Testing Conditional Independence with Deep Neural Network Based Binary Expansion Testing (DeepBET)

Yang Yang
University of Illinois Chicago

Kai Zhang
University of North Carolina at at Chapel Hill

Ping-Shou Zhong
University of Illinois Chicago

Abstract

This paper focuses on testing conditional independence between two random variables (X and Y) given a set of high-dimensional confounding variables (Z). The high dimensionality of these confounding variables presents a challenge, often resulting in inflated type-I errors or insufficient power in many existing tests. To address this issue, we leverage the power of Deep Neural Networks (DNNs) to handle complex, high-dimensional data while mitigating the curse of dimensionality. We propose a novel test procedure, DeepBET. First, a DNN is used on part of the data to estimate the conditional means of X and Y given Z. Then, binary expansion testing (BET) are applied to the predicted errors from the remaining data. Additionally, we implement a multiple-split procedure to further enhance the power of the test. DeepBET is computationally efficient and robust to the tuning parameters in DNNs.

Interestingly, the DeepBET statistic converges at a root-n rate despite the nonparametric and high-dimensional nature of the confounding effects.

Our numerical results demonstrate that the proposed method controls type-I error under various scenarios and enhances both power and interpretability for conditional dependence when present, making it a robust alternative for testing conditional independence in high-dimensional settings. When applied to dry eye disease data, DeepBET reveals

meaningful nonlinear relationships between the epithelial thickness and the tear production in the central region of eyes, given other regions.

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

1 Introduction

Conditional independence (CI) is a fundamental concept in statistics, machine learning, and artificial intelligence. Testing for CI has important applications in various statistical problems, such as causal inference [Pearl, 2009], feature selection [Cai et al., 2018], and dimension reduction [Waggoner, 2021]. The widespread demand for CI tests stems from the need to identify relationships between different objects, including understanding how they are connected, the mechanisms through which they interact, and how information flows between them. CI tests are now extensively used across numerous disciplines, including information extraction, speech recognition, computer vision, gene discovery, and disease diagnosis.

The primary objective of this article is to develop a testing procedure to assess whether two random variables, X and Y, are conditionally independent given a set of potentially high-dimensional confounding variables, Z. If the joint density of X, Y, Z exists, then we write $X \perp\!\!\!\perp Y|Z$ if $f(x, y|z) = f(x|z)f(y|z)$ for all x, y, z with $f(z) > 0$ where $f(x, y|z)$ is the joint density of X, Y given Z, $f(x|z)$ and $f(y|z)$ are, respectively, marginal densities of X and Y given Z (e.g.,[Shah and Peters, 2020]). We consider the following hypothesis testing problem:

$$H_0 : X \perp\!\!\!\perp Y|Z \quad \text{versus} \quad H_1 : X \not\perp\!\!\!\perp Y|Z,$$

where $\perp\!\!\!\perp$ denotes independence and '|' denotes conditioning. Our focus is on high-dimensional confounding variables Z, where the dimensionality d can potentially diverge to infinity as the sample size n increases. A key challenge arises from the "curse of dimensionality" in estimating the unknown and unstructured conditional distributions $f(X|Z)$ and $g(Y|Z)$, especially when the sample size is small relative to the dimensionality of Z.

Significant research has been conducted on conditional independence testing, with a comprehensive review available in [Li and Fan, 2020]. These methods

generally fall into four main categories: distance-based tests (e.g.,[Su and White, 2007, Su and White, 2014, Wang et al., 2015, Su and White, 2008]), kernel-based tests (e.g., [Fukumizu et al., 2007, Zhang et al., 2011, Strobl et al., 2017]), regression-based tests (e.g., [Hoyer et al., 2008, Zhang et al., 2023, Shah and Peters, 2020, Zhang et al., 2017]), and sampling-based tests (e.g., [Bellot and van der Schaar, 2019, Shi et al., 2021, Duong and Nguyen, 2022]). Distance-based methods typically compare the product of the conditional marginal distributions of X given Z and Y given Z with the conditional joint distribution of (X, Y) given Z. The performance of these methods depends on the choice of the distance measure and the estimators used for conditional distributions. Regression-based methods, on the other hand, rely on the conditional correlation between X and Y given Z, but it is well-known that uncorrelation does not imply independence, which means these methods may fail to detect conditional dependence in uncorrelated alternatives. Kernel-based tests extend the linear correlation of regression-based methods to non-linear correlations by utilizing the kernel trick. However, the effectiveness of detecting non-linear dependencies depends on the choice of kernel functions. Finally, sampling-based methods, such as those using Generative Adversarial Networks (GANs), generate synthetic samples from the conditional distributions given the confounding variables Z. GANs offer a flexible approach to approximating conditional distributions in high-dimensional settings.

The approaches mentioned above have at least one of the following limitations: inconsistency, meaning they lack power to detect certain alternatives; not appropriate to handle high-dimensional confounding variables, due to the "curse of dimensionality"; some sampling based methods are not computational efficient or too sensitive to tuning parameter choices. As a result, many existing tests suffer from inflated type-I errors, insufficient power to detect alternatives, or computational complexity and instability due to parameter tuning. Furthermore, a limitation of existing methods is their lack of interpretability. They often function as a black box, providing statistical significance without identifying the nature of the relationship. These challenges highlight the need for more robust, consistent, scalable, and interpretable testing procedures capable of handling complex, high-dimensional datasets. Our proposed method, DeepBET, addresses all of these challenges by leveraging the power of deep neural networks (DNNs) in combination with innovative binary expansion testing (BET) statistical techniques.

Our proposed DeepBET procedure combines the strength of three key components: DNN, BET, and multiple-splitting. First, to overcome the "curse of dimensionality," we implement deep neural network (DNN) models to estimate the conditional means of X given Z and Y given Z, respectively. DNNs [Polson and Sokolov, 2018] have proven effective in capturing nonlinear relationships in high-dimensional and complex data. Recent studies (e.g., [Lu et al., 2021, Bauer and Kohler, 2019, Schmidt-Hieber, 2020]) have explored the consistency and convergence rates of certain types DNN estimators for nonparametric functions involving high-dimensional confounding variables Z when appropriate optimization strategies are chosen (e.g.,[Jentzen and Riekert, 2022]). Second, we use the innovative nonparametric binary expansion testing (BET) [Zhang, 2019] to construct test statistics. BET has been shown [Zhang, 2019] to achieve uniform consistency and attain the minimax rate in terms of sample size requirements for reliable power. Notably, BET also offers clear insights into the conditional dependence relationships when independence is rejected. Finally, we employ a multi-split method [Guo and Shah, 2024] to enhance the power of our test. Our proposed method effectively controls type I error and demonstrates a strong ability to detect alternatives, making it a robust approach for testing conditional independence in the presence of high-dimensional confounding variables.

Our numerical studies demonstrate that the proposed DeepBET procedure outperforms existing methods in terms of power while effectively controlling the type I error, particularly in the presence of high-dimensional confounding variables. The performance of DeepBET is robust to the choice of tuning parameters in DNN fitting. More importantly, DeepBET is computationally efficient compared to existing sampling-based approaches (e.g., [Shi et al., 2021, Bellot and van der Schaar, 2019]). In addition to detecting dependence, DeepBET provides valuable insights into the form of the conditional dependence, which is helpful for subsequent modeling steps.

The remainder of the article is organized as follows: Section 2 provides a detailed explanation of the DeepBET testing procedure and establishes the asymptotic distribution of the proposed statistics. Section 3 presents the simulation results and includes a case study on corneal epithelial thickness data related to dry eye disease. Section 4 concludes the paper. The Appendix contains all the technical proofs.

2 A DeepBET procedure

Assume we observe independent and identically distributed samples (X_i, Y_i, Z_i) for $i \in \mathcal{S} = \{1, \ldots, n\}$, generated under additive noise models (ANM): $X_i = h(Z_i) + \epsilon_i$ and $Y_i = g(Z_i) + \nu_i$, where $h(\cdot)$ and $g(\cdot)$ are arbitrary unknown functions of the high-dimensional variables Z_i. The variables X_i and Y_i are univariate,

and ϵ_i and ν_i are random errors that are independent of Z_i. The ANM are assumed for the convenience of theoretical exploration of the asymptotic null distributions of the proposed DeepBET statistics. Our empirical simulation studies in Section 3.2 have investigated the performance of DeepBET when data are not generated from the ANM.

We begin by randomly splitting the indices $\mathcal{S}$ into two non-overlapping parts, denoted as $\mathcal{D}_1$ and $\mathcal{D}_2$, such that $\mathcal{D}_1 \cup \mathcal{D}_2 = \mathcal{S}$ and $\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$. Using the first subset, $\mathcal{D}_1$, we apply deep neural networks (DNN) [Lee, 2021] to estimate the conditional means of X given Z and Y given Z. Test statistics are then constructed using the second subset, $\mathcal{D}_2$.

Specifically, we employ DNNs to estimate the conditional expectations $h(Z_j) := E(Y_j|Z_j)$ and $g(Z_j) := E(X_j|Z_j)$ for $j \in \mathcal{D}_1$. A DNN is an artificial neural network architecture composed of multiple layers, with each layer containing interconnected nodes that transform input data using weighted sums and non-linear activation functions. This structure allows DNNs to effectively model complex relationships and capture hierarchical patterns in the data. Detailed architecture used in DeepBET is given in Section 3. Let the DNN estimators of the conditional expectations be denoted as $\hat{h}_{\mathcal{D}_1}(Z_j) = \hat{E}(X_j|Z_j)$ and $\hat{g}_{\mathcal{D}_1}(Z_j) = \hat{E}(Y_j|Z_j)$. We then use these estimators to predict the conditional expectations $X_i|Z_i$ and $Y_i|Z_i$ for $i \in \mathcal{D}_2$, denoted as $\hat{h}_{\mathcal{D}_1}(Z_i)$ and $\hat{g}_{\mathcal{D}_1}(Z_i)$, respectively.

We then compute the residuals $\hat{\epsilon}_i = X_i - \hat{h}_{\mathcal{D}_1}(Z_i)$ and $\hat{\nu}_i = Y_i - \hat{g}_{\mathcal{D}_1}(Z_i)$ for $i \in \mathcal{D}_2$, to approximate the random errors ϵ_i and ν_i. Next, we construct BET statistics to test the independence between $\hat{\epsilon}_i$ and $\hat{\nu}_i$. A graphical overview of our proposed testing procedure is provided in Figure 1.

