The Annals of Statistics
2021, Vol. 49, No. 1, 411–434
https://doi.org/10.1214/20-AOS1962

CLASSIFICATION ACCURACY AS A PROXY FOR TWO-SAMPLE TESTING

BY ILMUN KIM[1,*], ADITYA RAMDAS[1,†], AARTI SINGH[1,‡] AND LARRY WASSERMAN[1,§]

[1]*Department of Statistics & Data Science, Carnegie Mellon University,* [*]*ilmunk@stat.cmu.edu;* [†]*aramdas@stat.cmu.edu;* [‡]*aarti@cs.cmu.edu;* [§]*larry@stat.cmu.edu*

When data analysts train a classifier and check if its accuracy is significantly different from chance, they are implicitly performing a two-sample test. We investigate the statistical properties of this flexible approach in the high-dimensional setting. We prove two results that hold for all classifiers in any dimensions: if its true error remains ϵ-better than chance for some $\epsilon > 0$ as $d, n \to \infty$, then (a) the permutation-based test is consistent (has power approaching to one), (b) a computationally efficient test based on a Gaussian approximation of the null distribution is also consistent. To get a finer understanding of the rates of consistency, we study a specialized setting of distinguishing Gaussians with mean-difference δ and common (known or unknown) covariance Σ, when $d/n \to c \in (0, \infty)$. We study variants of Fisher's linear discriminant analysis (LDA) such as "naive Bayes" in a nontrivial regime when $\epsilon \to 0$ (the Bayes classifier has true accuracy approaching 1/2), and contrast their power with corresponding variants of Hotelling's test. Surprisingly, the expressions for their power match exactly in terms of n, d, δ, Σ, and the LDA approach is only worse by a constant factor, achieving an asymptotic relative efficiency (ARE) of $1/\sqrt{\pi}$ for balanced samples. We also extend our results to high-dimensional elliptical distributions with finite kurtosis. Other results of independent interest include minimax lower bounds, and the optimality of Hotelling's test when $d = o(n)$. Simulation results validate our theory, and we present practical takeaway messages along with natural open problems.

1. Introduction. The recent popularity of machine learning has resulted in the extensive teaching and utilization of prediction methods in theoretical and applied communities. When faced with a hypothesis testing problem in practice, data scientists sometimes opt for a prediction-based test-statistic. We study one example of this common practice in this paper, concerning arguably the most classical testing and prediction problems—*two-sample testing* (are the two underlying distributions the same?) and *classification* (learning a classifier that separates the two distributions, implicitly assuming they are not the same). Practitioners familiar with machine learning but not the hypothesis testing literature often find it intuitive to perform testing in the following way: first learn a classifier, and then see if its accuracy is significantly different from chance and if it is, then conclude that the distributions are different.

The central question that this paper seeks to answer is "*what are the pros and cons of the classifier-based approach to two-sample testing?*" As we shall detail in Section 2, the notion of *cost* or *price* that is appropriate for the Neyman–Pearson or Fisherian hypothesis testing paradigm, is the power achievable at a fixed false positive level α (in other words, the lowest possible type-2 error achievable at some prespecified target type-1 error). Indeed, we approach this question using the frequentist perspective of minimax theory. More formally, we can restate our question as "*when is the classifier-based test consistent, and how does its power compare to the minimax power?*"

Received May 2019; revised February 2020.
MSC2020 subject classifications. Primary 62H15; secondary 62E20.
Key words and phrases. Classification accuracy, two sample testing, high-dimensional asymptotics, Hotelling's T^2 test, linear discriminant analysis, permutation test.

1.1. *Practical motivation.* Before we delve into the details, it is worth mentioning that even though this paper is a theoretical endeavor, the question was initially practically motivated. Many scientific questions are naturally posed as two-sample tests—examples abound in epidemiology and neuroscience. As a hypothetical example from the latter, say we are interested in determining whether a particular brain region responds differently under two situations (say listening to loud harsh sounds versus soft smooth sounds), or for a person with a medical condition (patient) and a person without the condition (control). Often, one collects and analyzes brain data for the same patient under the two contrasting stimuli (to study the effect of change in that stimulus), or for different normal and ill patients under the same stimulus (to study effect of a medical condition). Since the work of Golland and Fischl (2003) where the authors examined permutation tests for classification with application to neuroimaging analysis, it has been increasingly common in the field of neuroscience (see Etzel, Gazzola and Keysers (2009), Pereira, Mitchell and Botvinick (2009), Stelzer, Chen and Turner (2013), Zhu et al. (2008)) to assess whether there is a significant difference between the two sets of data collected by learning a classifier to differentiate between them (because, for instance, they may be more familiar with classification than two-sample testing). Neuroscientists call this style of brain decoding as pattern discrimination and a positive answer can be seen as preliminary evidence that the mental process of interest might occur within the portion of the brain being studied; see Olivetti, Greiner and Avesani (2012) for a discussion of related issues. This classification approach to two-sample testing has been considered in other application areas including genetics (Yu et al. (2007)), speech analysis (Chen et al. (2009)), credit scoring (Xiao et al. (2014)), churn prediction (Xiao et al. (2015)) and video content analysis (Liu, Li and Póczos (2018)).

1.2. *Overview of the main results.* Our first contribution is to identify weak conditions on the classifier that suffice for both finite-sample or asymptotic type-1 error control, as well as for asymptotic consistency.

- *Asymptotic test (Proposition 9.1)*: We identify mild conditions under which the sample-splitting error of a general classifier (2.4) is asymptotically Gaussian as $n, d \to \infty$. We introduce a test based on this Gaussian approximation and prove its asymptotic type-1 error control. We also prove that a sufficient condition for its consistency (for its power to approach one) is that its true accuracy converges to $1/2 + \epsilon$ for any constant $\epsilon > 0$ as $n, d \to \infty$ at any relative rate.
- *Permutation test (Theorem 9.1)*: In addition to the asymptotic approach, we consider two types of random permutation procedures that yield a valid level α test in finite-sample scenarios. Under the same conditions made before, we find the minimum number of permutations that guarantees that the resulting permutation test is consistent.

For technical reasons, it is convenient to present these results last, after suitable notation, lemmas and assumptions are developed in earlier sections.

The above results leave two natural questions open: first, whether we can derive a rate of consistency in special cases, and second, whether testing can be consistent even when the classifier accuracy asymptotically approaches chance (is not bounded away from half). We answer both affirmatively; our second contribution is to rigorously analyze the asymptotic power of tests using classification accuracy for Gaussian and elliptical distributions in a high-dimensional setting when the error of the Bayes optimal classifier approaches half. In this direction, we have three main results:

- *Power of the accuracy of LDA for Gaussian distributions with known* Σ *(Theorem 6.1)*: The considered test statistic (6.1) is the centered and rescaled classification error of LDA estimated via sample splitting, when Σ is known. Under standard interpretable assumptions

(Section 5.1), this test statistic converges to a standard normal in the high-dimensional setting (Theorem 5.1) under both null and local alternative. Using this fact, we describe its local asymptotic power in expression (6.7). Comparing the latter with the minimax power (3.3), we highlight that the performance of the accuracy test is comparable to but worse than the minimax optimal test, achieving an asymptotic relative efficiency (ARE) of $1/\sqrt{\pi} \approx 0.564$ for balanced sample sizes.

- *Extensions to unknown Σ using naive Bayes and other variants (Theorem 7.1)*: We generalize the previous findings to other linear classifiers for unknown Σ, like naive Bayes. We again find that classifier-based tests are underpowered, achieving the same aforementioned ARE of $1/\sqrt{\pi}$ compared to corresponding variants of Hotelling's test such as Bai and Saranadasa (1996) and Srivastava and Du (2008).
- *Extensions to elliptical distributions (Theorem 8.1)*: We extend Theorem 6.1 to the class of (heavy-tailed) elliptical distributions with finite kurtosis, and prove that the asymptotic power expression matches the Gaussian setting up to an explicit constant factor, which is $\sqrt{2}$ times the marginal density evaluated at 0. Restricting our attention to multivariate t-distributions, we also find an interesting phenomenon that the classifier-based test becomes relatively more efficient when the underlying distributions have heavier tails (lower degrees of freedom).

As two side contributions, we formally study the fundamental minimax power of high-dimensional two-sample mean testing for Gaussians. In this direction, we have two main results.

- *Explicit and exact expression for asymptotic minimax power (Proposition 3.1)*: By building on prior work (Luschgy (1982)), we provide an explicit expression for the asymptotic minimax power of high-dimensional two-sample mean testing that is valid for any (shared) positive definite covariance matrix and unbalanced sample sizes when $d, n \to \infty$ at any relative rate.
- *Minimax optimality of Hotelling's T^2 test when $d = o(n)$ (Theorem 4.1)*: It is well known that Hotelling's test is minimax optimal when d is fixed and $n \to \infty$. In the high-dimensional setting, when the dimension d and the sample size n both increase to infinity with $d/n \to c \in (0, 1)$, Bai and Saranadasa (1996) show that Hotelling's test may have low power. Since then, Hotelling's test has been largely undervalued in the setting where d increases with n. In contrast to the aforementioned negative result, we prove that Hotelling's test remains asymptotically minimax optimal when $d \to \infty$ as long as $d/n \to 0$.

1.3. *Interpreting our results and practical takeaway messages.* There may be two somewhat contradictory ways that our results may be interpreted:

1. Practitioners may (possibly unjustly) use our results to reassure themselves that their utilization of prediction methods for testing, even in the high-dimensional setting, may not hurt their power too much.
2. For scientific disciplines in which data is not abundant, scientists may be wary of using prediction methods for hypothesis testing problems due to the loss in power.

After our manuscript appeared on arXiv in early 2016, a few different papers have cited our results to justify their choices in both of these above ways. We take the liberty to weigh in on this possible conundrum, using our intuition from this paper and from experiments in other followup papers (e.g., Hediger, Michel and Näf (2019), Lopez-Paz and Oquab (2016), Rosenblatt et al. (2019)) to instead propose complementary, noncontradictory takeaway messages:

1. If the data is relatively unstructured or not abundant, and if the alternative can be accurately specified in such a manner that is both practically meaningful and for which a provably powerful two-sample test statistic is available (or can be easily designed), then we recommend using such a well-tailored statistic.

2. Suppose the data is highly structured or abundant (say, images of two species of beetles), but the potential differences between the two distributions cannot be easily specified. In this case, constructing a refined test that has high power against an accurately prespecified alternative may be too hard, and thereby we recommend using a flexible two-sample test statistic like classification accuracy (say using a convolutional neural network classifier or random forests).

Of course, it seems very challenging to theoretically study these setups in their full generality to provide a thorough formal backing to such practical suggestions. However, we are hopeful that our work will spur others to extend our concrete results to new settings.

1.4. *Related work.* The idea of using binary classifiers for two-sample testing was conceptualized by Friedman (2004). However, Friedman's proposal was fundamentally different from the one proposed here: he suggested using training a classifier on all points, and using that classifier to assign a score to each point. Then he compared the scores in each class using a univariate two-sample test like Mann–Whitney or Kolmogorov–Smirnov. In other words, Friedman proposed using classifiers to reduce a multivariate two-sample test into a univariate one. A different classifier-based approach to the two-sample problem was proposed by Blanchard, Lee and Scott (2010). Although their test is built upon classification algorithms, it estimates the a priori probability of a contamination model, instead of classification accuracy.

In contrast, this paper considers held-out accuracy as the test statistic. The held-out accuracy of any classifier in any dimension can be used as the test statistic, and type-1 error can always be controlled nonasymptotically at the desired level using permutations (see Section 9.2). Hence, the main question of genuine mathematical interest is what we can prove about the power of such a test. To overcome the computational burden of permutations, if we instead use a Gaussian approximation to the null distribution, then it is unclear whether it remains valid in the high-dimensional setting and again its power is unclear.

To the best of our knowledge, our 2016 ArXiv manuscript was the first mathematical attempt to study the power of this general approach in a specialized setting. There has been a growing interest in this idea in both the statistics and the machine learning communities (Borji (2019), Gagnon-Bartsch and Shem-Tov (2019), Hediger, Michel and Näf (2019), Lopez-Paz and Oquab (2016), Rosenblatt et al. (2019)), most of which directly build on our work, but further provide valuable practical insight into the problem using various classifiers under different scenarios. However, most of these other works couple informal heuristic arguments with numerical experiments, motivating us to fully formalize and further generalize our earlier analysis.

