A Normality Test for High-dimensional Data Based on the Nearest Neighbor Approach

Hao Chen\textsuperscript{a} and Yin Xia\textsuperscript{b}

\textsuperscript{a}Department of Statistics, University of California at Davis, CA; \textsuperscript{b}Department of Statistics, School of Management, Fudan University

\textbf{ABSTRACT}

Many statistical methodologies for high-dimensional data assume the population is normal. Although a few multivariate normality tests have been proposed, to the best of our knowledge, none of them can properly control the Type I error when the dimension is larger than the number of observations. In this work, we propose a novel nonparametric test that uses the nearest neighbor information. The proposed method guarantees the asymptotic Type I error control under the high-dimensional setting. Simulation studies verify the empirical size performance of the proposed test when the dimension grows with the sample size and at the same time exhibit a superior power performance of the new test compared with alternative methods. We also illustrate our approach through two popularly used datasets in high-dimensional classification and clustering literatures where deviation from the normality assumption may lead to invalid conclusions.

\textbf{1. Introduction}

The population normality assumption is widely adopted in many classical statistical analysis (e.g., linear and quadratic discriminant analysis in classification, normal error linear regression models, and the Hotelling $T^2$-test), as well as many recently developed methodologies, such as network inference through Gaussian graphical models (Ma, Gong, and Bohnert 2007; Yuan and Lin 2007; Friedman, Hastie, and Tibshirani 2008; Rothman et al. 2008; Fan, Feng, and Wu 2009; Yuan 2010; Liu 2013; Xia, Cai, and Cai 2015), high-dimensional linear discriminant analysis (Bickel et al. 2004; Fan and Fan 2008; Cai and Liu 2011; Mai, Zou, and Yuan 2012), post-selection inference for regression models (Berk et al. 2013; Lee et al. 2016; Taylor and Tibshirani 2018), and change-point analysis for high-dimensional data (Xie and Siegmund 2013; Chan and Walther 2015; Wang and Samworth 2018; Liu, Zhang, and Mei 2019). When the data are univariate, there are many classical tools to check the normality assumption, such as the normal quantile-quantile plot and the Shapiro–Wilk test (Shapiro and Wilk 1965). However, many of the modern applications involve multivariate or even high-dimensional data and it constantly calls for multivariate normality testing methods with good theoretical performance.

In this article, we aim to address the following testing problem in the high-dimensional setting with a proper control of Type I error. Given a set of observations $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$, where $F$ is a distribution in $\mathbb{R}^d$, one wishes to test

$$H_0 : F \text{ is a multivariate Gaussian distribution},$$

versus the alternative hypothesis

$$H_a : F \text{ is not a multivariate Gaussian distribution}.$$
on multivariate skewness and kurtosis through the Bonferroni correction), “Ep” (an effective way of combining the multivariate skewness and kurtosis in Doornik and Hansen 2008), “Royston” (generalized Shapiro–Wilk test in Royston 1983), “HZ” (the test based on the characteristic function proposed in Henze and Zirkler 1990), “mvSW” (the multivariate Shapiro–Wilk’s test proposed in Villasenor Alva and Estrada 2009), and “eFR” (extended Friedman–Rafsky test in Smith and Jain 1988). In particular, the multivariate Shapiro–Wilk’s test requires smaller dimension than the sample size, and the extended Friedman–Rafsky test requires an estimate of the variance of the distribution while there is a lack of discussions on such estimations in their article. In the table, we use “mvSW0” and “eFR0” to, respectively, represent the test proposed in Villasenor Alva and Estrada (2009) and the extended Friedman–Rafsky test that are based on the sample covariance matrix, and use “mvSW” and “eFR” to, respectively, represent the tests that are based on a newly developed covariance matrix estimation method, the adaptive thresholding approach proposed in Cai and Liu (2011). We observe from the table that, except for the improved tests “mvSW” and “eFR,” all other existing methods are either not applicable to the cases when the dimension is larger than the sample size, that is, \( d > n \), or cannot control the Type I error well when the dimension is high.

The extended Friedman–Rafsky test is based on an edge-count two-sample test proposed in Friedman and Rafsky (1979). Due to the curse of dimensionality, it was shown in the recent work, Chen and Friedman (2017), that the edge-count two-sample test would suffer from low or even trivial power under some commonly appeared high-dimensional alternatives with typical sample sizes (ranging from hundreds to millions). The same problem also exists in the extended Friedman–Rafsky test for testing normality in the high-dimensional setting. Furthermore, the extended Friedman–Rafsky test can no longer properly control the Type I error when the dimension is much larger than the sample size, and similarly for the improved multivariate Shapiro–Wilk’s test. We refer the details to the size and power comparisons in Section 3.

In this article, we take into consideration the findings in Chen and Friedman (2017) and propose a novel nonparametric multivariate normality testing procedure based on nearest neighbor information. Through extensive simulation studies, we observe that the new test has good performance on the Type I error control, even when the dimension of the data is larger than the number of observations. It also exhibits much higher power than “mvSW” and “eFR” under the high-dimensional setting. Moreover, we provide theoretical results in controlling the Type I error for the new test when the dimension grows with the sample size. As far as we know, there is a paucity of systematic and theory-guaranteed hypothesis testing solutions developed for such type of problems in the high-dimensional setting, and our proposal offers a timely response. We also apply our test, respectively, to two datasets, a popularly used lung cancer dataset in the linear discriminant analysis literatures (Fan and Fan 2008; Cai and Liu 2011) where normality is a key assumption, and a colon cancer dataset that was used in high-dimensional clustering literature (Jin and Wang 2016) where the data are assumed to follow the normal assumption. The testing results provide useful prerequisites for such analyses that are based on the normality assumption.

The rest of the article is organized as follows. In Section 2, we propose a new nonparametric procedure to test the normality of the high-dimensional data and introduce the theoretical properties of the new approach. The performance of the proposed method is examined through simulation studies in Section 3 and the method is applied to two datasets in Section 4. Section 5 discusses a related statistic, possible extensions of the current proposal, and some sensitivity analyses. The main theorem is proved in Section 6 with technical lemmas collected and proved in Section 7.

2. Method and Theory

We propose in this section a novel nonparametric algorithm to test the normality of the high-dimensional data. We start with the intuition of the proposed method, and then study the error control of the new approach based on the asymptotic equivalence of two events for searching the nearest neighbors under the null hypothesis.

2.1. Intuition

A key fact of the Gaussian distribution is that it is completely determined by its mean and variance. Suppose that the mean \( \mu \) and covariance matrix \( \Sigma \) of the distribution \( F \) are known, then testing whether \( F \) is a multivariate Gaussian
distribution is the same as testing whether $F = G$, where $G = \mathcal{N}_d(\mu, \Sigma)$. For this purpose, one may consider goodness-of-fit tests, such as Bartoszyński, Pearl, and Lawrence (1997) and the approach proposed in Liu, Lee, and Jordan (2016) for high-dimensional data. We could also generate a new set of observations $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} G$, and apply the two-sample tests, such as Jurečková and Kalina (2012) and Marozzi (2015) and the graph-based two-sample tests (Friedman and Ráský 1979; Chen and Friedman 2017; Chen, Chen, and Su 2018), to examine $F = G$ for arbitrary dimensions.

However, in practice, the parameters $\mu$ and $\Sigma$ are unknown in general. To compromise, we use the mean ($\mu_x$) and covariance matrix ($\Sigma_x$) estimated from the set of observations $\{X_1, X_2, \ldots, X_n\}$ as substitutes. We could again generate a new set of observations $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} G_x = \mathcal{N}_d(\mu_x, \Sigma_x)$, but unfortunately, now the original testing problem is no longer equivalent to testing whether $F = G_x$.

To address this issue, we use the same combination of $\mu_x$ and $\Sigma_x$ to generate another set of independent observations $X_1^*, X_2^*, \ldots, X_n^* \overset{iid}{\sim} G_x = \mathcal{N}_d(\mu_x, \Sigma_x)$. Then we estimate the mean and covariance matrix of these new observations and denote them by $\mu_{x^*}$ and $\Sigma_{x^*}$, respectively. Based on them, we further generate a new set of independent observations from the normal distribution with mean $\mu_{x^*}$ and covariance matrix $\Sigma_{x^*}$, that is, $Y_1^*, Y_2^*, \ldots, Y_n^* \overset{iid}{\sim} \mathcal{N}_d(\mu_{x^*}, \Sigma_{x^*})$. Intuitively, if the null hypothesis $H_0$ is true, that is, the original distribution $F$ is multivariate Gaussian, then the relationship between $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_n\}$ would be similar to that of $\{X_1^*, X_2^*, \ldots, X_n^*\}$ and $\{Y_1^*, Y_2^*, \ldots, Y_n^*\}$. Henceforth, we shall test whether these two relationships are similar enough to decide whether $F$ is close enough to a Gaussian distribution.

