Ensemble Subsampling for Imbalanced Multivariate Two-Sample Tests

Lisha Chen
Department of Statistics
Yale University, New Haven,CT 06511
email: lisha.chen@yale.edu
Winston Wei Dou
Department of Financial Economics
MIT, Cambridge, MA 02139
email: wdou@mit.edu
Zhihua Qiao
Model Risk and Model Development
JPMorgan Chase, New York, NY 10172
email: zhihua.qiao@gmail.com

22 April 2013

Author's Footnote:

Lisha Chen (Email: lisha.chen@yale.edu) is Assistant Professor, Department of Statistics, Yale University, 24 Hillhouse Ave, New Haven, CT 06511. Winston Wei Dou (Email: wdou@mit.edu) is PhD candidate, Department of Financial Economics, MIT, 100 Main St, Cambridge,MA, 02139. Zhihua Qiao (Email: zhihua.qiao@gmail.com) is associate, Model Risk and Model Development, JPMorgan Chase, New York, 277 Park Avenue, New York, NY, 10172. The authors thank Joseph Chang and Ye Luo for helpful discussions. Their sincere gratitude also goes to three anonymous reviewers, an AE and the co-editor Xuming He for many constructive comments and suggestions.

Abstract

Some existing nonparametric two-sample tests for equality of multivariate distributions perform unsatisfactorily when the two sample sizes are unbalanced. In particular, the power of these tests tends to diminish with increasingly unbalanced sample sizes. In this paper, we propose a new testing procedure to solve this problem. The proposed test, based on a nearest neighbor method by Schilling (1986*a*), employs a novel ensemble subsampling scheme to remedy this issue. More specifically, the test statistic is a weighted average of a collection of statistics, each associated with a randomly selected subsample of the data. We derive the asymptotic distribution of the test statistic under the null hypothesis and show that the new test is consistent against all alternatives when the ratio of the sample sizes either goes to a finite limit or tends to infinity. Via simulated data examples we demonstrate that the new test has increasing power with increasing sample size ratio when the size of the smaller sample is fixed. The test is applied to a real data example in the field of Corporate Finance.

KEYWORDS: Corporate Finance, ensemble methods, imbalanced learning, Kolmogorov-Smirnov test, nearest neighbors methods, subsampling methods, multivariate two-sample tests.

1. INTRODUCTION

In the past decade, imbalanced data have drawn increasing attention in the machine learning community. Such data commonly arise in many fields such as biomedical science, financial economics, fraud detection, marketing, and text mining. The imbalance refers to a large difference between the sample sizes of data from two underlying distributions or from two classes in the setting of classification. In many applications, the smaller sample or the minor class is of particular interest. For example, the CoIL Challenge 2000 data mining competition presented a marketing problem where the task is to predict the probability that a customer will be interested in buying a specific insurance product. However, only 6% of the customers in the training data actually owned the policy. A more extreme example is the well-cited Mammography dataset (Woods et al. 1994), which contains 10,923 healthy patients but only 260 patients with cancer. The challenge in learning from these data is that conventional algorithms can obtain high overall prediction accuracy by classifying all data points to the majority class while ignoring the rare class that is often of greater interest. For the imbalanced classification problem, two main streams of research are sampling methods and cost-sensitive methods. He and Garcia (2009) provide a comprehensive review of existing methods in machine learning literature.

We tackle the challenges of imbalanced learning in the setting of the long-standing statistical problem of multivariate two-sample tests. We identify the issue of unbalanced sample sizes in the well-known multivariate two-sample tests based on nearest neighbors (Henze 1984; Schilling 1986*a*) as well as in two other nonparametric tests. We propose a novel testing procedure using ensemble subsampling based on the nearest neighbor method to handle the unbalanced sample sizes. We demonstrate the strong power of the testing procedure via simulation studies and a real data example, and provide asymptotic analysis for our testing procedure.

