stats

MDPI

Article

A New Depth-Based Test for Multivariate Two-Sample Problems

My Luu [1], Yuejiao Fu [1], Augustine Wong [1,*] and Xiaoping Shi [2]

[1] Department of Mathematics and Statistics, York University, Toronto, ON M3J 1P3, Canada; luumy@yorku.ca (M.L.); yuejiao@yorku.ca (Y.F.)
[2] Department of Computer Science, Mathematics, Physics and Statistics, University of British Columbia, Kelowna, BC V1V 1V7, Canada; xiaoping.shi@ubc.ca
* Correspondence: august@yorku.ca

Abstract

Statistical depth provides a center–outward ordering of multivariate observations and is widely used in nonparametric inference. We study depth-based tests for multivariate two-sample problems and examine the behaviour of different depth notions using the DD plot (data-depth plot) across a variety of distributional space. The DD plot illustrates that depth functions differ in their sensitivity to distributional differences, emphasizing the importance of depth selection in two-sample testing. We propose a new two-sample test statistic, log DDR, constructed from ratios of numerical depth values rather than depth-induced ranks. Simulation studies under multiple scenarios and for three representative depth functions indicate that log DDR achieves improved power relative to several competing depth-based nonparametric tests. The results further demonstrate that the performance of log DDR and existing methods depends strongly on the chosen depth function, consistent with insights from the DD plot. These findings support a two-stage testing approach in which the DD plot is used to guide the choice of depth notion before applying log DDR for homogeneity testing.

Keywords: multivariate hypothesis testing; two-sample test; data depth

1. Introduction

Multivariate data analysis plays a central role in modern scientific research, where observations are often high-dimensional and structurally complex. Two fundamental problems frequently arise in this setting: assessing whether a multivariate dataset follows a specified distribution and conducting two-sample tests to identify whether two multivariate samples originate from the same underlying distribution.

Classical parametric inference procedures depend on the assumptions about the underlying probability distributions. When such assumptions are misspecified or fail to hold in practice, the resulting parameter estimates and associated inferential statements may suffer from bias and loss of efficiency. This has motivated the development of nonparametric methods. The nonparametric methods were well-studied for univariate cases [1–3]. However, extending these procedures to multivariate settings remains challenging, mainly because data in higher dimensions lack a natural ordering.

Whereas univariate ranks order observations along a line, statistical depth extends this idea to multivariate data by measuring how deeply a point lies within a data cloud. By providing a meaningful notion of centrality relative to a given distribution or sample, data depth offers a powerful framework for handling multivariate problems. Observations

Check for updates

Academic Editor: Wei Zhu

Received: 2 February 2026
Revised: 28 March 2026
Accepted: 1 April 2026
Published: 3 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

with larger depth values are viewed as more central, while those with smaller depth values are considered more outlying, thereby enabling a natural center–outward ordering in high-dimensional spaces. Importantly, depth-based methods do not require distributional assumptions and retain their effectiveness across different dimensions, making them well-suited for nonparametric inference tasks such as goodness-of-fit assessment, outlier detection, and multivariate two-sample homogeneity testing [4].

Niyogi and Dhar [5] introduced Data-Depth Discrepancy (DDD), founded on Tukey's half-space depth, which can be efficiently computed using software such as R. They proposed unified, distribution-free, and computationally efficient tests for both goodness-of-fit and two-sample problems, applicable to data of any dimension. These tests were shown to be consistent with asymptotic properties established under contiguous alternatives. Their work is the first work to employ the numerical depth values in the test statistic instead of the rank-based approach. However, they focused exclusively on half-space depth and did not evaluate the performance of the proposed DDD statistic when using alternative depth functions such as Mahalanobis, projection, or spatial depth. In this paper, one of our aims is to explore whether the DDD framework can be extended to incorporate these alternative depth functions and assess the performance of the DDD-based tests under various scenarios involving different depth measures. In addition, we propose a bootstrap-based approach to enhance the performance of the DDD statistic. Specifically, we compare the efficiency of bootstrap methods with permutation methods.

Unlike the DDD statistic, which is based on depth differences, we propose two alternative test statistics that rely on relative depth comparisons: the data depth ratio (DDR) and its logarithmic transformation (log DDR). The testing procedure incorporates both nonparametric bootstrap and permutation methods to assess robustness and reliability. We compare the performance of the DDD and DDR-based statistics under various alternative hypotheses and evaluate their effectiveness using three distinct depth functions: Mahalanobis, projection, and spatial depth. Simulation results indicate that the log DDR statistic consistently outperforms the others, demonstrating greater stability and power. The choice of depth function has a substantial impact on test performance. While the Mahalanobis depth consistently delivers superior results across most scenarios, its effectiveness diminishes when comparing mixture normal distributions, where projection depth functions perform better. When both distributions are mixtures, the projection depth demonstrates exceptional performance, achieving near-perfect power for both the DDD and log DDR statistics. This result highlights its effectiveness and reliability in detecting differences between complex or multimodal distributions. In contrast, the DDD statistic performs poorly with spatial depth as it loses power when dealing with non-normal distributions. However, the log DDR statistic maintains both stability and power even when spatial depth is used.

The paper is structured as follows: Section 2 reviews several well-known notions of data depth, including half-space depth, Mahalanobis depth, projection depth, and spatial depth, and uses the DD plot (data-depth plot) [6] to visualize distributional differences and demonstrate the impact of the depth notions in hypothesis testing. Subsequently, we present our proposed two-sample testing methodology, formally defining the DDR and constructing two DDR-based multivariate test statistics. Section 3 presents a comprehensive simulation study that compares our proposed method with existing approaches and confirms that the results are consistent with the patterns observed in the DD plots. Section 4 illustrates the methodology using real data examples, highlighting its practical applicability. Finally, Section 5 discusses the results, provides final remarks and outlines potential directions for future research.

https://doi.org/10.3390/stats9020039

2. Test Statistics Based on Data Depth Ratios

2.1. Common Notions of Data Depth

Let $x \in \mathbb{R}^d$ be an observation from a d-dimensional distribution F. We consider four commonly used measures of data depth.

- Half-space depth [7] at $x \in R^d$ is the minimum probability attached to any half-space containing x:

$$HD(x;F) = \inf\{P(H) : H \text{ is a closed half-space}, x \in H\}.$$

  Tukey's half-space depth performs optimally for elliptical or symmetric distributions, where its depth contours naturally coincide with the underlying shape of the data. However, it also maintains robustness and interpretability in non-normal or heavy-tailed distributions, offering a reliable measure of centrality and dispersion even in the presence of outliers.
- Mahalanobis depth [8]:

$$MD(x;F) = \frac{1}{1+(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

  where μ and Σ are the mean and covariance matrix of the random vector X. Mahalanobis depth measures how far a data point is from the center of a distribution, taking into account the spread and correlation of the data. It performs particularly well when data follow elliptically symmetric distributions such as the multivariate normal distribution, where its assumptions align closely with the true structure of the data. However, it may perform poorly for non-elliptical or multimodal distributions.
- Spatial depth [9–12]:

$$SD(x;F) = 1 - \left\| \mathbb{E}\left[\frac{x-X}{\|x-X\|} \right] \right\|$$

  where $X \sim F$, $\|\cdot\|$ denotes the Euclidean norm, and $\frac{x-X}{\|x-X\|}$ is the spatial sign vector. Spatial depth offers robustness, broad applicability, and simplicity without requiring strong distributional assumptions. It is well-suited for skewed, multimodal, or irregularly shaped data distributions.
- Projection depth [13,14]:

$$PD(x;F) = \frac{1}{1+O(x;F)}$$

  where $O(x;F)$ is defined as

$$O(x;F) = \sup_{\|\nu\|=1} \frac{|\nu^T x - Med(\nu^T X)|}{Med|\nu^T X - Med(\nu^T X)|}$$

  with *Med* denotes the median function and $X \sim F$. Projection depth is highly robust to outliers, as it considers the worst-case deviation of a point from the center of the distribution across all projections. It also adapts well to non-elliptical, skewed, or multimodal distributions.

The above four measures of data depth are implemented in the R package *ddalpha* [8–10,13–15].

2.2. Review of Depth-Based Rank Statistics

A unifying treatment of depth functions was provided in [13], which proposed general structural frameworks for their construction. Within this framework, four fundamental properties were identified. These four properties are affine invariance, maximality at the

center, monotonicity relative to the deepest point and vanishing at infinity. These properties are now standard criteria for evaluating depth notions. Extending these ideas beyond multivariate data, Nieto-Reyes and Battey [16] proposed an axiomatic definition of depth for functional data based on six properties, offering a principled framework for selecting appropriate depth functions in infinite-dimensional settings.

Using statistical depth as a ranking mechanism, Liu and Singh [8] introduced a depth-based multivariate rank-sum test motivated by applications in quality control and experimental studies. Let F denote a reference population and G a future "check" population. They defined the quality index

$$Q(F, G) = P\{D(X; F) \leq D(Y; F) \mid X \sim F, Y \sim G\},$$

which measures the overall outlyingness of G relative to F. When F and G are unknown, the index is estimated by

$$Q(F_m, G_n) = \frac{1}{n} \sum_{i=1}^{n} R(y_i; F_m),$$

where F_m and G_n are the empirical distributions and $R(y_i; F_m)$ is the proportion of observations x_j such that

$$D(x_j; F_m) \leq D(y_i; F_m).$$

This construction yielded a natural multivariate extension of the Wilcoxon rank-sum statistic.

Liu and Singh [17] proposed nonparametric, depth-based rank tests for detecting multivariate scale differences by using data depth to induce a center-outward ordering of observations. The resulting tests are affine invariant, distribution-free under the null hypothesis, and consistent against a wide range of scale alternatives, providing a robust and effective extension of classical univariate scale tests to multivariate data.

