Supplementary material for AUGUST

Benjamin Brown[*], Kai Zhang[†]

December 6, 2023

0 Supplementary Materials

0.1 Example and Visualization

An interpretation example. Let $d = 3$. Given $\boldsymbol{X}$ and $\boldsymbol{Y}$ data, suppose that $\mathbf{H}_8\boldsymbol{P}_{\boldsymbol{X}}$ is explicitly computed to be

$$\begin{pmatrix} 1.00 \\ 0.00 \\ -0.10 \\ 0.02 \\ -0.02 \\ -0.02 \\ -0.08 \\ 0.00 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix} \begin{pmatrix} 0.10 \\ 0.10 \\ 0.14 \\ 0.15 \\ 0.13 \\ 0.12 \\ 0.13 \\ 0.13 \end{pmatrix}.$$

Recall that $\boldsymbol{S}_{\boldsymbol{X}}$ consists of all but the first coordinate of the vector on the left. In this

[*]Benjamin Brown is a Ph.D. student (E-mail: brownb1@live.unc.edu), Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599.

[†]Kai Zhang is an Associate Professor (E-mail: zhangk@email.unc.edu), Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599.

case, the vector $\boldsymbol{S_X}$ has two notable entries, which decompose the non-uniformity of $\boldsymbol{P_X}$ into two orthogonal signals. The largest entry of $\boldsymbol{S_X}$ in absolute value is -0.10, corresponding to the third row of $\mathbf{H}_8$:

$$\begin{pmatrix} 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \end{pmatrix}.$$

We can interpret this correspondence in the following way: the distribution of $\boldsymbol{X}$ has a coarse Venetian blind pattern relative to $\boldsymbol{Y}$. The second largest entry of $\boldsymbol{S_X}$ in absolute value is -0.08, due to the seventh row of $\mathbf{H}_8$:

$$\begin{pmatrix} 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix}.$$

From this, we see that the $\boldsymbol{X}$ points are centrally concentrated relative to the points of $\boldsymbol{Y}$. We would expect the interquartile region of $\boldsymbol{Y}$ to contain over half of the points of $\boldsymbol{X}$.

An example visualization. In practice, we may wish to visualize the largest imbalance recorded in $\boldsymbol{S_X}$ – one possible visualization at a depth of $d = 3$ is given in Figure 1. The rationale for this plot is as follows: let $R_1, \ldots, R_8$ be real intervals such that $1/2^d = 1/8$ of the $\boldsymbol{Y}$ sample is contained in each R_i. These eight intervals correspond to the cells of $\boldsymbol{P_X}$: if $\boldsymbol{P}_{\boldsymbol{X},i}$ is large (small), we would expect R_i to contain more (less) than $1/8$ of the $\boldsymbol{X}$ sample. In the context of testing, the largest asymmetry in $\boldsymbol{S_X}$ can be thought of as the primary reason for rejection of the null. For the simulated $\boldsymbol{X}$ and $\boldsymbol{Y}$ data of Figure 1, the largest asymmetry recorded in $\boldsymbol{S_X}$ corresponds to the regions R_3, R_4, R_7, and R_8.

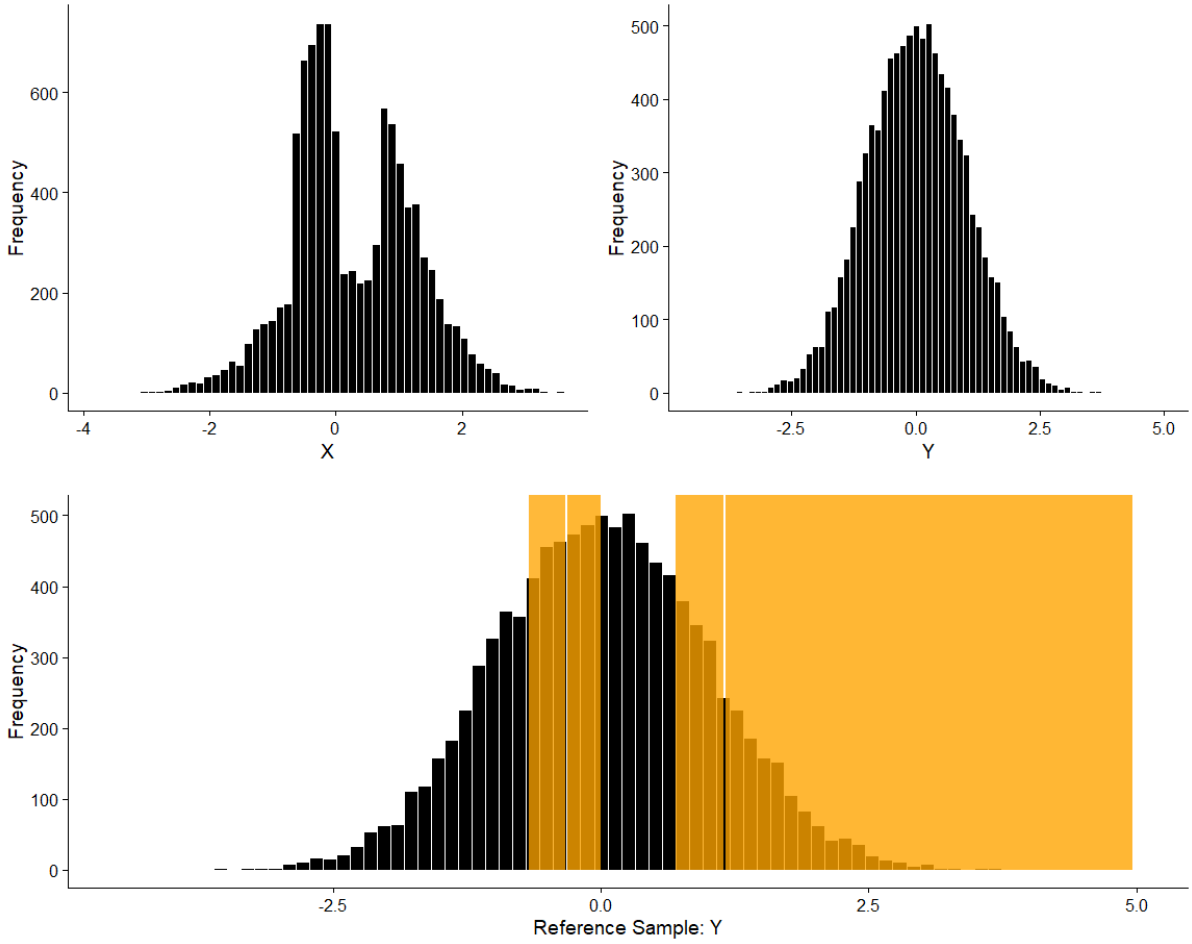

Figure 1: Visualization of a symmetry statistic.

0.2 Proofs on Subsampling and Computation

Here, we prove Theorem 1, Proposition 1, and Theorem 2 from the main paper.