We first briefly review the problem and existing works. Two-sample tests are commonly used when we want to determine whether the two samples come from the same underlying distribution, which is assumed to be unknown. For univariate data, the standard test is the nonparametric Kolmogorov-Smirnov test. Multivariate two-sample tests have been of continuous interest to the statistics community. Chung and Fraser (1958) proposed several randomization tests. Bickel (1969) constructed a multivariate two-sample test by conditioning on the empirical distribution function

of the pooled sample. Friedman and Rafsky (1979) generalized some univariate two-sample tests, including the runs test (Wald and Wolfowitz 1940) and the maximum deviation test (Smirnoff 1939), to the multivariate setting by employing the minimal spanning trees of the pooled data. Several tests were proposed based on nearest neighbors, including Weiss (1960), Henze (1984) and Schilling (1986*a*). Henze (1988) and Henze and Penrose (1999) gave insights into the theoretical properties of some existing two-sample test procedures. More recently Hall and Tajvidi (2002) proposed a nearest neighbors-based test statistic that is particularly useful for high-dimensional problems. Baringhaus and Franz (2004) proposed a test based on the sum of interpoint distances. Rosenbaum (2005) proposed a cross-match method using distances between observations. Aslan and Zech (2005) introduced a multivariate test based on the energy between the observations in the two samples. Zuo and He (2006) provided theoretical justification for the Liu-Singh depth-based rank sum statistic (Liu and Singh 1993). Gretton et al. (2007) proposed a powerful kernel method for two-sample problem based on the maximum mean discrepancy.

Some of these existing methods for multivariate data, particularly including the tests based on nearest neighbors, the multivariate runs test, and the cross-match test, are constructed using the interpoint closeness of the pooled sample. The effectiveness of these tests assumes the two samples to be comparable in size. When the sample sizes become unbalanced, as is the case in many practical situations, the power of these tests decreases dramatically (Section 4). This near-balance assumption has also been crucial for theoretical analyses of consistency and asymptotic power of these tests.

Our new test is designed to address the problem of unbalanced sample sizes. It is built upon the nearest neighbor statistic (Henze 1984; Schilling 1986*a*), calculated as the mean of the proportions of nearest neighbors within the pooled sample belonging to the same class as the center point. A large statistic indicates a difference between the two underlying distributions. When the two samples become more unbalanced, the nearest neighbors tend to belong to the dominant sample, regardless of whether there is a difference between the underlying distributions. Consequently the power of the test diminishes as the two samples become more imbalanced. In order to eliminate the dominating effect of the larger sample, our method uses a *subsample* that is randomly drawn from the dominant sample and is then used to form a pooled sample together with the smaller

sample. We constrain the nearest neighbors to be chosen within the pooled sample resulted from subsampling.

Our test statistic is then a weighted average of a collection of statistics, each associated with a subsample. More specifically, after a subsample is drawn for each data point, a corresponding statistic is evaluated. Then these pointwise statistics are combined via averaging with appropriate weights. We call this subsampling scheme *ensemble subsampling*. Our ensemble subsampling is different from the random undersampling for the imbalanced classification problem, where only one subset of the original data is used and a large proportion of data is discarded. The ensemble subsampling enables us to make full use of the data and to achieve stronger power as the data become more imbalanced.

Ensemble methods such as bagging and boosting have been widely used for regression and classification (Hastie et al. 2009). The idea of ensemble methods is to build a model by combining a collection of simpler models which are fitted using bootstrap samples or reweighted samples of the original data. The composite model improves upon the base models in prediction stability and accuracy. Our new testing procedure is another manifestation of ensemble methods, adapting to a novel unique setting of imbalanced multivariate two-sample tests.

Moreover, we provide asymptotic analysis for our testing procedure, as the ratio of the sample sizes goes to either a finite constant or infinity. We establish an asymptotic normality result for the test statistic that does not depend on the underlying distribution. In addition, we show that the test is consistent against general alternatives and that the asymptotic power of the test increases and approaches a nonzero limit as the ratio of sample sizes goes to infinity.

The paper is organized as follows. In Section 2 we introduce notations and present the new testing procedure. Section 3 presents the theoretical properties of the test. Section 4 provides thorough simulation studies. In Section 5 we demonstrate the effectiveness of our test using a real data example. In Section 6 we provide summary and discussion. Proofs of the theoretical results are sketched in Section 7, and the detailed proofs are provided in the supplemental material.

2. THE PROPOSED TEST

In this section, we first review the multivariate two-sample tests based on nearest neighbors proposed by Schilling (1986*a*) and discuss the issue of sample imbalance. Then we introduce our new test which combines ensemble subsampling with the nearest neighbor method to resolve the issue. Lastly, we show how the ensemble subsampling can be adapted to two other nonparametric two-sample tests.

