axioms

MDPI

Article

A Robust High-Dimensional Test for Two-Sample Comparisons

Hasan Bulut [1,*], Soofia Iftikhar [2], Nosheen Faiz [3] and Olayan Albalawi [4]

[1] Department of Statistics, Ondokuz Mayıs University, Samsun 55200, Turkey
[2] Department of Statistics, Shaheed Benazir Bhutto Women University, Peshawar 25000, Pakistan
[3] Department of Statistics, Abdul Wali Khan University, Mardan 23200, Pakistan
[4] Department of Statistics, University of Tabuk, Tabuk 47713, Saudi Arabia
* Correspondence: hasan.bulut@omu.edu.tr

Abstract: The Hotelling T^2 statistic is used to compare the mean vectors of two independent multivariate Gaussian distributions. Nevertheless, this statistic is highly sensitive to outliers and is not suitable for high-dimensional datasets where the number of variables exceeds the sample size. This study introduces a robust permutation test based on the minimum regularized covariance determinant (MRCD) estimator to address these limitations of the two-sample Hotelling T^2 statistic. Simulation studies were performed to evaluate the proposed test's empirical size, power, and robustness. Additionally, the test was applied to both uncontaminated and contaminated Alzheimer's Disease datasets. The findings from the simulations and real data examples provide clues that the proposed test can be effectively used with high-dimensional data without being impacted by outliers. Finally, an R function within the "MVTests" package was developed to implement the proposed test statistic on real-world data.

Keywords: robust Hotelling test statistic; high-dimensional data; minimum regularized covariance estimators; permutation test; MVTests

MSC: 62G10; 62G35; 62H15; 62P10

check for updates

Citation: Bulut, H.; Iftikhar, S.; Faiz, N.; Albalawi, O. A Robust High-Dimensional Test for Two-Sample Comparisons. *Axioms* **2024**, *13*, 585. https://doi.org/10.3390/axioms13090585

Academic Editor: Hans J. Haubold

Received: 12 July 2024
Revised: 12 August 2024
Accepted: 26 August 2024
Published: 28 August 2024

Copyright: © 2024 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 (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Let $\boldsymbol{X}_1 \sim N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $\boldsymbol{X}_2 \sim N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ be the distribution of two independent groups. We decide to test the hypothesis given in (1) to see whether the mean parameters of these groups, which have multivariate Gaussian distributions, are equal.

$$\begin{aligned} H_0 &: \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \\ H_1 &: \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2 \end{aligned} \tag{1}$$

If the scatter parameters of these multivariate distributions are homogeneous ($\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$) and the common covariance matrix is unknown, two samples Hotelling T^2 statistics which are given in Equation (2) is used to test the hypothesis defined in (1):

$$T^2_{Hotelling} = \left(\frac{n_1 n_2}{n_1 + n_2}\right)(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2)^T \mathbf{S}^{-1}(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2) \sim T^2_{n_1+n_2-2} \tag{2}$$

where n_1 and n_2 are sample sizes of the first and second groups; moreover $\overline{\boldsymbol{X}}_1$ and $\overline{\boldsymbol{X}}_2$ are the sample mean vectors of the first and second groups, respectively. Also $(.)^T$ is the transpose operator. $\boldsymbol{S}$ is the pooled covariance matrix and is calculated as below:

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2}((n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2) \tag{3}$$

where $\mathbf{S}_1$ is the sample covariance matrix of the first group and $\mathbf{S}_2$ is the sample covariance matrix of the second group. Let T_C^2 be the calculated value of the $T_{Hotelling}^2$ statistic given in Equation (2). When $T_C^2 > T_{n_1+n_2-2;\alpha}^2$, we reject the null hypothesis [1,2].

The Hotelling T^2 statistic is based on the classical mean vectors and covariance matrices. For this reason, it is heavily sensitive to outliers in the data. The term sensitive means that the estimators are affected by outliers in the data, which means that outliers in the data will cause their estimations to fail. The common solution to this problem is to use robust mean vectors and covariance matrices in the literature. Willems et al. [3] used the minimum covariance determinant (MCD) estimators, Çetin and Aktaş [4] used the minimum volume ellipsoid (MVE) estimators, Mokhtar et al. [5] used the M-estimators for a one-sample Hotelling T^2 statistic instead of classical ones. Similarly, Todorov and Filzmoser [6] used the MCD estimators for one-way MANOVA.

Another important problem is that the $T_{Hotelling}^2$ statistic given in Equation (2) can only be used for low-dimensional datasets, which are $p < n_1 + n_2 - 2$. Otherwise, the matrix $\mathbf{S}$ is singular and the precision matrix $\left(\mathbf{S}^{-1}\right)$ cannot be obtained. Therefore, the $T_{Hotelling}^2$ cannot be calculated. There are many test statistics in the literature to overcome this problem.

Bai and Saranadasa [7] proposed the statistic in Equation (4). Because there is no precision matrix in the test statistics given in Equation (4), this statistic can be used in high-dimensional data. We can perform this statistic on any real data via the *apval_Bai1996()* function in the R package called "highmean" [8].

$$T_{BS}^2 = \frac{\left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2\right)^T\left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2\right) - tr(\boldsymbol{S})}{\sqrt{\frac{2n(n+1)}{(n-1)(n+2)}\left\{tr\left(\boldsymbol{S}^2\right) - \frac{(tr(\boldsymbol{S}))^2}{n}\right\}}} \tag{4}$$

where $tr(.)$ is a trace operator.

Srivastava and Du [9] suggested using a diagonal matrix instead of a covariance matrix. The diagonal elements of this diagonal matrix are variances in matrix $\boldsymbol{S}$. Therefore, Srivastava and Du [9] proposed a test statistic that can be used in high-dimensional data by removing the covariance information. This statistic is given in Equation (5). The p-value for this statistic can be obtained by using the *apval_Sri2008()* function in the R package called "highmean" [8].

$$T_{SD}^2 = \frac{\left(\frac{n_1 n_2}{n_1+n_2}\right)\left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2\right)^T\boldsymbol{D}_S^{-1}\left(\overline{\boldsymbol{X}}_1 - \overline{\boldsymbol{X}}_2\right) - \frac{np}{n-2}}{\sqrt{2\left(tr\left(\boldsymbol{R}^2\right) - p^2 n^{-1}\right)c_{p,n}}} \tag{5}$$

where $\boldsymbol{D}_S = diag\left(s_{11}, s_{22}, \ldots, s_{pp}\right)$ is a diagonal matrix, and s_{ii} is the i_{th} diagonal element of matrix $\boldsymbol{S}$. Moreover, $\boldsymbol{R} = \boldsymbol{D}_S^{-1/2}\boldsymbol{S}\boldsymbol{D}_S^{-1/2}$ is the correlation matrix, and $c_{p,n} = 1 + \frac{tr\left(\boldsymbol{R}^2\right)}{\sqrt{p^3}}$.

