Electronic Journal of Statistics
Vol. 18 (2024) 4767–4794
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2315

Conditional independence testing for discrete distributions: Beyond χ^2- and G-tests

Ilmun Kim[1], **Matey Neykov**[2], **Sivaraman Balakrishnan**[3] and **Larry Wasserman**[4]

[1] *Department of Statistics and Data Science, Yonsei University,* *e-mail:* ilmun@yonsei.ac.kr

[2] *Department of Statistics and Data Science, Northwestern University,* *e-mail:* mneykov@northwestern.edu

[3] *Department of Statistics and Data Science, Carnegie Mellon University,* *e-mail:* siva@stat.cmu.edu

[4] *Department of Statistics and Data Science, Carnegie Mellon University,* *e-mail:* larry@stat.cmu.edu

Abstract: This paper is concerned with the problem of conditional independence testing for discrete data. In recent years, researchers have shed new light on this fundamental problem, emphasizing finite-sample optimality. The non-asymptotic viewpoint adapted in these works has led to novel conditional independence tests that enjoy certain optimality under various regimes. Despite their attractive theoretical properties, the considered tests are not necessarily practical, relying on a Poissonization trick and unspecified constants in their critical values. In this work, we attempt to bridge the gap between theory and practice by reproving optimality without Poissonization and calibrating tests using Monte Carlo permutations. Along the way, we also prove that classical asymptotic χ^2- and G-tests are notably sub-optimal in a high-dimensional regime, which justifies the demand for new tools. Our theoretical results are complemented by experiments on both simulated and real-world datasets. Accompanying this paper is an `R` package `UCI` that implements the proposed tests.

MSC2020 subject classifications: 62C20, 62G99, 62H17.
Keywords and phrases: Depoissonization, negative association, permutation tests, conditional independence, sample complexity.

Received October 2023.

1. Introduction

Conditional independence (CI) is the backbone of diverse fields in statistics, including graphical models [18, 31] and causal inference [44, 37, 22]. Among several benefits, this fundamental assumption allows us to simplify the structure of a model, thereby increasing interpretability and reducing computational costs. To justify the use of CI assumption, it is of considerable interest to test whether

arXiv: 2308.05373

two random variables X and Y are independent after accounting for the effect of another random variable Z. Due to its important role, the problem of CI testing has received much attention in the past decade, resulting in numerous exciting new developments [e.g., 12, 42, 8, 35, 40, 33]. See [32] and [16] for recent reviews. However, most of the recent work is dedicated to continuous data and the importance of discrete CI testing is relatively overlooked.

In discrete settings, two commonly used methods are the χ^2-test [38] and the G-test [34], and their asymptotic equivalence is well-known under regularity conditions [e.g., Chapter 14 of 10]. When X and Y are binary, the Cochran–Mantel–Haenszel test [2] is another popular method for CI testing. Despite their popularity, these methods are asymptotic in nature, frequently calibrated by their limiting null distributions. Therefore their finite-sample validity remains questionable. This miscalibration issue becomes more serious in high-dimensional regimes where the number of categories can be significantly larger than the sample size. Besides, the power of these methods is not well-understood except in classical fixed-dimensional settings.

For discrete CI testing, [13] put forward two testing algorithms and analyzed their sample complexity from a non-asymptotic perspective. Their sample complexity results are further complemented by matching lower bounds, demonstrating optimality of their procedures in some regimes. In spite of these technical advances, their approach poses several practical challenges. First, their results rely on a Poissonization trick where the sample size is treated as a Poisson random variable. This assumption greatly simplifies the theoretical analysis, but is untenable in practice. Another issue worth highlighting is the dependence of the test on unspecified constants in their critical values. In many statistical applications, the type I error is a greater concern than the type II error. It is therefore desirable to set a critical value in such a way as to maximize the power, while *tightly* controlling the type I error. However, it is unclear from [13] how to modify their tests to meet this criteria, thereby leaving room for improvement from a practical perspective. Indeed, this issue was the main motivation of recent work of [28, 29] that advocates the use of permutation methods in two-sample and (both unconditional and conditional) independence testing problems.

With these issues in mind, our work makes the following contributions: (i) In Theorem 4.1, we depoissonize the sample complexity results of [13] and establish the same theoretical guarantees under the standard sampling setting. On a technical level, the challenge lies in dealing with the complicated dependence structure of multinomial samples. We overcome this difficulty using the negative association property of multinomial distributions [24]. (ii) In Section 3.2, we introduce a refined version of the general CI tester described in [13, Section 5]. This refinement reduces the number of splits from three to two, thereby utilizing the data more efficiently while maintaining the same theoretical guarantee. (iii) We further make the algorithms of [13] practical by leveraging the permutation method to calibrate test statistics. This resampling approach completely removes the issue arising from unspecified constants, and provably controls the finite-sample type I error. In Theorem 4.2, we prove that Monte Carlo permutation tests achieve the same sample complexity as the theoretical tests of [13].

This result is achieved by leveraging the general recipe for analyzing Monte Carlo permutation tests introduced in [28]. (iv) The considered test statistics are linear combinations of fourth order U-statistics, which can be daunting computationally. We address this computational concern by presenting alternative linear time expressions in Proposition 1. (v) We also prove an independent result that demonstrates sub-optimality of asymptotic χ^2- and G-tests in their power performance. This negative result naturally inspires efforts to develop new CI tests that perform better than the classical ones. (vi) Finally, we provide extensive simulation results that demonstrate the practical value of the proposed methods in Section 5, and the algorithms are available in the `R` package `UCI`.[1] To the best of our knowledge, this work is the first to investigate the optimal sample complexity for CI testing without Poissonization, supported by empirical results.

Our work is related to [45] who warn about the risk of asymptotic calibration for χ^2- and G-tests, and further highlight benefits of the permutation procedure in type I error control. The risk of asymptotic calibration has also been discussed in other testing problems, such as those studied in [4, 5, 28]. In line with this research, we prove the negative result of asymptotic χ^2- and G-tests, and demonstrate attractive properties of the permutation method both in type I and II error control. Another related work is [7] where the authors propose a permutation test based on a U-statistic for unconditional independence testing. Concurring with our view, [7] put an emphasis on the permutation approach for practical calibration and demonstrate the competitive performance of their proposal, coined `USP` test, over χ^2- and G-tests. In fact, when the conditional variable is degenerate (i.e., Z takes a single value), one of our practical proposals becomes exactly the same as that of [7]. In this sense, our work can be considered as an extension of [7] to CI testing. We also refer to [1, 49] that discuss exact inference methods for contingency tables. It is worth pointing out that the current paper builds on our prior work [29], which proves that the sample complexity results of [13] continue to hold using permutation tests. However, the analysis of [29] relies on Poissonization and also makes use of (computationally expensive) full permutation tests. The current work deviates from [29] by removing Poissonization and employing a more computationally efficient permutation test via Monte Carlo sampling. As part of this effort, we depoissonize several lemmas of [13] in Appendix B [30], which may be useful in other contexts. We also propose a new permutation test, called `wUCI`-test, that avoids sample splitting and achieves the optimal sample complexity in certain regimes. We illustrate its competitive finite sample performance under a variety of settings.

Organization The rest of this paper is organized as follows. In Section 2, we set the stage by presenting some background information on sample complexity and Poissonization. Section 3 describes the test statistics that we study, and verifies that they can be computed in linear time. Section 4 contains our main theoretical results including depoissonization and sub-optimality of χ^2- and G-

[1] Publicly available at github repository: https://github.com/ilmunk/UCI

tests. In Section 5, we demonstrate the empirical performance of the proposed methods based on simulated and real-world datasets, before concluding in Section 6. All the proofs of our results are relegated to the appendix [30].

Notation For a positive integer a, we use the shorthand $[a] = \{1, \ldots, a\}$. The conditional independence of X and Y given Z is denoted as $X \perp\!\!\!\perp Y \mid Z$. Given two discrete distributions p and q, we write the L_1 distance between p and q as $\|p - q\|_1$. We say that random variables $X_1, \ldots, X_n$ are i.i.d. when they are independent and identically distributed. For two positive sequences a_n and b_n, we write $a_n \asymp b_n$ if it holds that $C_1 \leq a_n/b_n \leq C_2$ for some positive constants C_1 and C_2, and for all n. We also write $a_n = O(b_n)$ or $a_n \lesssim b_n$ to indicate that $a_n \leq Cb_n$ for some positive constant C independent of n.

2. Background

Before presenting our main results, we start by building some background knowledge on sample complexity and Poissonization.

2.1. *Setting the stage*

Consider the set of discrete distributions of (X, Y, Z) on a domain $[\ell_1] \times [\ell_2] \times [d]$, denoted by $\mathcal{P}$. Let $\mathcal{P}_0$ be the subset of $\mathcal{P}$ such that $X \perp\!\!\!\perp Y \mid Z$. Given n i.i.d. random vectors $\{(X_i, Y_i, Z_i)\}_{i=1}^n$ drawn from $p_{X,Y,Z} \in \mathcal{P}$, our goal is to distinguish

$$H_0 : p_{X,Y,Z} \in \mathcal{P}_0 \quad \text{versus} \quad H_1 : p_{X,Y,Z} \in \mathcal{P}_1(\varepsilon) = \Big\{ p \in \mathcal{P} : \inf_{q \in \mathcal{P}_0} \|p - q\|_1 \geq \varepsilon \Big\}, \tag{1}$$

where $\varepsilon > 0$ is a distance parameter (Figure 1 for a pictorial description).

We point out that the class of alternatives $\mathcal{P}_1(\varepsilon)$ is defined with respect to the L_1 distance or equivalently twice the total variation (TV) distance:

$$\mathrm{TV}(p, q) = \sup_A |p(A) - q(A)| = \frac{1}{2}\|p - q\|_1,$$

where the supremum is taken over all possible measurable sets. The TV distance has a clear probabilistic interpretation as being the maximum absolute difference of the probabilities assigned to the same event by two distributions. Moreover, it is a bounded metric and remains invariant under bijective transformations. Lastly, as an f-divergence, the TV distance comes with the data-processing inequality, which ensures that transformations cannot introduce additional information. All of these desirable properties have prompted the use of the TV distance in distribution testing, which we also follow in the present work.