Zuo and He [18] established the asymptotic theory for the Liu–Singh statistic under general regularity conditions. The quality index is generally asymmetric due to the choice of the reference distribution

$$Q(F_m, G_n) \neq Q(G_n, F_m).$$

However, under the null hypothesis of identical distributions, this asymmetry disappears asymptotically. Specifically,

$$\left[\frac{1}{12}\left(\frac{1}{m} + \frac{1}{n}\right)\right]^{-1} \left[Q(F_m, G_n) - \frac{1}{2}\right]^2 \xrightarrow{d} \chi^2(1),$$

with an analogous result holding for $Q(G_n, F_m)$. Moreover, under H_0,

$$Q(F_m, G_n) - \tfrac{1}{2} \approx \tfrac{1}{2} - Q(G_n, F_m),$$

which provides the basis for a two-sample test. Their results further showed that the depth-based rank-sum test can be more powerful than several commonly used multivariate rank tests.

Building on the asymptotic behavior of the Liu–Singh quality index Q, Shi et al. [19] proposed test statistics that utilize the two directional measures $Q(F_m, G_n)$ and $Q(G_n, F_m)$. The maximum statistic is defined as

$$M_{m,n} = \left[\frac{1}{12}\left(\frac{1}{m} + \frac{1}{n}\right)\right]^{-1} \max\left\{\left(Q(F_m, G_n) - \tfrac{1}{2}\right)^2, \left(Q(G_n, F_m) - \tfrac{1}{2}\right)^2\right\}.$$

The maximum statistic resolves the asymmetry inherent in the quality index Q, ensuring that the resulting two-sample test is no longer affected by the choice of reference distribution.

2.3. Review of Data-Depth Discrepancy

The Q statistic-based approach and the data depth–based approach offer two different ways of using depth information in two-sample testing. The Q statistic-based approach is attractive because its asymptotic null distribution is well understood and it tends to show similar power across different depth notions. However, since it is fundamentally based on rank-sum principles, it mainly reflects relative ordering and does not fully account for how large the depth differences are. In contrast, data depth–based approaches work directly with numerical depth discrepancies, which allows them to capture more subtle distributional differences in certain settings. This flexibility comes at a cost: their performance can depend strongly on the choice of depth function, and a general asymptotic theory under the null for some statistics is not well-studied. As a result, selecting an appropriate depth notion becomes an important practical consideration when aiming for good power and accuracy.

By explicitly incorporating numerical values of data depth, Niyogi and Dhar [5] proposed the data-depth discrepancy (DDD) statistic to quantify differences between two distributions. The DDD between the two distributions at a point $x \in \mathbb{R}^d$ is defined as

$$\mathrm{DDD}(x; F, G) = D(x; F) - D(x; G),$$

where the $D(\cdot;\cdot)$ function is a measure of data depth. Their work represented the first to leverage numerical depth values, rather than depth ranks, for multivariate two-sample homogeneity testing, although it is restricted to Tukey's half-space depth. The two-sample homogeneity test based on the DDD framework is carried out using a bootstrap procedure. The empirical p-value is obtained by bootstrap from the pooled sample $F_m \cup G_n$. At each bootstrap iteration, $m + n$ observations are drawn with replacement, with the first m observations assigned to the first group and the remaining n observations assigned to the second group, thereby mimicking the null hypothesis of identical distributions. For each resampled dataset, the DDD test statistic is computed as the sum of squared differences between depth values evaluated with respect to the two empirical distributions,

$$S_{m,n} = \sum_{i=1}^{m+n} \{D(z_i; F_m) - D(z_i; G_n)\}^2, \qquad z_i \in F_m \cup G_n.$$

The observed value of $S_{m,n}$ is then compared with its bootstrap distribution, and the p-value is computed as the proportion of bootstrap statistics that exceed the observed test statistic from the original data. The DDD statistic demonstrates strong power relative to the competing test statistics. In addition, the theoretical properties of the DDD statistic, such as consistency and its behavior under local alternatives, were thoroughly investigated and well-established.

Bootstrap and permutation methods are widely used resampling techniques, each with different properties in two-sample testing [20–24]. Bootstrap methods are primarily used to estimate variability, construct confidence intervals, and assess uncertainty. While it can be applied to hypothesis testing, its sensitivity to subtle distributional differences may sometimes be limited, and bootstrap tests are generally not exact under the null hypothesis. Permutation tests, which resample without replacement, provide exact inference under the assumption of exchangeability and can be considered as an alternative resampling strategy [25,26]. They may offer advantages in certain scenarios, although their computational cost can be higher for large datasets. Motivated by these considerations, we investigate the proposed test statistics across alternative depth functions beyond half-space depth and explore whether incorporating a permutation-based approach could help improve the detection of distributional differences while maintaining robustness across different depth notions.

2.4. Data Depth Ratio-Based Statistics

The DDD statistic measures the discrepancy in data depth between two distributions, with larger deviations from zero indicating greater dissimilarity. Inspired by this concept, we introduce a new approach to quantify differences between distributions based on their depth values. The underlying idea is that similar distributions yield comparable depth measurements, while substantial variations in depth reflect divergence in key characteristics such as location, scale, or shape. This approach provides a more intuitive and flexible framework for assessing distributional differences.

Instead of comparing two distributions through the difference of the data depth measure, we consider two related depth-based quantities:

(i) Data depth ratio (DDR), which is defined as

$$DDR(x; F, G) = \frac{D(x; F)}{D(x; G)}.$$

(ii) Logarithmic Data Depth Ratio (log DDR), defined as

$$\log(DDR(x; F, G)) = \log\left(\frac{D(x; F)}{D(x; G)}\right) = \log(D(x; F)) - \log(D(x; G)).$$

The log DDR can be viewed as a logarithmic-scale analogue of the DDD statistic, in which relative depth information is expressed additively rather than multiplicatively.

Using ratios to quantify differences offers a relative and unit-independent way to compare data. Unlike absolute differences, ratios reflect proportional changes, making results more interpretable and comparable. This normalization improves robustness and ensures that conclusions remain valid under transformations or variations in data magnitude. As a result, ratio-based measures such as the F-ratio or likelihood ratio are commonly employed to highlight meaningful discrepancies between groups and enhance interpretability in hypothesis testing and statistical analysis.

Depth-based measures can have large variances, especially in high-dimensional data, and applying the logarithm helps to compress large values while expanding smaller ones, making differences between distributions more discernible. The logarithmic transformation also stabilizes the variance and reduces the influence of extreme values or outliers, improving the sensitivity of DDR to detect distributional differences. Applying a logarithmic transformation to the DDR yields a more stable and interpretable test statistic, especially under skewed or heavy-tailed distributions. For a depth ratio $DDR(x; F, G) = D(x; F)/D(x; G)$, variability can be unstable, particularly when the denominator is small. This brings an issue of how to choose the denominator as $D(x; F)/D(x; G) \neq D(x; G)/D(x; F)$. Applying a logarithmic transformation preserves the symmetry of the statistic, as taking the absolute value or square of the log-transformed ratio ensures that the measure remains consistent regardless of which distribution is used as the reference (i.e., $|\log(DDR(x; F, G))| = |\log(DDR(x; G, F))|$. This property makes the statistic invariant to the ordering of the distributions, thereby enabling a fair and balanced comparison. The logarithmic transformation also helps mitigate computational issues that arise when the depth ratio approaches 1. In such cases, numerical precision can be lost because standard computing software may truncate or round off decimal values once they exceed the maximum number of representable digits. By applying the logarithmic transformation, ratios near 1 are mapped to values near zero, preserving small differences and reducing the risk of numerical instability during computation. Therefore, we apply a logarithmic transformation to the DDR to enhance its interpretability and improve the stability of the resulting test statistics.

Since ratio calculations are undefined for zero depth values, half-space depth cannot be applied in the DDR-based statistic. In finite samples, this depth often yields zero values for observations near the data boundary because it depends on the smallest probability mass contained within any half-space. To avoid this limitation and achieve more stable ratio computations, the log DDR statistic incorporates alternative depth functions such as Mahalanobis, projection, and spatial depths.

2.5. Data-Depth Plot

The DD plot, introduced by Liu, Parelius and Singh [6], is a visualization tool for comparing two multivariate distributions using data depth. Given a depth function $D(\cdot;\cdot)$, each observation x is represented in a two-dimensional space by the pair

$$\big(D(x;F),\, D(x;G)\big),$$

where F and G denote the two distributions. This representation enables a robust, nonparametric assessment of relative centrality: observations with larger depth with respect to one distribution are closer to that population, while points near the origin indicate outlying or atypical observations.

Our goal is to study the behavior of various depth functions and assess their sensitivity to distributional differences using data depth ratio–based statistics, thereby enabling a comparison of their effectiveness and improving test performance when an appropriate depth notion is employed. We consider each sample consists of 100 independent and identically distributed observations, and the two samples are from the following distributions:

- Model 1: Standard bivariate normal distribution vs. standard bivariate normal distribution.
- Model 2: Standard bivariate normal distribution vs. standard bivariate Cauchy distribution.
- Model 3: Standard bivariate normal distribution vs. standard bivariate t_6.
- Model 4: $F = 0.8N(\binom{0}{0}, \mathbf{I}) + 0.2N(\binom{15}{15}, \mathbf{I})$ vs. $G = 0.8N(\binom{7}{7}, \mathbf{I}) + 0.2N(\binom{-8}{-8}, \mathbf{I})$.