Chen and Qin [10] suggested the test statistic given in Equation (6). The p-value for the T_{CQ}^2 statistic can be obtained by using the *apval_Chen2010()* function in the R package called "highmean" [8].

$$T_{CQ}^2 = \frac{\sum_{i\neq j}^{n_1} X_{1i}^T X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i\neq j}^{n_2} X_{2i}^T X_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{1i}^T X_{2j}}{n_1 n_2} \tag{6}$$

These test statistics lose power when the $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ vector contains many zero elements [11,12]. For this reason, Cai, Liu and Xia [11] proposed a supremum test statistic. The p-value for this supremum statistic can be calculated by using the *apval_Cai2014()* function in the R package called "highmean" [8].

$$T_{CLX}^2 = \left(\frac{n_1 n_2}{n_1+n_2}\right) max_{\,1\leq i\leq p}\left(\overline{\boldsymbol{X}}_1^{(i)} - \overline{\boldsymbol{X}}_2^{(i)}\right)^2 / s_{ii} \tag{7}$$

The test statistics (T^2_{BS}, T^2_{SD}, T^2_{CQ}, T^2_{CLX}) given in Equations (4)–(7) can be used to test the hypothesis in (2) with high-dimensional data. However, these test statistics are still sensitive to outliers because they are based on classical estimators.

As mentioned before, Willems, Pison, Rousseeuw and Van Aelst [3] and Todorov and Filzmoser [6] used the MCD estimators to obtain robust tests. However, it is well known that the MCD estimators can only be used for low-dimensional data. To overcome this lack of MCD estimators, Boudt et al. [13] developed the minimum regularized covariance determinant (MRCD) estimators. MRCD estimators can be obtained robustly in high-dimensional data, unlike MCD ones. Detailed information about MRCD estimators is given in Section 2. Bulut [14] used the MRCD estimators to propose a robust Hoteling T^2 statistic for one sample.

The main purpose of this study is to propose a test that can be used in high-dimensional data and is not sensitive to outliers in the data. For this aim, we extend the approach developed by Bulut (2023b) to two-sample cases. Moreover, we aim to use our test procedure on real data by generating R functions in the "MVTests" package [15].

The remainder of the study is designed as follows. In Section 2, the MRCD estimators are introduced. Section 3 defines the proposed robust Hotelling T^2 statistic and test procedure. In Section 4, a robust version of T^2_{CLX} proposed by Cai, Liu and Xia [11] is introduced. We perform a simulation study to compare the performance of our test procedure and alternative tests in Section 5. In Section 6, we perform tests on real example data to compare the performance of the tests. We introduce the R functions, which can be used to perform our test statistic in Section 7. Finally, we provide conclusions in Section 8.

2. Minimum Regularized Covariance Determinant (MRCD) Estimators

Willems, Pison, Rousseeuw and Van Aelst [3] and Todorov and Filzmoser [6] used the MCD estimators to obtain robust test statistics. However, these estimators cannot be used when $n < p$.

Boudt, Rousseeuw, Vanduffel and Verdonck [13] proposed Minimum Regularized Covariance Determinant (MRCD) estimators of location and scatter parameters without affecting outliers in high-dimensional data. The MRCD estimator has good breakdown point properties of the MCD estimator [13,14].

To calculate MRCD estimations, firstly, we standardize data by using the median and Q_n as the univariate location and scatter estimators [16]; then, we use the $\boldsymbol{T}$ target matrix. This matrix $\boldsymbol{T}$ is symmetric and positive definite. The regularized covariance matrix of any subset H, which is obtained from standardized $\boldsymbol{Z}$ data, is calculated as below:

$$\boldsymbol{K}(H) = \rho\boldsymbol{T} + (1-\rho)c_\alpha \boldsymbol{S}_Z(H) \tag{8}$$

where ρ is the regularization parameter, c_α is the consistency factor defined by Croux and Haesbroeck [17], and

$$\boldsymbol{S}_Z(H) = \frac{1}{h-1}(\boldsymbol{Z}_H - \boldsymbol{\mu}_Z(H))^T(\boldsymbol{Z}_H - \boldsymbol{\mu}_Z(H)), \boldsymbol{\mu}_Z(H) = \frac{1}{h}\boldsymbol{Z}_H^T\boldsymbol{1}_h \tag{9}$$

MRCD estimations are obtained from the subset H_{MRCD} which is obtained by solving the minimization problem given in Equation (10).

$$H_{MRCD} = \underset{H\in\mathcal{H}}{\operatorname{argmin}}\left[det(\boldsymbol{K}(H))^{1/p}\right] \tag{10}$$

where $\mathcal{H}$ is the set, which consists of all subsets with size h in data. Finally, the MRCD location and scatter estimators are obtained as given in Equations (11) and (12).

$$\hat{\boldsymbol{\mu}}_{MRCD} = \boldsymbol{V}_X + \boldsymbol{D}_X\,\boldsymbol{\mu}_Z(H_{MRCD}\,) \tag{11}$$

$$\hat{\boldsymbol{\Sigma}}_{MRCD} = \boldsymbol{D}_X\,\boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}[\rho\boldsymbol{I} + (1-\rho)c_\alpha\boldsymbol{S}_w(H_{MRCD})]\boldsymbol{\Lambda}^{1/2}\boldsymbol{Q}'\boldsymbol{D}_X \tag{12}$$

where $\mathbf{\Lambda}$ and $\boldsymbol{Q}$ are eigenvalues and eigenvectors matrices of $\boldsymbol{T}$, respectively. Also, $\boldsymbol{S}_w(H_{MRCD})$ is calculated as below:

$$\boldsymbol{S}_w(H_{MRCD}) = \mathbf{\Lambda}^{-1/2}\boldsymbol{Q}'\boldsymbol{S}_Z(H_{MRCD})\boldsymbol{Q}\mathbf{\Lambda}^{-1/2} \tag{13}$$

More detailed information about MRCD estimators is available in [13]. In this study, we used the "rrcov" package in the R software for calculations regarding the MRCD estimators [18]. While we use the package, we assume that we know the outlier rate of the data. Moreover, we prefer to use the default values of the regularization parameter (rho) and target matrix. The function automatically calculates these values from the dataset.

