Sankhyā A : The Indian Journal of Statistics
https://doi.org/10.1007/s13171-023-00333-7

Extreme Value Theory for Binary Expansion Testing

Siqi Xiang, Wan Zhang, Kai Zhang and J. S. Marron
University of North Carolina at Chapel Hill, Chapel Hill, USA

Abstract

Binary expansion testing (BET) provides powerful detection of interesting nonlinear dependence among pairs of variables in the exploratory data analysis of large-scale data sets. However, the Bonferroni adjusted p-values can be overly conservative when used to determine the significant testing pairs. A novel contribution of this paper is the extreme value theory analysis of BET. This results in a potentially powerful new significance threshold for the maximal BET z-statistics.

AMS (2000) subject classification. Primary 62G32; Secondary 62G10.
Keywords and phrases. Binary expansion testing, extreme value, nonlinear dependence, nonparametric dependence testing.

1 Introduction

The Binary Expansion Testing (BET) framework (Zhang, 2019) is a powerful nonparametric test of independence of two continuous random variables. In particular, it has wide use in the detection of interesting nonlinear relationships in pairwise applications, especially for large scale data sets. For example, Xiang et al. (2022) uses BET to find many important nonlinear patterns among the expression of pairs of genes from the breast cancer subset of The Cancer Genome Atlas (TCGA) data set (The Cancer Genome Atlas Network, 2012). Many of these seem to have interesting biological interpretations.

Although BET works well in testing independence between pairwise variables, there is room for improvement in the exploratory data analysis of large-scale data sets. As discussed in Section 3.2 of Xiang et al. (2022), the significantly dependent pairs are determined by the pairwise application of BET using Bonferroni adjusted p-values. However, that Bonferroni adjustment can be overly conservative, especially in large-scale data sets such as TCGA. To address this issue and improve the statistical power, the goal of this paper is to study the distribution of the maximal BET z-statistics when doing pairwise testing. Specifically, we plan to use extreme value theory to

Published online: 06 December 2023

derive the limiting distribution of the maximal BET z-statistics under the null hypothesis that all variables are independent. In Section 2, we briefly introduce the ideas and procedure of BET. In Section 3, we calculate the distribution of the BET z-score for pairwise testing and the limiting distribution of the maximal BET z-statistics. In Section 4, we propose a new method to select the significantly dependent pairs by using the maximal BET z-statistics.

2 BET Framework: Background and Definitions

First, we give a brief introduction to binary expansion as developed in Zhang (2019). This BET framework uses symmetry statistics that are complete and sufficient statistics for dependence. By Basu's theorem (Basu, 1958), bounded complete sufficient statistics are independent of any ancillary statistic. Thanks to this important theoretical insight, the analysis of dependence can be focused on these symmetry statistics.

Suppose $X_1, X_2, ..., X_p$ are p independent continuous variables, and n independent replications are observed. For testing the independence of X_i and X_j using BET, first, the copula transformation is applied to get $U = F_{X_i}(X_i)$ and $V = F_{X_j}(X_j)$, where U, V have a uniform marginal distribution over $[0, 1]$, and the ordering relationship of X_i and X_j is preserved. For example, in Fig. 1, the left panel shows a scatter plot of expression for the genes FAM171A1 and SPARCL1 from TCGA (The Cancer Genome Atlas Network, 2012) in the normalized log count scale, which has an approximately parabolic relationship; the right panel shows the scatter plot of the same two genes after the copula transformation with the BET diagnostic plot explained below.

The binary expansions of U and V are expressed as $U = \sum_{k=1}^{\infty} A_k/2^k$ and $V = \sum_{k'=1}^{\infty} B_{k'}/2^{k'}$, where $A_k \overset{i.i.d}{\sim} Bernoulli(1/2), B_{k'} \overset{i.i.d}{\sim} Bernoulli(1/2)$. Next, in the BET procedure, we truncate the expansions at given depths d_1 and d_2 separately. To achieve common approximation error, we assume $d_1 = d_2 = d$, then the truncated expansions $U_d = \sum_{k=1}^{d} A_k/2^k$ and $V_d = \sum_{k'=1}^{d} B_{k'}/2^{k'}$ are approximations of U and V, and are discrete variables taking 2^d possible values.

Such A_k and $B_{k'}$ generate $m = (2^d - 1)^2$ cross interactions, each of which is a new binary variable and reflects a dependence relationship between U_d and V_d. To express each cross interaction in the form of products, we first define new binary variables: $\dot{A}_k = 2A_k - 1$ and $\dot{B}_{k'} = 2B_{k'} - 1$, where $\dot{A}_k$ and $\dot{B}_{k'}$ take the values -1 and 1. Hence the interaction between each pair

