x, r) \right]^2 d[F_n(x) - F(x)] = O_p(n^{-11/8}).$$

Without loss of generality, let $F(x)$ be the uniform distribution function, since we can make the transformation $X' = F(X)$ for the continuous random variable X . And

$$I_n(r) = \frac{1}{n} \sum_{j=1}^n \left[\frac{1}{n} \sum_{i=1}^n f_{i,n}(X_j, r) \right]^2 - \int_0^1 \left[\frac{1}{n} \sum_{i=1}^n f_{i,n}(x, r) \right]^2 dx.$$

For any $x, y \in (0, 1)$, it can be easily proved that

$$E f_{i,n}(x, r) f_{j,n}(y, r) = \frac{x \wedge y - xy}{\sqrt{x^{(n)}(1-x)^{(n)} y^{(n)}(1-y)^{(n)}}} (p_r - p_r^2) I(i=j),$$

where $x^{(n)} = x \vee n^{-1/8}$ and $(1-x)^{(n)} = (1-x) \vee n^{-1/8}$. Then

$$\begin{aligned} EI_n^2(r) &= E \left\{ \int_0^1 \left[\frac{1}{n} \sum_{j=1}^n [\bar{f}_n(X_j)^2 - \bar{f}_n(x)^2] \right] dx \right\}^2 \\ &= E \left\{ \int_0^1 \int_0^1 \left[\frac{1}{n} \sum_{j=1}^n [\bar{f}_n(X_j)^2 - \bar{f}_n(x)^2] \right] \left[\frac{1}{n} \sum_{j=1}^n [\bar{f}_n(X_j)^2 - \bar{f}_n(y)^2] \right] dx dy \right\} \\ &= \frac{1}{n} \int_0^1 \int_0^1 E \left\{ [\bar{f}_n(X_1)^2 - \bar{f}_n(x)^2] [\bar{f}_n(X_1)^2 - \bar{f}_n(y)^2] \right\} dx dy \\ &\quad + \frac{n-1}{n} \int_0^1 \int_0^1 E \left\{ [\bar{f}_n(X_1)^2 - \bar{f}_n(x)^2] [\bar{f}_n(X_2)^2 - \bar{f}_n(y)^2] \right\} dx dy \\ &= \int_0^1 \int_0^1 E \left\{ [\bar{f}_n(X_1)^2 - \bar{f}_n(x)^2] [\bar{f}_n(X_2)^2 - \bar{f}_n(y)^2] \right\} dx dy \\ &\quad + \frac{1}{n} [E[\bar{f}_n(X_1)^2 \bar{f}_n(X_1)^2] - E[\bar{f}_n(X_1)^2 \bar{f}_n(X_2)^2]], \end{aligned}$$

where $\bar{f}_n(x) = n^{-1} \sum_{i=1}^n f_{i,n}(x, r)$. And be careful here that $\bar{f}_n(x)$ is different from $\bar{f}_n(x, r)$ defined above.

Since $E f_{i,n}(x, r) = 0$ under H_0 , we have

$$E[f_{i,n}(x, r) f_{j,n}(x, r) f_{k,n}(y, r) f_{l,n}(y, r)] = 0$$

under H_0 if one of $\{i, j, k, l\}$ is different from the other three. Then we have

$$\begin{aligned} E[\bar{f}_n(x)^2 \bar{f}_n(y)^2] &= \frac{1}{n^4} \sum_{ij} \sum_{kl} E[f_{i,n}(x, r) f_{j,n}(x, r) f_{k,n}(y, r) f_{l,n}(y, r)] \\ &= \frac{1}{n^3} E[f_{1,n}(x, r)^2 f_{1,n}(y, r)^2] + \frac{n-1}{n^3} E[f_{1,n}^2(x, r)] E[f_{2,n}^2(y, r)] \end{aligned}$$

$$\begin{aligned}
 & + \frac{2(n-1)}{n^3} \{E[f_{1,n}(x, r)f_{1,n}(y, r)]\}^2 \\
 & = \frac{1}{n^3} \frac{1}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} E[f_1(x, r)^2 f_1(y, r)^2] \\
 & \quad + \frac{n-1}{n^3} \frac{1}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} E[f_1^2(x, r)] E[f_1^2(y, r)] \\
 & \quad + \frac{2(n-1)}{n^3} \frac{1}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} \{E[f_1(x, r)f_1(y, r)]\}^2 \\
 & = O(n^{-11/4}) + \frac{(p_r - p_r^2)^2}{n^2} \frac{[xy(1-x)(1-y) + 2(x \wedge y - xy)]^2}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}}.
 \end{aligned}$$