3. Proposed Hotelling T^2 Test

The Hotelling T^2 statistic, which is used to compare mean vectors of two independent groups, is given in Equation (2). As mentioned before, however, this statistic cannot be used in high-dimensional data and is sensitive to outliers in data. In this study, we propose a test statistic that can be used in high-dimensional data without being affected by outliers in data. This statistic is given in Equation (14).

$$T^2_{MRCD} = \left(\frac{n_1 n_2}{n_1 + n_2}\right)(\boldsymbol{M}_1 - \boldsymbol{M}_2)^T\mathbf{C}^{-1}(\boldsymbol{M}_1 - \boldsymbol{M}_2) \tag{14}$$

where n_1 and n_2 are the sample size of the first and second groups, respectively. $\boldsymbol{M}_1$ and $\mathbf{M}_2$ are MRCD estimations of location parameters of the first and second groups, respectively. The matrix C is a robust pooled covariance matrix and it is calculated as in Equation (15):

$$\mathbf{C} = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)\boldsymbol{S}_1^{\boldsymbol{MRCD}} + (n_2 - 1)\boldsymbol{S}_2^{\boldsymbol{MRCD}}\right) \tag{15}$$

where $\boldsymbol{S}_1^{\boldsymbol{MRCD}}$ and $\boldsymbol{S}_2^{\boldsymbol{MRCD}}$ are the MRCD covariance matrices of the first and second samples, respectively.

Because the finite-sample distribution of MRCD estimators is unknown, Bulut [14] used an approach distribution for a one-sample case. This approach is used by Willems, Pison, Rousseeuw and Van Aelst [3] and Todorov and Filzmoser [6] in low-dimensional data. However, we prefer a permutation test instead of any asymptotic distribution because the asymptotic distribution approach needs too much calculation time. As a result, we propose the robust permutation test to compare the mean vectors of two independent groups in high-dimensional data as below:

1. Calculate the MRCD estimations of the first and second groups.
2. Calculate the T^2_{MRCD} value based on Equation (14). Let $T^2_{MRCD\ (c)}$ be this calculated value.
3. Compound all observations as one sample with size $n = n_1 + n_2$.
4. Under the null hypothesis, randomly separate observations such that n_1 observations are in the first group and n_2 observations are in the second group.
5. Calculate the $T^2_{MRCD(i)}$ $(i = 1, 2, \ldots, N)$ value based on Equation (14) for the generated synthetic groups.
6. Repeat steps (iii–v) N times. Here, N is the permutation number.
7. Calculate the p-value as given in Equation (16):

$$p\ value = \frac{\#\left(T^2_{MRCD\ (c)} > T^2_{MRCD(i)}\right)}{N} \tag{16}$$

According to the test algorithm, we can calculate the p-value directly without any distribution assumption. Moreover, we can calculate this p-value in high-dimensional data. We reject the null hypothesis when the p-value is less than the significance level. This test

procedure can be performed on any real example data with the R function introduced in Section 7.

Like other permutation tests, the performance of the proposed test statistics is based on the permutation number, and once the permutation number increases, the results will be more stable.

4. Robust CLX Test

Wang et al. [19] proposed a robust alternative to the test procedure suggested by Cai, Liu and Xia [11]. We show T^2_{RCLX} this test statistic for consistence. T^2_{RCLX} is defined in Equation (17):

$$T^2_{RCLX} = \left(\frac{n_1 n_2}{n_1 + n_2}\right) max\left\{\frac{\left(\overline{Z}_{\eta_1,1}\right)^2}{w_{11}}, \ldots, \frac{\left(\overline{Z}_{\eta_p,p}\right)^2}{w_{pp}}\right\} \tag{17}$$

where $\overline{\boldsymbol{Z}}_\eta = \left(\overline{Z}_{\eta_1,1}, \ldots, \overline{Z}_{\eta_p,p}\right)^T = \boldsymbol{\Omega_\eta}\left(\overline{\boldsymbol{X}}_\eta - \overline{\boldsymbol{Y}}_\eta\right)$, and $\boldsymbol{\Omega_\eta} = \left(w_{ij}\right)_{p\times p} = \boldsymbol{\Sigma}_\eta^{-1}$ is the common precision matrix for trimmed mean vectors $\sqrt{n_1}\,\overline{\boldsymbol{X}}_\eta$ and $\sqrt{n_2}\,\overline{\boldsymbol{Y}}_\eta$. $\overline{\boldsymbol{X}}_\eta$ and $\overline{\boldsymbol{Y}}_\eta$ are the trimmed mean vectors for the first and second groups, respectively.

Let a random sample be $X_1, \ldots, X_n$ and $X_{(i)}$ is the i_{th} order statistic of this sample. For the trimming level $\eta \in (0, 0.5)$, we can calculate the trimmed mean of this sample given by:

$$\overline{X}_\eta = \frac{1}{n - 2n\eta} \sum_{i=n\eta+1}^{n-n\eta} X_{(i)} \tag{18}$$

Therefore, the trimmed mean vector can be obtained by calculating all the trimmed means of the p-variables for each group. According to this, we can define the trimmed mean vectors for the first and second groups as $\overline{\boldsymbol{X}}_{\boldsymbol{\eta}} = \left(\overline{\boldsymbol{X}}_{\boldsymbol{\eta_1,1}}, \ldots, \overline{\boldsymbol{X}}_{\boldsymbol{\eta_p,p}}\right)^{\boldsymbol{T}}$ and $\overline{\boldsymbol{Y}}_{\boldsymbol{\eta}} = \left(\overline{\boldsymbol{Y}}_{\boldsymbol{\eta_1,1}}, \ldots, \overline{\boldsymbol{Y}}_{\boldsymbol{\eta_p,p}}\right)^{\boldsymbol{T}}$, respectively.

As a result, we can reject the null hypothesis when $\boldsymbol{T^2_{RCLX}} > \mathbf{2ln}(\boldsymbol{p}) - \boldsymbol{ln}[\mathbf{ln}(\boldsymbol{p})] + \boldsymbol{q_\alpha}$. Here, $\boldsymbol{q_\alpha}$ is the $(\mathbf{1} - \boldsymbol{\alpha})$th quantile of the Gumbel distribution. More detailed information about the $\boldsymbol{T^2_{RCLX}}$ test statistic is available from Wang, Lin and Tang [19].

