NON-PARAMETRIC ESTIMATION OF INFORMATION-THEORETIC QUANTITIES IN ENTROPIC PERSPECTIVE

by

Jialin Zhang

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics

Charlotte

2019

Approved by:

_____
Dr. Zhiyi Zhang

_____
Dr. Jiancheng Jiang

_____
Dr. Jun Song

_____
Dr. Arindam Mukherjee

ABSTRACT

JIALIN ZHANG. NON-PARAMETRIC ESTIMATION OF
INFORMATION-THEORETIC QUANTITIES IN ENTROPIC PERSPECTIVE.
(Under the direction of DR. ZHIYI ZHANG)

Introduced by Shannon [1], mutual information is a fundamental brick of information theory for its essential role in measuring association on non-ordinal alphabets. Mutual information being zero is a golden property as it indicates a probabilistic independent between the distributions. This article offers asymptotic chi-square distributions for the plug-in estimator and a non-parametric estimator of mutual information. The established distributions allow new tests of independence in entropic perspective.

ACKNOWLEDGEMENTS

The author would like to express the deepest appreciation to his committee chair Professor Zhiyi Zhang, who has continually and convincingly conveyed a spirit of bold adventure in regard to research. Without his uncountable, persistent, and invaluable guidance this dissertation would not have been possible. Special thanks also go to the members in the committee, Prof. Jiancheng Jiang, Prof. Jun Song, and Prof. Arindam Mukherjee, for their insightful and abundant assistance during this study.

In addition, the author would like to thank the University of North Carolina at Charlotte for financial assistance via GASP.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

1.1 Challenges in Big Data

Recent advancement in information technology increases the capability to obtain, exchange, and store data, hence the term big data. In analyzing these data, at least two fundamental issues immediately present themselves: 1) high dimensionality, and 2) discrete and non-ordinal nature.

Specifically, the vastly complex data space suggests that a data observation can only be appropriately registered in a very high-dimensional space, so much so that the dimensionality could essentially be infinite. On such spaces, the usual statistical methodologies quickly run into estimation and fundamental conceptual problems. Besides, the generality of the data space suggests that possible data values may not have an inherent order among themselves, for example, different gene types in the human genome, different words in a text, and various species in an ecological population.

The absence of inherent order immediately challenges the concept of a random variable. Namely, random variable, a function from the sample space to the real space, does not naturally exist on non-ordinal data spaces. Consequently, many familiar and fundamental concepts of Statistics and Probability no longer exist, for example,

moments, correlation, tail, characteristic function. As a result, to characterize the probability distributions on non-ordinal alphabets, inter-disciplinary research between Statistics, Probability, and Information Theory is needed.

1.2 Entropy, Mutual Information, and their Estimation

In 1948, Shannon introduced the concept of entropy and mutual information in his landmark paper [1], where he defined Shannon's entropy as:

$$H = \sum_k p_k \ln p_k.$$

Compared to classical concepts (*e.g.*, moments), entropy is calculated by probabilities (or ordered probabilities), and it does not rely on metric information, and thus it could exist in non-ordinal alphabets. Without using any metric information, entropy describes the level of dispersion in the probability distribution. In general, the larger the entropy, the heavier the dispersion. For example, a probability distribution with an effective cardinality $K = 4$ can produce a maximum possible entropy of $\ln 4$, and the maximum is achieved when the population distribution is uniform. And thus entropy can be considered as the moments on non-ordinal alphabets. Based on entropy, various information-theoretic quantities were proposed, for example, mutual information, Kullback-Leibler divergence, and entropic moments. These quantities characterize the information from a non-ordinal perspective, and they could be useful under much broader conditions. For example, mutual information could capture the

associations among both non-ordinal and ordinal random elements (or random variables) no matter if the relationship between the ordinal random variables is linear or nonlinear. For another example, entropic moments could serve as a characterization of distributions in place of characteristic functions in classic Statistics, and entropic moments exist in both non-ordinal and ordinal spaces where the characteristic functions only exist in ordinal spaces.

Most of the existing information-theoretic quantities are linear functions of entropy. As a result, the estimation of entropy plays a central role in practice in Information Theory. However, the estimation of entropy is technically difficult problems due to the curse of "High Dimensionality" and "Discrete and Non-ordinal Nature". For about 50 years since [1], advances in this area have been slow to come. Naively, people have been using the plug-in estimator (or the maximum likelihood estimator)

$$\hat{H} = \sum \hat{p}_k \ln \hat{p}_k.$$

When the K is finite, [2] showed that the bias of $\hat{H}$ is

$$\mathrm{E}(\hat{H}) - H = -\frac{K-1}{2n} + \frac{1}{12n^2}\left(1 - \sum_{k=1}^{K} \frac{1}{p_k}\right) + \mathcal{O}\left(n^{-3}\right).$$

Plentiful estimators tried to add bias corrections to the plug-in estimators when K is finite. Such attempts include Miller-Madow (1955) estimator ($\hat{H}_{MM}$) and the

Jackknife (1977) estimator ($\hat{H}_{JK}$). let $\hat{K}$ be the number of categories observed in the sample, and

$$\hat{H}_{MM} = \hat{H} + \frac{\hat{K} - 1}{2n}.$$

It can be shown that, for finite K, the bias of $\hat{H}_{MM}$ is

$$\mathrm{E}(\hat{H}_{MM}) - H = \frac{1}{12n^2}\left(1 - \sum_{k=1}^{K}\frac{1}{p_k}\right) + \mathcal{O}\left(n^{-3}\right).$$

$\hat{H}_{JK}$ is calculated in three steps:

1. for each $i \in \{1, 2, \ldots, n\}$ construct $\hat{H}^{(i)}$, which is a plug-in estimator based on a sub-sample of size $n - 1$ obtained by leaving the ith observation out;

2. obtain $\hat{H}_{(i)} = n\hat{H} - (n-1)\hat{H}^{(i)}$ for $i = 1, \cdots, n$; and then

3. compute the jackknife estimator

$$\hat{H}_{JK} = \frac{\sum_{i=1}^{n}\hat{H}_{(i)}}{n}. \tag{1.2.1}$$

Equivalently, (1.2.1) can be written as

$$\hat{H}_{JK} = n\hat{H} - (n-1)\frac{\sum_{i=1}^{n}\hat{H}^{(i)}}{n}.$$

The jackknife estimator of entropy was first described by [3]. When $K < \infty$, it can

be shown that the bias of $\hat{H}_{JK}$ is

$$\mathrm{E}\left(\hat{H}_{JK}\right) - H = \mathcal{O}\left(n^{-2}\right).$$

$\hat{H}_{MM}$ and $\hat{H}_{JK}$ reduce the rate of bias to a higher order power-decaying. While researchers were seeking for unbiased estimators, [4] proved that for finite K, an unbiased estimator for entropy does not exist. As a result, it is only possible to reduce the bias to a smaller extent. Inspired by Turing's formula, [5] showed that Shannon's entropy is algebraically equivalent to a function of entropic moments (ζ_v), which the uniformly minimum-variance unbiased estimators (UMVUEs) were established in [6] for the first $(n-1)$ moments, $\zeta_1, \cdots, \zeta_{n-1}$. Based on these UMVUEs, $\hat{H}_z$, the state-of-the-art entropy estimator, is developed as

$$\hat{H}_z(\cdot) = \sum_{v=1}^{n-1} \frac{1}{v} \frac{n^{1+v}[n-(1+v)]!}{n!} \sum_k \left[\hat{p}_k \prod_{j=0}^{v-1} \left(1 - \hat{p}_k - \frac{j}{n}\right) \right]. \tag{1.2.2}$$

The bias of $\hat{H}_z$ is

$$\mathrm{E}(\hat{H}_z) - H = \mathcal{O}\left(\frac{(1-p_\wedge)^n}{n}\right),$$

where $p_\wedge = \min\{p_k > 0\}$. When K is finite, $\hat{H}_z$ reduces the bias of entropy estimation from power decaying to exponentially decaying. Compared to $\hat{H}$ and $\hat{H}_{MM}$, of which the biases are infinity with a finite sample, $\hat{H}_z$ has an exponentially decaying bias under the same condition.