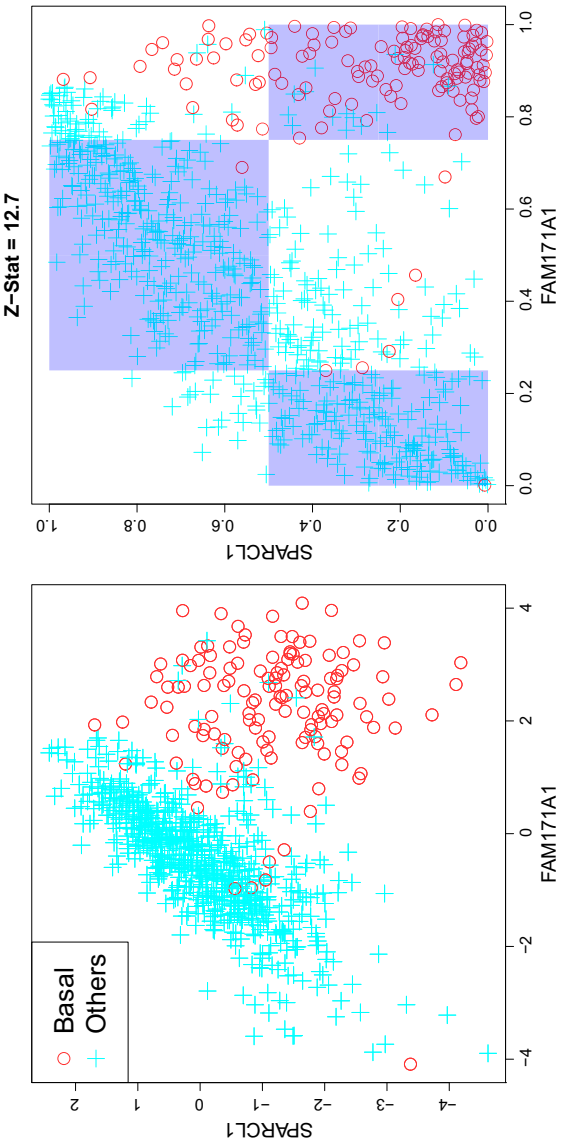

Figure 1: Left: The scatter plot of expression of two genes in TCGA breast cancer data in the normalized log count scale; Right: The scatter plot of the same two genes after the copula transformation with the nonlinear dependence pattern from BET. Strong statistical significance is indicated by the BET Z-statistic of 12.7

of A_k and $B_{k'}$ can be presented as the product $\dot{A}_k\dot{B}_{k'}$. Thus, each cross interaction results from the product of at least one $\dot{A}_k$ and $\dot{B}_{k'}$, which has the form $\dot{A}_{k_1}...\dot{A}_{k_r}\dot{B}_{k'_1}...\dot{B}_{k'_t}$ with $r, t > 0$. Each cross interaction is a binary variable taking the values -1 and 1, dividing the unit square $[0,1]^2$ into 2 regions. These two regions are colored white (1) and shaded (-1). For example, the right panel of Fig. 1 shows one cross interaction when given depth parameter $d = 2$. This cross interaction captures a particular parabolic dependence pattern in terms of many more points in the shaded region than in the white region. This relationship is clear in the scatter plot on the left panel.

Each cross interaction defines one Binary Interaction Design (BID). For each BID, a symmetry statistic S_k is defined as the difference of point counts of n pairs (U, V) in white and shaded regions partitioned by the BID, as shown in the right panel of Fig. 1 where S_k is -363. If U_d and V_d are independent, the points should be uniformly distributed on the unit squares, and for each BID, the symmetry statistic S_k should be close to 0. Thus, when the absolute value of one S_k is far from 0, we could reject the independent null hypothesis. The corresponding BID could explain the potential dependence pattern between U and V, i.e., the relationship between X_i and X_j.

For m BIDs used in the BET procedure, the z-score of each BID is defined as $|S_k/\sqrt{n}|$. The title of the right panel of Fig. 1 indicates the z-score of the corresponding BID (12.7).

Now we get a complete BET procedure for testing a pair of variables. First we choose depth parameters $d_1 = d_2 = d$ and get the $m = (2^d - 1)^2$ BIDs. After computing all symmetry statistics S_k of m BIDs for this pair, we find the BID with the largest absolute value of S_k, and also record the corresponding p-value and z-statistic. We use this largest BID z-score as the z-statistic of BET, i.e., when given $d_1 = d_2 = d$, we denote $Z_{ij,d}$ as the z-statistic of BET between X_i and X_j:

$$Z_{ij,d} = \max_{k\in 1,\ldots,m} \frac{|S_k|}{\sqrt{n}}. \tag{1}$$

As discussed above, the BET diagnostic plot in the right panel of Fig. 1 (white and shaded regions) represents the BID detecting a parabolic relationship, where the absolute value of the corresponding symmetry statistic S_k of this BID for this pair of genes is the largest and is far from 0. Thus, the z-score of this BID shown in the title of the right panel (z-stat = 12.7) is also the z-statistic of BET between this pair of genes, and this BID gives a reasonable explanation of the dependence pattern. Based on the symbols

of the breast cancer subtypes, we notice that this relationship tends to be driven by a separation between the Basal (dark ∘) subgroup and other breast cancer (light +) subgroups, which is biologically meaningful.

Xiang et al. (2022) did a careful comparison with more conventional measures of dependence, and found some overall improvement in identifying non-linear dependent pairs of variables, at a much reduced computational cost.