Theorem 0.1. *Let $\boldsymbol{Y}$ be a fixed vector of observed data, and let x be a real number. Consider the following bootstrap method for computing a vector $\boldsymbol{P}^*(x)$ using K subsamples from $\boldsymbol{Y}$.*

1. *Take bootstrap subsamples $\boldsymbol{Y}_k^*$ of size $2^{d+1}-1$ from $\boldsymbol{Y}$ without replacement, for subsamples $1 \leq k \leq K$.*
2. *Compute $\widehat{F}_{\boldsymbol{Y}_k^*}(x)$, for subsamples $1 \leq k \leq K$.*
3. *Set $\boldsymbol{P}_i^*(x) = \#\left\{k : \widehat{F}_{\boldsymbol{Y}_k^*}(x) \in \left[\frac{i-1}{2^d}, \frac{i}{2^d}\right)\right\}/K$, for coordinates $1 \leq i \leq 2^d$.*

It follows that

$$pr\left(\lim_{K\to\infty} \boldsymbol{P}^*(x) = \boldsymbol{P}(x)\right) = 1,$$

where the probability is taken over the randomness of the subsampling.

Proof. As $\boldsymbol{Y}$ is assumed to be deterministic for the purposes of this theorem, all randomness is due to resampling. Referring to the discussion above the theorem statement, we know

$$\boldsymbol{P}_i(x) = \mathbf{P}\left(\widehat{F}_{\boldsymbol{Y}^*}(x) \in \left[\frac{i-1}{2^d}, \frac{i}{2^d}\right)\right)$$

for a random resample $\boldsymbol{Y}^*$ of length $2^{d+1}-1$ from $\boldsymbol{Y}$. We can rewrite $\boldsymbol{P}_i^*(x)$ as an average of K independent indicator variables

$$\boldsymbol{P}_i^*(x) = \frac{\sum_{k=1}^K I\left(\widehat{F}_{\boldsymbol{Y}_k^*}(x) \in \left[\frac{i-1}{2^d}, \frac{i}{2^d}\right)\right)}{K}.$$

By the law of large numbers, $\boldsymbol{P}_i^*(x) \to \boldsymbol{P}_i(x)$ almost surely as $K \to \infty$. As $\boldsymbol{P}(x)$ has finitely many coordinates, this convergence holds for every $1 \leq i \leq 2^d$ almost surely. □

Proposition 0.1. *Suppose* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *satisfy* $\max_i\{\boldsymbol{X}_i\} < \min_j\{\boldsymbol{Y}_j\}$. *Then* $\cos(\theta) = -(2^d - 1)^{-1}$, *where* θ *is the angle between* $\boldsymbol{S_X}$ *and* $\boldsymbol{S_Y}$ *as vectors in* $\mathbb{R}^{2^d-1}$.

Proof. As every $\boldsymbol{X}$ value is smaller than every $\boldsymbol{Y}$ value, $\widehat{F}_{\boldsymbol{Y}^*}(\boldsymbol{X}_i) = 0$ and $\widehat{F}_{\boldsymbol{X}^*}(\boldsymbol{Y}_j) = 1$ for every resample $\boldsymbol{X}^*$ and $\boldsymbol{Y}^*$ and every i and j. Hence, $\boldsymbol{P_X} = \boldsymbol{e}_1$ and $\boldsymbol{P_Y} = \boldsymbol{e}_{2^d}$, where $\boldsymbol{e}_k$ is the vector of length 2^d with a 1 in coordinate k and a 0 everywhere else.

Let $\widetilde{\mathbf{H}}_{2^d}$ denote $\mathbf{H}_{2^d}$ without its first row $\begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$. In particular, this means that $\boldsymbol{S_X} = \widetilde{\mathbf{H}}_{2^d}\boldsymbol{P_X}$ and $\boldsymbol{S_Y} = \widetilde{\mathbf{H}}_{2^d}\boldsymbol{P_Y}$. The $2^d \times 2^d$ matrix $(\widetilde{\mathbf{H}}_{2^d})^T\widetilde{\mathbf{H}}_{2^d}$ is equal to $2^d - 1$ along the diagonal and -1 everywhere else. In addition, $\widetilde{\mathbf{H}}_{2^d}\boldsymbol{P_X}$ and $\widetilde{\mathbf{H}}_{2^d}\boldsymbol{P_Y}$ are equal to either 1 or -1 in every coordinate. As a result, we have

$$\begin{aligned}\cos(\theta) &= \frac{\boldsymbol{e}_1(\widetilde{\mathbf{H}}_{2^d})^T\widetilde{\mathbf{H}}_{2^d}\boldsymbol{e}_{2^d}}{\|\widetilde{\mathbf{H}}_{2^d}\boldsymbol{P_X}\|_2\|\widetilde{\mathbf{H}}_{2^d}\boldsymbol{P_Y}\|_2} \\ &= \frac{-1}{2^d - 1}.\end{aligned}$$

□

Theorem 0.2. *There exists an algorithm for calculating the test statistic* S *that requires* $O((m+n)\log(m+n))$ *elementary operations.*

Proof. Referring to Algorithm 3, note that lines 1-4 take $O(m+n)$ operations. With any efficient sorting algorithm, line 5 has an average case of $O((m+n)\log(m+n))$ operations. Line 6 is constant in n and m, and the loop spanning lines 7-21 iterates $m+n$ times and executes each iteration in constant time. The vector and matrix operations in lines 22-24 have running time independent of n and m. Thus, line 5 is the bottleneck, and the overall running time has leading term $O((m+n)\log(m+n))$. □

Algorithm 3 AUGUST+$(\boldsymbol{X}, \boldsymbol{Y}, d)$

Define $m = \text{length}(\boldsymbol{X})$, $n = \text{length}(\boldsymbol{Y})$, and $r = 2^{d+1} - 1$
Initialize empty matrix $\mathbf{M}$ of dimension $2 \times (m + n)$
Assign the first row of $\mathbf{M}$ to the concatenated vector $(\boldsymbol{X}^T, \boldsymbol{Y}^T)$
Assign the second row of $\mathbf{M}$ to a row vector with m entries equal to 1 followed by n entries equal to 0
Sort the columns of $\mathbf{M}$ ascending by the entries in the first row of $\mathbf{M}$
Initialize integers $c_x, c_y = 0$ and vectors $\boldsymbol{P_X}, \boldsymbol{P_Y} = \mathbf{0}_{2^d}$
for $i = 1$ to $(m + n)$ **do**
  if $\mathbf{M}_{2,i} = 1$ **then**
    $c_x = c_x + 1$
    for $j = 1$ to 2^d **do**
      $k = 2j - 2$
      $\boldsymbol{P_X}_{,j} = \boldsymbol{P_X}_{,j} + \frac{\binom{c_y}{k}\binom{n-c_y}{r-k}}{\binom{n}{r}} + \frac{\binom{c_y}{k+1}\binom{n-c_y}{r-k-1}}{\binom{n}{r}}$
    end for
  else
    $c_y = c_y + 1$
    for $j = 1$ to 2^d **do**
      $k = 2j - 2$
      $\boldsymbol{P_Y}_{,j} = \boldsymbol{P_Y}_{,j} + \frac{\binom{c_x}{k}\binom{m-c_x}{r-k}}{\binom{n}{r}} + \frac{\binom{c_x}{k+1}\binom{m-c_x}{r-k-1}}{\binom{m}{r}}$
    end for
  end if
end for
Assign $\boldsymbol{P_X} = \boldsymbol{P_X}/m$ and $\boldsymbol{P_Y} = \boldsymbol{P_Y}/n$
Assign $\boldsymbol{S_X} = (\mathbf{H}_{2^d}\boldsymbol{P_X})_{-1}$ and $\boldsymbol{S_Y} = (\mathbf{H}_{2^d}\boldsymbol{P_Y})_{-1}$
Return the test statistic $S = -\boldsymbol{S_X}^T\boldsymbol{S_Y}$

0.3 Proofs on Limiting Distributions

The remainder of the supplementary materials is dedicated to proving Theorem 3 from the main paper, which appears here as Theorem 0.7.

