JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 528, 1620–1637: Theory and Methods
https://doi.org/10.1080/01621459.2018.1537921

BET on Independence*

Kai Zhang

Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC

ABSTRACT

We study the problem of nonparametric dependence detection. Many existing methods may suffer severe power loss due to nonuniform consistency, which we illustrate with a paradox. To avoid such power loss, we approach the nonparametric test of independence through the new framework of binary expansion statistics (BEStat) and binary expansion testing (BET), which examine dependence through a novel binary expansion filtration approximation of the copula. Through a Hadamard transform, we find that the symmetry statistics in the filtration are complete sufficient statistics for dependence. These statistics are also uncorrelated under the null. By using symmetry statistics, the BET avoids the problem of nonuniform consistency and improves upon a wide class of commonly used methods (a) by achieving the minimax rate in sample size requirement for reliable power and (b) by providing clear interpretations of global relationships upon rejection of independence. The binary expansion approach also connects the symmetry statistics with the current computing system to facilitate efficient bitwise implementation. We illustrate the BET with a study of the distribution of stars in the night sky and with an exploratory data analysis of the TCGA breast cancer data. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received December 2017
Revised October 2018

KEYWORDS
Binary expansion;
Contingency table;
Hadamard transform;
Multiple testing;
Nonparametric inference

1. Introduction

Independence is one of the most foundational concepts in statistics. It is also one of the most common assumptions in statistical literature. Thus, verifying independence is one of the most important testing problems. If we are not able to check this crucial condition, then we are "betting on independence" at the risk of losing the validity of our conclusions. In this article, we study the dependence detection problem in a distribution-free setting, in which we do not make any assumption on the joint distribution. We focus on the test of independence between two continuous variables, though the approach can be generalized for more variables. Without loss of generality, we consider n iid observations from the copula (U, V) whose marginal distributions are uniform over $[0, 1]$. This copula can be obtained by transformations with marginal cumulative distribution functions (CDFs) when they are known. In this case, U and V are independent if and only if their joint distribution $\mathbf{P}_{(U,V)}$ is the bivariate uniform distribution over $[0, 1]^2$, denoted by $\mathbf{P}_0$. We also study the case when the marginal CDFs are unknown. In this case, we can use the empirical CDFs, and the test is about the independence of observed ranks. The theory and procedures are shown to be similar.

Tests of independence have been extensively studied in statistics and information theory. One of the most classical parametric methods is based on the Pearson correlation, which can be interpreted as a measure of linear relationship. Classical results in Rényi (1959) connect correlation and independence. Recent tests based on robust versions of correlation include Han, Chen, and Liu (2017). Existing nonparametric testing procedures can be roughly categorized into three main classes:

(a) The CDF approach, which compares the joint CDF and the product of marginal CDFs: this pioneer approach includes variants of the Kolmogorov–Smirnov test such as Hoeffding (1948) and Romano (1989).

(b) The distance and kernel based approach, which can be regarded as a generalization of the correlation: one important recent development on dependence measures is the distance correlation (Székely et al. 2007; Székely and Rizzo 2009), which possesses the crucial property that a zero distance correlation implies independence. Tests based on sample versions of the distance correlation (Székely and Rizzo 2013a, 2013b) have since been popular methods. Other important methods include the generalized measures of correlation (GMC) by Zheng et al. (2012) and the Hilbert Schmidt independence criterion (HSIC) by Gretton et al. (2007), Sejdinovic et al. (2013), and Pfister et al. (2016) who study dependence through distances between embedding of distributions to reproducing kernel Hilbert spaces (RKHS).

(c) The binning approach, which generalizes the comparison of the joint density and the product of marginal ones: by discretizing X and Y into finite many categories, classical statistical or information theoretical methods such as the χ^2 tests and Fisher's exact tests can be applied to study the dependence. Miller and Siegmund (1982) studied the maximal χ^2 statistic from forming 2×2 tables through partitions of data. Reshef et al. (2011, 2015a, 2015b) introduced the maximal

CONTACT Kai Zhang zhangk@email.unc.edu Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599.

*This article is dedicated to Professor Lawrence D. Brown with the author's deepest appreciation for his teaching and mentoring.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

information coefficient (MIC) by aggregating information from optimal partitions of the scatterplot for different partition sizes. This approach was further studied by the k-nearest neighbor mutual information (KNN-MI) approach as described in Kraskov, Stögbauer, and Grassberger (2004) and Kinney and Atwal (2014). Heller, Heller, and Gorfine (2012), Heller et al. (2016), and Heller and Heller (2016) studied optimal permutation tests over partitions to improve the power. Filippi and Holmes (2015) took a Bayesian nonparametric approach to the partitions. Wang, Jiang, and Liu (2016) considered a generalized R^2 to detect piecewise linear relationships, a compromise between the distance approach and the binning approach that takes advantages of both. A very recent paper on Fisher exact scanning (FES) by Ma and Mao (2019) constructed multiscale scan statistics that are particularly effective at detecting local dependency through Fisher's exact tests over rectangle scanning windows.

Most of the above nonparametric tests enjoy the property of universal consistency against any particular form of dependence. Formally, this universality means that for any specific copula distribution $\mathbf{P}_1 \neq \mathbf{P}_0$, as $n \to \infty$, the test for the problem $H_0 : \mathbf{P}_{(U,V)} = \mathbf{P}_0$ versus $H_1 : \mathbf{P}_{(U,V)} = \mathbf{P}_1$ has an asymptotic power of 1. However, one important problem in many distribution-free tests is the lack of uniformity. To see this, we consider the total variation (TV) distance $\mathrm{TV}(\cdot,\cdot)$, which is defined by $\mathrm{TV}(\mathbf{P},\mathbf{Q}) = \sup_{S\in\mathcal{F}} |\mathbf{P}(S) - \mathbf{Q}(S)|$, where $\mathcal{F}$ is a σ-algebra of the sample space. The uniform consistency of nonparametric dependence detection w.r.t. the TV distance is to be consistent for any alternative which is certain distant from independence, that is,

$$H_0 : \mathbf{P}_{(U,V)} = \mathbf{P}_0 \text{ versus } H_1 : \mathrm{TV}(\mathbf{P}_{(U,V)}, \mathbf{P}_0) \geq \delta \qquad (1.1)$$

for some $0 < \delta \leq 1$. For the testing problem in (1.1), although many tests are universally consistent, we show in Section 2 and Theorem 2.2 the nonexistence of a test that is uniformly consistent w.r.t. the TV distance. The uniformity issue is due to the fact that the space of H_1 is large. Said another way: when two variables are not independent, there are so many ways they can be dependent. In practice, having this nonuniform consistency problem means having "blind spots" in dependence detection for a given sample size, that is, having very low power for many forms of dependency, especially nonlinear ones. Note that nonlinear forms of dependence are ubiquitous in sciences, for example, laws in physics defined by differential equations. Therefore, avoiding the power loss due to the nonuniform consistency problem in nonparametric dependence detection means having robust power against a large class of alternatives and improving the ability of discovering novel relationships in many areas of science.

Because of the impossibility of testing (1.1) with uniform consistency w.r.t. the TV distance (Theorem 2.2), to avoid such power loss, we propose to test approximate independence through a *filtration approach.* Such a filtration is constructed by the σ-fields generated by binary variables from marginal binary expansions which jointly approximate the copula distribution. Similar filtration ideas are nicely described in Liu and Meng (2014, 2016) in studying the Simpson's paradox. The approximation idea is also related to the "probably approximately correct" (PAC) approach in machine learning (Valiant 1984). We explain the details in Section 3.1.

We note here that although many other ways of filtration approximations are available, there are a few important advantages of the proposed binary expansion filtration that facilitate studies of dependence.

(a) The σ-field generated by binary variables is *finite.*

(b) Two binary variables are independent if and only if they are *uncorrelated.*

We call the statistics that are functions of the Bernoulli variables from the above filtration approximation binary expansion statistics (BEStat), and we call the testing framework on the corresponding approximate independence the binary expansion testing (BET) framework. This approach leads to studies of contingency tables from discretizations. Although classical tests such as the χ^2 tests (Lehmann and Romano 2006) are readily available, they have some drawbacks: (a) the exponentially growing degrees of freedom that would affect the power and (b) the unclear interpretability of dependence when the independence hypothesis is rejected. To improve on these two issues, we consider reparameterization of the likelihood of the contingency tables through a novel binary interaction design (BID) equation (Theorem 3.4), which connects the study of dependence to the Hadamard transform in signal processing. Through this connection, the interactions of binary variables in the filtration are shown to be complete sufficient statistics for dependence. By using these interactions, we convert the dependence detection problem to a multiple testing problem. Statistically speaking, the benefits of the above approach are summarized below:

(a) The Hadamard transform provides new insights for the analysis of any contingency table whose size is a power of 2. Compared to the conventional parameterization, the novel parameters marginal interaction odds ratios (MIOR) and cross interaction odds ratios (CIOR) separate the marginal and joint information, and CIORs being 1 is equivalent to independence. As an analogy, the CIORs are to contingency tables as the correlations are to multivariate normal distributions. See Theorems 3.7 and 3.8.

(b) The symmetry statistics from the reparameterization are shown to be complete sufficient statistics for dependence. They are identically distributed and are uncorrelated under the null. See Theorems 4.1–4.3.

(c) As a consequence of the above properties, the multiple testing procedure is shown to be minimax in the sample size requirement for reliable power. See Theorem 4.4.

(d) Upon rejection of independence, the largest absolute symmetry statistic and the corresponding cross interaction provide clear interpretation of the dependency.

Although theories for copula and contingency tables are well-developed, we are not aware of similar approach or results in statistical literature.

We also note that the BEStat approach is closely related to computing. In current computing systems, each decimal number is coded as a sequence of binary bits, which is exactly the binary expansion of that number. This connection means that one can carry out the BEStat procedures by operating directly

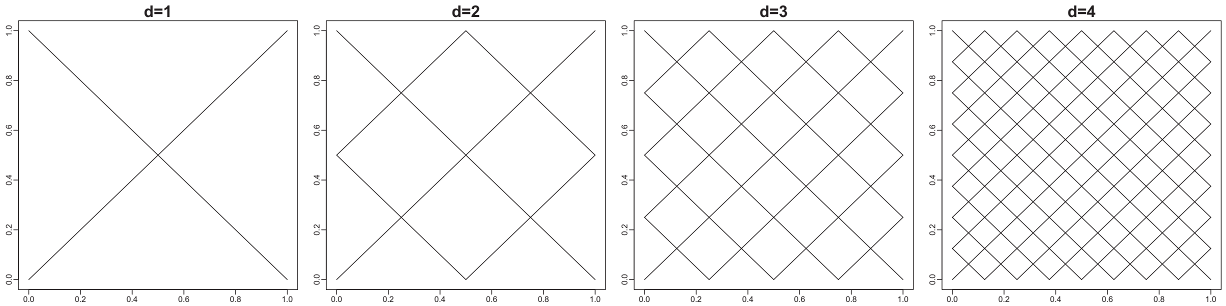


Figure 1. The bisection expanding cross (BEX) at level $d = 1, \ldots, 4$.

over bits. Since bitwise operations are one of the most efficient operations in current computing systems, we are able to develop computationally efficient implementations of the proposed method. The detailed algorithm is described in a separate paper (Zhao et al. 2019), and it improves the speed of existing methods by orders of magnitude.

This article is organized in as follows: Section 2 explains the problem of nonuniform consistency. Section 3 introduces the concept and basic theory in the framework of BEStat and BET. Section 4 studies the Max BET procedure and its properties. Section 5 connects the BEStat framework to current computing system. Sections 6–8 illustrate the procedure with simulated and real data studies. Section 9 concludes the article with discussions of future work. The proofs can be found in the supplementary materials.

2. Motivation: Nonuniform Consistency

To explain the problem of nonuniform consistency, we develop the following example of the bisection expanding cross (BEX). Many existing methods suffer substantial power loss under this example due to this problem, which can be avoided through the BEStat proposed in Sections 3 and 4.

We call the following sequence of one-dimensional manifolds in $[0, 1]^2$ the BEX. These manifolds can be defined through the implicit function $\gamma_d(x, y) = 0$ for every integer $d > 0$: $\mathrm{BEX}_d = \{(x, y) \in [0, 1]^2 : \gamma_d(x, y) = 0\}$, where

$$\gamma_d(x,y)=\sum_{i=1}^{2^{d-1}}\sum_{j=1}^{2^{d-1}}\left(\left|x-\frac{i}{2^{d-1}}+\frac{1}{2^d}\right|-\left|y-\frac{j}{2^{d-1}}+\frac{1}{2^d}\right|\right)$$
$$\times\, \mathrm{I}\left(\left|x-\frac{i}{2^{d-1}}+\frac{1}{2^d}\right|\le\frac{1}{2^d}\right)\mathrm{I}\left(\left|y-\frac{j}{2^{d-1}}+\frac{1}{2^d}\right|\le\frac{1}{2^d}\right).$$

The BEX structure is illustrated in Figure 1, where the first four levels are plotted. Graphically, this grid can be regarded as a space-filling fractal by recursively expanding the bisector of the four "arms" of BEX_1 until intersection.

Now we consider the random variables (X_d, Y_d) that are uniformly distributed over BEX_d whose joint distribution is denoted by $\mathbf{P}_d$. The properties of these distributions are summarized in the following proposition.

Proposition 2.1.

(a) X_d and Y_d are marginally Uniform$[0, 1]$ for any d.

(b) $\gamma_d(X_d, Y_d) = 0$ for any d, that is, the joint distribution of (X, Y) is degenerate. In particular, $\mathrm{TV}(\mathbf{P}_d, \mathbf{P}_0) = 1$ for any d.

(c) $\forall (x, y) \in [0, 1]^2$, as $d \to \infty$, $|\mathbf{P}_d(X_d \le x, Y_d \le y) - \mathbf{P}_d(X_d \le x)\mathbf{P}_d(Y_d \le y)| \to 0$.

Part (b) and part (c) of Proposition 2.1 seem to contradict each other: part (b) says that the joint distribution of X_d and Y_d is far away from independence in the TV distance, thus they are strongly nonindependent. Yet, part (c) claims that when d is large, X_d and Y_d are nearly independent. Indeed, the BEX shows that despite a TV distance of 1, degenerate distributions can be arbitrarily close to independence. We shall explain this paradox in Section 4.3. This paradox also lead to a challenge to testing methods: given a finite sample, can we effectively distinguish any form of dependency from independence?

Unfortunately, for any testing method, the answer is negative. Intuitively speaking, this is because for any given test with a given samples size n, one can keep expanding the BEX until it is so close to independence that this test becomes powerless. This example thus illustrates the problem of nonuniform consistency of the test in (1.1): no test can be uniformly consistent against all forms of dependence, not even all levels of the BEX, for which $\delta = 1$ in (1.1). See Theorem 2.2.

The power loss due to nonuniform consistency can be severe. For example, simulations (see Section 1.1 in the supplementary materials) show that many CDF based and kernel based tests are powerless in detecting BEX at level 4 even when the sample size is as high as 20,000. Note that with such a large sample, the BEX structure and the dependency can be clearly observed in the scatterplot by naked eyes. However, many existing tests cannot distinguish it from independence.

We make a few remarks about the BEX example before proceeding.