In an orthogonal work, Scott and Nowak (2005) proposed a Neyman–Pearson classification framework within which one would like to minimize the probability of classification error for one class, subject to a bound on the probability of classification error for the other class. Their problem is a variant of classification in which the classifier is judged by a different error metric, but it is quite different from our goal of two-sample testing. Other connections between classification and two-sample testing have also been explored by Ben-David et al. (2007), Sriperumbudur et al. (2009) and Gretton et al. (2012), but none of them set out to solve our problem.

Another class of two-sample tests is based on geometric graphs; examples include the k-nearest neighbor (NN) graph (Henze (1988), Schilling (1986)), the minimum spanning tree (Friedman and Rafsky (1979)) and the cross-matching (Rosenbaum (2005)). Recently,

Bhattacharya (2020) presented general asymptotic properties of graph-based tests under the fixed dimensional setting. Comparing the performance of the k-NN graph test and the k-NN classifier test (based on its held-out classification accuracy, as studied in this paper) may be interesting to explore in future work.

There is of course a very large body of work that just analyzes classifiers, or just analyzes two-sample tests (e.g., Arias-Castro, Pelletier and Saligrama (2018), Hu and Bai (2016), and the references therein), but without connecting the two.

Paper outline. The rest of this paper is organized as follows. In Section 2, we formally define both testing and classification problems. In Section 3, we discuss a minimax lower bound for two-sample testing in high-dimensional settings and in Section 4, we prove that Hotelling's T^2 test achieves this lower bound when $d/n \to 0$. Section 5 studies the limiting distribution of Fisher's LDA accuracy in the high-dimensional setting. Building on this limiting distribution, Section 6 presents the asymptotic power of Fisher's LDA for two-sample mean testing under known Σ. Section 7 extends this asymptotic power expression to other linear classifiers with unknown Σ, like naive Bayes. Generalizations to elliptical distributions are in Section 8. In Section 9, we examine the type-1 error control and consistency of the asymptotic test as well as the permutation test for *any* classifier. In Section 10, we provide simulation results that confirm our theoretical analysis, before concluding in Section 11. The proofs of all the results along with the discussion on open problems are provided in the Supplementary Material (Kim et al. (2020)).

Notation. Let $\mathcal{N}_d(\mu, \Sigma)$ refer to the d-variate Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and $d \times d$ positive definite covariance matrix Σ. With a slight abuse of notation, we sometimes use $\mathcal{N}_d(z; \mu, \Sigma)$ to denote the corresponding density evaluated at z. The symbol $\|\cdot\|$ refers to the L_2 norm. Let $\mathbb{I}[\cdot]$ denote the standard 0-1 indicator function. Let $\Phi(\cdot)$ denote the standard Gaussian CDF, and let z_α be its upper $1-\alpha$ quantile. For a square matrix A, let $\operatorname{diag}(A)$ denote the diagonal matrix formed by zeroing out the off-diagonal entries of A, and let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ be the minimum and the maximum eigenvalues of A. We write the identity matrix as I. For sequences of constants a_n and b_n, we write $a_n = O(b_n)$ if there exists a universal constant c such that $|a_n/b_n| \le c$ for all n larger than some n_0, and we write $a_n = o(b_n)$ if $a_n/b_n \to 0$. Similarly, for a sequence of random variables X_n and constants a_n, we write $X_n = O_P(a_n)$ if $a_n^{-1}X_n$ is stochastically bounded and $X_n = o_P(a_n)$ if $a_n^{-1}X_n$ converges to zero in probability.

2. Background. In this section, we introduce two-sample testing, including the special case of two-sample mean testing using Hotelling-type statistics and Fisher's linear discriminant analysis (LDA). We only introduce the basic versions here, later introducing variants like naive Bayes. We will be working in the high-dimensional setting where the number of samples n and dimension d can both increase to infinity simultaneously.

2.1. *Two-sample testing.* Suppose that $X_1, \dots, X_{n_0}, Y_1, \dots, Y_{n_1}$ are independent random vectors in $\mathbb{R}^d$ such that $\mathcal{X}_1^{n_0} \stackrel{\text{def}}{=} \{X_1, \dots, X_{n_0}\}$ are identically distributed with the distribution $\mathbb{P}_0$ and $\mathcal{Y}_1^{n_1} \stackrel{\text{def}}{=} \{Y_1, \dots, Y_{n_1}\}$ are identically distributed with the distribution $\mathbb{P}_1$. Given these samples, the two-sample problem aims at testing whether

$$H_0 : \mathbb{P}_0 = \mathbb{P}_1 \quad \text{vs.} \quad H_1 : \mathbb{P}_0 \neq \mathbb{P}_1. \tag{2.1}$$

While some of our results are on general classifiers and distributions (Section 9), we often focus on the specific case where $\mathbb{P}_0$ and $\mathbb{P}_1$ are d-variate Gaussian distributions with densities $p_0(x) \stackrel{\text{def}}{=} \mathcal{N}_d(x; \mu_0, \Sigma)$ and $p_1(y) \stackrel{\text{def}}{=} \mathcal{N}_d(y; \mu_1, \Sigma)$, respectively. We discuss the extension to

heavy-tailed elliptical distributions in Section 8. When the Gaussians have equal covariance, the previous problem boils down to testing whether two distributions have the same mean vector or not. This two-sample mean testing is a fundamental decision-theoretic problem, having a long history in statistics; for example, the past century has seen a wide adoption of the T^2-statistic by Hotelling (1931) to decide if two-samples have different population means (see Hu and Bai (2016) for a review). Given the sample mean vectors $\widehat{\mu}_0 \stackrel{\text{def}}{=} \sum_{i=1}^{n_0} X_i/n_0$ and $\widehat{\mu}_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n_1} Y_i/n_1$ and the pooled sample covariance matrix

$$\widehat{\Sigma} \stackrel{\text{def}}{=} \frac{1}{n_0+n_1-2}\left[\sum_{i=1}^{n_0}(X_i-\widehat{\mu}_0)(X_i-\widehat{\mu}_0)^\top + \sum_{i=1}^{n_1}(Y_i-\widehat{\mu}_1)(Y_i-\widehat{\mu}_1)^\top\right],$$

Hotelling's T^2-statistic is given by

$$T_H = (\widehat{\mu}_0-\widehat{\mu}_1)^\top \widehat{\Sigma}^{-1}(\widehat{\mu}_0-\widehat{\mu}_1).$$

Hotelling's T^2 test based on T_H was introduced for Gaussians, but it has been generalized to non-Gaussian settings as well (e.g., Kariya (1981)).

2.2. *Held-out classification accuracy.* Consider the same distributional setting described in the previous section. Given the samples $\mathcal{X}_1^{n_0}$ and $\mathcal{Y}_1^{n_1}$, classification is the problem of predicting to which class a new observation Z belongs, that is, we want to predict whether Z came from $\mathbb{P}_0$ or $\mathbb{P}_1$. Let the samples from $\mathbb{P}_0$ and $\mathbb{P}_1$ be given labels 0 and 1, respectively. A classifier C is a function that maps a datapoint Z to $\{0,1\}$. Define the *conditional* error of a classifier C trained on the labeled data as

$$\begin{aligned}
\mathcal{E} &\stackrel{\text{def}}{=} (\mathcal{E}_0+\mathcal{E}_1)/2 \quad \text{where}\\
\mathcal{E}_0 &\stackrel{\text{def}}{=} \Pr_{Z\sim\mathbb{P}_0}\left(C(Z)=1 \mid \mathcal{X}_1^{n_0}, \mathcal{Y}_1^{n_1}\right),\\
\mathcal{E}_1 &\stackrel{\text{def}}{=} \Pr_{Z\sim\mathbb{P}_1}\left(C(Z)=0 \mid \mathcal{X}_1^{n_0}, \mathcal{Y}_1^{n_1}\right).
\end{aligned} \tag{2.2}$$

Clearly, $\mathcal{E}$ is a random variable that depends on the input data. Next, define the *unconditional* error of C as

$$\begin{aligned}
E &\stackrel{\text{def}}{=} (E_0+E_1)/2 \quad \text{where}\\
E_0 &\stackrel{\text{def}}{=} \mathbb{E}_{n_0,n_1}\left[\Pr_{Z\sim\mathbb{P}_0}\left(C(Z)=1 \mid \mathcal{X}_1^{n_0}, \mathcal{Y}_1^{n_1}\right)\right],\\
E_1 &\stackrel{\text{def}}{=} \mathbb{E}_{n_0,n_1}\left[\Pr_{Z\sim\mathbb{P}_1}\left(C(Z)=0 \mid \mathcal{X}_1^{n_0}, \mathcal{Y}_1^{n_1}\right)\right],
\end{aligned} \tag{2.3}$$

where $\mathbb{E}_{n_0,n_1}$ denotes the expectation with respect to the n_0 and n_1 labeled datapoints. Note that E, E_0, E_1 do not depend on the input data and are only functions of d, δ, Σ, $n \stackrel{\text{def}}{=} n_0+n_1$. Importantly, if $\mathbb{P}=\mathbb{Q}$, chance performance is always $E=1/2$, no matter the ratio of sample sizes from each class (hence predicting the dominant label also achieves accuracy half).

Even though E is unknown, one can estimate E in a few different ways. One simple way is via sample splitting where the samples are split into training and test sets. Let us denote the number of samples of each class in the training (or test) set by $n_{0,\text{tr}}$ and $n_{1,\text{tr}}$ (or $n_{0,\text{te}}$ and $n_{1,\text{te}}$). In other words, there are $n_{\text{tr}} \stackrel{\text{def}}{=} n_{0,\text{tr}}+n_{1,\text{tr}}$ samples in the training set and $n_{\text{te}} \stackrel{\text{def}}{=} n_{0,\text{te}}+n_{1,\text{te}}$ samples in the test set. We then learn a classifier $\widehat{C}$ using n_{tr} samples, and

estimate its sample-splitting error using the remaining n_{te} samples as

$$\begin{aligned}
\widehat{E}^S &\stackrel{\text{def}}{=} (\widehat{E}_0^S + \widehat{E}_1^S)/2 \quad \text{where} \\
\widehat{E}_0^S &\stackrel{\text{def}}{=} \frac{1}{n_{0,\text{te}}} \sum_{i=1}^{n_{0,\text{te}}} \mathbb{I}\big[\widehat{C}(X_{n_{0,\text{tr}}+i}) = 1\big], \\
\widehat{E}_1^S &\stackrel{\text{def}}{=} \frac{1}{n_{1,\text{te}}} \sum_{i=1}^{n_{1,\text{te}}} \mathbb{I}\big[\widehat{C}(Y_{n_{1,\text{tr}}+i}) = 0\big].
\end{aligned} \tag{2.4}$$

It is clear that the classifier will have a true accuracy significantly above half only if $\mathbb{P} \neq \mathbb{Q}$. Hence one can use $\widehat{E}^S$ as a test statistic for two-sample testing, by checking whether $\widehat{E}^S$ is significantly less than half. The power of this approach is examined in Section 9, but we begin with the special case of mean-testing using linear discriminant analysis.

2.3. *Fisher's linear discriminant classifier.* In the Gaussian setting, the optimal classifier is given by Bayes rule:

$$\mathbb{I}\left[\log \frac{p_1(Z)}{p_0(Z)} > 0\right] = \mathbb{I}\left[(\mu_1 - \mu_0)^\top \Sigma^{-1}\left(Z - \frac{(\mu_0 + \mu_1)}{2}\right) > 0\right].$$

We denote $\delta \stackrel{\text{def}}{=} \mu_1 - \mu_0$ and $\mu_{\text{pool}} \stackrel{\text{def}}{=} (\mu_0 + \mu_1)/2$ so that we can succinctly write the Bayes rule as

$$C_{\text{Bayes}}(Z) \stackrel{\text{def}}{=} \mathbb{I}\big[\delta^\top \Sigma^{-1}(Z - \mu_{\text{pool}}) > 0\big]. \tag{2.5}$$

Then, by plugging in the estimators $\widehat{\delta} \stackrel{\text{def}}{=} \widehat{\mu}_1 - \widehat{\mu}_0$, $\widehat{\mu}_{\text{pool}} \stackrel{\text{def}}{=} (\widehat{\mu}_0 + \widehat{\mu}_1)/2$, and some appropriate choice of $\widehat{\Sigma}$, the linear discriminant analysis (LDA) rule is given by

$$\text{LDA}(Z) \stackrel{\text{def}}{=} \mathbb{I}\big[\widehat{\delta}^\top \widehat{\Sigma}^{-1}(Z - \widehat{\mu}_{\text{pool}}) > 0\big].$$

This classifier was derived by Fisher (1936, 1940) from a generalized eigenvalue problem (hence also called Fisher's LDA) and was later developed further by Wald (1944) and Anderson (1951). We will show that the held-out accuracy of Fisher's LDA in the high-dimensional Gaussian setting is asymptotically Gaussian, and derive its power when used for two-sample testing (for various choices of $\widehat{\Sigma}$). We later extend these results to heavy-tailed ellpitical distributions. However, we begin by understanding the fundamental minimax lower bounds for two-sample mean testing.