For finite K, the asymptotic normalities for $\hat{H}, \hat{H}_{MM}$, and $\hat{H}_{JK}$ are straight forward, and the asymptotic normality for $\hat{H}_z$ was provided in [7]. When K is countable infinite, asymptotic distributions with certain conditions for $\hat{H}, \hat{H}_{MM}$, $\hat{H}_{JK}$, and $\hat{H}_z$ are established recently in [8], [9], and [10]. Additional discussions on entropy estimation could also be found in [11] and [12].

The progress in entropy estimation allows further investigations on the estimation of mutual information. In Chapter 2, two mutual information estimators and their asymptotic properties are discussed.

CHAPTER 2: Asymptotic Distributions for Mutual Information Estimators

Let $\mathscr{X} = \{x_i; i = 1, \cdots, K_1\}$ and $\mathscr{Y} = \{y_j; j = 1, \cdots, K_2\}$ be two finite alphabets with cardinalities $K_1 < \infty$ and $K_2 < \infty$ respectively. Consider the Cartesian product $\mathscr{X} \times \mathscr{Y}$ with a joint probability distribution $\mathbf{p} = \{p_{i,j}\}$. Let the two marginal distributions be respectively denoted by $\mathbf{p}_x = \{p_{i,\cdot}\}$ and $\mathbf{p}_y = \{p_{\cdot,j}\}$ where $p_{i,\cdot} = \sum_j p_{i,j}$ and $p_{\cdot,j} = \sum_i p_{i,j}$. Assume that $p_{i,\cdot} > 0$ and $p_{\cdot,j} > 0$ for all $1 \le i \le K_1$ and $1 \le j \le K_2$, and that there are $K = \sum_{i,j} 1[p_{i,j} > 0]$ non-zero entries in $\{p_{i,j}\}$. We re-enumerate these K positive probabilities in one sequence and denote it as $\{p_k; k = 1, \cdots, K\}$.

Shannon's entropies for $\mathscr{X}$, $\mathscr{Y}$, and $\mathscr{X} \times \mathscr{Y}$, and mutual information between $\mathscr{X}$ and $\mathscr{Y}$ are defined as

$$\begin{aligned} H(X) &= -\sum_i p_{i,\cdot} \ln p_{i,\cdot}, \\ H(Y) &= -\sum_j p_{\cdot,j} \ln p_{\cdot,j}, \\ H(X,Y) &= -\sum_i \sum_j p_{i,j} \ln p_{i,j} = -\sum_{k=1}^{K} p_k \ln p_k, \\ MI(X,Y) &= H(X) + H(Y) - H(X,Y). \end{aligned} \tag{2.0.1}$$

For every pair of i and j, let $f_{i,j}$ be the observed frequency of the random pair (X, Y) taking value (x_i, y_j), where $i = 1, \cdots, K_1$ and $j = 1, \cdots, K_2$, in an *iid* sample of size n from $\mathscr{X} \times \mathscr{Y}$ under $\mathbf{p}$; and let $\hat{p}_{i,j} = f_{i,j}/n$ be the corresponding relative frequency. Consequently we write $\hat{\mathbf{p}} = \{\hat{p}_{i,j}\}$, $\hat{\mathbf{p}}_x = \{\hat{p}_{i,\cdot}\}$ and $\hat{\mathbf{p}}_y = \{\hat{p}_{\cdot,j}\}$ as the sets of observed joint and marginal relative frequencies. The objective of interest is to estimate the mutual information MI.

Let

$$\widehat{MI} = \widehat{MI}(X, Y) = \hat{H}(X) + \hat{H}(Y) - \hat{H}(X, Y) \tag{2.0.2}$$

where $\hat{H}(X) = -\sum_i \hat{p}_{i,\cdot} \ln \hat{p}_{i,\cdot}$, $\hat{H}(Y) = -\sum_j \hat{p}_{\cdot,j} \ln \hat{p}_{\cdot,j}$, and $\hat{H}(X, Y) = -\sum_{i,j} \hat{p}_{i,j} \ln \hat{p}_{i,j}$. $\widehat{MI}$ is the so-called plugin estimator of mutual information, or maximum likelihood estimator when K is finite.

Let

$$\begin{aligned}
\widehat{MI}_z = \widehat{MI}_z(X, Y) &= \hat{H}_z(X) + \hat{H}_z(Y) - \hat{H}_z(X, Y) \\
&= \textstyle\sum_{v=1}^{n-1} \frac{1}{v} \left\{ \frac{n^{v+1}[n-(v+1)]!}{n!} \sum_{i=1}^{K_1} \left[\hat{p}_{i,\cdot} \prod_{k=0}^{v-1} \left(1 - \hat{p}_{i,\cdot} - \frac{k}{n}\right) \right] \right\} \\
&\quad + \textstyle\sum_{v=1}^{n-1} \frac{1}{v} \left\{ \frac{n^{v+1}[n-(v+1)]!}{n!} \sum_{j=1}^{K_2} \left[\hat{p}_{\cdot,j} \prod_{k=0}^{v-1} \left(1 - \hat{p}_{\cdot,j} - \frac{k}{n}\right) \right] \right\} \\
&\quad - \textstyle\sum_{v=1}^{n-1} \frac{1}{v} \left\{ \frac{n^{v+1}[n-(v+1)]!}{n!} \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \left[\hat{p}_{i,j} \prod_{k=0}^{v-1} \left(1 - \hat{p}_{i,j} - \frac{k}{n}\right) \right] \right\}.
\end{aligned} \tag{2.0.3}$$

$\widehat{MI}_z$ is the mutual information estimator in Turing's perspective. It is showed in [13] that $\widehat{MI}$ has a power decaying bias and $\widehat{MI}_z$ has an exponentially decaying bias. We then introduce the asymptotic properties for the two mutual information estimators. The demonstration is given in two parts, depending on if the underlying mutual information is zero.

2.1 When Mutual Information is Not Zero

When $MI \neq 0$, [13] provided the asymptotic normality of $\widehat{MI}$ and $\widehat{MI}_z$. Let $g(v)$ and $\Sigma(v)$ be as defined in [13], their main theoretical results are summarized into the well-proven Proposition 1 and Theorem 1 below.

Proposition 1 *Provided that* $g^\tau(v)\Sigma(v)g(v) > 0$,

$$\sqrt{n}\left(\widehat{MI} - MI\right)\left[g^\tau(v)\Sigma(v)g(v)\right]^{-\frac{1}{2}} \xrightarrow{L} N(0,1). \tag{2.1.1}$$

Theorem 1 *Provided that* $g^\tau(v)\Sigma(v)g(v) > 0$,

$$\sqrt{n}\left(\widehat{MI}_z - MI\right)\left[g^\tau(v)\Sigma(v)g(v)\right]^{-\frac{1}{2}} \xrightarrow{L} N(0,1). \tag{2.1.2}$$

The normality of Proposition 1 and Theorem 1 are useful in providing confidence intervals for $MI > 0$ or in testing $H_0 : MI = c > 0$ for any $c > 0$. It is much more often of interest in the practice to test $H_0 : MI = 0$. However, Proposition 1 and Theorem 1 cannot be used to test the hypothesis $H_0 : MI = 0$. The normality of

Proposition 1 is based on a first-order delta method which requires $g^{\tau}(v)\Sigma(v)g(v) > 0$. It is demonstrated in Proposition 3 that this condition does not hold when $MI = 0$. In the following sub-section, we offer two chi-square tests for $H_0 : MI = 0$ based on $\widehat{MI}$ and $\widehat{MI}_z$ respectively, to complement what is not covered by the normality of Proposition 1 and Theorem 1.