3 Distribution of the Maximal Z-statistic of BET

In this section, we discuss the approximate distribution of the maximal BET z-score among all pairs of variables in a $n \times p$ data set. Recall in Section 2, we define $Z_{ij,d}$ to be the BET z-score for testing dependence between X_i and X_j. Let $Z_{(pn),d} = \max_{1 \le i < j \le p} Z_{ij,d}$, which denotes the maximal BET z-statistic of the pairwise application to the $n \times p$ data set. We have two steps to calculate the approximate distribution of $Z_{(pn),d}$: first, in Section 3.1, we calculate a tail bound of the z-score of BET between two independent variables $Z_{ij,d}$; second, in Section 3.2, we derive an approximate distribution of the maximal BET z-score among p independent variables $Z_{(pn),d}$.

3.1 Distribution of the BET Z-statistic In the first step, the tail bound of the BET z-statistic is stated in the following theorem:

Theorem 1 *Suppose X_i, X_j are two independent variables and n independent replications are observed. When given BET depths $d_1 = d_2 = d$, let $\boldsymbol{A}$ represent the set of indices of m BIDs, where $m = (2^d - 1)^2$, and let S_k denote the corresponding symmetry statistic for a given $k \in \boldsymbol{A}$. Then the z-statistic of BET between X_i and X_j is $Z_{ij,d} = \max_{k \in \boldsymbol{A}} |S_k|/\sqrt{n}$, which satisfies as $n \to \infty$,*

$$1 - e^{-\lambda_1(z,n)} - \lambda_1(z,n)^2/m \le P(Z_{ij,d} > z) \le 1 - e^{-\lambda_1(z,n)} + \lambda_1(z,n)^2/m, \quad (2)$$

where

$$\begin{aligned}
\lambda_1(z,n) &= 2mg(a,n)(1 + o(1)),\\
g(z,n) &= \frac{1}{1-r}\frac{1}{\sqrt{2\pi a(1-a)n}}e^{-nH},\\
a(z,n) &= \frac{z\sqrt{n} + n + 1}{2n}\\
r(z,n) &= \frac{1-a}{a},\\
H(z,n) &= a\log(2a) + (1-a)\log(2-2a).
\end{aligned}$$

Proof of Theorem 1 To get the tail bound of the z-statistic of BET, first we calculate the approximate distribution of $Z_{ij,d}$:

$$P(Z_{ij,d} \leq z) = P(\max_{k \in \boldsymbol{A}} \frac{|S_k|}{\sqrt{n}} \leq z) = P(\frac{|S_1|}{\sqrt{n}} \leq z, ..., \frac{|S_m|}{\sqrt{n}} \leq z),$$

where $S_1, ..., S_m$ represent the symmetry statistics S for all m BIDs.

Before calculating the distribution of the z-statistic of BET, we need to calculate the tail bound of the z-score of a given BID, which is $p_1 = P(\frac{|S_k|}{\sqrt{n}} > z)$:

$$\begin{aligned} p_1 &= P(\frac{|S_k|}{\sqrt{n}} > z) = P(|S_k| > z\sqrt{n}) = P(S_k > z\sqrt{n} \ or \ -S_k < -z\sqrt{n}) \\ &= P(S_k \geq z\sqrt{n} + 1 \ or \ -S_k \leq -z\sqrt{n} - 1) = 2P(S_k \geq z\sqrt{n} + 1). \end{aligned}$$

This equation reflects that we first need to compute the probability distribution of the symmetry statistic S_k. □

To obtain $P(S_k \geq s)$, we base our calculation on the following observations:

- $S_1, ..., S_m$ are pairwise independent from Theorem 4.3. in Zhang (2019).
- $B_{n,k} = (S_k + n)/2 \sim Binomial(n, 1/2)$ from Theorem 4.1. in Zhang (2019).
- The following lemma is Theorem 2 from Arratia and Gordon (1989) which gives a large deviation for the binomial distribution:

Lemma 2 *For $B_n \sim B(n, p')$, if the $P(B_n \geq k)$ satisfies $p' < a = k/n < 1$, then:*

$$P(B_n \geq an) \sim \frac{1}{1-r} \frac{1}{\sqrt{2\pi a(1-a)n}} e^{-nH} \ as \ n \ \to \infty,$$

where

$$\begin{aligned} r &\equiv \frac{p'}{a} \frac{1-a}{1-p'}, \\ H &\equiv a \log(\frac{a}{p'}) + (1-a) \log(\frac{1-a}{1-p'}). \end{aligned}$$

According to the second observation, the distribution of $B_{n,k}$ is written as:

$$P(B_{n,k} \geq an) = P((S_k + n)/2 \geq an) = P(S_k \geq 2an - n) = P(S_k \geq (2a-1)n).$$