In Smith and Jain (1988), the Friedman–Ráský’s two-sample test was used for this purpose. Unfortunately, as will be shown later in Section 3, this test was unable to properly control the Type I error when the dimension is growing with the number of observations.

To guarantee the error control in the high-dimensional setting, we use the nearest neighbor information in this article. To be specific, we pool $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_n\}$ together, and for each observation, we find its nearest neighbor, which is defined under the Euclidean distance in the current article. Similarly, we pool $\{X_1^*, X_2^*, \ldots, X_n^*\}$ and $\{Y_1^*, Y_2^*, \ldots, Y_n^*\}$ together, and again find the nearest neighbor for each observation.

Nearest neighbor information has been employed in hypothesis testing that can be applied to high-dimensional data (Schilling 1986; Henze et al. 1988; Chen and Zhang 2015; Chen 2019). However, in these work, nearest neighbors were used for two-sample testing, while in contrast, we only have one sample at the beginning of the current setup and then generate a second sample that depends on the original one. Hence, we need to develop a completely different set of technical tools to investigate the theoretical properties of the current construction. Let $YY$ be the event that an observation in $\{Y_1, Y_2, \ldots, Y_n\}$ finds its nearest neighbor in $\{Y_1, Y_2, \ldots, Y_n\}$, and let $Y^*Y^*$ be the event that an observation in $\{Y_1^*, Y_2^*, \ldots, Y_n^*\}$ finds its nearest neighbor in $\{Y_1^*, Y_2^*, \ldots, Y_n^*\}$. We will show in Theorem 1 that the events $Y^*Y^*$ and $YY$ are asymptotic equivalent under some suitable conditions. As a result, we can estimate the empirical distribution of the test statistic based on $YY$ through the distribution of the statistic associated with $Y^*Y^*$. Consequently, the Type I error of the proposed approach can be properly controlled at some prespecified significance level.

### 2.2. Theorem on Asymptotic Equivalence

Before studying the main result on the asymptotic equivalence between two events of searching nearest neighbors, we first introduce some notation. Denote by $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ the smallest and largest eigenvalues of $\Sigma$. For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$, denote by $a_n = O(b_n)$ if there exist constants $C > c > 0$ such that $c b_n \leq |a_n| \leq C b_n$ for all sufficiently large $n$. We also remark here that, when $d = 1$ or $d = 2$, the aforementioned univariate and conventional multivariate methods in the introduction can be easily applied to test the normality assumption, and we shall focus in our work on the cases when the dimension $d$ is larger than 2.

We next introduce two assumptions.

(A1) The eigenvalues of $\Sigma$ satisfy $C_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq C_2$ for some constants $C_1, C_2 > 0$.

(A2) There exists an estimator of $\mu$ such that $\|\mu_x - \mu\|_2 \leq O_p(1)$, and an estimator of $\Sigma$ such that $\|\Sigma_x - \Sigma\|_2 = o_p(n^{-\frac{1}{2}(2 + \log d + \epsilon)/\log n})$ with $\epsilon = 1 - \frac{1}{2} \log |\Sigma| - 2$ and $a = \begin{cases} 0 & \text{if } d \log d \leq \log n \\ \frac{\log n}{\epsilon d \log d} & \text{if } d \log d > \log n \text{ and } n = o(\log n) \\ 1/\epsilon d & \text{otherwise} \end{cases}$.

Under the above two conditions, Theorem 1 studies the asymptotic equivalence between the events $YY$ and $Y^*Y^*$ under the null hypothesis, which in turn guarantees the Type I error control of the proposed method.

**Theorem 1.** Assume (A1) and (A2). Then it follows that, under $H_0$, as $n \to \infty$, $P(YY) - P(Y^*Y^*) \to 0$.

The proof of the theorem is provided in Section 6.

**Remark 1.** Assumption (A1) is mild and is widely used in the high-dimensional literature (Bickel et al. 2008; Rothman et al. 2008; Yuan 2010; Cai, Liu, and Xia 2014). Assumption (A2) implies the relationship between the dimension $d$ and the sample size $n$. Specifically, $\|\mu_x - \mu\|_2 \leq O_p(1)$ can be easily satisfied when $d = O(n^\gamma)$, $\gamma \leq 1$. For the condition $\|\Sigma_x - \Sigma\|_2 = o_p(n^{-\frac{1}{2}(2 + \log d + \epsilon)/\log n})$, when $d \geq 3$ and $d = O(n^\gamma)$, $\gamma < 1/2$, it can be satisfied by many estimators under some regularity conditions. For example, if we apply the adaptive thresholding estimator in Cai and Liu (2011), and assume that $\Sigma$ is $s_0$ sparse in the sense that there are at most $s_0$ nonzero entries in each row of $\Sigma$, then we have $\|\Sigma_x - \Sigma\|_2 = O_p(s_0 \sqrt{\log d/n})$. So the condition holds if $s_0 = o(n^{\frac{1}{2} - \frac{\epsilon}{2} - \gamma})$ for some $\xi > (1 + \frac{\epsilon}{2})\gamma$, where $a$ is either equal or tending to zero as defined in detail in (A2). When $d = O(n^\gamma)$, $\gamma \geq 1/2$, simulation results show
that the conclusion holds well when \( d > n, d = O(n) \). There is potential to relax the condition on \( \| \Sigma_x - \Sigma \|_2 \) in the theorem. In the current proof, we made big relaxations from Equation (1) to Equation (2) and from Equation (3) to Equation (4) (see Section 6). More careful examinations could lead to tighter conditions. This requires nontrivial efforts and we save it for future work.

**Remark 2.** The theory based on nearest neighbor information in the high-dimensional setting has so far received little attention in the literature. We provide in this article a novel proof for the condition number is around 80 in typical simulation runs. Notethat Model 3 considers the nearly singular scenario where \( \min \Sigma \) is around 80 in the implementation.

2.3. Algorithm and Theoretical Error Control

Based on Theorem 1, we could adopt the following algorithm to test the multivariate normality of the data. To be specific, because of the asymptotic equivalence between the events \( Y^*Y^* \) and \( Y^*Y^* \), we repeatedly generate the data from the multivariate normal distribution with estimated mean and covariance matrix, and use the empirical distribution of the test statistics based on \( Y^*Y^* \) to approximate the empirical distribution of the test statistic based on \( Y^*Y^* \) under the null hypothesis.

Denote by \( r(YY) \) the percent of \( Y^* \)'s that find their nearest neighbors in \( \{Y_1,...,Y_n\} \), and \( r(Y^*Y^*) \) is defined similarly for \( Y^* \)'s. Let \( m(r(Y^*Y^*)) \) be the average of the \( r(Y^*Y^*) \)'s from Step 3 of the algorithm. We then propose a nonparametric normality test based on nearest neighbor information as the following.

**Algorithm 1**

1. Generate \( Y_1,\ldots,Y_n \overset{iid}{\sim} N_\mu(\mu, \Sigma) \), calculate \( r(YY) \).
2. Generate \( X_1,\ldots,X_n \overset{iid}{\sim} N_d(\mu, \Sigma) \), estimate its mean \( \mu \) and covariance matrix \( \Sigma_n \). Generate \( Y_1^*,\ldots,Y_n^* \overset{iid}{\sim} N_d(\mu^*, \Sigma_n) \), calculate \( r(Y^*Y^*) \).
3. Repeat Step 2 for \( B \) times to get an empirical distribution of \( r(YY) \) under \( H_0 \).
4. Compute the two-sided sampling \( p \)-value, \( p(YY) \), i.e., the percentage of \( |r(Y^*Y^*) - m(r(Y^*Y^*))| \) (out of \( B \)) that are larger than or equal to \( |r(YY) - m(r(Y^*Y^*))| \), where \( | \cdot | \) is the absolute value.
5. For a given significance level \( 0 < \alpha < 1 \), define \( \Psi_{\alpha} = I(p(YY) \leq \alpha) \). Reject the null hypothesis whenever \( \Psi_{0.05} = 1 \).