And also, we have

$$\begin{aligned}
 E[\bar{f}_n(X_1)^2 \bar{f}_n(y)^2] & = \frac{1}{n^4} \sum_{i,j} \sum_{k,l} E[f_{i,n}(X_1, r)f_{j,n}(X_1, r)f_{k,n}(y, r)f_{l,n}(y, r)] \\
 & = O(n^{-11/4}) + \frac{(p_r - p_r^2)^2}{n^2} \int_0^1 \frac{xy(1-x)(1-y) + 2(x \wedge y - xy)^2}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} dx, \\
 E[\bar{f}_n(x)^2 \bar{f}_n(X_2)^2] & = \frac{1}{n^4} \sum_{i,j} \sum_{k,l} E[f_{i,n}(x, r)f_{j,n}(x, r)f_{k,n}(X_2, r)f_{l,n}(X_2, r)] \\
 & = O(n^{-11/4}) + \frac{(p_r - p_r^2)^2}{n^2} \int_0^1 \frac{xy(1-x)(1-y) + 2(x \wedge y - xy)^2}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} dy, \\
 E[\bar{f}_n(X_1)^2 \bar{f}_n(X_2)^2] & = \frac{1}{n^4} \sum_{i,j} \sum_{k,l} E[f_{i,n}(X_1, r)f_{j,n}(X_1, r)f_{k,n}(X_2, r)f_{l,n}(X_2, r)] \\
 & = O(n^{-11/4}) + \frac{(p_r - p_r^2)^2}{n^2} \int_0^1 \frac{xy(1-x)(1-y) + 2(x \wedge y - xy)^2}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} dx dy,
 \end{aligned}$$

and

$$\begin{aligned}
 E[\bar{f}_n(X_1)^2 \bar{f}_n(X_1)^2] & = \frac{1}{n^4} \sum_{i,j} \sum_{k,l} E[f_{i,n}(X_1, r)f_{j,n}(X_1, r)f_{k,n}(X_1, r)f_{l,n}(X_1, r)] \\
 & = O(n^{-11/4}) + \frac{(p_r - p_r^2)^2}{n^2} \int_0^1 \frac{x^2(1-x)^2 + 2(x-x^2)^2}{(x^{(n)}(1-x)^{(n)})^2} dx.
 \end{aligned}$$

Hence,

$$\begin{aligned}
 E[I_n(r)^2] & = \int_0^1 \int_0^1 E[\bar{f}_n(X_1)^2 \bar{f}_n(X_2)^2] dx dy - \int_0^1 \int_0^1 E[\bar{f}_n(X_1)^2 \bar{f}_n(y)^2] dx dy \\
 & \quad - \int_0^1 \int_0^1 E[\bar{f}_n(x)^2 \bar{f}_n(X_1)^2] dx dy + \int_0^1 \int_0^1 E[\bar{f}_n(x)^2 \bar{f}_n(y)^2] dx dy \\
 & \quad + \frac{1}{n} [E[\bar{f}_n(X_1)^2 \bar{f}_n(X_1)^2] - E[\bar{f}_n(X_1)^2 \bar{f}_n(X_2)^2]] \\
 & = O(n^{-11/4}).
 \end{aligned}$$

So,

$$\begin{aligned}
 \widehat{\text{IPC}}_n - \widetilde{\text{IPC}}_n & = \widehat{\text{IPC}}_n - \widetilde{\text{IPC}}_{1n} + \widetilde{\text{IPC}}_{1n} - \widetilde{\text{IPC}}_n \\
 & = O_p(Rn^{-9/8}) + \frac{R}{\min_{1 \leq r \leq R} p_r} O_p(n^{-11/8}) + \widetilde{\text{IPC}}_n o_p(n^{-1/8}).
 \end{aligned}$$

A.2.4 Part 3

Now, we will complete the proof of Theorem 3.1.

