The New England Journal of Statistics in Data Science Volume 2, 357–367 (2024)
DOI: https://doi.org/10.51387/23-NEJSDS54

AUGUST: An Interpretable, Resolution-based Two-sample Test

Benjamin BROWN* and Kai ZHANG

Abstract

Two-sample testing is a fundamental problem in statistics. While many powerful nonparametric methods exist for both the univariate and multivariate context, it is comparatively less common to see a framework for determining which data features lead to rejection of the null. In this paper, we propose a new nonparametric two-sample test named AUGUST, which incorporates a framework for interpretation while maintaining power comparable to existing methods. AUGUST tests for inequality in distribution up to a predetermined resolution using symmetry statistics from binary expansion. Designed for univariate and low to moderate-dimensional multivariate data, this construction allows us to understand distributional differences as a combination of fundamental orthogonal signals. Asymptotic theory for the test statistic facilitates p-value computation and power analysis, and an efficient algorithm enables computation on large data sets. In empirical studies, we show that our test has power comparable to that of popular existing methods, as well as greater power in some circumstances. We illustrate the interpretability of our method using NBA shooting data.

keywords and phrases: Distributional difference, Interpretability, Power, Symmetry, Visualization.

1. INTRODUCTION

1.1 Addressing the Two-sample Testing Problem

Two-sample tests are one of the most frequently used methods for statistical inference. While rooted in classical statistics, the two-sample problem is relevant to numerous cutting-edge applications, including high-energy physics [11], computer vision [14], and genome-wide expression analysis [36].

We begin with two samples $\boldsymbol{X}$ and $\boldsymbol{Y}$, which may be either univariate or multivariate. In the nonparametric setting, we make minimal assumptions regarding the distributions F and G used to generate $\boldsymbol{X}$ and $\boldsymbol{Y}$, as we test the null hypothesis $F = G$. In Section 1.2, we discuss the landscape of existing methods.

While we face no shortage of effective two-sample tests, certain factors may hinder their real-world applicability. For one, we find some nonparametric tests to be more parsimonious than others against the range of potential alternatives. Relatively speaking, a method may excel at detecting location and scale shifts but struggle to catch bimodality when mean and variance are held constant, as one example. We explore this phenomenon in Section 5, showing that the relative performance of well-known univariate tests at detecting a location shift can be reversed by a suitable choice of distribution family. This is unintuitive, as one might expect power against location alternatives to be nearly independent of family.

Furthermore, many tests offer non-transparent rejections of the null hypothesis. While data visualizations and summary statistics offer some degree of explanation for a test, such analyses do not quantify the contribution of various data features to the test's rejection. For multivariate tests, this problem is compounded, as distributional alternatives may easily exceed human intuition for higher dimensions.

Here, we formulate a new nonparametric two-sample test called AUGUST, an abbreviation of Augmented CDF for Uniform Statistic Transformation. Our method explicitly tests for multiple orthogonal sources of distributional inequality up to a predetermined resolution d, giving it power against a wide range of alternatives. Upon rejection of the null, both resolution control and decomposition into orthogonal signals allow for interpretation of how equality in distribution between $\boldsymbol{X}$ and $\boldsymbol{Y}$ has failed. To promote ease of use, we provide asymptotic theory as well as algorithmic optimizations.

1.2 Relatives and Further Reading

Some well-known rank-based tests are designed for the univariate context, including [13, 27, 33]. Other approaches explicitly refer to a distance between the empirical cumulative distribution functions of $\boldsymbol{X}$ and $\boldsymbol{Y}$. For instance, [1, 12, 15, 26] are all widely known. The recent test of [16] somewhat combines [1] and [15].

For nonparametric multivariate methods, the range of approaches is quite broad. Tests based on geometric graphs, including [9, 10, 18, 39], have had considerable success [5]. Ball divergence [3, 36] and energy distance [2, 42] are also popular names. Among the family of kernel-based tests are

[*] Corresponding author.

[11, 20, 21, 22, 25, 41, 44], while [4, 24, 35, 40] use generalized ranks. Additional recent work includes [6, 7, 28, 31].

As for interpretable methods, one line of work [25, 44] proposes feature selection in the framework of maximum mean discrepancy [21]. Similar in principle is [34]. The test of [31] inherits the interpretability of the classifier on which the test is based. These methods provide a global type of interpretability, compared to the study of local significant differences [17, 23]. To put AUGUST in context, we propose that our method is more geometrical than feature-selecting tests, but more global than local significant difference methods.

Regarding methodological relatives, the use of subsampling has been considered for the two-sample location and scale problems [32, 37, 38]. In addition, our method builds on the binary expansion framework [45, 46], which has applications to resolution-based nonparametric models of dependency [8]. Using an underlying binary expansion framework, we furnish substantial methodological and algorithmic innovations to yield a practical test in the two-sample context.

2. DERIVATION OF A STATISTIC

2.1 Motivation from the Probability Integral Transformation

We begin by introducing our procedure in the context of univariate data. Given independent samples $\{\boldsymbol{X}_i\}_{i=1}^m$ and $\{\boldsymbol{Y}_i\}_{i=1}^n$, where $\boldsymbol{X}_i \sim G$ and $\boldsymbol{Y}_i \sim F$, we are interested in testing

$$H_0 : F = G \text{ vs. } H_a : F \neq G.$$

For our purposes, we will assume that F and G are absolutely continuous functions. We adopt boldface type in $\{\boldsymbol{X}_i\}_{i=1}^m$ and $\{\boldsymbol{Y}_i\}_{i=1}^n$ to indicate that these are collections of random quantities. In addition, we use blackboard bold $\mathbb{P}(A)$ to refer to the probability of an event A.

To begin, imagine a one-sample setting where we test $H_0 : \boldsymbol{X}_i \sim F$, with F known. It is a well-known result that $\boldsymbol{X}_i \sim F$ if and only if the transformed variables $\{F(\boldsymbol{X}_i) : i \in [m]\}$ follow a Uniform$(0,1)$ distribution. Given this fact, we can test the goodness-of-fit of F by testing of the uniformity of the collection $\{F(\boldsymbol{X}_i) : i \in [m]\}$. Moreover, examining how $\{F(\boldsymbol{X}_i) : i \in [m]\}$ fails to be uniform indicates why F does not fit the distribution of $\boldsymbol{X}$.

Returning to the two-sample setting, the same intuition holds true: we might construct transformed variables that are nearly uniform in $[0,1]$ when $F = G$, and that are not uniform otherwise. When the distributions of the two samples are different, the way that uniformity fails should be informative.

Given the fact that the transformed variables $\{G(\boldsymbol{X}_i) : i \in [m]\}$ follow a uniform distribution, an intuitive choice would be $\{\hat{F}_{\boldsymbol{Y}}(\boldsymbol{X}_i) : i \in [m]\}$, where $\hat{F}_{\boldsymbol{Y}}$ is the empirical cumulative distribution function of $\boldsymbol{Y}$:

$$\hat{F}_{\boldsymbol{Y}}(t) = \frac{1}{n}\sum_{i=1}^{n} I(\boldsymbol{Y}_i \leq t).$$

The binary expansion testing framework introduced in [45] provides a way to test $\{\hat{F}_{\boldsymbol{Y}}(\boldsymbol{X}_i) : i \in [m]\}$ for uniformity up to a given binary depth d, which is equivalent to testing multinomial uniformity over dyadic fractions $\{1/2^d, \ldots, 1\}$. In particular, we define the random vector $\boldsymbol{P}$ of length 2^d such that, for $1 \leq i \leq 2^d$,

$$\boldsymbol{P}_i = \frac{\#\left\{k : \hat{F}_{\boldsymbol{Y}}(\boldsymbol{X}_k) \in \left[\frac{i-1}{2^d}, \frac{i}{2^d}\right)\right\}}{m}.$$

That is, $\boldsymbol{P}$ counts the number of transformed observations falling in dyadic intervals of width $1/2^d$. The associated vector $\boldsymbol{S} = \mathbf{H}_{2^d}\boldsymbol{P}$ is said to contain *symmetry statistics*, where $\mathbf{H}_{2^d}$ is the Hadamard matrix of size 2^d according to Sylvester's construction. As the top row of $\mathbf{H}_{2^d}$ contains only ones, the first coordinate of $\boldsymbol{S}$ is always equal to $\sum_{i=1}^{2^d} \boldsymbol{P}_i = 1$, and we may as well restrict our attention to $\boldsymbol{S}_{-1}$, dropping the first component. As shown in [45], $\boldsymbol{S}_{-1}$ is a sufficient statistic for uniformity in the one sample setting, and the binary expansion test based on $\boldsymbol{S}_{-1}$ achieves the minimax rate in sample size required for power against a wide variety of alternatives.