3. Lower bounds for two-sample mean testing. We first introduce some notation. Let $\mathcal{P}$ be a set that consists of all pairs of d-dimensional multivariate normal density functions whose covariance matrices coincide, and is positive definite. Let $\mathcal{P}_0$ be the subset of $\mathcal{P}$ such that each pair also has the same mean. For a given $\alpha \in (0, 1)$, let us write a level α test based on $\mathcal{X}_1^{n_0}$ and $\mathcal{Y}_1^{n_1}$ by φ_α and the collection of all level α tests by

$$\mathcal{T}_\alpha \stackrel{\text{def}}{=} \Big\{\varphi_\alpha : \mathcal{X}_1^{n_0} \cup \mathcal{Y}_1^{n_1} \mapsto \{0, 1\} : \sup_{p_0, p_1 \in \mathcal{P}_0} \mathbb{E}_{p_0, p_1}[\varphi_\alpha] \leq \alpha\Big\}.$$

Additionally, we define a class of two multivariate normal density functions p_0 and p_1 whose distance is measured in terms of Mahalanobis distance parameterized by $\rho > 0$ as

$$\mathcal{P}_1(\rho) \stackrel{\text{def}}{=} \big\{(p_0, p_1) \in \mathcal{P} : (\mu_0 - \mu_1)^\top \Sigma^{-1}(\mu_0 - \mu_1) \geq \rho^2\big\}.$$

The use of Mahalanobis distance is conventional and has been considered in Giri, Kiefer and Stein (1963), Giri and Kiefer (1964) and Salaevskii (1971) to study the minimax character of Hotelling's one-sample test. The "oracle" Hotelling's two sample test is defined as

$$\varphi_H^* = \mathbb{I}\left[\frac{n_0 n_1}{n_0+n_1}(\widehat{\mu}_0-\widehat{\mu}_1)^\top \Sigma^{-1}(\widehat{\mu}_0-\widehat{\mu}_1) \geq c_{\alpha,d}\right],$$

where $c_{\alpha,d}$ is the $1-\alpha$ quantile of the chi-squared distribution with d degrees of freedom, and "oracle" signifies that Σ is known. Luschgy (1982) extends the previous one-sample results and shows that φ_H^* is minimax optimal over $\mathcal{P}_1(\rho)$, or more explicitly,

$$\sup_{\varphi_\alpha \in \mathcal{T}_\alpha} \inf_{p_0,p_1 \in \mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_\alpha] = \inf_{p_0,p_1 \in \mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_H^*], \tag{3.1}$$

for any finite n and d. However, this result does not clearly show how the underlying parameters (e.g., n, d, ρ) interact to determine the power. To shed light on this, we study the asymptotic expression for the minimax power. Denote the sample size ratio by $\lambda_1 = \lambda_{1,n} \stackrel{\text{def}}{=} n_1/n$. Recalling that Φ is the standard normal CDF and z_α its $1-\alpha$ quantile, we prove the following.

PROPOSITION 3.1. *Consider a high-dimensional regime where $n, d \to \infty$ (at any rate). Then the minimax power for Gaussian two-sample mean testing is*

$$\begin{aligned}&\sup_{\varphi_\alpha \in \mathcal{T}_\alpha} \inf_{p_0,p_1 \in \mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_\alpha] \\ &= \Phi\left(-\frac{\sqrt{2d}}{\sqrt{2d+n\lambda_1(1-\lambda_1)\rho^2}} z_\alpha + \frac{n\lambda_1(1-\lambda_1)\rho^2}{\sqrt{2d+4n\lambda_1(1-\lambda_1)\rho^2}}\right) + o(1).\end{aligned} \tag{3.2}$$

The proof is based on the central limit theorem and can be found in Appendix C.2. Notably, the expression (3.2) is asymptotically precise including all constant terms and is valid without any restrictions on d/n and λ_1. The way to interpret the bound in (3.2) is as follows. The first term inside the parentheses is not of interest for our purposes, its magnitude being bounded by the constant z_α. The second term is what determines the rate at which the power approaches one. When $\rho = 0$, the power reduces to $\Phi(-z_\alpha) = \alpha$ and if d and n are thought of as fixed, larger ρ leads to larger power. The key in high dimensions, however, is how the power depends jointly on the signal to noise ratio (SNR) ρ, the dimension d and the sample size n. To see this clearer, in the low SNR regime where $\rho^2 = o(d/n)$ and $\lambda_1 \to \lambda \in (0,1)$, the minimax lower bound simplifies to

$$\Phi\left(-z_\alpha + \frac{n\lambda(1-\lambda)\rho^2}{\sqrt{2d}}\right) + o(1). \tag{3.3}$$

It can be already seen that at constant SNR, n only needs to scale faster than $\sqrt{d}$ for test power to asymptotically approach unity—this $\sqrt{d}/n$ scaling is unlike the d/n scaling typically seen in prediction problems (for prediction error or classifier recovery, see Raudys and Young (2004)). Next, we prove that this lower bound is tight even when Σ is unknown, as long as $d = o(n)$.

4. Minimax optimality of Hotelling's test when $d = o(n)$. When Σ is unknown, φ_H^* is not implementable, and thus it remains unclear whether the previous asymptotic lower bound is tight. In other words, we do not know whether there exists a test that has the same asymptotic minimax power as φ_H^* in all high-dimensional regimes with unknown Σ. Below, we partially close this gap by showing that Hotelling's test with unknown Σ can achieve

the same asymptotic minimax power as φ_H^* when $d/n \to 0$. By letting $q_{\alpha,n,d}$ be the $1-\alpha$ quantile of the F distribution with parameters d and $n-1-d$, Hotelling's two-sample test with unknown Σ is given by

$$\varphi_H = \mathbb{I}\left[\frac{n_0 n_1 (n-d-1)}{n(n-2)d}(\widehat{\mu}_0 - \widehat{\mu}_1)^\top \widehat{\Sigma}^{-1}(\widehat{\mu}_0 - \widehat{\mu}_1) \geq q_{\alpha,n,d}\right].$$

For Gaussians, it is well known that φ_H satisfies $\sup_{p_0,p_1\in\mathcal{P}_0} \mathbb{E}_{p_0,p_1}[\varphi_H] \leq \alpha$ (e.g., Anderson (1958)). The next theorem studies the power of φ_H.

THEOREM 4.1. *Consider an asymptotic regime where $d/n \to 0$. Then the uniform power of φ_H is asymptotically the same as that of φ_H^* for Gaussian two-sample mean testing. In other words, as $n, d \to \infty$ with $d/n \to 0$, we have that* $\inf_{p_0,p_1\in\mathcal{P}_1(\rho)} \mathbb{E}_{p_0,p_1}[\varphi_H]$ *is equal to*

$$\Phi\left(-\frac{\sqrt{2d}}{\sqrt{2d + n\lambda_1(1-\lambda_1)\rho^2}} z_\alpha + \frac{n\lambda_1(1-\lambda_1)\rho^2}{\sqrt{2d + 4n\lambda_1(1-\lambda_1)\rho^2}}\right) + o(1).$$

The proof can be found in Appendix C.3. When $d > n$, T_H is not even well defined, but Bai and Saranadasa (1996) demonstrate that even when $d/n \to c \in (0,1)$ the power of φ_H is poor. Due to its limitations, Hotelling's test has been largely neglected when d increases with n. Unlike the previous negative results, Theorem 4.1 shows that it is minimax optimal when d is allowed to grow with n, but $d/n \to 0$. We also provide empirical support for our asymptotic results in Figure 4 of Section 10.3.

REMARK 4.1. Combining the previous theorem with Bai and Saranadasa (1996) and our simulation results in Section 10.3, we may describe the phase transition behavior of Hotelling's test with unknown Σ as:

- optimal regime (same power as φ_H^*): $d/n \to 0$,
- suboptimal regime (lower power than φ_H^*): $d/n \to c \in (0,1)$,
- not applicable: $d/n \to c \geq 1$.

Even though Hotelling's test is suboptimal when $d = O(n)$, it is still an open problem to determine whether the lower bound is achievable by some other test, or whether a stronger lower bound can be proved.

5. Asymptotic normality of the accuracy of generalized LDA. Here, we investigate the high-dimensional limiting distribution of the sample-splitting error in (2.4). Building on the results developed in this section, we will present the power of the classification test in Section 6. Our main interest is in the setting where the dimension is comparable to or potentially much larger than the sample size. In this high-dimensional scenario, Bickel and Levina (2004) prove that Fisher's LDA performs poorly in classification problems. When $d > n$, Fisher's LDA classifier is not even well defined since $\widehat{\Sigma}$ is not invertible. Thus, Bickel and Levina (2004) consider the naive Bayes (NB) classification rule by replacing $\widehat{\Sigma}^{-1}$ with the inverse of $\mathrm{diag}(\widehat{\Sigma})$ and show that it outperforms Fisher's LDA in the high-dimensional setting. In the context of two-sample testing, we encounter the same issue on $\widehat{\Sigma}$ as mentioned earlier. To simplify our analysis, we start by assuming that Σ is *known* and analyze the asymptotic behavior of the corresponding Fisher's LDA statistic. Later in Section 7, we extend the results to *unknown* Σ by considering the NB classifier and others.

5.1. *Assumptions*. Recalling that we work in the high-dimensional Gaussian setting with common covariance, let us detail some assumptions that facilitate our analysis. We assume that as $n = n_0 + n_1 \to \infty$, we have:

(A1) *High-dimensional asymptotics*: $\exists c \in (0, \infty)$ such that $d/n \to c$.
(A2) *Local alternative*: $\delta^\top \Sigma^{-1} \delta = O(n^{-1/2})$.
(A3) *Sample size ratio*: there exists $\lambda \in (0, 1)$ such that $n_0/n \to \lambda$.
(A4) *Sample splitting ratio*: there exists $\kappa \in (0, 1)$ such that $n_{\mathrm{tr}}/n \to \kappa$.

The asymptotic regime in *(A1)* is called *Raudys–Kolmogorov double asymptotics* (e.g., Zollanvari, Braga-Neto and Dougherty (2011)) and assumes that d increases linearly with n. In (A2), we assume that $\delta^\top \Sigma^{-1} \delta$ is close to zero such that a minimax test has nontrivial power. Note that under (A1), the low SNR regime $\delta^\top \Sigma^{-1} \delta = o(d/n)$ is implied by (A2). It is also interesting to note that the classification error of the Bayes optimal classifier (2.5) is computed as

$$\frac{1}{2} \Pr_{Z \sim \mathbb{P}_0} \{C_{\mathrm{Bayes}}(Z) = 1\} + \frac{1}{2} \Pr_{Z \sim \mathbb{P}_1} \{C_{\mathrm{Bayes}}(Z) = 0\} = 1 - \Phi\left(\frac{\sqrt{\delta^\top \Sigma^{-1} \delta}}{2} \right),$$

which means that the classification error of the Bayes classifier, and hence *any* classifier, approaches chance under (A2). Assumption (A3) rules out highly imbalanced cases and is common in the two-sample literature (e.g., Bai and Saranadasa (1996), Chen and Qin (2010), Srivastava, Katayama and Kano (2013)). (A4) assumes that the user-chosen sample-splitting ratio is within $(0, 1)$. We show in Theorem 6.1 that the asymptotic power of the test based on held-out classification accuracy is maximized when $\kappa = 1/2$ for the balanced case of $\lambda = 1/2$. In other cases, Theorem 6.1 may serve as a guideline for choosing κ that maximizes the asymptotic power. For any $d \times d$ symmetric positive definite matrix A, we define the generalized LDA classifier by

$$\mathrm{LDA}_{A, n_0, n_1}(Z) \stackrel{\mathrm{def}}{=} \mathbb{I}\big[\widehat{\delta}^\top A (Z - \widehat{\mu}_{\mathrm{pool}}) > 0\big]. \tag{5.1}$$

Its sample-splitting error can be calculated using expression (2.4):

$$\widehat{E}^S_A \equiv \text{classification error of } \mathrm{LDA}_{A, n_{0,\mathrm{tr}}, n_{1,\mathrm{tr}}}(Z),$$

emphasizing the dependency on the user-chosen matrix A. In terms of Σ and A, we assume that:

(A5) Σ *has bounded eigenvalues*: there exist constants c_1, c_2 such that $0 < c_1 \le \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) \le c_2 < \infty$.
(A6) A *has bounded eigenvalues*: there exist constants c_1', c_2' such that $0 < c_1' \le \lambda_{\min}(A) \le \lambda_{\max}(A) \le c_2' < \infty$.