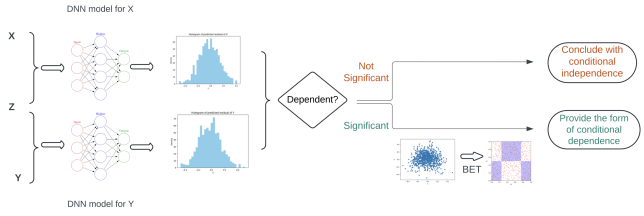

Figure 1: The flowchart of the proposed DeepBET procedure. First, we apply DNN to estimating the conditional mean of X and Y given Z with part of the data. Then, we obtain the residuals $\hat{\epsilon}$ and $\hat{\nu}$ from the remaining data. We implement BET in these residuals to verify conditional independence.

2.1 Binary expansion statistics

BET was first developed in [Zhang, 2019] to assess the independence between two random variables. The strength of BET lies in two key aspects. First, it achieves uniform consistency in distribution-free dependence detection and attains minimax optimal power. Second, the BET procedure can provide insights into the form of dependence if independence is rejected. However, BET is not directly applicable to conditional independence testing. This paper substantially extends the scope and applicability of BET to test conditional independence. One advantage of using binary expansion statistics is that it is computationally simple to obtain a binary expansion of a random variable. More importantly, for binary variables, uncorrelatedness is equivalent to independence, which is crucial for ensuring the consistency of our test procedure.

The following outlines the steps to construct marginal binary expansions. Assume we have n observations from the bivariate random vectors (U_i, V_i) whose marginal distributions are uniform over $[0, 1]$. In practice, if the marginal distributions are not uniform. We can apply transformations so that $\tilde{U} = F_\epsilon(\epsilon)$ and $\tilde{V} = G_\nu(\nu)$ are uniformly distributed over $[0, 1]$, where $F_\epsilon(\cdot)$ and $G_\nu(\cdot)$ are marginal distributions of ϵ and ν. If marginal distributions are unknown, we can use the empirical CDFs so that for every observation i, $U_i = F_\epsilon(\epsilon_i)$ and $V_i = G_\nu(\nu_i)$ are each uniformly distributed over $\{1/n, \cdots, 1\}$. In this paper, we focus on testing the dependence of two continuous variables $\hat{\epsilon}$ and $\hat{\nu}$ and use the empirical CDFs to perform transformation, since we do not have information about the marginal distributions of $\hat{\epsilon}$ and $\hat{\nu}$.

Consider the binary expansions of U_i and V_i giving by the following:

$$\hat{U}_i = \sum_{k=1}^{\infty}(\hat{A}_i^{(k)} + 1)/2^{k+1}, \hat{V}_i = \sum_{k=1}^{\infty}(\hat{B}_i^{(k)} + 1)/2^{k+1}$$

where $\hat{A}_i^{(k)}, \hat{B}_i^{(k)}$ are random coefficients that are Bernoulli distributed taking values in $\{-1, +1\}$. We truncate the expansion of U_i and V_i at some certain given depth d_1 and d_2 and define

$$\hat{U}_{i,d} = \sum_{k=1}^{d_1}(\hat{A}_i^{(k)} + 1)/2^{k+1} \text{ and } \hat{V}_{i,d} = \sum_{k=1}^{d_2}(\hat{B}_i^{(k)} + 1)/2^{k+1}.$$

For each d_1 and d_2, $\hat{U}_{d_1}$ and $\hat{V}_{d_2}$ are discrete multinomial variables taking 2^{d_1} and 2^{d_2} values, respectively. As $d_1, d_2 \to \infty$, $\hat{U}_{i,d_1} \to \hat{U}_i$ and $\hat{V}_{i,d_2} \to \hat{V}_i$ in probability. We could approximate the joint distribution of $(\hat{U}, \hat{V})$ through $(\hat{U}_{d_1}, \hat{V}_{d_2})$. We call the statistics that are functions of $\hat{A}_k$ and $\hat{B}_k$ BET, and we call the framework to test the independence of $(\hat{U}_{d_1}, \hat{V}_{d_2})$ the BET at depth d_1 and d_2.

Note that $(\hat{U}_{i,d_1}, \hat{V}_{i,d_2})$ can only take at most $2^{(d_1+d_2)}$ possible values, which lead to a partition of the range space $[0,1]\times[0,1]$ into a $2^{d_1} \times 2^{d_2}$ grids points. With this approach, truncating the binary expansion transforms the conditional dependence test problem into a problem defined over grid points (a contingency

table), where the cell probabilities are identifiable parameters.

To formally define the BET statistic, we introduce a filtration generated by $\{\hat{A}^{(k)}\}_{k=1}^{d_1}$ and $\{\hat{B}^{(k)}\}_{k=1}^{d_2}$ for every d_1 and d_2, which is $\sigma(U_{d_1}, V_{d_2})$, a σ-field formed by the binary variables $\{\hat{A}^{(1)}, \ldots, \hat{A}^{(d_1)}, \hat{B}^{(1)}, \ldots, \hat{B}^{(d_2)}\}$. Let $\mathbf{a}$ and $\mathbf{b}$ be indicator vectors with length d_1 and d_2, respectively. Let $\mathbf{ab}$ be one of the sets in the σ-field $\sigma(\hat{U}_{d_1}, \hat{V}_{d_2})$ and the entries are 0s and 1s in vectors $\mathbf{a}$ and $\mathbf{b}$ where 1 represents that the corresponding binary variable is selected in the set $\mathbf{ab}$. Define $\hat{\mathcal{A}}_i^{\mathbf{a}} = \prod_{k=1}^{d_1}\{\hat{A}_i^{(k)}\}^{a_k}$ and $\hat{\mathcal{B}}_i^{\mathbf{b}} = \prod_{k=1}^{d_2}\{\hat{B}_i^{(k)}\}^{b_k}$ for $i \in \mathcal{D}_2$. Then we obtain

$$\hat{S}_{(\mathbf{ab}),n} = \textstyle\sum_{i\in\mathcal{D}_2} \hat{\mathcal{A}}_i^{\mathbf{a}}\hat{\mathcal{B}}_i^{\mathbf{b}},$$

as a symmetry statistic which counts the difference between the number of data points with $\hat{\mathcal{A}}_i^{\mathbf{a}}\hat{\mathcal{B}}_i^{\mathbf{b}} = 1$ and $\hat{\mathcal{A}}_i^{\mathbf{a}}\hat{\mathcal{B}}_i^{\mathbf{b}} = -1$, for $\mathbf{a} \neq 0$ and $\mathbf{b} \neq 0$.

Consider the following example: the residual plots below are based on data generated from the nonlinear model described in Section 3.2. Visually detecting the dependence between X and Y in these plots is challenging.

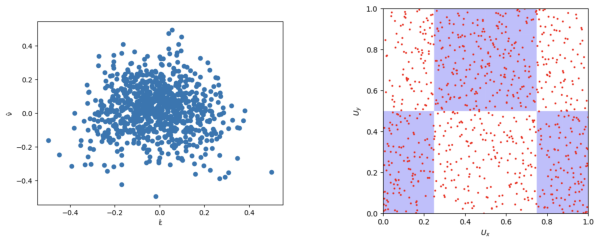

Figure 2: Left: Residual plot from $\mathcal{D}_2$, representing the curvilinear relationship between X and Y given Z. Right: The BET plot of residuals shows a significantly higher concentration of points in the blue region, corresponding to the interaction of the first bit from the Y residuals and the first two bits from the X residuals.

In addition to detecting dependence, the BET plot highlights a non-linear relationship, as more points concentrate in the blue area compared to the white area, revealing the conditional relationship between X and Y given Z. The DeepBET not only effectively captures the relationship between X and Y but also provides a clear visual representation of this dependence.

Theorem 1 below provides the asymptotic distribution of the estimated BET statistic $\hat{S}_{(\mathbf{ab}),n}$.

Theorem 1. *Assume the functions $h(\cdot)$ and $g(\cdot)$ are (p, C)-smooth and satisfy the generalized hierarchical model of order d_* defined in the Appendix. Assume the conditions in Lemma 1 in the Appendix hold. The CDFs $F(\cdot)$ and $G(\cdot)$ have bounded derivatives. If $d_1 = o(n^{p/(2p+d_*)})$ and $d_2 = o(n^{p/(2p+d_*)})$, then $(\hat{S}_{(\mathbf{ab}),n} + n_{\mathcal{D}_2})/4$ and Hypergeometric$(n_{\mathcal{D}_2}, n_{\mathcal{D}_2}/2, n_{\mathcal{D}_2}/2)$ for $\mathbf{a} \neq 0$ and $\mathbf{b} \neq 0$ converge to the same limiting distribution where $n_{\mathcal{D}_2}$ is the sample size of $\mathcal{D}_2$.*

The significance of Theorem 1 is in justifying the application of DNN in the proposed procedure. Despite the convergence rate of DNN based estimators is slower than $n^{-1/2}$ due to its nonparametric nature and high dimensionality of the confounding variables Z, the asymptotic distribution of $S_{(\mathbf{ab}),n}$ is still root-n consistent, which is mainly due to the binary nature of the random variables $\hat{\mathcal{A}}_i$ and $\hat{\mathcal{B}}_i$. As a comparison, the generative conditional independence test (GCIT) of [Bellot and van der Schaar, 2019] requires the total variation distance between the estimated conditional distribution and the underlying true conditional distribution converges at a faster rate than $n^{-1/2}$. The double GANs for conditional independence test (DGCIT) of [Shi et al., 2021] relaxes the condition and requires that the product of two total variation distances (for conditional distributions for $X|Z$ and $Y|Z$) is a small order of $n^{-1/2}$. However, these conditions required for GCIT and DGCIT may not be satisfied if the underlying dimension d_* in Theorem 1 is large and no specific smoothness or sparsity assumptions on the nonparametric functions. See [Shah and Peters, 2020] for more discussion of using nonparametric functions of high-dimensional confounding variables in conditional independence tests.

Based on Theorem 1, it is not difficult to see that $(n_{\mathcal{D}_2} - 1)^{1/2}\hat{S}_{(\mathbf{ab}),n}/n_{\mathcal{D}_2}$ is asymptotically normal. To maximize the power, we then construct a maximal normalized BET test statistic as following:

$$T_n = \max_{\mathbf{ab}\in\sigma(\hat{U}_{d_1},\hat{V}_{d_2}),\mathbf{a}\neq 0,\mathbf{b}\neq 0} (n_{\mathcal{D}_2} - 1)^{1/2}\hat{S}_{(\mathbf{ab}),n}/n_{\mathcal{D}_2}.$$

Theorem 2. *Assume the conditions in Theorem 1 hold and $d_1 + d_2 = o\{\log_2(n^{p/(2p+d_*)})\}$. As $n \to \infty$, the statistic $\hat{T}_n$ has the same asymptotic distribution as the maximum of a Gaussian distribution with mean zero and identity covariance under the H_0.*

The complete proof is given in the Appendix.