When the sample size n is too small, no valid test can reliably differentiate the null from the alternative. On the other hand, when the sample size n is too

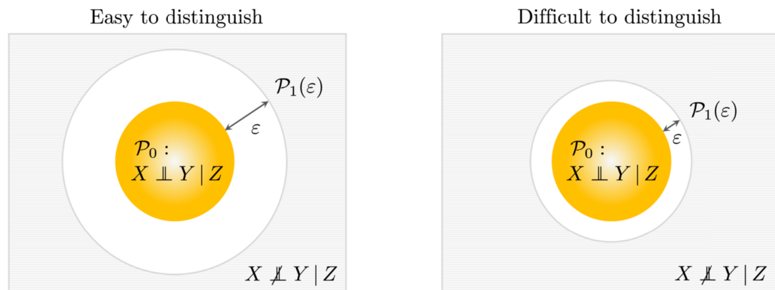

FIG 1. *Schematic of the hypotheses of CI testing. The space of null distributions is ε far from the space of alternative distributions where the distance parameter ε controls the difficulty of the problem.*

large, the problem becomes trivial, resulting in many successful tests. A natural question is then to determine the smallest n required to achieve the desired level of testing accuracy for an optimal test. This concept is known as the optimal sample complexity formally defined below.

Optimal sample complexity We define a test ϕ which maps from the samples, potentially augmented with external randomness R, to a binary outcome as $\phi : \{(X_i, Y_i, Z_i)\}_{i=1}^n \cup R \mapsto \{0, 1\}$. For some fixed $\alpha \in (0, 1)$, let Φ_α denote the set of level α tests such that for each $\phi \in \Phi_\alpha$,

$$\sup_{p \in \mathcal{P}_0} \mathbb{P}_p(\phi = 1) \leq \alpha.$$

The minimax risk is the worst-case type II error of an optimal level α test defined as

$$R_n(\varepsilon, \alpha) = \inf_{\phi \in \Phi_\alpha} \sup_{p \in \mathcal{P}_1(\varepsilon)} \mathbb{P}_p(\phi = 0).$$

The difficulty of the problem can be characterized as the minimum number of samples that makes the minimax risk bounded by some fixed constant $\beta \in (0, 1 - \alpha)$.[2] This minimum number of samples is called the optimal sample complexity given as

$$n^\star = \inf\Big\{ n : R_n(\varepsilon, \alpha) \leq \beta \Big\}.$$

We say that a level α test ϕ is *rate-optimal* in sample complexity if

$$n^\star \asymp \inf\Big\{ n : \sup_{p \in \mathcal{P}_1(\varepsilon)} \mathbb{P}_p(\phi = 0) \leq \beta \Big\}.$$

In practice, an optimal test whose risk is exactly equal to $R_n(\varepsilon, \alpha)$ is mostly inaccessible. For this reason, we instead aim to find a rate-optimal test.

[2]Throughout this paper, we treat the target type I and II errors α and β as universal constants, e.g., $\alpha = \beta = 0.05$.

Remark 1.

- (*Finite-sample optimality*) We highlight that the finite-sample minimax optimality considered in this paper is distinct from traditional asymptotic optimality. In classical asymptotic theory, the performance of a test is typically measured against contiguous alternatives, which converge to the null hypothesis at a $\sqrt{n}$-rate. While this type of local asymptotic analysis enables precise power comparisons, it does not provide useful insight when the underlying distribution is outside of regular, fixed-dimensional models. This limitation has been emphasized in [3] and [4] that advocate for the finite-sample minimax framework as in our work. This framework provides quantitative information in both low- and high-dimensional cases, offering a valuable alternative to traditional asymptotic approaches.
- (*Choice of metrics*) The optimal sample complexity depends crucially on the choice of metrics in the definition of the alternative hypothesis (1). We focus on the L_1 distance throughout this paper, due to its desirable properties mentioned earlier, and because our primary goal is to depoissonize previous results derived under the L_1 distance. Nevertheless, alternative metrics may be more suitable in certain contexts. For example, if p and q are known to differ in a few sparse bins, the L_∞ distance could offer a more efficient sample complexity than the L_1 distance. Alternatively, the Wasserstein distance is often preferred for mixed-type data due to its ability to effectively measure differences between diverse distributions. Depending on the metric, the construction of an optimal test needs to change by incorporating geometric properties of the chosen metric. As a concrete example, we refer the reader to the recent work of [36] that investigates the optimal sample complexity for CI testing in terms of the Wasserstein distance using a weighted multi-resolution U-statistic.

2.2. Poissonization

Poissonization is now a standard technique to study the sample complexity in hypothesis testing [e.g., 48, 15, 47, 5]. The idea itself is old in statistics and probability theory, dating back at least to [25]. See also Chapter 3.5 of [46]. It is used in theoretical works on hypothesis testing as a trick to simplify several calculations in dealing with categorical data. In particular, it is well-known that when the data are generated by Poisson sampling (Algorithm 1), the number of samples falling into disjoint sets are mutually independent. This independence property lies at the heart of deriving various existing results of the sample complexity in distribution testing. See [14] for a recent review.

2.3. General recipe and related issue

If constant factors are not the main concern, there are straightforward ways of transferring the sample complexity obtained from Poisson sampling to the usual

Algorithm 1 Poisson Sampling

Input: For a fixed $n \in \mathbb{N}$ and a distribution $p_{X,Y,Z}$

1. Draw $\tilde{N} \sim \text{Poisson}(n)$.
2. Generate i.i.d. $(X_1, Y_1, Z_1), \ldots, (X_{\tilde{N}}, Y_{\tilde{N}}, Z_{\tilde{N}})$ random variables from $p_{X,Y,Z}$.

Return: $\{(X_i, Y_i, Z_i)\}_{i=1}^{\tilde{N}}$

sampling scenario with a fixed sample size. One concrete procedure, described in [35], is as follows. Given n i.i.d. copies of (X, Y, Z),

1. Draw $\tilde{N} \sim \text{Poisson}\big(\frac{n}{2}\big)$.
2. If $\tilde{N} > n$, then accept the null hypothesis.
3. If $\tilde{N} \leq n$, perform a test based on $(X_1, Y_1, Z_1), \ldots, (X_{\tilde{N}}, Y_{\tilde{N}}, Z_{\tilde{N}})$ and return a result.

This procedure leverages the concentration property of a Poisson random variable to its mean, which implies that the probability of observing $\tilde{N} > n$ in the second step is small especially for large n. As a result, one can proceed to the third step with high probability and analyze the sample complexity under Poisson sampling of size $\tilde{N}$. To make this idea concrete, let ϕ be a generic test function using $(X_1, Y_1, Z_1), \ldots, (X_{\tilde{N}}, Y_{\tilde{N}}, Z_{\tilde{N}})$ and then the resulting test from the above procedure can be concretely written as $\phi^* = \mathbb{1}(\tilde{N} \leq n)\phi$. Suppose that the test ϕ has the type I error as well as the type II error bounded by α and β, respectively. Then the corresponding test ϕ^* has the type I and II error bounds as

$$\sup_{p \in \mathcal{P}_0} \mathbb{E}_p[\phi^*] \leq \alpha \quad \text{and} \quad \sup_{p \in \mathcal{P}_1(\varepsilon)} \mathbb{E}_p[1 - \phi^*] \leq \beta + \mathbb{P}[\tilde{N} > n].$$

Since a Poisson random variable is tightly concentrated around its mean, the additional term in the type II error of ϕ^* can be made small when n is relatively large compared to β.[3] Therefore one can transfer the sample complexity obtained under Poissonization to the usual sampling scheme up to a constant factor.

Nevertheless, due to an inefficient use of the data as well as a non-trivial chance of accepting the null irrelevant to the data, practitioners may not necessarily follow this general recipe. To address this concern, there has been an effort to depoissonize the results of distribution testing under Poisson sampling. For instance, [26] revisits the truncated χ^2 test for goodness-of-fit testing proposed by [5] and establishes the same optimality without Poissonization. We also refer to [23] and Chapter 2.5 of [39] that present useful depoissonization tools, but mostly for asymptotic results. In this work, we analyze the tests proposed by [13] under the usual i.i.d. sampling model with a fixed sample size and show that the same sample complexity can be achieved without any further assumption.

[3]More precisely, an exponential tail bound for a Poisson random variable ensures that $\mathbb{P}(\tilde{N} > n) \leq e^{-n/8}$ [e.g., Theorem A.8 of 14] so that if $n \geq 8\log(1/\beta)$, the type II error is bounded above by 2β.

At the heart of our technique is the negative association of multinomial distributions [24], which would be potentially useful for depoissonizing other sample complexity results.

Multinomial sampling. To fix terminology, we refer to the usual sampling with fixed sample size n as *multinomial sampling* in what follows.

3. Test statistics

To describe our main results, we first need to recall the test statistics introduced by [13]. As noted by [35], their test statistics can be viewed as linear combinations of U-statistics constructed using the observations of (X, Y) in the same category of Z. We describe two kinds of U-statistics considered in [13] and our modifications.

3.1. Unweighted U-statistic

Suppose that we observe $N \geq 4$ i.i.d. observations of (X, Y) supported on $[\ell_1] \times [\ell_2]$. We let $p_{X,Y}$ denote the joint discrete distribution of (X, Y) and let p_X and p_Y denote the marginal distribution of X and Y, respectively. The first U-statistic is an unbiased estimator of the squared L_2 distance between $p_{X,Y}$ and $p_X p_Y$. In more detail, mostly borrowing the notation from [35], let

$$\phi_{ij}(q,r) = \mathbb{1}(X_i = q)\mathbb{1}(Y_i = r) - \mathbb{1}(X_i = q)\mathbb{1}(Y_j = r).$$

For four distinct observations indexed by $i, j, k, l \in [N]$, the kernel of the unweighted U-statistic is defined as

$$h_{ijkl} = \frac{1}{4!} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4)\in\Pi} \sum_{q\in[\ell_1],r\in[\ell_2]} \phi_{\pi_1\pi_2}(q,r)\phi_{\pi_3\pi_4}(q,r),$$

where Π is the set of all possible permutations of (i, j, k, l). By the linearity of expectations, we see that h_{ijkl} is an unbiased estimator of the squared L_2 distance between $p_{X,Y}$ and $p_X p_Y$. Given this kernel and the dataset $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, the unweighted U-statistic is computed as

$$U(\mathcal{D}) = \binom{N}{4}^{-1} \sum_{i<j<k<l:(i,j,k,l)\in[N]} h_{ijkl}. \tag{2}$$

It is worth pointing out that $U(\mathcal{D})$ is equivalent to the U-statistic for unconditional independence testing considered in [6, 7, 28]. Denote the datasets $\mathcal{D}_m = \{(X_i, Y_i) : Z_i = m, i \in [n]\}$ and write the sample size within $\mathcal{D}_m$ by N_m. The final test statistic for CI testing based on $U(\mathcal{D})$ is then constructed as

$$T = \sum_{m\in[d]} \mathbb{1}(N_m \geq 4) N_m U(\mathcal{D}_m). \tag{3}$$

3.2. Weighted U-statistic

The previous unweighted U-statistic may suffer from high variance especially when the L_2 norms of $p_{X,Y}$ and $p_X p_Y$ are large, resulting in sub-optimal performance (see the simulation result in Figure 2 under Scenario 1). To address this issue, [13] employ the flattening idea proposed by [19]. This approach involves partitioning heavier bins into multiple smaller pieces in a way to reduce the L_2 norms of the transformed distributions, leading to a smaller variance of the test statistic. As noted in [4] and further elaborated by [35], the flattening procedure is equivalent to using a carefully designed weighted kernel for a U-statistic.