Each DD plot contains 200 points, each representing a pair of data depth values. To clearly distinguish observations from the two samples, different colours are used to indicate sample membership. The blue points correspond to observations from distribution F, while the red points correspond to those from distribution G. When the two distributions are identical, the points tend to lie close to the 45-degree line. A noticeable deviation from this diagonal line indicates differences between the distributions F and G. Liu, Parelius and Singh [6] identified several characteristic patterns in DD plots that correspond to specific types of distributional differences. The DD plot, with its clear departure from the diagonal and absence of a common maximum depth point, indicates a location shift between the two distributions. The half-moon pattern, oriented toward the distribution with a larger scale, reflects greater variability. Slight asymmetry around the diagonal suggests differences in higher-order moments, such as skewness or kurtosis, indicating that the distributions differ beyond location and scale. Figure 1 illustrates that all four depth functions exhibit similar behavior when both samples are from the same distribution, as the points are largely concentrated around the diagonal line. Among them, projection depth shows the smallest deviation, with points lying very close to the diagonal, while spatial and Mahalanobis depths produce a more elliptical pattern. To quantify these observations, the total distances of all points from the diagonal line are computed and reported in Table 1. Under the null hypothesis, the degree of departure is greatest when using half-space depth, whereas projection depth yields the smallest total distance. Figure 2 indicates that all four depth functions successfully distinguish the standard normal and standard bivariate Cauchy distributions, since all points are far from the diagonal line. In particular, the Mahalanobis depth displays the strongest

discrepancy, with points lying far from the diagonal. Table 1 also shows that the Mahalanobis depth produces the largest total distance, whereas the smallest one belongs to the projection depth. Figure 3 exposes a limitation of the projection and half-space depth, since most points align closely with the diagonal, suggesting that these two depth functions are less sensitive to distributional differences in this setting, whereas the remaining depth functions continue to capture the discrepancy. Table 1 indicates that projection depth results in the smallest total distance, whereas Mahalanobis depth exhibits the largest. Figure 4 further demonstrates the weakness of the Mahalanobis depth for multimodal data, where it fails to detect differences between two normal mixture distributions, while the projection depth shows superior performance in identifying such discrepancies. The results in Table 1 highlight the relative weakness of Mahalanobis depth and the superior performance of projection depth, as the total distance is largest for Mahalanobis depth and smallest for projection depth. The patterns observed in DD plots can indicate specific types of distributional discrepancies. As a visualization tool, DD plots, together with quantitative measures such as the distance of points from the diagonal, provide informative assessments of the behavior of depth functions. These findings guide the selection of appropriate depth functions for specific scenarios to enhance the performance of the two-sample tests. They also suggest a practical pathway for visualizing data depth and identifying which depth notion should be employed to achieve the most effective testing performance. The simulation results in Section 3 demonstrate that different depth functions reflect different features of the underlying distribution, thereby influencing the performance of the associated test statistics.

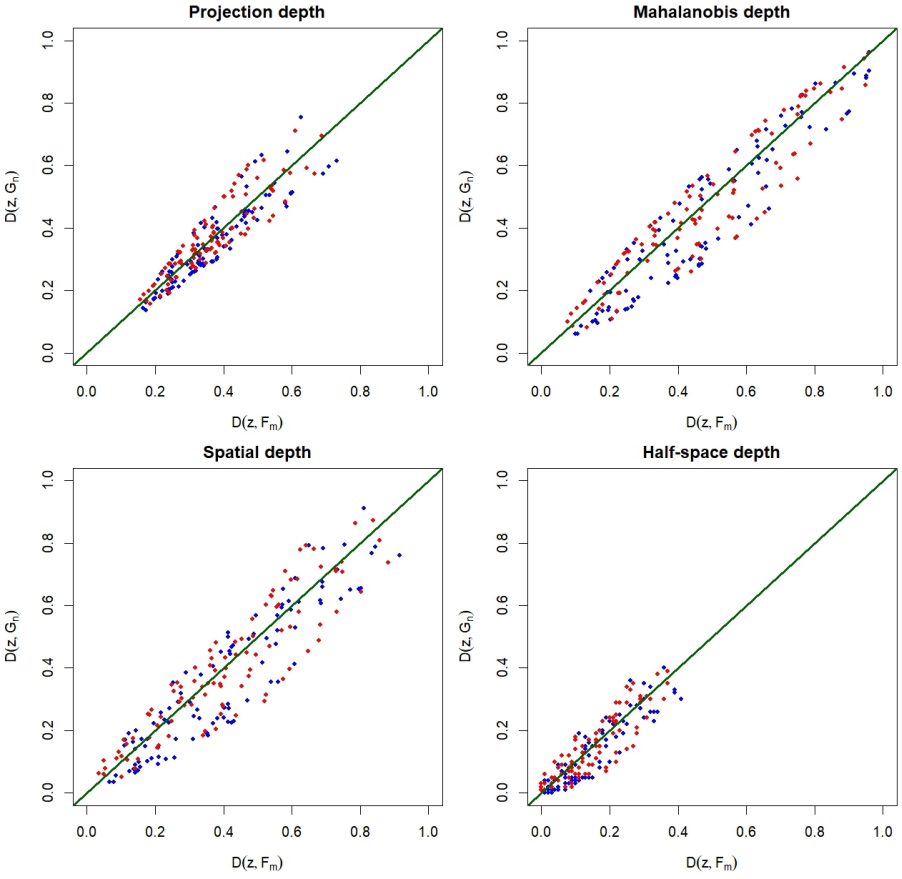


Figure 1. DD plot of $X \sim N(\binom{0}{0}, \mathbf{I})$ vs. $Y \sim N(\binom{0}{0}, \mathbf{I})$.

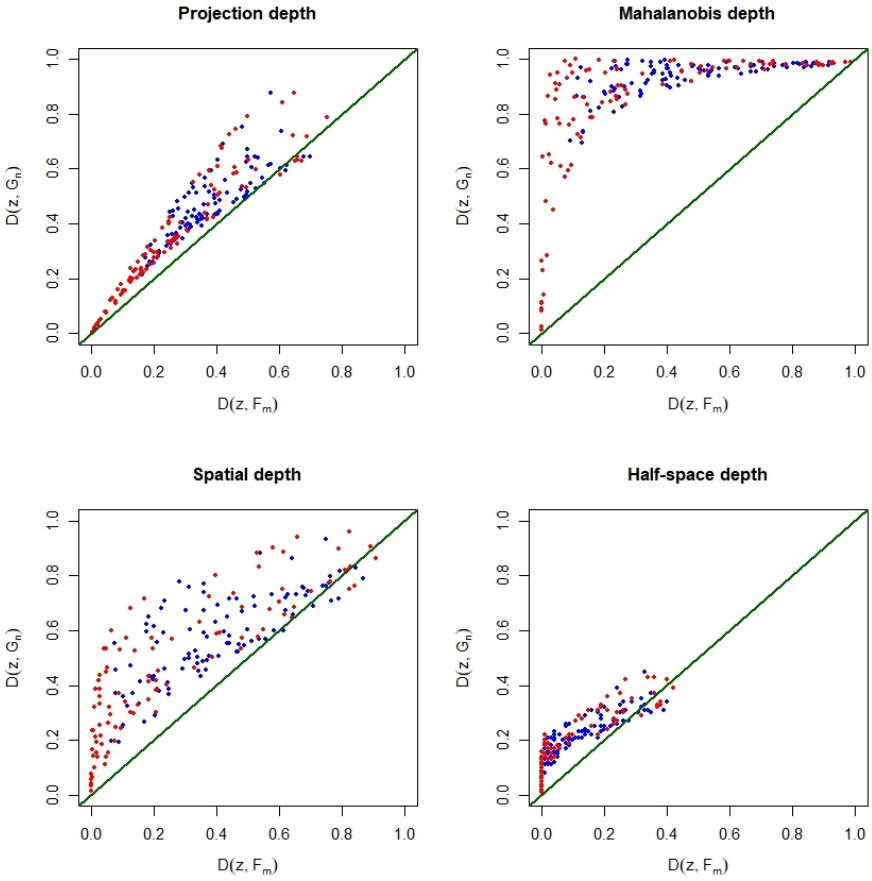


Figure 2. DD plot of $X \sim N(\binom{0}{0}, \mathbf{I})$ vs. $Y \sim$ Standard bivariate Cauchy.

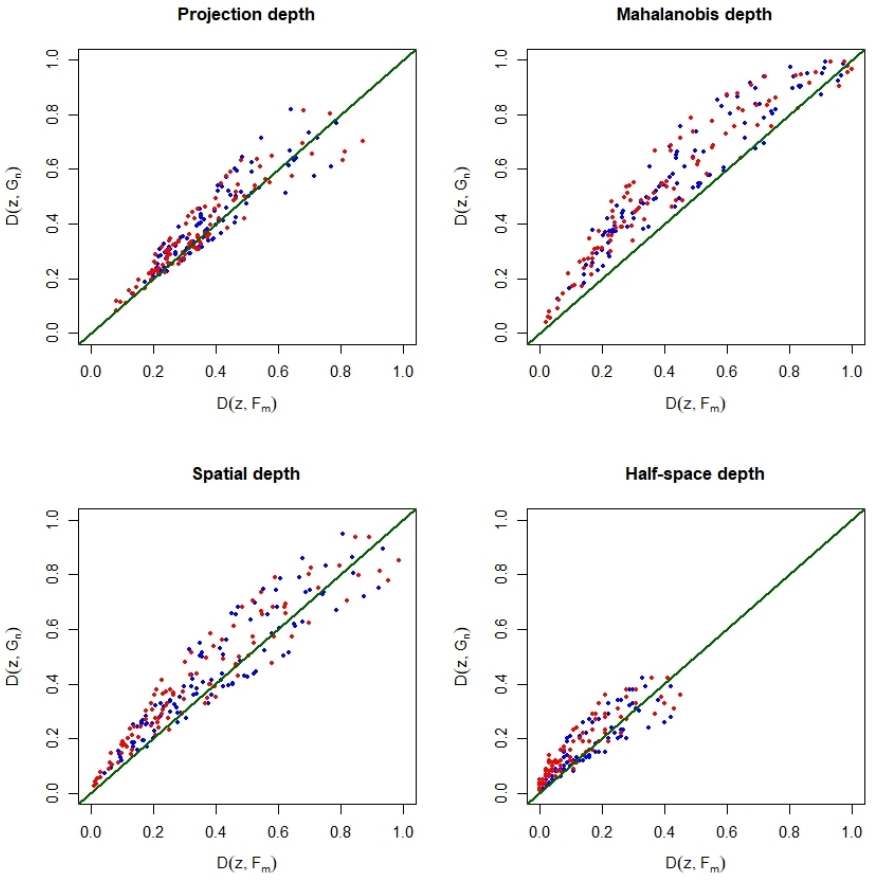


Figure 3. DD plot of $X \sim N(\binom{0}{0}, \mathbf{I})$ vs. $Y \sim t(\binom{0}{0}, \mathbf{I}, 6)$.