We first introduce some notation. Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_{\tilde{n}}$ be independent random samples in $\mathbb{R}^d$ generated from unknown distributions F and G, respectively. The distributions are assumed to be absolutely continuous with respect to Lebesgue measure. Their densities are denoted as f and g, respectively. The hypotheses of two-sample test can be stated as the null $H: \ F = G$ versus the alternative $K: \ F \neq G$.

We denote the two samples by $\mathcal{X} := \{X_1, \cdots, X_n\}$ and $\mathcal{Y} := \{Y_1, \cdots, Y_{\tilde{n}}\}$, and the pooled sample by $\mathcal{Z} = \mathcal{X} \cup \mathcal{Y}$. We label the pooled sample as $Z_1, \cdots, Z_m$ with $m = n + \tilde{n}$ where

$$Z_i = \begin{cases} X_i, & \text{if } i = 1, \cdots, n; \\ Y_{i-n}, & \text{if } i = n+1, \cdots, m. \end{cases}$$

For a finite set of points $\mathcal{A} \subset \mathbb{R}^d$ and a point $x \in \mathcal{A}$, let $NN_r(x, \mathcal{A})$ denote the r-th nearest neighbor (assuming no ties) of x within the set $\mathcal{A} \setminus \{x\}$. For two mutually exclusive subsets $\mathcal{A}_1, \mathcal{A}_2$ and a point $x \in \mathcal{A}_1 \cup \mathcal{A}_2$, we define an indicator function

$$I_r(x, \mathcal{A}_1, \mathcal{A}_2) = \begin{cases} 1, & \text{if } x \in \mathcal{A}_i \text{ and } NN_r(x, \mathcal{A}_1 \cup \mathcal{A}_2) \in \mathcal{A}_i, \ i = 1 \text{ or } 2 \\ 0, & \text{otherwise.} \end{cases}$$

The function $I_r(x, \mathcal{A}_1, \mathcal{A}_2)$ indicates whether x and its r-th nearest neighbor in $\mathcal{A}_1 \cup \mathcal{A}_2$ belong to the same subset.

2.1 Nearest Neighbor Method and the Problem of Imbalanced Samples

Schilling (1986*a*) proposed a class of tests for the multivariate two-sample problem based on nearest neighbors. The tests rely on the following quantity and its generalizations:

$$S_{k,n} = \frac{1}{mk} \left[\sum_{i=1}^{m} \sum_{r=1}^{k} I_r(Z_i, \mathcal{X}, \mathcal{Y}) \right]. \tag{1}$$

The test statistic $S_{k,n}$ is the proportion of pairs containing two points from the same sample, among all pairs formed by a sample point and one of its nearest neighbors in the pooled sample. Intuitively $S_{k,n}$ is small under the null hypothesis when the two samples are mixed well, while $S_{k,n}$ is large when the two underlying distributions are different. Under near-balance assumptions, Schilling (1986*a*) derived the asymptotic distribution of the test statistic under the null and showed that the test is consistent against general alternatives. The test statistic $S_{k,n}$ was further generalized by weighting each point differently based on either its rank or its value in order to improve the power of the test.

We consider the two-sample testing problem when the two sample sizes can be extremely imbalanced with $n << \tilde{n}$. We observe that the power of the test based on $S_{k,n}$ diminishes dramatically as the two sample sizes become disparate. In Figure 1, we show the numerical study of the power of the test for different sample size ratios and neighborhood sizes k in various simulation settings discussed in Section 4. We observe that the power of the test decreases as the sample size ratio increases while holding the size of $\mathcal{X}$ constant. This is obviously an undesirable property of the test because the power of a test should ideally increase with sample size. The decrease in power associated with unbalanced sample sizes can also be observed in other two-sample tests such as the run test (Friedman and Rafsky 1979) and the cross-match test (Rosenbaum 2005) (Section 4). This phenomenon can be explained by the dominance of the larger sample in the common support of the two samples. More specifically for the test based on $S_{k,n}$, when the pooled sample $\mathcal{Z}$ is dominated by the sample $\mathcal{Y}$, the summand over the sample $\mathcal{X}$, $\sum_{i=1}^{n}\sum_{r=1}^{k} I_r(Z_i, \mathcal{X}, \mathcal{Y})$, tends to be small, and the summand over the sample $\mathcal{Y}$, $\sum_{i=n+1}^{m}\sum_{r=1}^{k} I_r(Z_i, \mathcal{X}, \mathcal{Y})$, tends to be large, under both the null H and the alternative K. Therefore the test has weak power in detecting the difference between the two underlying distributions.