Lemma 0.1 (Orthogonality of the projection)**.** *Let $\boldsymbol{U} \in \mathbb{R}^p$ be a random vector, and let $\{W_i\}_{i=1}^N$ be a collection of N independent observations. Define the projection $\widehat{\boldsymbol{U}} = \mathbf{E}[\boldsymbol{U}] + \sum_{i=1}^N \mathbf{E}[\boldsymbol{U} - \mathbf{E}[\boldsymbol{U}]|W_i]$. Then*

$$\mathbf{E}[(\boldsymbol{U} - \widehat{\boldsymbol{U}})\widehat{\boldsymbol{U}}^T] = \mathbf{0}_{p\times p}$$

Proof. This follows from expressing $(\boldsymbol{U} - \widehat{\boldsymbol{U}})\widehat{\boldsymbol{U}}^T$ entry-wise and cancelling terms using properties of conditional expectation. □

Lemma 0.2 (Closeness of the projection)**.** *Let $\{W_i\}_{i=1}^\infty$ be an independent collection of random variables. Let $\{\boldsymbol{U}_N\}_{N=1}^\infty$ be a sequence of non-degenerate random vectors of length p. For each N, define the projection $\widehat{\boldsymbol{U}}_N = \mathbf{E}[\boldsymbol{U}_N] + \sum_{i=1}^N \mathbf{E}[\boldsymbol{U}_N - \mathbf{E}[\boldsymbol{U}_N]|W_i]$. Let $\boldsymbol{\Sigma}_{1,N} = Cov(\boldsymbol{U}_N)$ and $\boldsymbol{\Sigma}_{2,N} = Cov(\widehat{\boldsymbol{U}}_N)$. If $\boldsymbol{\Sigma}_{1,N}\boldsymbol{\Sigma}_{2,N}^{-1} \to I$ as $N \to \infty$, then*

$$\boldsymbol{\Sigma}_{1,N}^{-\frac{1}{2}}(\boldsymbol{U}_N - \mathbf{E}[\boldsymbol{U}_N]) - \boldsymbol{\Sigma}_{2,N}^{-\frac{1}{2}}\left(\widehat{\boldsymbol{U}}_N - \mathbf{E}[\widehat{\boldsymbol{U}}_N]\right) \xrightarrow{p} 0.$$

Proof. Note that the difference above has expectation 0. To show convergence in L^2, it is enough to show that the covariance matrix of this difference converges to $\mathbf{0}$. We have

$$\begin{aligned}
&\mathrm{Cov}\left(\boldsymbol{\Sigma}_{1,N}^{-\frac{1}{2}}(\boldsymbol{U}_N - \mathbf{E}[\boldsymbol{U}_N]) - \boldsymbol{\Sigma}_{2,N}^{-\frac{1}{2}}\left(\widehat{\boldsymbol{U}}_N - \mathbf{E}[\widehat{\boldsymbol{U}}_N]\right)\right)\\
&= 2I - \boldsymbol{\Sigma}_{1,N}^{-\frac{1}{2}}\mathrm{Cov}(\boldsymbol{U}_N, \widehat{\boldsymbol{U}}_N)\boldsymbol{\Sigma}_{2,N}^{-\frac{1}{2}} - \boldsymbol{\Sigma}_{2,N}^{-\frac{1}{2}}\mathrm{Cov}(\widehat{\boldsymbol{U}}_N, \boldsymbol{U}_N)\boldsymbol{\Sigma}_{1,N}^{-\frac{1}{2}}\\
&= 2I - \boldsymbol{\Sigma}_{1,N}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{2,N}\boldsymbol{\Sigma}_{2,N}^{-\frac{1}{2}} - \boldsymbol{\Sigma}_{2,N}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{2,N}\boldsymbol{\Sigma}_{1,N}^{-\frac{1}{2}} \qquad \text{(Lemma 1 and } \mathbf{E}[\boldsymbol{U}_N] = \mathbf{E}[\widehat{\boldsymbol{U}}_N])\\
&\to \mathbf{0}.
\end{aligned}$$

□

Theorem 0.3. *Let $N = n + m$, and assume that $n, m \to \infty$ in such a way that $m/N \to \lambda$ for some $\lambda \in (0, 1)$. Define the cross-covariance matrices*

$$\xi_{i,j} = Cov\Bigg(\boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_s), \\ \boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \boldsymbol{X}'_{i+1}, \ldots, \boldsymbol{X}'_r, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_j, \boldsymbol{Y}'_{j+1}, \ldots, \boldsymbol{Y}'_s)\Bigg).$$

If $\boldsymbol{\Sigma} = r^2\xi_{1,0}/\lambda + s^2\xi_{0,1}/(1-\lambda)$ is invertible, then

$$\sqrt{N}\,(\boldsymbol{U} - \boldsymbol{\theta}) \xrightarrow{d} N(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Proof. We use as a template the notation and technique of Van der Vaart (2000), which handles the case where $\boldsymbol{k}$ is a scalar function. We are unable to find a reference for the case of a two–sample vector–valued U-statistic, though such work is very likely to have been done somewhere. Define

$$\boldsymbol{k}_{1,0}(x) = \mathbf{E}[\boldsymbol{k}(x, \ldots, \boldsymbol{X}_r, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_s)] - \boldsymbol{\theta}$$
$$\boldsymbol{k}_{0,1}(y) = \mathbf{E}[\boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_r, y, \ldots, \boldsymbol{Y}_s)] - \boldsymbol{\theta}.$$

Simplifying conditional expectations, the projection of $\boldsymbol{U} - \boldsymbol{\theta}$ has form

$$\begin{aligned}\widehat{\boldsymbol{U}} &:= \sum_{i=1}^{m} \mathbf{E}[\boldsymbol{U} - \boldsymbol{\theta}|\boldsymbol{X}_i] + \sum_{j=1}^{n} \mathbf{E}[\boldsymbol{U} - \boldsymbol{\theta}|\boldsymbol{Y}_j] \\ &= \sum_{i=1}^{m} \frac{\binom{m-1}{r-1}}{\binom{m}{r}} \boldsymbol{k}_{1,0}(\boldsymbol{X}_i) + \sum_{j=1}^{n} \frac{\binom{n-1}{s-1}}{\binom{n}{s}} \boldsymbol{k}_{0,1}(\boldsymbol{Y}_j) \\ &= \frac{r}{m} \sum_{i=1}^{m} \boldsymbol{k}_{1,0}(\boldsymbol{X}_i) + \frac{s}{n} \sum_{j=1}^{n} \boldsymbol{k}_{0,1}(\boldsymbol{Y}_j).\end{aligned}$$

Define the auto-covariance matrices $\boldsymbol{\Sigma}_1 := \mathrm{Cov}(\boldsymbol{U})$ and $\boldsymbol{\Sigma}_2 := \mathrm{Cov}(\widehat{\boldsymbol{U}})$. Observe that $\mathrm{Cov}(\boldsymbol{k}_{1,0}(\boldsymbol{X}_i)) = \xi_{1,0}$ and $\mathrm{Cov}(\boldsymbol{k}_{0,1}(\boldsymbol{Y}_j)) = \xi_{0,1}$. Then based on the expression above and the

mutual independence of all $\boldsymbol{X}_i$ and $\boldsymbol{Y}_j$, we can compute $\boldsymbol{\Sigma}_2 = \frac{r^2}{m}\xi_{1,0} + \frac{s^2}{n}\xi_{0,1}$. From this, we see that $N\boldsymbol{\Sigma}_2$ converges to $r^2\xi_{1,0}/\lambda + s^2\xi_{0,1}/(1-\lambda)$ as $N \to \infty$.