Then according to Lemma 2, we let

$$\begin{aligned}(2a-1)n &= z\sqrt{n} + 1,\ i.e.,\ a = \frac{z\sqrt{n} + n + 1}{2n} > p' = 1/2,\\ r &= \frac{1-a}{a},\\ H &= a\log(2a) + (1-a)\log(2-2a),\end{aligned}$$

and define

$$g(z,n) = \frac{1}{1-r}\frac{1}{\sqrt{2\pi a(1-a)n}}e^{-nH},$$

the tail bound of the symmetry statistic S_k is given as follows:

$$P(S_k \geq z\sqrt{n}+1) = P(S_k \geq (2a-1)n) = P(B_{n,k} \geq an) = g(a,n)(1+o(1))\ as\ n \to \infty.$$

Thus we have the following expression:

$$p_1 = 2P(S_k \geq z\sqrt{n}+1) = 2g(a,n)(1+o(1))\ as\ n \to \infty.$$

To calculate the tail bound of the z-statistic of BET, we first define a useful notation for a Poisson approximation: let I be a finite or countable index set. For each $\alpha \in I$, let Y_α be a Bernoulli random variable with $p_\alpha = P(Y_\alpha = 1) > 0$, and let $\{N_\alpha, \alpha \in I\}$ be a set of subsets of I, i.e., $N_\alpha \subseteq I$ with $\alpha \in N_\alpha$. The set N_α is thought of as a neighborhood of α consisting of the set of the indices β such that Y_α and Y_β are dependent. Let

$$W = \sum_{\alpha \in I} Y_\alpha\ and\ \lambda = EW,$$

and define:

$$b_1 = \sum_{\alpha \in I}\sum_{\beta \in N_\alpha} p_\alpha p_\beta, \tag{3}$$

$$b_2 = \sum_{\alpha \in I}\sum_{\beta : \alpha \neq \beta \in N_\alpha} p_{\alpha\beta},\ where\ p_{\alpha\beta} = E[Y_\alpha Y_\beta], \tag{4}$$

$$b_3 = \sum_{\alpha \in I} E|E\{Y_\alpha - p_\alpha | \sigma(Y_\beta : \beta \notin N_\alpha)\}|. \tag{5}$$

Based on the above notation, we use the following lemma giving a Poisson approximation for the maximum of dependent variates using the first and second moments, which is from Arratia et al. (1990):

Lemma 3 *If b_1, b_2 and b_3 defined by (3), (4) and (5) are all small, then the probability of the event $\{W = 0\}$ has a Poisson approximation:*

$$|P(W = 0) - e^{-\lambda}| \leq (b_1 + b_2 + b_3)(1 - e^{-\lambda})/\lambda < (1 \wedge \lambda^{-1})(b_1 + b_2 + b_3).$$

In our calculation of $P(Z_{ij,d} \leq z) = P(\frac{|S_1|}{\sqrt{n}} \leq z, ..., \frac{|S_m|}{\sqrt{n}} \leq z)$, we take $I = I_1 = \{1, ..., m\}$. Let $\alpha_1 = k \in I_1$, and since the S_k are pairwise independent, we define $N_{\alpha_1} = N_k = \{k\}$, and we have:

$$\begin{aligned}
Y_{\alpha_1} &= Y_k = \mathbf{1}\{\frac{|S_k|}{\sqrt{n}} > z\}, \\
p_{\alpha_1} &= p_k = p_1 = 2g(a, n)(1 + o(1)) \ as \ n \to \infty, \\
W_1 &= \sum_{k=1}^{m} Y_k, \\
\lambda &= \lambda_1 = mEY_k = mp_{\alpha_1} = 2mg(a, n)(1 + o(1)) \ as \ n \to \infty,
\end{aligned}$$

and the distribution of the z-statistic of BET is written as:

$$\begin{aligned}
P(Z_{ij,d} \leq z) &= P(\frac{|S_1|}{\sqrt{n}} \leq z, ..., \frac{|S_m|}{\sqrt{n}} \leq z) \\
&= P(\sum_{k=1}^{m} Y_k = 0) \\
&= P(W_1 = 0).
\end{aligned}$$

Thus in our proof, we calculate b_1, b_2 and b_3 from (3), (4) and (5) as follows:

$$\begin{aligned}
b_1 &= \sum_{k \in I} p_k^2 = mp_1^2 = \lambda_1^2/m, \\
b_2 &= 0, \\
b_3 &= 0,
\end{aligned}$$

since Y_α is independent of the sigma field $\sigma(Y_\beta : \beta \notin N_\alpha)$ and $E\{Y_\alpha|\sigma(Y_\beta : \beta \notin N_\alpha)\} = p_\alpha$, $E\{Y_\alpha - p_\alpha|\sigma(Y_\beta : \beta \notin N_\alpha)\} = 0$.

Finally according to Lemma 3, we get:

$$\begin{aligned} e^{-\lambda_1} - \lambda_1^2/m < P(Z_{ij,d} \le z) < e^{-\lambda_1} + \lambda_1^2/m \\ 1 - e^{-\lambda_1} - \lambda_1^2/m \le P(Z_{ij,d} > z) \le 1 - e^{-\lambda_1} + \lambda_1^2/m, \end{aligned}$$

where $\lambda_1 = 2mg(a,n)(1+o(1))$ as $n \to \infty$.