Proof of Theorem 3.1.: Let $\tilde{T}_n = n\widetilde{\text{IPC}}_n$. Without loss of generality, we assume that $X \sim \text{Unif}(0, 1)$. Then $F(x) = x$ for $0 \leq x \leq 1$. According to Lemma A.4, we have

$$T_n - \tilde{T}_n = O_p(Rn^{-1/8}) + O_p\left(\frac{Rn^{-3/8}}{\min_{1 \leq r \leq R} p_r}\right) + o_p(\tilde{T}_n n^{-1/8}).$$

Then under the condition $\sqrt{R}/\min_{1 \leq r \leq R} p_r = o(n^{3/8})$, we have $R = o(n^{1/4})$, and thus $T_n - \tilde{T}_n = o_p(\sqrt{R}) + \tilde{T}_n o_p(n^{-1/8})$, i.e.,

$$\begin{aligned} \frac{T_n - (R - 1)}{\sqrt{2\left(\frac{\pi^2}{3} - 3\right)(R - 1)}} - \frac{\tilde{T}_n - (R - 1)}{\sqrt{2\left(\frac{\pi^2}{3} - 3\right)(R - 1)}} &= o_p(1) + \frac{\tilde{T}_n - (R - 1)}{\sqrt{2\left(\frac{\pi^2}{3} - 3\right)(R - 1)}} o_p(n^{-1/8}) + o_p\left(\sqrt{R}n^{-1/8}\right) \\ &= \frac{\tilde{T}_n - (R - 1)}{\sqrt{2\left(\frac{\pi^2}{3} - 3\right)(R - 1)}} o_p(n^{-1/8}) + o_p(1). \end{aligned}$$

Hence, we only need to prove that

$$\frac{\tilde{T}_n - (R - 1)}{\sqrt{2\left(\frac{\pi^2}{3} - 3\right)(R - 1)}} \xrightarrow{d} N(0, 1),$$

as $n \rightarrow \infty$.

Recall that $f_{i,n}(x, r) = \frac{(I(X_i \leq x) - x)(I(Y_i = r) - p_r)}{\sqrt{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}}$, where $x^{(n)} = x \vee n^{-1/8}$ and $(1-x)^{(n)} = (1-x) \vee n^{-1/8}$. We first give some important facts:

- (i) $E[f_{i,n}(x, r)f_{i,n}(y, s)] = \frac{(x \wedge y - xy)(p_r \delta_{rs} - p_r p_s)}{\sqrt{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}}$;
- (ii) $E[f_{i,n}^2(x, r)f_{i,n}^2(y, s)] \leq Cn^{1/8}(p_r \delta_{rs} + p_r p_s(p_r + p_s))$,

for all $1 \leq i \leq n, 1 \leq r, s \leq R$, where C is a constant and $\delta_{rs} = 1$ if $r = s$ and $\delta_{rs} = 0$, otherwise.

We prove (ii). Without loss of generality, we assume that $x \leq y$.

$$E[f_{i,n}^2(x, r)f_{i,n}^2(y, s)] = [p_r \delta_{rs} + p_r p_s(p_r + p_s)] E \frac{[I(X_i \leq x) - x]^2 [I(X_i \leq y) - y]^2}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}}.$$

And

$$\begin{aligned} E \frac{[I(X_i \leq x) - x]^2 [I(X_i \leq y) - y]^2}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} &= \frac{E[I(X_i \leq x) - 2xI(X_i \leq x) + x^2][I(X_i \leq y) - 2yI(X_i \leq y) + y^2]}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} \\ &= \frac{x(1-y)(1-y-2x+3xy)}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} \\ &\leq \frac{x(1-y)}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} \\ &\leq 4n^{1/8}. \end{aligned}$$

The last inequality is because, if $1/2 \leq x \leq y$, then $\frac{x(1-y)}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} \leq 4\frac{x}{(1-x)^{(n)}} \leq 4n^{1/8}$; if $x \leq y \leq 1/2$, then $\frac{x(1-y)}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} \leq 4\frac{1-y}{y^{(n)}} \leq 4n^{1/8}$; if $x \leq 1/2 \leq y$, then $\frac{x(1-y)}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} \leq 4$.