Note that Algorithm 1 is a simplified version of a more sophisticated algorithm that generates \( n \) independent sets of \( \{Y_1,\ldots,Y_n\} \) and \( \{Y_1^*,\ldots,Y_n^*\} \) in Steps 1 and 2, with \( r(YY) \) and \( r(Y^*Y^*) \) respectively representing the percent of \( Y_i \)'s and \( Y_i^* \)'s that find their nearest neighbors in their corresponding sets. The resulting test \( \Psi_{\alpha} \) of such an algorithm guarantees the Type I error control based on Theorem 1, that is, \( P(\text{Type I error}) = P_{H_0}(\Psi_{\alpha} = 1) \rightarrow \alpha \), as \( n,B \rightarrow \infty \). However, this algorithm is computationally much more expensive and it has asymptotically the same size performance as Algorithm 1, and hence we mainly focus on Algorithm 1 in the current article.

In the implementation, we use the sample mean to obtain \( \mu_x \) and \( \mu_{e^x} \) and use the adaptive thresholding method in Cai and Liu (2011) to compute \( \Sigma_x \) and \( \Sigma_{e^x} \). For the selection of \( B \), the empirical distribution can be more precisely estimated when \( B \) is larger. We choose \( B = 500 \) in the implementation and it provides well error control as shown in Section 3. It is worthwhile to note that, for faster and easier implementation of the method, the \( p \)-value \( p(YY) \) we obtain in Algorithm 1 is random, and we hence call it “sampling \( p \)-value.” To improve the power performance of the method, we can further increase the number of such sampling procedure, and the details are discussed in Section 5.3.

3. Simulation Studies

We analyze in this section the numerical performance of the newly developed algorithm. As we studied in the introduction, the existing methods “Skewness,” “Kurtosis,” “Bonferroni,” “Ep,” “Rosston,” and “mvSWt” all suffer from serious size distortion or are not applicable when the dimension is relatively large. We thus consider in this section the size and power comparisons of our approach with the method “eFR,” in which the covariances are estimated by the adaptive thresholding method in Cai and Liu (2011), the multivariate Shapiro–Wilk’s test (mvSW) proposed in Villasenor Alva and Estrada (2009) applying to the transformed statistics standardized by the adaptive thresholding covariance estimators, as well as the Fisher’s test (Fisher) by combining the \( p \)-values for each dimension of the aforementioned adaptive thresholding covariance standardized Shapiro–Wilk’s test. As suggested by Cai and Liu (2011), we use the fivefold cross-validation to choose the tuning parameter. Once we obtain an estimator \( \hat{\Sigma} \), we let \( \hat{\Sigma} = (\hat{\Sigma} + \delta I) / (1 + \delta) \) with \( \delta = \max\{-\lambda_{\min}(\Sigma), 0\} \) + 0.05 to guarantee the positive definiteness of the estimated covariance matrix.

The following matrix models are used to generate the data. Note that Model 3 considers the nearly singular scenario where the condition number is around 80 in typical simulation runs when \( d = 100 \).

- Model 1: \( \Sigma^{(1)} = I \).
- Model 2: \( \Sigma^{(2)} = (\sigma_{ij}^{(2)}) \) where \( \sigma_{ij}^{(2)} = 0.5|\cdot| \) for \( i \leq j \).
- Model 3: \( \Sigma^{(3)} = (\sigma_{ij}^{(3)}) \) where \( \sigma_{ij}^{(3)} = 1, \sigma_{ij}^{(3)} = \text{Unif}(1) \) Bernoulli(1, 0.02) for \( i < j \) and \( \sigma_{ij}^{(3)} = \sigma_{ij}^{(3)} \). \( \Sigma^{(3)} = (\Sigma^{(3)} + \delta I) / (1 + \delta) \) with \( \delta = \max\{-\lambda_{\min}(\Sigma^{(3)}), 0\} \) + 0.05 to ensure positive definiteness.

The sample sizes are taken to be \( n = 100 \) and 150, while the dimension \( d \) varies over the values 20, 100, and 300. For each model, data are generated from multivariate distribution with mean zero and covariance matrix \( \Sigma \). Under the null hypothesis, the distribution is set to be multivariate normal, while under the alternative hypothesis, the distribution is set to be one of the following distributions.

- Distribution 1: Multivariate \( t \) distribution with degrees of freedom \( v = d/2 \).
• Distribution 2: Mixture Gaussian distribution \(0.5N(0, (1 - a) \Sigma) + 0.5N(0, (1 + a) \Sigma)\) with \(a = \frac{\lambda}{\sqrt{d}}\).

We set the size of the tests to be 0.05 under all settings, and choose \(B = 500\) in the algorithm. We run 1000 replications to summarize the empirical size and power. The empirical size results are reported in Table 2 and the power results of Distributions 1 and 2 are reported in Tables 3 and 4.

From Table 2, we observe that the new test can control the size reasonably well under all settings, while the extended Friedman–Rafsky test has some serious size distortion for Model 3 when the dimension is larger than the sample size. In addition, both of the multivariate Shapiro–Wilk's test and the Fisher's test have some size inflation for Model 2 when \(d = 300\) and \(n = 100\).

For power comparison, we first studied the annoying heavy tail scenario—multivariate \(t\)-distribution. It can be seen from Table 3 that, the new test can capture the signal very well, while the extended Friedman–Rafsky test suffers from much lower power. In the meanwhile, both of the multivariate Shapiro–Wilk's test and the Fisher's test have competitive power performance under the low-dimensional settings, but have fast decaying power performance (much lower than the proposed method) as \(d\) increases. We also studied the scenario that the distribution is a mixture of two multivariate Gaussian distributions and we observed similar phenomena in Table 4 that the new test has much higher power than the extended Friedman–Rafsky test under all settings and has better performance than “mvSW” and “Fisher” for \(d = 100\) and 300.

The empirical size and power performance of all four methods are also illustrated in the empirical cumulative distribution function (ecdf) plots as shown in Figures 1 and 2, for Model 1 and \(d = n = 100\). We observe similar patterns for the other models. In summary, for all scenarios studied above, our newly proposed algorithm provides superior performance in both empirical size as well as empirical power comparing with the existing methods.

### Table 2. Empirical size (in percents) of the proposed algorithm (NEW), extended Friedman–Rafsky test (eFR), multivariate Shapiro–Wilk's test (mvSW) and the Fisher's test (Fisher).

| Model | \(n\) | \(d\) | 100 | 150 | 200 | 300 |
|-------|------|------|-----|-----|-----|-----|
| Model 1 | NEW | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 |
| | eFR | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 |
| | mvSW | 6.4 | 6.4 | 6.4 | 6.4 | 6.4 |
| | Fisher | 6.0 | 6.0 | 6.0 | 6.0 | 6.0 |
| Model 2 | NEW | 5.9 | 5.9 | 5.9 | 5.9 | 5.9 |
| | eFR | 4.3 | 4.3 | 4.3 | 4.3 | 4.3 |
| | mvSW | 5.6 | 5.6 | 5.6 | 5.6 | 5.6 |
| | Fisher | 5.3 | 5.3 | 5.3 | 5.3 | 5.3 |
| Model 3 | NEW | 5.9 | 5.9 | 5.9 | 5.9 | 5.9 |
| | eFR | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 |
| | mvSW | 4.7 | 4.7 | 4.7 | 4.7 | 4.7 |
| | Fisher | 5.2 | 5.2 | 5.2 | 5.2 | 5.2 |

### Table 3. Empirical power (in percents) of the proposed algorithm (NEW), extended Friedman–Rafsky test (eFR), multivariate Shapiro–Wilk's test (mvSW) and the Fisher's test (Fisher) for multivariate \(t\)-distribution.

| Model | \(n\) | \(d\) | 100 | 150 | 200 | 300 |
|-------|------|------|-----|-----|-----|-----|
| Model 1 | NEW | 45.8 | 81.7 | 81.6 | 66.3 | 95.8 |
| | eFR | 6.6 | 3.9 | 5.3 | 7.4 | 3.7 | 5.7 |
| | mvSW | 56.9 | 12.5 | 8.0 | 68.7 | 17.1 | 10.5 |
| | Fisher | 56.0 | 12.7 | 8.0 | 69.8 | 16.8 | 10.3 |
| Model 2 | NEW | 15.3 | 61.9 | 71.9 | 26.7 | 67.2 | 81.9 |
| | eFR | 6.3 | 5.6 | 4.9 | 10.4 | 6.4 | 6.5 |
| | mvSW | 46.3 | 17.7 | 19.2 | 63.3 | 19.7 | 12.7 |
| | Fisher | 47.3 | 18.3 | 19.2 | 63.8 | 19.8 | 12.3 |
| Model 3 | NEW | 45.0 | 75.4 | 86.9 | 64.0 | 90.8 | 94.7 |
| | eFR | 6.5 | 4.5 | 21.7 | 7.5 | 3.8 | 14.9 |
| | mvSW | 53.2 | 19.0 | 10.5 | 70.1 | 18.9 | 13.9 |
| | Fisher | 55.0 | 19.3 | 9.8 | 70.5 | 18.5 | 13.2 |