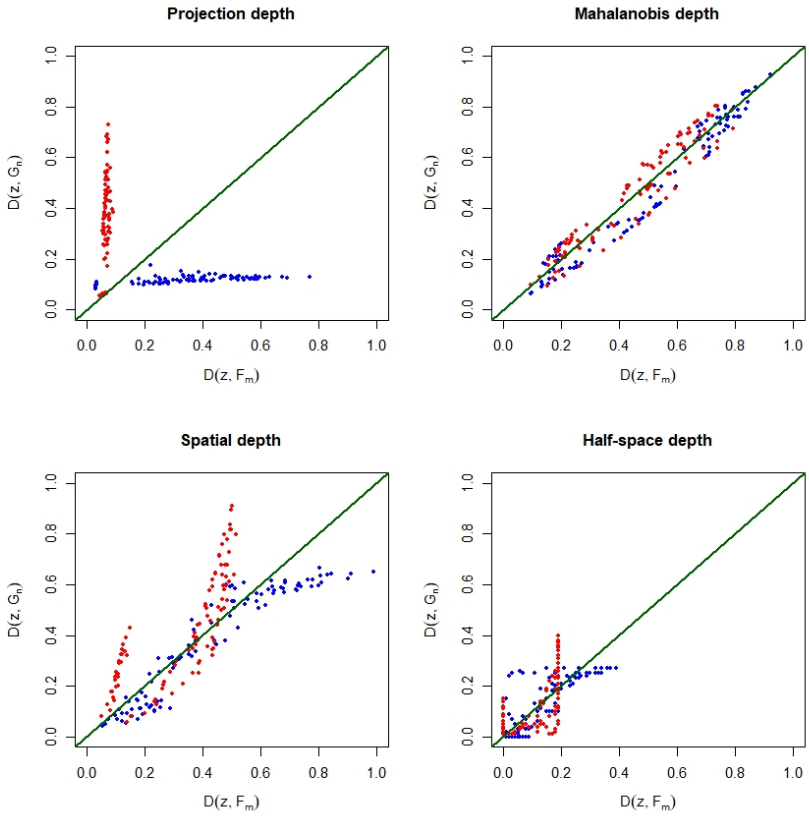


Figure 4. DD plot of $X \sim 0.8N((\begin{smallmatrix}0\\0\end{smallmatrix}), \mathbf{I}) + 0.2N((\begin{smallmatrix}15\\15\end{smallmatrix}), \mathbf{I})$ vs. $Y \sim 0.8N((\begin{smallmatrix}7\\7\end{smallmatrix}), \mathbf{I}) + 0.2N((\begin{smallmatrix}-8\\-8\end{smallmatrix}), \mathbf{I})$.

Table 1. Approximate total distances of the points on DD plot to the diagonal line for four depth functions for sample sizes $m = n = 1000$ and 1000 replications.

Model	Mahalanobis	Projection	Spatial	Half-Space
1	11.22	8.64	11.50	**12.54**
2	**97.19**	16.06	34.75	32.12
3	**17.53**	8.87	12.98	14.18
4	15.12	**51.74**	24.70	23.22

Note: The maximum values are written in bold font.

2.6. Testing the Homogeneity of Two Multivariate Samples

In this section, we present a significance test for the two-sample problem. Specifically, to test $H_0 : F = G$ against $H_1 : F \neq G$, we define the log DDR statistic, denoted by $T_{m,n}$, as follows

$$T_{m,n} = \sum_{i=1}^{m+n} [\mathbf{log}(DDR(z_i; F_m, G_n))]^2 \text{ for } z_i \in F_m \cup G_n,$$

where F_m, G_n are the empirical distributions of $\mathcal{X}$ and $\mathcal{Y}$, respectively. Note that the proposed test statistic $T_{m,n}$ involves empirical depth functionals combined with nonlinear ratio transformations, which induce complex dependence structures and render a complete asymptotic analysis technically challenging. A full derivation of the asymptotic null distribution would require substantial additional development and is beyond the scope of the present work.

In modern nonparametric two-sample methodology, it is common for new procedures to combine structural motivation with permutation calibration and extensive empirical evaluation. Within this framework, simulation studies play an important role in examining power, robustness, and stability across a variety of alternatives. We provide an algorithm

to compute an empirical p-value for the proposed test statistic $T_{m,n}$ using a bootstrap procedure with a sufficiently large number of replications B.

The procedure begins by computing the data depth of the observations with respect to the samples $\mathcal{X}$ and $\mathcal{Y}$, and evaluating the test statistic $T_{m,n}$. The two samples are then combined to form a pooled sample $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$. For each of the B bootstrap replicates, a sample of size $(m+n)$ is drawn from $\mathcal{Z}$ with replacement, where the first m observations are assigned to $\mathcal{X}^*$ and the remaining n observations to $\mathcal{Y}^*$. Using these resampled datasets, the data depth values are recomputed and the corresponding test statistic $T^*_{m,n}$ is obtained. The empirical p-value is computed as the proportion of bootstrap replicates for which the resampled test statistic $T^*_{m,n}$ exceeds the observed test statistic $T_{m,n}$. An alternative version of this procedure is based on a permutation strategy, in which, at each replicate, a sample of size $(m+n)$ is drawn from $\mathcal{Z}$ without replacement. The first m observations are taken as $\mathcal{X}^*$ and the remaining n observations as $\mathcal{Y}^*$, and the test statistic is computed in the same manner. The pseudo-code for this procedure is given below (Algorithm 1).

Algorithm 1 Empirical p-value computation.

Input: Samples $\mathcal{X}, \mathcal{Y}$; number of replicates B
Output: $\hat{p}$

Compute $T_{m,n} \leftarrow T(\mathcal{X}, \mathcal{Y})$
$\mathcal{Z} \leftarrow \mathcal{X} \cup \mathcal{Y}$
for $b = 1$ to B **do**
- Sample $\mathcal{Z}^{*(b)}$ of size $m+n$ from $\mathcal{Z}$
- Partition $\mathcal{Z}^{*(b)} = \left(\mathcal{X}^{*(b)}, \mathcal{Y}^{*(b)}\right)$ with sizes m and n
- Compute $T^{*(b)}_{m,n} \leftarrow T\left(\mathcal{X}^{*(b)}, \mathcal{Y}^{*(b)}\right)$

end for
$\hat{p} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left(T^{*(b)}_{m,n} > T_{m,n}\right)$
return $\hat{p}$

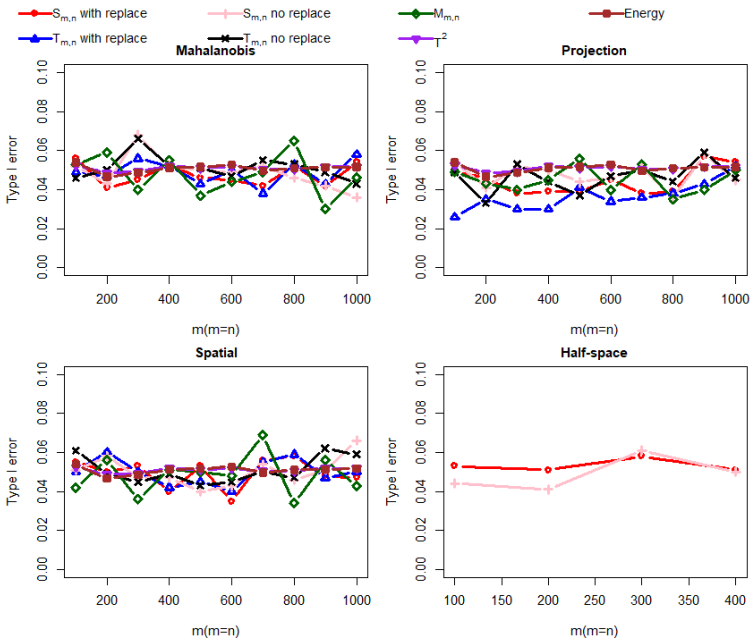
3. Simulation Studies

We consider power comparisons among the DDD statistics $S_{m,n}$, the log DDR statistics $T_{m,n}$, the maximum statistics $M_{m,n}$, the Hotelling's T^2 test (denoted by T^2) and the Energy test for the two-sample test. Two random samples $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$ and $\mathcal{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ are randomly drawn from the distributions F and G, respectively. For the sample sizes, we consider $m = n$ and $m = 2n$ where $m = 100, 200, 300, \ldots, 1000$. We assume F is the standard bivariate normal distribution. The distribution G is taken to be either a standard bivariate normal distribution, a standard bivariate Cauchy distribution, or a standard bivariate t_6 distribution. The inclusion of these alternatives allows us to assess test performance under varying tail behaviors and distributional deviations from normality. In particular, the multivariate t_6 distribution is commonly used to model small-sample multivariate data, as its heavier tails account for increased variability and potential outliers. This makes it particularly useful in simulations and methodological studies for evaluating the robustness of statistical procedures, including tests of means, covariances [27]. With six degrees of freedom, the distribution provides a realistic balance between capturing small-sample behaviour, maintaining robustness against extreme values, and remaining computationally manageable, making it a practical choice in both theoretical and applied multivariate analyses.