We can think of $\boldsymbol{S}_{-1}$ in a signal-processing context: the Hadamard transform maps the vector of cell probabilities $\boldsymbol{P}$ in the physical domain to the vector of symmetries $\boldsymbol{S}_{-1}$ in the frequency domain. This transformation is advantageous since, in the one sample setting, the entries of $\boldsymbol{S}_{-1}$ have mean zero and are pairwise uncorrelated under the null. As a result, fluctuations of $\boldsymbol{S}_{-1}$ away from $\mathbf{0}_{2^d-1}$ unambiguously support the alternative, and the coordinates of $\boldsymbol{S}_{-1}$ are interpretable as orthogonal signals of nonuniformity. Moreover, the vector $\boldsymbol{P}$ always satisfies $\sum_{i=1}^{2^d} \boldsymbol{P}_i = 1$, meaning that the mass of $\boldsymbol{P}$ is constrained to a $(2^d - 1)$-dimensional hyperplane in $\mathbb{R}^{2^d}$. In contrast, the vector $\boldsymbol{S}_{-1}$ is non-degenerate and summarizes the same information about non-uniformity with greater efficiency. We elaborate on the interpretability of $\boldsymbol{S}_{-1}$ in Section 2.4.

One possible choice of test statistic is the quantity $S = \|\boldsymbol{S}_{-1}\|_2^2$. A test based on S is essentially a χ^2 test and has decent power at detecting $F \neq G$. However, we can substantially improve the power by modifying our construction of $\boldsymbol{P}$.

2.2 An Augmented Cumulative Distribution Function

For our testing purposes, recall that we are only interested in the uniformity of $\{\hat{F}_{\boldsymbol{Y}}(\boldsymbol{X}_i), i \in [m]\}$ up to binary depth d. The range of $\hat{F}_{\boldsymbol{Y}}(x)$ as a function of x comprises $n + 1$ possible values, namely, $0, 1/n, \ldots, 1$. However, in our construction of cell counts $\boldsymbol{P}$, the collection

$\{\hat{F}_{\boldsymbol{Y}}(\boldsymbol{X}_i), i \in [m]\}$ is binned across 2^d-many dyadic intervals of depth d. Whenever $2^d < m$, some distinct values in the range of $\hat{F}_{\boldsymbol{Y}}(\cdot)$ correspond to the same dyadic interval by the pigeonhole principle, which indicates that a coarser transformation than $\hat{F}_{\boldsymbol{Y}}(\cdot)$ should work at least as well, and possibly better. Our approach is to consider transformed variables $\hat{F}_{\boldsymbol{Y}^*}(\boldsymbol{X}_i)$ based on a small, random subsample $\boldsymbol{Y}^*$ of some size r from $\boldsymbol{Y}$. The following discussion makes this alternate process explicit. In addition, we comment on the success of this approach in Section 4, and we include empirical power comparisons against the non-subsampled transformation $\hat{F}_{\boldsymbol{Y}}(\cdot)$ in the supplementary materials.

Let $\boldsymbol{Y}^*$ be a random subsample from $\boldsymbol{Y}$ of size $r = 2^{d+1} - 1$. We explain this choice of r momentarily. For any $x \in \mathbb{R}$ and integer $1 \leq k \leq 2^d$, let $\boldsymbol{p}_k(x)$ be the probability, conditional on $\boldsymbol{Y}$, that either $2k-2$ or $2k-1$ elements of $\boldsymbol{Y}^*$ are less than or equal to x. The probabilities $\boldsymbol{p}_k(x)$ are essentially hypergeometric and simple to compute:

$$\boldsymbol{p}_k(x) = \frac{\binom{\#\{i:\boldsymbol{Y}_i \leq x\}}{2k-2}\binom{\#\{i:\boldsymbol{Y}_i > x\}}{2^{d+1}-1-(2k-2)}}{\binom{n}{2^{d+1}-1}} + \frac{\binom{\#\{i:\boldsymbol{Y}_i \leq x\}}{2k-1}\binom{\#\{i:\boldsymbol{Y}_i > x\}}{2^{d+1}-1-(2k-1)}}{\binom{n}{2^{d+1}-1}}.$$

Using the scalar function $\boldsymbol{p}_k(\cdot)$, we define the $\boldsymbol{P} : \mathbb{R} \to \mathbb{R}^{2^d}$ such that, for each coordinate k,

$$\boldsymbol{P}_k(x) = \boldsymbol{p}_k(x), \text{ for } 1 \leq k \leq 2^d.$$

It holds that $\hat{F}_{\boldsymbol{Y}^*}(x) \in \left[(k-1)/2^d, k/2^d\right]$ when exactly $2k-2$ or $2k-1$ subsampled elements in $\boldsymbol{Y}^*$ are less than or equal to x. Therefore, we could equally say that for $1 \leq k \leq 2^d$,

$$\boldsymbol{P}_k(x) = \mathbb{P}\left(\hat{F}_{\boldsymbol{Y}^*}(x) \in \left[\frac{k-1}{2^d}, \frac{k}{2^d}\right] \middle| \boldsymbol{Y}\right). \tag{2.1}$$

It is in precisely this sense that $\boldsymbol{P}(x)$ can be considered an *augmented cumulative distribution function*: instead of mapping x to a single value in the unit interval, $x \mapsto \boldsymbol{P}(x)$ maps x to a distribution. Moreover, this characterization explains the choice of subsample size $r = 2^{d+1} - 1$. Any r satisfying $r = 2^q - 1$, $q \geq d$, guarantees that the discrete random variable $\hat{F}_{\boldsymbol{Y}^*}(x)$ has the same number of point masses inside every interval of the form $\left[(k-1)/2^d, k/2^d\right]$. In Section 4, we give further intuition behind the meaning of q, and in the supplementary materials, we asses our default choice of $q = d + 1$ empirically.

To collect information about every $\boldsymbol{X}_i$, we define the vector $\boldsymbol{P}_{\boldsymbol{X}}$ to be the average of all $\boldsymbol{P}(\boldsymbol{X}_i)$:

$$\boldsymbol{P}_{\boldsymbol{X}} = \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{P}(\boldsymbol{X}_i).$$

Given that the formula for $\boldsymbol{p}_k(x)$ is computed from hypergeometric probabilities, we refer the coordinates of $\boldsymbol{P}_{\boldsymbol{X}}$ as *hypergeometric cell probabilities*. Just as we expect the distribution of the transformed variables $\{\hat{F}_{\boldsymbol{Y}}(\boldsymbol{X}_i) : i \in [m]\}$ to be uniform under the null, we expect the mass of $\boldsymbol{P}_{\boldsymbol{X}}$ to be nearly uniform over its coordinates. The vector of symmetry statistics $\boldsymbol{S}_{\boldsymbol{X}} = (\mathbf{H}_{2^d}\boldsymbol{P}_{\boldsymbol{X}})_{-1}$ quantifies non-uniformity in $\boldsymbol{P}_{\boldsymbol{X}}$.

Importantly, the cell probabilities in $\boldsymbol{P}(x)$ are computed with reference to a subsampling procedure, but without actually subsampling. As the discussion above suggests, these probabilities could indeed be approximated by a bootstrap procedure: take many subsamples $\boldsymbol{Y}^*$ of size $2^{d+1}-1$ from $\boldsymbol{Y}$, compute $\hat{F}_{\boldsymbol{Y}^*}(x)$ each time, and bin the results as cell counts at intervals of $1/2^d$. The exact cell probabilities $\boldsymbol{P}(x)$ derived above are the limiting values of this bootstrap procedure as the number of subsamples tends to infinity. The following theorem makes this result explicit.

Theorem 1. *Let Y be a fixed vector of length at least 2^{d+1}, and let $x \in \mathbb{R}$. Consider the following bootstrap method for computing a vector $\boldsymbol{P}^*(x)$ using K subsamples from Y.*

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

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

where the probability is taken over the randomness of the subsampling, and $\boldsymbol{P}(x)$ is the augmented cumulative distribution function based on Y.