(a) The BEX is closely related to many research problems such as the chessboard detection in computer vision (Forsyth and Ponce 2002).

(b) The BEX is not the first example that a sequence of degenerate distributions converges to independence. The earliest example we could find is in Kimeldorf and Sampson (1978). There are also other interesting and useful fractal applications in statistics such as Craiu and Meng (2005, 2006). The basis of the BEX example is a classical result in Vitale (1990). We construct the BEX paradox due to its fractal structure which explains the problem of nonuniform consistency.

(c) The nonuniform consistency shown with the BEX is specifically for our choice of the TV distance between distributions. There are many other distances (Tsybakov 2008), and a different choice of distance could lead to a different test statistic and different results on uniform consistency. We choose the TV distance because (1) it is a widely used distance in literature, (2) it is equivalent to many other distances, and (3) it is convenient for the analysis in our binary expansion approach. Therefore, throughout this article, we focus on the TV distance, and all results about uniform consistency are w.r.t. the TV distance. In particular, we provide a formal statement of the problem of nonuniform consistency w.r.t. the TV distance below:

Theorem 2.2. Consider the testing problem in (1.1). For any finite number of iid observations n, for any test that has a Lebesgue measurable critical region $C_n \subset \mathbb{R}^{2n}$ with $\mathbf{P}_{H_0}(\partial C_n) = 0$ and $\mathbf{P}_{H_0}(C_n) \leq \alpha, \forall \epsilon > 0$, there exists a bivariate distribution $F_n \in H_1$ and $\mathbf{P}_{F_n}(C_n) \leq \alpha + \epsilon$.

The message of Theorem 2.2 is that in a distribution-free setting without any assumption on the joint distribution, dependence is not a tractable target. The intractability comes from the fact that without a model of the joint distribution, there is no parameter to characterize and identify the underlying form of dependency. Therefore, there is no target for inference about dependence from a test or any other statistical method. Although one can develop good measures of dependence such as distance correlation, GMC, HSIC, and MIC, such measures cannot make the joint distribution identifiable. Therefore, they can never replace the role of parameters in statistical inference about dependence. This fact motivates the following three key elements in the BEStat approach and the BET framework:

(a) Rather than one test of independence, we will study dependence through a carefully designed sequence of tests based on a filtration to achieve *universality*.
(b) For every test statistic in the sequence, there is an explicit well-defined set of parameters as the target for inference to achieve *identifiability*.
(c) At every step in the sequence, the test is consistent against all alternatives which are δ-away from independence in the TV distance to achieve *uniformity*.

The above BET framework can help explain the seeming paradox in the BEX example, and the proposed test can have high power against this dependency. See Section 4.3.

3. The Basic Theory of Binary Expansion Statistics

3.1. Binary Expansion Filtration

The considerations in Section 2 necessitate a multiscale binning approach to study dependence. For the dependence detection problem, this multiscale approach means to test some approximate independence rather than the exact hypothesis in (1.1). We study the known marginal CDF case first, for which we develop such a multiscale framework through the following classical result on the binary expansion of a uniform random variable (Kac 1959):

Theorem 3.1. If $U \sim \text{Uniform}[0,1]$, then $U = \sum_{k=1}^{\infty} \frac{A_k}{2^k}$ where $A_k \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$.

The binary expansion in Theorem 3.1 decomposes the information about U into information from independent Bernoulli A_ks. A_ks can be also regarded as indicator functions of U. For example, $A_1 = \text{I}(U \in (1/2, 1])$, $A_2 = \text{I}(U \in (1/4, 1/2] \cup (3/4, 1])$, see Kac (1959). To study the dependence between U and V, we consider the binary expansion of both U and V: $U = \sum_{k=1}^{\infty} \frac{A_k}{2^k}$ and $V = \sum_{k=1}^{\infty} \frac{B_k}{2^k}$ where $A_k \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$ and $B_k \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$.

Note that if we truncate the binary expansions of U and V at some finite depths d_1 and d_2, respectively, $U_{d_1} = \sum_{k=1}^{d_1} \frac{A_k}{2^k}$ and $V_{d_2} = \sum_{k=1}^{d_2} \frac{B_k}{2^k}$, then U_{d_1} and V_{d_2} are two discrete variables that can take 2^{d_1} and 2^{d_2} possible values, respectively. Moreover, as $d_1, d_2 \to \infty$, $|U_{d_1} - U| = O_p(2^{-d_1})$ and $|V_{d_2} - V| = O_p(2^{-d_2})$. In particular,

$$\|(U_{d_1}, V_{d_2}) - (U, V)\|_2 = O_p(2^{-\min\{d_1, d_2\}}). \qquad (3.1)$$

The above considerations are apparent if one regards the truncations as a filtration generated by $\{A_k\}_{k=1}^{d_1}$ and $\{B_k\}_{k=1}^{d_2}$ for each $d_1, d_2 \geq 1$. Indeed, the filtration idea is a consequence of George Box's aphorism "All models are wrong, but some are useful." At every d_1 and d_2, the probability model of (U_{d_1}, V_{d_2}) is a "wrong" model for the joint distribution (U, V). However, the "wrong" model of (U_{d_1}, V_{d_2}) can be very useful in many ways. In particular, we show below how the three key elements described at the end of Section 2 are achieved from this approach:

(a) *Universality*: The important message from (3.1) is that one can approximate the joint distribution of and hence the dependence in (U, V) through that in (U_{d_1}, V_{d_2}). Although the dependence in the joint distribution of (U, V) can be arbitrarily complicated, when d_1 and d_2 are large, we expect a good approximation from discrete variables (U_{d_1}, V_{d_2}) where the approximation error is exponentially small. In terms of testing independence, this means although the joint distribution of (U, V) can be arbitrarily close to independence, due to the filtration feature of the sequence, one can always detect the dependence when d_1 and d_2 are large to achieve universality.

(b) *Identifiability*: As we explained in Section 2, one crucial challenge in distribution-free dependence detection is identifiability. Without models and parameters, dependence is not a tractable target. On the other hand, (U_{d_1}, V_{d_2}) can only take a finite $2^{d_1+d_2}$ possible values, which leads to a partition of the scatterplot of data into a $2^{d_1} \times 2^{d_2}$ contingency table. With this consideration, the truncation of the binary expansions turns the problem on dependence, which is unidentifiable under the distribution-free setting, into a problem over a contingency table, which is fully identifiable. In terms of testing, when we begin without any assumptions about the joint distribution, there is no explicit way to write out the alternative likelihood under dependence. However, at each depths d_1 and d_2, due to the discreteness, the class of alternative distributions is restricted to those over the contingency table, which has an explicit distribution and has cell probabilities as identifiable parameters for inference (Agresti and Kateri 2011; Fienberg 2007).

(c) *Uniformity*: As a consequence of identifiability, we can avoid the problem of nonuniform consistency described in Section 2. At any depths d_1 and d_2, one can write out the TV distance between an alternative distribution and the null distribution in terms of the cell probabilities in the contingency table model. We are thus able to show the consistency and optimality of the proposed Max BET procedure in Section 4.2 for alternative distributions whose TV distances from the independence null is at least δ, for any $\delta > 0$.

The above considerations motivate us to propose the BEStat in studying the dependence between U and V in a distribution-free setting. Formally, we define BEStat as follows:

Definition 3.2. We call statistics as functions of finitely many Bernoulli variables from marginal binary expansions the BEStat.

Similarly, for the problem of detecting dependence from independence in a distribution-free setting, we define the BET framework as follows.

Definition 3.3. We call the testing framework based on the binary expansion filtration approximation up to certain depth the BET.

In the context of testing independence in bivariate distributions, the BET at depths d_1 and d_2 is to test the independence of U_{d_1} and V_{d_2}, which we refer to as (d_1, d_2)-independence and which is equivalently defined in Ma and Mao (2019) for scanning statistics. Formally, denote the bivariate uniform distribution over $\{\frac{0}{2^{d_1}}, \dots, \frac{2^{d_1}-1}{2^{d_1}}\} \times \{\frac{0}{2^{d_2}}, \dots, \frac{2^{d_2}-1}{2^{d_2}}\}$ by $\mathbf{P}_{0,d_1,d_2}$. For some $0 < \delta \le 1$, we consider

$$H_{0,d_1,d_2} : \mathbf{P}_{(U_{d_1},V_{d_2})} = \mathbf{P}_{0,d_1,d_2} \;\; v.s. \;\; H_{1,d_1,d_2} : \mathrm{TV}(\mathbf{P}_{(U_{d_1},V_{d_2})}, \mathbf{P}_{0,d_1,d_2}) \ge \delta. \qquad (3.2)$$

Not rejecting the null hypothesis in the BET at depths (d_1, d_2) thus indicates that there is no strong evidence against the null hypothesis of independence between U and V up to depths d_1 and d_2 in the binary expansions. Note that this interpretation is weaker than claiming independence between U and V: the dependence can occur at some larger (d_1, d_2) in the $O_p(2^{-\min\{d_1,d_2\}})$ remainder term in (3.1). However, as described in Section 2, claiming exact independence with finite samples and without any restriction on the alternative is impossible. On the other hand, this weaker hypothesis of approximate independence helps us to avoid the uniform consistency problem in the dependence detection under the distribution-free setting and provides reliable power for a large class of alternatives. To see the gains from this trade-off, one can compare our results in Section 4.2 with those in Section 2.

We remark here that the filtration in approximating dependence is not unique. For example, one can consider the filtration corresponding to orthogonal polynomials rather than the binary expansion. However, the σ-field in the binary expansion filtration has a few important advantages to facilitate studies of dependence.

(a) Finiteness of σ-fields: For the σ-field at each depths d_1 and d_2, the number of events is $2^{d_1+d_2} - 1$, which is finite. This is because interactions of binary variables are at most binary. If we consider some other filtration (e.g., orthogonal polynomials) for the approximation of dependence, then the σ-field might not be of finitely many events and can be much more complicated.

(b) Uncorrelatedness implying independence: Although uncorrelatedness usually does not imply independence, it is well known that it does for two binary variables. This property can greatly simplify studies of dependence in filtration. Again, if we consider some other filtration (e.g., orthogonal polynomials) for the approximation of dependence, then quantifying the dependence between variables in the σ-field can be much more complicated.

The above considerations also work similarly for the case when the marginal distributions are unknown. To study the binary expansion in this case, suppose the sample size is $n = 2^K$ for some $K > 0$ for easy explanation. With the marginal empirical CDF transformations, the ith observation in the empirical copula are $\widehat{U}_i$ and $\widehat{V}_i$ whose marginal distribution is Uniform$\{\frac{1}{2^K}, \dots, \frac{2^K}{2^K}\}$. Now let $\widehat{A}_{1,i} = \mathrm{I}(\widehat{U}_i \in (1/2, 1]), \dots, \widehat{A}_{K,i} = \mathrm{I}(\widehat{U}_i \in \cup_{k'=1}^{2^{K-1}}(\frac{2k'-1}{2^K}, \frac{2k'}{2^K}])$. It is easy to see that for each fixed i, $\widehat{A}_{k,i}$s are independent, and $\widehat{U}_i = \frac{1}{2^K} + \sum_{k=1}^{K} \frac{\widehat{A}_{k,i}}{2^k}$. Therefore, the binary expansion filtration can be similarly defined, and the BET at depths d_1 and d_2 is to test the independence of $\widehat{U}_{d_1,i} = \sum_{k=1}^{d_1} \frac{\widehat{A}_{k,i}}{2^k}$ and $\widehat{V}_{d_2,i} = \sum_{k=1}^{d_2} \frac{\widehat{B}_{k,i}}{2^k}$

$$H_{0,d_1,d_2} : \text{For each } i, \widehat{U}_{d_1,i} \text{ and } \widehat{V}_{d_2,i} \text{ are independent.} \qquad (3.3)$$

The interpretation of this null hypothesis is that for each observation, the row assignment and column assignment to the contingency table are independent, as in classical categorical data analysis (Agresti and Kateri 2011; Fienberg 2007). When $\widehat{U}_{K,i}$ and $\widehat{V}_{K,i}$ are independent for each i, the observed ranks are independent.

We explain the details of these tests in Sections 3.2 and 4. We remark here that although copula theory is well developed (Nelsen 2007), we are not aware of any filtration approach in the literature. We also remark here that tests of approximate independence are also considered in a very recent paper (Ma and Mao 2019) for scanning purposes, in which a filtration idea is implicitly described. In this article, our goal is to formally develop the framework of BEStat. We shall compare the theory and methods in both papers in Section 4.4.

3.2. *Revisiting the Classical Theory for Contingency Tables*

We start our analysis by first revisiting the model and theory of a general contingency table with r rows and c columns of n iid samples. The parameters of interest are $\boldsymbol{p} = \{p_{ij}, i = 1, \dots, r, j = 1, \dots, c\}$, and the cell counts are $\boldsymbol{n} = \{n_{ij}\}$. The only constraint is on the totals $\sum_{i,j} p_{ij} = 1$ and $\sum_{i,j} n_{ij} = n$. Two most important models for the likelihood are as follows (Agresti and Kateri 2011; Fienberg 2007):

(a) When there is no restriction on marginal totals, the joint distribution of the cell count vector $\boldsymbol{N}$ is multinomial (with the convention $0^0 = 1$): with $C_1(\boldsymbol{n}) = \frac{n!}{\prod_{i,j} n_{ij}!}$,

$$p(\boldsymbol{N} = \boldsymbol{n} | \boldsymbol{p}) = C_1(\boldsymbol{n}) \prod_{i,j} p_{ij}^{n_{ij}}. \qquad (3.4)$$

(b) Condition on positive row and column totals $\boldsymbol{n}_r = \{n_{i\cdot} = \sum_j n_{ij}, i = 1, \dots, r\}$ and $\boldsymbol{n}_c = \{n_{\cdot j} = \sum_i n_{ij}, j = 1, \dots, c\}$, for $i < r$ and $j < c$, with the reparameterization $\theta_{ij} =$

$\frac{p_{ij}p_{rc}}{p_{ic}p_{rj}}$ and normalizing constant $h_1(\boldsymbol{n}_r, \boldsymbol{n}_c, \boldsymbol{\theta})$, we have $p(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{\theta}, \boldsymbol{n}_r, \boldsymbol{n}_c) = C_1(\boldsymbol{n})h_1(\boldsymbol{n}_r, \boldsymbol{n}_c, \boldsymbol{\theta}) \prod_{i,j} \theta_{ij}^{n_{ij}}$ (Cornfield 1956). Note that under independence $\theta_{ij} = 1$, and the distribution is (central) multivariate hypergeometric

$$p(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{n}_r, \boldsymbol{n}_c) = C_1(\boldsymbol{n})h_1(\boldsymbol{n}_r, \boldsymbol{n}_c) = \frac{\prod_i n_{i\cdot}! \prod_j n_{\cdot j}!}{n! \prod_{i,j} n_{ij}!}. \quad (3.5)$$

With the above distributions, tests of independence for a contingency table can be done through classical methods such as χ^2 tests, Fisher's exact tests, and likelihood ratio tests (LRT). For the nonparametric dependence detection problem, the BET with these tests are uniformly consistent for any depths d_1 and d_2. However, these classical methods have two important limitations on power and interpretability:

(a) The minimal sample size for classical tests to have reliable power is known (Agresti and Kateri 2011; Fienberg 2007) to be about the size of the contingency table $O(2^{d_1+d_2})$. However, recent developments (Acharya, Daskalakis, and Kamath 2015) show that the optimal lower bound of this sample size requirement is $O(2^{\frac{d_1+d_2}{2}})$. This result indicates that classical tests may suffer substantial power loss in dependence detection, especially when d_1 and d_2 are large. For a well-known example, when the contingency table contain many empty cells, LRT and χ^2 tests will fail to work.