The same eigenvalue condition for Σ was used by Bickel and Levina (2004). Assumption (A6) is satisfied when A is diagonal with uniformly bounded entries, and when $A = \Sigma^{-1}$ under (A5).

5.2. *Asymptotic normality for nonrandom A*. Given the previous assumptions, we study the asymptotic distribution of the sample-splitting error of the generalized LDA classifier when A is nonrandom. Since Fisher's LDA with known Σ is a special case of generalized LDA, it is straightforward to derive the limiting distribution of $\widehat{E}^S_{\Sigma^{-1}}$ from the general result.

We first observe that the sample-splitting error of the generalized LDA classifier can be viewed as the average of independent observations when conditioning on the training set. Therefore, it is natural to expect that the sample-splitting error is asymptotically normally distributed. To make this statement formal, we define $\mathcal{E}_{i,A}$ and $E_{i,A}$ similarly as $\mathcal{E}_i$ and E_i

for $i = 1, 2$ from definitions (2.2) and (2.3), but by replacing the LDA classifier with the generalized LDA classifier with a given A. Then let us write the standardized test statistic as

$$W_A \stackrel{\text{def}}{=} \frac{\widehat{E}_A^S - \mathcal{E}_{0,A}/2 - \mathcal{E}_{1,A}/2}{\sqrt{\mathcal{E}_{0,A}(1-\mathcal{E}_{0,A})/(4n_{0,\text{te}}) + \mathcal{E}_{1,A}(1-\mathcal{E}_{1,A})/(4n_{1,\text{te}})}}. \tag{5.2}$$

In the next proposition, we present both *conditional* and *unconditional* limiting distributions of W_A in the high-dimensional setting.

PROPOSITION 5.1. *Suppose that the assumptions* (A1)–(A6) *hold. Then W_A converges to a standard Gaussian conditional on the training set*:

$$\sup_{t\in\mathbb{R}} \left|\Pr\left(W_A \le t \,|\, \mathcal{X}_1^{n_{0,\text{tr}}}, \mathcal{Y}_1^{n_{1,\text{tr}}}\right) - \Phi(t)\right| = O_P\left(n^{-1/2}\right).$$

Moreover, under the same assumptions, W_A converges to the standard normal distribution unconditional on the training set:

$$\sup_{t\in\mathbb{R}} \left|\Pr(W_A \le t) - \Phi(t)\right| = o(1).$$

The proof is given in Appendix C.4. Although the limiting distribution of W_A is known from the previous lemma, it is quite challenging to determine the power of a test based classification accuracy by analyzing W_A. The reason is that $\mathcal{E}_{0,A}$ and $\mathcal{E}_{1,A}$ are random since they depend on the training set. To address this issue, we shall present a tractable approximation of W_A that replaces $\mathcal{E}_{0,A}$ and $\mathcal{E}_{1,A}$ with nonrandom quantities. To ease notation, let us denote $V_{0,A} \stackrel{\text{def}}{=} \widehat{\delta}^\top A(\mu_0 - \widehat{\mu}_{\text{pool}})$, $V_{1,A} \stackrel{\text{def}}{=} \widehat{\delta}^\top A(\widehat{\mu}_{\text{pool}} - \mu_1)$ and $U_A \stackrel{\text{def}}{=} \widehat{\delta}^\top A \Sigma A \widehat{\delta}$. We would like to stress that $\widehat{\delta}$ and $\widehat{\mu}_{\text{pool}}$ are computed based only on the training set. Using this fact, $\mathcal{E}_{0,A}$ and $\mathcal{E}_{1,A}$ can be written as

$$\mathcal{E}_{0,A} = \Phi\left(\frac{V_{0,A}}{\sqrt{U_A}}\right) \quad \text{and} \quad \mathcal{E}_{1,A} = \Phi\left(\frac{V_{1,A}}{\sqrt{U_A}}\right). \tag{5.3}$$

Further write the expectations of $V_{0,A}$, $V_{1,A}$ and U_A by $\mathbb{E}[V_{0,A}] = \Psi_{A,n,d} + \Xi_{A,n,d}$, $\mathbb{E}[V_{1,A}] = \Psi_{A,n,d} - \Xi_{A,n,d}$ and $\mathbb{E}[U_A] = \Lambda_{A,n,d}$ where

$$\begin{aligned}
\Psi_{A,n,d} &\stackrel{\text{def}}{=} -\frac{1}{2}\delta^\top A \delta, \\
\Lambda_{A,n,d} &\stackrel{\text{def}}{=} \delta^\top A \Sigma A \delta + \left(\frac{1}{n_{0,\text{tr}}} + \frac{1}{n_{1,\text{tr}}}\right)\text{tr}\{(A\Sigma)^2\} \quad \text{and} \\
\Xi_{A,n,d} &\stackrel{\text{def}}{=} \frac{1}{2}\left(\frac{1}{n_{0,\text{tr}}} - \frac{1}{n_{1,\text{tr}}}\right)\text{tr}(A\Sigma).
\end{aligned} \tag{5.4}$$

Here, the first two terms $\Psi_{A,n,d}$ and $\Lambda_{A,n,d}$ can be viewed as signal and noise terms, respectively, which ultimately determine the asymptotic power of the accuracy test. The third term $\Xi_{A,n,d}$ is an extra variance that comes from unbalanced sample sizes. Finally, we define a scaling factor

$$\gamma_{A,n,d} \stackrel{\text{def}}{=} 2\sqrt{\frac{n_{0,\text{te}} n_{1,\text{te}}}{n_{0,\text{te}} + n_{1,\text{te}}}} \frac{1}{\sqrt{\Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\{1 - \Phi(\Xi_{A,n,d}/\sqrt{\Lambda_{A,n,d}})\}}}. \tag{5.5}$$

With this notation in hand and letting $\phi(\cdot)$ be the standard normal density function, we now introduce an approximation of W_A defined as

$$W_A^\dagger \stackrel{\text{def}}{=} \gamma_{A,n,d} \cdot \left\{\widehat{E}_A^S - \frac{1}{2} - \phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right\}.$$

It is clear that $W_A^\dagger$ is centered and scaled by explicit and nonrandom quantities. Next, we show that the difference between W_A and $W_A^\dagger$ is asymptotically negligible and, therefore, $W_A^\dagger$ is also asymptotically standard normal.

THEOREM 5.1. *Suppose that the assumptions* (A1)–(A6) *hold. Then we have that* $W_A = W_A^\dagger + o_P(1)$ *and thus the distribution of* $W_A^\dagger$ *converges to a standard normal*:

$$\sup_{t\in\mathbb{R}} |\Pr(W_A^\dagger \le t) - \Phi(t)| = o(1).$$

The proof of Theorem 5.1 can be found in Appendix C.5. The asymptotic normality, established in the above theorem, holds under the null as well as under the local alternative (A2). This enables us to explore the asymptotic power of the generalized LDA test with known Σ in the next section, and we deal with unknown Σ in the following section.

6. Asymptotic power of generalized LDA with nonrandom A. Here, we study the asymptotic power of the generalized LDA test for known Σ. Since a smaller value of $\widehat{E}_A^S - 1/2$ (or equivalently a larger value of the average per-class accuracy $1 - \widehat{E}_A^S$) is in favor of $H_1 : \mu_0 \neq \mu_1$, we define the test function by

$$\varphi_A \stackrel{\text{def}}{=} \mathbb{I}\left[\gamma_{A,n,d}\left(\widehat{E}_A^S - \frac{1}{2}\right) < -z_\alpha\right]. \tag{6.1}$$

It is then clear from Theorem 5.1 that φ_A has an asymptotic type-1 error controlled by α. Now under the local alternative hypothesis, φ_A has power given by

$$\begin{aligned}\mathbb{E}[\varphi_A] &= \Pr\left(W_A^\dagger < -z_\alpha - \gamma_{A,n,d}\cdot\phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\\ &= \Phi\left(-z_\alpha - \gamma_{A,n,d}\cdot\phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right) + o(1),\end{aligned} \tag{6.2}$$

where the second equality uses Theorem 5.1. Let us write

$$\beta_{A,\lambda,\kappa} \stackrel{\text{def}}{=} \frac{\lambda - 1/2}{\sqrt{\lambda(1-\lambda)\kappa}}\frac{n^{-1}\operatorname{tr}(A\Sigma)}{\sqrt{n^{-1}\operatorname{tr}\{(A\Sigma)^2\}}}. \tag{6.3}$$

Using assumptions (A1)–(A6), the main term in the power function (6.2) simplifies as

$$\begin{aligned}&-\gamma_{A,n,d}\cdot\phi\left(\frac{\Xi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\right)\frac{\Psi_{A,n,d}}{\sqrt{\Lambda_{A,n,d}}}\\ &\quad= \frac{\sqrt{2\kappa(1-\kappa)}\phi(\beta_{A,\lambda,\kappa})}{\sqrt{\Phi(\beta_{A,\lambda,\kappa})\{1-\Phi(\beta_{A,\lambda,\kappa})\}}}\cdot\frac{n\lambda(1-\lambda)\delta^\top A\delta}{\sqrt{2\operatorname{tr}\{(A\Sigma)^2\}}} + o(1).\end{aligned}$$

Resubstituting the above into expression (6.2), we finally infer that

$$\mathbb{E}[\varphi_A] = \Phi\left(-z_\alpha + \frac{\sqrt{2\kappa(1-\kappa)}\phi(\beta_{A,\lambda,\kappa})}{\sqrt{\Phi(\beta_{A,\lambda,\kappa})\{1-\Phi(\beta_{A,\lambda,\kappa})\}}}\cdot\frac{n\lambda(1-\lambda)\delta^\top A\delta}{\sqrt{2\operatorname{tr}\{(A\Sigma)^2\}}}\right) + o(1). \tag{6.4}$$

Since $\sup_{x\in\mathbb{R}}\phi(x)/\sqrt{\Phi(x)\{1-\Phi(x)\}} = \sqrt{2/\pi}$ and its maximum is achieved at $x = 0$, the asymptotic power (6.4) is maximized when $\lambda = 1/2$ and $\kappa = 1/2$, further supported by simulations in Appendix D. However it is unknown whether the same result continues to hold

for a random A (e.g., $A = \widehat{\Sigma}^{-1}$). In this balanced setting, the asymptotic power is further simplified as

$$\Phi\left(-z_\alpha + \frac{n\delta^\top A\delta}{\sqrt{32\pi\,\mathrm{tr}\{(A\Sigma)^2\}}}\right) + o(1). \tag{6.5}$$

For ease of reference, we summarize our discussion as a theorem.

THEOREM 6.1. *Suppose that the assumptions* (A1)–(A6) *hold. Then the generalized LDA test* (6.1) *asymptotically controls type-*1 *error at level* α *and its power for Gaussian two-sample mean testing is given by*

$$\mathbb{E}[\varphi_A] = \Phi\left(-z_\alpha + \frac{\sqrt{2\kappa(1-\kappa)}\phi(\beta_{A,\lambda,\kappa})}{\sqrt{\Phi(\beta_{A,\lambda,\kappa})\{1-\Phi(\beta_{A,\lambda,\kappa})\}}} \cdot \frac{n\lambda(1-\lambda)\delta^\top A\delta}{\sqrt{2\,\mathrm{tr}\{(A\Sigma)^2\}}}\right) + o(1). \tag{6.6}$$

Furthermore, keeping other parameters fixed, the asymptotic power is maximized when $\lambda = \kappa = 1/2$ (*corresponding to a balanced train/test split*).