Theorem 1 shows that the hypergeometric cell probabilities are equivalent to the limiting values of a certain bootstrap procedure. Effectively, one could say that actual subsampling is a valid way to approximate $\boldsymbol{P}_{\boldsymbol{X}}$. In practice, it is much faster to directly compute the limiting hypergeometric probabilities. While the procedure described in this subsection is more complicated than the original approach from Section 2.1, we achieve superior power using the augmented cumulative distribution function introduced here, which we illustrate empirically in the supplementary materials. In Section 5, we provide comparisons of empirical power against well-known nonparametric tests.

2.3 Distributional Difference as a Scalar Quantity

To combine information on all forms of asymmetry, we propose the statistic $S = -\boldsymbol{S}_{\boldsymbol{X}}^T\boldsymbol{S}_{\boldsymbol{Y}}$, with $\boldsymbol{S}_{\boldsymbol{Y}}$ defined analogously to $\boldsymbol{S}_{\boldsymbol{X}}$ by reversing the roles of the two samples. First, this choice of statistic has the advantage of treating the $\boldsymbol{X}$ and $\boldsymbol{Y}$ samples symmetrically. This is desirable because it would be counterintuitive for the value of S to change when the roles of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are switched. In addition, this statistic is a continuous function of the concatenated vector $(\boldsymbol{S}_{\boldsymbol{X}}^T, \boldsymbol{S}_{\boldsymbol{Y}}^T)^T$, and in Theorem 3, we state the asymptotic distribution of $(\boldsymbol{S}_{\boldsymbol{X}}^T, \boldsymbol{S}_{\boldsymbol{Y}}^T)^T$ in the case of univariate data. For the multivariate AUGUST test, which is described in Section 3.2, we use permutation for p-value calculation.

The negative sign in $-\boldsymbol{S}_{\boldsymbol{X}}^T\boldsymbol{S}_{\boldsymbol{Y}}$ comes from the fact that $\boldsymbol{S}_{\boldsymbol{X}}$ and $\boldsymbol{S}_{\boldsymbol{Y}}$ typically have opposite signs in the case of distributional difference, and we wish the critical values of S to be positive. The proposition below gives intuition for this phenomenon in the context of a location shift.

Proposition 1. *Let $m, n \geq 2^{d+1}$, and suppose $\{\boldsymbol{X}_i\}_{i=1}^m$ and $\{\boldsymbol{Y}_j\}_{j=1}^n$ satisfy $\max_i\{\boldsymbol{X}_i\} < \min_j\{\boldsymbol{Y}_j\}$. Then $\cos(\theta) = -(2^d-1)^{-1}$, where θ is the angle between $\boldsymbol{S}_{\boldsymbol{X}}$ and $\boldsymbol{S}_{\boldsymbol{Y}}$ as vectors in $\mathbb{R}^{2^d-1}$.*

Informally, we could say the following: if $\boldsymbol{X}$ is to the left of $\boldsymbol{Y}$, then $\boldsymbol{Y}$ is to the right of $\boldsymbol{X}$, and the symmetry statistic detecting left/right imbalance will be positive in $\boldsymbol{S}_{\boldsymbol{X}}$ and negative in $\boldsymbol{S}_{\boldsymbol{Y}}$. As shown in Section 5, the negative inner product $-\boldsymbol{S}_{\boldsymbol{X}}^T\boldsymbol{S}_{\boldsymbol{Y}}$ gives good power against a wide range of distributional alternatives. In addition, see Section 4 for exploration of the asymptotic properties of $(\boldsymbol{S}_{\boldsymbol{X}}^T, \boldsymbol{S}_{\boldsymbol{Y}}^T)^T$.

2.4 Interpretation of the Results

One may use entries of $\boldsymbol{S}_{\boldsymbol{X}}$ and $\boldsymbol{S}_{\boldsymbol{Y}}$ to interpret the outcome of the AUGUST test based on S. With respect to $\boldsymbol{S}_{\boldsymbol{X}}$, the sample $\boldsymbol{Y}$ serves as the *reference sample*, meaning that information from $\boldsymbol{S}_{\boldsymbol{X}}$ allows us to make statements about how points of $\boldsymbol{X}$ fall relative to the distribution of $\boldsymbol{Y}$. Each entry in the vector $\boldsymbol{S}_{\boldsymbol{X}}$ describes the non-uniformity of $\boldsymbol{P}_{\boldsymbol{X}}$ with respect to a row of $\mathbf{H}_{2^d}$. In particular, the largest entries of $\boldsymbol{S}_{\boldsymbol{X}}$ in absolute value tell us the sources of greatest asymmetry in $\boldsymbol{P}_{\boldsymbol{X}}$.

Before performing the test based on S, we must first choose some resolution d, which determines the scale on which the test will be sensitive. For convenience, let $\tilde{\mathbf{H}}_{2^d}$ denote the Hadamard matrix of size 2^d according to Sylvester's construction, without the first row, which is a row of all ones. Now, the depth $d = 1$ is sensitive primarily left/right imbalance. When $d = 1$, the $2^1 = 2$ entries of $\boldsymbol{P}_{\boldsymbol{X}}$ roughly correspond to the fraction of the $\boldsymbol{X}$ sample falling above or below the median of $\boldsymbol{Y}$. In this case, the only symmetry statistic is the product of $\tilde{\mathbf{H}}_2 = \begin{pmatrix}1 & -1\end{pmatrix}$ with $\boldsymbol{P}_{\boldsymbol{X}}$, namely the difference between the two components of $\boldsymbol{P}_{\boldsymbol{X}}$.

For $d = 2$, both $\begin{pmatrix}1 & 1 & -1 & -1\end{pmatrix}$ and $\begin{pmatrix}1 & -1 & -1 & 1\end{pmatrix}$ are (necessarily orthogonal) rows of $\tilde{\mathbf{H}}_4$. When multiplied by $\boldsymbol{P}_{\boldsymbol{X}}$, the first of these rows produces a statistic for left/right imbalance, similar to the $d = 1$ case, while the latter row detects differences in scale. Larger values of d detect more granular varieties of imbalance. We use a depth of $d = 2$ in our real data example, and $d = 3$ in simulated power comparisons. (See the supplementary materials for an empirical comparison across depths.) In [46], it is shown that a depth of $d = 3$ is sufficient for a symmetry statistic-based test of independence to outperform both distance correlation and the F-test, which are known to be optimal, in detecting correlation in bivariate normal distributions.

Higher depths $d > 3$ can be useful for alternatives that are extremely close in the Kolmogorov–Smirnov metric but have densities that are bounded apart in the uniform norm. As one example, we may have $\boldsymbol{X}$ sampled from Uniform$(0, 1)$ and $\boldsymbol{Y}$ sampled from a high frequency square wave distribution with the same support. In Section 6, we use symmetry statistics in visualizations of NBA shooting data. As an additional example, a step-by-step interpretation on simulated data is provided in the supplementary materials.

3. COMPUTATIONAL CONSIDERATIONS

3.1 Algorithms for the Univariate Statistic

Below, Algorithms 1 and 2 formalize the steps to calculating the AUGUST statistic outlined in earlier sections. In terms of prior notation, Algorithm 1 computes the vector $\boldsymbol{P}(x)$, and Algorithm 2 calculates the overall test statistic $S = -\boldsymbol{S}_{\boldsymbol{X}}^T\boldsymbol{S}_{\boldsymbol{Y}}$. Recall that we use $\mathbf{H}_{2^d}$ to refer to the Hadamard matrix of size 2^d according to Sylvester's construction, and for a matrix $\mathbf{M}$, we use $(\mathbf{M})_{-1}$ to refer to $\mathbf{M}$ without its first row.