2.2 Ensemble Subsampling for the Imbalanced Multivariate Two-Sample Test Based on Nearest Neighbors

In order to solve the problem of unbalanced sample sizes, one could consider two simple approaches to balance the samples: one is to randomly subsample n points from $\mathcal{Y}$ to match the size of $\mathcal{X}$, and the other is to oversample $\tilde{n}$ points from $\mathcal{X}$ to match the size of $\mathcal{Y}$. In the simple subsampling, a large proportion of data points are discarded and some data information is lost. In the simple

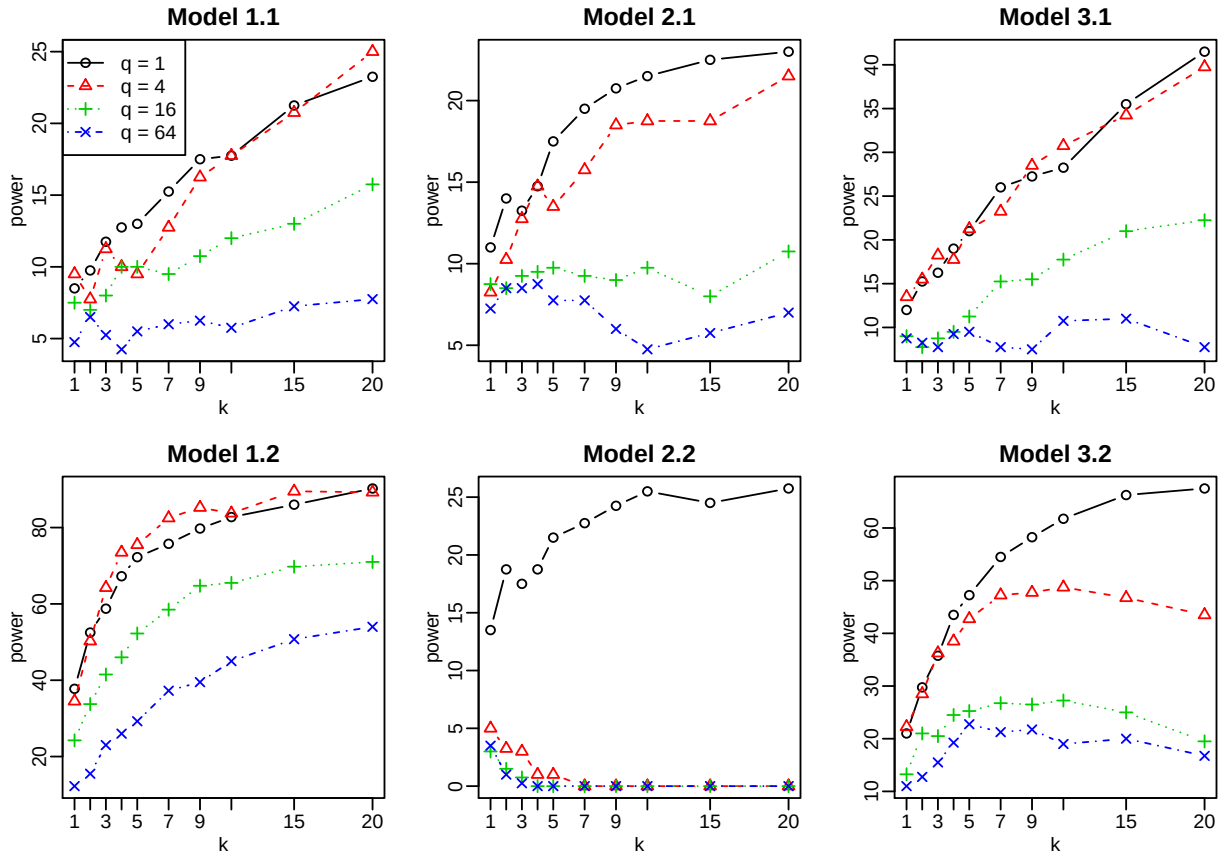

Figure 1: Simulation results representing the decreasing power of the original nearest neighbor test (1) as the ratio of the sample sizes q increases, $q = 1, 4, 16, 64$. The two samples are generated from the six simulation settings in Section 4. Power is approximated by the proportion of rejections over 400 runs of the testing procedure. A sequence of different neighborhood sizes k are used.

oversampling, the data is augmented with repeated data points and the augmented data no longer comprises of an i.i.d. sample from the true underlying distribution. There is a large amount of literature in the area of imbalanced classification regarding subsampling, oversampling and their variations (He and Garcia 2009). More sophisticated sampling methods have been proposed to improve the simple subsampling and oversampling methods, specifically for classification. However, there is no research on sampling methods for the two-sample test problem in the existing literature.