To describe the weighted U-statistic considered in [13], assume that the sample size $N = 4 + 4t$ for some $t \in \mathbb{N}$, and let $t_1 = \min(t, \ell_1)$ and $t_2 = \min(t, \ell_2)$. The construction of the weighted U-statistic involves three-fold splitting where the dataset $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ is split into $\tilde{\mathcal{D}}_X = \{X_i : i \in [t_1]\}$, $\tilde{\mathcal{D}}_Y = \{Y_i : t_1 + 1 \leq i \leq t_1 + t_2\}$ and $\tilde{\mathcal{D}}_{X,Y} = \{(X_i, Y_i) : 2t + 1 \leq i \leq N\}$ of sizes t_1, t_2 and $4 + 2t$, respectively. Define $\tilde{a}_q$ (and $\tilde{a}'_r$) as the number of observations equal to q (and r) in $\tilde{\mathcal{D}}_X$ (and $\tilde{\mathcal{D}}_Y$). For four distinct observations indexed by i, j, k, l in $\tilde{\mathcal{D}}_{X,Y}$, consider a weighted kernel given as

$$h^{\tilde{\mathbf{a}}}_{ijkl} = \frac{1}{4!} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4) \in \Pi} \sum_{q \in [\ell_1], r \in [\ell_2]} \frac{\phi_{\pi_1\pi_2}(q, r)\phi_{\pi_3\pi_4}(q, r)}{(1 + \tilde{a}_q)(1 + \tilde{a}'_r)}.$$

Given this kernel, the weighted U-statistic proposed in [13] is computed as

$$U^{\tilde{\mathbf{a}}}_W(\mathcal{D}) = \binom{2t + 4}{4}^{-1} \sum_{i<j<k<l : (i,j,k,l) \in \mathcal{D}_{X,Y}} h^{\tilde{\mathbf{a}}}_{ijkl},$$

where $(i, j, k, l) \in \mathcal{D}_{X,Y}$ indicates taking four observations indexed by (i, j, k, l) from the dataset $\mathcal{D}_{X,Y}$. It is worth noting that, due to the independence among the split datasets, the conditional expectation of $h^{\tilde{\mathbf{a}}}_{ijkl}$ taken over $\mathcal{D}_{X,Y}$ is the square of the L_2 distance between $p_{X,Y}$ and $p_X p_Y$ weighted by $(1+\tilde{a}_q)(1+\tilde{a}'_r)$.

As before, denote the datasets $\mathcal{D}_m = \{(X_i, Y_i) : Z_i = m, i \in [n]\}$ and write the sample size within $\mathcal{D}_m$ by N_m. Introducing $\omega_m = \sqrt{\min(N_m, \ell_1)\min(N_m, \ell_2)}$, the final test statistic proposed in [13] is given as

$$T_W = \sum_{m \in [d]} \mathbb{1}(N_m \geq 4) N_m \omega_m U^{\tilde{\mathbf{a}}}_W(\mathcal{D}_m). \tag{4}$$

However, the construction of T_W uses the dataset rather inefficiently. It relies on three-fold splitting and discards a portion of samples when $\ell_1, \ell_2 < t$ in the computation of $U^{\tilde{\mathbf{a}}}_W(\mathcal{D})$. This motivates our proposal below.

Refined weighted U-statistic We now enhance the previous weighted U-statistic by reducing the three-fold splits to two-fold splits. Additionally, our approach does not discard the samples $\{(X_i, Y_i) : t_1 + t_2 + 1 \leq i \leq 2t\}$ when

$t_1, t_2 < t$. As we will demonstrate later, this improvement can be achieved while maintaining the same theoretical guarantees.

To elaborate, consider the dataset $\mathcal{D} = \{(X_1, Y_1), \dots, (X_N, Y_N)\}$ with $N \geq 4$ and split it into $\mathcal{D}_{(1)} = \{(X_i, Y_i) : 1 \leq i \leq N_{(1)}\}$ and $\mathcal{D}_{(2)} = \{(X_i, Y_i) : N_{(1)} + 1 \leq i \leq N\}$ where $N_{(1)} = N - N_{(2)}$ and $N_{(2)} = 4 + \lfloor (N-4)/2 \rfloor$. In contrast to $\tilde{a}_q$ and $\tilde{a}'_r$, which are based on two independent splits, we define a_q and a'_r based on the single dataset $\mathcal{D}_{(1)}$ as

$$a_q = \frac{\ell_1}{N_{(1)}} \sum_{i=1}^{N_{(1)}} \mathbb{1}(X_i = q) \quad \text{and} \quad a'_r = \frac{\ell_2}{N_{(1)}} \sum_{i=1}^{N_{(1)}} \mathbb{1}(Y_i = r),$$

when $N_{(1)} \geq 1$. If $N_{(1)} = 0$, we set $a_q = a'_r = 0$. For four distinct observations indexed by i, j, k, l in $\mathcal{D}_{(2)}$, define a refined weighted kernel as

$$h^{\boldsymbol{a}}_{ijkl} = \frac{1}{4!} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4)\in\Pi} \sum_{q\in[\ell_1], r\in[\ell_2]} \frac{\phi_{\pi_1\pi_2}(q,r)\phi_{\pi_3\pi_4}(q,r)}{(1+a_q)(1+a'_r)}.$$

Given this kernel, we compute the weighted U-statistic as

$$U^{\boldsymbol{a}}_W(\mathcal{D}) = \binom{N_{(2)}}{4}^{-1} \sum_{i<j<k<l:(i,j,k,l)\in\mathcal{D}_{(2)}} h^{\boldsymbol{a}}_{ijkl}. \tag{5}$$

Recalling $\mathcal{D}_m = \{(X_i, Y_i) : Z_i = m, i \in [n]\}$, a refined version of T_W is then constructed as

$$T_{W^*} = \sum_{m\in[d]} \mathbb{1}(N_m \geq 4) N_m \omega_m U^{\boldsymbol{a}}_W(\mathcal{D}_m). \tag{6}$$

Plug-in approach We also propose another version of the test statistic that does not involve sample splitting. As demonstrated in Section 5, the test based on T_{W^*} may experience a loss of practical power in small sample size regimes because the main body of the statistic and weights are computed using separate datasets. To mitigate this issue, we consider a weight kernel computed without sample splitting:

$$h^{\boldsymbol{b}}_{ijkl} = \frac{1}{4!} \sum_{(\pi_1,\pi_2,\pi_3,\pi_4)\in\Pi} \sum_{q\in[\ell_1], r\in[\ell_2]} \frac{\phi_{\pi_1\pi_2}(q,r)\phi_{\pi_3\pi_4}(q,r)}{(1+b_q)(1+b'_r)},$$

where $b_q = \ell_1 N^{-1} \sum_{i=1}^N \mathbb{1}(X_i = q)$ and $b'_r = \ell_2 N^{-1} \sum_{i=1}^N \mathbb{1}(Y_i = r)$. The resulting weighted U-statistic is then computed as

$$U^{\boldsymbol{b}}_W(\mathcal{D}) = \binom{N}{4}^{-1} \sum_{i<j<k<l:(i,j,k,l)\in[N]} h^{\boldsymbol{b}}_{ijkl}. \tag{7}$$

Similarly as before, the final CI test statistic based on $U^{\boldsymbol{b}}_W(\mathcal{D})$ is given as

$$T_{W,\mathrm{plug}} = \sum_{m\in[d]} \mathbb{1}(N_m \geq 4) N_m \omega_m U^{\boldsymbol{b}}_W(\mathcal{D}_m). \tag{8}$$

Remark 2 (Connection with the truncated χ^2-test)**.** As pointed out by several authors [21, 5, 26], the classical χ^2-test for goodness-of-fit testing can easily break down for sparse multinomial data. To address this problem, [5] introduce a modification of the χ^2-test by using a truncated weight function and prove its minimax optimality. Interestingly, the weight $(1+b_q)(1+b'_r)$ (and also $(1+a_q)(1+a'_r)$) that we consider is closely connected to the truncated weight of [5] and may be regarded as an empirical counterpart for independence testing. To explain, let us simply focus on the first term in the product weight and notice that

$$\ell_1^{-1}(1+b_q) = \frac{1}{\ell_1} + \widehat{p}_X(q) \asymp \max\left\{\frac{1}{\ell_1}, \widehat{p}_X(q)\right\},$$

where $\widehat{p}_X(q) = N^{-1}\sum_{i=1}^N \mathbb{1}(X_i = q)$. The right-hand side of the above equation is exactly the same as the truncated weight in [5] for goodness-of-fit testing.

3.3. Linear time expression

The original forms of the aforementioned U-statistics take $O(N^4\ell_1\ell_2)$ time to compute, which can be prohibitive for large N, ℓ_1, ℓ_2. Luckily, this computational complexity can be reduced to $O(N)$ by exploiting a contingency table representation. A computationally convenient form of the unweighted U-statistic $U(\mathcal{D})$ is already given by [13, 6, 7]. We also note that an alternative form of $U_W^{\boldsymbol{a}}(\mathcal{D})$ is provided in [13], but a naive calculation of their expression takes at least $O(N\ell_1^2\ell_2^2)$ time. Here we present a general expression for the U-statistics and explain that it can be run in linear time on average *independent of* ℓ_1 and ℓ_2.