(iii) $\sum_{r,s,t,q=1}^R \frac{(p_r \delta_{rs} - p_r p_s)(p_r \delta_{rt} - p_r p_t)(p_t \delta_{tq} - p_t p_q)(p_s \delta_{sq} - p_s p_q)}{p_r p_s p_t p_q} = O(R)$. This result can be found in Cui and Zhong (2018) and Cui and Zhong (2019).

Write

$$\tilde{T}_n = \frac{1}{n} \sum_{r=1}^R \frac{1}{p_r} \int_0^1 \left[\sum_{i=1}^n f_{i,n}(x, r) \right]^2 dx =: J_{1n} + J_{2n},$$

where

$$J_{1n} = \frac{1}{n} \sum_{i=1}^n \sum_{r=1}^R \frac{1}{p_r} \int f_{i,n}^2(x, r) dx,$$

and

$$J_{2n} = \frac{1}{n} \sum_{i \neq j} \sum_{r=1}^R \frac{1}{p_r} \int f_{i,n}(x, r) f_{j,n}(x, r) dx.$$

Note that,

$$EJ_{1n} = \sum_{r=1}^R \frac{1}{p_r} \int E \frac{(I(X_i \leq x) - x)^2 (I(Y_i = r) - p_r)^2}{x^{(n)}(1-x)^{(n)}} dx$$

$$\begin{aligned}
 &= \sum_{r=1}^R (1 - p_r) \int_0^1 \frac{x(1-x)}{x^{(n)}(1-x)^{(n)}} dx \\
 &= (R-1)(1 - n^{-1/8}),
 \end{aligned}$$

and

$$\begin{aligned}
 \text{Var}(J_{1n}) &= \frac{1}{n} \text{Var} \left(\sum_{r=1}^R \frac{1}{p_r} \int f_{1,n}^2(x, r) dx \right) \leq \frac{1}{n} E \left(\sum_{r=1}^R \frac{1}{p_r} \int f_{1,n}^2(x, r) dx \right)^2 \\
 &= \frac{1}{n} \left(\sum_{r,s} \frac{1}{p_r p_s} \int E[f_{1,n}^2(x, r) f_{1,n}^2(y, s)] dx dy \right) \\
 &\leq \frac{Cn^{1/8}}{n} \sum_{r,s} \frac{p_r \delta_{rs} + p_r p_s (p_r + p_s)}{p_r p_s} \\
 &\leq \frac{C}{n^{7/8}} \left(\frac{R}{\min p_r} + R \right) = O(n^{-3/8}) = o(1).
 \end{aligned}$$

Hence,

$$\begin{aligned}
 E \left(\frac{J_{1n} - (R-1)}{\sqrt{2 \left(\frac{\pi^2}{3} - 3 \right) (R-1)}} \right)^2 &= C \{ \text{Var}(J_{1n}) / (R-1) + [EJ_{1n} - (R-1)]^2 / (R-1) \} \\
 &= C \text{Var}(J_{1n}) / (R-1) + C(R-1)n^{-1/4} = o(1),
 \end{aligned}$$

where C is a constant. Next, we only need to show that

$$\frac{J_{2n}}{\sqrt{2 \left(\frac{\pi^2}{3} - 3 \right) (R-1)}} \xrightarrow{d} N(0, 1).$$

Note that $EJ_{2n} = 0$, and

$$\begin{aligned}
 \text{Var}(J_{2n}) &= E(J_{2n}^2) \\
 &= \frac{1}{n^2} \sum_{i \neq j} \sum_{k \neq l} \sum_{r,s} \frac{1}{p_r p_s} \int E[f_{i,n}(x, r) f_{j,n}(x, r) f_{k,n}(y, s) f_{l,n}(y, s)] dx dy \\
 &= \frac{2n(n-1)}{n^2} \sum_{r,s} \frac{1}{p_r p_s} \int \{E[f_{1,n}(x, r) f_{1,n}(y, s)]\}^2 dx dy \\
 &= \frac{2n(n-1)}{n^2} \sum_{r,s} \frac{(p_r \delta_{rs} - p_r p_s)^2}{p_r p_s} \int \frac{(x \wedge y - xy)^2}{x^{(n)}(1-x)^{(n)}y^{(n)}(1-y)^{(n)}} dx dy \\
 &= \left(1 - \frac{1}{n}\right) (R-1) \left[2 \int \frac{(x \wedge y - xy)^2}{x(1-x)y(1-y)} dx dy + O(n^{-1/8}) \right] \\
 &= \left(1 - \frac{1}{n}\right) (R-1) [2(\pi^2/3 - 3) + O(n^{-1/8})].
 \end{aligned}$$