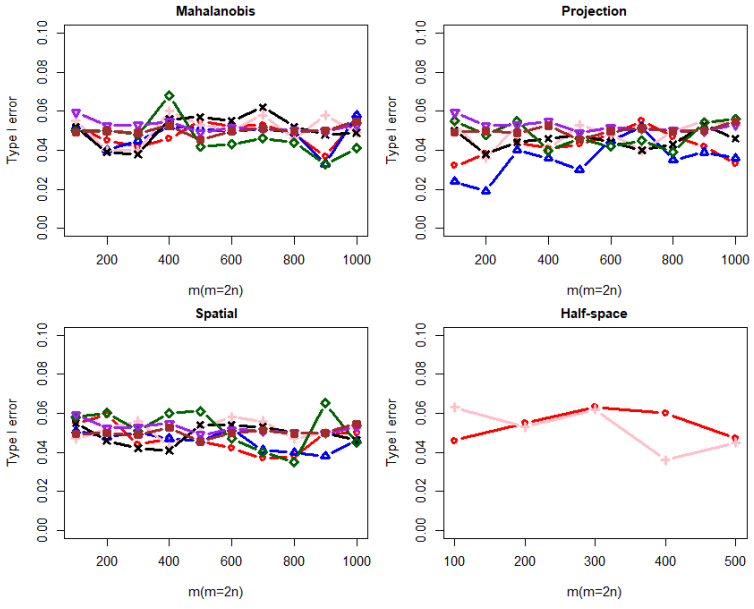
The empirical power is studied with 1000 repetitions, 500 bootstrap replications, and examined with Mahalanobis depth, projection depth, spatial depth for all three test statistics and half-space depth for $S_{m,n}$. Power is computed as the proportion of the

empirical p-values that is less than the lower 5% quantile of the simulated p-values under the null hypothesis.

We first assess the Type I error of all test statistics at the nominal significance level $\alpha = 0.05$. As shown in Figure 5, all statistics reject the null hypothesis at close to the expected 5% rate across 1000 simulation replications. However, the speed at which the Type I error approaches the nominal level depends on the chosen depth function. The maximum statistic shows pronounced fluctuations in Type I error for sample sizes below 1000, especially under Mahalanobis and spatial depths, while the other statistics exhibit relatively minor variation regardless of the depth notion. The log DDR statistic with the permutation strategy displays particularly stable performance, converging consistently to the nominal level at a sample size of 1000 across all depth functions. By comparison, the remaining statistics are slightly more sensitive to the choice of depth and exhibit less stable convergence behavior.



(**a**)



(**b**)

Figure 5. Simulated Type I error for $m = 100, 200, \cdots, 1000$ with (**a**) $m = n$, (**b**) $m = 2n$.

We assess the performance of the log DDR statistic under non-normal alternatives using the standard bivariate Cauchy distribution. As illustrated in Figure 6, the log DDR test exhibits a rapid gain in power as the sample size increases from 100 to 200 and maintains stable performance even when the sample sizes are unequal. Overall, all test statistics achieve their highest power when Mahalanobis depth is employed. In contrast, the DDD statistic is generally less powerful than log DDR and shows particularly poor performance when spatial depth is used. The maximum statistic displays relatively consistent performance across different depth notions, whereas both log DDR and DDD are strongly influenced by the choice of depth function. This behavior is consistent with the patterns observed in the DD plots presented in the previous section. Consequently, we recommend using DD plots as a practical tool to guide the selection of an appropriate depth notion. In this setting, the Energy test demonstrates essentially perfect power, while Hotelling's T^2 test is unable to identify the distributional difference.

Similarly, for the t_6 distribution, Figure 7 shows that log DDR quickly approaches full power as the sample size increases. However, the performance of both DDD and log DDR strongly depends on the depth function: projection depth and half-space depth tend to underperform, whereas Mahalanobis and spatial depths consistently provide higher power. This result is consistent with the findings in the DD plot, where the projection depth illustrates its drawback in dealing with a heavy-tailed distribution such as standard bivariate t_6. Log DDR drastically outperforms DDD when using the spatial depth function. Figure 7 shows the poor performance of DDD statistics when the spatial depth function is used. The power of DDD statistics tests using spatial depth in these two cases does not reach 1.0 even when the sample sizes are 1000. Again, bootstrap with replacement is generally less powerful than the permutation strategy. In this scenario, the maximum statistic, the Energy test, and Hotelling's T^2 test fail to detect the distributional discrepancy, with the power of all three tests remaining below 1.0.

We also investigate the performance of all statistics when comparing

$$F = 0.8N(\binom{0}{0}, \mathbf{I}) + 0.2N(\binom{15}{15}, \mathbf{I}) \text{ vs. } G = 0.8N(\binom{7}{7}, \mathbf{I}) + 0.2N(\binom{-8}{-8}, \mathbf{I}).$$

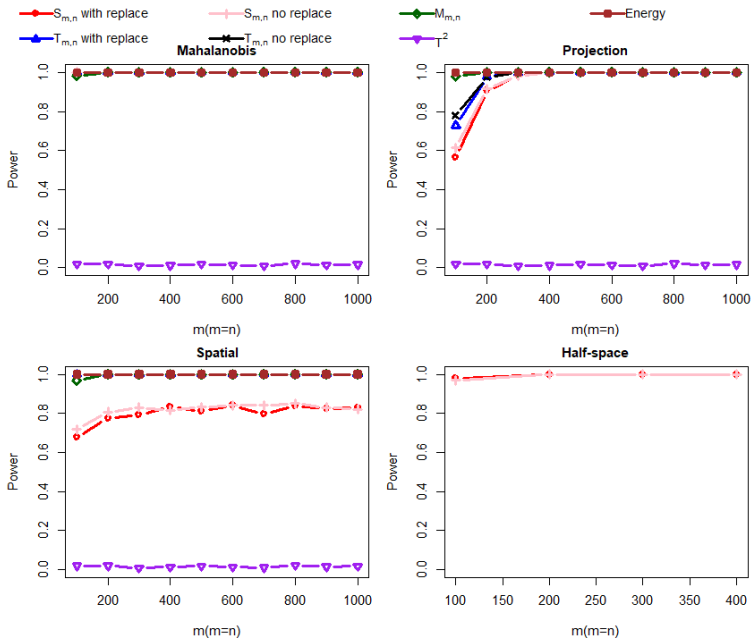
As shown in Figure 8, both the DDD and log DDR statistics exhibit reduced power when the Mahalanobis depth is used, whereas the projection depth demonstrates outstanding performance, achieving a test power of 1.0 even for the smallest sample sizes. The statistics also perform well when paired with spatial depth, highlighting the substantial impact of depth function choice on the ability to distinguish between distributions. In particular, projection depth is especially effective for multimodal distributions, such as mixture models, while Mahalanobis depth is less suitable in these scenarios. These simulation results confirm the usefulness of the DD plot as a diagnostic tool for identifying the most appropriate depth function to employ in the test statistics. While the Energy test performs exceptionally well alongside other statistics with projection depth, Hotelling's T^2 test shows reduced power in the mixture Gaussian scenario.

The performance of all statistics is also investigated when comparing

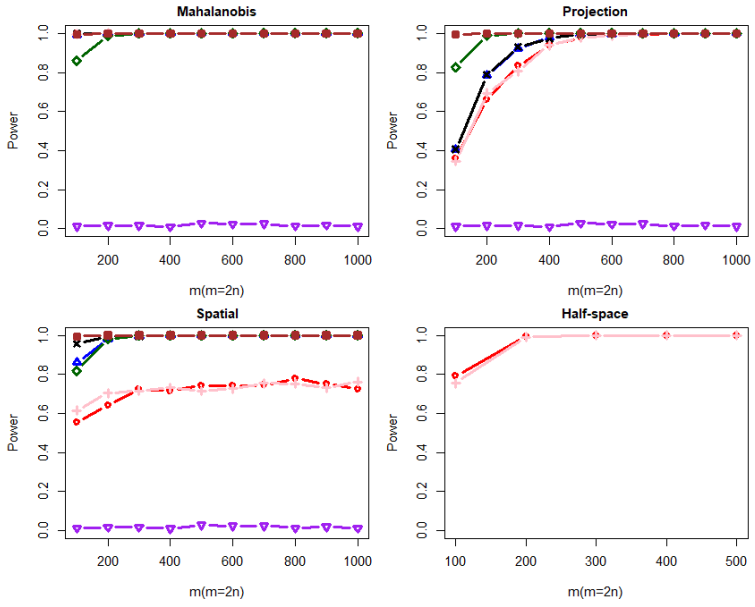
$$F \sim N(\mathbf{0}_{10}, \mathbf{I}) \text{ vs. } G = N(0.3\,\mathbf{1}_{10}, \mathbf{I}).$$

All test statistics demonstrate a strong ability to detect distributional discrepancies as the dimensionality increases. Figure 9 shows that both the DDD and log DDR statistics achieve performance comparable to the Energy test and Hotelling's T^2 test when the sample size exceeds 200 for balanced samples and 300 for unbalanced samples. In this setting, the choice of depth function has minimal impact on the performance of $S_{m,n}$ and $T_{m,n}$ statistics. In contrast, the maximum statistic combined with projection depth exhibits

weaker performance, requiring larger sample sizes to achieve full power. Moreover, the permutation strategy further enhances the ability of the test to distinguish between the two distributions in this scenario.