The proof of the above theorem follows immediately from the preceding discussion and so is omitted. As a direct consequence of Theorem 6.1, when $\lambda = 1/2$ and $\kappa = 1/2$, the power of the "oracle" Fisher's LDA test that uses $A = \Sigma^{-1}$ (again, "oracle" is used because it uses Σ^{-1}) becomes

$$\mathbb{E}[\varphi^*_{\Sigma^{-1}}] = \Phi\left(-z_\alpha + \frac{n\delta^\top\Sigma^{-1}\delta}{\sqrt{32\pi d}}\right) + o(1). \tag{6.7}$$

Comparing the above power with the minimax lower bound expression (3.3) with $\lambda = 1/2$, we may conclude that the classification accuracy test can achieve essentially minimax optimal power, up to the small constant factor $1/\sqrt{\pi} \approx 0.564$. In other words, we pay a constant factor by performing a two-sample test via classification. However, this conclusion should be treated with caution as emphasized below:

- First, Theorem 6.1 is a pointwise result. That means, the result holds for any sequence of distributions satisfying the assumptions, but not uniformly over a class of distributions. Hence, conceptually, this is weaker than the uniform power achieved by φ^*_H in Theorem 4.1. However, this drawback actually applies to almost every published result on high-dimensional two-sample testing that we are aware of (or certainly all those that we cite), and it is a much broader open problem to prove that the power guarantees for these tests hold uniformly over the relevant classes.
- Second, although a constant factor is not of major concern in determining the minimax rate, it may have a significant effect on power in practice. To see this, let n_{Fisher} and $n_{\mathrm{Hotelling}}$ be the sample sizes needed for $\varphi^*_{\Sigma^{-1}}$ and φ^*_H to obtain the same power against the local alternative considered in Theorem 6.1. Then the asymptotic relative efficiency (ARE) of $\varphi^*_{\Sigma^{-1}}$ with respect to φ^*_H is defined as the limit of the ratio $n_{\mathrm{Hotelling}}/n_{\mathrm{Fisher}}$ (e.g., Chapter 14 of van der Vaart (1998)). Based on the asymptotic power expressions (3.3) and (6.6), a simple closed-form expression of the ARE is available as

$$\mathrm{ARE}(\varphi^*_{\Sigma^{-1}}; \varphi^*_H) = \frac{\sqrt{2\kappa(1-\kappa)}\phi(\beta^*)}{\sqrt{\Phi(\beta^*)\{1-\Phi(\beta^*)\}}} \le \frac{1}{\sqrt{\pi}} \approx 0.564, \tag{6.8}$$

  where $\beta^* = \lim_{n,d\to\infty} \beta_{\Sigma^{-1},\lambda,\kappa}$ if it exists. This ARE expression implies that $\varphi^*_{\Sigma^{-1}}$ requires (at least) $\sqrt{\pi} \approx 1.77$ more samples to attain approximately the same power as φ^*_H. In this context, Hotelling's test should be preferred over the classifier-based test to obtain higher power against the Gaussian mean shift alternative.

In the following sections, we extend the results on the oracle Fisher's LDA classifier to it variants with unknown Σ and also to elliptical distributions.

REMARK 6.1. As mentioned in Section 5.1, the accuracy of the Bayes optimal classifier approaches half under the considered asymptotic regime, meaning that no classifier can have accuracy better than a random guess in the limit. In contrast, under the same asymptotic regime, two-sample testing based on generalized LDA can have nontrivial power (strictly greater than α) as shown in Theorem 6.1. These two results not only demonstrate that testing is easier than classification, but also that the local alternative (A2) is conceptually interesting—it corresponds to a regime where the LDA classifier performs as poorly as a random guess for classification, but is essentially optimal for testing.

7. Naive Bayes: Power of generalized LDA with unknown Σ. For low-dimensional Gaussians with unknown Σ, there are strong reasons to prefer Hotelling's test; it is well known that it is *uniformly most powerful* among all tests that are invariant with respect to nonsingular linear transformations (e.g., Anderson (1958)). We also refer to Giri and Kiefer (1964), Giri, Kiefer and Stein (1963), Kariya (1981), Luschgy (1982), Salaevskii (1971), Simaika (1941) for other optimality properties of Hotelling's test in finite d and n settings. Moreover, our result in Theorem 4.1 says that φ_H is asymptotically minimax optimal among all level α tests as long as $d/n \to 0$. Unfortunately, when d is linearly comparable to or larger than n, these optimal properties of Hotelling's test becomes highly nontrivial. In particular, φ_H has asymptotic power tending to the (trivial) value of α in the high-dimensional setting, when $d, n \to \infty$ with $d/n \to 1 - \epsilon$ for small $\epsilon > 0$ (Bai and Saranadasa (1996) for details). The problem becomes even worse when the dimension is larger than the sample size as T_H is not well defined.

The aforementioned issue on T_H has motivated the study of alternative two-sample mean test statistics in the high-dimensional setting. For instance, Bai and Saranadasa (1996) show that dropping $\widehat{\Sigma}$ from the Hotelling test statistic (i.e., replacing $\widehat{\Sigma}$ with the identity matrix) entirely leads to a test that does have asymptotic power tending to one in the high-dimensional setting where Hotelling's test fails. The test statistic proposed by Bai and Saranadasa (1996) can be essentially written as

$$T_{\mathrm{BS}} \stackrel{\mathrm{def}}{=} (\widehat{\mu}_0 - \widehat{\mu}_1)^\top (\widehat{\mu}_0 - \widehat{\mu}_1).$$

Following that, Srivastava and Du (2008) propose (in a similar spirit) the test statistic

$$T_{\mathrm{SD}} \stackrel{\mathrm{def}}{=} (\widehat{\mu}_0 - \widehat{\mu}_1)^\top \operatorname{diag}(\widehat{\Sigma})^{-1} (\widehat{\mu}_0 - \widehat{\mu}_1), \tag{7.1}$$

by replacing $\widehat{\Sigma}$ with $\operatorname{diag}(\widehat{\Sigma})$ in Hotelling's statistic. They show that T_{SD} also leads to high-dimensional consistency.

As mentioned earlier, the idea of using $\operatorname{diag}(\widehat{\Sigma})$ in place of $\widehat{\Sigma}$ has also been justified in the high-dimensional classification problem (Bickel and Levina (2004)). In particular, the naive Bayes classifier (corresponding to T_{SD}) outperforms Fisher's LDA classifier (corresponding to T_H) in terms of the worst-case classification error in the high-dimensional setting. We note that this relatively understated connection between two-sample testing and classification has important implications for extending our previous results to other linear classifiers. Specifically, as we shall see, the power of the classifier-based tests is only worse by a constant factor than the variants of Hotelling's test when both the classifier and the two-sample test use the same substitute for Σ^{-1}.

To start, let us consider two classifiers with unknown Σ. The first one is the naive Bayes classifier and the other is the generalized LDA classifier with the identity matrix, that is,

$A = I$. We then compare the power of the corresponding classification accuracy tests with the two-sample mean tests based on T_{SD} and T_{BS}. Throughout this section, we assume that $n_0 = n_1$, $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$ and $n_{\mathrm{tr}} = n_{\mathrm{te}}$ for simplicity.

From Theorem 6.1, the asymptotic power of the test based on $\widehat{E}_I^S$ is already available as

$$\mathbb{E}[\varphi_I] = \Phi\left(-z_\alpha + \frac{n\delta^\top \delta}{\sqrt{32\pi\,\mathrm{tr}(\Sigma^2)}}\right) + o(1). \tag{7.2}$$

Under more general conditions than the assumptions (A1)–(A6), Bai and Saranadasa (1996) show that the asymptotic power of the test based on T_{BS}, denoted by φ_{BS}, is

$$\mathbb{E}[\varphi_{\mathrm{BS}}] = \Phi\left(-z_\alpha + \frac{n\delta^\top \delta}{\sqrt{32\,\mathrm{tr}(\Sigma^2)}}\right) + o(1). \tag{7.3}$$

Now by comparing two power expressions in (7.2) and (7.3), we arrive at the same conclusion as before that the classification accuracy test is less powerful than the corresponding two-sample test φ_{BS} by the constant factor $1/\sqrt{\pi} \approx 0.564$.

Next, we focus on the naive Bayes classifier and compute the asymptotic power of the resulting test. Although the analysis proceeds similar to the previous one, we now need to deal with the randomness from the inverse diagonal matrix, which requires extra nontrivial work. By putting $\widehat{D}^{-1} \overset{\mathrm{def}}{=} \mathrm{diag}(\widehat{\Sigma})^{-1}$ and $D^{-1} = \mathrm{diag}(\Sigma)^{-1}$, the asymptotic power of the naive Bayes classifier is provided as follows.

THEOREM 7.1. *Consider the case where $n_0 = n_1$, $n_{0,\mathrm{tr}} = n_{1,\mathrm{tr}}$ and $n_{\mathrm{tr}} = n_{\mathrm{te}}$. Then under the assumptions* (A1), (A2) *and* (A5), *the power of the naive Bayes classifier test for Gaussian two-sample mean testing is*

$$\mathbb{E}[\varphi_{\widehat{D}^{-1}}] = \Phi\left(-z_\alpha + \frac{n\delta^\top D^{-1}\delta}{\sqrt{32\pi\,\mathrm{tr}\{(D^{-1}\Sigma)^2\}}}\right) + o(1). \tag{7.4}$$

The proof of Theorem 7.1 can be found in Appendix C.5. Srivastava and Du (2008) study the asymptotic power of the test φ_{SD} based on T_{SD} (7.1). One can also check that their conditions are fulfilled under the assumptions (A1)–(A5). Using $\lambda = 1/2$, the power of φ_{SD} is given by

$$\mathbb{E}[\varphi_{\mathrm{SD}}] = \Phi\left(-z_\alpha + \frac{n\delta^\top D^{-1}\delta}{\sqrt{32\,\mathrm{tr}\{(D^{-1}\Sigma)^2\}}}\right) + o(1).$$

Comparing this with the asymptotic power of $\varphi_{\widehat{D}^{-1}}$ in (7.4), we see that the power of the accuracy test based on the naive Bayes classifier is worse than the corresponding two-sample test φ_{SD}, once again achieving an ARE of exactly $1/\sqrt{\pi}$.

8. Extension to elliptical distributions. In this section, we extend our main result (Theorem 6.1) to the class of elliptical distributions and show that the asymptotic power expression remains the same up to a constant factor. Let μ be a d-dimensional vector, S be a $d \times d$ positive semidefinite matrix, $\xi(\cdot)$ be a nonnegative function. A random vector Z in $\mathbb{R}^d$ is said to have an elliptical distribution with location parameter μ, scale matrix S and generator $\xi(\cdot)$ if its characteristic function satisfies

$$\mathbb{E}\big[e^{it^\top Z}\big] = e^{it^\top \mu}\xi\big(t^\top S t\big) \quad \text{for all } t \in \mathbb{R}^d.$$

When the second moment exists, it can be verified that μ corresponds to the mean vector of Z and S is proportional to the covariance matrix of Z, denoted by Σ. More specifically,

by letting $\xi'(0)$ be the first derivative of ξ evaluated at zero, S is explicitly linked to Σ as $-2\xi'(0)S = \Sigma$. Notable examples of elliptical distributions include the multivariate normal, the multivariate student t, the multivariate Laplace and the multivariate logistic distribution. We refer to Fang, Kotz and Ng (2018), Frahm (2004), Gómez, Gómez-Villegas and Marín (2003) for further properties and examples of elliptical distributions. To have an explicit power expression, we make two extra assumptions on Z described as follows:

(A7) *Condition on kurtosis parameter*: let ζ_{kurt} be the kurtosis parameter of Z defined as

$$\zeta_{\text{kurt}} \stackrel{\text{def}}{=} \frac{\mathbb{E}[\{(Z-\mu)^\top \Sigma^{-1}(Z-\mu)\}^2]}{d(d+2)} - 1.$$

We assume that there exists a positive constant M such that $\zeta_{\text{kurt}} < M$ for all n, d.

(A8) *Condition on density function*: assume that the standardized first coordinate of Z, that is, $e_1^\top(Z-\mu)/(e_1^\top \Sigma e_1)^{1/2}$ where $e_1 = (1, 0, \ldots, 0)^\top$, has the density function $f_\xi(\cdot)$ with respect to the Lebesgue measure. We further assume that f_ξ is bounded and continuously differentiable.