5. Simulation Study

In this section, we design a simulation study to compare $\boldsymbol{T^2_{BS}}$, $\boldsymbol{T^2_{SD}}$, $\boldsymbol{T^2_{CQ}}$, $\boldsymbol{T^2_{CLX}}$, $\boldsymbol{T^2_{RCLX}}$, and $\boldsymbol{T^2_{MRCD}}$, which is our proposed test regarding the empirical size, power and robustness properties. We test the null hypothesis $\boldsymbol{H_0} : \boldsymbol{\mu_1} = \boldsymbol{\mu_2}$ in all simulation designs, and we take $\boldsymbol{\mu_2} = \mathbf{0}$ in all cases. Here, $\mathbf{0}$ is a zero vector.

We randomly sample observations from the multivariate Gaussian distributions $\boldsymbol{N_p}(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$ for the first group and $\boldsymbol{N_p}(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$ for the second group. The common and homogeneity covariance matrix of these distributions is generated based on three different models used by Cai, Liu and Xia [11] as below:

- **Model 1:** $\boldsymbol{\Sigma} = \sigma_{i,j}\ (i, j = 1, 2, \ldots p)$, where

$$\sigma_{i,j} = \begin{cases} 1 & , \quad i = j \\ 0.8 & , \quad 2(k-1)+1 \le i \ne j \le 2k, \ where\ k = 1, \ldots, [p/2] \\ 0 & , \quad otherwise \end{cases}$$

- **Model 2:** $= \sigma_{i,j}\ (i, j = 1, 2, \ldots p)$, where $\sigma_{i,j} = 0.6^{|i-j|}$ for $1 \le i, j \le p$.
- **Model 3:** $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$, and $\boldsymbol{\Omega} = w_{i,j}(i, j = 1, 2, \ldots p)$, where $w_{i,i} = 2\ (i = 1, 2, \ldots, p)$, $w_{i,i+1} = 0.8\ (i = 1, 2, \ldots, p-1)$, $w_{i,i+2} = 0.4\ (i = 1, 2, \ldots, p-2)$, $w_{i,i+3} = 0.4\ (i = 1, 2, \ldots, p-3)$, $w_{i,i+4} = 0.2\ (i = 1, 2, \ldots, p-4)$, $w_{i,j} = 0$ otherwise.

For all the simulation studies, we use sample size values $n_1 = n_2 = n = 10, 20, 30$, and variable numbers $p = 25, 50, 75$. Therefore, we study high-dimensional data in all cases.

5.1. Empirical Size (Type-1) of Tests

In this subsection, we perform a simulation study to compare the empirical size performance of T^2_{MRCD}, T^2_{RCLX}, T^2_{CLX}, T^2_{BS}, T^2_{SD}, and T^2_{CQ}. We generate $m = 3000$ random samples from the multivariate Gaussian distributions $N_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu_2}, \boldsymbol{\Sigma})$, respectively. We set $\boldsymbol{\mu_1} = \boldsymbol{\mu_2} = \mathbf{0}$ such that the null hypothesis is true in all cases. We obtain $\boldsymbol{\Sigma}$ based on Models 1–3. We test the null hypothesis $(H_0 : \boldsymbol{\mu_1} = \boldsymbol{\mu_2})$ with T^2_{MRCD}, T^2_{RCLX}, T^2_{CLX}, T^2_{BS}, T^2_{SD}, and T^2_{CQ} for each sample, and we calculate the empirical size of all the tests as the ratio of rejecting the true null hypothesis. We reject the null hypothesis when $p < 0.05$. The results are given in Table 1.

Table 1. Empirical sizes of test statistics.

Models	n	p	T^2_{MRCD}	T^2_{RCLX}	T^2_{CLX}	T^2_{BD}	T^2_{SD}	T^2_{CQ}
Model-1	10	25	4.20	1.40	10.43	6.73	6.30	6.73
	20	50	4.40	0.90	7.33	6.03	5.30	6.03
	30	75	5.67	0.33	6.43	5.33	4.37	5.33
Model-2	10	25	4.90	4.20	10.90	6.63	6.60	6.63
	20	50	5.90	1.70	7.53	6.37	4.97	6.37
	30	75	4.00	1.33	6.53	6.63	5.17	6.63
Model-3	10	25	5.20	3.00	11.50	7.13	6.73	7.13
	20	50	5.60	4.40	8.26	6.63	5.60	6.63
	30	75	5.67	4.20	7.23	5.90	5.40	5.90

Table 1 compares the empirical sizes of the test statistics. We can see that the empirical sizes of T^2_{MRCD} are reasonably close to the nominal level of 5%. As the sample size increases, the empirical sizes of T^2_{CLX} become close to the nominal value, while the empirical sizes of T^2_{RCLX} go away from the nominal value. We can see in Table 1 that the T^2_{BS}, T^2_{SD}, and T^2_{CQ} statistics have similar empirical size values.

5.2. Power of Tests

We perform a simulation study to compare the powers of T^2_{MRCD}, T^2_{RCLX}, T^2_{CLX}, T^2_{BS}, T^2_{SD}, and T^2_{CQ} in this subsection. We generate $m = 3000$ random samples from the multivariate Gaussian distributions $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, respectively. We set $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ such that the null hypothesis is false in all cases. In all cases, $\boldsymbol{\mu}_2 = 0$. We define $\boldsymbol{\mu}_1$ with two different approaches as below:

- **Fixed magnitude:** $\boldsymbol{\mu}_1 = \mu_{1j}$ $(j = 1, 2, \ldots, p)$, $\mu_{1j} = \mp\sqrt{2\log(p)/n}$ for $j = 1, \ldots, m$ with the equal probability, and $\mu_{1j} = 0$ $(j = m+1, \ldots, p)$.
- **Varied magnitude:** $\boldsymbol{\mu}_1 = \mu_{1j}$ $(j = 1, 2, \ldots, p)$, μ_{1j} is generated from $Uniform\left(-\sqrt{\frac{4\log(p)}{n}}, \sqrt{\frac{4\log(p)}{n}}\right)$ for $j = 1, \ldots, m$, and $\mu_{1j} = 0$ $(j = m+1, \ldots, p)$.

Here, we use two different m values of $m = \lceil 0.05\, p \rceil$ and $m = \lceil \sqrt{p} \rceil$. Also, $\lceil x \rceil$ denotes the minimum integer which is greater than x. We obtain $\boldsymbol{\Sigma}$ based on Models 1–3. We test the null hypothesis $(H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2)$ with T^2_{MRCD}, T^2_{RCLX}, T^2_{CLX}, T^2_{BS}, T^2_{SD}, and T^2_{CQ} for each sample. Then, we calculate the power of all the tests as the ratio of rejecting the false null hypothesis. We reject the null hypothesis when $p < 0.05$. The results are given in Table 2. According to Table 2, the power loss of the T^2_{MRCD} is acceptable because the datasets generated are not contaminated.