(b) The rejections from classical tests are not very interpretable. Even if we can claim significant dependence with a classical test, the test does not provide information about how the variables are dependent.

One intuition of the above limitations in classical tests is that each cell in a contingency table is considered in an isolated manner, thus the information between cells is somehow lost. To improve classical tests, we consider grouping the cells together to improve the power and interpretability. Such grouping process is effectively achieved through the BID described in Section 3.3.

3.3. Binary Interaction Design: Reparameterization of the $2^{d_1} \times 2^{d_2}$ Contingency Table Likelihood

We now turn to the case when the contingency table is generated by the binary expansion up to depths d_1 and d_2 as described in Section 3.1, so that the table has 2^{d_2} rows and 2^{d_1} columns (assuming U on the horizontal axis and V on the vertical axis). To provide a general theory for contingency tables, in this subsection we *do not* restrict the total probability of each row and column being the same (which happens when A_is and B_js are both iid Bernoulli(1/2)). However, in this subsection, we shall assume that all cell probabilities are positive.

To combine the cell information, we consider the σ-field generated from the binary expansion filtration. We explain in the known marginal distribution case first since it is similar for the unknown marginal distribution case. With d_1 Bernoulli variables $A_k, k = 1, \dots, d_1$ and another d_2 Bernoulli variables $B_k, k = 1, \dots, d_2$ (again in this subsection we *do not* assume them to be independent and symmetric), consider two general discrete variables defined by $U_{d_1} = \sum_{k=1}^{d_1} \frac{A_k}{2^k}$ and $V_{d_2} = \sum_{k=1}^{d_2} \frac{B_k}{2^k}$. The σ-field here is $\sigma(U_{d_1}, V_{d_2}) = \sigma(A_1, \dots, A_{d_1}, B_1, \dots, B_{d_2})$ and is generated by $2^{d_1+d_2} - 1$ Bernoulli variables resulting from interactions between A_is and B_js. We shall use the equivalent binary variables $\dot{A}_i = 2A_i - 1$ and $\dot{B}_j = 2B_j - 1$ since the interaction between them can be conveniently written as products. For example, the event $\{A_1 = 1, B_1 = 1\} \cup \{A_1 = 0, B_1 = 0\}$ is equivalent to the event $\{\dot{A}_1\dot{B}_1 = 1\}$.

Note that each of these binary interaction variables leads to a partition of the unit square $[0,1]^2$ and two groups of cells according to whether the interaction is positive. Moreover, for each interaction in the σ-field, the number of cells in the regions where it takes value 1 (and -1) is exactly $2^{d_1+d_2-1}$. This fact can be explained by the BID equation (Theorem 3.4), and it facilitates the definition of interaction odds ratio (IOR) as in Definition 3.6 as well as the reparameterization with IOR. The IORs group the cell information together and separate the marginal and joint information in the multinomial likelihood. See Figure 2.

Note also that the $2^{d_1+d_2} - 1$ binary variables in the σ-field can be categorized into two classes: the variables of the form $\dot{A}_{k_1} \dots \dot{A}_{k_r}$ or $\dot{B}_{k'_1} \dots \dot{B}_{k'_t}$ will be referred to as *marginal interactions* since they only involve the marginal distributions. On the other hand, the variables of the form $\dot{A}_{k_1} \dots \dot{A}_{k_r}\dot{B}_{k'_1} \dots \dot{B}_{k'_t}$ with $r, t > 0$ will be referred to as *cross-interactions* since they contain information of both U_{d_1} and V_{d_2}.

In explanation of the theory, we use the following binary integer indexing for related quantities: denote the Bernoulli random vectors in the binary expansion by $\boldsymbol{A} = (A_1, \dots, A_{d_1})$ and $\boldsymbol{B} = (B_1, \dots, B_{d_2})$, and denote vectors of length d_1 and d_2 with entries 0s and 1s by $\boldsymbol{a}$ and $\boldsymbol{b}$. The probability of each of the $2^{d_1+d_2}$ cells can then be written as $p_{(\boldsymbol{ab})} = \mathbf{P}(\boldsymbol{A} = \boldsymbol{a}, \boldsymbol{B} = \boldsymbol{b})$ with $(\boldsymbol{ab})$ being the concatenation of $\boldsymbol{a}$ and $\boldsymbol{b}$. Now let the integer c determined by $c = \sum_{i=1}^{d_1} a_i 2^{d_1+d_2-i} + \sum_{j=1}^{d_2} b_j 2^{d_2-j}$. Let $\boldsymbol{p}$ be the $2^{d_1+d_2}$-dimensional vector of probabilities whose $(2^{d_1+d_2} - c)$th entry is $p_{(\boldsymbol{ab})}$.

For the binary variables in $\sigma(\dot{A}_1, \dots, \dot{A}_{d_1}, \dot{B}_1, \dots, \dot{B}_{d_2})$, we also denote their expected values with binary integer index as follows. For $\mathbf{E}[\dot{A}_{k_1} \dots \dot{A}_{k_r}\dot{B}_{k'_1} \dots \dot{B}_{k'_t}], r = 1, \dots, d_1, t = 1, \dots, d_2$, we denote it by $E_{(\boldsymbol{ab})}$ where $\boldsymbol{a}$ is a d_1-dimensional binary vector with 1s at $k_1, \dots, k_r$ and are 0s otherwise, and $\boldsymbol{b}$ is a d_2-dimensional binary vector with 1s at $k'_1, \dots, k'_t$ and are 0s otherwise. Note here that $E_{(\mathbf{00})} = \mathbf{E}[1] = 1$. We also write the interaction as a product of binary variables $\dot{A}_{k_1} \dots \dot{A}_{k_r}\dot{B}_{k'_1} \dots \dot{B}_{k'_t}$ as $\dot{A}_{\boldsymbol{a}}\dot{B}_{\boldsymbol{b}}$. With c defined in the previous paragraph, let $\boldsymbol{E}$ be the $2^{d_1+d_2}$-dimensional vector of expected values whose $(c+1)$th entry is $E_{(\boldsymbol{ab})}$.

The above notation also applies to observed quantities: with the total n observations, the cell counts are denoted by $n_{(\boldsymbol{ab})}$. The collection of all $n_{(\boldsymbol{ab})}$s is denoted by $\boldsymbol{N}$ and is indexed as in $\boldsymbol{p}$. We also denote the sum of observed binary interaction variables by $S_{(\boldsymbol{ab})} = \sum_{i=1}^{n} \dot{A}_{\boldsymbol{a},i}\dot{B}_{\boldsymbol{b},i}$ with $S_{(\mathbf{00})} = n$. The collection of all $S_{(\boldsymbol{ab})}$s is denoted by $\boldsymbol{S}$ and is indexed as in $\boldsymbol{E}$. We shall refer $S_{(\boldsymbol{ab})}$ as *the symmetry statistic* for $\dot{A}_{\boldsymbol{a}}\dot{B}_{\boldsymbol{b}}$ as they can be regarded as the differences between the numbers of points in positive and negative regions. Thus, $S_{(\boldsymbol{ab})}$ is a statistic about symmetry. See Figure 2.

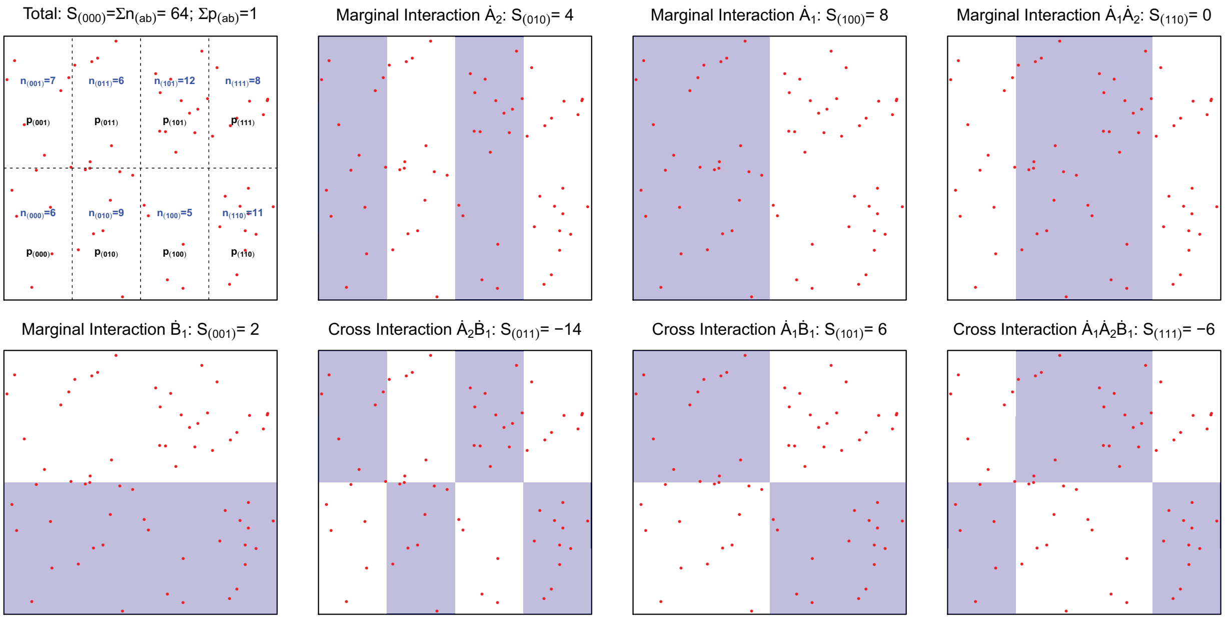


Figure 2. The binary interaction design (BID) at depths $d_1 = 2$ and $d_2 = 1$ with $n = 64$ observations. The number of observations in each cell is presented in the top left plot. There are seven nontrivial binary variables in the σ-field, whose positive regions are in white and whose negative regions are in blue. Symmetry statistics $S_{(\boldsymbol{ab})}$ are calculated for these four marginal interactions and three cross interactions. For example, $S_{(011)} = n_{(111)} - n_{(110)} - n_{(101)} + n_{(100)} + n_{(011)} - n_{(010)} - n_{(001)} + n_{(000)} = -14$.

With the above notation, we establish the equation connecting the contingency table distribution and the interactions of binary variables in the σ-field. The equation is established through $\mathbf{H} = \mathbf{H}_{2^{d_1+d_2}}$ being the Sylvester's construction of Hadamard matrix (Sylvester 1867). We shall refer this equation as the BID equation (name coined in Zhao et al. 2019).

Theorem 3.4.

(a) Population version of the BID equation: $\boldsymbol{E} = \mathbf{H}\boldsymbol{p}$.
(b) Sample version of the BID equation: $\boldsymbol{S} = \mathbf{H}\boldsymbol{N}$.

The Hadamard matrix $\mathbf{H}$ is referred to as Walsh matrix in literature of signal processing, where a linear transformation with $\mathbf{H}$ as in Theorem 3.4 is referred to as the Hadamard transform (Lynn 1973; Golubov et al. 2012; Harmuth 2013). The earliest referral to the Hadamard matrix we found in statistical literature is Pearl (1971). The Hadamard matrix is also closely related to the orthogonal full factorial design (Box, Hunter, and Hunter 2005; Cox and Reid 2000). In the context of dependence detection, this transform maps the cell domain (in $\boldsymbol{p}$ or $\boldsymbol{N}$) to the interaction domain (in $\boldsymbol{E}$ or $\boldsymbol{S}$). Thus, the information in individual cells can be grouped together to provide information about global dependency. Although theory and methods for contingency tables are well-developed, we are not aware of similar approach in related literature.

To see the importance of the BID equation and the symmetry statistic $S_{(\boldsymbol{ab})}$, we introduce some more notation here. We label the first to $2^{d_1+d_2}$th row (and column) of $\mathbf{H}$ with binary integer indices from $(\mathbf{0}_{d_1+d_2})$ to $(\mathbf{1}_{d_1+d_2})$. Denote $\overline{(\boldsymbol{ab})} = (\mathbf{11}) - (\boldsymbol{ab})$ to be the *binary conjugate*, or logical negation of $(\boldsymbol{ab})$, that is, $\overline{(010)} = (101)$. With the above notation, we summarize some useful properties of the Hadamard matrix $\mathbf{H}_{2^{d_1+d_2}}$ in the following proposition (Golubov et al. 2012).

Proposition 3.5.

(a) $\mathbf{H}_{2^{d_1+d_2}}$ is symmetric. The entry in $\mathbf{H}_{2^{d_1+d_2}}$ at the $(\boldsymbol{a}'\boldsymbol{b}')$th row and $(\boldsymbol{ab})$th column is $(-1)^{(\boldsymbol{a}'\boldsymbol{b}')^T(\boldsymbol{ab})}$.
(b) $\mathbf{H}_{2^{d_1+d_2}}$ has orthogonal columns: $\mathbf{H}^{-1}_{2^{d_1+d_2}} = \frac{1}{2^{d_1+d_2}}\mathbf{H}_{2^{d_1+d_2}}$.
(c) Hadamard matrices can be defined recursively: $\mathbf{H}_{2^{d_1+d_2+1}} = \mathbf{H}_{2^{d_1+d_2}} \otimes \mathbf{H}_2$.

Part (b) of Proposition 3.5 implies that $\boldsymbol{N} = \frac{1}{2^{d_1+d_2}}\mathbf{H}\boldsymbol{S}$, that is, $n_{(\boldsymbol{ab})} = \frac{1}{2^{d_1+d_2}}\mathbf{H}^T_{\overline{(\boldsymbol{ab})}}\boldsymbol{s}$ where $\mathbf{H}_{\overline{(\boldsymbol{ab})}}$ is the $\overline{(\boldsymbol{ab})}$th column of $\mathbf{H}$. With the above notation and transformation of variables, and by part (a) of Proposition 3.5, the multinomial distribution in the contingency table (3.4) can be written as

$$p(\boldsymbol{N} = \boldsymbol{n}|\boldsymbol{p}) = \frac{n!}{\prod_{\boldsymbol{a},\boldsymbol{b}} n_{(\boldsymbol{ab})}!} \prod_{\boldsymbol{a},\boldsymbol{b}} \left(\prod_{\boldsymbol{a}',\boldsymbol{b}'} p_{(\boldsymbol{a}'\boldsymbol{b}')}^{(-1)^{\overline{(\boldsymbol{a}'\boldsymbol{b}')}^T(\boldsymbol{ab})}} \right)^{\frac{s_{(\boldsymbol{ab})}}{2^{d_1+d_2}}}. \tag{3.6}$$

We are now ready to introduce the IOR:

Definition 3.6. We call $\lambda_{(\boldsymbol{ab})} = \prod_{\boldsymbol{a}',\boldsymbol{b}'} p_{(\boldsymbol{a}'\boldsymbol{b}')}^{(-1)^{\overline{(\boldsymbol{a}'\boldsymbol{b}')}^T(\boldsymbol{ab})}}$ the IOR with respect to the interaction $\dot{A}_{\boldsymbol{a}}\dot{B}_{\boldsymbol{b}}$. Denote the vector of $\lambda_{(\boldsymbol{ab})}$s by $\boldsymbol{\lambda}$ and order the entries in the same way as in $\boldsymbol{E}$.