Algorithm 1 Augmented CDF(V, d, x).

```
Initialize zero vector $P$ of length $2^d$
Set $N = \text{length}(V)$, $n = 2^{d+1} - 1$, $K = \#\{i : V_i \leq x\}$, $k = 0$
for $i = 1, \dots, 2^d$ do
    $k \leftarrow 2i - 2$
    $P_i \leftarrow \binom{K}{k}\binom{N-K}{n-k}/\binom{N}{n} + \binom{K}{k+1}\binom{N-K}{n-k-1}/\binom{N}{n}$
end for
Return $P$
```

Algorithm 2 AUGUST(X, Y, d).

```
Initialize zero vectors $P_X$, $P_Y$, $V$ of length $2^d$
for $i = 1, \dots, \text{length}(X)$ do
    $V \leftarrow$ Algorithm 1$(Y, d, X_i)$
    $P_X \leftarrow P_X + V/\text{length}(X)$
end for
for $i = 1, \dots, \text{length}(Y)$ do
    $V \leftarrow$ Algorithm 1$(X, d, Y_i)$
    $P_Y \leftarrow P_Y + V/\text{length}(Y)$
end for
$S_X \leftarrow (\mathbf{H}_{2^d} P_X)_{-1}$
$S_Y \leftarrow (\mathbf{H}_{2^d} P_Y)_{-1}$
Return $S = -S_X^T S_Y$
```

The two samples $\boldsymbol{X}$ and $\boldsymbol{Y}$ have sizes m and n, respectively. Treating d as a constant, Algorithm 2 requires $O(mn)$ elementary operations. This is due to the calculation of $K = \#\{i : \boldsymbol{V}_i \leq x\}$ in the Algorithm 1, which necessitates iterating over all entries of $\boldsymbol{V}$ each time that Algorithm 1$(\boldsymbol{V}, d, x)$ is called.

However, by first sorting the concatenated $\boldsymbol{X}$ and $\boldsymbol{Y}$ samples, it is possible to reduce the running time to $O((m+n)\log(m+n))$ operations.

Theorem 2. *There exists an algorithm for calculating the exact test statistic S that requires $O((m+n)\log(m+n))$ elementary operations.*

This improved algorithm, named AUGUST+, is stated explicitly in the supplementary materials. The time complexity is asymptotically equivalent to that of an efficient sorting algorithm applied to the concatenated data. The constant factor in this comparison depends on the resolution d, which is assumed constant in Theorem 2. In terms of storage, the AUGUST+ algorithm defines only one array whose length depends on m and n. This array has dimension $2 \times (m+n)$, meaning the space requirement is linear in the combined sample size. In Section 6, we use this algorithm to perform our two-sample test on a large data set on the order of 10^6 observations.

3.2 Multivariate Extension

With an appropriate transformation, we can extend the univariate test to the problem of multivariate two-sample testing. For the purposes of this subsection, let $\boldsymbol{X} = \{\boldsymbol{X}_i\}_{i=1}^m$ be an independent sample from multivariate distribution G, and let $\boldsymbol{Y} = \{\boldsymbol{Y}_j\}_{j=1}^n$ be an independent sample from multivariate distribution F, with F and G defined on $\mathbb{R}^k$ and $k \geq 2$. We adapt an approach that could be appropriately named *mutual Mahalanobis distance.*

Given a mean $\mu \in \mathbb{R}^k$ and invertible $k \times k$ covariance matrix Σ, recall that the Mahalanobis distance of $\boldsymbol{x} \in \mathbb{R}^k$ from μ with respect to Σ is

$$MD(x; \mu, \Sigma) = \left[(x-\mu)^T \Sigma^{-1} (x-\mu)\right]^{1/2}.$$

Let $\hat{\mu}_{\boldsymbol{X}}$ and $\hat{\Sigma}_{\boldsymbol{X}}$ be the sample mean and sample covariance matrix of $\boldsymbol{X}$, where we assume $\hat{\Sigma}_{\boldsymbol{X}}$ is nonsingular. Consider the transformed collections

$$\tilde{\boldsymbol{X}}^{(\boldsymbol{X})} = \left\{MD(\boldsymbol{X}_i; \hat{\mu}_{\boldsymbol{X}}, \hat{\Sigma}_{\boldsymbol{X}}) : 1 \leq i \leq m\right\}$$
$$\tilde{\boldsymbol{Y}}^{(\boldsymbol{X})} = \left\{MD(\boldsymbol{Y}_j; \hat{\mu}_{\boldsymbol{X}}, \hat{\Sigma}_{\boldsymbol{X}}) : 1 \leq j \leq n\right\},$$

where the superscript $(\boldsymbol{X})$ indicates that means and covariances are estimated using the $\boldsymbol{X}$ sample. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ come from the same multivariate distribution, then the collections $\tilde{\boldsymbol{X}}^{(\boldsymbol{X})}$ and $\tilde{\boldsymbol{Y}}^{(\boldsymbol{X})}$ should have similar univariate distributions. As a result, at a given depth d, it is reasonable to test the univariate samples $\tilde{\boldsymbol{X}}^{(\boldsymbol{X})}$ and $\tilde{\boldsymbol{Y}}^{(\boldsymbol{X})}$ as an assessment of the distributional equality of the multivariate samples $\boldsymbol{X}$ and $\boldsymbol{Y}$. Our choice of Mahalanobis distance is also motivated by the wide class of nonparametric tests based on data depth, particularly Mahalanobis depth [5, 30].

From the univariate method, recall that vectors of symmetry statistics quantify regions of imbalance between the univariate samples, as imbalances in distribution appear as non-uniformity in the vector of cell probabilities. Under a Mahalanobis distance transformation, cells in the domain of $\tilde{\boldsymbol{X}}^{(\boldsymbol{X})}$ and $\tilde{\boldsymbol{Y}}^{(\boldsymbol{X})}$ correspond to nested elliptical rings centered on $\hat{\mu}_{\boldsymbol{X}}$. This principle extends interpretability of symmetry statistics to the multivariate case.

As in the univariate case, it is desirable for the test statistic to be invariant to the transposition of $\boldsymbol{X}$ and $\boldsymbol{Y}$. To achieve this, we can use the statistic

$$S_{multi} = \max\left(\text{AUGUST}\left(\tilde{\boldsymbol{X}}^{(\boldsymbol{X})}, \tilde{\boldsymbol{Y}}^{(\boldsymbol{X})}, d\right),\right.$$
$$\left.\text{AUGUST}\left(\tilde{\boldsymbol{X}}^{(\boldsymbol{Y})}, \tilde{\boldsymbol{Y}}^{(\boldsymbol{Y})}, d\right)\right)$$

wherein we use both possible Mahalanobis distance transformations for $\boldsymbol{X}$ and $\boldsymbol{Y}$, compute two test statistics, and take the maximum. Aside from ensuring transposition invariance, the simultaneous use of the two test statistics is important for detecting some asymmetric alternatives. As a contrived example, take $k = 2$; suppose $\boldsymbol{X}$ comprises m independent samples of the bivariate standard normal $(Z_1, Z_2)^T \sim N_2(0, I_2)$, while $\boldsymbol{Y}$ comprises n independent samples of $2^{-1/2}(\chi_2, \chi_2)^T$, where χ_2 follows a chi distribution with two degrees of freedom. In this case, both $MD((Z_1, Z_2)^T; 0, I_2)$ and $MD(2^{-1/2}(\chi_2, \chi_2)^T; 0, I_2)$ follow a chi distribution with two degrees of freedom, which indicates that the test based solely on $\text{AUGUST}\left(\tilde{\boldsymbol{X}}^{(\boldsymbol{X})}, \tilde{\boldsymbol{Y}}^{(\boldsymbol{X})}, d\right)$ is powerless. (In this highly degenerate situation, the transposed statistic $\text{AUGUST}\left(\tilde{\boldsymbol{X}}^{(\boldsymbol{Y})}, \tilde{\boldsymbol{Y}}^{(\boldsymbol{Y})}, d\right)$ is technically undefined, because $\boldsymbol{Y}$ has no variance in the $(1, -1)^T$ direction.)

In practice, we calculate p-values for the multivariate statistic using permutation. While this current necessity sacrifices the computational advantage of the asymptotic result Theorem 3, the multivariate method may nonetheless take advantage of Theorem 2, as it is built upon the univariate procedure. As we show in Section 5, a depth of $d = 2$ is sufficiently large to detect common multivariate alternatives with empirical power comparable to existing tests.