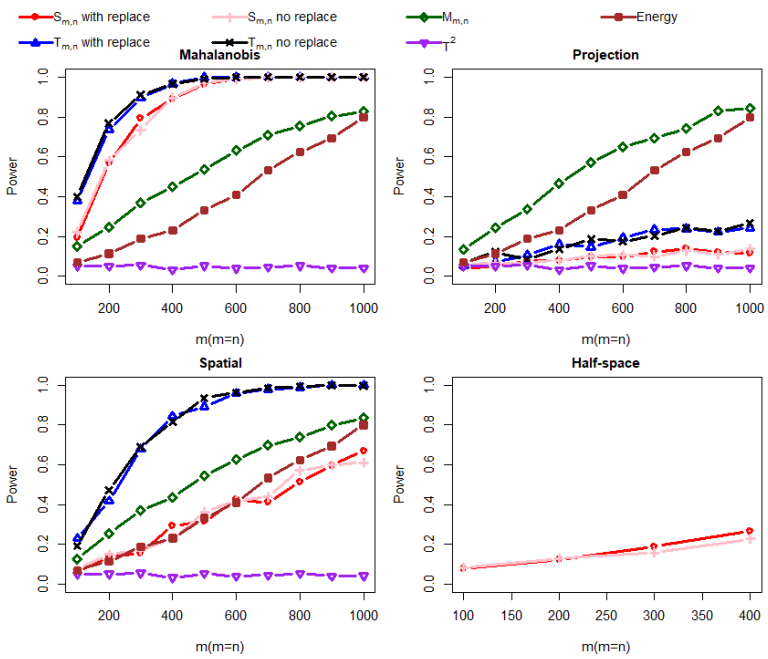
(**a**)



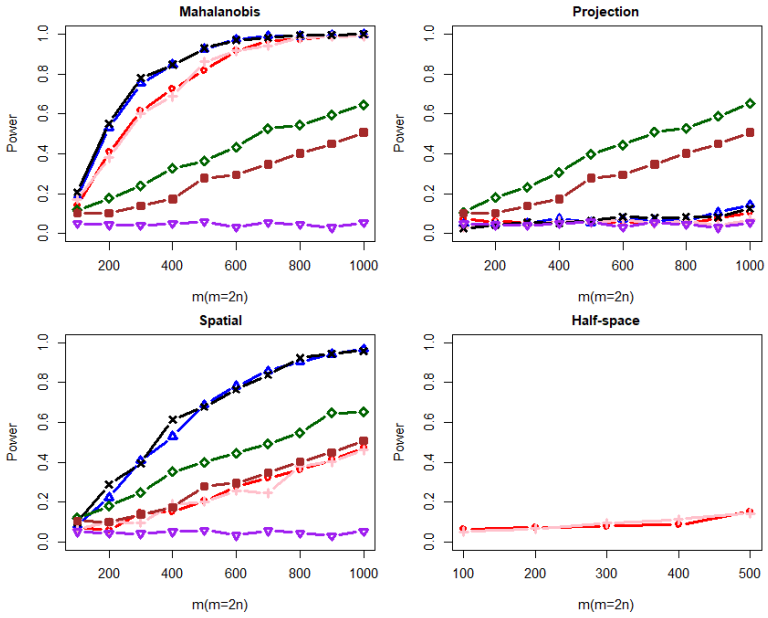
(**b**)

Figure 6. Power comparison under alternative hypothesis $H_1 : F \sim N(\binom{0}{0}, \mathbf{I})$ vs. $G \sim$ standard bivariate Cauchy for $m = 100, 200, \cdots, 1000$ with (**a**) $m = n$, (**b**) $m = 2n$.

The observed reduction in power across all settings under unbalanced sample sizes ($m = 2n$) suggests that most configurations present increased difficulty for all test statistics in detecting distributional differences. This decline in performance may arise from the unequal contribution of the samples, where the larger sample tends to dominate the estimation process, potentially masking subtle discrepancies between the underlying distributions.



(**a**)

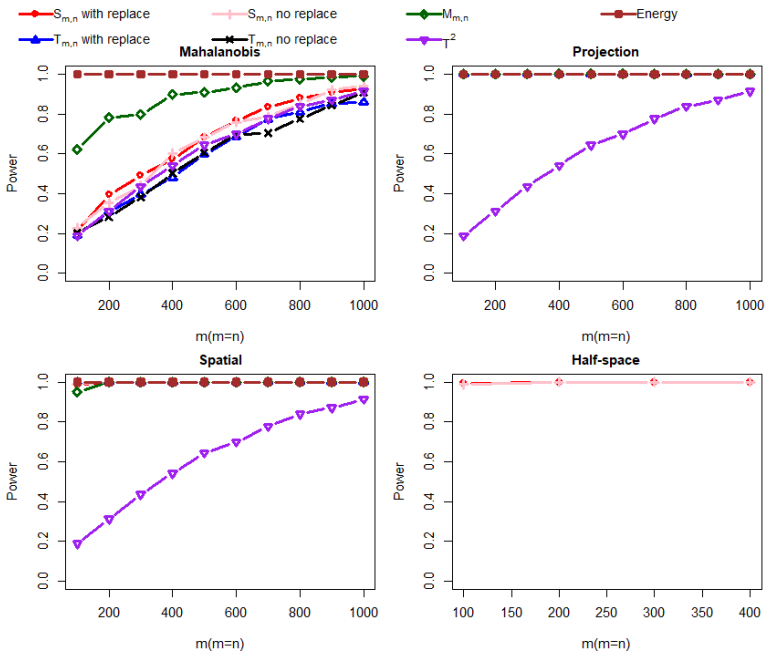


(**b**)

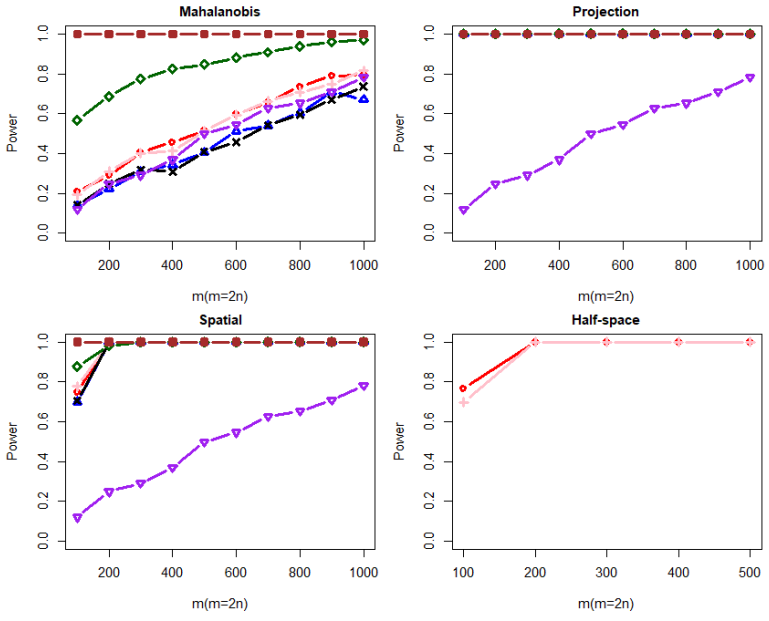
Figure 7. Power comparison under alternative hypothesis $H_1 : F \sim N(\binom{0}{0}, \mathbf{I})$ vs. $G \sim t(\binom{0}{0}, \mathbf{I}, 6)$ for $m = 100, 200, \cdots, 1000$ with (**a**) $m = n$, (**b**) $m = 2n$.

More generally, the simulation study provides evidence that the log DDR statistic outperforms the DDD statistic across a wide range of settings. Bootstrap method exhibits lower power compared with the permutation approach. The strongest performance of both statistics is observed when Mahalanobis depth is applied to elliptical distributions. Although projection depth shows limitations for comparisons involving non-normal distributions, it outperforms other depth functions when analyzing multimodal distributions such as mixtures of normal distributions. In addition, the DDD statistic exhibits instability when used with spatial depth and half-space depth, whereas the log DDR statistic maintains stable and strong performance across all cases when paired with spatial depth. Overall, these findings emphasize that the choice of depth function plays a critical role in

test performance, as different depth notions vary in their ability to capture distributional discrepancies. The DD plot thus provides valuable guidance for selecting the most effective depth function in practice.



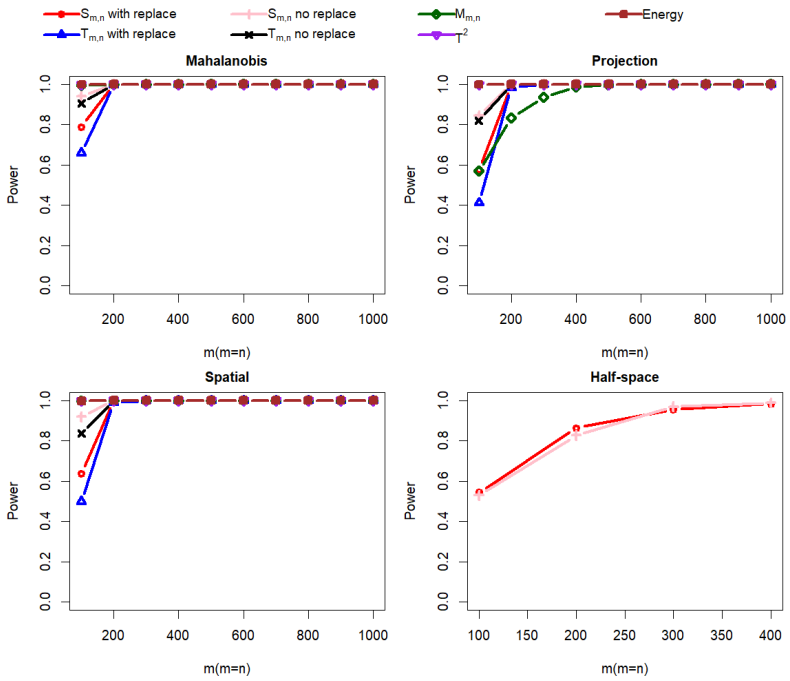
(**a**)