Table 2. Power of test statistics.

Models	n	p	Signal	m	T^2_{MRCD}	T^2_{RCLX}	T^2_{CLX}	T^2_{BD}	T^2_{SD}	T^2_{CQ}
Model-1	10	25	fixed	$\lceil 0.05\,p \rceil$	18.20	18.60	29.00	19.33	18.17	19.33
				$\lceil \sqrt{p} \rceil$	50.00	34.30	48.10	49.17	44.30	49.17
			varied	$\lceil 0.05\,p \rceil$	13.60	13.00	23.40	16.00	14.93	16.00
				$\lceil \sqrt{p} \rceil$	31.60	20.40	42.43	33.47	31.13	33.47
	20	50	fixed	$\lceil 0.05\,p \rceil$	25.40	37.20	30.73	24.27	20.83	24.27
				$\lceil \sqrt{p} \rceil$	80.40	75.40	57.20	69.63	63.93	69.63
			varied	$\lceil 0.05\,p \rceil$	21.40	23.80	26.80	17.57	15.23	17.57
				$\lceil \sqrt{p} \rceil$	53.20	52.60	49.80	45.43	40.10	45.43
	30	75	fixed	$\lceil 0.05\,p \rceil$	41.40	64.60	34.40	29.63	24.80	29.63
				$\lceil \sqrt{p} \rceil$	83.80	89.80	59.20	71.67	67.03	71.67
			varied	$\lceil 0.05\,p \rceil$	27.40	38.80	29.63	21.10	17.60	21.10
				$\lceil \sqrt{p} \rceil$	54.60	62.40	50.40	46.90	41.57	46.90
Model-2	10	25	fixed	$\lceil 0.05\,p \rceil$	17.70	15.40	28.20	18.07	16.03	18.07
				$\lceil \sqrt{p} \rceil$	51.50	34.20	49.70	43.77	39.30	43.77
			varied	$\lceil 0.05\,p \rceil$	12.60	12.00	23.60	14.27	13.30	14.27
				$\lceil \sqrt{p} \rceil$	34.60	30.60	43.60	27.40	24.83	27.40
	20	50	fixed	$\lceil 0.05\,p \rceil$	29.40	27.40	30.60	19.80	16.40	19.80
				$\lceil \sqrt{p} \rceil$	83.00	56.60	58.00	61.47	54.33	61.47
			varied	$\lceil 0.05\,p \rceil$	19.00	19.20	28.00	16.17	13.50	16.17
				$\lceil \sqrt{p} \rceil$	56.80	42.80	51.60	40.07	34.73	40.07
	30	75	fixed	$\lceil 0.05\,p \rceil$	40.00	41.80	36.57	24.77	20.80	24.77
				$\lceil \sqrt{p} \rceil$	84.80	73.00	58.00	64.57	58.07	64.57
			varied	$\lceil 0.05\,p \rceil$	26.40	29.60	30.80	16.77	13.47	16.77
				$\lceil \sqrt{p} \rceil$	57.00	57.00	52.40	40.10	33.90	40.10
Model-3	10	25	fixed	$\lceil 0.05\,p \rceil$	25.90	46.20	44.40	31.83	31.83	31.83
				$\lceil \sqrt{p} \rceil$	58.40	47.30	69.07	77.20	74.60	77.20
			varied	$\lceil 0.05\,p \rceil$	16.60	20.80	36.40	22.33	22.87	22.33
				$\lceil \sqrt{p} \rceil$	41.40	38.80	56.80	52.13	50.97	52.13
	20	50	fixed	$\lceil 0.05\,p \rceil$	36.80	40.60	53.60	41.53	41.00	41.53
				$\lceil \sqrt{p} \rceil$	83.00	72.40	81.80	94.23	92.60	94.23
			varied	$\lceil 0.05\,p \rceil$	22.60	32.40	43.33	28.43	27.87	28.43
				$\lceil \sqrt{p} \rceil$	56.20	56.60	69.20	70.90	68.90	70.90
	30	75	fixed	$\lceil 0.05\,p \rceil$	41.40	55.00	62.60	51.80	49.60	51.80
				$\lceil \sqrt{p} \rceil$	86.80	82.20	84.43	95.33	94.23	95.33
			varied	$\lceil 0.05\,p \rceil$	28.60	45.80	50.40	32.10	31.10	32.10
				$\lceil \sqrt{p} \rceil$	59.00	68.80	73.43	73.33	71.03	73.33

5.3. Robustness of Tests

In this subsection, we perform a simulation study to investigate the robustness performance of T^2_{MRCD}, T^2_{RCLX}, T^2_{CLX}, T^2_{BS}, T^2_{SD}, and T^2_{CQ}. We generate 90% of random samples from the multivariate Gaussian distributions $N_p(0, \mathbf{\Sigma})$ and $N_p(0, \mathbf{\Sigma})$, respectively. Moreover, we generate the remaining 10% of samples from the multivariate Gaussian distributions $N_p(\boldsymbol{\mu}_{1.out}, \mathbf{\Sigma})$ and $N_p(\boldsymbol{\mu}_{2.out}, \mathbf{\Sigma})$, respectively. Here, $\boldsymbol{\mu}_{1.out} = \left[\sqrt{p}, 0, \ldots, 0\right]^T$ and $\boldsymbol{\mu}_{2.out} = \left[-\sqrt{p}, 0, \ldots, 0\right]^T$. We obtain $\mathbf{\Sigma}$ based on Models 1–3. Therefore, we contaminate samples with a 10% outlier ratio.

We draw a plot to show the effect of outliers in Figure 1. We generate 27 random observations from the first population $N_p(0, \mathbf{\Sigma})$ and 27 observations from the second population $N_p(0, \mathbf{\Sigma})$ such that $p = 25$. Moreover, we generate three random observations for the first

group from the population $N_p(\boldsymbol{\mu}_{1.out}, \boldsymbol{\Sigma})$, and three observations for the second group from the population $N_p(\boldsymbol{\mu}_{2.out}, \boldsymbol{\Sigma})$. $\boldsymbol{\Sigma}$ is generated based on Model-1 in each case in Figure 1. Also, $\boldsymbol{\mu}_{1.out} = [5, 0, \ldots, 0]^T$ and $\boldsymbol{\mu}_{2.out} = [-5, 0, \ldots, 0]^T$. We used the multidimensional scaling method to reduce the dimension from 25 to 2.