For each corresponding interaction, the IOR can be regarded as the ratio of the product of all white cell probabilities to the product of all blue cell probabilities. There are three cases for the IOR $\lambda_{(\boldsymbol{ab})}$:

(a) When $\boldsymbol{a} = \mathbf{0}$ and $\boldsymbol{b} = \mathbf{0}$, $\lambda_{(\mathbf{00})} = \prod_{\boldsymbol{a}',\boldsymbol{b}'} p_{(\boldsymbol{a}'\boldsymbol{b}')}$. Note that the term $\lambda_{(\mathbf{00})}^{\frac{n}{2^{d_1+d_2}}}$ does not involve $\boldsymbol{N}$ and is constant.

(b) When $\boldsymbol{a} = \mathbf{0}$ but $\boldsymbol{b} \neq \mathbf{0}$ (or when $\boldsymbol{b} = \mathbf{0}$ but $\boldsymbol{a} \neq \mathbf{0}$), then $\lambda_{(\boldsymbol{ab})}$ is an *MIOR* quantifying the balance in the marginal interaction variable $\dot{A}_{\boldsymbol{a}}$ (or $\dot{B}_{\boldsymbol{b}}$). For example, when $d_1 = 2$ and $d_2 = 1$, $\lambda_{(110)} = \frac{p_{(111)}p_{(110)}p_{(001)}p_{(000)}}{p_{(101)}p_{(100)}p_{(011)}p_{(010)}}$ which is related to the distribution of $\dot{A}_1\dot{A}_2$. Note also that there are $2^{d_1}+2^{d_2}-2$ MIORs at depths d_1 and d_2.

(c) When $\boldsymbol{a} \neq \mathbf{0}$ and $\boldsymbol{b} \neq \mathbf{0}$, then $\lambda_{(\boldsymbol{ab})}$ is a *CIOR* quantifying the balance in the cross interaction variable $\dot{A}_{\boldsymbol{a}}\dot{B}_{\boldsymbol{b}}$. For example, when $d_1 = 2$ and $d_2 = 1$, $\lambda_{(111)} = \frac{p_{(111)}p_{(100)}p_{(010)}p_{(001)}}{p_{(110)}p_{(101)}p_{(011)}p_{(000)}}$ which is related to the distribution of $\dot{A}_1\dot{A}_2\dot{B}_1$. Note also that there are $(2^{d_1} - 1)(2^{d_2} - 1)$ CIORs at depths d_1 and d_2, which matches the degree of freedom for the χ^2 test.

An important observation is that with the IOR, (3.6) becomes

$$p(\boldsymbol{S} = \boldsymbol{s}|\boldsymbol{\lambda}) = C_2(\boldsymbol{s})h_2(\boldsymbol{\lambda}) \exp\left(\sum_{\boldsymbol{a}\neq\mathbf{0}} \frac{s_{(\boldsymbol{a}\mathbf{0})} \log \lambda_{(\boldsymbol{a}\mathbf{0})}}{2^{d_1+d_2}} + \sum_{\boldsymbol{b}\neq\mathbf{0}} \frac{s_{(\mathbf{0}\boldsymbol{b})} \log \lambda_{(\mathbf{0}\boldsymbol{b})}}{2^{d_1+d_2}} + \sum_{\substack{\boldsymbol{a}\neq\mathbf{0} \\ \boldsymbol{b}\neq\mathbf{0}}} \frac{s_{(\boldsymbol{ab})} \log \lambda_{(\boldsymbol{ab})}}{2^{d_1+d_2}}\right), \tag{3.7}$$

where $C_2(\boldsymbol{s}) = \frac{n!}{\prod_{\boldsymbol{a},\boldsymbol{b}} n_{(\boldsymbol{ab})}!}$ and $h_2(\boldsymbol{\lambda}) = \lambda_{(\mathbf{00})}^{\frac{n}{2^{d_1+d_2}}}$. Therefore, we reparameterize the distribution in (3.4) as a $(2^{d_1+d_2} - 1)$-dimensional exponential family with log-IORs as natural parameters, and the symmetry statistics are *complete sufficient statistics* for log-IORs. This fact is the basis of the binary expansion approach.

Similarly to the BID equations, we have a logarithm version of the BID equation:

Theorem 3.7. Denote the vectors of the logarithm of entries in $\boldsymbol{\lambda}$ and $\boldsymbol{p}$ by $\boldsymbol{\lambda}_l$ and $\boldsymbol{p}_l$, respectively. We have $\boldsymbol{\lambda}_l = \mathbf{H}\boldsymbol{p}_l$.

One important implication of (3.7) and Theorem 3.7 is that all information about dependence is contained in CIOR:

Theorem 3.8. U_{d_1} and V_{d_2} are independent if and only if $\lambda_{(\boldsymbol{ab})} = 1$ for all CIORs.

Theorem 3.8 shows that the null hypothesis of the test (3.2) is equivalent to

$$H_{0,d_1,d_2}: \text{ For all CIORs at depths } d_1 \text{ and } d_2, \lambda_{(\boldsymbol{ab})} = 1. \tag{3.8}$$

We summarize the advantages of the reparameterization in (3.7) and the test (3.8):

(a) Compared to the conventional parameterization in (3.4), the reparameterization in (3.7) is much more interpretable: note that the cell probabilities in $\boldsymbol{p}$ carry both marginal and joint information. On the other hand, the parameterization with $\boldsymbol{\lambda}$ extracts all dependence information in CIORs and separates it from the marginal information in MIORs. Thus, CIORs are to contingency tables as correlations are to multivariate normal distributions. Tests of independence can therefore focus on CIORs, as we study in details in Section 4.

(b) The sufficient statistics in the conventional parameterization are the cell counts $n_{(\boldsymbol{ab})}$s, whose distribution is Binomial$(n, p_{(\boldsymbol{ab})})$. This means that when n is small, one often has $n_{(\boldsymbol{ab})} = 0$ for many cells. These empty cells cause problems in the conventional tests. However, with the reparameterization (3.7), the sufficient statistics $S_{(\boldsymbol{ab})}$s instead have (after a linear transformation) a binomial distribution whose probability of success is the sum of $2^{d_1+d_2} - 1$ cell probabilities. Therefore, by grouping the cells, $S_{(\boldsymbol{ab})}$s provide much more information than $n_{(\boldsymbol{ab})}$s and avoid the well-known problem of insufficient samples in many binning methods.

(c) Note that each cross interaction in the filtration corresponds to a unique CIOR, which measures some form of dependency. In Section 4, we show that this consideration together with the number of CIORs $(2^{d_1} - 1)(2^{d_2} - 1)$ lead to an orthogonal decomposition of the χ^2 test.

(d) The BID equation in Theorem 3.4 can be generalized for any three-way or multiway contingency table whose size is a power of 2. This fact allows extensions of the IOR reparametrization and the BET for testing independence of random vectors.

When the marginal distributions are unknown, for each observation i, we can similarly define $\widehat{\dot{A}}_{k,i} = 2\widehat{A}_{k,i} - 1$, $\widehat{\dot{B}}_{k,i} = 2\widehat{B}_{k,i} - 1$, and $\widehat{S}_{(\boldsymbol{ab})} = \sum_{i=1}^{n} \widehat{\dot{A}}_{\boldsymbol{a},i}\widehat{\dot{B}}_{\boldsymbol{b},i}$ for the cross interaction $\widehat{\dot{A}}_{\boldsymbol{a}}\widehat{\dot{B}}_{\boldsymbol{b}}$. Now note the following simple corollaries from Theorem 3.4: (a) $\boldsymbol{n}_r$ and $\boldsymbol{n}_c$ are invertible functions of $\widehat{S}_{(\boldsymbol{a}\mathbf{0})}$s and $\widehat{S}_{(\mathbf{0}\boldsymbol{b})}$s through a univariate BID equation, and (b) the bivariate sample BID equation holds for $\widehat{\boldsymbol{S}}$ and $\boldsymbol{n}$. With these facts, by using $\boldsymbol{\theta}$ and the proof of Theorem 3.8, as well as conditioning on $\widehat{S}_{(\boldsymbol{a}\mathbf{0})}$ and $\widehat{S}_{(\mathbf{0}\boldsymbol{b})}$ in (3.4), we have

$$p(\widehat{\boldsymbol{S}}_{(\boldsymbol{ab})} = \widehat{\boldsymbol{s}}_{(\boldsymbol{ab})}|\boldsymbol{\lambda}_{(\boldsymbol{ab})}, \widehat{S}_{(\boldsymbol{a}\mathbf{0})}, \widehat{S}_{(\mathbf{0}\boldsymbol{b})}) = C_2(\widehat{\boldsymbol{s}}_{(\boldsymbol{ab})})h_3(\boldsymbol{\lambda}_{(\boldsymbol{ab})}) \exp\left(\sum_{\substack{\boldsymbol{a}\neq\mathbf{0} \\ \boldsymbol{b}\neq\mathbf{0}}} \frac{\widehat{s}_{(\boldsymbol{ab})} \log \lambda_{(\boldsymbol{ab})}}{2^{d_1+d_2}}\right) \tag{3.9}$$

for some function $h_3(\boldsymbol{\lambda}_{(\boldsymbol{ab})})$ as a normalizing constant.

Note that by conditioning on the counts of marginal interactions, the MIORs are eliminated, and we can focus on the CIORs for the analysis of dependence. Indeed, either by comparing (3.5) and (3.9) or by the proof of Theorem 3.8, we see that $\widehat{U}_{d_1,i}$ and $\widehat{V}_{d_2,i}$ are independent for each i if and only if $\lambda_{(\boldsymbol{ab})} = 1$ for all $\boldsymbol{a} \neq \mathbf{0}$ and $\boldsymbol{b} \neq \mathbf{0}$. Therefore, the tests of independence are unified in both of the cases of known and unknown marginal distributions to be (3.8).

We remark here that reparameterization of the contingency table likelihood into odds ratios has been extensively studied in the past Agresti (1992). The very recent paper Ma and Mao (2019) also considered a factorization under the null hypothesis of independence. However, we are not aware of similar ideas of the connection to the Hadamard transform and the concept of IOR. Compared to existing analyses of contingency tables, the new reparameterization is more global to use all the observations. See a detailed discussion in Section 4.4.

We also remark here that we are able to take advantage of the Hadamard transform only because the size of the contingency table is a power of 2, which is a result of $\dot{A}_i$'s and $\dot{B}_j$'s in the binary expansions. If we were to take a different approach or to partition $[0,1]^2$ into different sizes, then we might not be able to have similar theory. This advantage is an important motivation of the binary expansion approach.

4. The Max BET Procedure and Its Properties

4.1. BET as an Multiple Testing Problem

In this section, we return to the dependence detection problem, where we partition $[0,1]^2$ at the binary fractions based on Theorem 3.1. Therefore, the row and column total probabilities in the $2^{d_1} \times 2^{d_2}$ contingency table are 2^{-d_1} and 2^{-d_2}, respectively when the marginal distributions are known, and the row and column total counts in the contingency table are $n2^{-d_1}$ and $n2^{-d_2}$, respectively when the marginal distributions are unknown and when n is a multiple of $2^{\max\{d_1,d_2\}}$.

The discussions in Section 3 suggest test statistics based on interactions $S_{(\boldsymbol{ab})}$ or $\widehat{S}_{(\boldsymbol{ab})}$. Direct application of the MLE of $\lambda_{(\boldsymbol{ab})}$ can result in similar disadvantages as χ^2 tests as we discuss later. We instead construct a simple but optimal test statistic with the maximal symmetry statistics $\max|S_{(\boldsymbol{ab})}|$ or $\max|\widehat{S}_{(\boldsymbol{ab})}|$ for $\boldsymbol{a} \neq \mathbf{0}$ and $\boldsymbol{b} \neq \mathbf{0}$.

The key observations of $S_{(\boldsymbol{ab})}$ are summarized below.

Theorem 4.1. The following are equivalent:

(a) U_{d_1} and V_{d_2} are independent.
(b) $\mathbf{E}[\dot{A}_{\boldsymbol{a}}\dot{B}_{\boldsymbol{b}}] = 0$ for $\boldsymbol{a} \neq \mathbf{0}$ and $\boldsymbol{b} \neq \mathbf{0}$.
(c) $(S_{(\boldsymbol{ab})} + n)/2 \sim \text{Binomial}(n, 1/2)$ for $\boldsymbol{a} \neq \mathbf{0}$ and $\boldsymbol{b} \neq \mathbf{0}$.
(d) $\mathbf{E}[S_{(\boldsymbol{ab})}] = 0$ for $\boldsymbol{a} \neq \mathbf{0}$ and $\boldsymbol{b} \neq \mathbf{0}$.
(e) $\boldsymbol{E} = \boldsymbol{e}_{\mathbf{00}}$ where $\boldsymbol{e}_{\mathbf{00}}$ is the $2^{d_1+d_2}$-dimensional standard basis $(1, 0, \ldots, 0)^T$.

Note here that in Theorem 4.1, the homogeneity in the distribution of $S_{(\boldsymbol{ab})}$ is due to the symmetry in $\dot{A}_{\boldsymbol{a}}$ and $\dot{B}_{\boldsymbol{b}}$ in the binary expansion. Indeed, the main intuition of Theorem 4.1 is the symmetry of independence: when U_{d_1} and V_{d_2} are independent, the counts of observations in the positive and negative regions should be similar for any cross interaction. On the other hand, when U_{d_1} and V_{d_2} are not independent, we expect some strong asymmetry between the numbers of points in white or blue.

When the marginal distributions are unknown, we have similar results on symmetry assuming n is a multiple of $2^{\max\{d_1,d_2\}}$. When $\widehat{U}_{d_1,i}$ and $\widehat{V}_{d_2,i}$ are independent for each $i = 1, \ldots, n$, the distribution of $(\widehat{S}_{(\boldsymbol{ab})} + n)/4$ is Hypergeometric$(n, n/2, n/2)$. An intuitive way to understand this is that if we assign all n observations into a 2×2 table according to $\widehat{\dot{A}}_{\boldsymbol{a},i} = \pm 1$ and $\widehat{\dot{B}}_{\boldsymbol{b},i} = \pm 1$, $\widehat{S}_{(\boldsymbol{ab})}$ is the difference in counts of the interaction $\widehat{\dot{A}}_{\boldsymbol{a},i}\widehat{\dot{B}}_{\boldsymbol{b},i}$ being $+1$ or -1. We show below that the converse is also true.

Theorem 4.2. When n is a multiple of $2^{\max\{d_1,d_2\}}$, the following are equivalent:

(a) For each i, $\widehat{U}_{d_1,i}$ and $\widehat{V}_{d_2,i}$ are independent.
(b) $(\widehat{S}_{(\boldsymbol{ab})} + n)/4 \sim \textit{Hypergeometric}(n, n/2, n/2)$ for $\boldsymbol{a} \neq \mathbf{0}$ and $\boldsymbol{b} \neq \mathbf{0}$.
(c) $\mathbf{E}[\widehat{S}_{(\boldsymbol{ab})}] = 0$ for $\boldsymbol{a} \neq \mathbf{0}$ and $\boldsymbol{b} \neq \mathbf{0}$.