4. DISTRIBUTIONAL INSIGHTS

To simplify p-value calculation and simulation analysis, we provide theoretical results regarding the univariate procedure outlined in Section 3.1. These asymptotic insights follow from the adaptation of classical U-statistic theory. For each $k \in \{1, \ldots, 2^d\}$, define the function $p_k^F : \mathbb{R} \to [0, 1]$ by

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

with $p_k^G : \mathbb{R} \to [0, 1]$ defined analogously. These functions can be thought of as theoretical analogs of the data-dependent probabilities $\boldsymbol{p}_k(x)$ from Section 2.2. Further, define the integrated quantities

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

For reasons that will soon be apparent, we refer to $p_k^{F:G}$ and $p_k^{G:F}$ as the *limiting cell probabilities* of AUGUST.

Theorem 3. *Suppose that* $\{\boldsymbol{X}_i\}_{i=1}^m$ *and* $\{\boldsymbol{Y}_j\}_{j=1}^n$ *are independent univariate observations, where* $\boldsymbol{X}_i \sim G$ *and* $\boldsymbol{Y}_j \sim F$ *for continuous distributions* G, 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*

$$N^{1/2}\left(\begin{pmatrix}\boldsymbol{S_X}\\ \boldsymbol{S_Y}\end{pmatrix} - \mu\right) \to N(0_{2^{d+1}-2}, \Sigma)$$

in distribution, where

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

for $p^{F,G} = \left(p_1^{F:G}, \dots, p_{2^d}^{F:G}, p_1^{G:F}, \dots, p_{2^d}^{G:F}\right)^T$, *and* Σ *is a matrix depending on* λ, d, F, *and* G.

Because the exact form of Σ is notation-intensive, we state it in the supplementary materials. In light of the above result, given distributions F, G, it is possible to compute the limit in probability of the symmetry statistics $(\boldsymbol{S_X}, \boldsymbol{S_Y})^T$. This limit μ encodes asymmetry at the population level, analogous to how $(\boldsymbol{S_X}, \boldsymbol{S_Y})^T$ encodes asymmetry between the finite samples $\boldsymbol{X}$ and $\boldsymbol{Y}$. Moreover, using this theorem, one can efficiently simulate the test statistic S under the alternative, yielding a benchmark against a predetermined $F \neq G$ in large samples. In applications that require an *a priori* power analysis, this approach could simplify the process of determining the sample size necessary for detecting a given effect.

Building on these ideas, the limit μ indicates a small framework for understanding how symmetry statistics encode distributional differences. This principle helps explain AUGUST's power against alternatives at each depth d.

For convenience with inverse functions, we assume the CDFs F and G are differentiable and strictly increasing on $\mathbb{R}$, though similar reasoning applies under weaker conditions. Let E denote the CDF of the uniform distribution on $[0, 1]$. By a change of variables,

$$p_K^{F:G} = \int_{-\infty}^{\infty} p_k^F(x)g(x)dx = \int_0^1 p_k^E(u)\frac{g(F^{-1}(u))}{f(F^{-1}(u))}du,$$

wherein all information about F and G is contained in the likelihood ratio on the right. Moreover, for any other differentiable and strictly increasing CDF H, we have

$$\frac{g \circ F^{-1}}{f \circ F^{-1}} = (G \circ F^{-1})' = \left((G \circ H^{-1}) \circ (F \circ H^{-1})^{-1}\right)',$$

which shows that the limit μ is invariant to such transformations of F and G. That is, (F, G) and $(F \circ H^{-1}, G \circ H^{-1})$ are in an equivalence class of distribution pairs whose symmetry statistics have the same limit. Defining $Q = G \circ F^{-1}$, we arrive at

$$p_K^{F:G} = \int_0^1 p_k^E(u)q(u)du.$$

Asymmetry for the equivalence class of (F, G) is characterized by the deviation of Q from the uniform distribution E. Similar transformation invariance properties (perhaps phrased differently) are common among distribution-free tests; for AUGUST, we can be quite specific about the implications for Q. From the beginning of this section, recall that the population-level probability $p_k^F(x)$ is the sum of two binomial-like terms, and compare with the following result.

Theorem 4. *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* $k = 0, \dots, 2^t - 1$.
2. $\int u^k dQ(u) = \mathbf{E}[U^k]$ *for all integers* $k = 0, \dots, 2^t - 1$, *where* $U \sim$ *Unif*$[0, 1]$.

The importance of this theorem is as follows. In Section 2.2, we define the data-dependent function $\boldsymbol{p}_k(x)$ to be the probability that $2k - 2$ or $2k - 1$ elements of the subsampled $\boldsymbol{Y}^*$ are less than or equal to x, which reflects the choice of resample size $r = 2^{d+1} - 1$. In the language of equation (2.1), both $\boldsymbol{p}_k(x)$ and $p_k^F(x)$ are a sum of two probability terms because the interval $[(k-1)/2^d, k/2^d]$ contains exactly two point masses of $\hat{F}_{\boldsymbol{Y}^*}(x)$. If we instead select $r = 2^d - 1$, then each dyadic interval at depth d contains only one point mass of $\hat{F}_{\boldsymbol{Y}}(x)$, and the limiting cell probabilities are instead

$$\begin{aligned} p_k^{F:G} &= \int \binom{2^d-1}{k} F(x)^k(1-F(x))^{2^d-1-k}dG(x) \\ &= \int \binom{2^d-1}{k} u^k(1-u)^{2^d-1-k}dQ(u). \end{aligned}$$

Symmetry statistics are nonzero precisely when the underlying cell probabilities are imbalanced. By Theorem 4, when $r = 2^d - 1$, these limiting probabilities are balanced exactly when the first $2^d - 1$ raw moments of Q match the corresponding raw moments of the uniform distribution.

From this perspective, fixing d while increasing r involves higher moments of Q while nonetheless performing inference at a binary expansion depth of d. This would suggest that with d fixed, increasing r gives superior performance on more peculiar alternatives while overcomplicating simple cases, like location shift. We observe exactly this phenomenon in empirical studies, which we include in the supplementary materials. Our standard choice of $r = 2^{d+1} - 1$ represents a compromise between these two competing forces.

We conclude by remarking that as a heuristic, this discussion is relevant to interpreting the multivariate version of AUGUST, whose symmetry statistics measure imbalance in the transformed collections $\tilde{\boldsymbol{X}}^{(\boldsymbol{X})}$ and $\tilde{\boldsymbol{Y}}^{(\boldsymbol{X})}$. Alternative transformations to Mahalanobis distance may yield more suitable information for some applications, though building off of the univariate test yields convenient intuition for the parameters d and r.

5. EMPIRICAL PERFORMANCE

5.1 Univariate Performance

Here, we compare AUGUST to a sampling of other nonparametric two-sample tests: Kolmogorov–Smirnov distance [26], Wasserstein distance [15], energy distance [42], and the recent DTS [16]. For these simulations, we use a sample size of $n = m = 128$, and for our resolution-based test, we set a depth of $d = 3$. For all tests, we use a p-value cutoff of $\alpha = 0.05$. Simulation results are graphed in Fig. 1.

The first two plots of Fig. 1 correspond to normal and Laplace location alternatives, situations where differences in the first distributional moment are most diagnostic. Third, we have a symmetric beta versus asymmetric beta alternative, and fourth, we include a Laplace scale family. The last two plots of Fig. 1 focus on families with identical first and second moments: normal versus mean-centered gamma in the fifth position, and standard normal versus symmetric, variance-scaled normal mixture in the sixth. For this final alternative distribution, samples are generated by first drawing from a symmetric mixture of normals with unit variance and then dividing by the theoretical standard deviation of the mixture distribution.

For the location alternatives, the power of each method depends on the shape of the distribution. DTS, Wasserstein, and energy distance tests perform slightly better than ours for normal and beta distributions, and ours in turn outperforms Kolmogorov–Smirnov. In contrast, for a Laplace location shift, Kolmogorov–Smirnov outperforms every test, with our test in second place and DTS last. For the Laplace scale family, Kolmogorov–Smirnov performs poorly, with DTS and our test leading. DTS has the edge on the gamma skewness family, while we outperform all other tests at detecting normal versus symmetric normal mixture.