Figure 1. Scatter plots for instance outliers problem in high-dimensional data.

According to Figure 1, we can see that the outliers in the two groups move the equal mean vectors away from each other. For this reason, a non-robust test may reject the null hypothesis, which is true in real data, while a robust test must fail to reject the null hypothesis without being affected by outliers. In this section, we compare the robustness performance of each test.

We test the null hypothesis $(H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2)$ with T^2_{MRCD}, T^2_{RCLX}, T^2_{CLX}, T^2_{BS}, T^2_{SD}, and T^2_{CQ} for each sample, and we reject the null hypothesis when $p < 0.05$. Before the datasets are contaminated, the null hypothesis is true. If any test statistic is robust to outliers in data, it must reject the null hypothesis with a percentage close to the nominal significance level (5%). For this reason, we calculate the robustness performance as the ratio of rejecting the true null hypothesis. The results are given in Table 3.

Table 3. Robustness of test statistics.

Models	n	p	T^2_{MRCD}	T^2_{RCLX}	T^2_{CLX}	T^2_{BD}	T^2_{SD}	T^2_{CQ}
Model-1	10	25	5.60	1.00	10.47	9.73	7.57	9.73
	20	50	4.67	1.67	7.67	26.37	8.13	26.37
	30	75	6.33	0.00	6.33	59.47	8.40	59.47
Model-2	10	25	4.00	3.00	10.40	9.47	6.97	9.47
	20	50	5.67	1.00	8.33	22.37	7.70	22.37
	30	75	4.33	0.33	6.83	52.90	7.63	52.90
Model-3	10	25	5.20	2.20	12.40	9.93	7.57	9.93
	20	50	6.00	2.33	11.67	37.00	6.93	37.00
	30	75	6.33	4.33	7.07	84.13	8.57	84.13

According to Table 3, it is abundantly clear that the T^2_{MRCD} statistic is robust to outliers, while other statistics are heavily sensitive to outliers. The T^2_{RCLX} statistic has similar performance with contaminated data and uncontaminated data but this performance is unacceptable because the empirical size values are lower than the nominal value of 5%.

6. Real Example

Alzheimer's disease (AD) is a cognitive impairment disorder marked by memory loss and a decline in functional abilities beyond what is typical for a given age, making it the most prevalent cause of dementia in the elderly. Biologically, Alzheimer's disease is linked to amyloid-β (Aβ) brain plaques and brain tangles associated with a form of the Tau protein [20].

While medical imaging can be useful in predicting the onset of the disease, there is also a growing interest in potential cost-effective fluid biomarkers that can be extracted from plasma or cerebrospinal fluid (CSF). Currently, there are several acknowledged non-imaging biomarkers, including protein levels of specific forms of the Aβ and Tau proteins and the Apolipoprotein E genotype. The Apolipoprotein E genotype has three primary variants: E2, E3, and E4, with the E4 allele most closely associated with AD [21,22].

In a clinical study conducted by Craig-Schapiro et al. [23] involving 333 patients, including those with mild but well-characterized cognitive impairment and healthy individuals, CSF samples were collected from all subjects. The study aimed to determine if individuals in the early stages of impairment could be distinguished from cognitively healthy individuals. They included in the analysis demographic attributes like age and gender, the Apolipoprotein E genotype, protein measurements of Aβ, Tau, and a phosphorylated version of Tau (referred to as pTau), protein measurements of 124 exploratory biomarkers, and clinical dementia scores [20]. This data is available with the name *AlzheimerDisease* in the "AppliedPredictiveModeling" package in R software [24].

In this subsection, we use only 124 protein measurements as biomarkers of AD. Therefore, we have $p = 124$ variables in our data. Moreover, we use a subset that contains 106 high-risk patients with Apolipoprotein E genotype E3E4, as Wang, Lin and Tang [19] do. $n_1 = 41$ subjects of them are "impaired", and $n_2 = 65$ subjects of them are "healthy". Then, we standardized all variables. As a result, we have high-dimensional data, and we provide a correlation plot of this data in Figure 2.

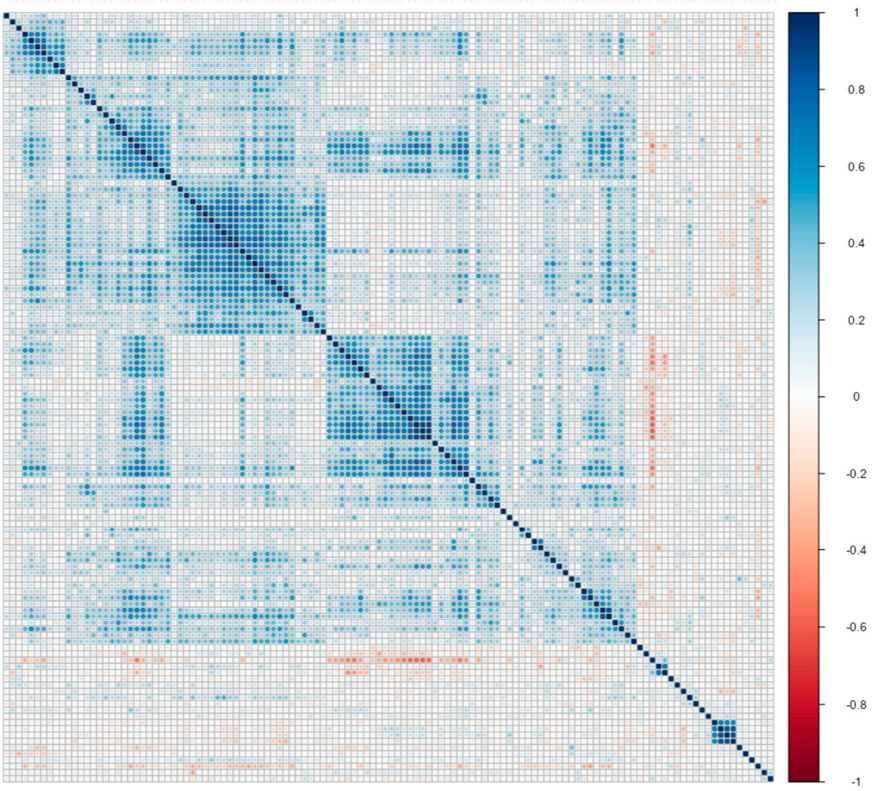


Figure 2. Correlation plot matrix (124×124). Columns and rows were sorted with clustering methods.