### Table 4. Empirical power (in percents) of the proposed algorithm (NEW), extended Friedman–Rafsky test (eFR), multivariate Shapiro–Wilk's test (mvSW) and the Fisher's test (Fisher) for mixture Gaussian distribution.

| Model | \(n\) | \(d\) | 20 | 100 | 300 | 20 | 100 | 300 |
|-------|------|------|-----|-----|-----|-----|-----|-----|
| Model 1 | NEW | 45.8 | 81.7 | 81.6 | 66.3 | 95.8 | 94.8 |
| | eFR | 6.6 | 3.9 | 5.3 | 7.4 | 3.7 | 5.7 |
| | mvSW | 55.9 | 12.5 | 8.0 | 68.7 | 17.1 | 10.5 |
| | Fisher | 56.0 | 12.7 | 8.0 | 69.8 | 16.8 | 10.3 |
| Model 2 | NEW | 15.3 | 61.9 | 71.9 | 26.7 | 67.2 | 81.9 |
| | eFR | 6.3 | 5.6 | 4.9 | 10.4 | 6.4 | 6.5 |
| | mvSW | 46.3 | 17.7 | 19.2 | 63.3 | 19.7 | 12.7 |
| | Fisher | 47.3 | 18.3 | 19.2 | 63.8 | 19.8 | 12.3 |
| Model 3 | NEW | 45.0 | 75.4 | 86.9 | 64.0 | 90.8 | 94.7 |
| | eFR | 6.5 | 4.5 | 21.7 | 7.5 | 3.8 | 14.9 |
| | mvSW | 53.2 | 19.0 | 10.5 | 70.1 | 18.9 | 13.9 |
| | Fisher | 55.0 | 19.3 | 9.8 | 70.5 | 18.5 | 13.2 |

### 4. Application

Classification is an important statistical problem that has been extensively studied both in the traditional low-dimensional setting and the recently developed high-dimensional setting. In particular, Fisher's linear discriminant analysis has been shown to perform well and enjoy certain optimality as the sample size tends to infinity while the dimension is fixed (Anderson 2003), and it has also been widely studied in the high-dimensional setting when the sample covariance matrix is no longer invertible, see, for example, Bickel et al. (2004), Fan and Fan (2008), Cai and Liu (2011) and Mai, Zou, and Yuan (2012). In all of those studies, normality of the data is a key assumption in order to obtain the linear discriminant rule and investigate the subsequent analysis of misclassification rate. We study in this section a lung cancer dataset, which was analyzed by Gordon et al. (2002) and is available at R documentation data(lung) with package propOverlap. This dataset was popularly used in the classification literature (Fan and Fan 2008; Cai and Liu 2011) where normality is a key assumption. In addition, we explore a dataset that was analyzed in Jin and Wang (2016) by their method IF-PCA for clustering, where data normality is assumed.
4.1. Lung Cancer Data

The lung cancer dataset has 181 tissue samples, including 31 malignant pleural mesothelioma (MPM) and 150 adenocarcinoma (ADCA), and each sample is described by 12,533 genes. This dataset has been analyzed in Fan and Fan (2008) by their methods FAIR and NSC, and in Cai and Liu (2011) by their LPD rule, for distinguishing MPM from ADCA, which is important and challenging from both clinical and pathological perspectives. However, before applying their proposed methods, none of them have checked the normality of the data, which is a fundamental assumption in the formulation of linear discriminants. If the normality fails to hold, then the misclassification rates can be affected and their results may no longer be valid.

In this section, we use our newly developed method to check the normality of the 150 ADCA samples in this lung cancer dataset. Note that, multivariate normality assumption for the 12,533 genes of the ADCA samples will be rejected if any subset for this large number of genes deviate from the normality. Thus, we randomly select a group of 200 genes, and applied our new method to test the multivariate normality assumption. By applying Algorithm 1 with $B = 500$, we obtain that, the sampling $p$-value is equal to 0, which gives sufficient evidence that the samples from this dataset have severe deviation from the multivariate normal distribution. We further repeat this procedure for 100 times. In each time, we randomly select a group of 200 genes and apply Algorithm 1 ($B = 500$) to the selected genes. It turns out that the sampling $p$-values are all 0 for these 100 times. Thus, it is not reasonable to assume the normality and directly apply the recent developed high-dimensional linear discriminant procedures to classify MPM and ADCA, as studied in Fan and Fan (2008) and Cai and Liu (2011). So our procedure serves as an important initial step for checking the normality assumption before applying any statistical analysis methods which assume such conditions.

4.2. Colon Cancer Data

Next, we study in this section a gene expression dataset on tumor and normal colon tissues that was analyzed and cleaned by Detting (2004). This dataset can be found at https://blog.nus.edu.sg/staww/softwarecode/. It has 40 tumor and 22 normal colon tissue samples, and each sample is described by 2000 genes. This dataset has been analyzed in Jin and Wang (2016) by their method IF-PCA for clustering, where they imposed normality assumption on the data, though they found the violation to such assumption in their analysis as the empirical null distribution of a test statistic they used was far from the theoretical null distribution derived from the normal assumption.

In this section, we use our newly developed method to check the normality of the 40 tumor samples in this colon cancer dataset. We compare the proposed method with eFR, the multivariate Shapiro–Wilk’s test and the Fisher’s test in this analysis. By applying Algorithm 1 with $B = 500$, we obtain that, the sampling $p$-value is equal to 0, which gives a sufficient evidence that the samples from this dataset have severe deviation from the multivariate normal distribution. This double confirms the deviation from the normality assumption noticed by the authors in Jin and Wang (2016). On the other hand, both the multivariate Shapiro–Wilk’s test and the Fishers test successfully reject the null while the eFR method reports a sampling $p$-value of 1 and fails to detect the violation to the normality assumption.
5. Discussion

We proposed in this article a nonparametric normality test based on the nearest neighbor information. It enjoys proper error control and is shown to have significant power improvement over the alternative approaches. We discuss in this section a related test statistic and some extensions and explorations of the current method.

5.1. Test Statistic Based on XX

Our proposed test statistic involves the event YY, that is, the event that an observation in \{Y_1, Y_2, \ldots, Y_n\} finds its nearest neighbor in \{X_1, X_2, \ldots, X_n\}. A straightforward alternative method could be based on the test statistics which involves the event XX, that is, the event that an observation in \{X_1, X_2, \ldots, X_n\} finds its nearest neighbor in \{X_1, X_2, \ldots, X_n\}, and a question is whether the XX-equivalent statistic could be incorporated to further enhance the power. Unfortunately, the XX version is not as robust as the YY version and does not have good performance in controlling the Type I error. Table 5 lists the empirical size of the XX version of the test under the same settings as in Table 2. We observe that this statistic has serious size distortion for Model 3 when the dimension is high. This also explains the bad performance of eFR in controlling Type I error under Model 3 because eFR partially uses the XX information.

5.2. Extension to Other Distributions in the Exponential Family

The idea of constructing this normality test could be extended to other distributions in the exponential family. As long as one has reasonably good estimators for the parameters of the distribution, a similar procedure as described in Section 2 can be applied. In particular, one could replace the multivariate normal distribution in Algorithm 1 by the distribution of interest, and replace the mean and covariance estimators by the estimators of the corresponding parameters. The conditions for the asymptotic equivalence between the events YY and Y^∗Y^∗ would need more careful investigations and warrant future research.