Thus, Theorem 3.1 gives a useful tail bound of the BET z-score between X_i and X_j from the limiting distribution. This tail bound sheds light on the limiting distribution of the maximal BET z-statistic, as described in Section 3.2.

3.2 Limiting Distribution of the Maximal BET Z-statistic Now we derive a limiting distribution of the maximal BET z-score among p independent variables, as stated in the following theorem:

Theorem 4 *Suppose $X_1, X_2, ..., X_p$ are p independent variables and n independent replications are observed. When given BET depths $d_1 = d_2 = d$, we have m BIDs. Then the maximal z-score of BET is $Z_{(pn),d} = \max_{1\le i<j\le p} Z_{ij,d}$, which satisfies as $p \to \infty$ and $p = O(n^\delta)$, where $\delta \in (0, 1/3)$,*

$$\lim_{n\to\infty} |P(Z_{(pn),d} \le \frac{p^4}{n^{4\delta-1/2}}t) - e^{f(t,p,n)}| = 0,$$

where

$$\begin{aligned} f(t,p,n) &= -mp(p-1)\frac{a}{2a-1}\frac{[2a^a(1-a)^{1-a}]^{-n}}{\sqrt{2\pi a(1-a)n}}, \\ a(t,p,n,\delta) &= \frac{p^4 t/n^{4\delta-1} + n + 1}{2n}. \end{aligned}$$

Proof of Theorem 4 First, for a given pair (X_i, X_j), from Theorem 1 we have the tail bound of the z-score between X_i and X_j:

$$\begin{aligned} 1 - e^{-\lambda_1} - \lambda_1^2/m \le P(Z_{ij,d} > z) \le 1 - e^{-\lambda_1} \\ + \lambda_1^2/m, \ \textit{where } \lambda_1 = 2mg(a,n)(1+o(1)), \end{aligned}$$

i.e,

$$P(Z_{ij,d} > z) = (1 - e^{-\lambda_1})(1 + o(1)) \ \textit{as } n \to \infty,$$

$$P(Z_{ij,d} > z) = [1 - e^{-2mg(a,n)}](1 + o(1)) \; as \; n \to \infty,$$

where

$$\begin{aligned} g(a,n) &= \frac{1}{1-r}\frac{1}{\sqrt{2\pi a(1-a)n}}e^{-nH}, \\ a &= \frac{z\sqrt{n}+n+1}{2n} \\ r &= \frac{1-a}{a}, \\ H &= a\log(2a) + (1-a)\log(2-2a). \end{aligned}$$

Now, to calculate the approximate distribution of $Z_{(pn),d} = \max_{1\le i<j\le p} Z_{ij,d}$, since $\{Z_{ij,d}, 1 \le i < j \le p\}$ are not mutually independent, we use the Poisson approximation theorem Lemma 3 again.

In this proof, we take the finite index set $I = I_2 = \{(i,j), 1 \le i < j \le p\}$. Let $\alpha_2 = (i,j) \in I_2$, we define the neighborhood set $N_{\alpha_2} = N_{ij} = \{(k,l) \in I_2; k = i \; or \; l = j\}$, and we have:

$$\begin{aligned} Y_{\alpha_2} &= Y_{ij} = \mathbf{1}\{Z_{ij,d} > z\}, \\ W &= W_2 = \sum_{1\le i<j\le p} Y_{ij}, \\ p_{\alpha_2} &= P(Z_{ij,d} > z), \\ \lambda &= \lambda_2 = \sum_{1\le i<j\le p} EY_{ij} = \frac{p(p-1)}{2}P(Z_{ij,d} > z) = \frac{p(p-1)}{2}p_{\alpha_2}. \end{aligned}$$

Based on the above definitions, the probability distribution of the maximal BET z-score can be presented as:

$$\begin{aligned} P(Z_{(pn),d} \le z) &= P(\max_{1\le i<j\le p} Z_{ij,d} \le z) \\ &= P\left(\sum_{1\le i<j\le p} Y_{ij} = 0\right) \\ &= P(W_2 = 0). \end{aligned}$$

In this proof, we have the following calculations for b_1, b_2 and b_3 defined by (3), (4) and (5):

$$
\begin{aligned}
b_1 &= \sum_{\alpha_2 \in I_2} \sum_{\beta_2 \in N_{\alpha_2}} p_{\alpha_2} p_{\beta_2} = \frac{p(p-1)}{2} \times (2p-1) p_{\alpha_2}^2 = \frac{4p-2}{p(p-1)} \lambda_2^2, \\
b_2 &= \sum_{\alpha_2 \in I_2} \sum_{\beta_2 : \alpha_2 \neq \beta_2 \in N_{\alpha_2}} p_{\alpha_2 \beta_2}, \\
b_3 &= 0,
\end{aligned}
$$

and to calculate an upper bound on b_2, we first calculate $p_{\alpha_2\beta_2} = E[Y_{\alpha_2} Y_{\beta_2}] = E[Y_{ij} Y_{ik}]$:

$$
\begin{aligned}
E[Y_{ij} Y_{ik}] &= E[I(Z_{ij} > z, Z_{ik} > z)] \\
&= P(\max_{i \in \{1,\ldots,m\}} \frac{|S_i|}{\sqrt{n}} > z, \max_{j \in \{m+1,\ldots,2m\}} \frac{|S'_j|}{\sqrt{n}} > z) \\
&= P([\cup_{i \in \{1,\ldots,m\}} \{\frac{|S_i|}{\sqrt{n}} > z\}] \cap [\cup_{j \in \{m+1,\ldots,2m\}} \{\frac{|S'_j|}{\sqrt{n}} > z\}]) \\
&= P(\cup_{i \in \{1,\ldots,m\}} [\{\frac{|S_i|}{\sqrt{n}} > z\} \cap [\cup_{j \in \{m+1,\ldots,2m\}} \{\frac{|S'_j|}{\sqrt{n}} > z\}]]) \\
&= P(\cup_{i \in \{1,\ldots,m\}} \cup_{j \in \{m+1,\ldots,2m\}} [\{\frac{|S_i|}{\sqrt{n}} > z\} \cap \{\frac{|S'_j|}{\sqrt{n}} > z\}]) \\
&\leq \sum_{i \in \{1,\ldots,m\}} \sum_{j \in \{m+1,\ldots,2m\}} P(\frac{|S_i|}{\sqrt{n}} > z, \frac{|S'_j|}{\sqrt{n}} > z) \\
&= \sum_{i \in \{1,\ldots,m\}} \sum_{j \in \{m+1,\ldots,2m\}} P(\frac{|S_i|}{\sqrt{n}} > z) P(\frac{|S'_j|}{\sqrt{n}} > z) \\
&= m^2 p_{\alpha_1}^2,
\end{aligned}
$$

where $p_{\alpha_1} = 2P(S_k \geq z\sqrt{n} + 1) = 2g(a,n)(1 + o(1))$ as $n \to \infty$. Thus we get upper bounds for b_2 and $b_1 + b_2$:

$$
\begin{aligned}
b_2 &= \sum_{\alpha_2 \in I_2} \sum_{\beta_2 : \alpha_2 \neq \beta_2 \in N_{\alpha_2}} p_{\alpha_2 \beta_2} \\
&\leq \sum_{\alpha_2 \in I_2} \sum_{\beta_2 : \alpha_2 \neq \beta_2 \in N_{\alpha_2}} m^2 p_{\alpha_1}^2
\end{aligned}
$$

$$= \frac{p(p-1)}{2} \times 2(p-1) \times m^2 p_{\alpha_1}^2,$$

hence

$$\begin{aligned} b_1 + b_2 &= \frac{p(p-1)}{2} \times (2p-1)p_{\alpha_2}^2 + \frac{p(p-1)}{2} \times 2(p-1) \times m^2 p_{\alpha_1}^2 \\ &< p^3(p_{\alpha_2}^2 + m^2 p_{\alpha_1}^2) \\ &\leq p^3((1-e^{-\lambda_1} + \lambda_1^2/m)^2 + \lambda_1^2) \\ &\leq 2p^3\lambda_1^2. \end{aligned}$$

After calculating the upper bound on the $b_1 + b_2 + b_3$, according to Lemma 3,

$$|P(W_2 = 0) - e^{-\lambda_2}| \leq (b_1 + b_2 + b_3)(1 - e^{-\lambda_2})/\lambda_2 < (1\Lambda\lambda_2^{-1})(b_1 + b_2 + b_3),$$

and since $P(Z_{(pn),d} \leq z) = P(W_2 = 0)$ from the above equation, we get

$$|P(Z_{(pn),d} \leq z) - e^{-\lambda_2}| < 2p^3\lambda_1^2,$$

where

$$\begin{aligned} \lambda_1 &= 2mg(a,n)(1+o(1)) \; when \; n \to \infty, \\ \lambda_2 &= \frac{p(p-1)}{2}[1 - e^{-2mg(a,n)}](1+o(1)) \; when \; n \to \infty, \\ g(a,n) &= \frac{1}{1-r}\frac{1}{\sqrt{2\pi a(1-a)n}}e^{-nH}, \\ a &= \frac{z\sqrt{n} + n + 1}{2n}, \\ r &= \frac{1-a}{a}, \\ H &= a\log(2a) + (1-a)\log(2-2a). \end{aligned}$$