We believe that the condition on ζ_{kurt} in (A7) is mild and satisfied for many elliptical distributions (e.g., Zografos (2008)). For example, the kurtosis parameter of the multivariate t-distribution with ν degrees of freedom is $2/(\nu-4)$ for $\nu > 4$, which in turn implies that ζ_{kurt} is zero for the Gaussian case. To interpret (A8), we note that each component of an elliptical random vector has the same distribution after standardization. Assumption (A8) then states that this common distribution has the density function f_ξ with some extra regularity conditions. Clearly, f_ξ corresponds to the standard normal density function for the Gaussian case that is bounded and continuously differentiable. But (A8) fails to hold for the Laplace distribution whose density function is not differentiable at zero. With these extra assumptions, we are now ready to present the main result of this section, which generalizes Theorem 6.1 to elliptical distributions.

THEOREM 8.1. *Suppose that $\mathbb{P}_0$ and $\mathbb{P}_1$ are elliptical distributions with parameters (μ_0, S, ξ) and (μ_1, S, ξ), respectively. Consider the case where $n_0 = n_1$, $n_{0,\text{tr}} = n_{1,\text{tr}}$ and $n_{\text{tr}} = n_{\text{te}}$, that is, $\lambda = \kappa = 1/2$, for simplicity. Then under the assumptions* (A1), (A2) *and* (A5)–(A8), *the generalized LDA test* (6.1) *asymptotically controls type-*1 *error at level α and has the asymptotic power for testing the hypothesis* (2.1) *as*

$$\mathbb{E}[\varphi_A] = \Phi\left(-z_\alpha + \frac{f_\xi(0) \cdot n\delta^\top A\delta}{\sqrt{16\,\text{tr}\{(A\Sigma)^2\}}}\right) + o(1). \tag{8.1}$$

The above result shows that the asymptotic power expression in Theorem 6.1 does not change in terms of n, d, Σ, A, δ, for elliptical distributions. To further illustrate the result, let us consider the specific case where $\mathbb{P}_0$ and $\mathbb{P}_1$ are multivariate t-distributions with ν degrees of freedom and the same scale matrix. We additionally assume that $\nu > 4$ under which the assumption (A7) is satisfied. In such a case, $f_\xi(0) = f_\xi(0; \nu)$ equals

$$f_\xi(0;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi(\nu-2)}\Gamma(\frac{\nu}{2})} \to \frac{1}{\sqrt{2\pi}} \approx 0.399 \quad \text{as } \nu \to \infty.$$

Hence, by taking $\nu \to \infty$, the asymptotic power (8.1) recovers the previous power expression (6.5) for the Gaussian case. Indeed $f_\xi(0;\nu)$ is a decreasing sequence of ν such that $f_\xi(0;\nu) < f_\xi(0;4) \approx 0.530$ for all $\nu > 4$. This fact demonstrates that the generalized LDA test becomes relatively more efficient when the underlying t-distributions have heavier tails, which is also validated by simulations (see Figure 2 in Section 10.1).

9. Results on general classifiers. So far we have focused on the accuracy tests based on linear classifiers and derived their explicit asymptotic power against local alternatives under Gaussian or elliptical distribution assumptions. In this section, we turn to more general settings and examine two key properties, namely the type-1 error control and consistency, of the accuracy test based on a general classifier. The main result of this section shows that a classification accuracy test achieves asymptotic power equal to one, provided that the corresponding classifier has an accuracy higher than chance. This result naturally motivates questions about rate, for which more assumptions are needed, and also motivates studying a more challenging setting where the true accuracy approaches half, like the one we consider for the generalized LDA test.

Recall that for a generic classifier $\widehat{C}$ based on the training set, the per-class and total errors $\widehat{E}_0^S(\widehat{C})$, $\widehat{E}_1^S(\widehat{C})$ and $\widehat{E}^S(\widehat{C})$ are calculated using expression (2.4). To facilitate analysis, we assume the following asymptotic properties of $\widehat{E}_0^S(\widehat{C})$ and $\widehat{E}_1^S(\widehat{C})$:

(A9) *Asymptotic classification errors*: assume that $\widehat{E}_0^S(\widehat{C}) = E_0(C) + o_P(1)$ and $\widehat{E}_1^S(\widehat{C}) = E_1(C) + o_P(1)$ where $E_1(C)$ and $E_2(C)$ are constants in $(0, 1)$. Moreover, there exists a constant $\epsilon > 0$ such that $E_0(C)/2 + E_1(C)/2 = 1/2 - \epsilon$ under the alternative hypothesis.

To determine the significance threshold for deciding if the error is different from chance, we consider two methods: (1) the Gaussian approximation that underlies our theory in the preceding sections and (2) the permutation procedure with finite sample guarantees that has been common in practice.

9.1. *Asymptotic test.* As discussed before, the sample-splitting error can be viewed as the sum of independent random variables given the training set. Therefore, it is natural to expect that this empirical error follows closely a normal distribution even for a general classifier when the sample size is large. Building on this intuition, we define the asymptotic test as

$$\mathbb{I}\left[\frac{2\widehat{E}^S(\widehat{C}) - 1}{\sqrt{\widehat{E}_0^S(\widehat{C})\{1 - \widehat{E}_0^S(\widehat{C})\}/n_{0,\mathrm{te}} + \widehat{E}_1^S(\widehat{C})\{1 - \widehat{E}_1^S(\widehat{C})\}/n_{1,\mathrm{te}}}} < -z_\alpha\right]$$

and denote it by $\varphi_{\widehat{C},\mathrm{Asymp}}$. We note that the quantity inside of the indicator function is a studentized sample-splitting error under the null hypothesis. In the next proposition, we prove that the normal approximation is indeed accurate, and thus $\varphi_{\widehat{C},\mathrm{Asymp}}$ is a valid test at least asymptotically. Moreover, when the sequence of classification errors tends to a constant that is strictly less than chance level, we show that the power of the asymptotic test tends to one as $n \to \infty$ potentially with $d \to \infty$.

PROPOSITION 9.1. *Suppose that the assumptions* (A3), (A4) *and* (A9) *hold as* $n \to \infty$ *potentially with* $d \to \infty$ *at any relative rate. Then under the null hypothesis* $H_0 : \mathbb{P}_0 = \mathbb{P}_1$, *we have* $\lim_{n\to\infty} \mathbb{E}_{H_0}[\varphi_{\widehat{C},\mathrm{Asymp}}] \le \alpha$. *On the other hand, under the alternative hypothesis* $H_1 : \mathbb{P}_0 \ne \mathbb{P}_1$, *the asymptotic test is consistent as* $\lim_{n\to\infty} \mathbb{E}_{H_1}[\varphi_{\widehat{C},\mathrm{Asymp}}] = 1$.

Despite its simplicity, the asymptotic approach has no finite sample guarantee. Next, we prove consistency of permutation-based approaches.

9.2. *Permutation tests.* In practice, one often employs permutation tests that can offer exact control of the type-1 error rate. There are two possible ways of applying permutation testing within the classification via sample splitting framework. The methods below differ in the italicized text.

METHOD 1 (Half-permutation).

- Split data into two halves, X^1, Y^1 and X^2, Y^2. Train the classifier on X^1, Y^1, call it f^*. Evaluate accuracy of f^* on X^2, Y^2, call it a^*.
- Repeat P times: *Pool the samples X^2, Y^2 into one bag, randomly permute the samples, and then split it into two parts, X^p, Y^p. Here, each part of X^p, Y^p has the same sample size as the corresponding part of X^2, Y^2. Evaluate the accuracy of f^* on this permuted data, call this a^p.*
- Sort $a^*, a^1, \ldots, a^P$ and denote their order statistics by $a^{(1)} \leq \cdots \leq a^{(P+1)}$; Let $k \stackrel{\text{def}}{=} \lceil (1-\alpha)(1+P) \rceil$. If $a^* > a^{(k)}$, then reject the null.

METHOD 2 (Full-permutation).

- Split data into two halves, X^1, Y^1 and X^2, Y^2. Train the classifier on X^1, Y^1, call it f^*. Evaluate accuracy of f^* on X^2, Y^2, call it a^*.
- Repeat P times: *Pool all samples X^1, Y^1, X^2, Y^2 into one bag, randomly permute the samples, and then split it into 4 parts X^p, Y^p, X'^p, Y'^p. Here, each part of X^p, Y^p, X'^p, Y'^p has the same sample size as the corresponding part of X^1, Y^1, X^2, Y^2. Train a new classifier f^p on the first half, evaluate it on the second half, to get accuracy a^p.*
- Sort $a^*, a^1, \ldots, a^P$ and denote their order statistics by $a^{(1)} \leq \cdots \leq a^{(P+1)}$. Let $k \stackrel{\text{def}}{=} \lceil (1-\alpha)(1+P) \rceil$. If $a^* > a^{(k)}$, then reject the null.

It is worth noting that both methods yield a valid level α test under $H_0 : \mathbb{P}_0 = \mathbb{P}_1$ as a direct consequence of, for example, Theorem 1 in Hemerik and Goeman (2018). In terms of power, method 2 may potentially be more powerful than method 1 as it uses the data more efficiently to determine a threshold. In particular, permuted accuracies via method 1 can take fewer values than those via method 2, which may result in a more conservative threshold depending on the nominal level. However, method 1 has a computational advantage over method 2 since it only requires to refit a classifier on the second half of the dataset. Nevertheless the following theorem shows that both methods provide a consistent test under the same assumptions made in Proposition 9.1. Let us denote the permutation test by $\varphi_{\widehat{C},\text{Perm}}$ via either method 1 or method 2 based on classifier $\widehat{C}$.

THEOREM 9.1. *Consider the same assumptions made in Proposition 9.1. Then under the null hypothesis $H_0 : \mathbb{P}_0 = \mathbb{P}_1$, we have $\mathbb{E}_{H_0}[\varphi_{\widehat{C},\text{Perm}}] \leq \alpha$ for each n and d. Under the alternative hypothesis $H_1 : \mathbb{P}_0 \neq \mathbb{P}_1$, the (half or full) permutation test is consistent as $\lim_{n\to\infty} \mathbb{E}_{H_1}[\varphi_{\widehat{C},\text{Perm}}] = 1$ given that the number of random permutations P is greater than $(1-\alpha)/\alpha$.*

One interesting aspect of the above theorem is that consistency is guaranteed as long as the number of random permutations P is greater than $(1-\alpha)/\alpha$ (e.g., $P \geq 20$ for $\alpha = 0.05$), which is independent of the sample size. We would also like to point out that the permutation test relies on a data-dependent threshold and thus it is more difficult to analyze than the asymptotic test. In Appendix C.11, we bound this data-dependent threshold with a more tractable quantity using Markov's inequality with the first two moments of the permuted test statistic. Leveraging this preliminary result, we prove that the permutation critical value cannot exceed the true accuracy in the limit, and this is the critical fact that completes the proof.

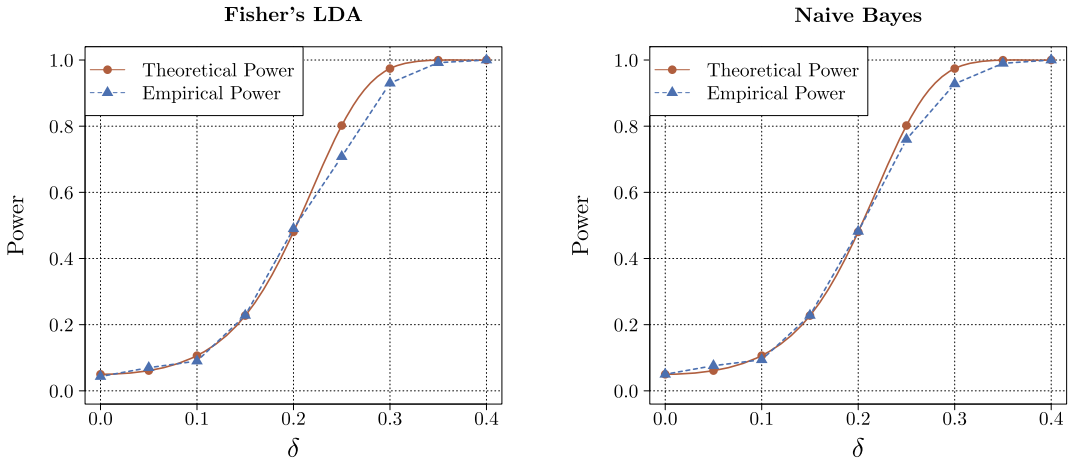

FIG. 1. *Comparisons of the empirical power to our theoretically derived expression for (asymptotic) power under the Gaussian setting. The curves are almost identical especially when the size of δ is not too big, which suggests that our theory under local alternatives accurately predicts power. See Section 10.1 for details.*

10. Experiments. In this section, we present several numerical results that support our theoretical analysis. Throughout our simulations (except in Section 10.3), we set the sample sizes and the dimension to be $n_0 = n_1 = d = 200$ and compare two multivariate Gaussian or multivariate t-distributions with the same identity covariance matrix, with means

$$\mu_0 = (0, \ldots, 0)^\top \quad \text{and} \quad \mu_1 = \frac{\delta}{d^{1/4}} \cdot (1, \ldots, 1)^\top$$

for $\delta \in \{0, 0.05, \ldots, 0.35, 0.40\}$. The simulations were repeated 500 times to estimate the power of each test at significance level $\alpha = 0.05$.