We aim to compare the mean vectors of two groups ("impaired" vs. "healthy"). For this purpose, firstly, we assume that our data does not contain outliers, and so we test the null hypothesis $\left(\boldsymbol{\mu}_{imparied} = \boldsymbol{\mu}_{healthy}\right)$ with T^2_{MRCD}, $T^2_{RCL\boldsymbol{X}}$, $T^2_{CL\boldsymbol{X}}$, T^2_{BS}, T^2_{SD}, and T^2_{CQ}. The results of the tests are given in the "Clean Data" part of Table 4. According to this part, all the tests reject the null hypothesis.

Table 4. Test results of Alzheimer's Disease data.

Test	Clean Data		Contaminated Data	
	Test Statistic	***p*-Value**	**Test Statistic**	***p*-Value**
T^2_{MRCD}	251.699	0.000	266.378	0.000
$T^2_{RCL\boldsymbol{X}}$	17.839	0.038	0.280	0.927
$T^2_{CL\boldsymbol{X}}$	26.923	0.000	7.702	0.492
T^2_{BS}	8.852	0.000	−0.612	0.730
T^2_{SD}	6.676	0.000	−0.262	0.603
T^2_{CQ}	8.864	0.000	−0.552	0.710

After analysis of the clean data, we contaminate 10% of the data such that the difference between the mean vectors decreases. For this aim, we determine the last four observations in Impaired group-1 and the last six observations in Healthy group-2 such that the mean values of all the variables in group-1 are equal to 2, and the mean values of all the variables in group-2 are equal to 1.8. Therefore, the difference between the mean vectors of the groups is 0.2 for all variables. Despite the mean vectors of uncontaminated data in each group initially differing, the introduction of outliers has mitigated this disparity. For this reason, the appropriate decision for any test is to reject the null hypothesis without being affected by outliers we added into the groups.

We test these contaminated data again to compare the mean vectors with T^2_{MRCD}, $T^2_{RCL\boldsymbol{X}}$, $T^2_{CL\boldsymbol{X}}$, T^2_{BS}, T^2_{SD}, and T^2_{CQ}. The results of the tests are given in the "Contaminated Data" part of Table 4. According to Table 4, T^2_{MRCD} still rejects the null hypothesis without being affected by outliers, whereas the other tests fail to reject the null hypothesis and are affected by outliers. These results show that T^2_{MRCD} is a robust test statistic to compare two mean vectors in high-dimensional datasets, while the other test statistics are not robust enough for this purpose.

7. Software Availability

We construct a function *RperT2()* to perform the proposed robust permutational test on real datasets in the R package called "MVTests" [15]. In this function, the data matrix of the first group must be entered into *X1*, and the data matrix of the second group must be entered into *X2*. The alpha argument determines the trimming ratio and must take a value from interval 0–1 (default; *alpha = 0.75*). The *N* argument determines the iteration number (default; *N = 100*). We give an example in Appendix A to demonstrate how to use this function to perform the proposed test on data in R.

8. Conclusions

In multivariate inference, the two-sample Hotelling T^2 statistic is popular. However, this statistic cannot be used for contaminated or high-dimensional data.

There are many studies in the literature that demonstrate that Hotelling T^2 statistics are not sensitive to outliers. The common approach in these studies is to use robust location and scatter estimation instead of classical ones [4–6]. However, these statistics cannot be used in high-dimensional data.

On the other hand, there are also many studies that obtain T^2 statistics, which can be used in high-dimensional data in the literature [9–12]. However, these statistics are sensitive to outliers in data.

Wang, Lin and Tang [19] proposed a robust test statistic that can be used in high-dimensional data. However, their statistic is useful for cell-wise contaminated data.

In this study, we propose a robust test statistic for high-dimensional data. This statistic is based on MRCD estimations. Because the finite-sample distribution of MRCD estimators is unknown, we propose using a permutation test without needing any asymptotic distribution.

We perform simulation studies to compare the empirical size, power and robustness performances of T^2_{MRCD}, $T^2_{RCL\boldsymbol{X}}$, $T^2_{CL\boldsymbol{X}}$, T^2_{BS}, T^2_{SD}, and T^2_{CQ} in clean and row-wise contaminated data. According to the simulation studies, the empirical sizes of T^2_{MRCD} are close to the nominal level (5%), the powers of T^2_{MRCD} are acceptable, and T^2_{MRCD} is not sensitive to outliers in data.

We perform tests on clean and contaminated Alzheimer's Disease data. In the clean case, all tests reject the null hypothesis. In the contaminated case, however, only T^2_{MRCD} rejects the null hypothesis, unlike other statistics.

Finally, we construct the *RperT2()* function to perform our proposed test on real data in the R package entitled MVTests [15]. We believe that our proposed statistic is a valuable contribution to multivariate inference for high-dimensional data.

In high-dimensional datasets, the missing values problem is another problem similar to the outlier one. Future studies will focus on how the proposed test statistic can cope with this problem in the case of missing observations in high-dimensional data.

Author Contributions: Conceptualization, H.B.; methodology, H.B.; software, H.B.; validation, S.I., N.F. and O.A.; formal analysis, H.B.; investigation, H.B. and N.F.; resources, S.I.; data curation, O.A.; writing—original draft preparation, H.B. and S.I.; writing—review and editing, H.B., N.F. and O.A.; visualization, H.B.; supervision, H.B.; funding acquisition, N.F. and O.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: We used synthetic datasets for simulations and Alzheimer's Disease datasets for real examples in this study. The Alzheimer's Disease data is available with the name *AlzheimerDisease* in the "AppliedPredictiveModeling" package in R software [24].

Acknowledgments: The authors thank the editors and four referees for their helpful, constructive comments, which have helped improve the quality and presentation of the paper.

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

Appendix A

`RperT2()` function in the MVTests package [15] performs the proposed test on real high-dimensional data. The general usage of the function is as below:

```
RperT2 (X1, X2, alpha = 0.75, N = 100)
```

Here, `X1` and `X2` is the data matrix for the first and second groups, respectively. Also, they must be matrix or data.frame. `alpha` is numeric parameter controlling the size of the subsets over which the determinant is minimized. The allowed values are between 0.5 and 1 and the default is 0.75. Finally, `N` is the permutation number and its default value is 100.