2.2 Multiple data splitting with rank-transformed subsampling

Although a single-split approach is simple and computationally efficient, it has its limitations. First, different random splits of the same dataset can lead to varying results. Second, the test tends to lose power because it does not fully utilize the entire sample. To address these issues, we adopt a multiple-splitting technique [Guo and Shah, 2024], which combines the results of multiple randomized tests to reduce randomness and enhance the power of the test.

2.2.1 Setup

Let $T_n^{(1)}, \cdots, T_n^{(L)}$ be test statistics that can be computed from samples in $\mathcal{D}_2$, where L is pre-specified ways

of splitting the testing sample. Consider the aggregated or "multiple-split" statistic

$$\Lambda_n = (|T_n^{(1)}| + \cdots + |T_n^{(L)}|)/L,$$

where $T_n^{(l)}$ is the l-th BET statistic based on the l-th sample split for $l = 1, \cdots, L$. By taking L reasonable large, we can expect that $\text{var}(\Lambda_n | X_1, \cdots, X_n)$ is small enough such that the aggregated test statistic is effectively de-randomized.

Based on Theorem 2, Λ_n converges to some distribution G_p under the null; Our aggregated test rejects H_0 for large values of Λ_n. We aim to mimic an oracle procedure that rejects whenever Λ_n exceeds an unknown upper α quantile of G_P. We use subsampling to compute $\tilde{G}_n$ to approximate G_p, and use its quantile to determine the critical values for Λ_n.

2.2.2 Rank-transformed subsampling

In this section, we describe our procedure when using a single aggregation function Λ_n. We start with our subsampling setup.

Subsampling Let $m < n$ be a subsample size. All the numerical experiments in this paper, we use $m = [n/\log_n]$. We randomly select a total of B sets of indices, each of size m, such that there is a sufficiently low degree of overlap among the sets. First of all, we choose a positive integer (e.g., $J = 50$) and let $B := J[n/m]$. Then our collection of set indices $\mathcal{B} := \{(i_{1,b}, \cdots, i_{m,b}) : b = 1, \cdots, B\}$ is formed using Algorithm 1 [Guo and Shah, 2024] below.

Algorithm 1 Generate ordered tuples

Input: Sample size n, subsample size m, positive integer J.

1: $\mathcal{B} \leftarrow \emptyset$.
2: **for** $j = 1, \cdots, J$ **do**
  $\pi \leftarrow$ a random permutation of $\{1, \cdots, n\}$.
  $\mathcal{B} \leftarrow \mathcal{B} \bigcup \{(\pi_1, \cdots, \pi_m), (\pi_{m+1}, \cdots, \pi_{2m}), \cdots, (\pi_{([n/m]-1)m+1}, \cdots, \pi_{[n/m]m})\}$.
3: **return** $\mathcal{B}$

We arrange the subsampled test statistics into a $B \times L$ matrix $\widehat{H} = (\widehat{H}_{b,l})$ consisting of rows:

$$\widehat{H}_{b,\cdot} := (T_m^{(1)}(X_{i_{1,b}}, \cdots, X_{i_{m,b}}), \cdots, T_m^{(L)}(X_{i_{1,b}}, \cdots, X_{i_{m,b}})), \qquad b = 1, \cdots, B.$$

Then, we apply the aggregation function Λ_n to each row of $\widehat{H}$, we could obtain $\widehat{\Lambda}_b := \Lambda_n(\widehat{H}_{b,1}, \cdots, \widehat{H}_{b,L})$, whose empirical distribution function $\widehat{G}_n(x) := \mathbb{F}_{\{\widehat{\Lambda}_b\}}(x)$ is the natural subsampling estimate for $G_p(x)$

Rank transform Let $\mathbb{F}_{\widehat{H}}$ denote the empirical distribution function based on entries of $\widehat{H}$. With this, we form a rank-transformed version of $\widehat{H}$ denoted $\tilde{H} = (\tilde{H}_{b,l})$ filled with entries

$$\tilde{H}_{b,l} := F_0^{-1}\{(R_{b,l} - 1/2)/(BL)\},$$

where $R_{b,l}$ be the rank of $T_m^{(l)}(\mathcal{D}_2^{(b)})$ among the statistics $\{T_m^{(l)}(\mathcal{D}_2^{(b)})\}$ for $b = 1, \cdots, B$ and $l = 1, \cdots, L$; F_0 is chosen as the CDF of a standard normal distribution. We then compute the aggregated statistics

$$\tilde{\Lambda}_b := \Lambda_n(\tilde{H}_{b,1}, \cdots, \tilde{H}_{b,L}),$$

and their resulting empirical distribution function $\tilde{G}_n := \mathbb{F}_{\{\tilde{\Lambda}_b\}}(x)$, which we then use to determine the critical value $\tilde{G}_n^{-1}(1-\alpha)$ for Λ_n. The full procedure is given in Algorithm 2 [Guo and Shah, 2024].

Algorithm 2 Aggregated multiple-split test

Input: Data $X_1, \cdots, X_n$, exchangeable single-split test statistics $(T_n^{(1)}, \cdots, T_n^{(L)})$, asymptotic null distribution function F_0, aggregation function Λ_n, significance level $\alpha \in (0, 1)$, positive integer J.

1: $m \leftarrow [n/\log_n], B \leftarrow J[n/m]$
2: Run Algorithm 1 to obtain $\mathcal{B} = \{(i_{1,b}, \cdots, i_{m,b} : b = 1, \cdots, B)\}$
3: Initialise $B \times L$ matrices $\widehat{H}, \tilde{H}$ and B-dimensional vector $\tilde{\Lambda}$
4: **for** $b = 1, \ldots, B$ **do**
  $\widehat{H}_{b,\cdot} \leftarrow (T_m^{(1)}(X_{i_{1,b}}, \ldots, X_{i_{m,b}}), \ldots, T_m^{(L)}(X_{i_{1,b}}, \ldots, X_{i_{m,b}})$
5: **for** $b = 1, \ldots, B$ **do**
  for $l = 1, \cdots, L$ **do**
    $\tilde{H}_{b,l} \leftarrow F_0^{-1}\{(R_{b,l} - 1/2)/(BL)\}$
  $\tilde{\Lambda}_b \leftarrow \Lambda_n(|\tilde{H}_{b,1}|, \cdots, |\tilde{H}_{b,L}|)$
6: $\tilde{G}_n \leftarrow \mathbb{F}_{\{\tilde{S}_b\}}$
7: Compute $\Lambda_n \leftarrow \Lambda_n(|T_n^{(1)}|, \cdots, |T_n^{(L)}|)$ from $X_1, \cdots, X_n$
8: Reject H_0 if $\Lambda_n > \tilde{G}_n^{-1}(1-\alpha/2)$ and report p-value $1 - \tilde{G}_n(\Lambda_n)$

3 Numerical studies

We begin by discussing key implementation details. Next, we conduct simulations to evaluate the empirical size and power of the proposed test, comparing its performance with several alternative methods.

3.1 Details of Implementing DNN

There is a trade-off when selecting the number of splits, L, in Algorithm 2. While L should be as large as possible to ensure strong power, increasing L also raises computational complexity. Based on our empirical investigations, we found that setting L between 30 to 50 strikes a favorable balance between power and computational cost. Therefore, we set L= 40. Existing research (e.g., [Chen and Shen, 1998]; [Schmidt-Hieber, 2020]) has demonstrated that DNNs can consistently estimate nonparametric functions of high-dimensional covariates,

effectively overcoming the curse of dimensionality. However, the feasibility and applicability of DNNs to statistical inference problems, such as conditional independence tests, remain largely unexplored. We carefully examined and tuned DNNs to ensure their suitability for statistical inference tasks. Additionally, we studied the robustness of DNNs and the impact of tuning parameters on the proposed conditional independence test. We constructed fully connected neural networks and experimented with different activation functions, including LeakyReLU, ReLU, and Tanh. Furthermore, we split the data into training and testing sets in a 1: 4 ratio. We set reasonable ranges for key parameters such as the learning rate (0.005), batch size (50), and the number of training epochs (15). The Adam optimizer was utilized for parameter optimization. To improve the network's generalization capability and reduce overfitting, we applied the dropout technique. Specifically, we used a dropout rate of 0.2, randomly dropping 20% of neurons in each layer.

3.2 Simulations

Simulation setting I

We generate the data following the post nonlinear noise model similarly as in [Shi et al., 2021]; [Bellot and van der Schaar, 2019], i.e.,

$$X = \sin(a_f^T Z) + \epsilon_f, \text{ and } \quad Y = \cos(a_g^T Z + bX) + \nu_g.$$

The entries of a_f, a_g are randomly and uniformly sampled from [0,1], then normalized to the unit ℓ_1 norm. The noise variables ϵ_g, ν_g are independently sampled from a normal distribution with mean zero and variance 0.25. In this model, the parameter b determines the degree of conditional dependence. When b=0, H_0 holds, and otherwise H_1 holds. We set the significance level at $\alpha = 0.05$. All results are based on 500 simulation replications.

The left panel of Figure 3 compares the empirical size and power of the BET with the generalized correlation measure (GCM) ([Shah and Peters, 2020]), using predicted residuals from a deep neural network model with a single split, where the dimension of Z is d_Z = 100 and sample size is n=1000. We vary the value of b = 0, 0.45, 0.6, 0.75, 0.9. The results in the left panel of Figure3 demonstrate that DeepBET outperforms DNN + GCM in this simulation setting. In addition, we conducted further simulations that compare the deep neural network (DNN) with the random forest (RF) model. The right panel of Figure 3 presents empirical size and power results for $d_Z = 500$ and $n = 500$, using the same values as b. The right panel of Figure 3 shows that DeepBET consistently outperforms RF+BET in this setting.

In Figure 4, we compare the empirical size and power of the proposed single-splitting procedure with the multiple-splitting procedure, the dimension of Z

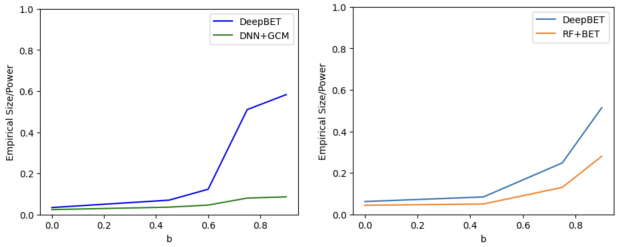

Figure 3: Left: The empirical size/power of the proposed test with DeepBET and DNN+GCM. Right: The empirical size/power of the DeepBET and RF+BET.

as $d_Z = 100$, and vary the value of b = 0, 0.45, 0.6, 0.75, 0.9. We note that multiple splits have controlled the type I error and significantly improved power. The power almost doubles when b approaches 0.9.