As expected, no single test performs best in all situations. Even for simple alternatives such as location families, the precise shape of the distribution is highly influential as to the tests' relative performance. In fact, the performance rankings of DTS, Wasserstein, energy distance, and Kolmogorov–Smirnov in the Laplace location trials are exactly reversed compared to the normal location trials. We theorize that because the symmetry statistics $\boldsymbol{S_X}$ and $\boldsymbol{S_Y}$

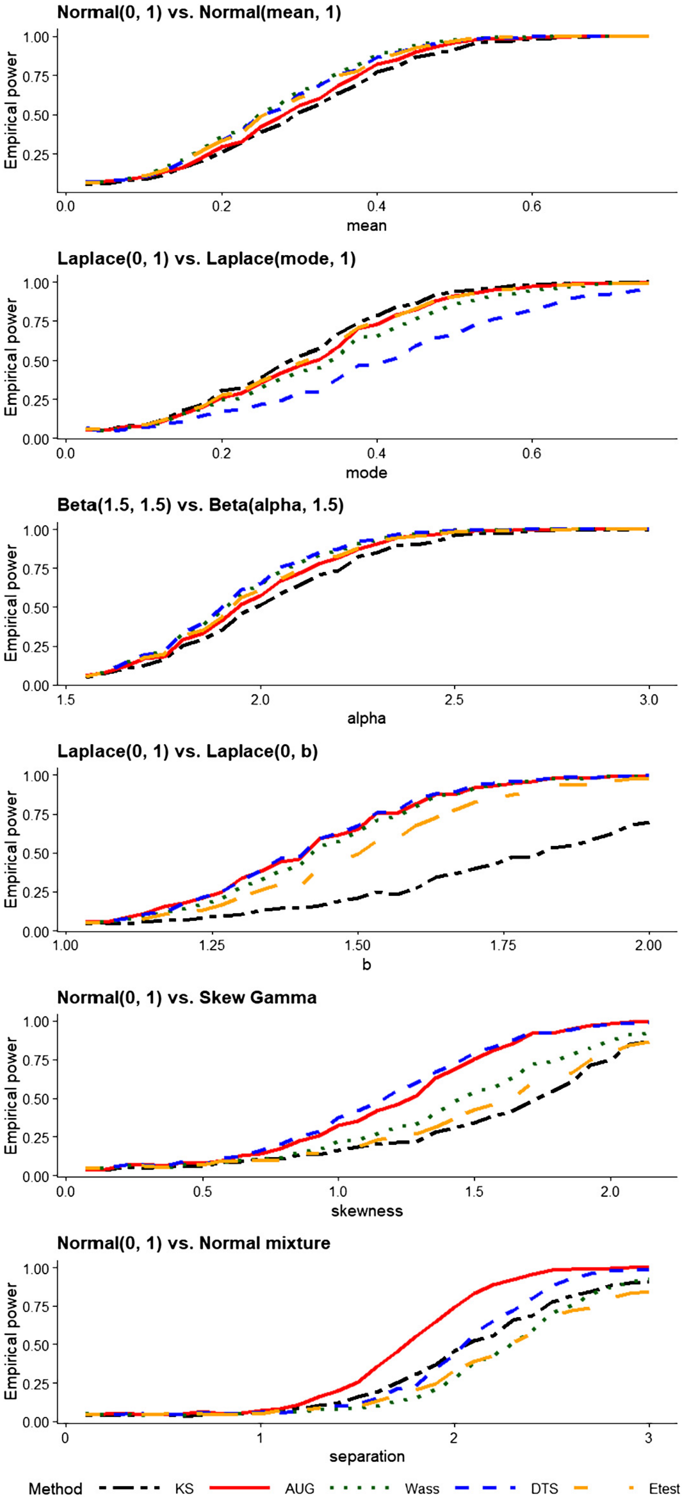

Figure 1: Univariate comparison of power between AUGUST in red, Kolmogorov–Smirnov distance in black, Wasserstein distance in green, DTS in blue, and energy distance in yellow. Our method performs comparably to existing approaches, with superior power in some circumstances.

are weighted equally in every coordinate, AUGUST is very parsimonious towards the range of potential alternatives. In contrast, other univariate methods show relatively greater sensitivity to location and scale shifts, but may be less robust against more obscure alternatives.

5.2 Multivariate Performance

In Fig. 2, we compare our multivariate resolution-based test at depth $d = 2$ to some other well-known nonparametric multivariate two-sample tests. We perform these simulations in a low-dimensional context with $k = 2$, using sample size $n = m = 128$ and cutoff $\alpha = 0.05$. In particular, we again consider the energy distance test of [42], as well as the classifier test of [31], the generalized edge-count method of [9], and the ball divergence test of [36], where the choice of these comparisons is inspired by [41]. For [9], we use a 5-minimum spanning tree based on Euclidean interpoint distance.

We consider a variety of alternatives. In order:

1. $N_2(0, I_{2\times2})$ vs. $N_2(\text{center} \times 1_2, I_{2\times2})$
2. $N_2(0, I_{2\times2})$ vs. $N_2(0, \text{scale} \times I_{2\times2})$
3. $N_2\left(0, \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\right)$ vs. $N_2\left(0, \begin{pmatrix}1 & \text{cov}\\ \text{cov} & 1\end{pmatrix}\right)$
4. $N_2\left(0, \begin{pmatrix}1 & 0\\ 0 & 9\end{pmatrix}\right)$ vs. $R_\theta N_2\left(0, \begin{pmatrix}1 & 0\\ 0 & 9\end{pmatrix}\right)$, where R_θ is the 2×2 rotation matrix through an angle θ
5. $\exp\left(N_2(0, I_{2\times2})\right)$ vs. $\exp\left(N_2(\mu \times 1_2, I_{2\times2})\right)$
6. $N_2(0, I_{2\times2})$ vs. (Z, B), where $Z \sim N(0, 1)$ and B independently follows the bimodal mixture distribution from Section 5.1.

In Fig. 2, we see that the energy and ball divergence tests dominate the other methods when mean shift is a factor (i.e. in the normal location and log-normal families). On a scale alternative, AUGUST has the best power, with ball divergence at a close second. In contrast, for correlation, rotation, and multimodal alternatives, the edge-count test has superior power, with ball divergence and energy distance coming at or near last place.

Overall, our test is robust against a wide range of possible alternatives, and it has particularly high performance against a scale alternative, where it outperforms all other methods considered. We theorize that, in part, this is because some of the other methods rely heavily on interpoint distances. The scale alternative does not result in good separation between $\boldsymbol{X}$ and $\boldsymbol{Y}$, meaning that interpoint distances are not as diagnostic as they would be in, say, a location shift.

In the supplementary materials, we include additional comparisons with $k = 5$ and $k = 20$, keeping the sample size fixed with $m = n = 128$. Performance follows the same general pattern as when $k = 2$: we lag against location alternatives but are very strong at scale, with no universal winner across all scenarios.