We propose a new testing procedure for multivariate two-sample tests that is immune to the unbalanced sample sizes. We use an ensemble subsampling method to make full use of the data. The idea is that for each point Z_i, $i = 1, \cdots, m$, a subsample is drawn from the larger sample $\mathcal{Y}$ and forms a pooled sample together with the smaller sample $\mathcal{X}$. We then evaluate a pointwise statistic,

the proportion of Z_i's nearest neighbors in the formed sample that belong to the same sample as Z_i. Lastly, we take average of the pointwise statistics over all Z_i's with appropriate weights. More specifically, for each Z_i, $i = 1, \cdots, m$, let $\mathcal{S}_i$ be a random subsample of $\mathcal{Y}$ of size n_s, which must contain Z_i if $Z_i \in \mathcal{Y}$. By constructions Z_i belongs to the pooled sample $\mathcal{X} \bigcup \mathcal{S}_i$, where $\mathcal{X} \bigcup \mathcal{S}_i$ is of size $n + n_s$. The pointwise statistic regarding Z_i is defined as

$$t_{k,n_s}(Z_i, \mathcal{X}, \mathcal{S}_i) = \frac{1}{k} \sum_{r=1}^{k} I_r(Z_i, \mathcal{X}, \mathcal{S}_i).$$

The statistic $t_{k,n_s}(Z_i, \mathcal{X}, \mathcal{S}_i)$ is the proportion of Z_i's nearest neighbors in $\mathcal{X} \bigcup \mathcal{S}_i$ that belong to the same sample as Z_i. The new test statistic is a weighted average of the pointwise statistics:

$$\begin{aligned} T_{k,n_s} &= \frac{1}{2n} \left[\sum_{i=1}^{n} t_{k,n_s}(Z_i, \mathcal{X}, \mathcal{S}_i) + \frac{1}{q} \sum_{i=n+1}^{m} t_{k,n_s}(Z_i, \mathcal{X}, \mathcal{S}_i) \right] \\ &= \frac{1}{2nk} \left[\sum_{i=1}^{n} \sum_{r=1}^{k} I_r(Z_i, \mathcal{X}, \mathcal{S}_i) + \frac{1}{q} \sum_{i=n+1}^{m} \sum_{r=1}^{k} I_r(Z_i, \mathcal{X}, \mathcal{S}_i) \right], \end{aligned} \tag{2}$$

where $q = \tilde{n}/n$ is the sample size ratio.

Compared with the original test statistic $S_{k,n}$ (1), this test statistic has three new features. First and most importantly, for each data point Z_i, $i = 1, \cdots, m$, a subsample $\mathcal{S}_i$ is drawn from $\mathcal{Y}$ and the nearest neighbors of Z_i are obtained in the pooled sample $\mathcal{X} \bigcup \mathcal{S}_i$. The size of subsample n_s is set to be comparable to n to eliminate the dominating effect of the larger sample $\mathcal{Y}$ in the nearest neighbors. A natural choice is to set $n_s = n$, which is the case we focus on in this paper. The second new feature is closely related to the first one, that is, a subsample is drawn separately and independently for each data point and the test statistic depends on an ensemble of all pointwise statistics corresponding to these subsamples. This is in contrast to the simple subsampling method in which only one subsample is drawn from $\mathcal{Y}$ and a large proportion of points in $\mathcal{Y}$ are discarded. The third new feature is that we introduce a weighting scheme so that the two samples contribute equally to the test. More specially, we downweight each pointwise statistic $t_{k,n_s}(Z_i, \mathcal{X}, \mathcal{S}_i)$ for $Z_i \in \mathcal{Y}$ by a factor of $1/q$ $(= n/\tilde{n})$ to balance the contributions of the two samples. The combination of these three features helps to resolve the issue of diminishing power due to the imbalanced sample sizes. We call our new test the *ensemble subsampling based on the nearest neighbor method* (ESS-NN).