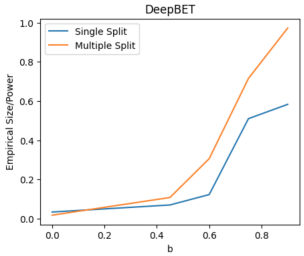

Figure 4: The empirical size/power of the proposed test with single- and multiple- splitting.

We aim to determine whether the choice of activation function significantly influences the results of our test. To investigate this, we experiment with various activation functions and compare their empirical performance. We vary the dimension of Z as d_Z = 50, 100, 150, 200, 250, generating Z from a standard normal distribution. Additional we compare the empirical power for d_Z= 100 while varying the value of b= 0.45, 0.6, 0.75, 0.9. Upon examining Figure 5, we observe minimal discrepancies in the results when using different activation functions. This suggests that our DeepBET is robust across various activation functions.

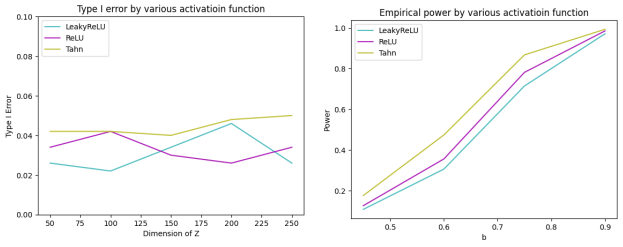

Figure 5: The empirical sizes and power of DeepBET are not sensitive to the choices of activation functions.

Figure 6 reports the empirical size when $b = 0$. We vary the dimension of Z as $d_Z = 50$, 100, 150, 200, 250, and consider two generation distributions of Z from a standard normal distribution and Laplace distribution. We compared two sample size, $n = 500$ and $n = 1000$. We compared our proposed test DeepBET with generative conditional independence test (GCIT) of [Bellot and van der Schaar, 2019], double generative adversarial networks for conditional independence test (DGCIT) of [Shi et al., 2021] and the classifier conditional independence test (CCIT) of [Sen et al., 2017]. We observe that DeepBET and GCIT effectively control the empirical size under the nominal level across all cases, while DGCIT exhibits inflated type-I errors in every instance. While DeepBET shows a slightly higher type-I error when the sample size is 500 and the dimension of $Z = 250$, CCIT experiences an initial increase in type-I error, followed by a decrease as the dimensionality of Z increases.

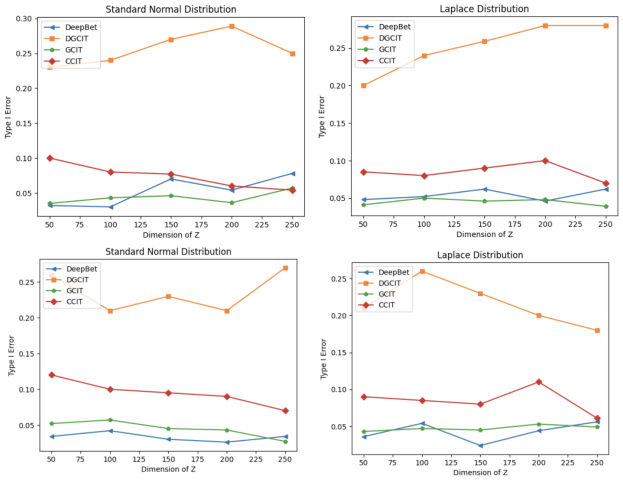

Figure 6: The empirical sizes of various tests under H_0. Left panels : Z is normal, right panels: Z is Laplacian. Top panels: sample size is 500, bottom panels: sample size is 1000. The size of DeepBET is around 0.05.

Figure 7 reports the empirical power with the dimension of Z as $d_Z = 100$, and vary the value of b = 0.45, 0.6, 0.75, 0.9. We compare two sample size, $n = 500$ and $n = 1000$. We observe that the empirical power of the DeepBET surpasses that of the competitors, converging to 1 as b increases to 0.9, demonstrating the consistency of the proposed test. In contrast, both GCIT and CCIT do not display sufficient power across all cases. Moreover, we also compare DeepBET with the kernel-based conditional distance correlation (KCDC) test in [Wang et al., 2015] when $n = 500$ and $d_Z = 100$. We found that the empirical size of KCDC was always 1, and was not able to control the type I error. This may indicate that KCDC is not applicable to conditional tests with high-dimensional confounding variables.

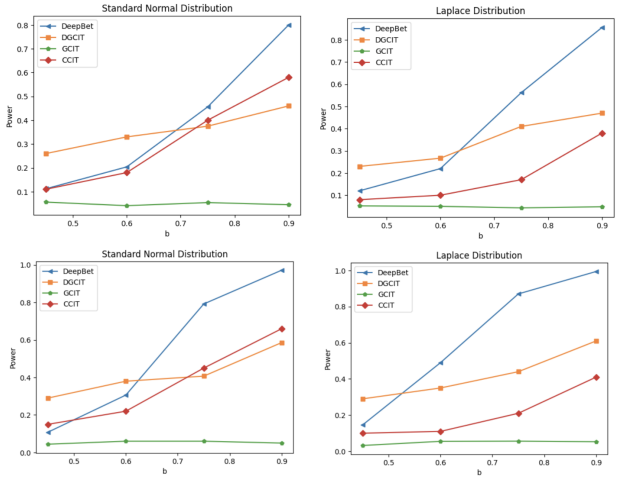

Figure 7: The empirical power of various tests under H_1. Left panels : Z is normal, right panels: Z is Laplacian. Top panels: sample size is 500, bottom panels: sample size is 1000. DeepBET has good power in all cases.

Lastly, we examine computational efficiency. All experiments were conducted on Google Colab using an Intel Xeon CPU with 2 vCPUs and a T4 GPU. The wall-clock time for executing the complete DeepBET test for one data replication was approximately 1.5 seconds. When employing multiple splits, the computation time increased to about 1 minute. In comparison, the running time for CCIT was around 2 minutes, for GCIT about 20 seconds, and for DGCIT, it extended to approximately 8 minutes.

Simulation setting II

We run additional simulation by generating data from another nonlinear noise model

$$X = \{1 + \exp(a_f^T Z)\}^{-1} + \epsilon_f, \text{ and}$$
$$Y = \tanh(a_g^T Z + bX) + \nu_g,$$

where the entries of a_f, a_g and noise variables ϵ_g, ν_g has same setting as setting I. The sample size is set at n=1000. The generation distribution of Z is from a standard normal distribution. We set the significance level at $\alpha = 0.05$. All results are based on 500 simulation replications.

We vary the dimension of Z as $d_Z = 50$, 100, 150, 200, 250. We compare our proposed DeepBET with generative conditional independence test (GCIT) of [Bellot and van der Schaar, 2019], and Deep Neural Network (DNN) with GCM of [Shah and Peters, 2020]. The left panel of Figure 8 reports the empirical size when $b = 0$. We find that DNN+GCM could potentially have inflated type I error when data dimension $d_Z = 50$.

In the right panel of Figure 8, we compare the empirical power of the DeepBET, DNN with GCM and CGIT when the dimension of Z is $d_Z = 100$. We vary the value of $b = 0.02, 0.08, 0.14$ that controls the magnitude of the alternative. We observe that DNN+GCM perform the best followed by DeepBET and then the GCIT.

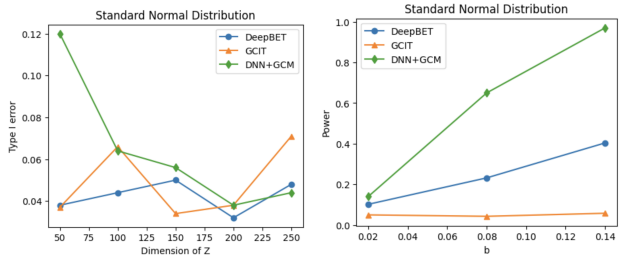

Figure 8: Left Panel: The empirical size of various tests under H_0. Right Panel: The empirical power of various tests under H_1.

Simulation setting III

We run additional simulation by generating data from another nonlinear noise model

$$X = \sin(a_f^T Z) + \epsilon_f, \text{ and } \; Y = (a_g^T Z + bX)^2 + \nu_g,$$

where the entries of a_f, a_g and noise variables ϵ_g, ν_g are all have the same setting with above model. The sample size is set at n=1000. The generation distribution of Z is from a standard normal distribution.

We vary the dimension of Z as $d_Z = 50$, 100, 150, 200, 250. We compare our proposed DeepBET with Deep Neural Network with GCM of [Shah and Peters, 2020]. The left panel of Figure 9 reports the empirical size when $b = 0$. We find that DNN+GCM have inflated type I errors in all instances. In the right panel of Figure 9, we compare the empirical power of the DeepBET and DNN with GCM when the dimension of Z is $d_Z = 100$. We vary the value of $b = 0.45$, 0.6, 0.75 and 0.9 that controls the magnitude of the alternative. We observe that DNN+GCM does not have sufficient power in this scenario.

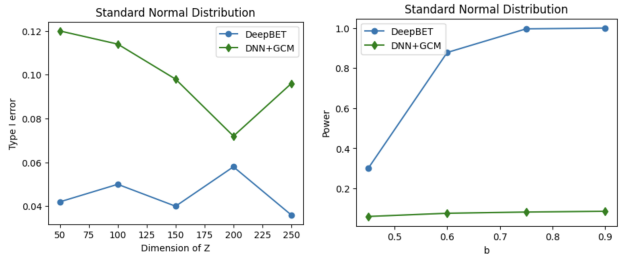

Figure 9: Left Panel: The empirical size of various tests under H_0. Right Panel: The empirical power of various tests under H_1.

3.3 Analysis of dry eye disease data

Dry eye disease (DED) encompasses a complex group of conditions resulting from dysfunction of the ocular system, significantly impacting patients' quality of life, financial resources, and US society as a whole. Symptoms of DED range from general discomfort to severe pain and burning sensations, which negatively affect sleep, productivity, and vision. It is estimated that DED affects over 16 million people in the US, with a prevalence of 6.8% among US adults.