5.3. A Power Enhanced Algorithm

To further improve the power performance of the method, especially when the sample size is limited, we can increase the number of sampling procedure in Algorithm 1 as detailed in the following algorithm.

| Algorithm 2 |
|-------------|
| 1. For \( i = 1 \) : \( L \), generate \( Y_{i1}, \ldots, Y_{in} \sim N_d(\mu, \Sigma) \), calculate \( r_i(YY) \). Let \( \overline{r}(YY) \) be the average of \( r_i(YY) \)'s.
| 2. Generate \( X_{1}^\ast, \ldots, X_{n}^\ast \sim N_d(\mu, \Sigma) \), estimate its mean \( \mu^\ast \) and covariance matrix \( \Sigma^\ast \). For \( i = 1 : L \), generate \( Y_{i1}^\ast, \ldots, Y_{in}^\ast \sim N_d(\mu^\ast, \Sigma^\ast) \), calculate \( r_i(Y^\ast Y^\ast) \). Let \( \overline{r}(Y^\ast Y^\ast) \) be the average of \( r_i(Y^\ast Y^\ast) \)'s.
| 3. Repeat Step 2 for \( B \) times to get an estimate of the empirical distribution of \( \overline{r}(YY) \) under \( H_0 \).
| 4. Compute the two-sided sampling p-value, \( p(YY) \), i.e., the percentage of \( \{r(Y^\ast Y^\ast) - m(\overline{r}(Y^\ast Y^\ast))\} \) (out of \( B \)) that are larger than or equal to \( \{|\overline{r}(YY) - m(\overline{r}(Y^\ast Y^\ast))|\} \), where \( m(\overline{r}(Y^\ast Y^\ast)) \) is the average of \( r_i(Y^\ast Y^\ast) \)'s in Step 3.
| 5. For a given significance level \( 0 < \alpha < 1 \), define \( \Psi_\alpha = I\{p(YY) \leq \alpha\} \). Reject the null hypothesis whenever \( \Psi_\alpha = 1 \).

Note that, when \( L = 1 \), Algorithm 2 is reduced to Algorithm 1 in Section 2.3. In the following Figure 3, we show the boxplots of the sampling p-values for Distribution 1 and Model 3 when \( L = 1, 2, \ldots, 10 \), for \( d = n = 100 \), with 100 replications. Similar patterns are observed for the other models. It can be seen that, as \( L \) increases, the power performance of the method can be significantly improved and will get stable when \( L \) is around 5. Also note that, the computation cost is growing as \( L \) increases. Hence, we mainly recommend Algorithm 1 in the article as it already shows reasonable well performance both in terms of empirical size and power as illustrated in Section 3.

5.4. Non-significant Results and Scale Sensitivity

When the testing results are nonsignificant, it means that the distribution is very close to the multivariate normal, whereas the unbiased property of the proposed test warrants future research. In this section, we perform additional simulation studies to explore the power performance of the proposed method as the distributions are approaching to normality. Specifically, we consider the following three sets of distributions: multivariate chi-squared distributions, multivariate t-distributions and multivariate Gaussian distribution with a certain proportion of the dimensions replaced by t distribution.

![Figure 3](image-url)

Figure 3. Boxplots of the sampling p-values for Distribution 1 and Model 3 in Section 3, for \( d = m = 100 \), with 100 replications.
• Multivariate chi-squared distribution with degrees of freedom $\nu = 3, 5, 10,$ and 20 (the larger $\nu$ is, the closer the distribution is to the multivariate normal distribution).

• Multivariate $t$-distribution with degrees of freedom $\nu = d/4, d/2, d$ and $2d$ (the larger $\nu$ is, the closer the distribution is to the multivariate normal distribution).

• Multivariate Gaussian distribution $N_d(0, \Sigma)$ with a certain proportion of the dimensions, ranging from 0.5 to 0.1, replaced by Multivariate $t$-distribution with degrees of freedom $\nu = d/4$ (the smaller the proportion is, the closer the distribution is to the multivariate normal distribution).

The power performance of the methods is summarized in Tables 6 and 7. It can be seen from the tables that, when the alternatives are getting closer to the multivariate normal, the testing results become more and more non-significant.

In addition, we explore the scale sensitivity of the proposed test by considering different covariance models with varying condition numbers as follows.

• Model 1A: $\Sigma = I$.

• Model 1B: $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d,d})$, where $\sigma_{i,i} = \text{Unif}(1,5)$ for $i = 1, \ldots, d$.

• Model 1C: $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_{d,d})$, where $\sigma_{i,i} = \text{Unif}(1,20)$ for $i = 1, \ldots, d$.

We see from Table 8 that, the empirical size and power performance (Distribution 1) of these three models are very similar to each other, which shows that the proposed method is not sensitive to the scale of the data.

### Table 6. Empirical power (in percent) of the proposed algorithm (NEW) and extended Friedman-Rafsky test (eFR) for Model 3, $\alpha = 0.05$, $d = 100$ and $n = 100$.

| Distribution | Chi-square | t |
|--------------|------------|---|
| DOF | 3 | 5 | 10 | 20 | $d/4$ | $d/2$ | 2d | 4d |
| NEW | 45.1 | 25.2 | 10.6 | 6.5 | 99.7 | 87.9 | 17.3 | 6.3 |
| eFR | 9.4 | 6.8 | 6.5 | 5.1 | 2.3 | 4.8 | 4.0 | 3.3 |

### Table 7. Empirical power (in percent) with varying proportion of non-Gaussian dimensions, ranging from 0.5 to 0.1, for Model 1, $\alpha = 0.05$, $d = 100$ and $n = 100$.

| non-Gaussian proportion | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 |
|-------------------------|-----|-----|-----|-----|-----|
| NEW | 48.1 | 25.9 | 12.9 | 7.0 | 4.2 |
| eFR | 3.7 | 5.2 | 4.8 | 5.0 | 4.1 |

### Table 8. Empirical size and power (in percent) of the proposed algorithm (NEW) and extended Friedman-Rafsky test (eFR) for varying condition numbers, $n = 100$.

| $d$ | 20 | 100 | 300 | 20 | 100 | 300 |
|-----|----|-----|-----|----|-----|-----|
| Model 1A | NEW | 4.3 | 4.3 | 5.1 | 58.5 | 91.3 | 93.0 |
| | eFR | 4.3 | 4.8 | 4.3 | 6.7 | 3.7 | 6.4 |
| Model 1B | NEW | 3.2 | 4.7 | 5.0 | 55.9 | 91.2 | 91.7 |
| | eFR | 4.0 | 3.6 | 5.6 | 7.6 | 3.9 | 4.4 |
| Model 1C | NEW | 5.1 | 5.1 | 5.3 | 49.8 | 87.0 | 92.0 |
| | eFR | 4.7 | 5.3 | 4.7 | 7.1 | 3.6 | 4.7 |

### 6. Proof of Theorem 1

Let $\Sigma = U \Lambda U^T$ and $\Sigma_x = U_x \Lambda_x U_x^T$ be, respectively, the eigen-decomposition of $\Sigma$ and $\Sigma_x$. Define $\Sigma^{1/2} = U \Lambda^{1/2} U^T$ and $\Sigma_x^{1/2} = U_x \Lambda_x^{1/2} U_x^T$. Then under the conditions of Theorem 1, by Lemma 1, we have

$$\|\Sigma^{1/2} - \Sigma_x^{1/2}\|_2 = \alpha_p(n^{-1/2}) = \frac{1}{\log n}.$$

Let $f(\cdot)$ be the density of $N_d(\mu, \Sigma)$, and $\tilde{f}(\cdot)$ be the density of $N_d(\mu_x, \Sigma_x)$. Then we have

$$P(YY) = \int P(YY|X_i = x_i)_{i=1,\ldots,n} \prod_{i=1}^{n} f(x_i) dx_i,$$

$$P(Y^*Y^*) = \int P(Y^*Y^*|X_i = x_i)_{i=1,\ldots,n} \prod_{i=1}^{n} \tilde{f}(x_i) dx_i.$$

By the construction of $\{Y_1, \ldots, Y_n\}$ and $\{Y^*_1, \ldots, Y^*_n\}$, we have

$$P(YY|X_i = x_i)_{i=1,\ldots,n} = P(Y^*Y^*|X_i = x_i^*)_{i=1,\ldots,n}.$$

Hence,

$$P(Y^*Y^*) = \int P(YY|X_i = x_i^*)_{i=1,\ldots,n} \prod_{i=1}^{n} \tilde{f}(x_i) dx_i.$$

By a change of measure, we have

$$P(Y^*Y^*) = \int P(YY|X_i = \Sigma_x^{1/2} \Sigma^{-1/2} x_i - \mu + \mu_x)_{i=1,\ldots,n} \prod_{i=1}^{n} f(x_i) dx_i.$$

It is not hard to see that if we shift the $x_i$'s all by a fixed value, the probability of $YY$ is unchanged. Hence,

$$P(Y^*Y^*) = \int P(YY|X_i = \Sigma_x^{1/2} \Sigma^{-1/2} x_i)_{i=1,\ldots,n} \prod_{i=1}^{n} f(x_i) dx_i.$$