In addition, expanding $\mathrm{Cov}(\boldsymbol{U})$ and counting terms gives

$$\begin{aligned}\boldsymbol{\Sigma}_1 &= \frac{1}{\binom{m}{r}^2\binom{n}{s}^2}\sum_{i=0}^{r}\sum_{j=0}^{s}\binom{m}{r}\binom{r}{i}\binom{m-r}{r-i}\binom{n}{s}\binom{s}{j}\binom{n-s}{s-j}\xi_{i,j} \\ &= \frac{1}{\binom{m}{r}\binom{n}{s}}\sum_{i=0}^{r}\sum_{j=0}^{s}\binom{r}{i}\binom{m-r}{r-i}\binom{s}{j}\binom{n-s}{s-j}\xi_{i,j}.\end{aligned}$$

Examining this expression of $\boldsymbol{\Sigma}_1$, we see that the terms of highest order correspond to $(i,j) = (1,0)$ and $(i,j) = (0,1)$. (In particular, $\boldsymbol{\xi}_{0,0} = \mathbf{0}$ by independence.) From the form of these two leading terms, it follows that $N\boldsymbol{\Sigma}_1$ converges to $r^2\xi_{1,0}/\lambda + s^2\xi_{0,1}/(1-\lambda)$, the same limit as $N\boldsymbol{\Sigma}_2$.

As a result, the hypothesis of Lemma 2 is satisfied. Lemma 2 and Slutsky imply that $\sqrt{N}(\boldsymbol{U} - \boldsymbol{\theta} - \widehat{\boldsymbol{U}}) \xrightarrow{p} \mathbf{0}$. By the multivariate CLT,

$$\sqrt{N}\widehat{\boldsymbol{U}} \xrightarrow{d} N\Big(\mathbf{0}, r^2\xi_{1,0}/\lambda + s^2\xi_{0,1}/(1-\lambda)\Big),$$

which gives the desired result. □

Theorem 0.4. *If $\boldsymbol{\Sigma} = r^2\xi_{1,0}/\lambda + s^2\xi_{0,1}/(1-\lambda)$ has rank $q \geq 1$, then*

$$\sqrt{N}\,(\boldsymbol{U} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

Proof. Recall from the proof of Theorem 0.3 that $\xi_{1,0}$ and $\xi_{0,1}$ can be written as auto-covariance matrices of certain random vectors. As a result, $\boldsymbol{\Sigma}$ is positive semi-definite, and $N(\mathbf{0}, \boldsymbol{\Sigma})$ is a well-defined distribution supported on a q-dimensional subspace $H \subseteq \mathbb{R}^p$.

Let $\mathbf{B}$ be a $p \times q$ matrix whose columns form an orthonormal basis for H. In particular, note that the nullspace of $\boldsymbol{\Sigma}$ is $H^\perp$, which implies $\mathbf{B}\mathbf{B}^T\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{B}\mathbf{B}^T = \boldsymbol{\Sigma}$. Now, observe that $\mathbf{B}^T\boldsymbol{Z}$ is a non-degenerate distribution in $\mathbb{R}^q$ for $\boldsymbol{Z} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$. As a result, $\mathrm{Cov}(\mathbf{B}^T\boldsymbol{Z}) =$

$\mathbf{B}^T\mathbf{\Sigma}\mathbf{B}$ is positive definite, and Theorem 0.3 gives

$$\mathbf{B}^T\left[\sqrt{N}\left(\boldsymbol{U}-\boldsymbol{\theta}\right)\right] \xrightarrow{d} N(\mathbf{0}, \mathbf{B}^T\mathbf{\Sigma}\mathbf{B}).$$

By the continuous mapping theorem, we have

$$\mathbf{B}\mathbf{B}^T\left[\sqrt{N}\left(\boldsymbol{U}-\boldsymbol{\theta}\right)\right] \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}).$$

To conclude the proof, we must show that $\mathbf{B}\mathbf{B}^T\left[\sqrt{N}\left(\boldsymbol{U}-\boldsymbol{\theta}\right)\right]$ is close to $\sqrt{N}\left(\boldsymbol{U}-\boldsymbol{\theta}\right)$. Let $\boldsymbol{V}_N = \sqrt{N}\left(\boldsymbol{U}-\boldsymbol{\theta}\right)$ and $\mathbf{\Sigma}_N = \mathrm{Cov}(\boldsymbol{V}_N)$. Note $\mathbf{E}[\boldsymbol{V}_N - \mathbf{B}\mathbf{B}^T\boldsymbol{V}_N] = \mathbf{0}$, and

$$\begin{aligned}
&\mathbf{E}[(\boldsymbol{V}_N - \mathbf{B}\mathbf{B}^T\boldsymbol{V}_N)(\boldsymbol{V}_N - \mathbf{B}\mathbf{B}^T\boldsymbol{V}_N)^T] \\
&= \mathbf{E}\left[\mathbf{V_N}\mathbf{V_N^T} - \mathbf{V_N}\mathbf{V_N^T}\mathbf{B}\mathbf{B^T} - \mathbf{B}\mathbf{B^T}\mathbf{V_N}\mathbf{V_N^T} + \mathbf{B}\mathbf{B^T}\mathbf{V_N}\mathbf{V_N^T}\mathbf{B}\mathbf{B^T}\right] \\
&= \mathbf{\Sigma}_N - \mathbf{\Sigma}_N^T\mathbf{B}\mathbf{B}^T - \mathbf{B}\mathbf{B}^T\mathbf{\Sigma}_N + \mathbf{B}\mathbf{B}^T\mathbf{\Sigma}_N\mathbf{B}\mathbf{B}^T \\
&\to \mathbf{0}.
\end{aligned}$$

Thus, $\boldsymbol{V}_N - \mathbf{B}\mathbf{B}^T\boldsymbol{V}_N \xrightarrow{L^2} \mathbf{0}$. □

Recall that $d \in \mathbb{N}$ is the fixed binary depth and $\boldsymbol{h} : \mathbb{R} \times \mathbb{R}^{2^{d+1}-1} \to \mathbb{R}^{2^d}$ be given by

$$\boldsymbol{h}_k(x, \boldsymbol{y}) = \begin{cases} 1 & \text{if } \#\{j : \boldsymbol{y}_j \le x\} = 2k-2 \text{ or } 2k-1 \\ 0 & \text{otherwise.} \end{cases}$$

The following key lemma explains how $\boldsymbol{P_X}$ can be expressed using $\boldsymbol{h}$.