To give an example, firstly, we generate random high dimensional data in R as below:

```
> x<-mvtnorm::rmvnorm (n = 10, sigma = diag(20), mean = rep(0,20))

> y<-mvtnorm::rmvnorm (n = 10, sigma = diag(20), mean = rep(1,20))
```

To generate random data from multivariate normal distribution, we use the `rmvnorm()` function in the package ***mvtnorm*** in R. It is seen that both samples have 10 observations and 20 variables. Therefore, we can say the data is high-dimensional. Moreover, we can see the mean vector of first group is $[0, 0, \ldots, 0]_{20\times 1}$ and the mean vector of second group is

$[1, 1, \ldots, 1]_{20\times 1}$. The covariance matrices of both groups are the same as the identity matrix. As a result, we expect to find a statistically significant difference between group means.

We can perform the proposed test on this data with default `alpha` and `N` values as below:

```
> install.packages("MVTests")
> library(MVTests)
> RperT2(X1 = x, X2 = y)
$T2
TR2
419.7017
$p.value
[1] 0
```

Function returns us $T^2_{MRCD\ (c)}$ and the p-value. Because the p-value is <0.001, we reject the null hypothesis with a 99.9% confidence level.

Finally, we can perform the proposed test on this data with different `alpha` and `N` values as below:

```
> install.packages("MVTests")
> library(MVTests)
> RperT2(X1 = x, X2 = y, alpha = 0.9, N = 300))
$T2
TR2
378.2798
$p.value
[1] 0
```

Function returns us $T^2_{MRCD\ (c)}$ and the p-value. Because the p-value is < 0.001, we reject the null hypothesis with a 99.9% confidence level.

References

1. Rencher, A.C. *Methods of Multivariate Analysis*; John Willey & Sons, Inc.: Montreal, QC, Canada, 2002.
2. Bulut, H. *Multivariate Statistical Methods with R Applications*, 2nd ed.; Nobel Academic Publishing: Ankara, Turkey, 2023.
3. Willems, G.; Pison, G.; Rousseeuw, P.J.; Van Aelst, S. A Robust Hotelling Test. *Metrika* **2002**, *55*, 125–138. [CrossRef]
4. Cetin, M.; Altunay, S.A. Hotelling's T2 Statistic Based on Minimum-Volume-Ellipsoid Estimator. *Gazi Univ. J. Sci.* **2003**, *16*, 691–695.
5. Mokhtar, M.A.A.M.; Yusoff, N.S.; Liang, C.Z. Robust Hotelling's T2 Statistic Based on M-Estimator. *J. Phys. Conf. Ser.* **2021**, *1988*, 012116. [CrossRef]
6. Todorov, V.; Filzmoser, P. Robust Statistic for the One-Way Manova. *Comput. Stat. Data Anal.* **2010**, *54*, 37–48. [CrossRef]
7. Bai, Z.; Saranadasa, H. Effect of High Dimension: By an Example of a Two Sample Problem. *Stat. Sin.* **1996**, *6*, 311–329.
8. Lin, L.; Pan, W. Highmean: Two-Sample Tests for High-Dimensional Mean Vectors; R Package Version 3.0; 2016. Available online: https://cran.r-project.org/web/packages/highmean/index.html (accessed on 23 August 2024).
9. Srivastava, M.S.; Du, M. A Test for the Mean Vector with Fewer Observations than the Dimension. *J. Multivar. Anal.* **2008**, *99*, 386–402. [CrossRef]
10. Chen, S.X.; Qin, Y.-L. A Two-Sample Test for High-Dimensional Data with Applications to Gene-Set Testing. *Ann. Stat.* **2010**, *38*, 808–835. [CrossRef]
11. Cai, T.T.; Liu, W.D.; Xia, Y. Two-Sample Test of High Dimensional Means under Dependence. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2014**, *76*, 349–372.
12. Xu, G.; Lin, L.; Wei, P.; Pan, W. An Adaptive Two-Sample Test for High-Dimensional Means. *Biometrika* **2016**, *103*, 609–624. [CrossRef] [PubMed]
13. Boudt, K.; Rousseeuw, P.J.; Vanduffel, S.; Verdonck, T. The Minimum Regularized Covariance Determinant Estimator. *Stat. Comput.* **2020**, *30*, 113–128. [CrossRef]
14. Bulut, H. A Robust Hotelling Test Statistic for One Sample Case in High Dimensional Data. *Commun. Stat.-Theory Methods* **2023**, *52*, 4590–4604. [CrossRef]
15. Bulut, H. An R Package for Multivariate Hypothesis Tests: Mvtests. *Technol. Appl. Sci.* **2019**, *14*, 132–138. [CrossRef]
16. Rousseeuw, P.J.; Croux, C. Alternatives to the Median Absolute Deviation. *J. Am. Stat. Assoc.* **1993**, *88*, 1273–1283. [CrossRef]

17. Croux, C.; Haesbroeck, G. Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *J. Multivar. Anal.* **1999**, *71*, 161–190. [CrossRef]
18. Todorov, V.; Filzmoser, P. An Object-Oriented Framework for Robust Multivariate Analysis. *J. Stat. Softw.* **2010**, *32*, 1–47. [CrossRef]
19. Wang, W.; Lin, N.; Tang, X. Robust Two-Sample Test of High-Dimensional Mean Vectors under Dependence. *J. Multivar. Anal.* **2019**, *169*, 312–329. [CrossRef]
20. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013; Volume 26.
21. Kim, J.; Basak, J.M.; Holtzman, D.M. The Role of Apolipoprotein E in Alzheimer's Disease. *Neuron* **2009**, *63*, 287–303. [CrossRef] [PubMed]
22. Bu, G. Apolipoprotein E and Its Receptors in Alzheimer's Disease: Pathways, Pathogenesis and Therapy. *Nat. Rev. Neurosci.* **2009**, *10*, 333–344. [CrossRef] [PubMed]
23. Craig-Schapiro, R.; Kuhn, M.; Xiong, C.; Pickering, E.H.; Liu, J.; Misko, T.P.; Perrin, R.J.; Bales, K.R.; Soares, H.; Fagan, A.M.; et al. Multiplexed Immunoassay Panel Identifies Novel Csf Biomarkers for Alzheimer's Disease Diagnosis and Prognosis. *PLoS ONE* **2011**, *6*, e18850. [CrossRef] [PubMed]
24. Kuhn, M.; Johnson, K. Appliedpredictivemodeling: Functions and Data Sets for 'Applied Predictive Modeling'; R Package Version 1. 2018. Available online: https://cran.r-project.org/web/packages/AppliedPredictiveModeling/index.html (accessed on 23 August 2024).

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.