Recent findings have shown thinning of the corneal epithelium in patients with DED (Figure 10 illustrates corneal epithelium thickness in normal individuals compared to three different types of DED). Epithelial mapping is conducted using anterior segment optical coherence tomography (AS-OCT), specifically the RTVue XR OCT Avanti System (Optovue Inc, Fremont, California, USA). This technique provides data on corneal epithelium thickness across 25 regions of the cornea, with darker colors indicating thinner areas of the cornea.

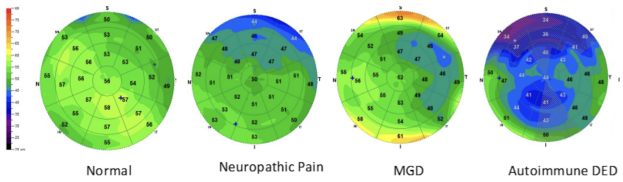

Figure 10: The example of Corneal Epithelial Thickness for Normal vs 3 types of Dry Eye Disease.

Our dataset includes measurements of corneal epithelium thickness across 25 regions and Schirmer's test results from 451 eyes (229 from right eyes and 222 from left eyes). First, we obtained the variable importance measures by fitting a random forest (RF) model. From the RF model results, we identified two key indicators for diagnosing DED: the Schirmer test, which measures tear production in five minutes, and the average corneal epithelial thickness of the central 10 regions of the cornea. These findings motivated us to implement the DeepBET method to determine which of the 25 regions of the corneal epithelium exhibits the strongest association with the Schirmer's test, given the results for the remaining 24 regions of corneal epithelial thickness.

We split the data into training and testing sets in a 1:4 ratio. The predictor variable X represents the corneal epithelium thickness in one of the regions, and the outcome variable Y is the Schirmer test result, with both X and Y conditioned on the remaining 24 regions of corneal epithelium thickness. We applied DeepBET to data from each participant's left eye to obtain p-values. Then, DeepBET was applied separately to

data from the right eyes. Our goal is to identify which region of the cornea is most strongly associated with tear production.

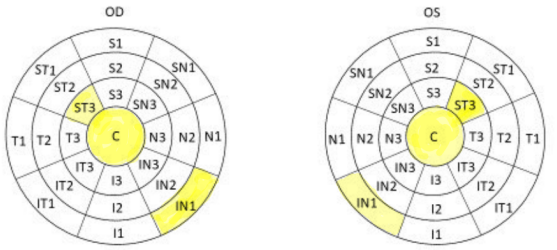

Figure 11: Three highlighted regions of epithelial thickness for both eyes are conditionally dependent with Shimmer test result.

Figure 11 reveals that 3 regions of corneal epithelium thickness for both eyes are significantly dependent on tear production (given other regions) after multiplicity corrections using conformal q-values proposed by [Zhao and Sun, 2024] with false discovery rate controlled at 5%.

Next, we examine the BET results for specific variables. Figure 12 shows two BET plots for the central region (C) of each eye, where both tests are significant. In both plots, U_x represents epithelial thickness, and U_y represents tear production. In the sigma field of binary variables, where $\mathbf{a} \neq 0$ and $\mathbf{b} \neq 0$, the max BET results reveal a significantly higher concentration of observations in the blue regions compared to the white regions, with the corresponding p-value being significant. This indicates that the conditional dependency between U_x and U_y arises from the interaction of the first bit from the U_y residuals and the first two bits from the U_x residuals.

Practically, this means that given other regions, patients with either thinner (below the first quartile) or thicker (above the third quartile) epithelial thickness in the center region tend to have lower (below median) tear production, as indicated by the clustering of points in the lower left and right regions. In contrast, patients with normal (between the first and third quartile) epithelial thickness exhibit higher (above median) tear production, as shown by the concentration of points in the upper central region. Notably again, this form of dependency is consistent for both eyes. Therefore, these BET plots provide some heuristic insights and explanations. Since other tests do not offer an immediate interpretability upon rejection if no further residual analysis is performed, DeepBET could be a valuable alternative in this sense.

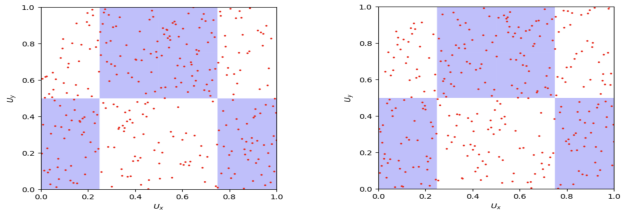

Figure 12: Left: BET plot for the central region (C) of the left eye. Right: BET plot for the central region (C) of the right eye. Both tests are significant, and both plots show significantly more points in the blue regions. These results reveal the same nonlinear relationship between the epithelial thickness and the tear production in the center region, given other regions.

4 Discussion

In this paper, we presented a novel method for conditional independence testing with high-dimensional confounding covariates. Our approach, called **DeepBET**, combines the strengths of advanced deep neural networks (DNNs), the non-parametric binary expansion testing (BET), and multiple-split techniques. The proposed method offers both theoretical and empirical advantages, including uniform consistency in detecting alternatives, fast computation, robustness to tuning parameter selection in DNNs, and the ability to handle high-dimensional confounding variables. Despite the nonparametric and high-dimensional nature of the confounding effects, DeepBET achieves a root-n consistent rate. This result is a complementary to a rapidly growing literature on estimation/statistical inference using machine learning methods, in particular, the semi-parametric estimation/statistical inference for low-dimensional parameters with nuisance functions of high-dimensional confounding variables (e.g., [Chernozhukov et al., 2018, Farrell et al., 2021]). However, there is limited research on using DNN for conditional independence tests.

We have also demonstrated the superior empirical performance of DeepBET and its computational efficiency compared to several existing methods. The gains in interpretability, as shown in the dry eye disease data analysis, suggest that DeepBET is a valuable alternative in practice.

Acknowledgments

The authors express sincere gratitude to Dr. Wan Zhang and Dr. Duo Zheng for their insightful discussions, which provided valuable inspiration for this research. The author also appreciates the constructive feedback of the anonymous reviewers, which has significantly enhanced the quality of this paper. In addition, special thanks are given to Sandeep Jain, MD, and

Zeenal G. Dabre for generously providing the data used in this study. The research was partially supported by NSF grants FRG 2152289 and FRG 2152070.

References

[Bauer and Kohler, 2019] Bauer, B. and Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261 – 2285.

[Bellot and van der Schaar, 2019] Bellot, A. and van der Schaar, M. (2019). Conditional independence testing using generative adversarial networks. pages 2199–2208.

[Bickel and Levina, 2008] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1).

[Cai et al., 2018] Cai, J., Luo, J., Wang, S., and Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, 300:70–79.

[Chen and Shen, 1998] Chen, X. and Shen, X. (1998). Sieve extremum estimates for weakly dependent data. *Econometrica*, 66(2):289–314.

[Chernozhukov et al., 2018] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

[Duong and Nguyen, 2022] Duong, B. and Nguyen, T. (2022). Conditional independence testing via latent representation learning. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 121–130.

[Farrell et al., 2021] Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

[Fukumizu et al., 2007] Fukumizu, K., Gretton, A., Sun, X., and Scholkopf, B. (2007). Kernel measures of conditional dependence. In *Neural Information Processing Systems*.

[Guo and Shah, 2024] Guo, F. R. and Shah, R. D. (2024). Rank-transformed subsampling: inference for multiple data splitting and exchangeable p-values.

[Hoyer et al., 2008] Hoyer, P. O., Janzing, D., Mooij, J., Peters, J., and Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, page 689–696, Red Hook, NY, USA. Curran Associates Inc.

[Jentzen and Riekert, 2022] Jentzen, A. and Riekert, A. (2022). A proof of convergence for the gradient descent optimization method with random initializations in the training of neural networks with relu activation for piecewise linear target functions. *Journal of Machine Learning Research*, 23(260):1–50.

[Lee, 2021] Lee, K. J. (2021). Chapter seven - architecture of neural processing unit for deep neural networks. In Kim, S. and Deka, G. C., editors, *Hardware Accelerator Systems for Artificial Intelligence and Machine Learning*, volume 122 of *Advances in Computers*, pages 217–245. Elsevier.

[Li and Fan, 2020] Li, C. and Fan, X. (2020). On nonparametric conditional independence tests for continuous variables. *WIREs Computational Statistics*, 12(3):e1489.

[Lu et al., 2021] Lu, J., Shen, Z., Yang, H., and Zhang, S. (2021). Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506.

[Pearl, 2009] Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

[Polson and Sokolov, 2018] Polson, N. G. and Sokolov, V. O. (2018). Deep learning.

[Schmidt-Hieber, 2020] Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897.

[Sen et al., 2017] Sen, R., Suresh, A. T., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017). Model-powered conditional independence test.

[Shah and Peters, 2020] Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3).

[Shi et al., 2021] Shi, C., Xu, T., Bergsma, W., and Li, L. (2021). Double generative adversarial networks for conditional independence testing. *The Journal of Machine Learning Research*, 22(1):13029–13060.

[Strobl et al., 2017] Strobl, E. V., Zhang, K., and Visweswaran, S. (2017). Approximate kernel-based conditional independence tests for fast non-parametric causal discovery.

[Su and White, 2007] Su, L. and White, H. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834.

[Su and White, 2008] Su, L. and White, H. (2008). A nonparametric hellinger metric test for conditional independence. *Econometric Theory*, 24(4):829–864.

[Su and White, 2014] Su, L. and White, H. L. (2014). Testing conditional independence via empirical likelihood. *Journal of Econometrics*, 182:27–44.

[Waggoner, 2021] Waggoner, P. D. (2021). *Modern Dimension Reduction*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

[Wang et al., 2015] Wang, X., Pan, W., Hu, W., Tian, Y., and Zhang, H. (2015). Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734.

[Zhang et al., 2023] Zhang, H., Xia, Y., Zhang, K., Zhou, S., and Guan, J. (2023). Conditional independence test based on residual similarity. *ACM Trans. Knowl. Discov. Data*, 17(8).

[Zhang et al., 2017] Zhang, H., Zhou, S., Zhang, K., and Guan, J. (2017). Causal discovery using regression-based conditional independence tests. In *AAAI Conference on Artificial Intelligence*.

[Zhang, 2019] Zhang, K. (2019). Bet on independence. *Journal of the American Statistical Association*, 114(528):1620–1637.

[Zhang et al., 2011] Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2011). Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, page 804–813, Arlington, Virginia, USA. AUAI Press.