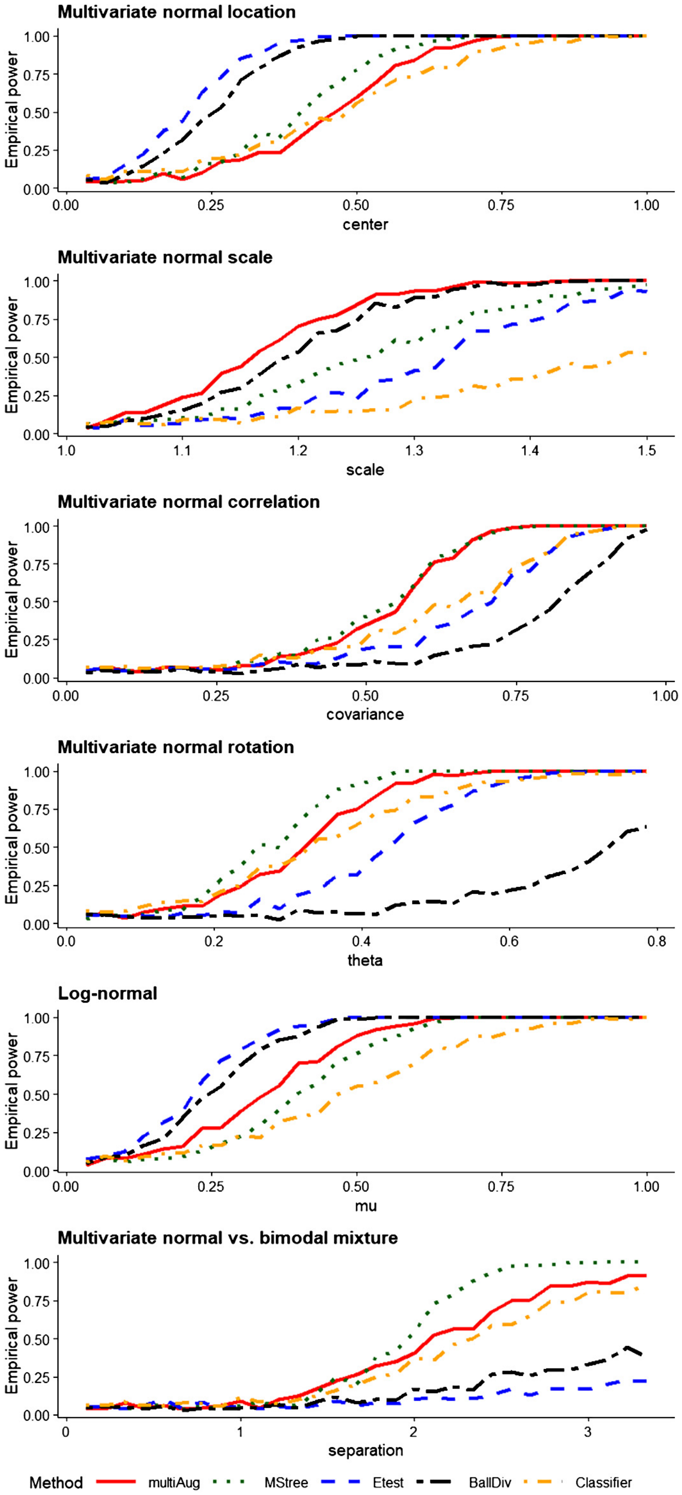

Figure 2: Multivariate comparison of power between AUGUST in red, [9] in green, energy distance in blue, [36] in black, and [31] in yellow for two dimensions. Our method has comparable power with existing methods, and it outperforms all others against a scale alternative.

6. A STUDY OF NBA SHOOTING DATA

We demonstrate the interpretability of AUGUST using 2015–2016 NBA play-by-play data.[1] Consider the distributions of throw distances and angles from the net. Are these distributions different for shots and misses? How about for the first two quarters versus the last two quarters? To address these questions, we acquired play-by-play data for the 2015-2016 NBA season. For each throw, the location of the throw was recorded as a pair of x, y coordinates. These coordinates were converted into a distance and angle from the target net, using knowledge of NBA court dimensions. This data processing yielded a data set on the order of 10^6 observations.

Data were split according to shots versus misses and early game versus late game. Four separate AUGUST tests at a depth of $d = 2$ were performed to analyze the distribution of throw distances and angles. For shot vs. miss distance, shot vs. miss angle, and early vs. late game distance, AUGUST reports $p < 0.001$, while for early vs. late game angle, AUGUST returns $p = 0.004$. For comparison, Kolmogorov-Smirnov yields the same result for the first three scenarios, giving $p = 0.086$ for the fourth. DTS produces $p = 0.033$ for the fourth.

To demonstrate interpretability, we provide visualizations in Fig. 3 as alluded to in Section 2.4. Each histogram corresponds to one of the two samples in the test: this reference sample is indicated on the x-axis. The shaded rectangles overlaid on these histograms illustrate the largest symmetry statistic from the corresponding AUGUST test. For example, the top plot corresponds to throw distance for shots versus misses. The histogram records the distribution of missed throw distances.

Each plot in Fig. 3 yields a specific interpretation as to the greatest distributional imbalance. From the top plot, we see that successful throws tend to be closer to the net than misses. Next, successful throws come from the side more often than misses. Following that, throws early in the game are more frequently from an intermediate distance than late game throws. Finally, throws early in the game come more frequently from the side than they do in the late game. The second of these four is perhaps most counterintuitive, as conventional wisdom suggests that throws from in front of the net are more accurate than throws from the sides. This apparent paradox comes from the fact that throws from the sides are typically at a much closer range.

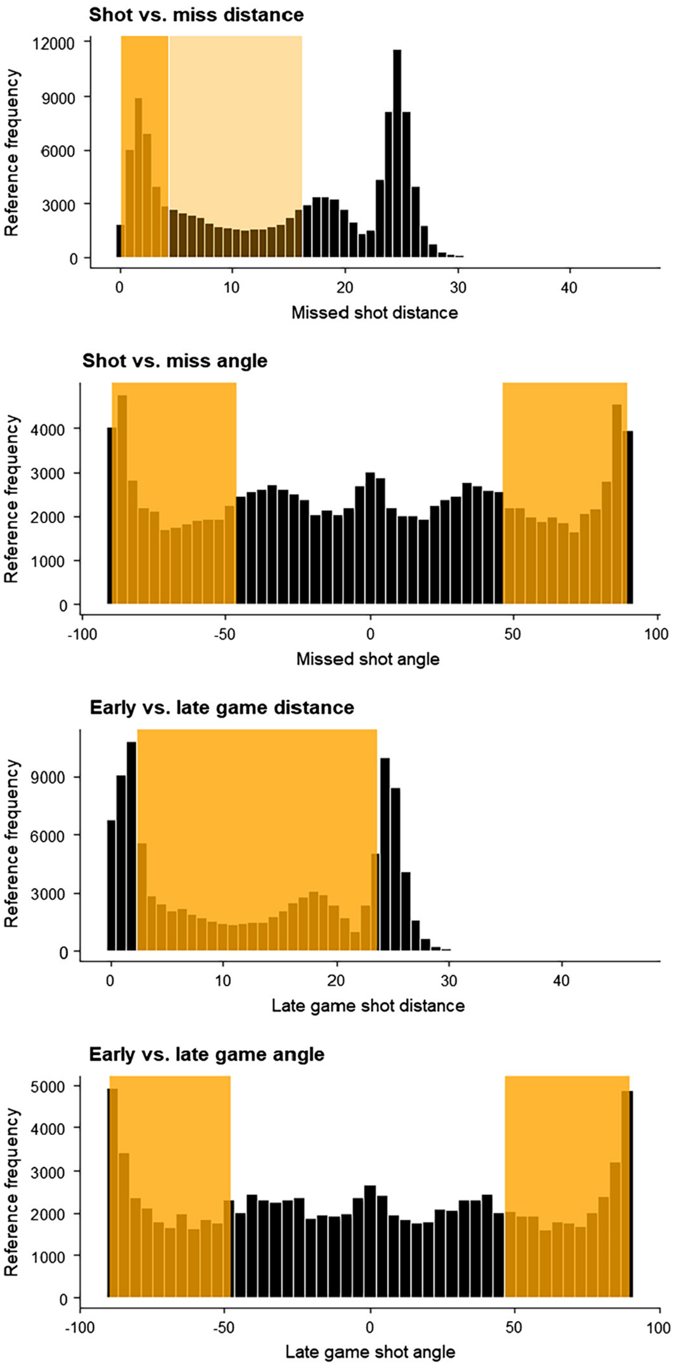

Figure 3: Greatest asymmetries in NBA data. Successful shots are closer to the net than missed shots and come from more extreme angles. Shots in the early game come from a more intermediate distance than in the late game, as well as from more extreme angles.

7. DISCUSSION

An important future direction involves refining the multivariate approach. The simulations of Section 2 speak solely to low-dimensional contexts. We emphasize that other multivariate tests such as [9] enjoy remarkable power properties in growing dimensions. As such, accurate estimation of covariance matrices remains a hindrance to the mutual Mahalanobis distance approach as the dimension k nears a significant fraction of n. In a multivariate context, the test of Section 3.2 serves as a useful starting point for future multi-resolution methods [29], and future work will fo-

[1] Source: nbastuffer.com

cus on extending asymptotic theory in light of the Mahalanobis distance transformation, or other transformations. Permutation, especially in the multivariate context, is feasible but still costly. Repeated evaluations of the hypergeometric probability mass function drive up the constant factor on the $O((n+m)\log(n+m))$ running time, compared to simpler methods of the same order, such as Kolmogorov-Smirnov. Computational burdens could be eased by performing inference across a carefully-selected range of binary depths. For example, as a multivariate test of dependence, the coarse-to-fine sequential adaptive method of [19] chooses a subset of available univariate tests at each resolution using spatial knowledge of dependency structures.