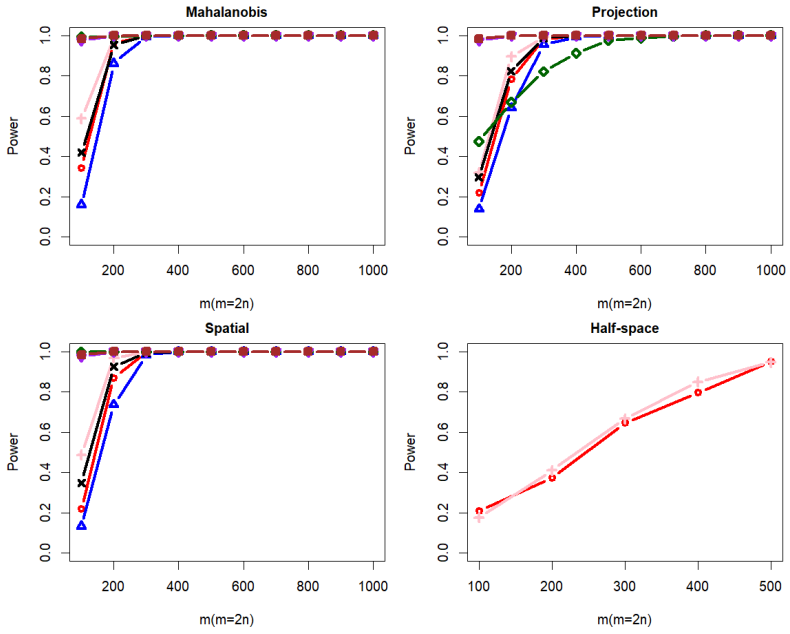
(**b**)

Figure 8. Power comparison under alternative hypothesis $H_1 : F = 0.8N((\begin{smallmatrix}0\\0\end{smallmatrix}), \mathbf{I}) + 0.2N((\begin{smallmatrix}15\\15\end{smallmatrix}), \mathbf{I})$ vs. $G = 0.8N((\begin{smallmatrix}7\\7\end{smallmatrix}), \mathbf{I}) + 0.2N((\begin{smallmatrix}-8\\-8\end{smallmatrix}), \mathbf{I})$ for $m = 100, 200, \cdots, 1000$ with (**a**) $m = n$, (**b**) $m = 2n$.

Half-space depth can be very slow to compute, resulting in high computational costs. Using alternative depth functions solves this problem, dramatically speeding up the calculations. In our study, all three alternative depths reduce the computation time by at least threefold. Table 2 shows the approximate computational times for two samples of dimension $d = 2$ for different sizes and four depth functions with 500 bootstrap replications. The difference is particularly noticeable for larger datasets when the sample size exceeds 500 observations.



(**a**)



(**b**)

Figure 9. Power comparison under alternative hypothesis $H_1 : F \sim N(\mathbf{0}_{10}, \mathbf{I})$ vs. $G \sim N(0.3\,\mathbf{1}_{10}, \mathbf{I})$ for $m = 100, 200, \cdots, 1000$ with (**a**) $m = n$, (**b**) $m = 2n$.

Table 2. Computational times (in seconds) for bivariate sample sizes $m = n = 100, 200, \cdots, 1000$ for four depth functions (500 bootstrap replications).

$m = n$	Mahalanobis	Projection	Spatial	Half-Space
100	2.38	3.16	3.46	3.86
200	2.48	3.99	5.05	8.47
500	2.78	6.91	13.07	83.51
1000	3.30	11.84	38.47	454.60

4. Real Data Analysis

In this section, we illustrate the use of the log DDR statistic in a two-sample test for real data applications and compare it with the DDD statistic.

4.1. Heart Failure Data

Heart failure is a significant health concern impacting millions globally. This complex condition arises when the heart struggles to pump blood efficiently, leading to a range of complications and a reduced quality of life for those affected. Despite significant advances in medical treatments and understanding of the disease, heart failure continues to rank as one of the leading causes of illness and death. Key risk factors were identified, such as high blood pressure, obesity, and diabetes. These factors play a critical role in early detection and treatment. By checking these health conditions, we can predict heart failure and provide early treatment. High blood pressure, obesity, and diabetes can be confirmed by conventional tests such as blood tests, body mass index (BMI), and usual check-ups. These variables are more likely to be unbiased and easily accessible. We want to explore whether the proposed test can differentiate healthy people and patients. The Statlog data [28] consists of 528 healthy people and 520 patients diagnosed with heart failure. The features to be used are resting blood pressure, cholesterol and maximum heart rate. The dataset was first partitioned into two subsets (healthy versus patients). Categorical variables and observations with missing values in the variables of interest were then removed. The selection of variables of interest was based on two main criteria. First, only numerical variables were retained, as the computation of depth functions is designed for continuous distributions. Second, the selected variables are clinically meaningful, considering their accessibility, reliability, and relevance for diagnosis or prognosis of the health condition. Mahalanobis depth, projection depth and spatial depth are used to compute the p-value of each statistics. We consider the permutation test versions of both DDD and log DDR statistics, as it has been shown that the permutation approach is more powerful in detecting distributional differences. The p-values are estimated under 10,000 repititions.

Figure 10 shows that all four depth functions are capable of revealing differences between the two samples, as most points are located far from the 45-degree line. However, the half-space depth appears to be less sensitive, since the points are more tightly clustered, which may reduce its ability to detect discrepancies in depth values. The total distances from points to the diagonal line in the DD plots are 44.58, 31.03, 43.84 and 39.90 for Mahalanobis, projection, spatial and half-space depth, respectively. This result indicates that the Mahalanobis depth and spatial depth can be more suitable than the others in distinguishing the distributions of the two groups. Table 3 presents the empirical p-values for the log DDR statistic and the DDD statistic computed using three depth functions over 10,000 repetitions. All p-values are extremely small, indicating very strong evidence against the null hypothesis. The Mahalanobis depth produces the smallest p-values, suggesting it is the most sensitive among the three depth measures in detecting differences. Notably, the log DDR statistic yields consistent p-values across all depth functions and remains significant even at stricter significance levels, highlighting its robustness in practical applications. The DDD statistic based on half-space depth yields substantially larger p-values, indicating reduced sensitivity.

Table 3. Empirical p-values of log DDR statistic, DDD statistic for heart failure data, under different depth functions (10,000 repetitions).

Depth Function	$T_{m,n}$	$S_{m,n}$
Mahalanobis	0.0000	0.0001
Projection	0.0007	0.0010
Spatial	0.0000	0.0001
Half-space		0.0112

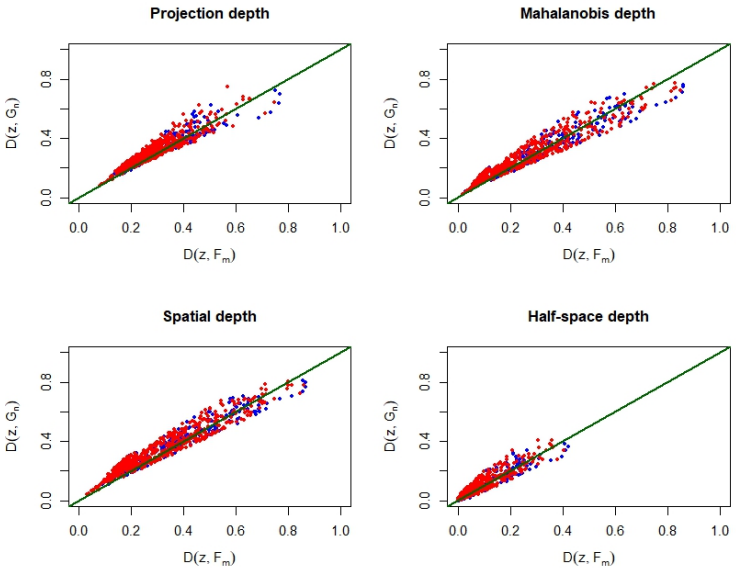


Figure 10. DD plot of healthy people vs. patients with heart failure.

4.2. Breast Cancer Data

Early detection is vital in breast cancer screening, as it substantially enhances the effectiveness of treatment and improves survival rates. In the study by Patrício et al., data from routine blood analyses including Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, MCP-1, Age, and BMI were used to predict breast cancer occurrence [29]. This dataset provides a valuable basis for evaluating the capability of the proposed test to differentiate between healthy controls and patients with breast cancer.

Blood samples [29] were collected from 64 breast cancer patients and 52 healthy controls after overnight fasting. Nine clinical variables were measured, including BMI (kg/m^2), Serum Glucose (mg/dL), Resistin (ng/mL), MCP-1 (pg/dL), Leptin (ng/mL), Adiponectin (µg/mL), and Plasma Insulin (µU/mL). Glucose was measured using an automatic analyzer, while the other biomarkers were quantified using ELISA kits. Insulin resistance was calculated using the HOMA index:

$$\text{HOMA} = \log\left(\frac{I_f \times G_f}{22.5}\right),$$

are fasting insulin and glucose levels, respectively. The data were divided into two subgroups corresponding to healthy controls and breast cancer patients. Only the variables of interest were retained, and the dataset contains no missing values.

Figure 11 illustrates that all four depth notions effectively capture distributional differences, with points widely dispersed and clearly deviating from the 45 degree line. In addition, most observations display distinct sample membership as the points consistently shift toward the distribution from which they originate. The total distances from all points to the diagonal line are 9.09, 4.78, 13.15 and 8.14 for Mahalanobis, projection, spatial and half-space depth. The largest total distance yields to the spatial depth function. Table 4 reports empirical p-values for the log DDR and DDD statistics across three depth functions over 10,000 repetitions. All results are highly significant. The DDD statistic shows greater variation in p-values when the projection depth is used, whereas the log DDR statistic exhibits only slight changes across the three depth functions. These results highlight the robustness and reliability of the log DDR statistic for practical data analysis.