[Zhao and Sun, 2024] Zhao, Z. and Sun, W. (2024). False discovery rate control for structured multiple testing: Asymmetric rules and conformal q-values.

Checklist

(I) For all models and algorithms presented, check if you include:

- (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable] Yes
- (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable] Yes
- (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]

(II) For any theoretical claim, check if you include:

- (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable] Yes
- (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable] Yes
- (c) Clear explanations of any assumptions. [Yes/No/Not Applicable] Yes

(III) For all figures and tables that present empirical results, check if you include:

- (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable] Yes
- (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable] Yes
- (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes/No/Not Applicable] Yes
- (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/No/Not Applicable] Yes

(IV) If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

- (a) Citations of the creator If your work uses existing assets. [Yes/No/Not Applicable] Yes
- (b) The license information of the assets, if applicable. [Yes/No/Not Applicable] Not Applicable
- (c) New assets either in the supplemental material or as a URL, if applicable. [Yes/No/Not Applicable] Not applicable
- (d) Information about consent from data providers/curators. [Yes/No/Not Applicable] Yes
- (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/Not Applicable] No

(V) If you used crowdsourcing or conducted research with human subjects, check if you include:

- (a) The full text of instructions given to participants and screenshots. [Yes/No/Not Applicable] Yes
- (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/Not Applicable] Not Applicable

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/Not Applicable]
No

Appendix A. Proofs

Technical proofs

For theoretical analysis, we impose some conditions on the family of the nonparametric functions $h(\cdot)$ and $g(\cdot)$, and the architecture of the neural network.

Definition 1. *A function $m(z)$ is (p, C)-smooth if the partial derivative $\partial^q m(z)/\partial z_1^{\alpha_1} \cdots \partial z_d^{\alpha_d}$ exists and satisfy*

$$\Big| \frac{\partial^q m(z)}{\partial z_1^{\alpha_1} \cdots \partial z_d^{\alpha_d}} - \frac{\partial^q m(x)}{\partial z_1^{\alpha_1} \cdots \partial z_d^{\alpha_d}} \Big| \leq C\|z - x\|^s,$$

for all $x, z \in R^d$, $\sum_{j=1}^d \alpha_j = q$ and $p = q + s$. A function $m(z)$ is a generalized hierarchical interaction model of order d_* and level 0, if there exist $a_1, \cdots, a_{d_*} \in R^d$ and $f : R^{d_*} \to R$ such that $m(z) = f(a_1^T z, \cdots, a_{d_*}^T z)$ for all z. Then, a function $m(z)$ is a generalized hierarchical interaction model of order d_* and level $l + 1$ if there exist K, $\theta_k(\cdot)$ and $f_{1,k}, \cdots, f_{d_*,k}$ such that $m(z) = \sum_{k=1}^K \theta_k\{f_{1,k}(z), \cdots, f_{d_*,k}(z)\}$ where $f_{1,k}, \cdots, f_{d_*,k}$ are generalized hierarchical models of order d_* and level l. If all the functions $f(\cdot)$, $\theta_k(\cdot)$, $f_{1,k}, \cdots, f_{d_*,k}$ are (p, C) smooth, then the function $m(z)$ is a (p, C)-smooth generalized hierarchical interaction model.

Definition 2. *A hierarchical neural network $\mathcal{H}^{(l)}$ is defined recursively given by the following:*

$$\mathcal{H}^{(l)} = \{h : h(z) = \sum_{k=1}^K \theta_k(f_{1,k}(z), \cdots, f_{d_*,k}(z)) \text{ for some } g_k \in \mathcal{F}^{NN}_{M^*,d_*,d,\alpha} \text{ and } f_{j,k} \in \mathcal{H}^{(l-1)}\}.$$

and $\mathcal{H}^{(0)} = \mathcal{F}^{NN}_{M^*,d_*,d,\alpha}$. Here $\mathcal{F}^{NN}_{M^*,d_*,d,\alpha}$ is the family of neural networks include the set of all functions of the form

$$f(z) = \sum_{j=1}^{M^*} \mu_i \sigma\Big(\sum_{j=1}^{4d_*} \lambda_{i,j} \sigma\Big(\sum_{v=1}^{d} \theta_{i,j,\nu} z_v + \theta_{i,j,0} \Big) + \lambda_{i,0} \Big) + \mu_0$$

for some $|\mu_i| \leq \alpha$, $|\lambda_{i,j}| \leq \alpha$ and $|\theta_{i,j,\nu}| \leq \alpha$, and an activation function $\sigma(\cdot)$.

We estimate the nonparametric function $h(\cdot)$ by minimizing the following objective function:

$$\hat{h} = \arg \min_{h \in \mathcal{H}^{(l)}} \sum_{i=1}^n (X_i - h(Z_i))^2, \tag{1}$$

where $\mathcal{H}^{(l)}$ is defined in Definition 2. In practice, we minimize the loss function in equation (1) using algorithms such as stochastic gradient descent algorithms. It is possible that the global minimization might not be achieved. In such cases, the results about the convergence rate in Lemma 1 might not be guaranteed. We made the following assumptions:

(C1) $z \in R^d$ has bounded support. Assume that both functions $h(z)$ and $g(z)$ satisfy the (p, C)-smooth generalized hierarchical interaction model of order d_* and finite level l with $p = q + s$ for some non-negative integers q and $s \in (0, 1]$.

(C2) Assume that $(X_1, Y_1, Z_1), \cdots, (X_n, Y_n, Z_n)$ are independent and identically distributed samples, $E\{\exp(c_1 X^2)\} < \infty$ and $E\{\exp(c_1 Y^2)\} < \infty$ for some constants c_1 and c_2.

The following results are proved in [Bauer and Kohler, 2019].

Lemma 1. *Assume (C1), (C2) and all the partial derivatives of order $\leq q$ of $\theta_k, f_{j,k}$ are bounded and θ_k is Lipschitz continuous. In addition, $M^* \asymp c n^{d_*/(2p+d_8)}$, $\alpha \asymp n^c$ for sufficient large constant c and $\sigma(\cdot)$ is N-admissible. Then, the estimated function from the optimization in (1) converges to $h(\cdot)$ with the the following rate:*

$$E\|\hat{h} - h\|_2^2 \leq c n^{-2p/(2p+d_*)} \log^3(n),$$

for sufficiently large n.

The following Lemma generalized the results the inequality in [Bickel and Levina, 2008] for the differences between products with l terms ($l \geq 2$).

Lemma 2. *Let* $I_a = \sum_{1\le v_1<\cdots<v_a\le l}\prod_{j=1}^{a}|C^{(v_j)}-\tilde{C}^{(v_j)}|\prod_{j\neq v_1,\cdots,v_a}^{a}|\tilde{C}^{(j)}|$. *Then, the following inequality holds:*

$$\left|\prod_{j=1}^{l}C^{(j)}-\prod_{j=1}^{l}\tilde{C}^{(j)}\right|\le\sum_{a=1}^{l}I_a.$$

Proof of Lemma 2: Note that we can write

$$\begin{aligned}\prod_{j=1}^{l}C^{(j)}&=\prod_{j=1}^{l}\{(C^{(j)}-\tilde{C}^{(j)})+\tilde{C}^{(j)}\}\\&=\prod_{j=1}^{l}\tilde{C}^{(j)}+\sum_{a=1}^{l}\sum_{1\le v_1<\cdots<v_a\le l}^{l}\prod_{j=1}^{a}(C^{(v_j)}-\tilde{C}^{(v_j)})\prod_{j\neq v_1,\cdots,v_a}^{a}\tilde{C}^{(j)}.\end{aligned}$$

It then follows that

$$\left|\prod_{j=1}^{l}C^{(j)}-\prod_{j=1}^{l}\tilde{C}^{(j)}\right|\le\sum_{a=1}^{l}\sum_{1\le v_1<\cdots<v_a\le l}^{l}\prod_{j=1}^{a}|C^{(v_j)}-\tilde{C}^{(v_j)}|\prod_{j\neq v_1,\cdots,v_a}^{a}|\tilde{C}^{(j)}|.$$

This completes the proof of Lemma 2.

Proof of Theorem 1: Since $n_{\mathcal{D}_2}$ is at the same order as n, we use n to replace $n_{\mathcal{D}_2}$ in this proof to simplify the notations. Theorem 4.2 of [Zhang, 2019] has shown that $(S_{(\mathbf{ab}),n}+n)/4\sim\text{Hypergeometric}(n,n/2,n/2)$. To prove the desired result, $(\hat{S}_{(\mathbf{ab}),n}+n)/4\sim\text{Hypergeometric}(n,n/2,n/2)$, we will need to show that $|\hat{S}_{(\mathbf{ab}),n}-S_{(\mathbf{ab}),n}|=o_p\{S_{(\mathbf{ab}),n}\}$.

We start with calculate the difference between $\hat{S}_{(\mathbf{ab}),n}$ and $S_{(\mathbf{ab}),n}$, given by $|\hat{S}_{(\mathbf{ab}),n}-S_{(\mathbf{ab}),n}|$:

$$\begin{aligned}|\hat{S}_{(\mathbf{ab}),n}-S_{(\mathbf{ab}),n}|&=|\hat{S}_{(\mathbf{ab}),n}-\hat{S}_{h_{(\mathbf{ab}),n}}+\hat{S}_{h_{(\mathbf{ab}),n}}-S_{h_{(\mathbf{ab}),n}}+S_{h_{(\mathbf{ab}),n}}-S_{(\mathbf{ab}),n}|\\&\le|\hat{S}_{(\mathbf{ab}),n}-\hat{S}_{h_{(\mathbf{ab}),n}}|+|\hat{S}_{h_{(\mathbf{ab}),n}}-S_{h_{(\mathbf{ab}),n}}|+|S_{h_{(\mathbf{ab}),n}}-S_{(\mathbf{ab}),n}|,\end{aligned}$$

where $S_{h_{(\mathbf{ab}),n}}$ is a smoothed version of $S_{(\mathbf{ab}),n}$, which will be defined below. We will show in three steps that each term on the right-hand side of the above inequality is a smaller order term of $S_{(\mathbf{ab}),n}$ so that $|\hat{S}_{(\mathbf{ab}),n}-S_{(\mathbf{ab}),n}|=o_p\{S_{(\mathbf{ab}),n}\}$.