Let $w_i = \Sigma_x^{1/2} \Sigma^{-1/2} x_i$. Then,

$$|P(YY) - P(Y^*Y^*)| = \left| \int \left( P(YY|X_i = x_i)_{i=1,\ldots,n} - P(YY|X_i = w_i)_{i=1,\ldots,n} \right) \prod_{i=1}^{n} f(x_i) dx_i \right| \leq \int \left| P(YY|X_i = x_i)_{i=1,\ldots,n} - P(YY|X_i = w_i)_{i=1,\ldots,n} \right| \prod_{i=1}^{n} f(x_i) dx_i.$$
By change of measure, we have that

\[
\begin{align*}
\int P(N_{\tilde{Y}_1} \in \{\tilde{Y}\} | Y_1 = \tilde{y}) g_{\tilde{Y}}(\tilde{y}) d\tilde{y} \\
= \int P(N_{\tilde{Y}_1} \in \{\tilde{Y}\}) \tilde{y} = y) g_Y(y) dy \\
- \int P(N_{\tilde{Y}_1} \in \{\tilde{Y}\} | Y_1 = \tilde{y}) g_{\tilde{Y}_2}(\tilde{y}) d\tilde{y}.
\end{align*}
\]

By similar arguments, we have

\[
\begin{align*}
\| (\Sigma_2^{-1/2} - \Sigma_1^{-1/2} / 2 \Sigma_1^{1/2}) \mu_2 \|_2 & = \| (\Sigma_2^{-1/2} - \Sigma_1^{-1/2} / 2 \Sigma_1^{1/2}) \Sigma_1^{-1/2} \mu_1 \|_2 \\
= \| (\Sigma_2^{-1/2} - \Sigma_1^{-1/2} / 2 \Sigma_1^{1/2}) \Sigma_1^{-1/2} \mu_1 \|_2 & = \| (\Sigma_2^{-1/2} - \Sigma_1^{-1/2} / 2 \Sigma_1^{1/2}) \mu_1 \|_2 \\
& \leq \| (\Sigma_2^{-1/2} - \Sigma_1^{-1/2} / 2 \Sigma_1^{1/2}) \mu_1 \|_2 \\
& = \| (\Sigma_2^{-1/2} - \Sigma_1^{-1/2} / 2 \Sigma_1^{1/2}) \mu_1 \|_2 \\
& = \| (\Sigma_2^{-1/2} - \Sigma_1^{-1/2} / 2 \Sigma_1^{1/2}) \mu_1 \|_2.
\end{align*}
\]

Thus, we have that \( \| y_w - y \|_2 = op(n^{\alpha_2}). \)

Let \( j_x = \arg \min \{ \| y - x_i \|_2 \}, \)

\( j_w = \arg \min \{ \| y_w - w_i \|_2 \}. \)

and \( D_{\min,x} = \| y - x_{j_x} \|_2, D_{\min,w} = \| y_w - w_{j_w} \|_2. \)

Suppose \( D_{\min,x} = op(n^{\alpha}). \) Notice that \( n^{\alpha} \geq n^{-\frac{1}{2}} d^{-1/2} e^{-\frac{x}{d}} \leq O(d^{-1/2}). \) When \( \alpha < \alpha^*, \) based on Lemma 3, the probability that \( D_{\min,x} = c_0 n^{\alpha} \) for some constant \( c_0 > 0 \) is of order \( n \times op(n^{\alpha} d^{-2} d^{-1/2} e^{-d/2} / \log n) \leq op(d^{-1/2}). \)

We thus focus on \( \alpha \geq \alpha^*. \) By definitions of \( D_{\min,x} \) and \( D_{\min,w} \), the facts that \( |x_i - w_i| \leq op(n^{\alpha_*}), \) and \( |y - y_w| \leq op(n^{\alpha_*}), \) we have that \( D_{\min,x} = D_{\min,w} + op(n^{\alpha_*}). \)

Let \( \rho \) be the probability that \( Y_k \sim \mathcal{N}(\mu_1, \Sigma_1) \) falls in the \( D_{\min,x} \)-ball of \( y, \) and \( p_{\omega} \) be the probability that \( \tilde{Y}_k \sim \mathcal{N}(\mu_2, \Sigma_2) \) falls in the \( D_{\min,w} \)-ball of \( y_w. \)

Let \( \alpha_0 = -\frac{1}{d} + \frac{1/2}{2 \log n} > \alpha^*. \) We consider two scenarios: (1) \( \alpha^* < \alpha < \alpha_0, \) and (2) \( \alpha \leq \alpha_0. \)

(a) When \( d \log d \leq \log n, \) we have \( n^{\alpha_0} \geq n^{-\frac{1}{2}} d^{-1/2} e^{-\frac{x}{d}} \leq O(d^{-1/2}). \) Since \( \mu_1 \) and \( \Sigma_1 \) satisfy the condition for Lemma 4, we have

\[
p_{\omega} = op(n^{\alpha_*} d^{-2} d^{-1/2} e^{-d/2} / \log n) \\
= op(n^{-1} \sqrt{\log 1 / |\Sigma_1|}) = op(n^{-1}),
\]

where \( \kappa_1 = 1 - \frac{\log |\Sigma_1|}{d} \leq 2. \)

Notice that \( \mu_2 \) and \( \Sigma_2 \) also satisfy the condition for Lemma 4, so

\[
p_{\omega} = op(n^{\alpha_*} d^{-2} d^{-1/2} e^{-d/2} / \log n) \\
= op(n^{-1} \sqrt{\log 1 / |\Sigma_2|}) = op(n^{-1}),
\]

where \( \kappa_2 = 1 - \frac{\log |\Sigma_2|}{d} \leq 2. \)

(b) When \( d \log d \geq \log n \), we have \( n^{\alpha_0} \leq O(\sqrt{d}), \) by Lemma 4, \( \log p_x \) is

\[
-\frac{1}{2} d \log d + d \log D_{\min,x} + \frac{1}{2} d (\kappa_1 + Op(1)) \\
= -\log n - \frac{d}{2} (\log d + Op(1)).
\]

Here, \( a = \frac{\log n}{\xi d_n, d \log d} \) with \( 1 \ll \xi d_n = \log (log n / d) \) a positive constant. We have \( \log p_x = -\log n - \frac{1}{2} \log n + Op(d) \ll -\log n. \) So \( p_x = op(n^{-1}). \) Similarly, \( p_{\omega} = op(n^{-1}). \)

(c) When \( d \) is of order \( \log n \) or higher, \( a = 1 / \xi d \) with \( 1 \ll \xi d_n = \log (log n / d) \) a positive constant. We have \( \log p_x = -\log n - \frac{1}{2} \log n + Op(d) \ll -\log n. \) So \( p_x = op(n^{-1}). \) Similarly, \( p_{\omega} = op(n^{-1}). \)
Under (a), (b), and (c), we all have \( p_x, p_w = O(p(n^{-1}) \). Then,
\[
|P(N_{Y_1} \in \{Y\} | Y_1 = y) - P(N_{\tilde{Y}_1} \in \{\tilde{Y}\} | Y_1 = y_w)| = |op(1) - op(1)| = op(1).
\]

2. \( \alpha \geq \alpha_0 \):

First we consider \( \alpha_0 \leq \alpha \leq \frac{\log d}{\log n} \). By the proof of Lemma 4 and the facts that
\[
|e^{(\epsilon_1 - \kappa d/2)} - \sqrt{\Sigma_1/\|\Sigma_1\|} - 1 + \alpha o(1),
\]
\[
e^{(\epsilon_2 - \kappa d/2)} - \sqrt{\Sigma_2/\|\Sigma_2\|} - 1 + \alpha o(1),
\]
and \( \|\Sigma_1^{-1} - \Sigma_1^{-1}\|_2 = op(n^{2\gamma}) \).

Then \( p_w \) is
\[
p_w \left( 1 + op(n^{2\gamma}) \right) e^{op(n^{2\gamma} e^{1/2})} = \sqrt{\Sigma_1/\|\Sigma_1\|} - 1 + \alpha o(1) + op(n^{2\gamma})
\]
\[
= p_w \left( 1 + \alpha o(1) \right) = p_w(1 + op(1)).
\]

Similarly, \( p_{w,2} = p_w(1 + op(d^{1/2}/D_{max,n})) = p_w(1 + op(1)) \). Then \( p_{x,2} \) and \( p_{w,2} \) are also of order \( Op(n^{-1}) \).