To this end, let us set some notation. Let $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_{\ell_1}\}$ and $\boldsymbol{\upsilon} = \{\upsilon_1, \ldots, \upsilon_{\ell_2}\}$ be some weight vectors with non-zero components. Given the dataset $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$, consider four distinct observations indexed by $i, j, k, l \in [N]$ and define the weighted kernel associated with $\boldsymbol{\eta}$ and $\boldsymbol{\upsilon}$ as

$$h_{ijkl}^{\boldsymbol{\eta},\boldsymbol{\upsilon}} = \frac{1}{4!}\sum_{(\pi_1,\pi_2,\pi_3,\pi_4)\in\Pi}\ \sum_{q\in[\ell_1], r\in[\ell_2]} \frac{\phi_{\pi_1\pi_2}(q,r)\phi_{\pi_3\pi_4}(q,r)}{\eta_q\upsilon_r}.$$

It is clear that the above kernel is equivalent to h_{ijkl} when $\boldsymbol{\eta}$ and $\boldsymbol{\upsilon}$ are vectors with each entry equal to one. Similarly, when $\boldsymbol{\eta}$ and $\boldsymbol{\upsilon}$ are defined with $1+a_q$ and $1+a'_r$, respectively, then the above kernel corresponds to $h_{ijkl}^{\boldsymbol{a}}$. The U-statistic based on $h_{ijkl}^{\boldsymbol{\eta},\boldsymbol{\upsilon}}$ is denoted by $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(\mathcal{D})$, which is similarly computed as in (2). For $q \in [\ell_1]$ and $r \in [\ell_2]$, define $o_{qr} = \sum_{i=1}^n \mathbb{1}(X_i = q)\mathbb{1}(Y_i = r)$, $o_{q+} = \sum_{r=1}^{\ell_2} o_{qr}$ and $o_{+r} = \sum_{q=1}^{\ell_1} o_{qr}$. With this notation in place, we give an alternative expression of $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(\mathcal{D})$ as follows.

Proposition 1 (Alternative expression)**.** *The U-statistic* $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(\mathcal{D})$ *can be written as*

$$U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(\mathcal{D}) \ = \ \frac{1}{N(N-3)}\left[A_1 + \frac{1}{(N-1)(N-2)}A_2 - \frac{2}{N-2}A_3\right]$$

where

$$A_1 = \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\left(\frac{o_{qr}^2 - o_{qr}}{\eta_q \upsilon_r}\right), \quad A_2 = \sum_{q=1}^{\ell_1}\left(\frac{o_{q+}^2 - o_{q+}}{\eta_q}\right)\cdot\sum_{r=1}^{\ell_2}\left(\frac{o_{+r}^2 - o_{+r}}{\upsilon_r}\right),$$

$$A_3 = \sum_{q=1}^{\ell_1}\sum_{r=1}^{\ell_2}\frac{o_{qr}(o_{q+}o_{+r} - o_{q+} - o_{+r} + 1)}{\eta_q \upsilon_r}.$$

We now discuss the average-case time complexity of $U_W^{\boldsymbol{\eta},\boldsymbol{\upsilon}}(\mathcal{D})$ by assuming that the weight vectors $\boldsymbol{\eta}, \boldsymbol{\upsilon}$ are given in advance. First of all, we note that the $\ell_1 \times \ell_2$ contingency table of $\mathcal{D}$ is sparse in a sense that it has at most N non-zero entries. Importantly, the zero entries do not affect the calculation of A_1, A_2, A_3. Hence we only focus on the non-zero entries of the contingency table, which can be computed in linear time by using hash tables [e.g., Chapter 11 of 17]. Similarly, the non-zero row sums and the non-zero column sums of the contingency table can be computed in linear time independent of ℓ_1 and ℓ_2. Given the non-zero entries of $\{o_{qr}, o_{q+}, o_{+r} : q \in [\ell_1], r \in [\ell_2]\}$ stored in hash tables, the computational complexity of the terms A_1, A_2, A_3 is linear since their non-zero summands are at most $O(N)$ and the average-case time complexity of retrieving entries from hash tables is $O(1)$. See Algorithm 5 in Appendix A for a more concrete description.

As mentioned, this linear-time complexity is an average-case guarantee. In worst-case scenarios, the time complexity can degrade to $O(N^2)$ due to hash collisions. However, with a well-designed hash function and a well-dimensioned hash table, one can mitigate worst-case scenarios, and the performance remains efficient in practice.

For the unweighted U-statistic $U(\mathcal{D})$, there is no additional cost for computing $\boldsymbol{\eta}, \boldsymbol{\upsilon}$ as they are vectors with each entry equal to one. For the weighted U-statistics $U_W^{\boldsymbol{a}}(\mathcal{D})$, $U_W^{\tilde{\boldsymbol{a}}}(\mathcal{D})$ and $U_W^{\boldsymbol{b}}(\mathcal{D})$, the weight vectors $\boldsymbol{\eta}, \boldsymbol{\upsilon}$ are functions of o_{q+} and o_{+r} (computed on a separate dataset for $U_W^{\boldsymbol{a}}(\mathcal{D})$) and they only require an additional $O(N)$ time to compute. Thus the overall time complexity is still linear on average.

4. Theoretical results

Having introduced the test statistics, we are now ready to provide the main theoretical results of this paper. In Section 4.1, we establish the sample complexity of the tests using T, T_W, T_{W^*} and $T_{W,\text{plug}}$ under multinomial sampling. In Section 4.2, we provide and analyze more practical tests based on permutation procedures.

4.1. Sample complexity without Poissonization

In this subsection, we revisit two main results of [13], namely Theorem 1.1 and Theorem 1.3 concerning with the sample complexity of a test using T in (3) and T_W in (4), respectively.

Sample complexity of a test based on T in (3): Suppose that the test statistic T is constructed using N i.i.d. samples from $p_{X,Y,Z}$ where $N \sim \text{Poisson}(n)$. We reject the null of CI when $T > \zeta\sqrt{\min(n,d)}$ for a sufficiently large constant $\zeta > 0$. Then for $\ell_1 = \ell_2 = 2$, Theorem 1.1 of [13] proves that the resulting test has the sample complexity

$$O\left(\max\left\{\frac{d^{1/2}}{\varepsilon^2},\ \min\left\{\frac{d^{7/8}}{\varepsilon},\ \frac{d^{6/7}}{\varepsilon^{8/7}}\right\}\right\}\right). \tag{9}$$

They also prove that this sample complexity is rate-optimal by presenting a matching lower bound.

Sample complexity of a test based on T_W in (4): The test based on T is not necessarily optimal in the high-dimensional regime where ℓ_1 and ℓ_2 can vary with other parameters. As shown in Theorem 1.3 of [13], a more general result of the sample complexity can be derived by using T_W. In particular, given N i.i.d. samples from $p_{X,Y,Z}$ where $N \sim \text{Poisson}(n)$, we reject the null of CI when $T_W > \zeta'\sqrt{\min(n,d)}$ for a sufficiently large constant $\zeta' > 0$. With $\ell_1 \geq \ell_2$, the sample complexity of the resulting test satisfies

$$\begin{aligned} O\Bigg(\max\Bigg\{&\min\left\{\frac{d^{7/8}\ell_1^{1/4}\ell_2^{1/4}}{\varepsilon},\ \frac{d^{6/7}\ell_1^{2/7}\ell_2^{2/7}}{\varepsilon^{8/7}}\right\}, \\ &\frac{d^{3/4}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon},\ \frac{d^{2/3}\ell_1^{2/3}\ell_2^{1/3}}{\varepsilon^{4/3}},\ \frac{d^{1/2}\ell_1^{1/2}\ell_2^{1/2}}{\varepsilon^2}\Bigg\}\Bigg). \end{aligned} \tag{10}$$

Moreover, this upper bound is shown to be optimal, up to constant factors, in a number of regimes. See [13] for a discussion.

We now depoissonize the previous results and establish the same sample complexity under multinomial sampling.

Theorem 4.1 (Multinomial sampling)**.** *Suppose that we observe $\mathcal{D}_n = \{(X_1,Y_1,Z_1),\ldots,(X_n,Y_n,Z_n)\}$ i.i.d. samples from $p_{X,Y,Z}$ with nonrandom sample size n. Then the following three statements hold:*

1. *Set T^* to be either T or $T_{W,\text{plug}}$. Compute T^* based on $\mathcal{D}_n$ and reject the null if $T^* > \zeta\sqrt{\min(n,d)}$ for a sufficiently large constant $\zeta > 0$. Then the resulting test has the sample complexity as in (9) when $\ell_1, \ell_2 = O(1)$.*
2. *Set T^* to be either T_W or T_{W^*}. Compute T^* based on $\mathcal{D}_n$ and reject the null if $T^* > \zeta'\sqrt{\min(n,d)}$ for a sufficiently large constant $\zeta' > 0$. Then the resulting test has the sample complexity as in (10).*

A few remarks are in order.

Remark 3.

- The above theorem essentially says that the tests based on T and T_W have the same performance in sample complexity under Poisson sampling and multinomial sampling. In particular, it means that they require the same number of samples, up to a constant factor, to achieve the desired testing error under both Poisson sampling and multinomial sampling. This result may not come as a surprise given that a Poisson random variable is strongly concentrated around its mean. See empirical evidence in [29]. However, the proof under multinomial sampling turns out to be nontrivial, requiring a careful analysis.