Theorems 4.1 and 4.2 reduce the test of independence to tests of marginal properties of $\mathbf{E}[S_{(\boldsymbol{ab})}]$ and $\mathbf{E}[\widehat{S}_{(\boldsymbol{ab})}]$. In particular, these results show the equivalence between the BET at depths d_1 and d_2 and a multiple testing problem: the testing problems in (3.2) and (3.3) are equivalent to testing if all cross interactions up to depths d_1 and d_2 are symmetric. The advantage of this consideration is two-folded: (a) we reduce the test of a joint distribution (difficult) to that of marginal ones (simple) and (b) we reduce the test of dependence (difficult) to that of symmetry (simple).

Note that the equivalent multiple testing problem is about controlling the family-wise error rate (FWER): rejecting any symmetry results in the rejection of independence. The simplest FWER control is the Bonferroni procedure, where the adjusted p-value is the minimum of 1 and the product of $(2^{d_1}-1)(2^{d_2}-1)$ and the smallest p-value of all marginal tests. We refer this procedure as the Max BET.

We illustrate the Max BET procedure at depths $d_1 = 2$ and $d_2 = 1$ with the 64 samples studied in Section 3.3. The procedure consists of the following steps, as shown in Figure 2:

Step 1: We count white and blue points for each cross interaction $\dot{A}_2\dot{B}_1$, $\dot{A}_1\dot{B}_1$, and $\dot{A}_1\dot{A}_2\dot{B}_1$ for $d_1 = 2$ and $d_2 = 1$.
Step 2: Among these three cross interactions, we look for the one with the strongest asymmetry, which is $\dot{A}_2\dot{B}_1$ with 25 in white and 39 in blue. The symmetry statistic is $S_{(011)} = -14$. The binomial p-value is 0.103.
Step 3: Use the Bonferroni adjustment to multiply 3 and get the overall p-value of the Max BET at depths $d_1 = 2$ and $d_2 = 1$ to be 0.310.

Would the Bonferroni procedure be overly conservative? Our observation is no because of the orthogonality of the symmetry statistics. A formal study of optimality of the Bonferroni procedure is in Section 4.2. Here, we state some results on the joint properties of symmetry statistics which provide some intuition.

Theorem 4.3.

(a) When the marginal distributions are known and U_{d_1} and V_{d_2} are independent, the symmetry statistics $S_{(\boldsymbol{ab})}$s are pairwise independent.
(b) When the marginal distributions are unknown and for each i, $\widehat{U}_{d_1,i}$ and $\widehat{V}_{d_2,i}$ are independent, $\widehat{S}_{(\boldsymbol{ab})}$s are uncorrelated.
(c) The classical χ^2 test statistic C is $C = \frac{1}{n}\sum_{\boldsymbol{a}\neq\mathbf{0},\boldsymbol{b}\neq\mathbf{0}} \widehat{S}^2_{(\boldsymbol{ab})}$.

Part (a) and (b) of Theorem 4.3 imply that due to the orthogonality in the BID, each symmetry statistic provides nonredundant information. Furthermore, part (b) and (c) of Theorem 4.3 imply that the $(2^{d_1}-1)(2^{d_2}-1)$ sample symmetry statistics $\widehat{S}_{(\boldsymbol{ab})}$s form an orthogonal decomposition of the χ^2 test statistic whose degrees of freedom is also $(2^{d_1}-1)(2^{d_2}-1)$. Therefore, instead of aggregating the information through sum of squares in the χ^2 statistic, we here take a divide-and-conquer approach. To follow up the discussions in Section 3.2, we summarize the advantages of our approach below and describe the details in Sections 4.2 and 4.3.

(a) In Arias-Castro, Candès, and Plan (2011) and Barnett, Mukherjee, and Lin (2017), it was noted that when the number of hypotheses is large and the signals are rare and weak, using a Bonferroni type of multiple comparison control can substantially outperform χ^2 tests. In our context, this means that when d_1 and d_2 are large and when the dependence is through only a few cross interactions, the χ^2 test is "wasting" many degrees of freedom. Instead, using the Max BET can help discover weaker dependence.

(b) Interpretability. One major advantage of using cross interactions over the χ^2 test is that the grouping arrangement of the white and blue cells for each interaction helps indicate the pattern of the dependence, as described earlier in Section 3.3. When the dependence is through only a few of cross interactions, with the rejection of the Max BET, we can identify the strongest interactions between the variables. These strongest interactions can in turn help describe the dependence.

4.2. *Power and Optimality of the Max BET*

In this section, we study the power of the Max BET when the marginal distributions are known. The uniform consistency of the Max BET at any depths d_1 and d_2 follows from classical analysis of contingency tables. Moreover, despite the conservative nature of the Bonferroni approach, we show below that the Max BET can be optimal in power for a large collection of alternative distributions:

Theorem 4.4. For any fixed $0 < \delta < 1/2$, denote by $\mathcal{H}^R_{1,d_1,d_2}$ the collection of alternative distributions $\mathbf{P}_{(U_{d_1},V_{d_2})}$ such that

1. $\mathrm{TV}(\mathbf{P}_{(U_{d_1},V_{d_2})}, \mathbf{P}_{0,d_1,d_2}) \geq \delta$;
2. $\|\boldsymbol{E} - \boldsymbol{e}_{(\mathbf{00})}\|_\infty \geq \sqrt{d_1+d_2}\,2^{-(d_1+d_2)/4}\|\boldsymbol{E} - \boldsymbol{e}_{(\mathbf{00})}\|_2$.

Consider the testing problem

$$H_{0,d_1,d_2} : \mathbf{P}_{(U_{d_1},V_{d_2})} = \mathbf{P}_{0,d_1,d_2} \;\; v.s. \;\; H_1 : \mathbf{P}_{(U_{d_1},V_{d_2})} \in \mathcal{H}^R_{1,d_1,d_2}. \tag{4.1}$$

For large d_1 and d_2, we have the following:

1. For any $\epsilon > 0$, the Max BET with size α needs $n = O(2^{(d_1+d_2)/2}/\delta^2)$ samples to have power $1-\epsilon$.
2. Let $\mathcal{T}_\alpha$ be the collection of all measurable size-α tests: $\mathcal{T}_\alpha = \{T_\alpha : \mathbf{P}_{0,d_1,d_2}(T_\alpha = 1) \leq \alpha\}$. If $n = o(2^{(d_1+d_2)/2}/\delta^2)$, then there $\exists 0 < \epsilon' < 1-\alpha$ such that

$$\inf_{T_\alpha \in \mathcal{T}_\alpha} \; \sup_{\mathbf{P}_{(U_{d_1},V_{d_2})} \in \mathcal{H}^R_{1,d_1,d_2}} \mathbf{P}_{(U_{d_1},V_{d_2})}(T_\alpha = 0) \geq 1 - \alpha - \epsilon'. \tag{4.2}$$

The magnitude of the minimal sample size requirement has been carefully studied in statistics, information theory, and machine learning. It describes the minimal number of samples to uniformly detect certain departure from the independence and in turn indicates the uniform power of the test. Part 1 of Theorem 4.4 states that such a requirement for Max BET is $O(2^{(d_1+d_2)/2}/\delta^2)$, which matches the optimal rate in Paninski (2008) and Acharya, Daskalakis, and Kamath (2015). Moreover, part 2 of Theorem 4.4 asserts that if the sample size grows at any smaller rate, then for any test, there exist alternatives such that the power of this test is strictly bounded away from 1. In this sense, the Max BET is minimax in the sample size requirement.

Note that the consistency of χ^2 tests is shown in Agresti and Kateri (2011) and Fienberg (2007) to require $n > 2^{d_1+d_2}$. This requirement is much higher than the magnitude $O(2^{(d_1+d_2)/2}/\delta^2)$ in Theorem 4.4 and indicates that the power of χ^2 test can be much less than that of the Max BET. One intuitive explanation of this fact is that χ^2 tests rely on good estimates of each cell probability in the table, while in the Max BET $S_{(\boldsymbol{ab})}$s are based on grouped cells to use all n observations.

The condition $\|\boldsymbol{E} - \boldsymbol{e}_{(\mathbf{00})}\|_\infty \geq \sqrt{d_1+d_2}\,2^{-(d_1+d_2)/4}\|\boldsymbol{E} - \boldsymbol{e}_{(\mathbf{00})}\|_2$ compares the strongest signal to the overall signal in the space of alternatives and indicates the signals to take on a spiky form. It can also be regarded as (but is more general than) a sparsity constraint, as it can be satisfied when at most $\frac{1}{d_1+d_2}2^{(d_1+d_2)/2}$ (out of $(2^{d_1}-1)(2^{d_2}-1)$) cross-interactions have nonzero means. Under this generalized form of sparsity, the Bonferroni approach is not overly conservative. In particular, Theorem 4.4 is consistent with the results in Arias-Castro, Candès, and Plan (2011) under the ANOVA setting that when the signals are square-root sparse, the max test has better power than the χ^2 test. Note also that such a condition over $\boldsymbol{E}$ does not imply sparsity in $\boldsymbol{p}$. Therefore, the optimal rate in Paninski (2008) and Acharya, Daskalakis, and Kamath (2015) still applies and is attained by the Max BET.

The sample size requirement in Theorem 4.4 also indicates that for a given sample size n, one can expect to detect dependence up to a depth of about $\log_2 n$. This result again explains the problem of nonuniform consistency: one cannot expect one test to uniformly detect all types of dependency, and with n samples one can only reliably detect dependence up to a depth of about $\log_2 n$ in the binary expansion filtration approximation. Note again that with the χ^2 test the depth can only go up to about $\frac{1}{2}\log_2 n$, which means it may not have good power for many forms of dependency.

4.3. *Interpretation of the Max BET*

In this section, we explain the interpretations of the BET, that is, we ask when the BET at depths d_1 and d_2 is rejected, where is the dependence? The BET can explain this question explicitly with the cross interactions, because it returns with the 50% area with significantly more points.

We will explain some common patterns of dependence in simulation studies in Section 6. We will also illustrate the interpretation of BET with real data in Sections 7 and 8. In what follows, we revisit the BEX as an example. See Figure 3. Note that with probability 1, samples of (X_d, Y_d) on BEX_d all fall in the positive region for $\dot{A}_d\dot{A}_{d+1}\dot{B}_d\dot{B}_{d+1}$. This is the strongest asymmetry of BEX_d, and the p-value for the Max BET at $d_1 = d_2 = d+1$ is $2(2^{d+1}-1)^2/2^n$ which can be very small when n is much larger than $2d$. Note that with the rejection of the Max BET at $d_1 = d_2 = d+1$, the cross interaction $\dot{A}_d\dot{A}_{d+1}\dot{B}_d\dot{B}_{d+1}$ is also found to present the dependency between X_d and Y_d.

With the above considerations, we explain the paradox following Proposition 2.1. For (X_d, Y_d) on BEX_d, let U_d and V_d be the truncated variables in the marginal binary expansion of X_d and Y_d, respectively. Note that U_d and V_d are independent. However, U_{d+1} and V_{d+1} are dependent, as is evidenced by the

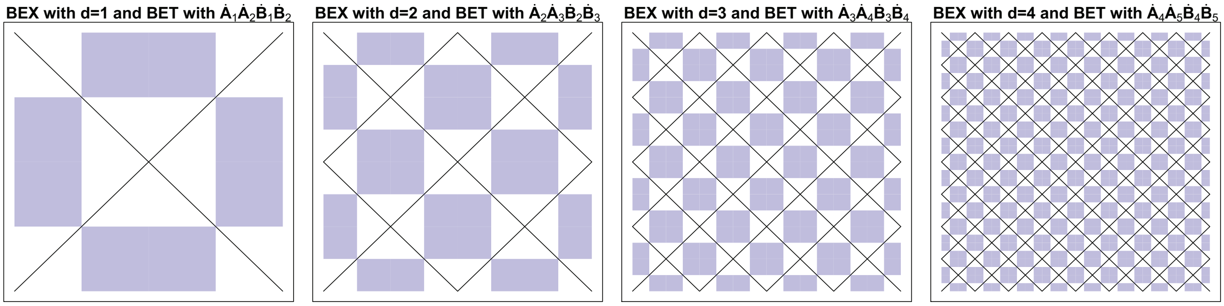

BEX with d=1 and BET with $\dot{A}_1\dot{A}_2\dot{B}_1\dot{B}_2$
BEX with d=2 and BET with $\dot{A}_2\dot{A}_3\dot{B}_2\dot{B}_3$
BEX with d=3 and BET with $\dot{A}_3\dot{A}_4\dot{B}_3\dot{B}_4$
BEX with d=4 and BET with $\dot{A}_4\dot{A}_5\dot{B}_4\dot{B}_5$

Figure 3. The bisection expanding cross (BEX) at $d = 1, \dots, 4$ captured in the positive regions of the BET, which illustrates the interpretation of dependency in the BET.

small p-value. These facts thus explain the seeming paradox: if we are at depths $d_1 = d_2 = d$, then the fact that U_d and V_d are independent implies that X_d and Y_d are (d, d)-independent, that is, nearly independent. On the other hand, if we are at depths $d_1 = d_2 = d + 1$, then the small p-value of the BET implies that X_d and Y_d are strongly nonindependent. Therefore, being strongly nonindependent or nearly independent depends on the choice of depth, and there is no contradiction in this example.

4.4. Relations to Other Binning Methods

Although the binary expansion approach leads to multiscale discretization, the BET is different from existing tests in the binning approach in several ways: (a) many existing binning methods such as Reshef et al. (2011) and Kinney and Atwal (2014) involve an optimization step in search of the optimal partition of data under some criteria such as mutual information. This step could be computationally expensive due to a search over many overlapping partitions which contain redundant information. Instead, the partitions based on interactions from the binary expansion filtration are created in a systematic manner with a natural hierarchy. The orthogonal design of interactions also saves much redundant information and improves the power, (b) many binning tests may have problems of insufficient observations in small bins, while in the BET all n samples are used repeatedly in an orthogonal manner which has advantages both for the level and power, (c) many binning tests return a p-value based on permutations, which can again be computationally more expensive than the BET.

We also compare the Max BET with recent work in scan statistics (Walther 2010; Ma and Mao 2019) which are based on rectangle scanning windows for local dependency. We note that some scanning method can be formulated in terms of the BEStat. For example, the FES in Ma and Mao (2019) up to $(2, 1)$-independence can be regarded as the following three tests of symmetry: $\mathbf{E}[\dot{A}_1\dot{B}_1] = 0, \mathbf{E}[\dot{A}_2\dot{B}_1|\dot{A}_1 = 1] = 0$, and $\mathbf{E}[\dot{A}_2\dot{B}_1|\dot{A}_1 = -1] = 0$. Compared to the three tests of symmetry in the Max BET $\mathbf{E}[\dot{A}_1\dot{B}_1] = 0$, $\mathbf{E}[\dot{A}_2\dot{B}_1] = 0$ and $\mathbf{E}[\dot{A}_1\dot{A}_2\dot{B}_1] = 0$, FES can be regarded as a conditional version of the BET. This conditional formulation can be advantageous in detecting local dependency, but may not have optimal power when the dependency is global and may have the insufficient sample problem discussed above. In the Max BET, the grouping of positive and negative regions does not necessarily result in a region of the rectangle shape but is more capable of detecting global dependency. Thus, each method has its advantageous scenarios.