Step 1: Let $\hat{\epsilon}_i$ and $\hat{\nu}_i$ be residuals obtained from the DNN fitting. Let F and G be, respectively, distribution functions of $\hat{\epsilon}_i$ and $\hat{\nu}_i$. Then we obtain the transformed random variables: $\hat{U}_i=F(\hat{\epsilon}_i)$ and $\hat{V}_i=G(\hat{\nu}_i)$. In the proposed BET procedure, the binary variables $\hat{A}_i^{(k)}$ and $\hat{B}_i^{(k)}$ are generated by the predicted residuals:

$$\hat{A}_i^{(k)}=\begin{cases}-1 & if\quad \frac{2k-2}{2^{d_1}}<\hat{U}_i\le\frac{2k-1}{2^{d_1}}\\+1 & if\quad \frac{2k-1}{2^{d_1}}<\hat{U}_i\le\frac{2k}{2^{d_1}}\end{cases}$$

and

$$\hat{B}_i^{(j)}=\begin{cases}-1 & if\quad \frac{2j-2}{2^{d_2}}<\hat{V}_i\le\frac{2j-1}{2^{d_2}}\\+1 & if\quad \frac{2j-1}{2^{d_2}}<\hat{V}_i\le\frac{2j}{2^{d_2}}\end{cases}$$

where $k=1,\cdots,2^{d_1-1}$ and $j=1,\cdots,2^{d_2-1}$.

Let us construct a smoothed version of $\hat{A}_{h_i}^{(k)}$ so that it transitions smoothly from -1 to +1. That is

$$\hat{A}_{h_i}^{(k)}=\begin{cases}-1 & if\quad \frac{2k-2}{2^{d_1}}<\hat{U}_i<\frac{2k-1}{2^{d_1}}-h\\-1+2(\int_{\frac{2k-1}{2^{d_1}}-h}^{\hat{U}_i}\psi(\frac{t-(\frac{2k-1}{2^{d_1}})}{h})dt/\int_{\frac{2k-1}{2^{d_1}}-h}^{\frac{2k-1}{2^{d_1}}+h}\psi(\frac{t-(\frac{2k-1}{2^{d_1}})}{h})dt) & if\quad \frac{2k-1}{2^{d_1}}-h\le\hat{U}_i\le\frac{2k-1}{2^{d_1}}+h\\+1 & if\quad \frac{2k-1}{2^{d_1}}+h<\hat{U}_i<\frac{2k}{2^{d_1}}\end{cases}\tag{2}$$

where $\psi(x)$ is a bump function which is infinitely differentiable defined by

$$\psi(x)=\begin{cases}e^{-1/(1-x^2)} & for-1<x<1\\0 & otherwise.\end{cases}\tag{3}$$

The difference between $\hat{A}_i^{(k)}$ and $\hat{A}_{h_i}^{(k)}$ occurs only in the interval $[\frac{2k-1}{2^{d_1}} - h, \frac{2k-1}{2^{d_1}} + h]$. Outside of this region, both functions are identical. Define the event where $\hat{A}_i^{(k)}$ and $\hat{A}_{h_i}^{(k)}$ differ:

$$\mathcal{E}_A := \left\{ \hat{U}_i \in \left(\frac{2k-1}{2^{d_1}} - h, \frac{2k-1}{2^{d_1}} + h \right) \right\}.$$

Then, for any $u > 0$,

$$P(|\hat{A}_i^{(k)} - \hat{A}_{h_i}^{(k)}| > u) \leq \mathbb{P}\left(\hat{A}_i^{(k)} \neq \hat{A}_{h_i}^{(k)} \right) = \mathbb{P}\left(\mathcal{E}_A \right) = \int_{\frac{2k-1}{2^{d_1}} - h}^{\frac{2k-1}{2^{d_1}} + h} f_{\hat{U}_i}(u) du,$$

where $f_{\hat{U}_i}(u)$ denote the probability density function of $\hat{U}_i$.
When h is small and the density $f_{\hat{U}_i}(u)$ is bounded, the above integration can be approximated by the following:

$$\mathbb{P}\left(\mathcal{E}_A \right) \approx f_{\hat{U}_i}\left(\frac{2k-1}{2^{d_1}} \right) \cdot \left(\frac{2k-1}{2^{d_1}} + h - \left(\frac{2k-1}{2^{d_1}} - h \right) \right) = 2h \cdot f_{\hat{U}_i}\left(\frac{2k-1}{2^{d_1}} \right) = \mathcal{O}(h).$$

Thus, we have:

$$|\hat{A}_i^{(k)} - \hat{A}_{h_i}^{(k)}| = \mathcal{O}_p(h).$$

Similarly, we can also construct a smoothed version of $\hat{B}_i^{(k)}$ using $\hat{B}_{h_i}^{(k)}$. Using the same approach, we can bound the difference between $\hat{B}_i^{(k)}$ and $\hat{B}_{h_i}^{(k)}$ by

$$|\hat{B}_i^{(k)} - \hat{B}_{h_i}^{(k)}| = \mathcal{O}_p(h).$$

Now, we will bound the difference $\hat{S}_{(\mathbf{ab}),n_i}$ and $\hat{S}_{h_{(\mathbf{ab}),n_i}}$. Recall that

$$\hat{S}_{(\mathbf{ab}),n_i} = \prod_{k=1}^{d_1} \{\hat{A}_i^{(k)}\}^{a_k} \prod_{k=1}^{d_2} \{\hat{B}_i^{(k)}\}^{b_k}, ; \text{and } \hat{S}_{h_{(\mathbf{ab}),n_i}} = \prod_{k=1}^{d_1} \{\hat{A}_{h_i}^{(k)}\}^{a_k} \prod_{k=1}^{d_2} \{\hat{B}_{h_i}^{(k)}\}^{b_k}.$$

Let $\hat{C}_i^{(l)}$ and $\hat{C}_{h_i}^{(l)}$ denote the l-th term in $\hat{S}_{(\mathbf{ab})}$ and $\hat{S}_{h_{(\mathbf{ab})}}$. Applying the Lemma 2, we have

$$\begin{aligned} |\hat{S}_{(\mathbf{ab}),n_i} - \hat{S}_{(\mathbf{ab}),n_i}| &= |\hat{C}_i^{(1)} \hat{C}_i^{(2)} \cdots \hat{C}_i^{(l)} - \hat{C}_{h_i}^{(1)} \hat{C}_{h_i}^{(2)} \cdots \hat{C}_{h_i}^{(l)}| \\ &\leq I_{i1} + I_{i2} + \cdots I_{i(d_1+d_2)}, \end{aligned}$$

where, for $a = 1, \cdots, d_1 + d_2$,

$$I_{ia} = \sum_{1 \leq v_1 < \cdots < v_a \leq l} \prod_{j=1}^{a} |\hat{C}_i^{(v_j)} - \hat{C}_{h_i}^{(v_j)}| \prod_{j \neq v_1, \cdots, v_a}^{a} |\hat{C}_i^{(j)}|.$$

By definition $|\hat{C}_i^{(j)}| = 1$, So we have

$$I_{ia} = \sum_{1 \leq v_1 < \cdots < v_a \leq l} \prod_{j=1}^{a} |\hat{C}_i^{(v_j)} - \hat{C}_{h_i}^{(v_j)}|.$$

Recall $\hat{C}_i^{(l)}$ and $\hat{C}_{h_i}^{(l)}$ are the l-th term in $\hat{S}_{(\mathbf{ab}),n_i}$ and $\hat{S}_{h_{(\mathbf{ab}),n_i}}$, so $\hat{C}^l$ is either $\hat{A}_i^{(k)}$ or $\hat{B}_i^{(k)}$ and $\hat{C}_{h_i}^l$ is either $\hat{A}_{h_i}^{(k)}$ or $\hat{B}_{h_i}^{(k)}$. Because $|\hat{A}_i^{(k)} - \hat{A}_{h_i}^{(k)}| = \mathcal{O}_p(h)$ and $|\hat{B}_i^{(k)} - \hat{B}_{h_i}^{(k)}| = \mathcal{O}_p(h)$, we have $|\hat{C}_i^{(v_j)} - \hat{C}_{h_i}^{(v_j)}| = \mathcal{O}_p(h)$. Thus, we have:

$$I_{ia} = \mathcal{O}_p\Big\{ \binom{d_1 + d_2}{a} h^a \Big\},$$

for $a = 1, \cdots, d_1 + d_2$. Then

$$\sum_{a=1}^{d_1+d_2} I_{ia} = \mathcal{O}_p\{(1+h)^{d_1+d_2} - 1\}.$$

Then, if we choose small enough h, then we could bound the difference $|\hat{S}_{(\mathbf{ab}),n} - \hat{S}_{(\mathbf{ab}),n}|$:

$$|\hat{S}_{(\mathbf{ab}),n} - \hat{S}_{(\mathbf{ab}),n}| = |\sum_{i=1}^{n}(\hat{S}_{(\mathbf{ab}),n_i} - \hat{S}_{(\mathbf{ab}),n_i})| = \sum_{i=1}^{n}\sum_{a=1}^{d_1+d_2} I_{ia} = \mathcal{O}_p\left(n\{(1+h)^{d_1+d_2} - 1\}\right) = o_p(S_{\mathbf{ab},n}).$$

Step 2: We will bound the difference between $\hat{S}_{h_{(\mathbf{ab}),n}}$ and $S_{h_{(\mathbf{ab}),n}}$, which is given by the following expression:

$$|\hat{S}_{h_{(\mathbf{ab}),n_i}} - S_{h_{(\mathbf{ab}),n_i}}| = |\hat{A}_{h_i}^{(1)}\hat{A}_{h_i}^{(2)}\cdots\hat{A}_{h_i}^{(d_1)}\hat{B}_{h_i}^{(1)}\hat{B}_{h_i}^{(2)}\cdots\hat{B}_{h_i}^{(d_2)} - A_{h_i}^{(1)}A_{h_i}^{(2)}\cdots A_{h_i}^{(d_1)}B_{h_i}^{(1)}B_{h_i}^{(2)}\cdots B_{h_i}^{(d_2)}|.$$

Let $\hat{C}_{h_i}^{(l)}$ and $C_{h_i}^{(l)}$ denote the l-th term in $\hat{S}_{h_{(\mathbf{ab})}}$ and $S_{h_{(\mathbf{ab})}}$, then

$$\begin{aligned}|\hat{S}_{h_{(\mathbf{ab})}} - S_{h_{(\mathbf{ab})}}| &= |\sum_{i=1}^{n}(\hat{C}_{h_i}^{(1)}\hat{C}_{h_i}^{(2)}\cdots\hat{C}_{h_i}^{(d_1+d_2)} - C_{h_i}^{(1)}C_{h_i}^{(2)}\cdots C_{h_i}^{(d_1+d_2)}| \\ &\leq \sum_{i=1}^{n}(I'_{i1} + I'_{i2} + \cdots I'_{i(d_1+d_2)}),\end{aligned}$$

where

$$I'_{ia} = \sum_{i=1}^{n}\Big(\prod_{j=0}^{a-1}|C_{h_i}^{(j)}|\Big)|\hat{C}_{h_i}^{(a)} - C_{h_i}^{(a)}|\Big(\prod_{j=a+1}^{d_1+d_2}|\hat{C}_{h_i}^{(j)}|\Big).$$