- The above result also presents theoretical guarantees for the tests based on the proposed statistics T_{W^*} and $T_{W,\text{plug}}$. First of all, we prove that the test utilizing T_{W^*} achieves the same sample complexity as T_W, while using the dataset more efficiently. Second, we establish that the test based on $T_{W,\text{plug}}$ achieves the optimal sample complexity when $\ell_1, \ell_2 = O(1)$. Based on our empirical results, it seems plausible that this plug-in approach maintains comparable sample complexity as T_{W^*} even when ℓ_1, ℓ_2 increase. However, proving this formally may require different techniques, and is thus warranted for future work.
- One of the main technical hurdles in the proof is to overcome a lack of independence between random variables in different bins when the sample size is no longer Poisson. The independence property is useful in analyzing the variance of the sum of U-statistics as it leads to zero covariance terms. We address the lack of independence by employing the negative association (NA) property of multinomial random vectors [24]. This NA property ensures that the covariance terms are non-positive, which turns out to be enough to ensure the same theoretical guarantees hold under multinomial sampling.
- Another technical challenge arises when analyzing the variance of a nonlinear function of a Binomial random variable, such as $N_m \mathbb{1}(N_m \geq 4)\omega_m$ in (6). Unlike a Poisson random variable, whose moments are fully determined by a single rate parameter, a Binomial random variable depends on two parameters, making the analysis significantly more complex. To simplify the variance analysis, we leverage the key observation that a Binomial random variable can be written as a sum of i.i.d. Bernoulli random variables. This allows us to employ the Efron–Stein inequality (Lemma B.10), which is suitable for bounding the variance of a non-linear function of independent random variables.
- Even though we remove Poissonization, the resulting tests are not necessarily practical. In particular, their critical values depend on unspecified constants ζ and ζ'. The choice of these constants, resulting in tight control of the type I error, is challenging in practice. We take a further step to address this issue via the permutation method in Section 4.2, and demonstrate their empirical performance in Section 5.

Algorithm 2 `UCI`: U-statistic CI test

1: **Input:** Sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, the number of permutations B, significance level α
2: **for** $j \in [B]$ **do**
3: **for** $m \in [d]$ **do**
4: Generate $\pi \sim \text{Uniform}(\Pi_{N_m})$ independent of everything else.
5: Compute $U(\mathcal{D}_m^\pi)$ as in (2) based on the permuted dataset $\mathcal{D}_m^\pi$.
6: **end for**
7: Set $T_j \leftarrow \sum_{m\in[d]} \mathbb{1}(N_m \geq 4) N_m U(\mathcal{D}_m^\pi)$.
8: **end for**
9: Set $T \leftarrow \sum_{m\in[d]} \mathbb{1}(N_m \geq 4) N_m U(\mathcal{D}_m)$ computed without permutations.
10: Compute the permutation p-value

$$p_{\text{perm}} = \frac{1}{B+1}\left[\sum_{j=1}^{B} \mathbb{1}(T_j \geq T) + 1\right].$$

11: **Output:** Reject H_0 if $p_{\text{perm}} \leq \alpha$; otherwise, accept H_0.

4.2. Calibration via permutations

As mentioned earlier, the tests used in Theorem 4.1 depend on unspecified constants, which raises the issue of practicality. This section attempts to address this problem by presenting more practical tests calibrated by the permutation method, and examine their sample complexity under multinomial sampling. In particular, we prove that their sample complexity remains the same as the corresponding (theoretical) tests in Theorem 4.1. As briefly mentioned earlier, a similar result was established in Theorem 5 of [29] but under Poisson sampling. In contrast, we do not assume that the sample size follows a Poisson distribution and therefore reduce the gap between theory and practice. We start by describing the testing procedures that we analyze.

Permutation test using T in (3) This first test compares the test statistic T with its permutation correspondences, and rejects the null if the resulting permutation p-value is less than or equal to significance level α. To further explain, let Π_{N_m} denote the set of all permutations of $[N_m]$ for each $m \in [d]$. Given π drawn from Π_{N_m}, we define $\mathcal{D}_m^\pi$ by rearranging Y values in $\mathcal{D}_m$ according to π. More specifically, suppose that we have $\mathcal{D}_m = \{(X_1, Y_1), \ldots, (X_{N_m}, Y_{N_m})\}$. Then the corresponding permuted dataset becomes $\mathcal{D}_m^\pi = \{(X_1, Y_{\pi_1}), \ldots, (X_{N_m}, Y_{\pi_{N_m}})\}$. Equipped with this notation, we implement Algorithm 2 and make a decision based on the output. We coin this test as `UCI`-test.

Permutation test using T_{W^*} in (6) The second test that we analyze calculates its p-value by comparing T_W with its permutation correspondences. The overall procedure is similar to the previous one except that it utilizes the half-permutation procedure for a technical reason. To explain the procedure, we decompose $\mathcal{D}_m$ into $\mathcal{D}_{(1),m}$ and $\mathcal{D}_{(2),m}$ of size $N_m - 4 - \lfloor (N_m-4)/2 \rfloor$ and $4 + \lfloor (N_m-4)/2 \rfloor$, respectively, as in Section 3.2. Unlike Algorithm 2, we only

Algorithm 3 `wUCI_split`: weighted U-statistic CI test using sample splitting

Input: Sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, the number of permutations B, significance level α
for $j \in [B]$ **do**
  for $m \in [d]$ **do**
    Generate $\pi \sim \text{Uniform}(\Pi_{4+\lfloor (N_m-4)/2 \rfloor})$ independent of everything else.
    Define $\mathcal{D}_m^\pi = \mathcal{D}_{(1),m} \cup \mathcal{D}_{(2),m}^\pi$.
    Compute $U_W^{\boldsymbol{a}}(\mathcal{D}_m^\pi)$ as in (5) based on the permuted dataset $\mathcal{D}_m^\pi$.
  end for
  Set $T_{j,W^*} \leftarrow \sum_{m\in[d]} \mathbb{1}(N_m \geq 4) N_m \omega_m U_W^{\boldsymbol{a}}(\mathcal{D}_m^\pi)$.
end for
Set $T_{W^*} \leftarrow \sum_{m\in[d]} \mathbb{1}(N_m \geq 4) N_m \omega_m U_W^{\boldsymbol{a}}(\mathcal{D}_m)$ computed without permutations.
Compute the permutation p-value

$$p_{\text{perm}} = \frac{1}{B+1}\left[\sum_{j=1}^{B} \mathbb{1}(T_{j,W^*} \geq T_{W^*}) + 1\right].$$

Output: Reject H_0 if $p_{\text{perm}} \leq \alpha$; otherwise, accept H_0.

permute Y values within $\mathcal{D}_{(2),m}$ for each $m \in [d]$, resulting in $\mathcal{D}_{(2),m}^\pi$, and then evaluate the significance of T_{W^*}. As mentioned in Remark 6 of [29], this half-permutation procedure greatly simplifies the analysis by preserving the independence structure between $\mathcal{D}_{(1),m}$ and $\mathcal{D}_{(2),m}^\pi$. Moreover it ensures that the weights of the U-statistic in (5) remain invariant under permutations, and thus removes the randomness that would arise from different weights under full permutations. Additionally, the half-permutation test offers a computational advantage over the full-permutation test since there is no need to recompute weights for each permutation. A more detailed procedure is described in Algorithm 3. We refer to this test as `wUCI_split`-test.

Permutation test using $T_{W,\text{plug}}$ in (8) The third test computes its p-value by comparing $T_{W,\text{plug}}$ with its permutation correspondences. The procedure is essentially the same as for `UCI`-test except that we utilize $T_{W,\text{plug}}$ as our test statistic. A more detailed procedure is described in Algorithm 4, and we refer to this test as `wUCI`-test.

Having outlined the permutation procedures, we now discuss their sample complexity. First of all, it is noteworthy that both permutation tests are exact level α in any finite sample scenarios. This simply follows by the fact that the original test statistic and their permutation correspondences are exchangeable under the null. Using this observation, it can be seen that that the resulting p-values in Algorithms 2, 3 and 4 are super-uniform [e.g., Lemma 1 of 41]. The next theorem turns to the type II error and establishes their sample complexity.

Theorem 4.2 (Permutation tests)**.** *Suppose that we observe $\mathcal{D}_n = \{(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n)\}$ i.i.d. samples from $p_{X,Y,Z}$ with nonrandom sample size n. We also assume that the number of random permutations B sat-*

Algorithm 4 `wUCI`: weighted U-statistic CI test

Input: Sample $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, the number of permutations B, significance level α
for $j \in [B]$ **do**
  for $m \in [d]$ **do**
    Generate $\pi \sim \text{Uniform}(\Pi_{N_m})$ independent of everything else.
    Compute $U_W^{\boldsymbol{b}}(\mathcal{D}_m^\pi)$ as in (7) based on the permuted dataset $\mathcal{D}_m^\pi$.
  end for
  Set $T_{j,W,\text{plug}} \leftarrow \sum_{m\in[d]} \mathbb{1}(N_m \geq 4) N_m \omega_m U_W^{\boldsymbol{b}}(\mathcal{D}_m^\pi)$.
end for
Set $T_{W,\text{plug}} \leftarrow \sum_{m\in[d]} \mathbb{1}(N_m \geq 4) N_m \omega_m U_W^{\boldsymbol{b}}(\mathcal{D}_m)$ computed without permutations.
Compute the permutation p-value

$$p_{\text{perm}} = \frac{1}{B+1}\left[\sum_{j=1}^{B} \mathbb{1}(T_{j,W,\text{plug}} \geq T_{W,\text{plug}}) + 1\right].$$

Output: Reject H_0 if $p_{\text{perm}} \leq \alpha$; otherwise, accept H_0.

isfies $B \geq \max\{4(1-\alpha)\alpha^{-1}, 8\alpha^{-2}\log(4\beta^{-1})\}$ *where* α *and* β *are pre-specified type I and II errors, respectively. Then the following two statements hold:*

1. *The test from Algorithm 2 or Algorithm 4 has the sample complexity as in* (9) *when* $\ell_1, \ell_2 = O(1)$.
2. *The test from Algorithm 3 has the sample complexity as in* (10).

We highlight that the above theorem studies the random permutation tests where B is not required to increase with the sample size n. This is in contrast to full permutation tests considered in [29] that enumerate all possible permutations. This random permutation approach imposes additional technical challenges in the theoretical analysis of permutation tests due to the extra randomness inherent in the Monte Carlo procedure. In particular, it is a non-trivial task to determine the minimum number of Monte Carlo repetitions that ensures an error guarantee nearly equivalent to that of the full permutation test. To tackle this problem, we need to carefully measure the discrepancy between the full permutation distribution and its Monte Carlo counterpart via sharp concentration inequalities. To this end, we leverage the recent result of [28] in our analysis that uses the Dvoretzky–Kiefer–Wolfowitz inequality. We also note that the constant factors in the condition on B are not tight and can be improved at the expense of inflating a constant factor in the sample complexity. See Remark 5 in Appendix B for a discussion.