Effect of Weighting and Subsampling The weighting scheme is essential to the nice properties of the new test. Alternatively, we could weigh all points equally and use the following unweighted statistic, i.e. the nearest neighbor statistic (NN) combined with subsampling without modification,

$$T^u_{k,n_s} = \frac{1}{mk}\left[\sum_{i=1}^{n}\sum_{r=1}^{k} I_r(Z_i, \mathcal{X}, \mathcal{S}_i) + \sum_{i=n+1}^{m}\sum_{r=1}^{k} I_r(Z_i, \mathcal{X}, \mathcal{S}_i)\right].$$

However our simulation study shows that, compared with T_{k,n_s}, the unweighted test T^u_{k,n_s} is less robust to general alternatives and to the choice of neighborhood sizes.

In Figure 2, we compare the power of the unweighted test (Column 3, NN+Subsampling) and the new (weighted) test (Column 4, ESS-NN) in three simulation settings (Models 1.2, 2.2, 3.2 in Section 4), where the two samples are generated from the same family of distributions with different parameters. Both testing procedures are based on the ensemble subsampling and therefore differences in results, if any, are due to the different weighting schemes. Note that the two statistics become identical when $q = 1$. The most striking contrast is in the middle row, representing the case in which we have two distributions generated from multivariate normal distributions differing only in scaling and the dominant sample has larger variance (Model 2.2). The test without weighting has nearly no power for $q = 4, 16$, and 64, while the new test with weighting improves on the power considerably. In this case the pointwise statistics of the dominant sample can, on average, have much lower power in detecting the difference between two distributions, and therefore downweighting them is crucial to the test. For the other two rows in Figure 2, even though the unweighted test seems to do well for smaller neighborhood sizes k, the weighted test outperforms the unweighted test for larger k's. Moreover, for the weighted test, the increasing trend of power versus k is consistent for all q in all simulation settings. In contrast, for the unweighted test, the trend of power versus k depends on q and varies in different settings.

Naturally, one might question the precise role played by weighting alone in the original nearest neighbor test without random subsampling. We compare NN (Column 1) with NN + Weighting (Column 2), without incorporating subsampling. The most striking difference is observed in the model 2.2 and 3.2, where the power of the weighted test improves from the original unweighted NN test. In particular, the power at $q = 4$ is smaller than that at $q = 1$ for the unweighted test but the opposite is true for the weighted test. This again indicates that the pointwise statistics of the

dominant sample on average have lower power in detecting the difference and downweighting them in the imbalanced case makes the test more powerful. However, weighting alone cannot correct the effect of the dominance of the larger sample on the pointwise statistics, which becomes more problematic at larger q's. We can see that the power of the test at $q = 16$ and 64 is lower than at $q = 4$ for NN+Weighting (Column 2). We can overcome this problem by subsampling from the larger sample and calculating pointwise statistics based on the balanced pooled sample. The role played by random subsampling alone is clearly demonstrated by comparing NN+Weighting (Column 2) and ESS-NN (Column 4).

The Size of Random Subsample The size of subsample n_s should be comparable to the smaller sample size n so that the power of the pointwise statistics (and consequently the power of the combined statistic) does not diminish as the two samples become increasingly imbalanced. Most of the work in this paper is focused on the perfectly balanced case where the subsampling size n_s is equal to n. As we will see in Section 3, the asymptotic variance formula of our test statistic is significantly simplified in this case. When $n_s \neq n$, the probability of sharing neighbors will be involved and the asymptotic variance will be more difficult to compute. Hence, $n_s = n$ seems to be the most natural and convenient choice. However, it is sensible for a practitioner to ask whether n_s can be adjusted to make the test more powerful. To answer this question, we perform simulation study for $n_s = n$, $2n$, $3n$, and $4n$ in the three multivariate settings (Models 1.2, 2.2, 3.2) considered in Section 4. See Figure 3. The results show that $n_s = n$ produces the strongest power on average and $n_s = 4n$ is the least favorable choice.

2.3 Ensemble Subsampling for Runs and Cross-match

The unbalanced sample sizes is also an issue for some other nonparametric two-sample tests such as the multivariate runs test (Friedman and Rafsky 1979) and the cross-match test (Rosenbaum 2005). In Section 4, we demonstrate the diminishing power of the multivariate runs test and the problem of over-rejection for the cross-match test as q increases. These methods are similar in that their test statistics rely on the closeness defined based on interpoint distances of the pooled sample. The dominance of the larger sample in the common support of the two samples makes these tests less powerful in detecting potential differences between the two distributions.