2.2 When Mutual Information is Zero

The said tests are summarized in Proposition 2 and Theorem 2 below.

Proposition 2 *Provided that* $MI = 0$,

$$\chi_1^2 = 2n\widehat{MI} \xrightarrow{L} \chi^2\left((K_1 - 1)(K_2 - 1)\right) \tag{2.2.1}$$

Theorem 2 *Provided that* $MI = 0$,

$$\chi_2^2 = 2n\widehat{MI}_z + (K_1 - 1)(K_2 - 1) \xrightarrow{L} \chi^2\left((K_1 - 1)(K_2 - 1)\right) \tag{2.2.2}$$

The chi-square test in Proposition 2 is a well-known test, and its proof is a direct application of the log likelihood ratio test established by [14]. The proof is given in Chapter 3.

The chi-square test in Theorem 2 is the focal point of this article. The proof of Theorem 2 is non-trivial and is a part of the main results given in Chapter 3.

CHAPTER 3: Main Results

3.1 Proof of the Assumption in First Order Delta Method does not Hold when $MI = 0$

We first demonstrate that the assumption of the first order delta method is violated when $MI = 0$.

Proposition 3 *If $MI = 0$, then $g^\tau(v)\Sigma(v)g(v) = 0$.*

To prove Proposition 3, it is necessary to recall several notations in [13],

1. a re-enumeration of $\{p_{i,j}; i = 1, \cdots, K_1 \text{ and } j = 1, \cdots, K_2\}$ in the form of $\{p_k; k = 1, \cdots, K\}$, where $K = K_1 \times K_2$ (note that if $MI = 0$, $K = K_1 \times K_2$ is equivalent to $K = \sum_{i,j} 1[p_{i,j} > 0]$), and

2. a partition of the index set $\{(i, j); i = 1, \cdots, K_1 \text{ and } j = 1, \cdots, K_2\}$, denoted as

   $\{S_1, \cdots, S_{K_1}\}$ and $\{T_1, \cdots, T_{K_2}\}$, where

   (a) $S_s = \{k; p_k \in \{p_{s,j}; j = 1, \cdots, K_2\}\}$ for each s, $s = 1, \cdots, K_1$; and

   (b) $T_t = \{k; p_k \in \{p_{i,t}; i = 1, \cdots, K_1\}\}$ for each t, $t = 1, \cdots, K_2$.

Let $v = (p_1, \cdots, p_{K-1})^\tau$, $G(v) = MI = H(X) + H(Y) - H(X, Y)$ and $g(v) =$

$\nabla G(v) = (\partial G(v)/\partial p_1, \cdots, \partial G(v)/\partial p_{K-1})^\tau$, it was shown in [13] that

$$\frac{\partial}{\partial p_k} G(v) =$$
$$\begin{cases} \ln[(p_{K_1,\cdot})(p_{\cdot,K_2})(p_k)] - \ln[(p_{i,\cdot})(p_{\cdot,j})(p_K)], & \text{if } k \in S_i \neq S_{K_1} \text{ and } k \in T_j \neq T_{K_2} \\ \ln[(p_{\cdot,K_2})(p_k)] - \ln[(p_{\cdot,j})(p_K)], & \text{if } k \in S_{K_1} \text{ and } k \in T_j \neq T_{K_2} \\ \ln[(p_{K_1,\cdot})(p_k)] - \ln[(p_{i,\cdot})(p_K)], & \text{if } k \in S_i \neq S_{K_1} \text{ and } k \in T_{K_2} \end{cases} \tag{3.1.1}$$

where $p_K = 1 - \sum_{k \neq K} p_k$.

Proof of Proposition 3. If $MI = 0$ then X and Y are independent, *i.e.*, $p_{i,j} = p_{i,\cdot} p_{\cdot,j}$ for all (i, j). Consider the three cases of (3.1.1) separately. If $k \in S_i \neq S_{K_1}$ and $k \in T_j \neq T_{K_2}$, then $p_k = p_{i,\cdot} p_{\cdot,j}$ and $p_K = p_{K_1,\cdot} p_{\cdot,K_2}$, and therefore $\partial G(v)/\partial p_k = 0$. If $k \in S_{K_1}$ and $k \in T_j \neq T_{K_2}$, then $p_K = p_{K_1,\cdot} p_{\cdot,K_2}$ and $p_k = p_{\cdot,j} p_{K_1,j}$, and therefore $\partial G(v)/\partial p_k = 0$. If $k \in S_i \neq S_{K_1}$ and $k \in T_{K_2}$, then $p_K = p_{K_1,\cdot} p_{\cdot,K_2}$ and $p_k = p_{i,\cdot} p_{\cdot,K_2}$, and therefore $\partial G(v)/\partial p_k = 0$. It follows that $g(v) = \nabla G(v) = 0$ and hence $g^\tau(v) \Sigma(v) g(v) = 0$. □

3.2 Proof of the Asymptotic Distribution for $\widehat{MI}$ when $MI = 0$

Proof of Proposition 2

Consider the test $H_0 : p_{i,j} = p_{i,\cdot} p_{\cdot,j}$; $\sum p_{i,\cdot} = 1, \sum p_{\cdot,j} = 1$. For a random sample of size n, let $f_{i,\cdot}$, $f_{\cdot,j}$, and f_k be the observed frequency of the i-th marginal category of

X, the j-th marginal category of Y, and the k-th joint category, respectively. The generalized likelihood-ratio is

$$\begin{aligned} L &= \frac{\sup_{\theta\in\Theta_0} L(\theta|x)}{\sup_{\theta\in\Theta} L(\theta|x)} \\ &= \frac{\frac{n!}{f_1!\cdots f_K!}\left(\hat{p}_{1,\cdot}^{f_{1,\cdot}}\cdots\hat{p}_{K_1,\cdot}^{f_{K_1,\cdot}}\right)\left(\hat{p}_{\cdot,1}^{f_{\cdot,1}}\cdots\hat{p}_{\cdot,K_2}^{f_{\cdot,K_2}}\right)}{\frac{n!}{f_1!\cdots f_K!}\hat{p}_1^{f_1}\cdots\hat{p}_K^{f_K}} \\ &= \frac{\left(\hat{p}_{1,\cdot}^{f_{1,\cdot}}\cdots\hat{p}_{K_1,\cdot}^{f_{K_1,\cdot}}\right)\left(\hat{p}_{\cdot,1}^{f_{\cdot,1}}\cdots\hat{p}_{\cdot,K_2}^{f_{\cdot,K_2}}\right)}{\hat{p}_1^{f_1}\cdots\hat{p}_K^{f_K}} \end{aligned}$$