We next turn our attention to χ^2- and G-tests and discuss their sub-optimality.

4.3. Sub-optimality of χ^2- and G-tests

Practitioners frequently use χ^2- and G-tests for conditional independence, which have nice asymptotic properties in fixed dimensional settings. In this subsection, we move beyond this fixed dimensional case and prove that these classical tests are markedly sub-optimal in terms of sample complexity. To define the χ^2-test and G-test formally, let us write $o_{qrs} = \sum_{i=1}^{n} \mathbb{1}(X_i = q)\mathbb{1}(Y_i = r)\mathbb{1}(Z_i = s)$ and $e_{qrs} = o_{q+s}o_{+rs}/o_{++s}$ where $o_{q+s} = \sum_{r\in[\ell_2]} o_{qrs}$, $o_{+rs} = \sum_{q\in[\ell_1]} o_{qrs}$ and $o_{++s} = \sum_{q\in[\ell_1],r\in[\ell_2]} o_{qrs}$, respectively, for $q \in [\ell_1], r \in [\ell_2], s \in [d]$. Given this notation, the χ^2-test and G-test are based on the following test statistics

$$\begin{aligned}\chi^2 &= \sum_{q\in[\ell_1],r\in[\ell_2],s\in[d]} \frac{(o_{qrs}-e_{qrs})^2}{e_{qrs}} \quad \text{and} \\ G &= 2\sum_{q\in[\ell_1],r\in[\ell_2],s\in[d]} o_{qrs}\log\left(\frac{o_{qrs}}{e_{qrs}}\right).\end{aligned} \tag{11}$$

These test statistics converge to a χ^2 distribution with $(\ell_1 - 1) \times (\ell_2 - 1) \times d$ degrees of freedom under the null of conditional independence and under some regularity conditions [45]. Based on this asymptotic result, χ^2-test and G-test reject the null when χ^2 and G are larger than the $1-\alpha$ quantile of the χ^2 distribution with $(\ell_1 - 1) \times (\ell_2 - 1) \times d$ degrees of freedom. We first emphasize that these classical tests do not control the type I error uniformly over the null distributions and their validity guarantee requires that the sample size go to infinity. This is even true for the simplest case where $d = 1$, which corresponds to the unconditional independence problem [see 7, for details]. Moreover, their asymptotic power can be exactly equal to zero in some regimes where the sample size is much larger than the bound in (9) as shown below.

Proposition 2 (Sub-optimality of χ^2- and G-tests)**.** *Assume that $\ell_1 = \ell_2 = 2$, $\varepsilon = 0.25$ and $\alpha \in (0,1)$ is some fixed constant. Further assume that $d = n \times r_n$ where r_n is an arbitrary positive sequence that increases to infinity as $n \to \infty$. In this scenario, the worst case power of χ^2- and G-tests approach zero as*

$$\lim_{n\to\infty} \inf_{p\in\mathcal{P}_1(\varepsilon)} \mathbb{P}_p(\chi^2 > q_{1-\alpha,d}) = 0 \quad \textit{and} \quad \lim_{n\to\infty} \inf_{p\in\mathcal{P}_1(\varepsilon)} \mathbb{P}_p(G > q_{1-\alpha,d}) = 0,$$

where $q_{1-\alpha,d}$ denotes the $1-\alpha$ quantile of the χ^2 distribution with d degrees of freedom.

We provide some remarks on this result.

Remark 4.

- As shown in Theorem 4.2, the proposed tests can achieve significant power (indeed rate-optimal when $\ell_1 = \ell_2 = 2$) under the same scenario and $r_n \lesssim n^{1/6}$. On the other hand, χ^2- and G-tests have asymptotically zero power for any r_n that increases to infinity, which highlights sub-optimality of these classical tests. Moreover, our {`UCI`, `wUCI`, `wUCI_split`} tests are valid over the entire class of null distributions unlike asymptotic χ^2- and G-tests.
- At a high-level, the reason behind this negative result is that the critical values of χ^2- and G-tests are not adaptive to the underlying distribution.

More specifically, we can think of a setting where most of conditional bins are empty with high probability, i.e., the intrinsic dimension of Z is much smaller than d. In this case, it is more natural to use a critical value that reflects the intrinsic dimension rather than the ambient dimension. However, the χ^2- and G-tests do not take this intuition into account, and their test statistics can be much smaller than $q_{1-\alpha,d}$ under the alternative. This leads to asymptotically zero power as we formally prove in Appendix C.4.

- This issue can be alleviated by using the permutation approach where empty bins are ignored in calibration [45]. Nevertheless, it is unknown whether the permutation-based χ^2- and G-tests are optimal or not in terms of sample complexity. Our numerical results in Section 5 indicate that the permutation-based χ^2-test becomes almost powerless in specific settings (e.g., Scenario 2), suggesting that it may not achieve optimal sample complexity. Conversely, the permutation-based G-test demonstrates more robust performance across various scenarios, although it is significantly outperformed by our methods in certain cases (e.g., Scenario 3 and Scenario 6). Future research is warranted to delve deeper into this topic and assess the optimality of these permutation-based tests.

5. Numerical analysis

In this section, we provide numerical results that compare the proposed tests (`UCI` in Algorithm 2, `UCI_split` in Algorithm 3 and `wUCI`-test in Algorithm 4) with χ^2-test and G-test under various scenarios. For a fair comparison, we calibrate both χ^2-test and G-test using the permutation method and reject the null when their permutation p-values are less than or equal to the significance level α. Throughout our simulations, we set $\alpha = 0.05$ and the number of permutations $B = 199$. All the power values reported in this section are estimated by Monte Carlo simulation with 10,000 repetitions.

5.1. Simulated data examples

We start by comparing the power of the considered tests based on synthetic datasets. We only focus on the power results given that all of the tests are calibrated by the permutation procedure, resulting in valid type I error control in any finite sample sizes. There are eight different scenarios that we consider under the alternative where the domain sizes of X, Y, Z are set to $\ell_1 = \ell_2 = 20$ and $d = 10$, respectively. The considered scenarios are described as follows.

- **Scenario 1.** Set p_Z to be uniform over $[d]$. For each $z \in [d]$, (i) first let $p_{X,Y|Z}(x, y \,|\, z) \propto x^{-2}y^{-2}$, (ii) then replace $p_{X,Y|Z}(\ell_1, \ell_2 \,|\, z)$ with 0.015, and (iii) finally normalize $p_{X,Y|Z}$ to have its sum to be one. This setting results in a strong signal in χ^2 divergence but relatively weaker signal in the L_2 distance over bins.

- **Scenario 2.** Set p_Z to be uniform over $[d]$. For each $z \in [d]$, (i) let $p_{X,Y|Z}(x,y\,|\,z) \propto x^{-2}y^{-2}$, (ii) set $\delta = \min\{p_{X,Y|Z}(x,y\,|\,z) : x \in [2], y \in [2]\}$, and (iii) perturb $p_{X,Y|Z}$ by replacing $p_{X,Y|Z}(x,y\,|\,z)$ with $p_{X,Y|Z}(x,y\,|\,z) + (-1)^{x+y}\delta$ for $x \in [2]$ and $y \in [2]$. The resulting probability vector has a small signal in χ^2 divergence, but relatively stronger signal in the L_2 distance over bins.
- **Scenario 3.** Set p_Z to be uniform over $[d]$. For each $z \in [d]$, (i) set $p_{X,Y|Z}(x,y\,|\,z) = 0$ for all $x \in [\ell_1]$ and $y \in [\ell_2]$, (ii) set $p_{X,Y|Z}(1,1\,|\,z) = (1-q)^2$, $p_{X,Y|Z}(1,y\,|\,z) = (1-q)q(\ell_1-1)^{-1}$ for $y \in [\ell_2]\setminus\{1\}$, $p_{X,Y|Z}(x,1\,|\,z) = (1-q)q(\ell_1-1)^{-1}$ for $x \in [\ell_1]\setminus\{1\}$, $p_{X,Y|Z}(x,y\,|\,z) = q^2(\ell_1-1)^{-1}$ for $x = y \in [\ell_1]\setminus\{1\}$ where $q = 0.2$. This simulation setting is borrowed from [50].
- **Scenario 4.** Set p_Z to be uniform over $[d]$. For each $z \in [d]$, let $p_{X,Y|Z}(x,y\,|\,z) = \{1+(-1)^{x+y}\}\ell_1^{-1}\ell_2^{-1}$ be a perturbed uniform distribution. This is the setting where χ^2-test, G-test and `UCI`-test perform similarly for unconditional independence testing. See Figure 5 of [7].
- **Scenario 5.** Set p_Z to be uniform over $[d]$. For $z = 1$, let $p_{X,Y|Z}(x,y\,|\,z) = 0.25$ for $x \in [2], y \in [2]$ and zero otherwise. In other words, there is no signal in the first category of Z. On the other hand, for $z \in [d]\setminus\{1\}$, set $p_{X,Y|Z}(x,y\,|\,z) = \{1+(-1)^{x+y}\}\ell_1^{-1}\ell_2^{-1}$ as in Scenario 4.
- **Scenario 6.** Set p_Z to be uniform over $[d]$. For $z = 1$, $p_{X,Y|Z}(1,1\,|\,z) = p_{X,Y|Z}(2,2\,|\,z) = 0.4$, $p_{X,Y|Z}(1,2\,|\,z) = p_{X,Y|Z}(2,1\,|\,z) = 0.1$ and zero otherwise. On the other hand, for $z \in [d]\setminus\{1\}$, set $p_{X,Y|Z}(x,y\,|\,z) = \ell_1^{-1}\ell_2^{-1}$, i.e., $X \perp\!\!\!\perp Y$ for $z \in [d]\setminus\{1\}$, resulting in a sparse alternative.
- **Scenario 7.** Set $p_Z(z) \propto z^{-1}$. For $z \in [d]$, let $p_{X,Y|Z}(x,y\,|\,z) = \{1+(-1)^{x+y}z^{-1}\}\ell_1^{-1}\ell_2^{-1}$. By construction, the signal becomes weaker as z increases and the sample size N_z tends to be smaller as z increases.
- **Scenario 8.** $p_Z(z) \propto z^{-1}$. For $z \in [d]$, let $p_{X,Y|Z}(x,y\,|\,z) = \{1+(-1)^{x+y}(d-z+1)^{-1}\}\ell_1^{-1}\ell_2^{-1}$. Note that the signal becomes stronger as z increases, and the sample size N_z tends to be smaller as z increases. To put it in another way, this is the reverse setting of Scenario 7.