Lemma 0.3. *With $\boldsymbol{h}$ as defined above, it holds that*

$$\frac{1}{\binom{n}{2^{d+1}-1}}\sum_{\beta}\boldsymbol{h}\left(x, \boldsymbol{Y}_{\beta_1}, \boldsymbol{Y}_{\beta_2}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}\right) = \textit{Algorithm 1}(\boldsymbol{Y}, d, x).$$

Consequently,

$$\frac{1}{m\binom{n}{2^{d+1}-1}} \sum_i \sum_\beta \boldsymbol{h}\left(\boldsymbol{X}_i, \boldsymbol{Y}_{\beta_1}, \boldsymbol{Y}_{\beta_2}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}\right) = \boldsymbol{P_X}.$$

Proof. To see the first equality, fix $k \in \{1, \ldots, 2^d\}$, and consider the kth coordinate of Algorithm 1$(\boldsymbol{Y}, d, x)$. By definition, this is equal to the probability that after choosing $2^{d+1}-1$ random elements from $\boldsymbol{Y}$ without replacement, exactly $2k-2$ or $2k-1$ of these elements will be less than or equal to x.

For each combination $\boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}$ of elements from $\boldsymbol{Y}$, the kth coordinate of

$$\boldsymbol{h}\left(x, \boldsymbol{Y}_{\beta_1}, \boldsymbol{Y}_{\beta_2}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}\right)$$

is an indicator of the event that $2k-2$ or $2k-1$ components of the combination $\boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}$ are less than or equal to x. Thus, the average

$$\frac{1}{\binom{n}{2^{d+1}-1}} \sum_\beta \boldsymbol{h}_k\left(x, \boldsymbol{Y}_{\beta_1}, \boldsymbol{Y}_{\beta_2}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}\right)$$

over all combinations β is precisely the probability that a randomly chosen combination will contain $2k-2$ or $2k-1$ elements that are less than or equal to x.

The second equality follows from the definition of the vector $\boldsymbol{P_X}$.

For intuition on this result, we can look to the classic urn model. Consider an urn with n balls: one red ball for each $\boldsymbol{Y}_i \leq x$, and one black ball for each $\boldsymbol{Y}_i > x$. Subsampling $2^{d+1}-1$ points from $\boldsymbol{Y}$ is equivalent to drawing $2^{d+1}-1$ balls from the urn. In this case, the kth coordinate of $\boldsymbol{h}$ is an indicator of the event that exactly $2k-2$ or $2k-1$ red balls were drawn. By averaging $\boldsymbol{h}_k$ over every possible combination of balls from the urn, we compute the probability of this event. Computing the probability this way is inefficient compared to the obvious hypergeometric approach, but this form ultimately allows us to write $\begin{pmatrix} \boldsymbol{S_X} \\ \boldsymbol{S_Y} \end{pmatrix}$ as

a U-statistic. $\square$

Theorem 0.5. *There exists a kernel function*

$$\boldsymbol{k} : \mathbb{R}^{2^{d+1}-1} \times \mathbb{R}^{2^{d+1}-1} \to \mathbb{R}^{2^d-1} \times \mathbb{R}^{2^d-1}$$

such that

$$\frac{1}{\binom{m}{2^{d+1}-1}\binom{n}{2^{d+1}-1}} \sum_{\alpha} \sum_{\beta} \boldsymbol{k}(\boldsymbol{X}_{\alpha_1}, \ldots, \boldsymbol{X}_{\alpha_{2^{d+1}-1}}, \boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}) = \begin{pmatrix} \boldsymbol{S_X} \\ \boldsymbol{S_Y} \end{pmatrix}.$$

Proof. Let $\boldsymbol{g}_1 : \mathbb{R}^{2^{d+1}-1} \times \mathbb{R}^{2^{d+1}-1} \to \mathbb{R}^{2^d}$ be given by

$$\boldsymbol{g}_1(x_1, \ldots, x_{2^{d+1}-1}, y_1, \ldots, y_{2^{d+1}-1}) = \sum_{i=1}^{2^{d+1}-1} \boldsymbol{h}(x_i, y_1, \ldots, y_{2^{d+1}-1}).$$

For each combination β, $\boldsymbol{g}_1$ has the important property that

$$\begin{aligned} &\sum_{\alpha} \boldsymbol{g}_1(\boldsymbol{X}_{\alpha_1}, \ldots, \boldsymbol{X}_{\alpha_{2^{d+1}-1}}, \boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}) \\ &= \binom{m-1}{2^{d+1}-2} \sum_{i=1}^{m} \boldsymbol{h}(\boldsymbol{X}_i, \boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}). \end{aligned}$$

This is because each $\boldsymbol{X}_i$ appears in $\binom{m-1}{2^{d+1}-2}$ combinations of length $2^{d+1}-1$ of elements from $\boldsymbol{X}$. Consequently, we have by Lemma 0.3

$$\begin{aligned} &\frac{1}{\binom{m}{2^{d+1}-1}\binom{n}{2^{d+1}-1}} \sum_{\alpha} \sum_{\beta} \boldsymbol{g}_1(\boldsymbol{X}_{\alpha_1}, \ldots, \boldsymbol{X}_{\alpha_{2^{d+1}-1}}, \boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}) \\ &= \frac{\binom{m-1}{2^{d+1}-2}}{\binom{m}{2^{d+1}-1}\binom{n}{2^{d+1}-1}} \sum_{i=1}^{m} \sum_{\beta} \boldsymbol{h}(\boldsymbol{X}_i, \boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}) \\ &= \frac{2^{d+1}-1}{m\binom{n}{2^{d+1}-1}} \sum_{i=1}^{m} \sum_{\beta} \boldsymbol{h}(\boldsymbol{X}_i, \boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}) \\ &= (2^{d+1}-1)\boldsymbol{P_X}. \end{aligned}$$

Defining $\boldsymbol{g}_2 : \mathbb{R}^{2^{d+1}-1} \times \mathbb{R}^{2^{d+1}-1} \to \mathbb{R}^{2^d}$ to be

$$\boldsymbol{g}_2(x_1, \ldots, x_{2^{d+1}-1}, y_1, \ldots, y_{2^{d+1}-1}) = \sum_{i=1}^{2^{d+1}-1} \boldsymbol{h}(y_i, x_1, \ldots, x_{2^{d+1}-1}),$$

an identical argument shows that

$$\begin{aligned}
&\frac{1}{\binom{m}{2^{d+1}-1}\binom{n}{2^{d+1}-1}} \sum_{\alpha} \sum_{\beta} \boldsymbol{g}_2(\boldsymbol{X}_{\alpha_1}, \ldots, \boldsymbol{X}_{\alpha_{2^{d+1}-1}}, \boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}) \\
&= (2^{d+1}-1)\boldsymbol{P}_{\boldsymbol{Y}}.
\end{aligned}$$

We define $\boldsymbol{g}$ to be the concatenation of $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, namely

$$\boldsymbol{g} = \begin{pmatrix} \boldsymbol{g}_1 \\ \boldsymbol{g}_2 \end{pmatrix}.$$