And

$$\begin{aligned} -2\ln L &= -2\ln\frac{\left(\hat{p}_{1,\cdot}^{f_{1,\cdot}}\cdots\hat{p}_{K_1,\cdot}^{f_{K_1,\cdot}}\right)\left(\hat{p}_{\cdot,1}^{f_{\cdot,1}}\cdots\hat{p}_{\cdot,K_2}^{f_{\cdot,K_2}}\right)}{\hat{p}_1^{f_1}\cdots\hat{p}_K^{f_K}} \\ &= -2\left(\sum_i f_{i,\cdot}\ln\hat{p}_{i,\cdot} + \sum_j f_{\cdot,j}\ln\hat{p}_{\cdot,j} - \sum_{i,j} f_{i,j}\ln\hat{p}_{i,j}\right) \\ &= 2n\left(-\sum_i \hat{p}_{i,\cdot}\ln\hat{p}_{i,\cdot} - \sum_j \hat{p}_{\cdot,j}\ln\hat{p}_{\cdot,j} + \sum_{i,j}\hat{p}_{i,j}\ln\hat{p}_{i,j}\right) \\ &= 2n\widehat{MI} \end{aligned}$$

By [14], $-2\ln L \sim \chi^2$ with degrees of freedom $(K_1-1)(K_2-1)$. □

3.3 Proof of the Asymptotic Distribution for $\widehat{MI}_z$ when $MI = 0$

The proof of Theorem 2 needs several additional notations and lemmas. Consider a single alphabet $\mathscr{X}$ and the associated probability distribution $\mathbf{p} = \{p_k; k = 1, \cdots, K\}$. Suppose an *iid* sample of size n results in letter frequencies $\{Y_k; k \geq 1\}$

and relative frequencies $\hat{\mathbf{p}} = \{\hat{p}_k; k \geq 1\}$. Let $\hat{H} = H(\hat{\mathbf{p}})$ and

$$\hat{H}_z = \hat{H}_z(\hat{\mathbf{p}}) = \sum_{v=1}^{n-1} \frac{1}{v} Z_v \tag{3.3.1}$$

where

$$Z_v = \sum_{k=1}^{\infty} \left[\hat{p}_k \prod_{j=0}^{v-1} \left(1 - \hat{p}_k - \frac{j}{n} \right) \right] = \sum_{k=1}^{\infty} \left\{ \hat{p}_k \prod_{j=0}^{v-1} \left[\left(1 - \hat{p}_k - \frac{j}{n} \right) \frac{1}{1 - \frac{j+1}{n}} \right] \right\}. \tag{3.3.2}$$

Lemma 1 *For any $p \in (0, 1)$,*

1. $\sum_{v=1}^{\infty} (1-p)^{v-1} = 1/p$,
2. $\sum_{v=1}^{\infty} v(1-p)^{v-1} = 1/p^2$,
3. $\sum_{v=1}^{\infty} v^2(1-p)^{v-1} = 2/p^3 - 1/p^2$, *and*
4. $\sum_{v=1}^{\infty} v^3(1-p)^{v-1} = 6/p^4 - 6/p^3 + 1/p^2$.

Proof of Lemma 1.

1.

$$\sum_{v=1}^{\infty} (1-p)^{v-1} = \frac{1}{1-(1-p)} = \frac{1}{p}$$

2.

$$\begin{aligned}
\sum_{v=1}^{\infty} v(1-p)^{v-1} =& \frac{1}{p}\left(\sum_{v=1}^{\infty} v(1-p)^{v-1} - (1-p)\sum_{v=1}^{\infty} v(1-p)^{v-1}\right) \\
=& \frac{1}{p}\left(\sum_{v=1}^{\infty} v(1-p)^{v-1} - \sum_{v=1}^{\infty} v(1-p)^{v}\right) \\
=& \frac{1}{p}\left(\sum_{v=0}^{\infty} (v+1)(1-p)^{v} - \sum_{v=1}^{\infty} v(1-p)^{v}\right) \\
=& \frac{1}{p}\left((1-p)^0 + \sum_{v=1}^{\infty} (v+1)(1-p)^{v} - \sum_{v=1}^{\infty} v(1-p)^{v}\right) \\
=& \frac{1}{p}\left((1-p)^0 + \sum_{v=1}^{\infty} (1-p)^{v}\right) \\
=& \frac{1}{p}\sum_{v=0}^{\infty} (1-p)^{v} \\
=& \frac{1}{p}\sum_{v=1}^{\infty} (1-p)^{v-1} \\
=& \frac{1}{p^2}
\end{aligned}$$

3.

$$\begin{aligned}
\sum_{v=1}^{\infty} v^2(1-p)^{v-1} =& \frac{1}{p}\left(\sum_{v=1}^{\infty} v^2(1-p)^{v-1} - (1-p)\sum_{v=1}^{\infty} v^2(1-p)^{v-1}\right) \\
=& \frac{1}{p}\left(\sum_{v=0}^{\infty} (v+1)^2(1-p)^{v} - \sum_{v=1}^{\infty} v^2(1-p)^{v}\right) \\
=& \frac{1}{p}\left((1-p)^0 + \sum_{v=1}^{\infty} (v+1)^2(1-p)^{v} - \sum_{v=1}^{\infty} v^2(1-p)^{v}\right) \\
=& \frac{1}{p}\left((1-p)^0 + \sum_{v=1}^{\infty} (2v+1)(1-p)^{v}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{p}\sum_{v=0}^{\infty}(2v+1)(1-p)^{v} \\
&= \frac{1}{p}\sum_{v=1}^{\infty}(2v-1)(1-p)^{v-1} \\
&= \frac{1}{p}\left(2\sum_{v=1}^{\infty}v(1-p)^{v-1} - \sum_{v=1}^{\infty}(1-p)^{v-1}\right) \\
&= \frac{1}{p}\left(\frac{2}{p^2} - \frac{1}{p}\right) \\
&= \frac{2}{p^3} - \frac{1}{p^2}
\end{aligned}
$$

4.

$$
\begin{aligned}
\sum_{v=1}^{\infty}v^3(1-p)^{v-1} &= \frac{1}{p}\left(\sum_{v=1}^{\infty}v^3(1-p)^{v-1} - (1-p)\sum_{v=1}^{\infty}v^3(1-p)^{v-1}\right) \\
&= \frac{1}{p}\left(\sum_{v=1}^{\infty}v^3(1-p)^{v-1} - \sum_{v=1}^{\infty}v^3(1-p)^{v}\right) \\
&= \frac{1}{p}\left(\sum_{v=0}^{\infty}(v+1)^3(1-p)^{v} - \sum_{v=1}^{\infty}v^3(1-p)^{v}\right) \\
&= \frac{1}{p}\left((1-p)^0 + \sum_{v=1}^{\infty}(v+1)^3(1-p)^{v} - \sum_{v=1}^{\infty}v^3(1-p)^{v}\right) \\
&= \frac{1}{p}\left((1-p)^0 + \sum_{v=1}^{\infty}\left((v+1)^3 - v^3\right)(1-p)^{v}\right) \\
&= \frac{1}{p}\left((1-p)^0 + \sum_{v=1}^{\infty}\left(3v^2+3v+1\right)(1-p)^{v}\right) \\
&= \frac{1}{p}\sum_{v=0}^{\infty}\left(3v^2+3v+1\right)(1-p)^{v} \\
&= \frac{1}{p}\sum_{v=1}^{\infty}\left(3(v-1)^2+3(v-1)+1\right)(1-p)^{v-1} \\
&= \frac{1}{p}\sum_{v=1}^{\infty}\left(3v^2-3v+1\right)(1-p)^{v-1}
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{1}{p}\left(3\sum_{v=1}^{\infty}v^2(1-p)^{v-1}-3\sum_{v=1}^{\infty}v(1-p)^{v-1}+\sum_{v=1}^{\infty}(1-p)^{v-1}\right)\\
=&\frac{1}{p}\left(\frac{6}{p^3}-\frac{3}{p^2}-\frac{3}{p^2}+\frac{1}{p}\right)\\
=&\frac{6}{p^4}-\frac{6}{p^3}+\frac{1}{p^2}
\end{aligned}
$$

□

Lemma 2 *(Weierstrass Product Inequality) For any set of real numbers,* $\{a_i; i = 1, \cdots, n\}$*, such that* $a_i \in [0,1]$ *for each* i*,* $1-\sum_{i=1}^{n} a_i \leq \prod_{i=1}^{n}(1-a_i) \leq 1/(1+\sum_{i=1}^{n} a_i)$.