Since $|C_{h_i}^{(j)}| \leq 1$ and $|\hat{C}_{h_i}^{(j)}| \leq 1$, we have

$$\sum_{i=1}^{n} I'_{ia} \leq \sum_{i=1}^{n}|\hat{C}_{h_i}^{(a)} - C_{h_i}^{(a)}|.$$

Recall that the definition of $\hat{C}_{h_i}^{(l)}$ and $C_{h_i}^{(l)}$ in $\hat{S}_{h_{(\mathbf{ab})}}$ and $S_{h_{(\mathbf{ab})}}$, we need to find out the difference between $|\hat{A}_{h_i}^{(k)} - A_{h_i}^{(k)}|$ and $|\hat{B}_{h_i}^{(k)} - B_{h_i}^{(k)}|$. Using Taylor's expansion, we have: $\hat{A}_{h_i}^{(k)} = A_{h_i}^{(k)} + A_{h_i}^{\prime(k)}(\hat{U}_i - U_i), \hat{B}_{h_i}^{(k)} = B_{h_i}^{(k)} + B_{h_i}^{\prime(k)}(\hat{V}_i - V_i)$ Then,

$$\begin{aligned}|\hat{A}_{h_i}^{(k)} - A_{h_i}^{(k)}| &\leq |A_{h_i}^{\prime(k)}(u_i)| \cdot |(\hat{U}_i - U_i)| \\ |\hat{B}_{h_i}^{(k)} - B_{h_i}^{(k)}| &\leq |B_{h_i}^{\prime(k)}(u_i)| \cdot |(\hat{V}_i - V_i)|\end{aligned}$$

for some $u_i \in (\min(\hat{U}_i, U_i), \max(\hat{U}_i, U_i))$ and $v_i \in (\min(\hat{V}_i, V_i), \max(\hat{V}_i, V_i))$. Since $|\hat{U}_i - U_i| = |F(\hat{\epsilon}_i) - F(\epsilon_i)|$ and $|\hat{V}_i - V_i| = |G(\hat{\nu}_i) - G(\nu_i)|$, using the first order Taylor expansion of F and G, we have

$$\begin{aligned}|\hat{U}_i - U_i| = |F(\hat{\epsilon}_i) - F(\epsilon_i)| = |F^{'}(c_i)(\hat{\epsilon}_i - \epsilon_i)|, \\ |\hat{V}_i - V_i| = |G(\hat{\nu}_i) - G(\nu_i)| = |G^{'}(d_i)(\hat{\nu}_i - \nu_i)|,\end{aligned}$$

where c_i and d_i is an intermediate point between $\hat{\epsilon}_i$ and ϵ_i, $\hat{\nu}$ and ν_i. Then, for a fix $1 \leq a \leq d_1$,

$$\sum_{i=1}^{n} I'_{ia} \leq \sum_{i=1}^{n}|\hat{C}_{h_i}^{(a)} - C_{h_i}^{(a)}| \leq \sum_{i=1}^{n}|A_{h_i}^{\prime(k)}(u_i)| \cdot |(\hat{U}_i - U_i)|^{a_1}.$$

By the Cauchy-Schwarz inequality:

$$\sum_{i=1}^{n} I'_{ia} \leq \sqrt{\sum_{i=1}^{n}|A_{h_i}^{\prime(k)}(u_i)F'(c_i)|^2} \cdot \sqrt{\sum_{i=1}^{n}|(\hat{\epsilon}_i - \epsilon_i)|^2}.$$

Assume $F^{'}(c_i)$ is bounded by some constant L and M, $|F^{'}(c_i)| \leq L$ and $|G^{'}(d_i)| \leq Q$. The derivative of $\hat{A}_{h_i}^{(k)}$ with respect to u_i is given by:

$$A_{h_i}^{\prime(k)}(\hat{U}_i) = \begin{cases} 0 & \text{if } \frac{2k-2}{2^{d_1}} < \hat{U}_i < \frac{2k-1}{2^{d_1}} - h, \\ \frac{2}{h} \cdot \frac{\psi\left(\frac{\hat{U}_i - \frac{2k-1}{2^{d_1}}}{h}\right)}{\int_{\frac{2k-1}{2^{d_1}}-h}^{\frac{2k-1}{2^{d_1}}+h} \psi\left(\frac{t-(\frac{2k-1}{2^{d_1}})}{h}\right) dt} & \text{if } \frac{2k-1}{2^{d_1}} - h \leq \hat{U}_i \leq \frac{2k-1}{2^{d_1}} + h, \\ 0 & \text{if } \frac{2k-1}{2^{d_1}} + h < \hat{U}_i < \frac{2k}{2^{d_1}}. \end{cases}$$

For any $c > 0$, we know that, as $h \to \infty$,

$$P\left(\left|A_{h_i}^{\prime(k)}(\hat{U}_i)\right| > c\right) \leq P\left(\frac{2k-1}{2^{d_1}} - h \leq \hat{U}_i \leq \frac{2k-1}{2^{d_1}} + h\right) \to 0.$$

Thus, we have: $|A_{h_i}^{\prime(k)}(\hat{U}_i)| = o_p(1)$. If we choose h small enough so that $nh \to 0$, then

$$\sqrt{\sum_{i=1}^{n} |A_{h_i}^{\prime(k)}(u_i)F'(c_i)|^2} = o_p(1).$$

By the ANM model assumption: $X_i = h(Z_i) + \epsilon_i$:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i - \epsilon_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\{(X_i - \hat{h}(Z_i)) - (X_i - h(Z_i))\}^2 \to \|\hat{h} - h\|^2.$$

Because of the Markov inequality and using Lemma 1, we obtain:

$$E\|\hat{h} - h\|_2^2 \leq cn^{-2p/(2p+d_*)} \log^3(n),$$

we have

$$P(\|\hat{h} - h\|^2 > c) \leq \frac{E\|\hat{h} - h\|^2}{c} = \mathcal{O}(n^{-2p/(2p+d_*)} \log^3(n)).$$

So, we have

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i - \epsilon_i)^2 = \mathcal{O}_p(n^{-2p/(2p+d_*)} \log^3(n)).$$

Thus, we can write:

$$\sqrt{\sum_{i=1}^{n}(\hat{\epsilon}_i - \epsilon_i)^2} = \mathcal{O}_p\{n^{1/2-\frac{p}{2p+d_*}} \log^{3/2}(n)\} = o_p(n^{1/2}).$$

If $d_1 = o(n^{\frac{p}{2p+d_*}})$ and $d_2 = o(n^{\frac{p}{2p+d_*}})$, then we have

$$|\hat{S}_{h_{(\mathbf{ab}),n}} - S_{h_{(\mathbf{ab}),n}}| = o_p(n^{1/2}) = o_p\{S_{h_{(\mathbf{ab}),n}}\}.$$

Step 3: For the binary variable A_i^k which is generated by the i.i.d random residual ϵ

$$A_i^{(k)} = \begin{cases} -1 & if \quad \frac{2k-2}{2^{d_1}} < U_i \leq \frac{2k-1}{2^{d_1}} \\ +1 & if \quad \frac{2k-1}{2^{d_1}} < U_i \leq \frac{2k}{2^{d_1}} \end{cases} \tag{4}$$

where $k = 1, \cdots, 2^{d_1-1}$. Following the same steps in the above step 1, we could conclude the result that

$$|S_{h_{(\mathbf{ab})}} - S_{(\mathbf{ab})}| = o_p(S_{h_{(\mathbf{ab}),n}}).$$

In summary of the above Steps 1-3, we have proved that $|\hat{S}_{(\mathbf{ab}),n} - S_{(\mathbf{ab}),n}| = o_p\{S_{(\mathbf{ab}),n}\}$. This completes the proof of Theorem 1.

Proof of Theorem 2: We construct a normalized BET test statistic

$$T_n = \max_{\mathbf{ab}\in\sigma(\hat{U}_{d_1},\hat{V}_{d_2})}(n_{\mathcal{D}_2}-1)^{1/2}S_{(\mathbf{ab}),n}/n_{\mathcal{D}_2}.$$

According to the results in [Zhang, 2019], the maximum of the standardized statistic T_n has the same asymptotic distribution with the asymptotic distribution of independent Gaussian distributions under the null. To show the desired results, we need to show that $|\hat{T}_n - T_n| = o_p(1)$. Based on the results in Theorem 1, we have

$$|(n_{\mathcal{D}_2}-1)^{1/2}\hat{S}_{(\mathbf{ab}),n}/n_{\mathcal{D}_2} - (n_{\mathcal{D}_2}-1)^{1/2}S_{(\mathbf{ab}),n}/n_{\mathcal{D}_2}| = o_p\{(d_1+d_2)n^{-p/(2p+d_*)}\}.$$

Note that the total number of elements in the σ-field $\sigma(\hat{U}_{d_1},\hat{V}_{d_2})$ is $2^{d_1+d_2}$. So, we have

$$\begin{aligned}
|\hat{T}_n - T_n| &\le \max_{\mathbf{ab}\in\sigma(\hat{U}_{d_1},\hat{V}_{d_2})} |(n_{\mathcal{D}_2}-1)^{1/2}\hat{S}_{(\mathbf{ab}),n}/n_{\mathcal{D}_2} - (n_{\mathcal{D}_2}-1)^{1/2}S_{(\mathbf{ab}),n}/n_{\mathcal{D}_2}| \\
&\le \sum_{\mathbf{ab}\in\sigma(\hat{U}_{d_1},\hat{V}_{d_2})} |(n_{\mathcal{D}_2}-1)^{1/2}\hat{S}_{(\mathbf{ab}),n}/n_{\mathcal{D}_2} - (n_{\mathcal{D}_2}-1)^{1/2}S_{(\mathbf{ab}),n}/n_{\mathcal{D}_2}| = o_p\{(d_1+d_2)2^{d_1+d_2}n^{-p/(2p+d_*)}\}.
\end{aligned}$$

If $d_1 + d_2 = o\{\log_2(n^{p/(2p+d_*)})\}$, then $|\hat{T}_n - T_n| = o_p(1)$. This completes the proof of Theorem 2.