By our work with $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$, we obtain

$$\begin{aligned}
&\frac{1}{\binom{n}{2^{d+1}-1}\binom{m}{2^{d+1}-1}} \sum_{\alpha} \sum_{\beta} \boldsymbol{g}(\boldsymbol{X}_{\alpha_1}, \ldots, \boldsymbol{X}_{\alpha_{2^{d+1}-1}}, \boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}) \\
&= (2^{d+1}-1) \begin{pmatrix} \boldsymbol{P}_{\boldsymbol{X}} \\ \boldsymbol{P}_{\boldsymbol{Y}} \end{pmatrix}.
\end{aligned}$$

Let $\widetilde{\mathbf{H}}_{2^d}$ denote $\mathbf{H}_{2^d}$ without its first row $\begin{pmatrix} 1 & 1 & \ldots & 1 \end{pmatrix}$. Define the kernel function $\boldsymbol{k}$ in terms of $\boldsymbol{g}$ with

$$\boldsymbol{k} = \frac{1}{2^{d+1}-1} \begin{pmatrix} \widetilde{\mathbf{H}}_{2^d} & \mathbf{0}_{(2^d-1)\times 2^d} \\ \mathbf{0}_{(2^d-1)\times 2^d} & \widetilde{\mathbf{H}}_{2^d} \end{pmatrix} \boldsymbol{g}.$$

As $\boldsymbol{S_X} = \widetilde{\mathbf{H}}_{2^d}\boldsymbol{P_X}$ and $\boldsymbol{S_Y} = \widetilde{\mathbf{H}}_{2^d}\boldsymbol{P_Y}$, it follows that

$$\frac{1}{\binom{n}{2^{d+1}-1}\binom{m}{2^{d+1}-1}} \sum_{\alpha}\sum_{\beta} \boldsymbol{k}(\boldsymbol{X}_{\alpha_1}, \ldots, \boldsymbol{X}_{\alpha_{2^{d+1}-1}}, \boldsymbol{Y}_{\beta_1}, \ldots, \boldsymbol{Y}_{\beta_{2^{d+1}-1}}) = \begin{pmatrix} \boldsymbol{S_X} \\ \boldsymbol{S_Y} \end{pmatrix}.$$

□

Theorem 0.6. *Suppose that we have univariate iid observations $\{\boldsymbol{X}_i\}_{i=1}^m$ and $\{\boldsymbol{Y}_j\}_{j=1}^n$ under the null. Let $N = n + m$, and assume that $n, m \to \infty$ in such a way that $m/N \to \lambda$ for some $\lambda \in (0, 1)$. Then*

$$\sqrt{N}\begin{pmatrix} \boldsymbol{S_X} \\ \boldsymbol{S_Y} \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}).$$

Defining the cross-covariance matrices

$$\xi_{i,j} = Cov\Bigg(\boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{d+1}-1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1}), \\ \boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \boldsymbol{X}'_{i+1}, \ldots, \boldsymbol{X}'_{2^{d+1}-1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_j, \boldsymbol{Y}'_{j+1}, \ldots, \boldsymbol{Y}'_{2^{d+1}-1})\Bigg),$$

the auto-covariance matrix $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = (2^{d+1} - 1)^2 \left(\xi_{1,0}/\lambda + \xi_{0,1}/(1 - \lambda)\right).$$

Proof. The asymptotic normality and limiting covariance are immediate consequences of Theorems 0.4 and 0.5, provided we argue that

$$\mathbf{E}[\boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{d+1}-1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1})] = \mathbf{0}$$

under the null.

Because the samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ come from the same distribution, the vectors $\boldsymbol{h}(\boldsymbol{X}_1, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1})$

and $\boldsymbol{h}(\boldsymbol{Y}_1, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{d+1}-1})$ are uniformly distributed over their support, which is the standard basis of $\boldsymbol{R}^{2^d}$. As a result,

$$\mathbf{E}[\boldsymbol{h}(\boldsymbol{X}_1, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1})] = \begin{pmatrix} \frac{1}{2^d} \\ \vdots \\ \frac{1}{2^d} \end{pmatrix}$$

and

$$\mathbf{E}[\boldsymbol{h}(\boldsymbol{Y}_1, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{d+1}-1})] = \begin{pmatrix} \frac{1}{2^d} \\ \vdots \\ \frac{1}{2^d} \end{pmatrix}.$$

From the definition of $\boldsymbol{k}$ and linearity of expectation, the equalities above imply

$$\mathbf{E}[\boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{d+1}-1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1})] = \mathbf{0}.$$

□

Finally, the following result appears as Theorem 3 in the main paper.

Theorem 0.7. *Suppose that we have univariate, independent observations* $\{\boldsymbol{X}_i\}_{i=1}^m$ *and* $\{\boldsymbol{Y}_j\}_{j=1}^n$*, where* $\boldsymbol{X}_i \sim G$ *and* $\boldsymbol{Y}_j \sim F$*. Let* $N = n + m$*, and assume that* $n, m \to \infty$ *in such a way that* $m/N \to \lambda$ *for some* $\lambda \in (0, 1)$*. Then*

$$\sqrt{N}\left(\begin{pmatrix} \boldsymbol{S_X} \\ \boldsymbol{S_Y} \end{pmatrix} - \boldsymbol{\mu}\right) \to N(\mathbf{0}, \boldsymbol{\Sigma})$$

in distribution, where $\boldsymbol{\mu}$ *is*

$$\boldsymbol{\mu} = \begin{pmatrix} \widetilde{\mathbf{H}}_{2^d} & \mathbf{0}_{(2^d-1)\times 2^d} \\ \mathbf{0}_{(2^d-1)\times 2^d} & \widetilde{\mathbf{H}}_{2^d} \end{pmatrix} \boldsymbol{p}^{F,G},$$

for $\boldsymbol{p}^{F,G} = \left(p_1^{F:G}, \ldots, p_{2^d}^{F:G}, p_1^{G:F}, \ldots, p_{2^d}^{G:F}\right)^T$. *Defining the cross-covariance matrices*

$$\xi_{i,j} = Cov\Bigg(\boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{d+1}-1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1}), \\ \boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \boldsymbol{X}'_{i+1}, \ldots, \boldsymbol{X}'_{2^{d+1}-1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_j, \boldsymbol{Y}'_{j+1}, \ldots, \boldsymbol{Y}'_{2^{d+1}-1})\Bigg),$$

where expectations are taken under the alternative, the auto-covariance matrix $\boldsymbol{\Sigma}$ *is*

$$\boldsymbol{\Sigma} = (2^{d+1}-1)^2 \left(\xi_{1,0}/\lambda + \xi_{0,1}/(1-\lambda)\right).$$

Proof. For each $k \in \{1, \ldots, 2^d\}$, define the function $p_k^F : \mathbb{R} \to [0,1]$ by

$$\begin{aligned} p_k^F(x) = &\binom{2^{d+1}-1}{2k-2} F(x)^{2k-2}(1-F(x))^{2^{d+1}-1-(2k-2)} \\ &+ \binom{2^{d+1}-1}{2k-1} F(x)^{2k-1}(1-F(x))^{2^{d+1}-1-(2k-1)}, \end{aligned}$$

and similarly define $p_k^G : \mathbb{R} \to [0,1]$ by

$$\begin{aligned} p_k^G(x) = &\binom{2^{d+1}-1}{2k-2} G(x)^{2k-2}(1-G(x))^{2^{d+1}-1-(2k-2)} \\ &+ \binom{2^{d+1}-1}{2k-1} G(x)^{2k-1}(1-G(x))^{2^{d+1}-1-(2k-1)}. \end{aligned}$$

Further, define the quantities

$$p_k^{F:G} = \int p_k^F(x) dG(x)$$

and

$$p_k^{G:F} = \int p_k^G(x) dF(x).$$

Theorems 0.4 and 0.5 do the heavy lifting, provided we show that

$$\boldsymbol{\mu} = \mathbf{E}[\boldsymbol{k}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{d+1}-1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1})]$$

under the alternative.