The idea of ensemble subsampling can also be applied to these tests to deal with the issue of imbalanced sample sizes. Here, we briefly describe how to incorporate the subsampling idea into runs and cross-match tests. The univariate runs test (Wald and Wolfowitz 1940) is based on the total number of runs in the sorted pooled sample where a run is defined as a consecutive sequence of observations from the same sample. The test rejects H for a small number of runs. Friedman and Rafsky (1979) generalized the univariate runs test to the multivariate setting by employing the minimal spanning trees of the pooled data. The analogous definition of number of runs proposed is the total number of edges in the minimal spanning tree that connect the observations from different samples, plus one. By omitting the constant 1, we can re-express the test statistic as follows,

$$\frac{1}{2}\sum_{i=1}^{m} E\left(Z_i, \mathcal{T}(\mathcal{X}\cup\mathcal{Y})\right),$$

where $\mathcal{T}(\mathcal{X}\cup\mathcal{Y})$ denotes the minimal spanning tree of the data $\mathcal{X}\cup\mathcal{Y}$, and $E(Z_i, \mathcal{T}(\mathcal{X}\cup\mathcal{Y}))$ denotes the number of observations that link to Z_i in $\mathcal{T}(\mathcal{X}\cup\mathcal{Y})$ and belong to the different sample from Z_i. The $1/2$ is a normalization constant because every edge is counted twice as we sum over the observations. As in Section 2.2, let $\mathcal{S}_i$ be a Z_i associated subsample of size n_s from $\mathcal{Y}$, which contains Z_i if $Z_i \in \mathcal{Y}$. Subsampling can be incorporated into the statistic by constructing the minimal spanning trees of the pooled sample formed by $\mathcal{X}$ and $\mathcal{S}_i$. The modified runs statistic with the ensemble subsampling can be expressed as follows:

$$\frac{1}{2}\left[\sum_{i=1}^{m} E(Z_i, \mathcal{T}(\mathcal{X}\cup\mathcal{S}_i)) + \frac{1}{q}\sum_{i=m+1}^{n} E(Z_i, \mathcal{T}(\mathcal{X}\cup\mathcal{S}_i))\right].$$

The cross-match test first matches the m observations into non-overlapping $m/2$ pairs (assuming that m is even) so that the total distance between pairs is minimized. This matching procedure is called "minimum distance non-bipartite matching". The test statistic is the number of cross-matches, i.e., pairs containing one observation from each sample. The null hypothesis would be rejected if the cross-match statistic is small. The statistic can be expressed as

$$\frac{1}{2}\sum_{i=1}^{m} C(Z_i, \mathcal{B}(\mathcal{X}\cup\mathcal{Y})),$$

where $\mathcal{B}(\mathcal{X}\cup\mathcal{Y})$ denotes the minimum distance non-bipartite matching of the pooled sample $\mathcal{X}\cup\mathcal{Y}$, and $C(Z_i, \mathcal{B}(\mathcal{X}\cup\mathcal{Y}))$ indicates whether Z_i and its paired observation in $\mathcal{B}(\mathcal{X}\cup\mathcal{Y})$ are from different

samples. Similarly the cross-match statistic can be modified as follows to incorporate the ensemble subsampling:

$$\frac{1}{2}\left[\sum_{i=1}^{n} C(Z_i, \mathcal{B}(\mathcal{X} \cup \mathcal{S}_i)) + \frac{1}{q}\sum_{i=n+1}^{m} C(Z_i, \mathcal{B}(\mathcal{X} \cup \mathcal{S}_i))\right].$$

In this subsection we have demonstrated how the ensemble subsampling can be adapted to other two-sample tests to potentially improve their power for imbalanced samples. Our theoretical and numerical studies in the rest of the paper remain focused on the ensemble subsampling based on the nearest neighbor method.

3. THEORETICAL PROPERTIES

There are some general desirable properties for an ideal two-sample test (Henze 1988). First, the ideal test has a type I error that is independent of the distribution F. Secondly, the limiting distribution of the test statistic under H is known and is independent of F. Thirdly, the ideal test is consistent against any general alternative $K: F \neq G$.

In this section, we discuss these theoretical properties of our new test in the context of imbalanced two-sample tests with possible diverging sample size ratio q. As we mentioned in Section 2.2, we focus on the case in which the subsample is of the same size as the smaller sample, that is, $n_s = n$. In the first theorem, we establish the asymptotic normality of the new test statistic (2) under the null hypothesis, which does not depend on the underlying distribution F, and we provide asymptotic values for mean and variance. In the second theorem, we show the consistency of our testing procedure.