Now we calculate the error bound as follows:

$$\begin{aligned} 2p^3\lambda_1^2 &= 8p^3m^2\frac{1}{(1-r)^2}\frac{e^{-2nH}}{2\pi a(1-a)n}(1+o(1)) \\ &= 8p^3m^2\frac{a^2}{(2a-1)^2}\frac{e^{-2nH}}{2\pi a(1-a)n}(1+o(1)) \end{aligned}$$

$$= \frac{4p^3m^2}{\pi}\frac{ae^{-2nH}}{(2a-1)^2(1-a)n}(1+o(1)).$$

Since the exponential term is:

$$\begin{aligned} e^{-2nH} &= e^{-2n(\log[(2a)^a]+\log[(2-2a)^{1-a}])} \\ &= e^{-2n\log[2a^a(1-a)^{1-a}]} \\ &= [2a^a(1-a)^{1-a}]^{-2n}, \end{aligned}$$

we have the error bound as follows:

$$2p^3\lambda_1^2 = \frac{4p^3m^2}{\pi}\frac{a}{(2a-1)^2n}\frac{[2a^a(1-a)^{1-a}]^{-2n}}{(1-a)}(1+o(1))$$

and since $\frac{[2a^a(1-a)^{1-a}]^{-2n}}{(1-a)} \in (0,2)$, we get the upper bound of the error term as:

$$\begin{aligned} 2p^3\lambda_1^2 &\le \frac{4p^3m^2}{\pi}\frac{a}{(2a-1)^2n}(1+o(1))\times 2 \\ &= \frac{4p^3m^2}{\pi}\frac{z\sqrt{n}+n+1}{2(z\sqrt{n}+1)^2}(1+o(1))\times 2 \\ &= \frac{4p^3m^2}{\pi}\frac{z\sqrt{n}+n+1}{(z\sqrt{n}+1)^2}(1+o(1)). \end{aligned}$$

Now we calculate, for the condition $p = O(n^\delta)$, the region of δ: consider

$$2p^3\lambda_1^2 \le \frac{4p^3m^2}{\pi}\frac{z\sqrt{n}+n+1}{(z\sqrt{n}+1)^2}(1+o(1))$$

and

$$z\sqrt{n}+n+1 < 2n,\ i.e.,\ z < \frac{n-1}{\sqrt{n}} < \sqrt{n},$$

we have that p satisfies: at least there exists a $\zeta > 0$,

$$\begin{aligned} p^3 n &= O(n^{2-\zeta}), \\ \Rightarrow p^3 &= O(n^{1-\zeta}), \\ \Rightarrow p &= O(n^{1/3-\zeta/3}). \end{aligned}$$

Thus, $p = O(n^{\delta}),\ where\ \delta \in (0, 1/3)$.

We take $z = \frac{p^4}{n^{4\delta-1/2}}t$, where t is a constant, then we get: for the error term,

$$\begin{aligned} 2p^3\lambda_1^2 &\leq \frac{4p^3m^2}{\pi}\frac{p^4t/n^{4\delta-1}+n+1}{p^8t^2/n^{8\delta-2}+2p^4t/n^{4\delta-1}+1}(1+o(1)) \\ &= \frac{4m^2}{\pi}\frac{p^7tn^{4\delta-1}+p^3n^{8\delta-1}+p^3n^{8\delta-2}}{p^8t^2+2p^4tn^{4\delta-1}+n^{8\delta-2}}(1+o(1)) \\ &= \frac{4m^2}{\pi}O(\frac{1}{p}) \end{aligned}$$

and

$$\begin{aligned} 2p^3\lambda_1^2 &\leq \frac{4p^3m^2}{\pi}\frac{p^4t/n^{4\delta-1}+n+1}{p^8t^2/n^{8\delta-2}+2p^4t/n^{4\delta-1}+1}(1+o(1)) \\ &= \frac{4m^2}{\pi}\frac{p^7tn^{4\delta-1}+p^3n^{8\delta-1}+p^3n^{8\delta-2}}{p^8t^2+2p^4tn^{4\delta-1}+n^{8\delta-2}}(1+o(1)) \\ &= \frac{4m^2}{\pi}\frac{O(n^{11\delta-1})+O(n^{11\delta-1})+O(n^{11\delta-2})}{O(n^{8\delta})+O(n^{8\delta-1})+O(n^{8\delta-2})}(1+o(1)) \\ &= \frac{4m^2}{\pi}O(\frac{1}{n^{1-3\delta}}). \end{aligned}$$

Finally we get when $p = O(n^{\delta})$, where $\delta \in (0, 1/3)$,

$$\lim_{n\to\infty}|P(Z_{(pn),d} \leq \frac{p^4}{n^{4\delta-1/2}}t) - e^{-mp(p-1)\frac{a}{2a-1}\frac{[2a^a(1-a)^{1-a}]^{-n}}{\sqrt{2\pi a(1-a)n}}}| = 0,$$

where

$$\begin{aligned} m &= (2^d-1)^2 \\ a(t,p,n) &= \frac{p^4t/n^{4\delta-1}+n+1}{2n}. \end{aligned}$$