Proof of Lemma 2.

Part 1: To show

$$1-\sum_{i=1}^{n} a_i \leq \prod_{i=1}^{n}(1-a_i).$$

When $n = 1$,

$$1 - a_1 \leq 1 - a_1$$

is true. For any positive integer k, when $n = k$, suppose

$$1-\sum_{i=1}^{k} a_i \leq \prod_{i=1}^{k}(1-a_i)$$

is true, then for $n = k + 1$

$$\begin{aligned} 1 - \sum_{i=1}^{k+1} a_i =& 1 - \sum_{i=1}^{k} a_i - a_{k+1} \\ &\leq \prod_{i=1}^{k}(1 - a_i) - a_{k+1} \\ &\leq \prod_{i=1}^{k}(1 - a_i) - a_{k+1}\prod_{i=1}^{k}(1 - a_i) \\ &\leq \prod_{i=1}^{k+1}(1 - a_i). \end{aligned}$$

Therefore Part 1 is true by induction.

Part 2: To show

$$\prod_{i=1}^{n}(1 - a_i) \leq \frac{1}{1 + \sum_{i=1}^{n} a_i}$$

When $n = 1$, since $(1 - a_1)(1 + a_1) = 1 - a_1^2 \leq 1$,

$$1 - a_1 \leq \frac{1}{1 + a_1}$$

is true. For any positive integer k, when $n = k$, suppose

$$\prod_{i=1}^{k}(1 - a_i) \leq \frac{1}{1 + \sum_{i=1}^{k} a_i}$$

is true, then for $n = k + 1$

$$\begin{aligned}\prod_{i=1}^{k+1}(1-a_i) =& \prod_{i=1}^{k}(1-a_i)\cdot(1-a_{k+1}) \\ \leq& \frac{1-a_{k+1}}{1+\sum_{i=1}^{k}a_i} \\ \leq& \frac{1-a_{k+1}+a_{k+1}}{1+\sum_{i=1}^{k}a_i+a_{k+1}} \\ =& \frac{1}{1+\sum_{i=1}^{k+1}a_i}.\end{aligned}$$

Therefore Part 2 is true by induction. □

Lemma 3 $2n(\hat{H}_z - \hat{H}) \xrightarrow{p} K - 1$.

Proof of Lemma 3. Consider

$$\begin{aligned}& 2n(\hat{H}_z - \hat{H}) \\ =& 2n\left(\sum_{k=1}^{K}\hat{p}_k \ln \hat{p}_k + \sum_{v=1}^{n-1}\frac{1}{v}Z_v\right) \\ =& 2n\sum_{k=1}^{K}\left\{\hat{p_k}\left\{-\sum_{v=1}^{\infty}\frac{1}{v}(1-\hat{p_k})^v + \sum_{v=1}^{n-1}\left\{\frac{1}{v}\prod_{j=0}^{v-1}\left[\left(1-\hat{p}_k-\frac{j}{n}\right)\frac{1}{1-\frac{j+1}{n}}\right]\right\}\right\}\right\} \\ =& 2n\sum_{k=1}^{K}\left\{\hat{p_k}\left\{-\sum_{v=1}^{\infty}\frac{1}{v}(1-\hat{p_k})^v + \sum_{v=1}^{\infty}\left\{\frac{1}{v}\prod_{j=0}^{v-1}\left[\left(1-\hat{p}_k-\frac{j}{n}\right)\frac{1}{1-\frac{j+1}{n}}\right]\right\}\right\}\right\} \\ =& : 2n\sum_{k=1}^{K}\{d_k\}.\end{aligned}$$

Note that

$$\sum_{v=1}^{\infty}\left\{\frac{1}{v}\prod_{j=0}^{v-1}\left[\left(1-\hat{p}_k-\frac{j}{n}\right)\frac{1}{1-\frac{j+1}{n}}\right]\right\}=\sum_{v=1}^{n-1}\left\{\frac{1}{v}\prod_{j=0}^{v-1}\left[\left(1-\hat{p}_k-\frac{j}{n}\right)\frac{1}{1-\frac{j+1}{n}}\right]\right\}$$

is because $\prod_{j=0}^{v-1}\left[\left(1-\hat{p}_k-\frac{j}{n}\right)\frac{1}{1-\frac{j+1}{n}}\right]=0$ when $v\geq n$.

For any pair of (k,v), let

$$f_1(k,v)=\frac{1}{v}(1-\hat{p}_k)^v=\frac{1}{v}\prod_{j=0}^{v-1}(1-\hat{p}_k)$$

$$f_2(k,v)=\begin{cases}\frac{1}{v}\prod_{j=0}^{v-1}\left[\left(1-\hat{p}_k-\frac{j}{n}\right)\frac{1}{1-\frac{j+1}{n}}\right], & \text{if } v\leq n(1-\hat{p})+1\\ 0, & \text{if } v\geq n(1-\hat{p}).\end{cases}$$

For any k, nd_k can be re-expressed as

$$\begin{aligned}nd_k&=n\hat{p}_k\left(-\sum_{v=1}^{\infty}f_1(k,v)+\sum_{v=1}^{\infty}f_2(k,v)\right)\\&=n\hat{p}_k\sum_{v=1}^{\infty}\left(-f_1(k,v)+f_2(k,v)\right)\\&=n\hat{p}_k\sum_{v=1}^{\infty}f_1(k,v)\left(\frac{f_2(k,v)}{f_1(k,v)}-1\right).\end{aligned}$$

The ratio of $f_2(k,v)$ to $f_1(k,v)$ can be written as, $f_2(k,v)/f_1(k,v)=0$ if $v\geq n(1-\hat{p})+2$, otherwise

$$\frac{f_2(k,v)}{f_1(k,v)}=\frac{\frac{1}{v}\prod_{j=0}^{v-1}\left[\left(1-\hat{p}_k-\frac{j}{n}\right)\frac{1}{1-\frac{j+1}{n}}\right]}{\frac{1}{v}\prod_{j=0}^{v-1}(1-\hat{p}_k)}$$

$$= \prod_{j=0}^{v-1} \left[\left(1 - \frac{j}{n(1-\hat{p}_k)} \right) \frac{1}{1 - \frac{j+1}{n}} \right]$$
$$= \frac{\prod_{j=0}^{v-1} \left[1 - \frac{j}{n(1-\hat{p}_k)} \right]}{\prod_{j=0}^{v-1} \left(1 - \frac{j+1}{n} \right)}.$$