Recall that we wrote $\boldsymbol{k}$ in terms of the function $\boldsymbol{g}$ in Theorem 0.5:

$$\boldsymbol{k} = \frac{1}{2^{d+1}-1} \begin{pmatrix} \widetilde{\mathbf{H}}_{2^d} & \mathbf{0}_{(2^d-1)\times 2^d} \\ \mathbf{0}_{(2^d-1)\times 2^d} & \widetilde{\mathbf{H}}_{2^d} \end{pmatrix} \boldsymbol{g}.$$

As such, it is enough to show that

$$\frac{1}{2^{d+1}-1}\mathbf{E}[\boldsymbol{g}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{d+1}-1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1})] = \begin{pmatrix} p_1^{F:G} \\ \vdots \\ p_{2^d}^{F:G} \\ p_1^{G:F} \\ \vdots \\ p_{2^d}^{G:F} \end{pmatrix}.$$

We proceed coordinate-wise. Fix $k \in \{1, \ldots, 2^d\}$. Using the definition of $\boldsymbol{g}$, note that

$$\begin{aligned}
&\frac{1}{2^{d+1}-1}\mathbf{E}\left[\boldsymbol{g}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{2^{d+1}-1}, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1})\right]_k \\
&= \frac{1}{2^{d+1}-1}\mathbf{E}\left[\sum_{i=1}^{2^{d+1}-1} \boldsymbol{h}_k(\boldsymbol{X}_i, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1})\right] \\
&= \mathbf{E}[\boldsymbol{h}_k(\boldsymbol{X}_1, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{2^{d+1}-1})] \\
&= \mathbf{P}(\#\{i : \boldsymbol{Y}_i \leq \boldsymbol{X}_1\} \in \{2k-2, 2k-1\}) \\
&= \mathbf{E}_{\boldsymbol{X}_1}\left[\mathbf{P}\left(\#\{i : \boldsymbol{Y}_i \leq \boldsymbol{X}_1\} \in \{2k-2, 2k-1\} \middle| \boldsymbol{X}_1\right)\right].
\end{aligned}$$

However, the conditional probability

$$\mathbf{P}\left(\#\{i : \boldsymbol{Y}_i \leq \boldsymbol{X}_1\} \in \{2k-2, 2k-1\} \middle| \boldsymbol{X}_1 = x\right)$$

is precisely $p_k^F(x)$, meaning that

$$\begin{aligned}
\mathbf{E}_{\boldsymbol{X}_1}\left[\mathbf{P}\left(\#\{i : \boldsymbol{Y}_i \leq \boldsymbol{X}_1\} \in \{2k-2, 2k-1\} \middle| \boldsymbol{X}_1\right)\right] &= \mathbf{E}_{\boldsymbol{X}_1}\left[p_k^F(\boldsymbol{X}_1)\right] \\
&= p_k^{F:G} \\
&= \boldsymbol{\mu}_k.
\end{aligned}$$

A similar argument holds for $k \in \{2^d + 1, \ldots, 2^{d+1}\}$ as well, giving the desired result. □

0.4 Proof of Sensitivity to Non-uniform Moments

Here, we prove Theorem 4 from the main paper.

Theorem 0.8. *Let $t \geq 1$ be an integer, and assume the CDFs F and G are differentiable and strictly increasing on $\mathbb{R}$, with $Q = G \circ F^{-1}$. The following are equivalent:*

1. $\int \binom{2^t-1}{k} u^k (1-u)^{2^t-1-k} dQ(u) = 2^{-t}$ *for all integers* $k = 0, \ldots, 2^t - 1$.

2. $\int u^k dQ(u) = \mathbf{E}[U^k]$ *for all integers* $k = 0, \ldots, 2^t - 1$, *where* $U \sim$ *Unif*$[0, 1]$.

Proof. $(2 \Rightarrow 1)$ For each $k = 0, \ldots, 2^t - 1$, we know that $\binom{2^t-1}{k} u^k (1-u)^{2^t-1-k}$ is a polynomial in u with degree $2^t - 1$. Then by assumption, the integral of each monomial in the expansion of $\binom{2^t-1}{k} u^k (1-u)^{2^t-1-k}$ is the same whether integrating with respect to $dQ(u)$ or with respect to the uniform measure on $[0, 1]$. Thus, for U following the uniform distribution,

$$\begin{aligned}
\int \binom{2^t-1}{k} u^k (1-u)^{2^t-1-k} dQ(u) &= \binom{2^t-1}{k} \mathbf{E}[U^k (1-U)^{2^t-1-k}] \\
&= \binom{2^t-1}{k} \text{Beta}(k+1, 2^t-k) \\
&= \frac{(2^t-1)!}{k!(2^t-1-k)!} \frac{k!(2^t-1-k)!}{2^t!} \\
&= 2^{-t}.
\end{aligned}$$

$(1 \Rightarrow 2)$ We proceed by strong induction in k, starting with the base case $k = 2^t - 1$ and working our way down. First, plugging in $k = 2^t - 1$, we immediately have $\int u^{2^t-1} dQ(u) = 2^{-t}$ by assumption. Now, let v be a positive integer less than $2^t - 1$, and suppose we have $\int u^w dQ(u) = \mathbf{E}[U^w]$ for all integers $v + 1 \leq w \leq 2^t - 1$. We wish to show that part 2 of the theorem holds for the vth raw moment.

By assumption, we have $\int \binom{2^t-1}{v} u^v (1-u)^{2^t-1-v} dQ(u) = 2^{-t}$. Expanding by the binomial theorem, this means

$$\sum_{i=v}^{2^t-1} \left[(-1)^{i-v} \binom{2^t-1}{v} \binom{2^t-1-v}{i-v} \int u^i dQ(u) \right] = 2^{-t}. \tag{0.1}$$

For $v + 1 \leq i \leq 2^t - 1$, we know $\int u^i dQ(u) = \mathbf{E}[U^i]$. Considering (0.1) as a linear system with variable $\int u^v dQ(u)$, we see that the coefficient $\binom{2^t-1}{v}$ of $\int u^v dQ(u)$ is nonzero, so the solution to (0.1) is unique. Moreover, by our work in the $(2 \Rightarrow 1)$ direction of this proof, we see that $\int u^v dQ(u) = \mathbf{E}[U^v]$ is a solution, and it must therefore be the unique solution. $\square$

0.5 Additional Simulation Results

In this section, we provide additional empirical studies mentioned in the main text.