4 Significant Pairs Selection using the BET Z-score

As described in Section 1, when applying BET in a large-scale data set, the Bonferroni adjustment based on the BET p-value could be a conservative method to select the interesting nonlinear dependence pairs. While the scope of this paper is purely theoretical, deeper investigation using simulations

would be very interesting; for example, one could explore the power comparison of the maximal BET z-scores with the Bonferroni correction and that with the proposed extreme value theory by replicating the common dependency structures such as linear, parabolic, circular, sine, checkerboard as in Zhang (2019). Based on the theorem of the extreme value distribution of the maximal BET z-statistic (Theorem 4), we reject the independence null hypothesis when the observed z-statistic is large compared to the extreme value distribution. In particular, the BET z-score is a meaningful statistic to select the pairs of variables which tend to be nonlinearly dependent. For a data set with p variables $X_1, X_2, .., X_p$ and n independent observations where $p = O(n^\delta), \delta \in (0, 1/3)$, after giving a fixed choice of the BET depth parameters $d_1 = d_2 = d$, we identify the set of variable pairs that are potentially nonlinearly dependent as:

$$T = \{(i, j) : i < j \ and \ Z_{ij,d} \geq z_\alpha\}, \tag{6}$$

where $Z_{ij,d}$ is the z-statistic of BET between X_i and X_j, and z_α is the $100(1-\alpha)\%$ quantile of the distribution given in Theorem 4 with the significance level α. Larger α results in a more conservative test which means fewer significant pairs are selected in T.

5 Discussion

In this paper, we have developed a limiting distribution theory for the maximal BET z-score under the condition $p = O(n^\delta), \delta \in (0, 1/3)$. A limitation of this theory is demonstrated by the TCGA data analyzed in Xiang et al. (2022), which is an example data set that has a large number of variables (16615 genes) and only 817 samples. Thus the theorem of the maximal BET z-statistic distribution (Theorem 4) is less relevant for this TCGA data set. An open problem is extending the condition $p = O(n^\delta), \delta \in (0, 1/3)$ to analyze the behavior of BET for larger data sets or even the high-dimension low-sample size domain (HDLSS) (Hall et al., 2005), i.e., finding the distribution of the maximal BET z-score under the condition $p = O(n^\delta),\ \delta \in (0, \infty)$. We would also like to mention that the fast algorithm for BET symmetry statistics with bitwise operations is proposed in Zhang et al. (2023). More specifically, these are different domains of asymptotic, and different high dimensional theories could be considered: first, we should consider the condition $p = O(n^\delta), \delta \in (0, 1)$; second, it would be interesting to pursue random matrix theory (Tracy and Widom (2002); for a useful introduction, see Bai and Silverstein (2010) and Tao (2012)), which helps solve this question under

the situation $p = n$; third, the HDLSS domain is another interesting open problem, i.e., we want to solve this extreme value distribution question when $p \gg n$ (Zhang, 2017).

Such generalizations could extend the application of the significant pairs selection using the BET z-score into various important real data situations. For additional future work, this significant pairs selection approach based on the BET z-scores can be extended to the variable selections in different models, such as regression models or neural networks, which involve the influence of nonlinear relationships between predictors.

Acknowledgements. The results published here are in whole or part based upon data from the Cancer Genome Atlas managed by the NCI and NHGRI (dbGaP accession phs000178).
Xiang's research was supported by NIH/NIAMS Grants P30AR072580 and R21AR074685, and DMS-2152289 from NSF.
Zhang's research was partially supported by DMS-1613112, IIS-1633212, DMS-1916237, and DMS-2152289 from NSF.
Marron's research was supported by NSF Grants IIS-1633074 and DMS-2113404.

Declarations

Competing interest Authors have no financial interest directly or indirectly related to this work.

References

Arratia, R. and Gordon, L. (1989). Tutorial on large deviations for the binomial distribution. *Bull. Math. Biol.*, **51**, 125–131. https://doi.org/10.1007/BF02458840

Arratia, R., Goldstein, L. and Gordon, L. (1990). Poisson approximation and the chen-stein method. *Stat. Sci.*, 403–424.

Bai, Z. and Silverstein, J.W. (2010). Spectral analysis of large dimensional random matrices, vol. 20. Springer.

Basu, D. (1958). On statistics independent of sufficient statistics. Sankhyā: *Indian J Statist. (1933-1960)*, **20**, 223–226. http://www.jstor.org/stable/25048393.

Hall, P., Marron, J.S. and Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **67**, 427–444.

Tao, T. (2012). Topics in random matrix theory, vol. 132. American Mathematical Society.

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Tracy, C.A. and Widom, H. (2002) Distribution functions for largest eigenvalues and their applications. https://doi.org/10.48550/ARXIV.MATH-PH/0210034.

Xiang, S., Zhang, W. and Liu, S. et al. (2022). Pairwise nonlinear dependence analysis of genomic data. arXiv:2202.09880

Zhang, K. (2017). Spherical cap packing asymptotics and rank-extreme detection. *IEEE Trans. Inform. Theory*, **63**, 4572–4584.

Zhang, K. (2019). BET on independence. *J Amer. Statist. Assoc.*, **114**, 1620–1637. https://doi.org/10.1080/01621459.2018.1537921.

Zhang, W., Zhao, Z. and Baiocchi, M. et al. (2023). SorbET: A fast and powerful algorithm to test dependence of variables

Siqi Xiang, Wan Zhang, Kai Zhang and J. S. Marron
Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

E-mail: xiangsiqi.unc@gmail.com
E-mail: wan.zhang@unc.edu
E-mail: zhangk@email.unc.edu
E-mail: marron@unc.edu

Paper received: 15 May 2023