The last equality holds because

$$\int_0^1 \int_0^1 \frac{(x \wedge y - xy)^2}{x(1-x)y(1-y)} dx dy = \frac{\pi^2}{3} - 3.$$

Let $\mathcal{F}_i = \sigma\{(X_1, Y_1), \dots, (X_i, Y_i)\}$ be the σ -field generated by a set of random variables $\{(X_1, Y_1), \dots, (X_i, Y_i)\}$, $i = 1, \dots, n$. We see that

$$\begin{aligned}
 \frac{J_{2n}}{\sqrt{2 \left(\frac{\pi^2}{3} - 3 \right) (R-1)}} &= \frac{\sum_{i=2}^n \left[\frac{2}{n} \sum_{j=1}^{i-1} \sum_{r=1}^R \frac{1}{p_r} \int f_{i,n}(x, r) f_{j,n}(x, r) dx \right]}{\sqrt{2 \left(\frac{\pi^2}{3} - 3 \right) (R-1)}} \\
 &=: \sum_{i=2}^n Z_{ni}
 \end{aligned}$$

is the summation of a martingale difference sequence with $E(Z_{ni}) = 0$ and $\text{Var}(\sum_{i=2}^n Z_{ni}) = (1 - \frac{1}{n})(1 + O(n^{-1/8})) \rightarrow 1$. According to Hall and Heyde (1980), we need to prove $\sum_{i=2}^n E[Z_{ni}^2 | \mathcal{F}_{i-1}] \xrightarrow{P} 1$.

$$E[Z_{ni}^2 | \mathcal{F}_i] = \frac{1}{2(\pi^2/3 - 3)(R - 1)} \left(\frac{2}{n}\right)^2 \times \sum_{j,k}^{i-1} \sum_{r,s}^R \frac{1}{p_r p_s} \int \int E[f_{i,n}(x, r) f_{i,n}(y, s)] f_{j,n}(x, r) f_{k,n}(y, s) dx dy.$$

Thus we have

$$\sum_{i=2}^n E[Z_{ni}^2 | \mathcal{F}_{i-1}] = J_{3n} + J_{4n},$$

where

$$J_{3n} = \frac{1}{2(\pi^2/3 - 3)(R - 1)} \left(\frac{2}{n}\right)^2 \times \sum_{j=1}^{n-1} (n - j) \sum_{r,s}^R \frac{1}{p_r p_s} \int \int E[f_{i,n}(x, r) f_{i,n}(y, s)] f_{j,n}(x, r) f_{j,n}(y, s) dx dy,$$

and

$$J_{4n} = \frac{2}{2(\pi^2/3 - 3)(R - 1)} \left(\frac{2}{n}\right)^2 \times \sum_{j < k \leq n} (n - k) \sum_{r,s}^R \frac{1}{p_r p_s} \int \int E[f_{i,n}(x, r) f_{i,n}(y, s)] f_{j,n}(x, r) f_{k,n}(y, s) dx dy.$$