The results are presented in Figure 2 and Figure 3. It is clear from the results that no test outperforms the others over all scenarios. In particular, χ^2-test has the highest power when the underlying distribution has a strong signal in χ^2 divergence such as Scenario 1, and similarly, `UCI`-test performs well when there is a strong signal in the L_2 distance such as Scenario 2. It is also interesting to observe that `wUCI`-test and `wUCI_split`-test show an impressive performance compared to the others in Scenario 3, and all of the tests except `wUCI_split`-test perform similarly in Scenario 4 where the corresponding null distribution of $(X,Y)\,|\,Z$ is uniform over bins. In Scenario 5 and Scenario 6, we are essentially in a situation where $\ell_1 = \ell_2 = 2$ for $z = 1$ and $\ell_1 = \ell_2 = 20$ for $z \in [d]\setminus\{1\}$. In these scenarios, the first bin of Z plays a less important role than the other bins in determining the value of χ^2 and G statistics (these statistics have higher variance when ℓ_1 and ℓ_2 are large). In contrast, the test statistic for the `UCI`-test is mostly

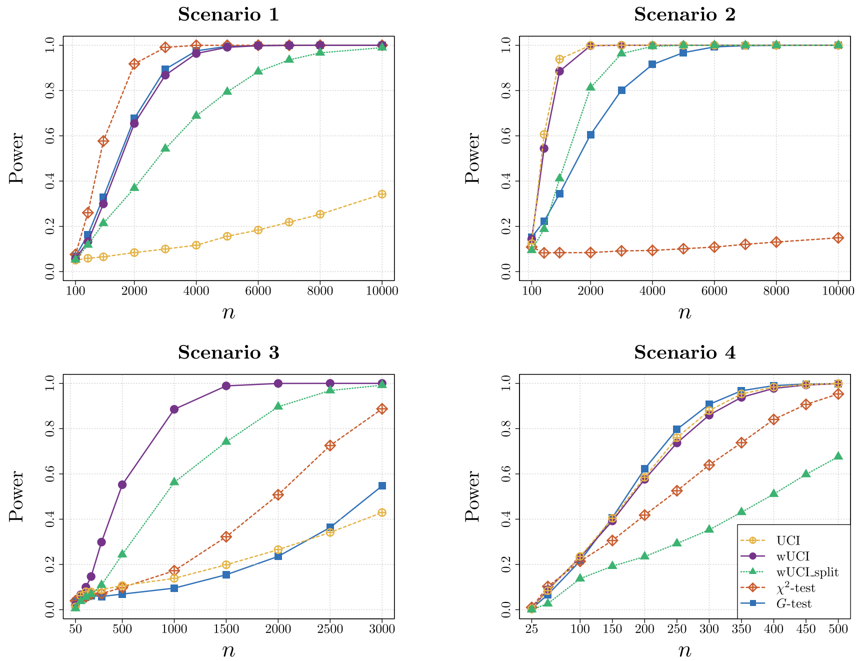

Fig 2. *Power comparisons of the considered tests in Scenarios 1–4 described in Section 5.1.*

dominated by the data from the first bin of Z in the same scenarios. This explains the outstanding performance of χ^2-test and G-test in Scenario 5 where signals are spread out over bins except the first one. On the other hand, χ^2-test and G-test have low power in Scenario 6 where only the first bin of Z has a signal. Under the same scenarios, `UCI`-test behaves in the opposite way, attaining high power when the first bin is significant. Another interesting observation is that `wUCI`-test performs as powerful as G-test in Scenario 5 whereas it outperforms both χ^2- and G-tests in Scenario 6 when the sample size is large. This may be explained by the choice of weights in its statistic that roughly interpolate χ^2 weights and uniform weights as explained in Remark 2. We also note that the test statistic for `UCI`-test is a linear combination of U-statistics weighted by sample sizes over bins. This explains the relatively lower power of `UCI`-test than χ^2- and G-tests in Scenario 7 where the bins with smaller sample sizes tend to have stronger signals. In contrast, we observe the opposite behavior in Scenario 8. On the other hand, `wUCI`-test performs the second best in both Scenario 7 and Scenario 8. Lastly, we highlight that the performance of `wUCI_split` closely follows that of `wUCI`-test when the sample sizes in bins with a strong signal are sufficiently large such as in Scenarios 1–3. However, its performance degrades when the sample size is small as observed in other scenarios. Based on these findings, we recommend using the `wUCI`-test over the `wUCI_split`-test, despite the additional theoretical guarantees provided by the latter.

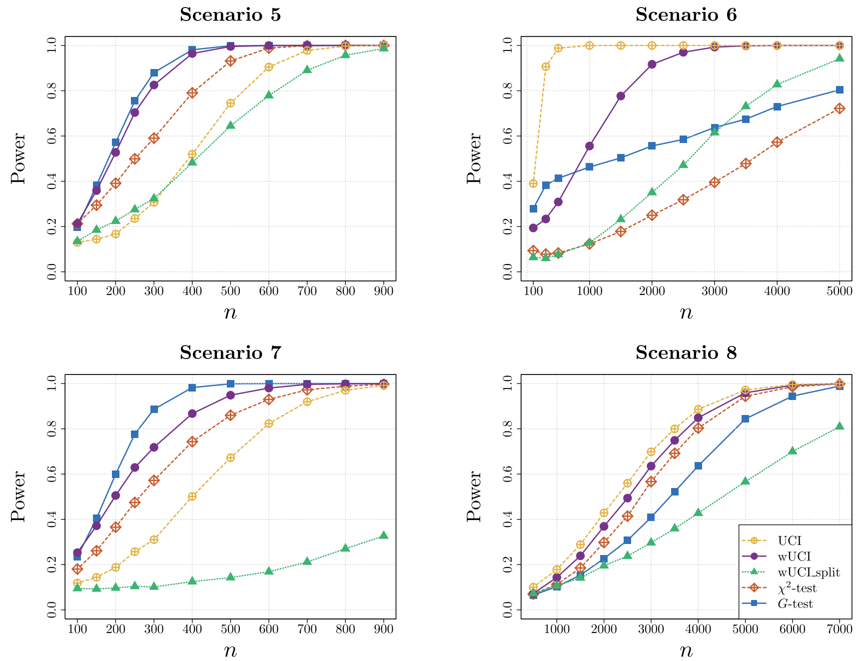

FIG 3. *Power comparisons of the considered tests in Scenarios 5–8 described in Section 5.1.*

To summarize, we observe that different tests perform better than the others under different scenarios. The proposed tests often dominate the classical ones when there are strong signals, especially in the L_2 distance, over bins with large sample sizes. On the other hand, it is possible to design situations such as Scenario 1 where the proposed tests attain lower power than the classical ones. Nevertheless, the classical tests, especially χ^2-test, can fail badly in terms of the worst-case performance. By contrast, our proposals, especially `wUCI`-test, demonstrate robust performance across different scenarios, indicating that they can work as practical tools that complement classical χ^2-test and G-test under various scenarios.

5.2. Real-world data examples

We next provide numerical illustrations based on real-world datasets.

Admission dataset The first dataset that we look at is the Berkeley admissions dataset, which is a well-known example of Simpson's paradox [9]. As summarized in Table 1, the dataset consists of 4,526 applications with 3 variables (X, Y, Z) where X and Y are binary variables, representing the gender (male or female) and the admission status (admitted or rejected), respectively. The conditional variable Z takes the department name among {A, B, C, D, E, F}. When the dataset is aggregated over the departments, it appears that male applicants are more likely to be admitted than woman applicants. However, as reported by [9], there seems to exist a bias in favor of women when looking at the individual departments, indicating the existence of conditional dependence. We assess this claim of conditional dependence by implementing the considered permutation tests. For this dataset, the corresponding p-values are computed as follows: 0.005 for both χ^2- and G-tests, 0.04 for `UCI`, 0.03 for both `wUCI` and 0.05 for `wUCI_split`, respectively. All the p-values are significant at level $\alpha = 0.05$, revealing evidence of conditional dependence.

Table 1
Admissions data at University of California, Berkeley from the six largest departments in 1973.

	Men		Women	
Major	Applicants	Admitted	Applicants	Admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7

Diamonds dataset Next we consider the diamonds dataset available in `R` package `ggplot2`. The dataset contains the information of 53,940 diamonds including their price, clarity, color, quality of the cut, etc. In our analysis, the price variable is partitioned into 100 intervals of equal size. We set the corresponding categorized price variable as X and set the clarity variable as Y. The clarity variable has 8 categories (I1, SI2, SI1, VS2, VS1, VVS2, VVS1, IF) and it measures the purity of a diamond. The conditional variable Z is chosen to be either the cut variable or the color variable in our analysis. Both variables are discrete with 5 (Fair, Good, Very Good, Premium, Ideal) and 7 (D, E, F, G, H, I, J) categories, respectively. In the experiments, we treat the entire dataset as the population (thereby the ground truth is known to us)[4] from which we randomly draw n observations without replacement. Based on this subsample of size n, we compute the permutation p-values of the considered tests and we repeat this process 10,000 times to estimate their power. The results can be found in Figure 4 where we collect the power of $\{\texttt{UCI}, \texttt{wUCI}, \texttt{wUCI\_split}, \chi^2, G\}$-tests by changing the sample size n. The left panel of Figure 4 provides the power results when the conditional variable is set to be the color variable. As can be seen, χ^2-test has the significantly lower power than the others. Among the other three tests, `wUCI` has the highest power followed by `UCI` while the difference is minor. We can see a similar pattern from the right panel of Figure 4 where the conditional variable is set to be the cut variable. These results highlight the practical value of the proposed tests in analyzing real-world datasets where classical tests potentially suffer from low power.

[4]We verified numerically that X and Y are conditionally dependent on Z by comparing the distribution of $p_{X,Y,Z}$ and $p_{X|Z}p_{Y|Z}p_Z$.