The interpretability of our two-sample test also sheds light on transformations of data from one distribution to the other. This problem is a fundamental subject in transportation theory [43]. We plan to study this problem with recent developments in multi-resolution nonparametric modeling [8] to provide insights on the optimal transportation.

SUPPLEMENTARY MATERIAL

Supplementary material for AUGUST.

ACKNOWLEDGEMENTS

The authors thank the editor, associate editor, and reviewers for their helpful feedback. The authors additionally thank Shankar Bhamidi, Hao Chen, Jan Hannig, Michael Kosorok, Xiao-Li Meng, and Richard Smith for valuable comments and suggestions.

FUNDING

This research is partially supported by NSF grants DMS-1613112, IIS-1633212, DMS-1916237, and DMS-2152289.

Accepted 4 September 2023

REFERENCES

[1] Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics* 193–212. https://doi.org/10.1214/aoms/1177729437. MR0050238

[2] Aslan, B. and Zech, G. (2005). New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation* **75**(2) 109–119. https://doi.org/10.1080/00949650410001661440. MR2117010

[3] Banerjee, B. and Ghosh, A. K. (2022). On high dimensional behaviour of some two-sample tests based on ball divergence. *arXiv preprint arXiv:2212.08566*.

[4] Baumgartner, W., Weiss, P. and Schindler, H. (1998). A nonparametric test for the general two-sample problem. *Biometrics* 1129–1135.

[5] Bhattacharya, B. B. (2019). A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **81**(3) 575–602. MR3961499

[6] Biswas, M. and Ghosh, A. K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis* **123** 160–171. https://doi.org/10.1016/j.jmva.2013.09.004. MR3130427

[7] Biswas, M., Mukhopadhyay, M. and Ghosh, A. K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika* **101**(4) 913–926. https://doi.org/10.1093/biomet/asu045. MR3286925

[8] Brown, B., Zhang, K. and Meng, X. -L. (2022). BELIEF in dependence: leveraging atomic linearity in data bits for rethinking generalized linear models. *arXiv preprint arXiv:2210.10852*.

[9] 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. https://doi.org/10.1080/01621459.2016.1147356. MR3646580

[10] Chen, H., Chen, X. and Su, Y. (2018). A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association* **113**(523) 1146–1155. https://doi.org/10.1080/01621459.2017.1307757. MR3862346

[11] Chwialkowski, K. P., Ramdas, A., Sejdinovic, D. and Gretton, A. (2015). Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems* **28** 1981–1989.

[12] Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal* **1928**(1) 13–74.

[13] Cucconi, O. (1968). Un nuovo test non parametrico per il confronto fra due gruppi di valori campionari. *Giornale degli Economisti e Annali di Economia* 225–248.

[14] DeCost, B. L. and Holm, E. A. (2017). Characterizing powder materials using keypoint-based computer vision methods. *Computational Materials Science* **126** 438–445.

[15] Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications* **15**(3) 458–486. MR0298716

[16] Dowd, C. (2020). A new ECDF two-sample test statistic. *arXiv preprint arXiv:2007.01360*.

[17] Duong, T. (2013). Local significant differences from nonparametric two-sample tests. *Journal of Nonparametric Statistics* **25**(3) 635–645. https://doi.org/10.1080/10485252.2013.810217. MR3174288

[18] Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests. *The Annals of Statistics* 697–717. MR0532236

[19] Gorsky, S. and Ma, L. (2022). Multi-scale Fisher's independence test for multivariate dependence. *Biometrika* **109**(3) 569–587. https://doi.org/10.1093/biomet/asac013. MR4472834

[20] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B. and Smola, A. J. (2007). A kernel statistical test of independence. In: *Advances in Neural Information Processing Systems* 585–592.

[21] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research* **13**(1) 723–773. MR2913716

[22] Harchaoui, Z., Bach, F. R. and Moulines, E. (2007). Testing for homogeneity with kernel Fisher discriminant analysis. In *NIPS* 609–616. Citeseer.

[23] Hazelton, M. L. and Davies, T. M. (2022). Pointwise comparison of two multivariate density functions. *Scandinavian Journal of Statistics* **49**(4) 1791–1810. MR4544820

[24] Hettmansperger, T. P., Möttönen, J. and Oja, H. (1998). Affine invariant multivariate rank tests for several samples. *Statistica Sinica* 785–800. MR1651508

[25] Jitkrittum, W., Szabó, Z., Chwialkowski, K. P. and Gretton, A. (2016). Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems* **29**.

[26] Kolmogorov, A. (1933). Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.* **4** 83–91.

[27] Lepage, Y. (1971). A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrika* **58**(1) 213–217. https://doi.org/10.1093/biomet/58.1.213. MR0408101

[28] Li, J. (2018). Asymptotic normality of interpoint distances for high-dimensional data with applications to the two-sample problem. *Biometrika* **105**(3) 529–546. https://doi.org/10.1093/biomet/asy020. MR3842883

[29] Li, X. and Meng, X. -L. (2021). A multi-resolution theory for approximating infinite-p-zero-n: Transitional inference, individualized predictions, and a world without bias-variance tradeoff. *Journal of the American Statistical Association* **116**(533) 353–367. https://doi.org/10.1080/01621459.2020.1844210. MR4227699

[30] Liu, R. Y. (1992). Data depth and multivariate rank tests. *L1-Statistical Analysis and Related Methods* 279–294. MR1214839

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

[32] Mahajan, K. K., Gaur, A. and Arora, S. (2011). A nonparametric test for a two-sample scale problem based on subsample medians. *Statistics & Probability Letters* **81**(8) 983–988. https://doi.org/10.1016/j.spl.2011.01.018. MR2803733

[33] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 50–60. https://doi.org/10.1214/aoms/1177730491. MR0022058

[34] Mueller, J. W. and Jaakkola, T. (2015). Principal differences analysis: Interpretable characterization of differences between distributions. In: *Advances in Neural Information Processing Systems* **28**.

[35] Oja, H. (2010) *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. Springer Science & Business Media. https://doi.org/10.1007/978-1-4419-0468-3. MR2598854

[36] Pan, W., Tian, Y., Wang, X. and Zhang, H. (2018). Ball divergence: nonparametric two sample test. *Annals of Statistics* **46**(3) 1109. https://doi.org/10.1214/17-AOS1579. MR3797998

[37] Pandit, P. V., Kumari, S. and Javali, S. (2014). Tests for two-sample location problem based on subsample quantiles. *Open Journal of Statistics* **2014**.

[38] Robert Stephenson, W. and Ghosh, M. (1985). Two sample nonparametric tests based on subsamples. *Communications in Statistics-Theory and Methods* **14**(7) 1669–1684. https://doi.org/10.1080/03610928508829003. MR0801632

[39] Rosenbaum, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(4) 515–530. https://doi.org/10.1111/j.1467-9868.2005.00513.x. MR2168202

[40] Rousson, V. (2002). On distribution-free tests for the multivariate two-sample location-scale model. *Journal of Multivariate Analysis* **80**(1) 43–57. https://doi.org/10.1006/jmva.2000.1981. MR1889832

[41] Song, H. and Chen, H. (2020). Generalized kernel two-sample tests. *arXiv preprint arXiv:2011.06127*.

[42] Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference* **143**(8) 1249–1272. https://doi.org/10.1016/j.jspi.2013.03.018. MR3055745

[43] Villani, C. (2009) *Optimal Transport: Old and New* **338**. Springer. https://doi.org/10.1007/978-3-540-71050-9. MR2459454

[44] Yamada, M., Wu, D., Tsai, Y. q. H. H., Takeuchi, I., Salakhutdinov, R. and Fukumizu, K. (2018). Post selection inference with incomplete maximum mean discrepancy estimator. *arXiv preprint arXiv:1802.06226*.

[45] Zhang, K. (2019). BET on Independence. *Journal of the American Statistical Association* **114**(528) 1620–1637. https://doi.org/10.1080/01621459.2018.1537921.

[46] Zhang, K., Zhao, Z. and Zhou, W. (2021). BEAUTY powered BEAST. *arXiv preprint arXiv:2103.00674*.

Benjamin Brown. Chapel Hill, North Carolina, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, USA. E-mail address: brownb1@live.unc.edu

Kai Zhang. Chapel Hill, North Carolina, Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, USA. E-mail address: zhangk@email.unc.edu