Since $E(J_{3n}) \rightarrow 1$, and

$$\begin{aligned} \text{Var}(J_{3n}) &= \frac{C}{(R - 1)^2 n^4} \sum_{j=1}^{n-1} (n - j)^2 \times \text{Var} \left(\sum_{r,s}^R \frac{1}{p_r p_s} \int \int E[f_{i,n}(x, r) f_{i,n}(y, s)] f_{j,n}(x, r) f_{j,n}(y, s) dx dy \right) \\ &\leq \frac{C}{(R - 1)^2 n^4} \sum_{j=1}^{n-1} (n - j)^2 \times E \left(\sum_{r,s}^R \frac{1}{p_r p_s} \int \int E[f_{i,n}(x, r) f_{i,n}(y, s)] f_{j,n}(x, r) f_{j,n}(y, s) dx dy \right)^2 \\ &= \frac{C}{(R - 1)^2 n^4} \sum_{j=1}^{n-1} (n - j)^2 \times E \left(\sum_{r,s}^R \frac{p_r \delta_{rs} - p_r p_s}{p_r p_s} \int \int \frac{x \wedge y - xy}{\sqrt{x^{(n)}(1-x)^{(n)} y^{(n)}(1-y)^{(n)}}} f_{j,n}(x, r) f_{j,n}(y, s) dx dy \right)^2 \\ &\leq \frac{C}{(R - 1)^2 n^4} \sum_{j=1}^{n-1} (n - j)^2 R^2 \\ &\quad \times E \left\{ \sum_{r,s}^R \left(\frac{p_r \delta_{rs} - p_r p_s}{p_r p_s} \right)^2 \left(\int \int \frac{x \wedge y - xy}{\sqrt{x^{(n)}(1-x)^{(n)} y^{(n)}(1-y)^{(n)}}} f_{j,n}(x, r) f_{j,n}(y, s) dx dy \right)^2 \right\} \\ &\leq \frac{C}{(R - 1)^2 n^4} \sum_{j=1}^{n-1} (n - j)^2 \\ &\quad \times R^2 E \left\{ \sum_{r,s}^R \left(\frac{p_r \delta_{rs} - p_r p_s}{p_r p_s} \right)^2 \int \int \frac{(x \wedge y - xy)^2}{x^{(n)}(1-x)^{(n)} y^{(n)}(1-y)^{(n)}} f_{j,n}^2(x, r) f_{j,n}^2(y, s) dx dy \right\} \\ &\leq \frac{C}{(R - 1)^2 n^4} \sum_{j=1}^{n-1} (n - j)^2 R^2 \sum_{r,s}^R \left(\frac{p_r \delta_{rs} - p_r p_s}{p_r p_s} \right)^2 \int \int E[f_{j,n}^2(x, r) f_{j,n}^2(y, s)] dx dy \\ &\leq \frac{C'}{(R - 1)^2 n^4} \sum_{j=1}^{n-1} (n - j)^2 R^2 \sum_{r,s}^R \left(\frac{p_r \delta_{rs} - p_r p_s}{p_r p_s} \right)^2 n^{1/8} (p_r \delta_{rs} + p_r p_s (p_r + p_s)) \\ &\leq \frac{C'}{(R - 1)^2 n^4} \sum_{j=1}^{n-1} (n - j)^2 R^2 n^{1/8} \frac{R}{\min p_r} \\ &= O\left(n^{-7/8} \frac{R}{\min p_r}\right) = O(n^{-3/8}), \end{aligned}$$

where C and C' are constants. Thus $J_{3n} \rightarrow 1$. And $E(J_{4n}) = 0$, and

$$\begin{aligned} \text{Var}(J_{4n}) &= \frac{C}{R^2 n^4} \sum_{j < k, l < m} (n - k)(n - m) \\ &\quad \times \sum_{r,s}^R \sum_{t,q}^R E \left\{ \frac{1}{p_r p_s p_t p_q} \int \int E[f_{i,n}(x, r) f_{i,n}(y, s)] f_{j,n}(x, r) f_{k,n}(y, s) dx dy \right. \end{aligned}$$