Based on the proof of Lemma 3, \( p_{x,2} \) and \( p_{w,2} \) differ by a factor of
\[
1 + Op(d\gamma) = 1 + Op(d^{\gamma - \log n}/\log n) - O(n^{-1}) = 1 + op(1).
\]

Notice that \( p_{x,2} - p_x = p_x(p_{x,2}/p_x - 1) = Op(n^{-1}) \). Similarly, \( p_{w,2} - p_w = Op(n^{-1}) \). Let \( \xi_{min} = \min (\delta d/D_{max,n}, d^{\gamma - \log n}/\log n) \), we have that \( |p_x - p_w| = Op(n^{-1} \xi_{min}) \). Let \( p_x = c_0 n^{-1} \). Then \( p_w = c_0 n^{-1} + c_1 n^{-1} \xi_{min} + op(n^{-1} \xi_{min}) \) for a constant \( c_1 \), then,
\[
|P(N_{Y_1} \in \{Y\} | Y_1 = y) - P(N_{\tilde{Y}_1} \in \{\tilde{Y}\} | \tilde{Y}_1 = y_w)| = (\xi_{min} - 1) + \alpha o(n^{-1} \xi_{min} + op(1) = Op(1).
\]

Thus, under all possibilities of Scenarios (1) and (2), we have
\[
|P(N_{Y_1} \in \{Y\} | Y_1 = y) - P(N_{\tilde{Y}_1} \in \{\tilde{Y}\} | \tilde{Y}_1 = y_w)| = Op(1).
\]

Hence,
\[
|P(YY) - P(Y^*Y^*)| \leq \int_{x_1, \ldots, x_n} |P(y \in \{Y\} | Y_1 = y) - P(\tilde{y} \in \{\tilde{Y}\} | \tilde{Y}_1 = y)| d\gamma_1(y) dy 
\]
\[
\times f(\gamma_1(y)) d\gamma_1 \sum_{i=1}^{n} \gamma_1(x_i) d\gamma_1 = o(1),
\]
and the conclusion of the theorem follows.

7. Technical Lemmas

**Lemma 1.** For independent observations \( X_1, \ldots, X_n \) \( \iid \) \( \Lambda_d(\mu, \Sigma) \), assume that \( \lambda_{min}(\Sigma) \geq C \) for some constant \( C > 0 \). If \( \|\Sigma_1 - \Sigma_2\|_2 = op(r_{n,d}) \) with \( r_{n,d} = O(1) \), then we have \( \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_2 = op(r_{n,d}) \).

**Proof.** Denote by \( v \in R^d \) an eigenvector of \( \Sigma_1^{1/2} - \Sigma_2^{1/2} \) of unit length, we have
\[
|\langle(\Sigma_1^{1/2} - \Sigma_2^{1/2})v, v\rangle| = \|v\|^2 \|\Sigma_1 - \Sigma_2\|_2 = op(r_{n,d}).
\]

Suppose that \( \langle(\Sigma_1^{1/2} - \Sigma_2^{1/2})v, v\rangle = \lambda v \), then we have that
\[
|\lambda v^T(\Sigma_1^{1/2} + \Sigma_2^{1/2})v| = op(r_{n,d}).
\]

By the condition that \( \lambda_{min}(\Sigma) \geq C \) and that \( \|\Sigma - \Sigma_2\|_2 = op(r_{n,d}) \), we have
\[
v^T(\Sigma_2^{1/2}v + v^T(\Sigma_1^{1/2} - \Sigma_2^{1/2})v \geq C - o(1)
\]
with probability going to 1. Hence, for some constant $C_0 > 0$, we have, with probability tending to 1,

$$v^T(\Sigma_1^{1/2} + \Sigma_1^{1/2})v \geq C_0.$$ 

It yields that $\lambda = \text{op}(r_{n,d})$. Since $v$ could be any eigenvector of $\Sigma_1^{1/2} - \Sigma_1^{1/2}$, we have

$$\|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2 = \text{op}(r_{n,d}).$$

\[\square\]

**Lemma 2.** Suppose $x \sim N_d(0, \Sigma)$. Then under the conditions of Theorem 1, we have

$$\|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2 = \text{op}(n^{-1/2}(\log d)(\log n)^{1/2}).$$ 

(7)

**Proof.**

$$\|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2 = \|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2$$

$$\|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2 = \|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2$$

Let $z = \Sigma_1^{1/2}x$, we have

$$\|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2 \leq \|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2 \leq \|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2.$$

Notice that the covariance matrix of $z$ is an identity matrix, $\|z\|^2/d$ converges to a constant almost surely as $d \to \infty$. By the condition that $\|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2 = \text{op}(n^{-1/2}(\log d)(\log n)^{1/2})$, we have that

$$\|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2 = \text{op}(n^{-1/2}(\log d)(\log n)^{1/2})\|z\|_2$$

$$\|\Sigma_1^{1/2} - \Sigma_1^{1/2}\|_2 = \text{op}(n^{-1/2}(\log d)(\log n)^{1/2}).$$

\[\square\]

**Lemma 3.** Let $X_1 \sim N_d(\mu, \Sigma)$, $Y$ independent of $X_1$’s and $Y \sim N_d(\mu, \Sigma)$, where $\mu$, $\Sigma$, $\mu_x$, and $\Sigma_x$ satisfy the conditions in Theorem 1.

1. When $d$ is fixed, for $r = \text{op}(1)$, the probability $Y$ falls in the $r$-ball centered at $X_1$ is of order $\text{Op}(r^d).$
2. When $d$ increases with $n$, for $r = \text{O}(d^\beta)$, $\beta \leq \frac{1}{2}$, the logarithm of the probability $Y$ falls in the $r$-ball centered at $X_1$ is $-\frac{1}{2}d \log d + d \log r + \frac{1}{2}d(\log(1 + \text{Op}(1))).$ More specifically, when $\beta \leq -0.5$, the probability $Y$ falls in the $r$-ball centered at $X_1$ is of order $\text{Op}(r^{d-2d^\beta}).$

**Proof.** Under a special case that $\mu_x = 0$, $\Sigma_x = I$ and $X_1 \equiv 0$, the probability is

$$\int_0^r \frac{d \tau^{d/2}}{\Gamma(d/2 + 1)} \tau^{d-1} \frac{1}{2^{d/2}} \sqrt{\tau} e^{-\tau/2} d\tau$$

$$= \frac{d}{2^{d/2} \Gamma(d/2 + 1)} \int_0^\tau \sqrt{\tau} e^{-\tau/2} d\tau,$$

which is of order $2^{-d/2}e^{-d/2 \log d} + d/2 \int_0^\tau \tau^{d-1} e^{-\tau/2} d\tau = d^{1-d/2}e^{d/2} \int_0^\tau \tau^{d-1} e^{-\tau/2} d\tau.$$

For generic $\mu_x$, $\Sigma_x$ and $X_1 \sim N_d(\mu, \Sigma)$, notice that

$$\frac{f_{\mu_x, \Sigma_x}(Y)}{f_{\mu_x, \Sigma_x}(X_1)} = \frac{e^{-\frac{1}{2}(Y - \mu)^T \Sigma^{-1}(Y - \mu)}}{e^{-\frac{1}{2}(X_1 - \mu)^T \Sigma^{-1}(X_1 - \mu)}}$$

$$= e^{(X_1 - Y)^T \Sigma^{-1}(X_1 - \mu) - \frac{1}{2}(X_1 - Y)^T \Sigma^{-1}(X_1 - Y)}.$$ 

When $\|Y - X_1\|_2 = t$, based on the conditions in Theorem 1, there exists a positive function $c_1(t)$ and a constant $c_2$ such that $\frac{f_{\mu_x, \Sigma_x}(Y)}{f_{\mu_x, \Sigma_x}(X_1)} = \text{Op}(c_1(t)(e\sqrt{d}c_2^t))$. Then, probability $Y$ falls in the $r$-ball of $X_1$ is of order

$$d^{1-d/2}e^{d/2} \int_0^r t^{d-1} e^{-\frac{1}{2}t^2} \int |\Sigma|^{-1/2} e^{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)}$$

$$\times (2\pi)^{-d/2} |\Sigma|^{1/2} e^{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)} dx dt.$$ 

**Under the conditions of Theorem 1,** we have

$$\int |\Sigma|^{-1/2} e^{-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)} dx$$

$$= O \left( \frac{1}{|\Sigma|^{1/2}} \right)$$

$$= O(\sqrt{d}e^{d/2}).$$

Thus, the probability $Y$ falls in the $r$-ball of $X_1$ is of order

$$d^{1-d/2}e^{d/2} \int_0^r t^{d-1} e^{-\frac{1}{2}t^2} dt \int_0^r t^{d-1} e^{-\frac{1}{2}t^2} dt$$

$$= \int_0^r t^{d-1} e^{-\frac{1}{2}t^2} e^{d/2} \int_0^r t^{d-1} e^{-\frac{1}{2}t^2} dt.$$ 

We first consider the cases when $d$ increases with $n$. Suppose $t = c_0d^\beta$, for some fixed $\beta$ and $0 < c_0 \leq C$ for some constant $C > 0$. Then the integrand is

$$e^{-d(1/\beta)(\log d + \log c_0) + \log d(e\sqrt{d})}.$$ 

We consider the following two scenarios.