4.5. Issues in Practice

In this section, we discuss issues of the Max BET that can happen in practice. The first issue is that we often do not know correct depths d_1 and d_2 where the dependency may be present. To address this issue, we propose a search over different depths and a second stage multiplicity control. This proposal is based on the observation that the approximation error in (3.1) is $O_p(2^{-\min\{d_1,d_2\}})$. Therefore, we can first test the hypotheses (3.8) for $d_1 = d_2 = d$ with $d = 1, \dots, d_{\max}$, where $d_{\max}$ reflects the desirable accuracy in the approximation. Then we can apply some further FWER multiplicity control procedure such as the Bonferroni method over the $d_{\max}$ tests to ensure the overall FWER.

In practice, note that from (3.1) $d_{\max} = 4$ provides good approximation to the true distribution. Note also that to avoid overlapping cross interactions in different depths, for each $d \geq 2$, one can test the symmetry of all added interactions involving $\dot{A}_d$ or $\dot{B}_d$, which are in $\sigma(U_d, V_d)$ but not in $\sigma(U_{d-1}, V_{d-1})$. We illustrate this procedure in Sections 6 and 7. The effect of such multiplicity control on power is studied in Section 1.2 of the supplementary materials.

Another practical issue for the empirical BET is that n might not be a multiple of $2^{\max\{d_1,d_2\}}$, that is, the column and row total counts might not be equal in the $2^{d_1} \times 2^{d_2}$ table. In this case, the reparameterization in Section 3.3 still applies, and the test for each cross interaction is still a Fisher's exact test for 2×2 tables. However, the distribution of a symmetry statistic (after a linear transformation) is not necessarily Hypergeometric$(n, n/2, n/2)$. In general, instead of $n/2$s, the parameters for the hypergeometric distribution are numbers of observations for which the marginal interactions are positive. Thus, symmetry and homogeneity might be lost in this case. Nonetheless, the BET still applies for any sample size $n \geq 2^{\max\{d_1,d_2\}}$ (otherwise there exist cross interactions for which all observations are positive). Moreover, when n is large, one can use the normal approximation in Kou and Ying (1996) for these tests.

5. Connection to Computing

The binary expansion approach is partially motivated by its close connections to the current computing system, which is based on a binary architecture. By turning an electrical circuit "on" (represented by "1") and "off" (represented by "0"), computers process information with unprecedented speed and power. In particular, each decimal number in computing is processed as a rounded version of its binary representation. For example, calculations of $0.1_{10} = 0.000110011\ldots_2$ are based on a rounded version of $0.000110011\ldots_2$ to certain bits (depending on a 32-bit or 64-bit computing system).

The key observation here is that *the binary representation of a decimal number is precisely its binary expansion!* The $\{A_k\}_{k=1}^{d_1}$ and $\{B_k\}_{k=1}^{d_2}$ in the BEStat approach directly correspond to the first d_1 and d_2 bits of U and V, respectively in current computing systems. This fact implies that as long as a statistician is processing data with a computing device (desktop, laptop, smartphone, hand-held calculator, etc.), the $\{A_k\}_{k=1}^{d_1}$ and $\{B_k\}_{k=1}^{d_2}$ are given to him/her automatically. These binary bits are hidden resources of data available for statisticians from computers. We often use bits for computing, but *bits are data!* We can construct statistics and make inference with bits, and the BET at depths d_1 and d_2 can be explicitly interpreted as testing whether the data are independent up to the first d_1 and d_2 bits.

Moreover, the BEStat approach provides statisticians the access to the most fundamental level of the computing system and enables direct operations over bits. For example, the cell locating process of a data point in the contingency table can be done through some bitwise Boolean operations over the a_ks and b_ks. Such bitwise operations are known to be computationally efficient. We develop such a bitwise algorithm of the BET in a separate paper (Zhao et al. 2019), where the procedure is shown to improve the speed of existing methods by orders of magnitude.

6. Simulation Studies

In this section, we use simulation studies to compare the Max BET and existing nonparametric methods. For the Max BET, we consider the empirical CDF transformation and consider the second stage multiplicity control over depths with the Bonferroni procedure with $d_{\max} = 4$, as discussed in Section 4.5. For comparison, we consider the Hoeffding's D test from the CDF approach, the distance correlation from the distance approach, the default KNN-MI method from the binning approach, and the very recent method of FES. We consider the χ^2 test for the same contingency table for the Max BET with $d_1 = d_2 = 4$ too.

We compare the power the above methods over common dependency structures such as linear, parabolic, circular, sine, and checkerboard, which are widely considered in evaluation of tests of independence (Reshef et al. 2011; Heller, Heller, and Gorfine 2012; Kinney and Atwal 2014; Filippi and Holmes 2015). We also consider the local dependency setting in Ma and Mao (2019). The scenarios are designed by adapting those in Ma and Mao (2019) with an emphasis on small sample performance with a fixed sample size 128. The level of the tests is set to be 0.1.

Table 1. Simulation scenarios: at each noise level $l = 1, \ldots, 10$, $\epsilon, \epsilon', \epsilon'' \overset{\text{iid}}{\sim} \mathcal{N}(0, (l/40)^2)$, and the following variables are all independent: $U \sim \text{Uniform}[0, 1]$, $\vartheta \sim \text{Uniform}[-\pi, \pi]$, $W \sim \text{Multi} - \text{Bern}(\{1, 2, 3\}, (1/3, 1/3, 1/3))$, $V_1 \sim \text{Bern}(\{2, 4\}, (1/2, 1/2))$, $V_2 \sim \text{Multi} - \text{Bern}(\{1, 3, 5\}, (1/3, 1/3, 1/3))$, $G_1, G_2 \overset{\text{iid}}{\sim} \mathcal{N}(0, 1/4)$.

Scenario	Generation of X	Generation of Y
Linear	$X = U$	$Y = X + 6\epsilon$
Parabolic	$X = U$	$Y = (X - 0.5)^2 + 1.5\epsilon$
Circular	$X = \cos\vartheta$ $+2.5\epsilon$	$Y = \sin\vartheta + 2.5\epsilon'$
Sine	$X = U$	$Y = \sin(4\pi X) + 8\epsilon$
Checkerboard	$X = W + \epsilon$	$Y = \begin{cases} V_1 + 4\epsilon' & \text{if } W = 2 \\ V_2 + 4\epsilon'' & \text{otherwise} \end{cases}$
Local	$X = G_1$	$Y = \begin{cases} X + \epsilon & \text{if } 0 \le G_1 \le 1 \text{ and } 0 \le G_2 \le 1 \\ G_2 & \text{otherwise} \end{cases}$

We simulate each of the scenarios at 10 different noise levels to present the whole range of power. The details of the setting are summarized in Table 1.

The power curves of the six nonparametric tests of independence are presented in Figure 4. Generally speaking, as is found similarly in Ma and Mao (2019) and many other papers, no test can uniformly dominate all others in all settings. In what follows, we separate the detailed discussions of the first five scenarios (linear, parabolic, circular, sine, and checkerboard) and the last scenario (local).

In the first five scenarios where the dependency is global, we notice that each existing method has shown some limitations: in the linear and parabolic setting, the χ^2 test provides the least power. In the circular setting, distance correlation provides the least power. In the sine setting, KNN-MI provides the least power. In the checkerboard setting, Hoeffding's D and FES provide the least power, which is partially due to the fact that observations in this setting are locally independent. On the other hand, the BET never provides the least power under these common relationships. One reason of such robustness of the BET is that the global dependency in these settings can be well explained through only a few cross interactions in the binary expansion, as can be seen in Figure 5 and in discussions below. Therefore, the minimaxity in Theorem 4.4 guarantees that the BET has reliable power against a large class of alternative distributions. We also note here that to echo with the discussions in Section 4.4, the BET has better power than FES in most of these global dependency settings because of its global grouping of cells. On the other hand, FES has better performance in the local dependency setting, as we discuss below.

We now turn to the setting of the local relationship. The BET does not perform well because observations in this setting are independent outside the area with the local dependency. Therefore, the global grouping of cells in the BET does not provide more information than a few local cells. In this case, the condition in Theorem 4.4 can be violated as many cross interactions are asymmetric with weak signals. As shown in Figure 4, this limitation of the BET can be remedied by scanning based binning methods such as FES, which focuses on local dependency, or clustering based binning methods such as KNN-MI, which performs well on mixtures of distributions.

One useful property of the BET is its interpretability of dependency based on the interactions of binary variables, which we illustrate in Figure 5. In each column, we present a simulated dataset in each scenario with noise level $l = 2$. In the first five scenarios, the global dependency in the data is well explained by a corresponding cross interaction: observations with linear dependency tend to fall in the positive region of $\widehat{\dot{A}}_1\widehat{\dot{B}}_1$, observations with the parabolic dependency tend to fall in the positive region of $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1$, observations with circular dependency tend to fall in the negative region of $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1\widehat{\dot{B}}_2$, observations with the sine dependency tend to fall in the negative region of $\widehat{\dot{A}}_2\widehat{\dot{B}}_1$, observations with the checkerboard dependency tend to fall in the positive region of $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1\widehat{\dot{B}}_2$. Since these common global dependency patterns can be well explained by a single cross interaction, Theorem 4.4 applies and the Max BET has good performance in terms of power as shown in Figure 4.

The local dependency in the last scenario is also well captured by the positive region of $\widehat{\dot{A}}_2\widehat{\dot{B}}_2$, particularly in the four upper right cells. However, outside this region the variables are independent, so the interpretation of dependency is rather explained by a local and conditional cross interaction $\widehat{\dot{A}}_2\widehat{\dot{B}}_2$ given $\{\widehat{\dot{A}}_1 = 1, \widehat{\dot{B}}_1 = 1\}$, than by the global cross interaction $\widehat{\dot{A}}_2\widehat{\dot{B}}_2$. In this case, scanning based binning methods such as FES provide better interpretation of the local dependency.

7. Are Stars Randomly Distributed in the Sky?

In this section, we study the curious question of whether stars in the night sky are randomly distributed. Despite a simple statement of this long standing question, we are not aware of any complete scientific theory that explains the phenomenon with a confirming or disconfirming answer. In what follows, we provide some statistical analysis of this problem.

To study this question, we collected the galactic coordinates of the 256 brightest stars in the night sky (Perryman et al. 1997). The galactic coordinates are essentially spherical coordinates with the Sun as the center. These coordinates consist of radius, longitude $\phi \in [0, 2\pi)$, and latitude $\varphi \in (-\pi/2, \pi/2]$. We ignore the radius information and focus on the unit sphere. Since the density of the uniform distribution over the unit sphere is proportional to $\cos\varphi d\phi d\varphi$, as long as $X = \phi$ and $Y = \sin\varphi$ of the stars are independent, the stars are uniformly distributed in the night sky.

We first consider some classical tests of independence. The sample correlation between X and Y is -0.07 with a p-value of 0.264, which is not significant. The distance correlation between X and Y is 0.137 with a p-value of 0.064. Hoeffding's D test returns with a p-value of 0.103. These p-values indicate some evidence against independence. The KNN-MI test provides a p-value of 0.02, which is strong evidence against independence. However, this p-value does not provide any information about the relationship between X and Y, and the dependence pattern is still unclear even when we rejected the null.

We now consider applying the two-stage empirical Max BET with $d_{\max} = 4$ on these data. The BET returns the strongest

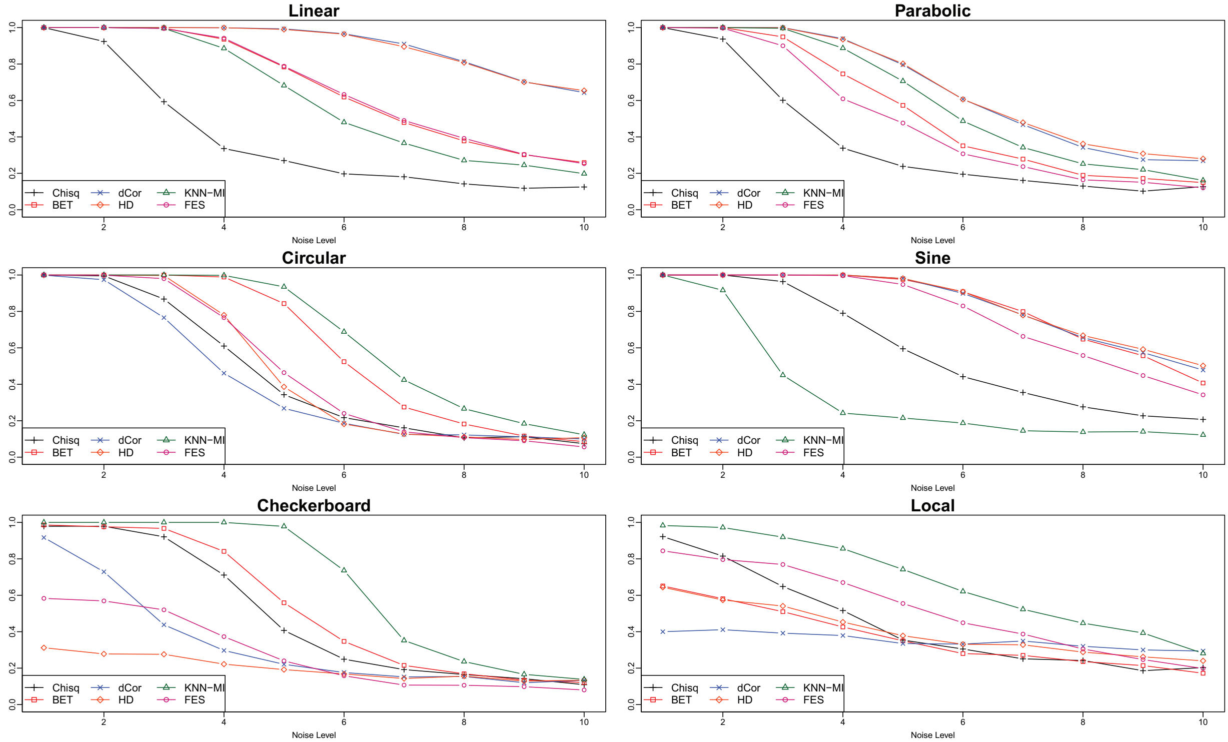


Figure 4. Comparison of powers from six nonparametric tests of independence: the two-stage Max BET with empirical CDF and with $d_{\max} = 4$ (BET), χ^2 test for the discretization when $d_1 = d_2 = 4$ (Chisq), distance correlation (dCor), Hoeffding's D (HD), k-nearest neighbor mutual information (KNN-MI), and Fisher exact scanning (FES).