$$\begin{aligned}
 & \times \int \int E [f_{i,n}(x', t) f_{i,n}(y', q)] f_{l,n}(x', t) f_{m,n}(y', q) dx' dy' \Big\} \\
 &= \frac{C}{R^2 n^4} \sum_{j < k, l < m} (n - k) (n - m) \sum_{r,s}^R \sum_{t,q}^R \frac{(p_r \delta_{rs} - p_r p_s) (p_t \delta_{tq} - p_t p_q)}{p_r p_s p_t p_q} \\
 & \quad \times \int \frac{(x \wedge y - xy) (x' \wedge y' - x' y')}{\sqrt{x^{(n)} (1 - x)^{(n)} y^{(n)} (1 - y)^{(n)} (x')^{(n)} (1 - x')^{(n)} (y')^{(n)} (1 - y')^{(n)}}} \\
 & \quad \times E [f_{j,n}(x, r) f_{k,n}(y, s) f_{l,n}(x', t) f_{m,n}(y', q)] dx dy dx' dy' \\
 &= \frac{C}{R^2 n^4} \sum_{j < k} (n - k) (n - k) \sum_{r,s}^R \sum_{t,q}^R \frac{(p_r \delta_{rs} - p_r p_s) (p_t \delta_{tq} - p_t p_q)}{p_r p_s p_t p_q} \\
 & \quad \times \int \frac{(x \wedge y - xy) (x' \wedge y' - x' y')}{\sqrt{x^{(n)} (1 - x)^{(n)} y^{(n)} (1 - y)^{(n)} (x')^{(n)} (1 - x')^{(n)} (y')^{(n)} (1 - y')^{(n)}}} \\
 & \quad \times E [f_{j,n}(x, r) f_{k,n}(y, s) f_{j,n}(x', t) f_{k,n}(y', q)] dx dy dx' dy' \\
 &\leq \frac{C}{R^2 n^4} \sum_{j < k} (n - k) (n - k) \times \sum_{r,s}^R \sum_{t,q}^R \frac{(p_r \delta_{rs} - p_r p_s) (p_t \delta_{tq} - p_t p_q) (p_r \delta_{rt} - p_r p_t) (p_s \delta_{sq} - p_s p_q)}{p_r p_s p_t p_q} \\
 &= \frac{C}{R^2 n^4} \sum_{k=2}^n (k - 1) (n - k)^2 O(R) = O(1/R).
 \end{aligned}$$

Thus, $J_{4n} \xrightarrow{P} 0$. On the other hand

$$\begin{aligned}
 \sum_{i=2}^n E(Z_{ni}^4) &\leq \sum_{i=2}^n \frac{C}{n^4 R^2} E \left[\sum_{j=1}^{i-1} \sum_{r=1}^R \frac{1}{p_r} \int f_{i,n}(x, r) f_{j,n}(x, r) dx \right]^4 \\
 &\leq \sum_{i=2}^n \frac{C}{n^4 R^2} \left(6 \binom{i-1}{2} + i - 1 \right) E \left[\sum_{r=1}^R \frac{1}{p_r} \int f_{1,n}(x, r) f_{2,n}(x, r) dx \right]^4 \\
 &\leq \frac{C'}{n R^2} E \left[\sum_{r=1}^R \frac{1}{p_r} \int f_{1,n}(x, r) f_{2,n}(x, r) dx \right]^4 \\
 &= \frac{C'}{n R^2} E \left[\sum_{r,s}^R \frac{1}{p_r p_s} \int f_{1,n}(x, r) f_{1,n}(y, s) f_{2,n}(x, r) f_{2,n}(y, s) dx dy \right]^2 \\
 &\leq \frac{C'}{n R^2} E \left[\sum_{r,s}^R \frac{1}{p_r p_s} \left(\int f_{1,n}^2(x, r) f_{1,n}^2(y, s) dx dy \right)^{1/2} \left(\int f_{2,n}^2(x, r) f_{2,n}^2(y, s) dx dy \right)^{1/2} \right]^2 \\
 &\leq \frac{C'}{n R^2} \left(\sum_{r,s}^R \frac{1}{p_r p_s} \int E [f_{1,n}^2(x, r) f_{1,n}^2(y, s)] dx dy \right)^2 \\
 &\leq \frac{C''}{n R^2} \left(\sum_{r,s}^R \frac{p_r \delta_{rs} + p_r p_s (p_r + p_s)}{p_r p_s} n^{1/8} \right)^2 \\
 &= \frac{C''}{n^{3/4} R^2} \left(\frac{R}{\min p_r} + R + 2 \right)^2 = O \left(\frac{1}{n^{3/4} (\min p_r)^2} \right) = o(1/R),
 \end{aligned}$$

where C, C' and C'' are constants. By the central limit theorem of the martingale difference (Hall & Heyde, 1980), we have

$$\frac{\tilde{T}_n - (R - 1)}{\sqrt{2(\pi^2/3 - 3)(R - 1)}} \xrightarrow{d} N(0, 1),$$

as $n \rightarrow \infty$. This completes the proof. ■