1. **If $\beta < 0$ or $0 < \beta \leq \frac{1}{2}$**, then $d \log d$ dominates the other terms. Furthermore, we have that

$$\frac{d \log d}{c_0d^\beta + 0.5} = \frac{d \log d}{c_0d} = \frac{1}{c_0}d^{1-2\beta} \log d \geq O(\log d).$$

2. **If $\beta = 0$**, we further consider the following two cases. Let $\varepsilon = \frac{\log d}{\sqrt{d}}$.

(a) **if $c_0 \geq 1 + \varepsilon$, then $d \log c_0$ dominates the other terms**, and we have that

$$\frac{d \log c_0}{c_0d^\beta} \geq O(\log d),$$

$$\frac{d \log c_0}{c_0} \geq O(\log d \sqrt{d}) \geq O(\log d).$$

(b) **if $c_0 \leq 1 - \varepsilon$, again $d \log c_0$ dominates the other terms**, and we have that

$$\frac{d \log c_0}{c_0d^\beta} \geq O(\log d),$$

$$\frac{d \log c_0}{c_0} \geq O(\log d \sqrt{d}) \geq O(\log d).$$
First of all, when $\beta \leq -0.5$, from Scenario (1), we have
\[
\frac{d \log d}{c_0 d^{\beta + 0.5}} \geq O(d \log d), \quad \text{and} \quad \frac{d \log d}{c_0 d^{\beta}} \geq O(d^2 \log d).
\]

Then,
\[
\begin{align*}
&d^{1-d/2} e^{d/2} \int_0^r t^{d-1} \phi(t) \sqrt{\beta} (c_2 + 0.5)t^2 \, dt \\
&= d^{1-d/2} e^{d/2} \int_0^r t^{d-1} \phi(t) \log (1 + O(\sqrt{\frac{1}{\sqrt{d}}})) \, dt \\
&= d^{1-d/2} e^{d/2} r^{d/2} \log \left( \frac{1}{\sqrt{d}} \right) = O(d^{1/2} d^{d/2}).
\end{align*}
\]

When $-0.5 < \beta < 0.5$, based on Scenarios (1) and (2), we have
\[
\begin{align*}
&d^{1-d/2} e^{d/2} \int_0^r t^{d-1} \phi(t) \log (1 + O(\sqrt{\frac{1}{\sqrt{d}}})) \, dt \\
&= d^{1-d/2} e^{d/2} \int_0^r t^{d-1} \phi(t) \log (1 + O(1/\log d)) \, dt \\
&= d^{1-d/2} e^{d/2} r^{d/2} O(\log d) = d^{-d/2} e^{d/2} (\kappa + O(\frac{\log d}{\log d})) \\
&= d^{-d/2} e^{d/2} (d + O(1)).
\end{align*}
\]

For Equation (9), the part of the integral from $1 - \epsilon_1$ to $1 + \epsilon_1$ is not an issue: Notice that $\int_{1-\epsilon_1}^{1+\epsilon_1} \sqrt{\beta} \, dt = (1 + \epsilon_1)\sqrt{\beta} - (1 - \epsilon_1)\sqrt{\beta} = e^{\phi(\epsilon_1)} - e^{\phi(-\epsilon_1)} = e^{\phi(\sqrt{\log d})}$, and $\int_{1-\epsilon_1}^{1+\epsilon_1} \sqrt{\beta} \, dt = \frac{(1 + \epsilon_1)\sqrt{\beta} - (1 - \epsilon_1)\sqrt{\beta}}{\sqrt{\beta}} = \frac{e^{\phi(\epsilon_1)} - e^{\phi(-\epsilon_1)}}{\sqrt{\beta}} = O(e^{\phi(\sqrt{\log d})})$ with $e^c = \sup_{1 - \epsilon_1 \leq t \leq 1 + \epsilon_1} \phi(t)$ a positive constant. Then, the difference between the two integrals is at most $e^{\phi(\sqrt{\log d})}$, which is much smaller than $e^{\phi(\sqrt{\log d})}$ and thus does not affect the above result.

When $d$ is fixed, the proofs are much simpler, and it is not hard to see that, when $r = o(1)$, the probability is of order $r^{d/2} e^{d/2}$.

\section*{Acknowledgments}
The authors thank the associate editor and two referees for their helpful constructive comments which have helped to improve the quality and presentation of the article. The authors also thank to the reviewer's suggestions on various numerical alternatives including the improved multivariate Shapiro–Wilk's test and the Fisher's test.

\section*{Funding}
The research of Hao Chen was supported in part by NSF Grants DMS-1513653 and DMS-1848579. The research of Yi Xu was supported in part by NSF grants 12022103, 11771094 and 11690013.

\section*{Supplementary Materials}
The supplementary materials contain the R codes used to reproduce the simulation results in the paper.

\section*{References}
Anderson, T. W. (2003), An Introduction To Multivariate Statistical Analysis (3rd ed.), New York: Wiley-Interscience. [723]
Baringhaus, L., and Henze, N. (1988), "A Consistent Test for Multivariate Normality Based on the Empirical Characteristic Function," Metrika, 35, 339–348. [719]
Bartoszyński, R., Pearl, D. K., and Lawrence, J. (1997), "A Multidimensional Goodness-of-Fit Test Based on Interpoint Distances," Journal of the American Statistical Association, 92, 577–586. [721]
Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L. (2013), "Valid Post-Selection Inference," The Annals of Statistics, 41, 802–837. [719]
Bickel, P. J., and Levina, E. (2004), "Some Theory for Fisher’s Linear Discriminant Function, Naive Bayes, and Some Alternatives When There are Many More Variables Than Observations," Bernoulli, 10, 989–1010. [719, 723]
Bickel, P. J., and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," The Annals of Statistics, 36, 199–227. [721]
Cai, T., and Liu, W. (2011), "A Direct Estimation Approach to Sparse Linear Discriminant Analysis," Journal of the American Statistical Association, 106, 1566–1577. [719, 720, 721, 722, 723, 724]
Cai, T. T., Liu, W., and Xia, Y. (2014), "Two-Sample Test of High Dimensional Means Under Dependency," Journal of Royal Statistical Society, Series B, 76, 349–372. [721]
Chan, H. P., and Walther, G. (2015), "Optimal Detection of Multi-Sample Aligned Sparse Signals," The Annals of Statistics, 43, 1865–1895. [719]
Chen, H. (2019), "Sequential Change-Point Detection Based on Nearest Neighbors," The Annals of Statistics, 47, 1381–1407. [721]
Chen, H., Chen, X., and Su, Y. (2018), "A Weighted Edge-Count Two-Sample Test for Multivariate and Object Data," Journal of the American Statistical Association, 113, 1146–1155. [721]
Chen, H., and Friedman, J. H. (2017), "A New Graph-Based Two-Sample Test for Multivariate and Object Data," Journal of the American Statistical Association, 112, 397–409. [720, 721]
Chen, H., and Zhang, N. (2015), "Graph-Based Change-Point Detection," The Annals of Statistics, 43, 139–176. [721]
Dettling, M. (2004), "Bagboosting for Tumor Classification With Gene Expression Data," Bioinformatics, 20, 3583–3593. [724]
Doornik, J. A., and Hansen, H. (2008), "An Omnibus Test for Univariate and Multivariate Normality," Oxford Bulletin of Economics and Statistics, 70, 927–939. [719, 720]
Fan, J., and Fan, Y. (2008), "High Dimensional Classification Using Features Annealed Independence Rules," Annals of Statistics, 36, 2605. [719, 720, 723, 724]
Fan, J., Feng, Y., and Wu, Y. (2009), "Network Exploration Via the Adaptive Lasso and Scad Penalties," The Annals of Applied Statistics, 3, 521. [719]