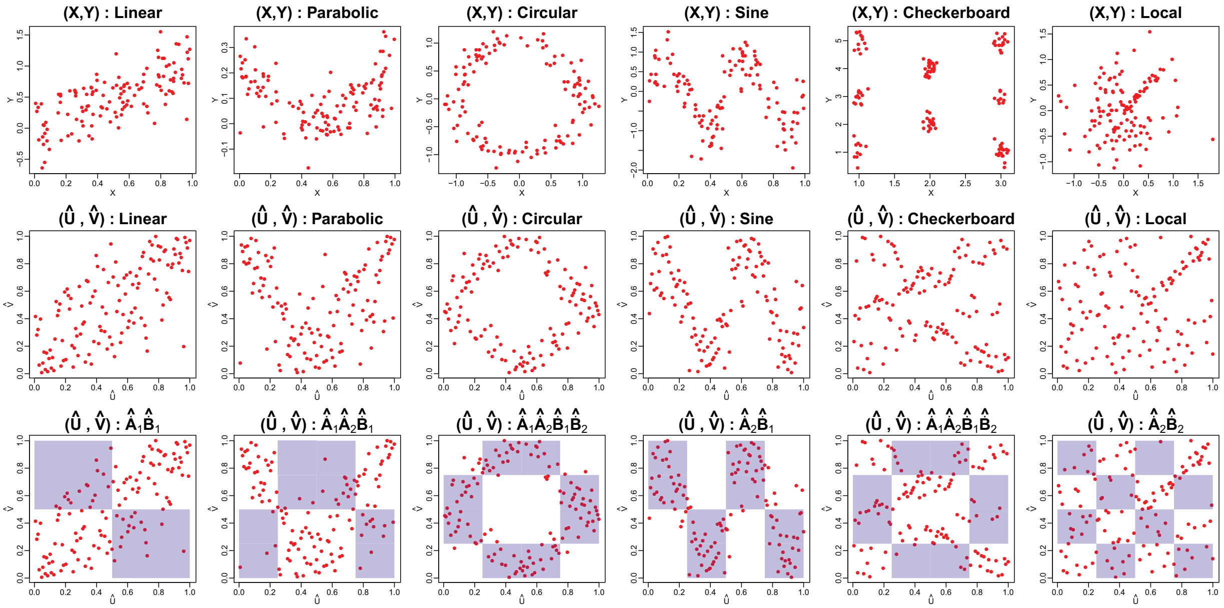


Figure 5. The BET interpretations of dependency patterns. The observations are generated as in Table 1 with noise level $l = 2$. The first row shows the scatterplots of original data (X, Y). The second row shows the corresponding empirical copula $(\widehat{U}, \widehat{V})$ for $i = 1, \ldots, 128$. The third row shows the cross interaction of the strongest asymmetry, which the BET returns with the rejection of independence null.

asymmetry $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1$, where 156 stars are in the positive region and 100 are in the negative region. Thus, $\widehat{S}_{(111)} = 56$ and the approximate z-statistics is 3.5 with the overall p-value 0.019. Besides the strong evidence against independence, one important advantage of the BET is that we can also visualize the dependency upon rejection. In part (c) of Figure 6, we transform the interaction in part (b) back to the original scale. Note that the labeled stars are well-known to be along the Milky Way in the night sky. Indeed, the Milky Way in the night sky is where stars in the galaxy cluster together, and its shape is captured by the positive region of $\dot{A}_1\dot{A}_2\dot{B}_1$. This fact explains the dependency in this data and the significance of the BET.

We note here that the application of FES to the star data returns with a p-value of 0.032 with the strongest local dependency in $\widehat{\dot{A}}_2\widehat{\dot{B}}_1$ given $\{\widehat{\dot{A}}_1 = 1\}$. Compared with the BET which uses all 256 observations to detect the dependency in $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1$, the p-value of FES is higher because it only uses 128 observations in the detection of local dependency when $\{\widehat{\dot{A}}_1 = 1\}$. In terms of interpretation, the FES only explains the dependency in the data with the "right arm" of the milky way, whereas the BET captures the entire milky way with an global cross interaction $\widehat{\dot{A}}_1\widehat{\dot{A}}_2$ $\widehat{\dot{B}}_1$.

A caveat here is that we regard the above analysis more as an illustration of the BET method rather than a scientific discovery, which requires a much more careful study. For example, the only strong assumption in the BET approach is the iid assumption on the observations. This assumption might be violated when the data points are stars. Moreover, we also note that the radius, which is excluded from this study, plays an important role in the location of stars. However, the interpretations from the BET can still be of immediate practical value: For example, it can help people find bright stars in the night sky.

8. Exploratory Data Analysis of TCGA Data

8.1. *Nonlinearity and Mixture of Subtype Distributions*

Conventional exploratory data analysis (EDA) of small multivariate datasets usually starts with a scatterplot matrix, see Buja and Tukey (1992) and Cleveland (1993) for good reviews. Pairwise scatterplots can help people find interesting dependency patterns among variables, which can in turn suggest further statistical or scientific investigation. However, for high-dimensional data, the scatterplot matrix is not feasible since there are too many pairwise plots to inspect (Sun and Zhao 2014). Common EDA tools in this situation such as principal component analysis, can only show high-level structure in the data, and focus mainly on linear relationships of variables. The BET can provide an alternative approach for such EDA due to the interpretability of its p-value. We illustrate this idea below in the context of breast cancer classification.

The TCGA lobular freeze breast cancer data in Cancer Genome Atlas Network (2012) and Ciriello et al. (2015) contain gene expression intensities of 817 subjects, about 2/3 of which, or 544 samples, are used here as a training set and the remaining 273 observations are used as a test set. This dataset is based on 16,615 genes. There are five subtypes groups indicated in this dataset. In what follows, we focus on basal-like breast cancer, which is known to be more aggressive, more difficult to treat, and have poorer prognosis compared to the other subtypes (Perou et al. 2000). Accurate classification of this subtype of breast cancer is thus very important for the health quality of patients.

The goal of this analysis is to use the BET as an EDA tool in the training dataset in search for nonlinear dependency between pairs of genes. Once a pair is identified in the EDA phase we look in the literature for mentions of the two corresponding genes

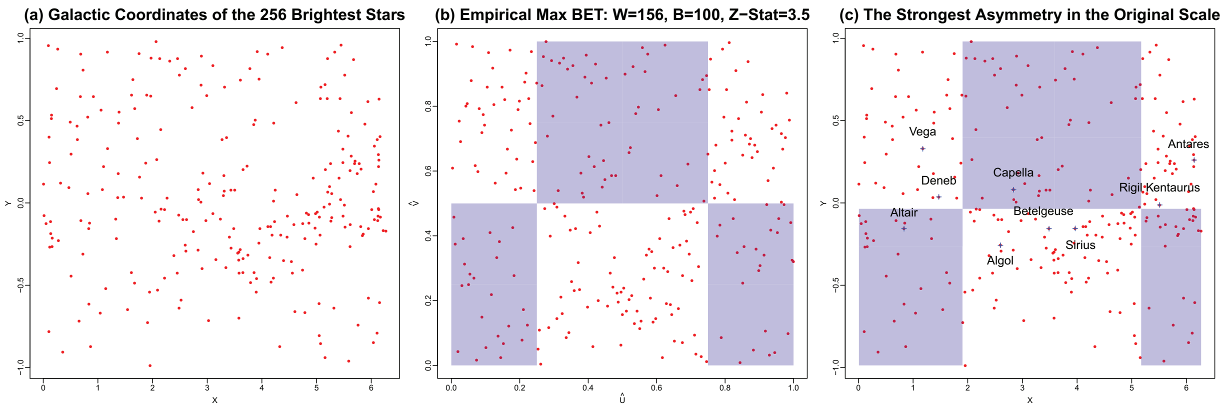


Figure 6. (a) The longitude and sine latitude of the 256 brightest stars in the night sky. (b) The strongest asymmetry for the BET at $d = 2$ is found to be the interaction $\widehat{\dot{A}}_1\widehat{\dot{B}}_1\widehat{\dot{A}}_2$. (c) The strongest asymmetry in the original scale and some famous stars along the Milky Way.

and study their connection to subgroup typing. We also use the test dataset for confirmatory analysis.

Why can nonlinear dependency be related to studies on subgroup typing? As we illustrate below, one source of nonlinearity could be mixture of different subtype distributions. Intuitively, some genes might have different joint behavior under different subtypes of cancer. Such distributional differences could be in location, scale, covariance, and other moments. When these different bivariate distributions are mixed together, some nonlinear dependency pattern could be created in the pooled joint distribution. Since the BET can capture nonlinear dependency patterns and indicate the form of nonlinearity, once a pair is identified by the BET, we hope to track back with the label information to find interesting pairs of genes that are related to different subtypes of breast cancer.

We first prepare the data by excluding genes which had nonunique entries in intensities. Such ties are results of the thresholding step in the preprocessing, and we exclude these genes here for simplicity. This filtering step results in 10,107 genes in the remaining data. In the EDA phase with the training dataset, we scan over all pairs of these 10,107 genes with the BET based on the empirical CDF transformation and depths $d_1 = d_2 = 2$, and the p-value are calculated based on the large sample normal approximation of hypergeometric distribution in Kou and Ying (1996). This approach leads to a total of $\binom{10,107}{2} = 51,070,671 \approx 5 \times 10^7$ comparisons. We control the multiplicity over these comparisons through the Bonferroni method. We use the level 0.1 threshold for multiplicity adjusted p-values to determine whether a pairing is interesting enough to follow up in the literature.

We emphasize here that many existing nonparametric dependence detection methods, such as Hoeffding's D, distance correlation, KNN-MI, and FES, are not suitable for this EDA task for the following reasons:

(a) Classical methods such as Hoeffding's D, distance correlation, and KNN-MI do not provide clear interpretation upon rejection of independence. For example, even if the tests based on them are significant, they cannot distinguish pairs of genes with nonlinear dependency from pairs of genes with linear dependency.

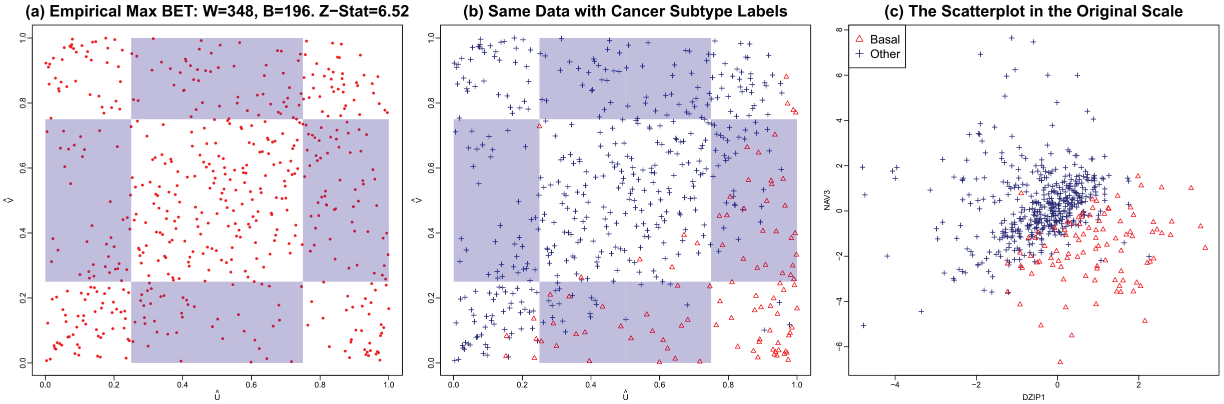
(b) Although mutual information based methods such as KNN-MI have good power against mixtures of distributions, the p-value of KNN-MI is obtained through permutations. With the Bonferroni control over 5×10^7 pairwise tests, we need at least 5×10^8 random permutations for each test to have a valid significance level of 0.1. The computational expense is prohibitive.

(c) Although FES provides interpretation of local dependency, it does not allow users to specify a global form of dependency in search of interesting relationships between variables. Thus, it cannot identify pairs of genes with global nonlinear dependency. See the discussions below.

8.2. Results From TCGA Data

In the EDA phase, the BET rejects independence over more than 10,000 pairings of genes out of 5×10^7. Out of these pairs of genes, we can focus on some particular form of dependency. For example, we can restrict on pairs of genes whose dependency can be explained by the cross interaction $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1\widehat{\dot{B}}_2$. This consideration results in only 84 pairs of genes. Note that this specification process of global dependency is not possible with FES, nor other existing methods. Of those 84 pairs of $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1\widehat{\dot{B}}_2$ dependency, we focus on DZIP1 and NAV3. For this pair of genes, there are 348 observations and 196 observations falling into the positive and negative regions of $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1\widehat{\dot{B}}_2$, respectively. See Figure 7(a). The symmetry statistic is $\widehat{S}_{(1111)} = 152$, and the z-statistic of the difference is 6.52, making the p-value of the BET to be 6.5×10^{-10}. After multiplying 5×10^7 for the Bonferroni control, the overall adjusted p-value is 0.033, which is strong evidence against the independence null. Furthermore, from the interaction $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1\widehat{\dot{B}}_2$ we could see interesting dependency patterns: in part of the data there exists strong monotone increasing dependency, while there is a cluster of observations above the third quartile of U and below the first quartile of V. These patterns make the overall dependency nonlinear, which is captured by $\widehat{\dot{A}}_1\widehat{\dot{A}}_2\widehat{\dot{B}}_1\widehat{\dot{B}}_2$.

The above EDA with the BET suggests an interesting question: What is the reason of this nonlinear dependency? By adding the label of basal-like breast cancer, the cluster of obser-


(a) Empirical Max BET: W=348, B=196. Z−Stat=6.52

(b) Same Data with Cancer Subtype Labels

(c) The Scatterplot in the Original Scale

Figure 7. (a) The BET with $d = 2$ for two genes in the TCGA data. There are 348 observations and 196 observations in the empirical copula distribution falling into the positive and negative regions of $\widehat{A}_1\widehat{A}_2\widehat{B}_1\widehat{B}_2$, respectively. The z-statistic of the difference is 6.52. (b) The same two genes with the labels shown. Basal-like breast cancer patients are marked with a red triangle. (c) The scatterplot of same gene expressions in the original scale.

vations in the lower right white box can be explained as a result of the joint distribution of the two genes under this subtype. From Figure 7(b), we see clearly that basal-like breast cancer patients tend to have higher DZIP1 intensity and lower NAV3 intensity. We also make the scatterplot of the same two genes in the original scale in Figure 7(c), and we see that the bivariate distribution of DZIP1 and NAV3 under the basal-like subtype has different location and scale and is almost disjoint from the rest of the data. This fact explains the reason of nonlinearity in the pooled distribution: when the bivariate distribution of this subtype is mixed together with those of other subtypes, some nonlinearity pattern is created. With the identification of this nonlinearity from the BET and with the label information, we can retrospectively extract such mixtures of different subtype distributions.

By searching the medical literature, we find both genes have been individually investigated and are confirmed to be highly related to basal-like breast cancer. For examples, the relationship between DZIP1 and basal-like cancer is studied by Kikuyama et al. (2012) and ShigunovShigunov et al. (2014), and similar studies for NAV3 are done in Maliniemi et al. (2011) and Cohen-Dvashi et al. (2015). However, we are not able to find results on the joint behavior of these two genes. The BET result indicates that this joint behavior could be scientifically important, as these two genes behave dramatically different under the basal-like subtype. This further suggests the possible existence of some biological functional relationships between these two genes and this subtype of cancer. This could be an interesting issue to investigate.