Noting that

$$\sum_{j=0}^{v-1} \frac{j}{n(1-\hat{p_k})} = \frac{v(v-1)}{2n(1-\hat{p_k})} \qquad \text{and} \qquad \sum_{j=0}^{v-1} \frac{j+1}{n} = \frac{v(v+1)}{2n},$$

by the Weierstrass Product Inequality, it follows that, for each $v \leq n(1-\hat{p}_k)+1$,

$$1 - \frac{v(v-1)}{2n(1-\hat{p}_k)} \leq \prod_{j=0}^{v-1} \left[1 - \frac{j}{n(1-\hat{p}_k)} \right] \leq \frac{1}{1 + \frac{v(v-1)}{2n(1-\hat{p}_k)}} \tag{3.3.3}$$

and

$$1 - \frac{v(v+1)}{2n} \leq \prod_{j=0}^{v-1} \left(1 - \frac{j+1}{n} \right) \leq \frac{1}{1 + \frac{v(v+1)}{2n}}. \tag{3.3.4}$$

By the first inequality of (3.3.3) and the second of (3.3.4), a lower bound for $f_2(k,v)/f_1(k,v)$, for each $v \leq n(1-\hat{p}_k)+1$, is established as

$$\frac{f_2(k,v)}{f_1(k,v)} \geq \frac{1 - \frac{v(v-1)}{2n(1-\hat{p}_k)}}{\frac{1}{1+\frac{v(v+1)}{2n}}}$$
$$= \left[1 - \frac{v(v-1)}{2n(1-\hat{p}_k)} \right] \left[1 + \frac{v(v+1)}{2n} \right]$$
$$= 1 - \frac{v(v-1)}{2n(1-\hat{p}_k)} + \frac{v(v+1)}{2n} - \frac{v^2(v^2-1)}{4n^2(1-\hat{p}_k)}. \tag{3.3.5}$$

In fact, noting that the expression on the far left of (3.3.3) takes on negative values for $v \geq n(1-\hat{p}_k)+2$, that the expression on the far right of (3.3.4) remains positive for all $v \geq 1$, and that $f_2(k,v)/f_1(k,v) \geq 0$, the inequality in (3.3.5) holds for all $v \geq 1$.

It therefore follows that, applying Lemma 1 whenever necessary,

$$
\begin{aligned}
nd_k &= n\hat{p}_k \sum_{v=1}^{\infty} f_1(k,v) \left(\frac{f_2(k,v)}{f_1(k,v)} - 1 \right) \\
&\geq n\hat{p}_k \sum_{v=1}^{\infty} f_1(k,v) \left[-\frac{v(v-1)}{2n(1-\hat{p}_k)} + \frac{v(v+1)}{2n} - \frac{v^2(v^2-1)}{4n^2(1-\hat{p}_k)} \right] \\
&= n\hat{p}_k \sum_{v=1}^{\infty} \left\{ \left[\frac{1}{v}(1-\hat{p}_k)^v \right] \left[-\frac{v(v-1)}{2n(1-\hat{p}_k)} + \frac{v(v+1)}{2n} - \frac{v^2(v^2-1)}{4n^2(1-\hat{p}_k)} \right] \right\} \\
&= n\hat{p}_k \sum_{v=1}^{\infty} \left\{ (1-\hat{p}_k)^v \left[-\frac{v-1}{2n(1-\hat{p}_k)} + \frac{v+1}{2n} - \frac{v(v^2-1)}{4n^2(1-\hat{p}_k)} \right] \right\} \\
&= \hat{p}_k \sum_{v=1}^{\infty} \left\{ (1-\hat{p}_k)^v \left[-\frac{v-1}{2(1-\hat{p}_k)} + \frac{v+1}{2} - \frac{v(v^2-1)}{4n(1-\hat{p}_k)} \right] \right\} \\
&= \hat{p}_k \sum_{v=1}^{\infty} \left\{ -\frac{1}{2} \left[v(1-\hat{p}_k)^{v-1} - (1-\hat{p}_k)^{v-1} \right] + \frac{1-\hat{p}_k}{2} \left[v(1-\hat{p}_k)^{v-1} + (1-\hat{p}_k)^{v-1} \right] \right. \\
&\qquad \left. - \frac{1}{4n} \left[v^3(1-\hat{p}_k)^{v-1} - v(1-\hat{p}_k)^{v-1} \right] \right\} \\
&= \hat{p}_k \left\{ -\frac{1}{2} \left[\sum_{v=1}^{\infty} v(1-\hat{p}_k)^{v-1} - \sum_{v=1}^{\infty} (1-\hat{p}_k)^{v-1} \right] \right. \\
&\qquad + \frac{1-\hat{p}_k}{2} \left[\sum_{v=1}^{\infty} v(1-\hat{p}_k)^{v-1} + \sum_{v=1}^{\infty} (1-\hat{p}_k)^{v-1} \right] \\
&\qquad \left. - \frac{1}{4n} \left[\sum_{v=1}^{\infty} v^3(1-\hat{p}_k)^{v-1} - \sum_{v=1}^{\infty} v(1-\hat{p}_k)^{v-1} \right] \right\} \\
&= \hat{p}_k \left\{ -\frac{1}{2} \left(\frac{1}{\hat{p}_k^2} - \frac{1}{\hat{p}_k} \right) + \frac{1-\hat{p}_k}{2} \left(\frac{1}{\hat{p}_k^2} + \frac{1}{\hat{p}_k} \right) - \frac{1}{4n} \left(\frac{6}{\hat{p}_k^4} - \frac{6}{\hat{p}_k^3} + \frac{1}{\hat{p}_k^2} - \frac{1}{\hat{p}_k^2} \right) \right\}
\end{aligned}
$$

$$= -\frac{1}{2}\left(\frac{1}{\hat{p}_k} - 1\right) + \frac{1-\hat{p}_k}{2}\left(\frac{1}{\hat{p}_k} + 1\right) - \frac{1}{4n}\left(\frac{6}{\hat{p}_k^3} - \frac{6}{\hat{p}_k^2}\right)$$
$$= -\frac{1}{2\hat{p}_k} + \frac{1}{2} + \frac{1-\hat{p}_k}{2\hat{p}_k} + \frac{1-\hat{p}_k}{2} - \frac{1}{4n}\left(\frac{6}{\hat{p}_k^3} - \frac{6}{\hat{p}_k^2}\right)$$
$$= \frac{1-\hat{p}_k}{2} - \frac{3}{2n}\left(\frac{1}{\hat{p}_k^3} - \frac{1}{\hat{p}_k^2}\right) =: A_{k,n}.$$

Therefore,

$$\begin{aligned} 2n(\hat{H}_z - \hat{H}) &= 2n\sum_k d_k \\ &\geq 2\sum_k A_{k,n} \\ &= K - 1 - \frac{3}{2n}\sum_k \left(\frac{1}{\hat{p}_k^3} - \frac{1}{\hat{p}_k^2}\right) \\ &\geq K - 1 \end{aligned} \tag{3.3.6}$$

On the other hand, it is proved in [5] and [2] that (recall that $p_\wedge = \min p_k$)

$$\mathrm{E}[\hat{H}_z - H] = \mathcal{O}\left(\frac{(1-p_\wedge)^n}{n}\right),$$

and

$$\mathrm{E}[H - \hat{H}] = \frac{K-1}{2n} - \frac{1}{12n^2}\left(1 - \sum_{k=1}^{K}\frac{1}{p_k}\right) + \mathcal{O}\left(n^{-3}\right),$$

hence

$$2n\,\mathrm{E}\left[\hat{H}_z - \hat{H}\right] \xrightarrow{p} K - 1. \tag{3.3.7}$$

Since (3.3.6) and (3.3.7), by Markov's inequality,

$$2n(\hat{H}_z - \hat{H}) \xrightarrow{p} K - 1.$$

□

Proof of Theorem 2. By Proposition 2 and noting that

$$\begin{aligned} 2n\widehat{MI} &= 2n\left(\hat{H}(X) + \hat{H}(Y) - \hat{H}(X,Y)\right) \\ &= 2n\left[-(\hat{H}_z(X) - \hat{H}_X) - (\hat{H}_z(Y) - \hat{H}(Y)) + (\hat{H}_z(X,Y) - \hat{H}(X,Y))\right] + 2n\widehat{MI}_z, \end{aligned}$$

and applying Lemma 3, it follows that

$$2n\widehat{MI}_z + (K_1 - 1)(K_2 - 1) \sim 2n\widehat{MI} \xrightarrow{L} \chi^2((K_1 - 1)(K_2 - 1)).$$

□

3.4 Examples

[13] gave three examples involving evaluation of gene-to-gene association. We re-work these examples to demonstrate the usage of the proposed new test. Three genes were under consideration, and they were coded as TMEM30A, MTCH2, and ENAH. In Example 1, readings of TMEM30A and MTCH2 were analyzed. In Examples 2 and 3, readings on two different probes, designed for the same gene ENAH, on the same microchip were analyzed.