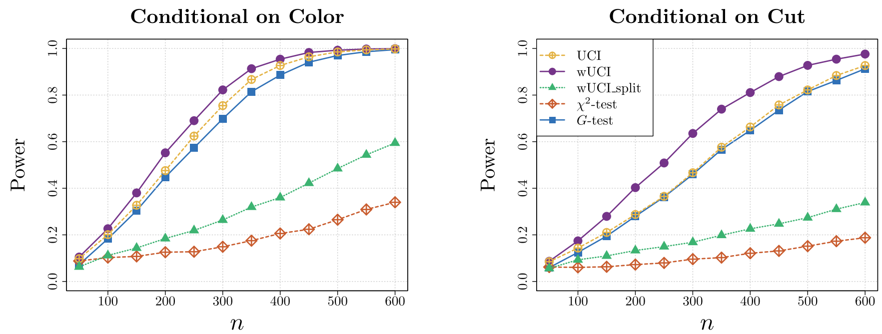

FIG 4. *Power comparisons of the considered tests based on the diamonds dataset. Both panels analyze independence between the (categorized) price and clarity variables conditional on the color variable and the cut variable, respectively. All of the tests have increasing power as the sample size increases. Markedly, χ^2-test and `wUCI_split`-test have significantly lower power than the other tests, whereas `wUCI`-test seems to perform the best for this dataset.*

6. Discussion

In this paper, we have revisited recent developments of CI testing for discrete data. Despite attractive theoretical properties, these recent tests have limited practical value, relying on Poissonization and unspecified constants in their critical values. In this work, we have made an attempt to bridge the gap between theory and practice by removing Poissonization and utilizing the Monte Carlo permutation method to calibrate test statistics. We have also complemented our theoretical results with a thorough numerical analysis and demonstrated certain benefits of the proposed tests over classical χ^2- and G-tests. Finally, we have developed `R` package `UCI` that implements the proposed methods.

Our work leaves several important avenues for future research. One prominent direction is to depoissonize other sample complexity results in the literature using the tools developed in this paper. For instance, one can reproduce the results of [35, 29] for continuous CI testing without Poissonization. Another direction which may be fruitful to pursue is to devise a CI test that incorporates prior information about potential alternative distributions. For example, suppose that we are in an alternative setting where only a handful of conditional categories are significant. In this case, it is possible to obtain a substantial power gain by using sparse weights in the proposed statistics, and one could analyze the resulting tests. Additionally, one could attempt to further bridge the gap between practice and theory. While our results suggest that `wUCI`-test achieves the optimal sample complexity when $\ell_1, \ell_2 = O(1)$, it is currently unknown whether it can achieve the optimal sample complexity in general regimes with increasing ℓ_1, ℓ_2. On the other hand, `wUCI_split`-test attains the general sample complexity as in (10), but it loses practical power due to inefficient use of the data. It is therefore interesting to explore the theoretical foundation of `wUCI`-test or other tests without sample splitting. We leave these interesting questions for future work.

Funding

We would like to thank the reviewers for their thoughtful comments that significantly improved our paper. This work was partially supported by funding from the NSF grants DMS-2113684 and DMS-2310632, as well as an Amazon AI and a Google Research Scholar Award to SB. MN acknowledges support from the NSF grant DMS-2113684. IK acknowledges support from the Yonsei University Research Fund of 2023-22-0419 as well as support from the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2022R1A4A1033384), and the Korea government (MSIT) RS-2023-00211073.

Supplementary Material

Supplement to "Conditional independence testing for discrete distributions: Beyond χ^2- and G-tests"
(doi: 10.1214/24-EJS2315SUPP; .pdf). This supplementary material includes the detailed proofs and technical lemmas omitted from the main text, as well as the pseudo-code for the proposed algorithm.

References

[1] Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science* **7** 131–153. MR1173420

[2] Agresti, A. (2003). *Categorical Data Analysis*. John Wiley & Sons. MR1914507

[3] Arias-Castro, E., Pelletier, B. and Saligrama, V. (2018). Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics* **30** 448–471. MR3794401

[4] Balakrishnan, S. and Wasserman, L. (2018). Hypothesis testing for high-dimensional multinomials: A selective review. *The Annals of Applied Statistics* **12** 727–749. MR3834283

[5] Balakrishnan, S. and Wasserman, L. (2019). Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *The Annals of Statistics* **47** 1893–1927. MR3953439

[6] Berrett, T. B., Kontoyiannis, I. and Samworth, R. J. (2021). Optimal rates for independence testing via U-statistic permutation tests. *The Annals of Statistics* **49** 2457–2490. MR4338371

[7] Berrett, T. B. and Samworth, R. J. (2021). USP: an independence test that improves on Pearson's chi-squared and the G-test. *Proceedings of the Royal Society A* **477** 20210549. MR4366500

[8] Berrett, T. B., Wang, Y., Barber, R. F. and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82** 175–197. MR4060981

[9] Bickel, P. J., Hammel, E. A. and O'Connell, J. W. (1975). Sex Bias in Graduate Admissions: Data from Berkeley: Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation. *Science* **187** 398–404.

[10] Bishop, Y. M., Fienberg, S. E. and Holland, P. W. (2007). *Discrete Multivariate Analysis: Theory and Practice*. Springer Science & Business Media. MR2344876

[11] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press. MR3185193

[12] Candés, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80** 551-577. MR3798878

[13] Canonne, C. L., Diakonikolas, I., Kane, D. M. and Stewart, A. (2018). Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*. *STOC 2018* 735–748. Association for Computing Machinery, New York, NY, USA. MR3826290

[14] Canonne, C. L. et al. (2022). Topics and techniques in distribution testing: a biased but representative sample. *Foundations and Trends® in Communications and Information Theory* **19** 1032–1198.

[15] Chan, S.-O., Diakonikolas, I., Valiant, P. and Valiant, G. (2014). Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms* 1193–1203. SIAM. MR3376448

[16] Chatterjee, S. (2022). A survey of some recent developments in measures of association. *arXiv preprint arXiv:2211.04702*.

[17] Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (2022). *Introduction to Algorithms*. MIT Press. MR2572804

[18] De Campos, L. M. and Huete, J. F. (2000). A new approach for learning belief networks using independence criteria. *International Journal of Approximate Reasoning* **24** 11–37. MR1759736

[19] Diakonikolas, I. and Kane, D. M. (2016). A new approach for testing properties of discrete distributions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)* 685–694. IEEE. MR3631031

[20] Gretton, A., Bousquet, O., Smola, A. and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory* 63–77. Springer. MR2255909

[21] Haberman, S. J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association* **83** 555–560. MR0971386

[22] Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press. MR3309951

[23] Jacquet, P. and Szpankowski, W. (1998). Analytical depoissonization

and its applications. *Theoretical Computer Science* **201** 1–62.

[24] Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *The Annals of Statistics* **11** 286–295. MR0684886

[25] Kac, M. (1949). On deviations between theoretical and empirical distributions. *Proceedings of the National Academy of Sciences* **35** 252–257. MR0029490

[26] Kim, I. (2020). Multinomial goodness-of-fit based on U-statistics: High-dimensional asymptotic and minimax optimality. *Journal of Statistical Planning and Inference* **205** 74–91. MR4011624

[27] Kim, I. (2022). Comments on "Testing conditional independence of discrete distributions". *arXiv preprint arXiv:2207.02819*.

[28] Kim, I., Balakrishnan, S. and Wasserman, L. (2022). Minimax optimality of permutation tests. *The Annals of Statistics* **50** 225–251. MR4382015

[29] Kim, I., Neykov, M., Balakrishnan, S. and Wasserman, L. (2022). Local permutation tests for conditional independence. *The Annals of Statistics* **50** 3388–3414. MR4524501

[30] Kim, I., Neykov, M., Balakrishnan, S. and Wasserman, L. (2024). Supplement to "Conditional independence testing for discrete distributions: Beyond χ^2- and G-tests". `https://doi.org/10.1214/24-EJS2315SUPP`.

[31] Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT Press. MR2816736

[32] Li, C. and Fan, X. (2020). On nonparametric conditional independence tests for continuous variables. *Wiley Interdisciplinary Reviews: Computational Statistics* **12** e1489. MR4098477

[33] Liu, M., Katsevich, E., Janson, L. and Ramdas, A. (2022). Fast and powerful conditional randomization testing via distillation. *Biometrika* **109** 277–293. MR4430958

[34] McDonald, J. H. (2014). G-test of goodness-of-fit. *Handbook of Biological Statistics* **487** 53–58.

[35] Neykov, M., Balakrishnan, S. and Wasserman, L. (2021). Minimax optimal conditional independence testing. *The Annals of Statistics* **49** 2151–2177. MR4319245

[36] Neykov, M., Wasserman, L., Kim, I. and Balakrishnan, S. (2023). Nearly minimax optimal Wasserstein conditional independence testing. *arXiv preprint arXiv:2308.08672*. MR4319245

[37] Pearl, J. (2014). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Elsevier. MR0965765

[38] Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **50** 157–175.

[39] Penrose, M. (2003). *Random Geometric Graphs* **5**. OUP Oxford. MR1986198

[40] Petersen, L. and Hansen, N. R. (2021). Testing conditional independence via quantile regression based partial copulas. *Journal of Machine Learning Research* **22** 1–47. MR4253763

[41] Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association* **100** 94–108. MR2156821

[42] Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics* **48** 1514–1538. MR4124333

[43] Song, L., Smola, A., Gretton, A., Bedo, J. and Borgwardt, K. (2012). Feature selection via dependence maximization. *Journal of Machine Learning Research* **13** 1393–1434. MR2930643

[44] Spohn, W. (1994). On the properties of conditional independence. In *Patrick Suppes: Scientific Philosopher* 173–196. Springer. MR1349350

[45] Tsamardinos, I. and Borboudakis, G. (2010). Permutation testing improves Bayesian network learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 322–337. Springer.

[46] Vaart, A. W. and Wellner, J. A. (1996). Weak convergence. In *Weak Convergence and Empirical Processes* 16–28. Springer. MR1385671

[47] Valiant, G. and Valiant, P. (2017). An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing* **46** 429–455. MR3614697

[48] Valiant, P. (2011). Testing symmetric properties of distributions. *SIAM Journal on Computing* **40** 1927–1968. MR2863200

[49] Yao, Q. and Tritchler, D. (1993). An exact analysis of conditional independence in several 2×2 contingency tables. *Biometrics* 233–236. MR1221407

[50] Zhang, J. and Zhang, Z. (2024). A normal test for independence via generalized mutual information. *Statistics & Probability Letters* **210** 110113. MR4723470