8.3. Improvements in Classification

Statistically, the above EDA with the BET suggests that DZIP1 and NAV3 could jointly be good predictors of basal-like breast cancer. We validate this conjecture with the test dataset of 273 subjects. We use the k-nearest neighbor classification method with $k = 1$. The classification accuracy in the test dataset is 91%. We assess this performance with cross-validation and observe similar results. Note that if we were to use DZIP1 or NAV3 alone for the classification task, the accuracy was 79% and 76%, respectively, that is, each of them is a good predictor but far from perfect. However, by combining these two genes and using the joint distribution for classification, we substantially improve the classification accuracy.

Existing classification studies are usually based on a selected set of many variables. One drawback of such studies is lack of interpretability. With some black box selection procedure over many variables, the effect of each variable is hard to scientifically interpret. On the other hand, the BET analysis can help identify pairs of variables which have high potential joint classification power, and explanations of the effects of variables can be obtained from the pattern of the nonlinear dependency. Therefore, the BET can be a useful EDA tool in practice: It provides p-values that we can see.

9. Summary and Discussions

Nonparametric dependence detection is an important problem in statistics. To avoid the power loss due to nonuniform consistency, we introduce the concept of BEStat, which combines four classical statistical wisdoms: copula, filtration, orthogonal design, and multiple testing. The proposed BET framework combines the strength from these wisdoms and enjoys the invariance property from the copula distribution, universality, identifiability and uniformity from the filtration, orthogonality and symmetry from the orthogonal design, and interpretability from multiple testing. The binary expansion approach also facilitates efficient bitwise computing implementation.

Two important potential generalizations are nonparametric tests of independence for general categorical variables and for random vectors. For general contingency tables, the filtration and the separation of marginal and joint information need to be developed carefully. For random vectors, the binary expansion filtration approximation in (3.1) , the BID equation in Theorem 3.4 and the IOR reparametrization can all be generalized. We welcome further thoughts on related topics for deeper understanding of dependence and useful procedures in practice.

Supplementary Materials

Online supplementary materials for this article include additional numerical studies, proofs of the results, and R functions used in the numerical studies.

Acknowledgments

The author thanks Richard Berk, Larry Brown, Andreas Buja, Edward George, Arun Kumar Kuchibhotla, Linda Zhao, and Zhigen Zhao for inspiring discussions that stimulated this research. The author also thanks Edoardo Airoldi, Mike Baiocchi, Shankar Bhamidi, Bhaswar Bhattacharya, Rong Chen, Jessi Cisewski, Bradley Efron, Jianqing Fan, Dean Foster, Andrew Gelman, Jan Hannig, Ruth Heller, Peter Hoff, Katherine Hoadley, Xiaoming Huo, Pierre Jacob, Vinay Kashyap, Michael Kosorok, S.C. Samuel Kou, Duyeol Lee, Michael Levine, Ping Li, Yun Li, Xihong Lin, Oliver Linton, Han Liu, Jun Liu, Mike Love, Li Ma, Zongming Ma, Steve Marron, Xiao-Li Meng, Joel Parker, Charles Perou, Vladas Pipiras, Richard Samworth, David Siegmund, Dylan Small, Robert Stine, William Strawderman, Weijie Su, Gábor Székely, Xinlu Tan, Yihong Wu, Han Xiao, Daniel Yekutieli, Ming Yuan, Yuan Yuan, Cun-Hui Zhang, Nancy Zhang, Tingting Zhang, and Harrison Zhou for valuable comments and suggestions. This research is completed while the author is visiting Princeton University. The author thanks Jianqing Fan and the Department of Operations Research and Financial Engineering at Princeton for the warm hospitality.

Funding

This research is partially supported by NSF DMS-1613112 and NSF IIS-1633212.

References

Acharya, J., Daskalakis, C., and Kamath, G. C. (2015), "Optimal Testing for Properties of Distributions," in *Advances in Neural Information Processing Systems*, pp. 3591–3599. [1625,1629]

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, 7, 131–153. [1627]

Agresti, A., and Kateri, M. (2011), *Categorical Data Analysis*, Berlin: Springer. [1623,1624,1625,1629]

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011), "Global Testing Under Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism," *The Annals of Statistics*, 39, 2533–2556. [1628,1629]

Barnett, I., Mukherjee, R., and Lin, X. (2017), "The Generalized Higher Criticism for Testing SNP-Set Effects in Genetic Association Studies," *Journal of the American Statistical Association*, 112, 64–76. [1628]

Box, G. E., Hunter, J. S., and Hunter, W. G. (2005), *Statistics for Experimenters: Design, Innovation, and Discovery* (Vol. 2), New York: Wiley-Interscience. [1626]

Buja, A., and Tukey, P. A. (1992), *Computing and Graphics in Statistics*, New York: Springer-Verlag New York, Inc. [1633]

Cancer Genome Atlas Network (2012), "Comprehensive Molecular Portraits of Human Breast Tumors," *Nature*, 490, 61. [1633]

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., and Bowlby, R. (2015), "Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer," *Cell*, 163, 506–519. [1633]

Cleveland, W. S. (1993), *Visualizing Data*, Summit, NJ: Hobart Press. [1633]

Cohen-Dvashi, H., Ben-Chetrit, N., Russell, R., Carvalho, S., Lauriola, M., Nisani, S., Mancini, M., Nataraj, N., Kedmi, M., Roth, L., and Köstler, W. (2015), "Navigator-3, a Modulator of Cell Migration, May Act as a Suppressor of Breast Cancer Progression," *EMBO Molecular Medicine*, 7, 299–314. [1635]

Cornfield, J. (1956), "A Statistical Problem Arising From Retrospective Studies," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 4), Berkeley, CA: University of California Press, pp. 135–148. [1625]

Cox, D. R., and Reid, N. (2000), *The Theory of the Design of Experiments*, Boca Raton, FL: CRC Press. [1626]

Craiu, R. V., and Meng, X.-L. (2005), "Multiprocess Parallel Antithetic Coupling for Backward and Forward Markov Chain Monte Carlo," *Annals of Statistics*, 33, 661–697. [1622]

——— (2006), "Meeting Hausdorff in Monte Carlo: A Surprising Tour With Antihype Fractals," *Statistica Sinica*, 16, 77–91. [1622]

Fienberg, S. E. (2007), *The Analysis of Cross-Classified Categorical Data*, New York: Springer Science & Business Media. [1623,1624,1625,1629]

Filippi, S., and Holmes, C. (2015), "A Bayesian Nonparametric Approach to Testing for Dependence Between Random Variables," arXiv no. 1506.00829. [1621,1631]

Forsyth, D. A., and Ponce, J. (2002), *Computer Vision: A Modern Approach*, Upper Saddle River, NJ: Prentice Hall Professional Technical Reference. [1622]

Golubov, B., Efimov, A., and Skvortsov, V. (2012), *Walsh Series and Transforms: Theory and Applications* (Vol. 64), Netherlands: Springer Science & Business Media. [1626]

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007), "A Kernel Statistical Test of Independence," in *Advances in Neural Information Processing Systems*, pp. 585–592. [1620]

Han, F., Chen, S., and Liu, H. (2017), "Distribution-Free Tests of Independence in High Dimensions," *Biometrika*, 104, 813–828. [1620]

Harmuth, H. (2013), *Transmission of Information by Orthogonal Functions*, Berlin, Heidelberg: Springer. [1626]

Heller, R., and Heller, Y. (2016). "Multivariate tests of association based on univariate tests," In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 29*, pp. 208–216. Curran Associates, Inc., *http://papers.nips.cc/paper/6220-multivariate-tests-of-association-based-on-univariate-tests.pdf*. [1621]

Heller, R., Heller, Y., and Gorfine, M. (2012), "A Consistent Multivariate Test of Association Based on Ranks of Distances," *Biometrika*, 100, 503–510. [1621,1631]

Heller, R., Heller, Y., Kaufman, S., Brill, B., and Gorfine, M. (2016), "Consistent Distribution-Free *k*-Sample and Independence Tests for Univariate Random Variables," *Journal of Machine Learning Research*, 17, 1–54. [1621]

Hoeffding, W. (1948), "A Non-parametric Test of Independence," *The Annals of Mathematical Statistics*, 19, 546–557. [1620]

Kac, M. (1959), *Statistical Independence in Probability, Analysis and Number Theory* (Vol. 134), Oberlin, OH: Mathematical Association of America. [1623]

Kikuyama, M., Takeshima, H., Kinoshita, T., Okochi-Takada, E., Wakabayashi, M., Akashi-Tanaka, S., Ogawa, T., Seto, Y., and Ushijima, T. (2012), "Development of a Novel Approach, the Epigenome-Based Outlier Approach, to Identify Tumor-Suppressor Genes Silenced by Aberrant DNA Methylation," *Cancer Letters*, 322, 204–212. [1635]

Kimeldorf, G., and Sampson, A. R. (1978), "Monotone Dependence," *Annals of Statistics*, 6, 895–903. [1622]

Kinney, J. B., and Atwal, G. S. (2014), "Equitability, Mutual Information, and the Maximal Information Coefficient," *Proceedings of the National Academy of Sciences of the United States of America*, 111, 3354–3359. [1621,1630,1631]

Kou, S., and Ying, Z. (1996), "Asymptotics for a 2× 2 Table With Fixed Margins," *Statistica Sinica*, 6, 809–829. [1630,1634]

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004), "Estimating Mutual Information," *Physical Review E*, 69, 066138. [1621]

Lehmann, E. L., and Romano, J. P. (2006), *Testing Statistical Hypotheses*, New York: Springer Science & Business Media. [1621]

Liu, K., and Meng, X.-L. (2014), "Comment: A Fruitful Resolution to Simpson's Paradox via Multiresolution Inference," *The American Statistician*, 68, 17–29. [1621]

——— (2016), "There Is Individualized Treatment. Why Not Individualized Inference?," *Annual Review of Statistics and Its Application*, 3, 79–111. [1621]

Lynn, P. A. (1973), *An Introduction to the Analysis and Processing of Signals*, London: Macmillan. [1626]

Ma, L., and Mao, J. (2019), "Fisher Exact Scanning for Dependency," *Journal of the American Statistical Association* (accepted). [1621,1624,1627,1630,1631]

Maliniemi, P., Carlsson, E., Kaukola, A., Ovaska, K., Niiranen, K., Saksela, O., Jeskanen, L., Hautaniemi, S., and Ranki, A. (2011), "Nav3 Copy Number Changes and Target Genes in Basal and Squamous Cell Cancers," *Experimental Dermatology*, 20, 926–931. [1635]

Miller, R., and Siegmund, D. (1982), "Maximally Selected Chi Square Statistics," *Biometrics*, 38, 1011–1016. [1620]

Nelsen, R. B. (2007), *An Introduction to Copulas*, New York: Springer Science & Business Media. [1624]

Paninski, L. (2008), "A Coincidence-Based Test for Uniformity Given Very Sparsely Sampled Discrete Data," *IEEE Transactions on Information Theory*, 54, 4750–4755. [1629]

Pearl, J. (1971), "Application of Walsh Transform to Statistical Analysis," in *IEEE Transactions on Systems, Man, and Cybernetics, SMC-1*, 111–119. [1626]

Perou, C. M., Sørile, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., and Fluge, Ø. (2000), "Molecular Portraits of Human Breast Tumours," *Nature*, 406, 747–752. [1633]

Perryman, M. A., Lindegren, L., Kovalevsky, J., Hoeg, E., Bastian, U., Bernacca, P., Crézé, M., Donati, F., Grenon, M., Grewing, M., and Van Leeuwen, F. (1997), "The HIPPARCOS Catalogue," *Astronomy and Astrophysics*, 323, L49–L52. [1632]

Pfister, N., Bühlmann, P., Schölkopf, B., and Peters, J. (2016), "Kernel-Based Tests for Joint Independence," arXiv no. 1603.00285. [1620]

Rényi, A. (1959), "On Measures of Dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, 10, 441–451. [1620]

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011), "Detecting Novel Associations in Large Data Sets," *Science*, 334, 1518–1524. [1620,1630,1631]

Reshef, D. N., Reshef, Y. A., Sabeti, P. C., and Mitzenmacher, M. M. (2015a), "An Empirical Study of Leading Measures of Dependence," arXiv no. 1505.02214. [1620]

——— (2015b), "Equitability, Interval Estimation, and Statistical Power," arXiv no. 1505.02212. [1620]

Romano, J. P. (1989), "Bootstrap and Randomization Tests of Some Nonparametric Hypotheses," *Annals of Statistics*, 17, 141–159. [1620]

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013), "Equivalence of Distance-Based and RKHS-Based Statistics in Hypothesis Testing," *Annals of Statistics*, 41, 2263–2291. [1620]

Shigunov, P., Sotelo-Silveira, J., Stimamiglio, M. A., Kuligovski, C., Irigoín, F., Badano, J. L., Munroe, D., Correa, A., and Dallagiovanna, B. (2014), "Ribonomic Analysis of Human DZIP1 Reveals Its Involvement in Ribonucleoprotein Complexes and Stress Granules," *BMC Molecular Biology*, 15, 12. [1635]

Sun, N., and Zhao, H. (2014), "Putting Things in Order," *Proceedings of the National Academy of Sciences of the United States of America*, 111, 16236–16237. [1633]

Sylvester, J. J. (1867), "LX. Thoughts on Inverse Orthogonal Matrices, Simultaneous Signsuccessions, and Tessellated Pavements in Two or More Colours, With Applications to Newton's Rule, Ornamental Tile-Work, and the Theory of Numbers," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 34, 461–475. [1626]

Székely, G. J., and Rizzo, M. L. (2009), "Brownian Distance Covariance," *The Annals of Applied Statistics*, 3, 1236–1265. [1620]

——— (2013a), "The Distance Correlation *t*-Test of Independence in High Dimension," *Journal of Multivariate Analysis*, 117, 193–213. [1620]

——— (2013b), "Energy Statistics: A Class of Statistics Based on Distances," *Journal of Statistical Planning and Inference*, 143, 1249–1272. [1620]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007), "Measuring and Testing Dependence by Correlation of Distances," *The Annals of Statistics*, 35, 2769–2794. [1620]

Tsybakov, A. (2008), *Introduction to Nonparametric Estimation*, Springer Series in Statistics, New York: Springer. [1622]

Valiant, L. G. (1984), "A Theory of the Learnable," *Communications of the ACM*, 27, 1134–1142. [1621]

Vitale, R. A. (1990), "On Stochastic Dependence and a Class of Degenerate Distributions," *Lecture Notes-Monograph Series*, 16, 459–469. [1622]

Walther, G. (2010), "Optimal and Fast Detection of Spatial Clusters With Scan Statistics," *The Annals of Statistics*, 38, 1010–1033. [1630]

Wang, X., Jiang, B., and Liu, J. S. (2016), "Generalized *R*-Squared for Detecting Non-independence," arXiv no. 1604.02736. [1621]

Zhao, Z., Baiocchi, M., and Zhang, K. (2019), "Fast, Flexible, and Powerful: Introducing a Scalable, Bitwise Framework for Non-parametric Testing for Dependence Structure," submitted. [1622,1626,1631]

Zheng, S., Shi, N.-Z., and Zhang, Z. (2012), "Generalized Measures of Correlation for Asymmetry, Nonlinearity, and Beyond," *Journal of the American Statistical Association*, 107, 1239–1252. [1620]