Example 1 TMEM30A and MTCH2. Using the results of [13], $\widehat{MI} = 0.1459$ and $\widehat{MI}_z = 0.0552$, $\chi_1^2 = 2n\widehat{MI} = 55.7338$ and $\chi_2^2 = 2n\widehat{MI}_z + (K_1 - 1)(K_2 - 1) = 102.0864$. With degrees of freedom 81, the respective p-values are 0.9856 and 0.0567. At $\alpha = 0.05$, neither test rejects $H_0 : MI = 0$.

Example 2 ENAH and ENAH with $K = K_1 \times K_2$. Using the results of [13], $\widehat{MI} = 0.2060$ and $\widehat{MI}_z = 0.1157$, $\chi_1^2 = 2n\widehat{MI} = 78.692$ and $\chi_2^2 = 2n\widehat{MI}_z + (K_1 - 1)(K_2 - 1) = 125.1974$. With degrees of freedom 81, the respective p-values are 0.5519 and 0.0012. χ_2^2 detects an association with strong evidence and χ_1^2 fails to do so by far. Since in this case it is known a priori that an association exists, this example illustrates the added utility of χ_2^2.

Example 3 ENAH and ENAH with $K \leq K_1 \times K_2$. In this example, [13] assumes that several cells in the join alphabet are associated with zero probabilities. This

assumption is invalid since, if it were the case, then under the null hypothesis of $H_0 : MI = 0$ (X and Y are independent) either some of the marginal probabilities of X or Y would have to be zeros. However every (of the ten) marginal categories is covered by observations, that is to say, none of the marginal probabilities can be zero. Without the assumption of zero probabilities, the chi-square tests of Proposition 2 and Theorem 2 give identical results as in Example 2.

CHAPTER 4: Simulation Study

To further explore and study the property of $\widehat{MI}_z$, various tests of independence were compared in the simulation study to evaluate the size and power of the tests under different sample sizes.

Particularly, five tests of independence were compared:

1. Pearson Chi-square Test:

   Test Statistic:

$$\chi^2_{Pearson} = \sum_{k=1}^{K^*} \frac{(O_k - E_k)^2}{E_k} \sim \chi^2\left((K_1^* - 1)(K_2^* - 1)\right). \tag{4.0.1}$$

   Calculating $\chi^2_{Pearson}$ requires all the denominators E_k's to be positive, where each E_k is calculated by n multiplies the corresponding $\hat{p}_{i,.}\hat{p}_{.,j}$. When the sample size is not sufficiently large, many $\hat{p}_{i,.}$'s and $\hat{p}_{.,j}$'s could be zero, and it makes Pearson's test invalid. Moreover, it is commonly suggested that each cell should have at least five observations to use Pearson's test. In order to fairly compare other tests with Pearson's test under each situation, all the original samples in the simulation were adjusted. Particularly, in each bivariate sample data frequency table, all rows (and columns) with total frequencies less than 5 were

combined to the row (and column) with the least total frequency among the rows (and columns) with at least five total frequency. If there are two or more such rows (or columns), one is selected randomly. For example, suppose the sample data frequency table is Table 4.1. The frequencies of $X = 3$ and 4 are less than 5, and the frequency of $X = 5$ is the category with the least frequency that is at least 5. As a result, the adjusted sample regarding the frequencies of X is described in Table 4.2. For the frequency of Y, the frequencies of $Y = 2$ and 4 are less than 5, and the frequencies of $Y = 1$ and 3 are the same. In the example, $Y = 1$ is randomly selected to be combined with low-frequency categories. And the adjusted sample is as described in Table 4.3. Note that after the adjustment, the cardinality of X, Y, and $X \times Y$ are reduced from 5, 4, and 20 to 3, 2, and 6. Because of the possible adjustment, K^*, K_1^*, and K_2^* are used instead of K, K_1, and K_2 in (4.0.1). In the following simulation settings, the cardinality of X and Y are 10 and 15, and sample sizes are started from 100. Therefore it is guaranteed that there is always at least one category in X and Y with a frequency of 5 or more.

Table 4.1: Original Sample Frequency Table

		Y			
		1	2	3	4
X	1	4	0	6	0
	2	1	1	3	1
	3	2	2	0	0
	4	0	0	0	0
	5	3	0	1	1

Table 4.2: Partially Adjusted Sample (X categories with low frequencies combined)

		Y			
		1	2	3	4
X	1	4	0	6	0
	2	1	1	3	1
	5	5	2	1	1

Table 4.3: Adjusted Sample (X and Y categories with low frequencies combined)

		Y	
		1	3
X	1	4	6
	2	3	3
	5	8	1

2. Test of independence using $\widehat{MI}$ and Proposition 2:

   Test Statistic:

$$\chi^2_{MIhat} = 2n\widehat{MI}(\text{original sample}) \sim \chi^2\left((K_1-1)(K_2-1)\right),$$

   and

$$\chi^{2*}_{MIhat} = 2n\widehat{MI}(\text{adjusted sample}) \sim \chi^2\left((K_1^*-1)(K_2^*-1)\right).$$

   Two tests of independence using $\widehat{MI}$ are examined. $\widehat{MI}$ does not require non-zero marginal sample probabilities as that of Pearson's test. To make a complete comparison, tests using $\widehat{MI}$ on both original sample and adjusted sample are included.

3. Test of independence using $\widehat{MI}_z$ [1] and Theorem 2:

[1] The computation of $\widehat{MI}_z$ has been implemented in R and can be found in R package "Entropy-

Test Statistic:

$$\chi^2_{\widehat{MI}z} = 2n\widehat{MI}_z(\text{original sample}) + (K_1-1)(K_2-1) \sim \chi^2\left((K_1-1)(K_2-1)\right),$$

and

$$\chi^{2*}_{\widehat{MI}z} = 2n\widehat{MI}_z(\text{adjusted sample}) + (K_1^*-1)(K_2^*-1) \sim \chi^2\left((K_1^*-1)(K_2^*-1)\right).$$

For the same reason (to make a complete comparison), tests using $\widehat{MI}_z$ on both original sample and adjusted sample are included.

4.1 Simulation Settings

The two marginal distributions are

$$p_{i,\cdot} = \frac{1}{10}, \qquad i = 1, 2, \cdots, 10;$$

and

$$p_{\cdot,j} = \frac{16-j}{120}, \qquad j = 1, 2, \cdots, 15.$$

When evaluating the size of tests,

$$p_{i,j} = p_{i,\cdot}p_{\cdot,j}.$$

Estimation".