10.1. *Empirical power versus theoretical power.* In the following experiment, we compare the empirical power of classification accuracy tests with the corresponding theoretical power. For the Gaussian case, we consider the accuracy tests $\varphi_{\Sigma^{-1}}$ and $\varphi_{\widehat{D}^{-1}}$ based on the Fisher's LDA classifier and the naive Bayes classifier, respectively. As specified in the definitions of $\varphi_{\Sigma^{-1}}$ and $\varphi_{\widehat{D}^{-1}}$, the critical values of both tests are based on a normal approximation. Here, we split the samples into training and test sets with equal sample sizes so that the power is asymptotically maximized. In this case, the asymptotic power expression for each test is presented in (6.7) and (7.4), respectively. For the case of multivariate t-distributions, we focus on the accuracy test $\varphi_{\Sigma^{-1}}$ and see whether the asymptotic power expression (8.1) approximates its empirical power over different values of degrees of freedom ν.

The results are given in Figure 1 and Figure 2. We see that the empirical power almost coincides with the theoretical counterpart especially when δ is not too big (i.e., low SNR regime), which confirms our theoretical analysis. We also see that the accuracy test has higher power when the underlying t-distributions have smaller degrees of freedom, an interesting and initially surprising fact that is again predicted by our theory.

10.2. *Sample-splitting versus resubstitution.* In the following experiment, we compare the performance of sample-splitting tests with resubstitution accuracy tests under the Gaussian setting. As their name suggests, the resubstitution accuracy tests use resubstitution accuracy estimates as their test statistic. The precise definition of a resubstitution estimate is given in Appendix B. We also consider Hotelling's test and its variant proposed by Srivastava and Du (2008) as reference points. The setup is almost the same as the previous experiment except for the choice of critical values. In particular, since the (asymptotic) null distribution

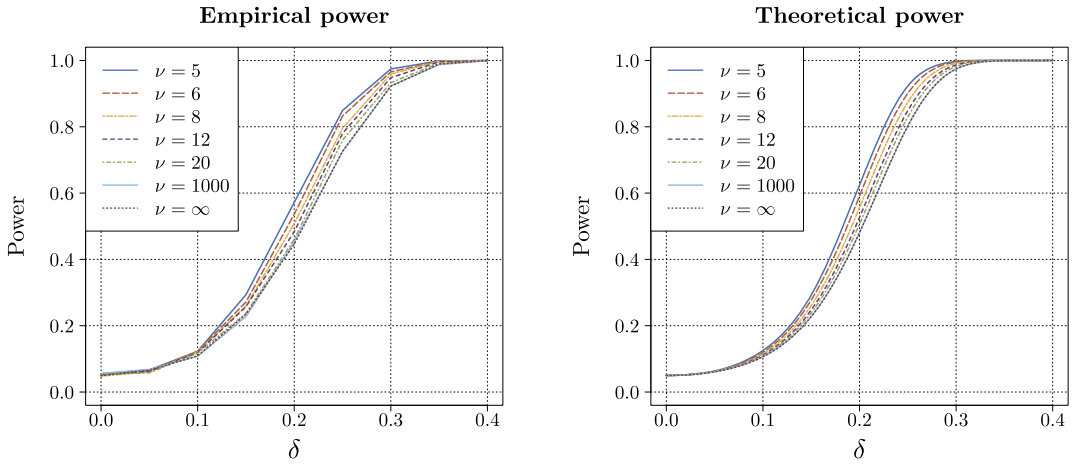

FIG. 2. *The empirical power and theoretical (asymptotic) power of the accuracy test based on Fisher's LDA classifier for comparing multivariate t-distributions with ν degrees of freedom. The curves are tightly matched across ν. Moreover, predicted by Theorem 8.1, the power* decreases *with ν. See Section 10.1 for details.*

of a resubstitution statistic is unknown, the critical values of all tests are determined using permutations for a fair comparison. Specifically, to calibrate critical values, we use the full permutation method from Section 9.2 with 200 random permutations.

In the first part, Fisher's LDA is considered as a base line classifier. Then the accuracy is estimated via (i) sample-splitting with $n_{\text{tr}} = n_{\text{te}}$ and (ii) resubstitution. As a reference point, we consider Hotelling's test as it shares the same weight matrix with Fisher's LDA. For both Hotelling's and Fisher's LDA tests, we assume that Σ is known. In the second part, the naive Bayes classifier is considered as a base line classifier with unknown Σ. We then perform tests based on sample-splitting and resubstitution accuracy statistics defined similarly as before. In this part, we consider T_{SD} given in (7.1) as a reference point since it relies on the inverse of diagonal sample covariance matrix as in the naive Bayes classifier.

From the results presented in Figure 3, it stands out that Hotelling's test and its high-dimensional variant are more powerful than the corresponding tests via classification accuracy as we expected. The results also show that the powers of the sample-splitting tests are slightly higher than those of the resubstitution tests in both Fisher's LDA and naive Bayes classifier examples. However, additional simulation studies, not presented here, suggest that resubstitution tests tend to be more powerful than sample-splitting tests in low-dimensional settings (or when the sample sizes are relatively small), and thus, at least empirically, neither of them is strictly better than the other under all scenarios. Similar empirical results were observed by Rosenblatt et al. (2019) where they conducted extensive simulation studies to compare the performance of the accuracy tests via resubstitution and 4-fold cross-validation and different versions of Hotelling's test. From their simulation results, one reaches the same conclusion that the accuracy tests tend to have lower power than Hotelling's test against Gaussian mean shift alternatives.

10.3. *Asymptotic power of Hotelling's test.* In this subsection, we provide numerical support for the asymptotic optimality of Hotelling's test under Gaussian settings with unknown Σ (Theorem 4.1). Here, we compare two multivariate Gaussian distributions with the mean vectors

$$\mu_0 = \frac{1}{d^{1/4} n_0^{1/2}} \cdot (1, \ldots, 1)^\top \quad \text{and} \quad \mu_1 = -\frac{1}{d^{1/4} n_0^{1/2}} \cdot (1, \ldots, 1)^\top$$

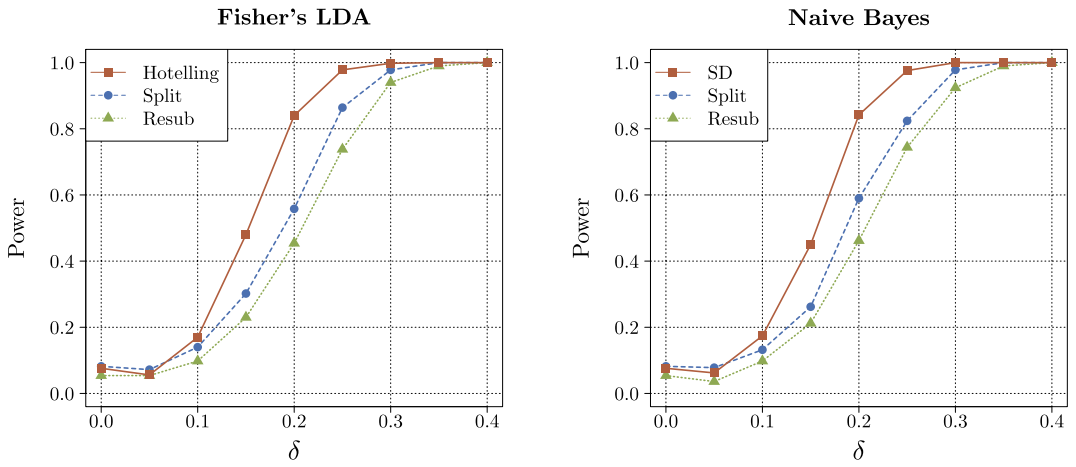

FIG. 3. *Comparisons between sample-splitting (Split) and resubstitution (Resub) tests using LDA and naive Bayes. As reference points, we also consider Hotelling's test and the test based on* T_{SD}. *Under the given scenarios, the sample-splitting tests have higher power than the resubstitution tests but lower power than Hotelling's and SD tests, the latter being predicted by our theory. See Section 10.2 for details.*

and the identity covariance matrix. In this case, by setting $n_0 = n_1$, the asymptotic minimax power tends to be constant as in (3.3). Now we consider six different asymptotic regimes: (i) $d = \lfloor n_0^{1/4} \rfloor$, (ii) $d = \lfloor n_0^{2/4} \rfloor$, (iii) $d = \lfloor n_0^{3/4} \rfloor$, (iv) $d = 0.5n_0$, (v) $d = 1.0n_0$ and (vi) $d = 1.5n_0$. According to Theorem 4.1, Hotelling's test with unknown Σ (denoted by φ_H) obtains asymptotically the same power as the minimax optimal test (denoted by φ_H^*) in the first three regimes. Whereas, in the last three regimes where d and n are linearly comparable, φ_H becomes less powerful than φ_H^* proved by Bai and Saranadasa (1996). To illustrate this numerically, we increase the sample size by $n_0 \in \{10^1, 10^2, \ldots, 10^6\}$ and compute the power of φ_H^* and φ_H for each n_0. To calculate the power, we use the fact that $\mathbb{E}[1 - \varphi_H^*]$ and $\mathbb{E}[1 - \varphi_H]$ are noncentral χ^2 and F distribution functions evaluated at their critical values, which are $c_{\alpha,d}$ and $q_{\alpha,n,d}$, respectively.

As can be seen in the first row of Figure 4, the power of φ_H becomes approximately the same as that of φ_H^* in the first three regimes as n increases. On the other hand, in the last three regimes where $d/n \to c \in (0, 1)$, we observe significantly different results. Specifically, from the second row of Figure 4, it is seen that the power of φ_H is much lower than that of φ_H^* and the gap does not decrease even in large n. This, thereby, supports our argument that φ_H is asymptotically comparable to the minimax optimal test in the case of $d/n \to 0$, but it is underpowered otherwise.

11. Conclusion. This paper provided analyses on the use of classification accuracy as a test statistic for two-sample testing. We started by presenting a fundamental minimax lower bound for high-dimensional two-sample mean testing and showed that Hotelling's test with unknown Σ can be optimal in high-dimensional settings as long as $d/n \to 0$. When $d = O(n)$, we found that two-sample tests via the classification accuracy of various versions of Fisher's LDA (including naive Bayes) have the same power as high-dimensional versions of Hotelling's test in terms of all problem parameters (n, d, δ, Σ), but having worse (but explicit) constants.