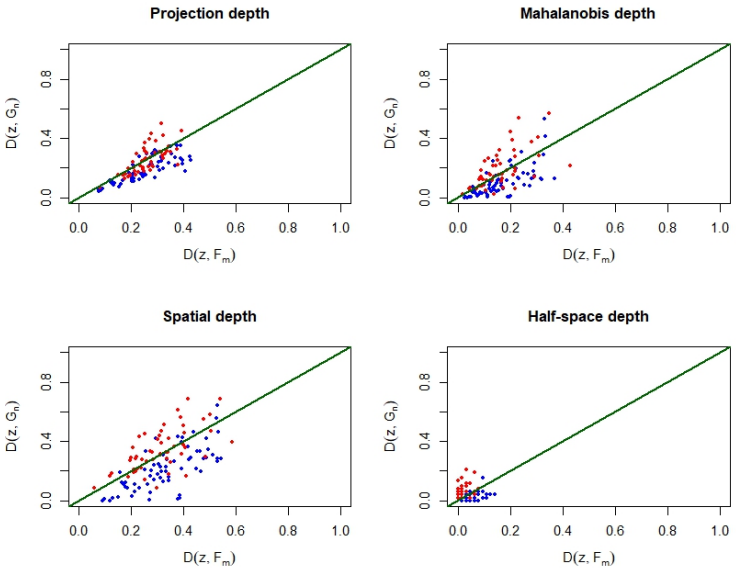


Figure 11. DD plot of blood samples of healthy people vs. cancer patients.

Table 4. Empirical p-values of log DDR statistic and DDD statistic for breast cancer data, under different depth functions (10,000 repetitions).

Depth Function	$T_{m,n}$	$S_{m,n}$
Mahalanobis	0.0008	0.0002
Projection	0.0018	0.0026
Spatial	0.0014	0.0001
Half-space		0.0005

5. Conclusions

In this paper, we investigate depth-based approaches for multivariate two-sample homogeneity testing and propose the log DDR statistic. The log DDR statistic consistently yields significant results across all considered depth functions, highlighting its robustness and reliability, whereas the DDD statistic shows some instability in terms of power. In addition, permutation testing is found to improve test performance relative to the bootstrap approach. Extensive simulation studies across different depth notions confirm that the log DDR statistic effectively captures distributional discrepancies. The results also indicate that test performance depends critically on the choice of depth function. Using alternative depth notions, rather than half-space depth, substantially reduces computational time and improves computational efficiency. Based on these findings, we propose a two-step procedure. First, the DD plot is used as a visualization tool, together with a numerical summary such as the total distance of points from the diagonal, to guide the selection of an appropriate depth function; second, the chosen depth notion is used to construct the test statistic. While the proposed method demonstrates strong empirical performance, it has some limitations. The permutation approach used for inference, while generally more powerful than bootstrap, can be computationally intensive for large sample sizes. In such cases, approximate permutation methods such as using a reduced number of random permutations or employing sequential stopping rules can be adopted to improve computational efficiency without substantial loss of accuracy. Exploring more scalable resampling strategies warrants further investigation [30,31]. Other important directions for future research include establishing the asymptotic null distributions of the log DDR statistic and developing additional tests formulated directly from numerical depth values.

Author Contributions: Conceptualization, A.W., M.L., X.S. and Y.F.; methodology, A.W., M.L., X.S. and Y.F.; software, M.L.; validation, A.W., M.L., X.S. and Y.F.; formal analysis, M.L.; investigation, A.W., M.L., X.S. and Y.F.; data curation, A.W., M.L., X.S. and Y.F.; writing—original draft preparation, M.L.; writing—review and editing, A.W., M.L., X.S. and Y.F.; visualization, M.L.; supervision, A.W., X.S. and Y.F.; project administration, A.W. and Y.F. All authors have read and agreed to the published version of the manuscript.

Funding: Dr. Shi's work was supported by the NSERC Discovery Grant RGPIN 2022-03264, the NSERC Alliance International Catalyst Grant ALLRP 590341-23, and the University of British Columbia Okanagan (UBC-O) Vice Principal Research in collaboration with the UBC-O Irving K. Barber Faculty of Science. Dr. Fu's research was supported by NSERC Discovery Grant RGPIN 2018-05846.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Raw data sets are publicly available and can be accessed via the provided weblinks. Statlog (Heart) data: https://archive.ics.uci.edu/dataset/145/statlog+heart (accessed on 20 August 2025). Breast cancer data: https://archive.ics.uci.edu/dataset/451/breast+cancer+coimbra (accessed on 20 August 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Anderson, T.W. On the distribution of the two-sample Cramér–von Mises criterion. *Ann. Math. Stat.* **1962**, *33*, 1148–1159. [CrossRef]
2. Anderson, T.W.; Darling, D.A. Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Ann. Math. Stat.* **1952**, *23*, 193–212. [CrossRef]
3. Daniel, W.W. *Applied Nonparametric Statistics*; PWS-Kent: Boston, MA, USA, 1990.
4. Chenouri, S.; Small, C.G. A nonparametric multivariate multisample test based on data depth. *Electron. J. Stat.* **2012**, *6*, 760–782. [CrossRef]
5. Niyogi, P.; Dhar, S. Inspecting discrepancy between multivariate distributions using half-space depth–based information criteria. *arXiv* **2024**, arXiv:2301.01345.
6. Liu, R.Y.; Parelius, J.M.; Singh, K. Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Ann. Stat.* **1999**, *27*, 783–858. [CrossRef]
7. Tukey, J.W. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*; Canadian Mathematical Congress: Vancouver, BC, Canada, 1975; Volume 2, pp. 523–531.
8. Liu, R.Y.; Singh, K. A quality index based on data depth and multivariate rank tests. *J. Am. Stat. Assoc.* **1993**, *88*, 252–260. [CrossRef]
9. Brown, B.M. Statistical uses of the spatial median. *J. R. Stat. Soc. Ser. B* **1983**, *45*, 25–30. [CrossRef]
10. Pokotylo, O.; Mozharovskyi, P.; Dyckerhoff, R. ddalpha: Depth-Based Classification and Calculation of Data Depth; R Package. 2023. Available online: https://CRAN.R-project.org/package=ddalpha (accessed on 20 August 2025).
11. Chaudhuri, P. On a geometric notion of quantiles for multivariate data. *J. Am. Stat. Assoc.* **1996**, *91*, 862–872. [CrossRef]
12. Vardi, Y.; Zhang, C.-H. The multivariate L1-median and associated data depth. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 1423–1426. [CrossRef]
13. Zuo, Y.; Serfling, R. General notions of statistical depth function. *Ann. Stat.* **2000**, *28*, 461–482. [CrossRef]
14. Liu, R.Y. Data depth and multivariate rank tests. In *L1-Statistical Analysis and Related Methods*; North-Holland: Amsterdam, The Netherlands, 1992; pp. 279–294.
15. Gower, J.C. Algorithm AS 78: The Mediancentre. *Appl. Stat.* **1974**, *23*, 466–470. [CrossRef]
16. Nieto-Reyes, A.; Battey, H. A topologically valid definition of depth for functional data. *Stat. Sci.* **2016**, *31*, 61–79. [CrossRef]
17. Liu, R.Y.; Singh, K. Rank tests for multivariate scale difference based on data depth. *DIMACS Ser. Discret. Math. Theor. Comput. Sci.* **2006**, *72*, 17–35.
18. Zuo, Y.; He, X. On the limiting distributions of multivariate depth-based rank-sum statistics and related tests. *Ann. Stat.* **2006**, *34*, 2879–2896. [CrossRef]
19. Shi, X.; Zhang, Y.; Fu, Y. Two-sample tests based on data depth. *Entropy* **2023**, *25*, 238. [CrossRef]
20. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1993. [CrossRef]

21. Lehmann, E.L.; Romano, J.P.; Casella, G. *Testing Statistical Hypotheses*, 3rd ed.; Springer: New York, NY, USA, 2005.
22. Westfall, P.H.; Young, S.S. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*; Wiley: Hoboken, NJ, USA, 1993.
23. Pitman, E.J.G. Significance tests which may be applied to samples from any populations. *Suppl. J. R. Stat. Soc.* **1937**, *4*, 119–130. [CrossRef]
24. Anderson, M.; Ter Braak, C. Permutation tests for multi-factorial analysis of variance. *J. Stat. Comput. Simul.* **2003**, *73*, 85–113. [CrossRef]
25. Bonnini, S.; Assegie, G.M.; Trzcinska, K. Review about the permutation approach in hypothesis testing. *Mathematics* **2024**, *12*, 2617. [CrossRef]
26. Dufour, J.M.; Flachaire, E.; Khalaf, L. Permutation tests for comparing inequality measures. *J. Bus. Econ. Stat.* **2018**, *37*, 457–470. [CrossRef]
27. Lange, K.; Little, R.J.A.; Taylor, J.M.G. Robust statistical modeling using the *t* distribution. *J. Am. Stat. Assoc.* **1989**, *84*, 881–896.
28. Statlog (Heart) [Dataset]. *UCI Machine Learning Repository*. Available online: https://archive.ics.uci.edu/dataset/145/statlog+heart (accessed on 20 August 2025).
29. Patrício, M.; Pereira, J.; Crisóstomo, J.; Matafome, P.; Gomes, M.; Seiça, R.; Caramelo, F. Using resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer* **2018**, *18*, 29 . [CrossRef]
30. Hemerik, J.; Goeman, J.J. Exact testing with random permutations. *TEST* **2018**, *27*, 811–825. [CrossRef]
31. Lock, R.H. A sequential approximation to a permutation test. *Commun. Stat. Simul. Comput.* **1991**, *20*, 341–363. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.