When evaluating the power of tests,

$$p_{i,j} = \frac{150 - 15(i-1) - j + 1}{11325} \qquad \text{(triangle distribution)}.$$

For both evaluation, the simulations were conducted with sample size $n = 100$, 200, 300, ..., 14900, 15000. And for each sample size, the simulation was iterated for 50000 times.

4.2 Simulation Results

The simulation results are presented in Figure 4.1 and 4.2. To help understand the legend, pearson.reject_vec, miz.reject_vec, miz.reject_adj_vec, mihat.reject_vec, and mihat.reject_adj_vec stand for testing using $\chi^2_{Pearson}$, χ^2_{MIz}, χ^{2*}_{MIz}, χ^2_{MIhat}, and χ^{2*}_{MIhat}, respectively.

The sizes of all tests reached the neighborhood of α when n is more than 4500; therefore Figure 4.1 did not include the simulation results when n is more than 6500. Based on Figure 4.1, Pearson's test on adjusted (combined) samples converged to α faster than other tests. The two sets of tests using estimators of mutual information converged to α at a similar rate. Although $\widehat{MI}_z$ has a smaller bias over $\widehat{MI}$, surprisingly that the simulation showed that testing independence using $\widehat{MI}$ has better size than that of using $\widehat{MI}_z$ when the sample size is relatively small.

The powers of all tests using mutual information estimators are consistently higher than the power of Pearson's test. It suggests that when the sample size is sufficiently

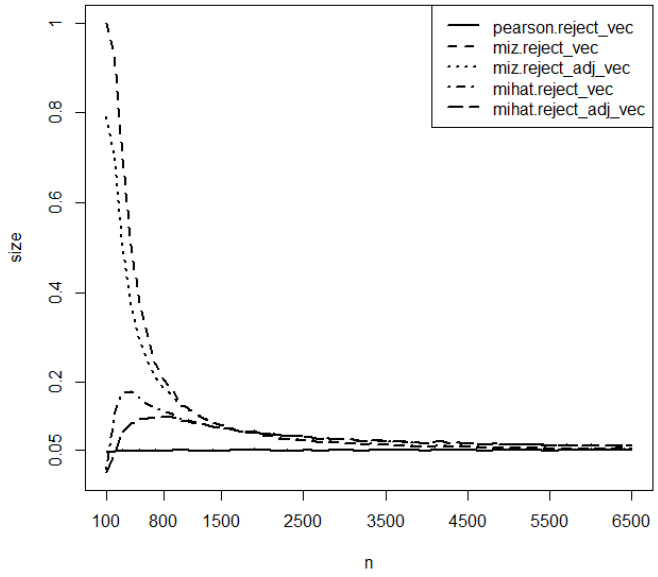

Figure 4.1: Size of different tests under different sample size when $\alpha = 0.05$.

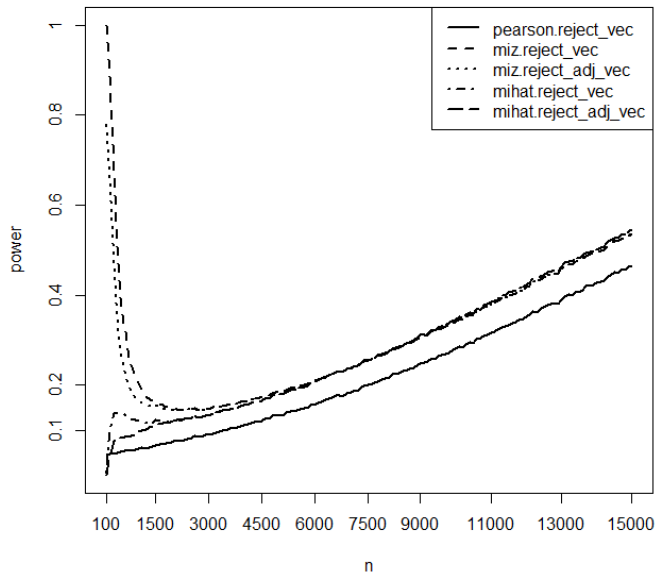

Figure 4.2: Power of different tests under different sample size when $\alpha = 0.05$ under a joint triangle distribution.

large, a test using either $\widehat{MI}$ or $\widehat{MI}_z$ should be adopted instead of Pearson's test.

Moreover, when the sample size is moderate (n from 1500 to 4500 in the designed

simulation), testing using $\widehat{MI}_z$ and $\widehat{MI}$ have similar size, but the power of using $\widehat{MI}_z$ is higher.

CHAPTER 5: Conclusion and Future Work

In conclusion, the asymptotic distributions for $\widehat{MI}$ and $\widehat{MI}_z$ are offered under the situation that $MI = 0$. Based on the simulation study, when the sample size is sufficiently large, testing independence using mutual information estimators are preferred because they have higher powers than Pearson's test. When the sample size is moderate, Pearson test's size has a faster convergence rate to α and is preferred. When the sample size is small, although Pearson's test also has a faster convergence rate of size, it is frequently incalculable without a merge of empty cells.

$\widehat{MI}_z$ is known to have a faster-decaying bias compared to $\widehat{MI}$, whereas the size of test of independence with $\widehat{MI}$ has a faster converging rate when the sample size is relatively small. It leads to my conjecture that additional bias correction terms are needed in Theorem 2 to improve its performance under small samples, which will be future work. Furthermore, it is worthy of investigating why Pearson's test has such a fast converging rate in the size of test. Finally, Theorem 2 could be generated as a new test of goodness-of-fit using Kullback-Leibler divergence $\widehat{KL}_z$ [15], one could study its property and compare it with other goodness-of-fit tests.

REFERENCES

[1] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] B. Harris, "The statistical estimation of entropy in the non-parametric case," tech. rep., WISCONSIN UNIV-MADISON MATHEMATICS RESEARCH CENTER, 1975.

[3] S. Zahl, "Jackknifing an index of diversity," *Ecology*, vol. 58, no. 4, pp. 907–913, 1977.

[4] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[5] Z. Zhang, "Entropy estimation in turing's perspective," *Neural computation*, vol. 24, no. 5, pp. 1368–1389, 2012.

[6] Z. Zhang and J. Zhou, "Re-parameterization of multinomial distributions and diversity indices," *Journal of Statistical Planning and Inference*, vol. 140, no. 7, pp. 1731–1738, 2010.

[7] Z. Zhang, "Asymptotic normality of an entropy estimator with exponentially decaying bias," *IEEE Transactions on Information Theory*, vol. 59, no. 1, pp. 504–508, 2013.

[8] Z. Zhang and X. Zhang, "A normal law for the plug-in estimator of entropy," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 2745–2747, 2012.

[9] X. Zhang, "Asymptotic normality of entropy estimators," 2013.

[10] C. Chen, M. Grabchak, A. Stewart, J. Zhang, and Z. Zhang, "Normal laws for two entropy estimators on infinite alphabets," *Entropy*, vol. 20, no. 5, p. 371, 2018.

[11] Z. Zhang, *Statistical Implications of Turing's Formula*. John Wiley & Sons, 2016.

[12] A. Antos and I. Kontoyiannis, "Convergence properties of functional estimates for discrete distributions," *Random Structures & Algorithms*, vol. 19, no. 3-4, pp. 163–193, 2001.

[13] Z. Zhang and L. Zheng, "A mutual information estimator with exponentially decaying bias," *Statistical applications in genetics and molecular biology*, vol. 14, no. 3, pp. 243–252, 2015.

[14] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[15] Z. Zhang and M. Grabchak, "Nonparametric estimation of küllback-leibler divergence," *Neural computation*, vol. 26, no. 11, pp. 2570–2593, 2014.