Beyond linear classifiers, we also proved that both the asymptotic test and the permutation test based on a general classifier are consistent if the limiting value of the true accuracy is higher than chance. This consistency result naturally motivated a more challenging setting in which the Bayes error approaches half while the corresponding accuracy-based test can

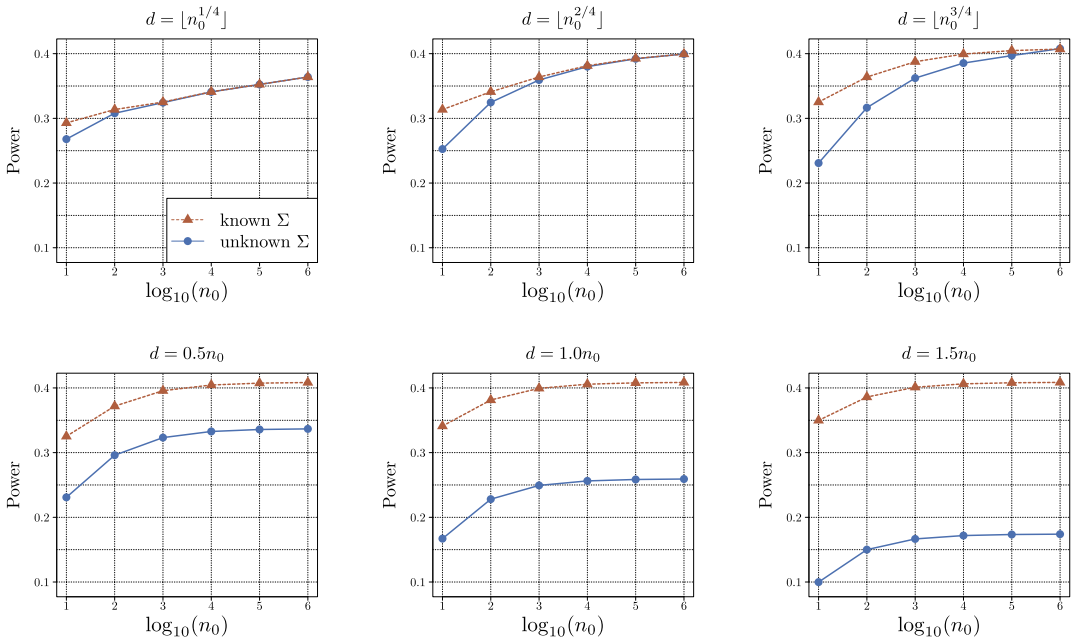

FIG. 4. *Comparisons of the power of* (1) *Hotelling's test* φ_H *with unknown* Σ *and* (2) *Hotelling's test* φ_H^* *with known* Σ *at* $\alpha = 0.05$ *in different asymptotic regimes. These results coincide with our theoretical results in Section* 4, *showing that* φ_H *has asymptotically the same power as* φ_H^* *when* $d/n \to 0$ (*first row*) *and it is less powerful when* $d/n \to c \in (0, 1)$ (*second row*). *See Section* 10.3 *for details.*

still have nontrivial power, which is the regime studied in most of this paper. Under such a challenging regime, it would be interesting to see whether explicit expressions of power can be derived for nonlinear classifiers. Characterizing the high-dimensional power (beyond consistency as we have shown) of permutation-based tests is also an important open problem.

Acknowledgments. We thank the Associate Editor and anonymous referees for their valuable comments that significantly improved the paper.

SUPPLEMENTARY MATERIAL

Supplement to "Classification accuracy as a proxy for two-sample testing" (DOI: 10.1214/20-AOS1962SUPP; .pdf). This supplemental file includes the technical proofs omitted in the main text and a discussion on open problems.

REFERENCES

ANDERSON, T. W. (1951). Classification by multivariate analysis. *Psychometrika* **16** 31–50. MR0041403 https://doi.org/10.1007/BF02313425

ANDERSON, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. *Wiley Publications in Statistics*. Wiley, New York. MR0091588

ARIAS-CASTRO, E., PELLETIER, B. and SALIGRAMA, V. (2018). Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *J. Nonparametr. Stat.* **30** 448–471. MR3794401 https://doi.org/10.1080/10485252.2018.1435875

BAI, Z. and SARANADASA, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6** 311–329. MR1399305

BEN-DAVID, S., BLITZER, J., CRAMMER, K. and PEREIRA, F. (2007). Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems* 137–144.

BHATTACHARYA, B. B. (2020). Asymptotic distribution and detection thresholds for two-sample tests based on geometric graphs. *Ann. Statist.* **40** 2879–2903. MR4152627 https://doi.org/10.1214/19-AOS1913

BICKEL, P. J. and LEVINA, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* **10** 989–1010. MR2108040 https://doi.org/10.3150/bj/1106314847

BLANCHARD, G., LEE, G. and SCOTT, C. (2010). Semi-supervised novelty detection. *J. Mach. Learn. Res.* **11** 2973–3009. MR2746544

BORJI, A. (2019). Pros and cons of GAN evaluation measures. *Comput. Vis. Image Underst.* **179** 41–65.

CHEN, S. X. and QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38** 808–835. MR2604697 https://doi.org/10.1214/09-AOS716

CHEN, N. F., SHEN, W., CAMPBELL, J. and SCHWARTZ, R. (2009). Large-scale analysis of formant frequency estimation variability in conversational telephone speech. In *Tenth Annual Conference of the International Speech Communication Association*.

ETZEL, J. A., GAZZOLA, V. and KEYSERS, C. (2009). An introduction to anatomical ROI-based fMRI classification analysis. *Brain Res.* **1282** 114–125.

FANG, K. T., KOTZ, S. and NG, K. W. (2018). *Symmetric Multivariate and Related Distributions*. Chapman and Hall/CRC.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annu. Eugen.* **7** 179–188.

FISHER, R. A. (1940). The precision of discriminant functions. *Annu. Eugen.* **10** 422–429. MR0003543

FRAHM, G. (2004). Generalized elliptical distributions: Theory and applications. Ph.D. thesis, Universität zu Köln.

FRIEDMAN, J. (2004). On multivariate goodness-of-fit and two-sample testing. Technical report, Stanford Linear Accelerator Center, Menlo Park, CA (US).

FRIEDMAN, J. H. and RAFSKY, L. C. (1979). Multivariate generalizations of the Wald–Wolfowitz and Smirnov two-sample tests. *Ann. Statist.* **7** 697–717. MR0532236

GAGNON-BARTSCH, J. and SHEM-TOV, Y. (2019). The classification permutation test: A flexible approach to testing for covariate imbalance in observational studies. *Ann. Appl. Stat.* **13** 1464–1483. MR4019146 https://doi.org/10.1214/19-AOAS1241

GIRI, N. and KIEFER, J. (1964). Local and asymptotic minimax properties of multivariate tests. *Ann. Math. Stat.* **35** 21–35. MR0159388 https://doi.org/10.1214/aoms/1177703730

GIRI, N., KIEFER, J. and STEIN, C. (1963). Minimax character of Hotelling's T^2 test in the simplest case. *Ann. Math. Stat.* **34** 1524–1535. MR0156408 https://doi.org/10.1214/aoms/1177703884

GOLLAND, P. and FISCHL, B. (2003). Permutation tests for classification: Towards statistical significance in image-based studies. In *Biennial International Conference on Information Processing in Medical Imaging* 330–341. Springer, New York.

GÓMEZ, E., GÓMEZ-VILLEGAS, M. A. and MARÍN, J. M. (2003). A survey on continuous elliptical vector distributions. *Rev. Mat. Complut.* **16** 345–361. MR2031887 https://doi.org/10.5209/rev_REMA.2003.v16.n1.16889

GRETTON, A., BORGWARDT, K. M., RASCH, M. J., SCHÖLKOPF, B. and SMOLA, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13** 723–773. MR2913716

HEDIGER, S., MICHEL, L. and NÄF, J. (2019). On the use of random forest for two-sample testing. arXiv preprint, arXiv:1903.06287.

HEMERIK, J. and GOEMAN, J. J. (2018). False discovery proportion estimation by permutations: Confidence for significance analysis of microarrays. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **80** 137–155. MR3744715 https://doi.org/10.1111/rssb.12238

HENZE, N. (1988). A multivariate two-sample test based on the number of nearest neighbor type coincidences. *Ann. Statist.* **16** 772–783. MR0947577 https://doi.org/10.1214/aos/1176350835

HOTELLING, H. (1931). The generalization of student's ratio. *Ann. Math. Stat.* **2** 360–378.

HU, J. and BAI, Z. (2016). A review of 20 years of naive tests of significance for high-dimensional mean vectors and covariance matrices. *Sci. China Math.* **59** 2281–2300. MR3578957 https://doi.org/10.1007/s11425-016-0131-0

KARIYA, T. (1981). A robustness property of Hotelling's T^2-test. *Ann. Statist.* **9** 211–214. MR0600550

KIM, I., RAMDAS, A., SINGH, A. and WASSERMAN, L. (2021). Supplement to "Classification accuracy as a proxy for two-sample testing." https://doi.org/10.1214/20-AOS1962SUPP

LIU, Y., LI, C.-L. and PÓCZOS, B. (2018). Classifier two-sample test for video anomaly detections. In *British Machine Vision Conference* 2018, *BMVC* 2018 71. Northumbria Univ., Newcastle, UK.

LOPEZ-PAZ, D. and OQUAB, M. (2016). Revisiting classifier two-sample tests. arXiv preprint, arXiv:1610.06545.

LUSCHGY, H. (1982). Minimax character of the two-sample χ^2-test. *Stat. Neerl.* **36** 129–134. MR0673752 https://doi.org/10.1111/j.1467-9574.1982.tb00784.x

OLIVETTI, E., GREINER, S. and AVESANI, P. (2012). Induction in neuroscience with classification: Issues and solutions. In *Machine Learning and Interpretation in Neuroimaging* 42–50. Springer, New York.

PEREIRA, F., MITCHELL, T. and BOTVINICK, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage* **45** S199–S209.

RAUDYS, Š. and YOUNG, D. M. (2004). Results in statistical discriminant analysis: A review of the former Soviet Union literature. *J. Multivariate Anal.* **89** 1–35. MR2041207 https://doi.org/10.1016/S0047-259X(02)00021-0

ROSENBAUM, P. R. (2005). An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 515–530. MR2168202 https://doi.org/10.1111/j.1467-9868.2005.00513.x

ROSENBLATT, J. D., BENJAMINI, Y., GILRON, R., MUKAMEL, R. and GOEMAN, J. J. (2019). Better-than-chance classification for signal detection. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxz035

SALAEVSKII, O. (1969). Minimax character of Hotelling's T^2 test. I. In *Investigations in Classical Problems of Probability Theory and Mathematical Statistics* 74–101. Springer, New York.

SCHILLING, M. F. (1986). Multivariate two-sample tests based on nearest neighbors. *J. Amer. Statist. Assoc.* **81** 799–806. MR0860514

SCOTT, C. and NOWAK, R. (2005). A Neyman–Pearson approach to statistical learning. *IEEE Trans. Inf. Theory* **51** 3806–3819. MR2239000 https://doi.org/10.1109/TIT.2005.856955

SIMAIKA, J. B. (1941). On an optimum property of two important statistical tests. *Biometrika* **32** 70–80. MR0003547 https://doi.org/10.1093/biomet/32.1.70

SRIPERUMBUDUR, B. K., FUKUMIZU, K., GRETTON, A., LANCKRIET, G. R. and SCHÖLKOPF, B. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems* 1750–1758.

SRIVASTAVA, M. S. and DU, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Multivariate Anal.* **99** 386–402. MR2396970 https://doi.org/10.1016/j.jmva.2006.11.002

SRIVASTAVA, M. S., KATAYAMA, S. and KANO, Y. (2013). A two sample test in high dimensional data. *J. Multivariate Anal.* **114** 349–358. MR2993891 https://doi.org/10.1016/j.jmva.2012.08.014

STELZER, J., CHEN, Y. and TURNER, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage* **65** 69–82. https://doi.org/10.1016/j.neuroimage.2012.09.063

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

WALD, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Stat.* **15** 145–162. MR0010371 https://doi.org/10.1214/aoms/1177731280

XIAO, J., WANG, R., TENG, G. and HU, Y. (2014). A transfer learning based classifier ensemble model for customer credit scoring. In 2014 *Seventh International Joint Conference on Computational Sciences and Optimization* 64–68. IEEE.

XIAO, J., XIAO, Y., HUANG, A., LIU, D. and WANG, S. (2015). Feature-selection-based dynamic transfer ensemble model for customer churn prediction. *Knowl. Inf. Syst.* **43** 29–51.

YU, K., MARTIN, R., ROTHMAN, N., ZHENG, T. and LAN, Q. (2007). Two-sample comparison based on prediction error, with applications to candidate gene association studies. *Ann. Hum. Genet.* **71** 107–118.

ZHU, C.-Z., ZANG, Y.-F., CAO, Q.-J., YAN, C.-G., HE, Y., JIANG, T.-Z., SUI, M.-Q. and WANG, Y.-F. (2008). Fisher discriminative analysis of resting-state brain function for attention-deficit/hyperactivity disorder. *NeuroImage* **40** 110–120.

ZOGRAFOS, K. (2008). On Mardia's and Song's measures of kurtosis in elliptical distributions. *J. Multivariate Anal.* **99** 858–879. MR2405095 https://doi.org/10.1016/j.jmva.2007.05.001

ZOLLANVARI, A., BRAGA-NETO, U. M. and DOUGHERTY, E. R. (2011). Analytic study of performance of error estimators for linear discriminant analysis. *IEEE Trans. Signal Process.* **59** 4238–4255. MR2865981 https://doi.org/10.1109/TSP.2011.2159210