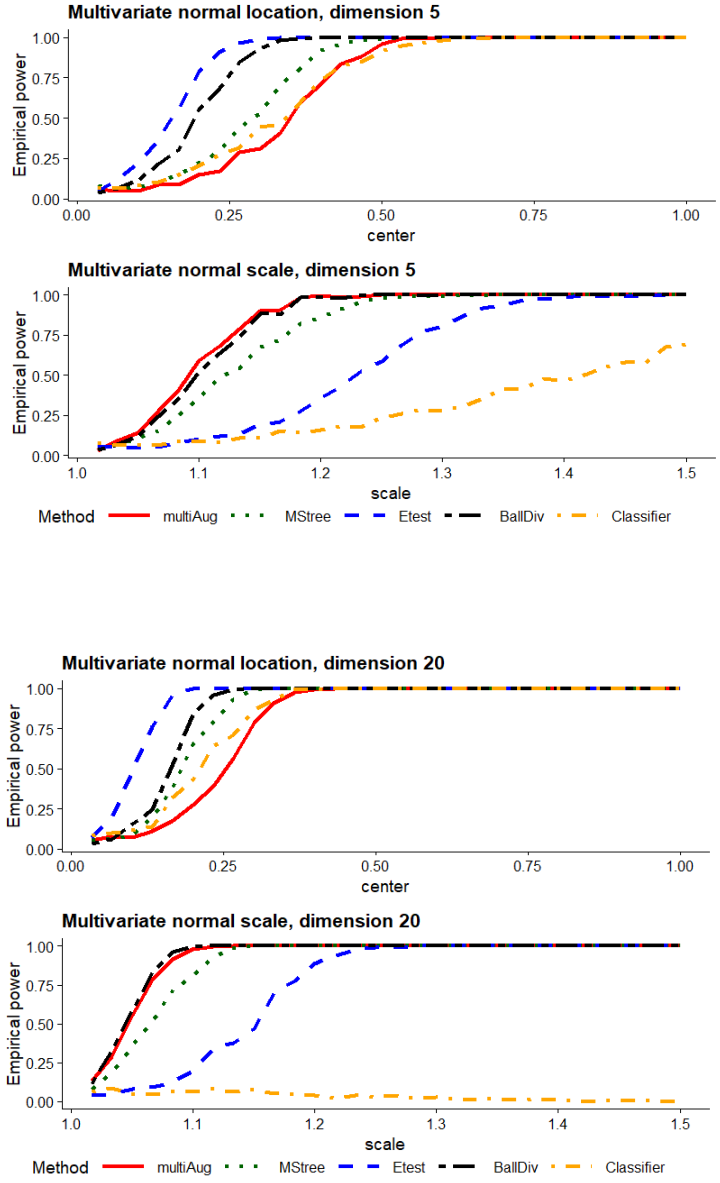

Figure 2: Comparison of power between AUGUST in red, Chen and Friedman (2017) in green, energy distance in blue, Pan et al. (2018) in black, and Lopez-Paz and Oquab (2016) in yellow. As found in lower dimensional studies, AUGUST's power is generally lower at location shift but higher at scale difference. Each method represents a tradeoff.

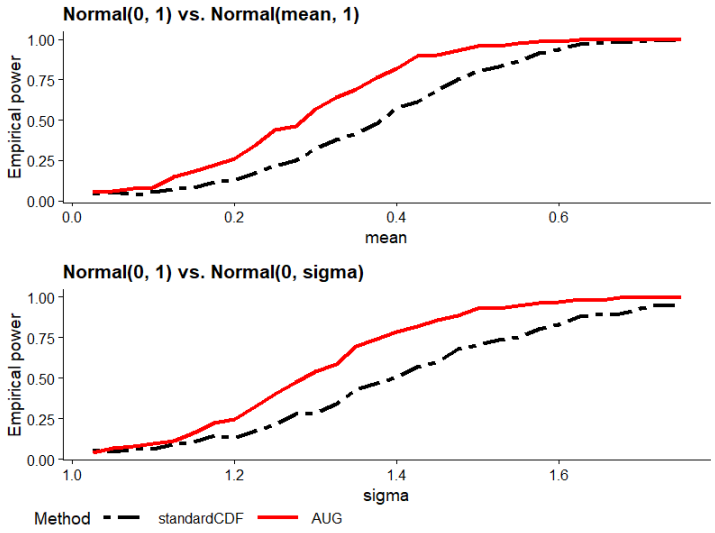

Figure 3: Comparison of power between augmented CDF transformation versus standard CDF transformation, both at a binary expansion depth of three.

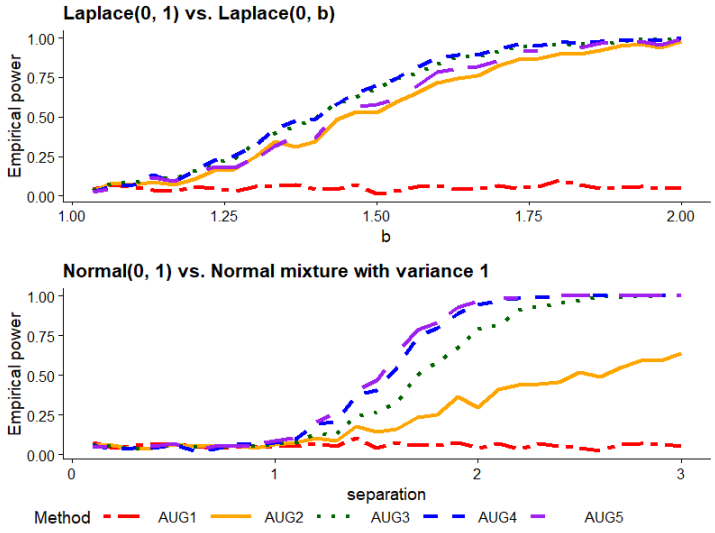

Figure 4: Comparison of power by depth for two alternatives.

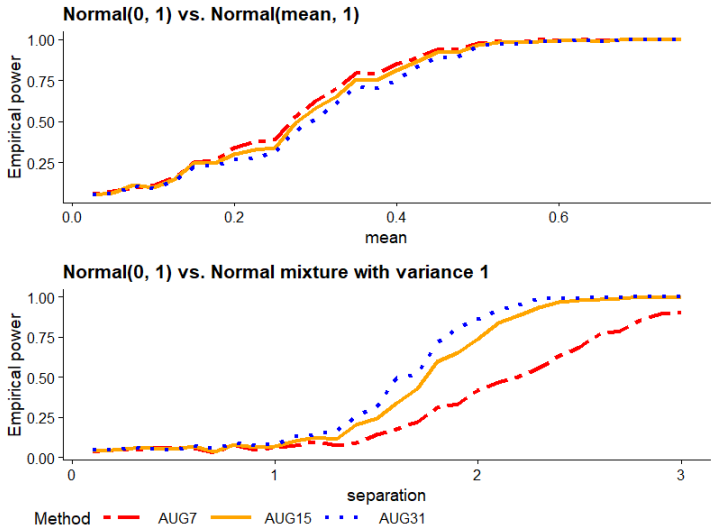

Figure 5: Comparison of power for AUGUST across varying subsample sizes ($r = 7, 15, 31$), all at a depth of $d = 3$.

References

Chen, H. and Friedman, J. H. (2017). A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association*, 112(517):397–409.

Lopez-Paz, D. and Oquab, M. (2016). Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*.

Pan, W., Tian, Y., Wang, X., and Zhang, H. (2018). Ball divergence: nonparametric two sample test. *Annals of statistics*, 46(3):1109.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.