We would like to emphasize that our results include two cases, in which the ratio of the sample sizes $q(n) = \tilde{n}/n$ goes to either a finite constant or infinity as $n \to \infty$. Let λ be the limit of the sample size ratio, $\lambda = \lim_{n\to\infty} q(n)$, with $\lambda < \infty$ and $\lambda = \infty$ representing the two cases respectively. The asymptotic power of the new testing procedure can be measured by an efficacy coefficient. We show that this coefficient increases as λ becomes larger, and approaches a nonzero limit as λ tends to infinity. This is in contrast to the original test based on $S_{k,n}$ (1) whose efficacy coefficient decreases to zero as λ goes to infinity. In the following, we first present the asymptotic mean and variance of our test statistic. In order to derive these quantities, we introduce three types of mutual and shared neighbors probabilities, and then derive two propositions regarding their relationship.

3.1 Mutual and Shared Neighbors

We consider three types of events characterizing mutual neighbors. All three types are needed here because the samples $\mathcal{X}$ and $\mathcal{Y}$ play asymmetric roles in the test and therefore need to be treated separately.

(i) mutual neighbors in $\mathcal{X}$: $NN_r(Z_1, \mathcal{X} \cup \mathcal{S}_1) = Z_2, NN_s(Z_2, \mathcal{X} \cup \mathcal{S}_2) = Z_1$;

(ii) mutual neighbors in $\mathcal{Y}$: $NN_r(Z_{n+1}, \mathcal{X} \cup \mathcal{S}_{n+1}) = Z_{n+2}, NN_s(Z_{n+2}, \mathcal{X} \cup \mathcal{S}_{n+2}) = Z_{n+1}$;

(iii) mutual neighbors between $\mathcal{X}$ and $\mathcal{Y}$: $NN_r(Z_1, \mathcal{X} \cup \mathcal{S}_1) = Z_{n+1}, NN_s(Z_{n+1}, \mathcal{X} \cup \mathcal{S}_{n+1}) = Z_1$.

Similarly we consider three types of events indicating neighbor-sharing:

(i) neighbor-sharing in $\mathcal{X}$: $NN_r(Z_1, \mathcal{X} \cup \mathcal{S}_1) = NN_s(Z_2, \mathcal{X} \cup \mathcal{S}_2)$;

(ii) neighbor-sharing in $\mathcal{Y}$: $NN_r(Z_{n+1}, \mathcal{X} \cup \mathcal{S}_{n+1}) = NN_s(Z_{n+2}, \mathcal{X} \cup \mathcal{S}_{n+2})$;

(iii) neighbor-sharing between $\mathcal{X}$ and $\mathcal{Y}$: $NN_r(Z_1, \mathcal{X} \cup \mathcal{S}_1) = NN_s(Z_{n+1}, \mathcal{X} \cup \mathcal{S}_{n+1})$.

The null probabilities for the three types of mutual neighbors are denoted by $p_{x,1}(r,s)$, $p_{y,1}(r,s)$, and $p_{xy,1}(r,s)$ and those for neighbor-sharing are denoted by $p_{x,2}(r,s)$, $p_{y,2}(r,s)$, and $p_{xy,2}(r,s)$. The following two propositions describe the values of these probabilities for large samples.

Proposition 1. *We have the following relationship between the null mutual neighbor probabilities,*

$$p_1(r,s) := \lim_{n\to+\infty} np_{x,1}(r,s) = \lim_{n\to+\infty} q(n)np_{xy,1}(r,s) = \lim_{n\to+\infty} q(n)^2 np_{y,1}(r,s),$$

where the analytical form of limit $p_1(r,s)$ (4) is given at the beginning of Section 7.

The proof is given in Section 7. The relationship between the mutual neighbor probabilities $p_{xy,1}$ and $p_{x,1}$ can be easily understood by noting that $p_{xy,1}$ involves the additional subsampling of $\mathcal{Y}$, and the probability of Z_i $(i = n+1 \cdots m)$ being chosen by subsampling is $1/q(n)$. Similar arguments apply to $p_{y,1}$ and $p_{xy,1}$. The limit $p_1(r,s)$ depends on r and s, as well as the dimension d and the limit of sample size ratio λ. $\lambda = 1$ is a special case of Schilling (1986*a*), where there is no subsampling involved and the three mutual neighbor probabilities are all equal. With $\lambda > 1$,

subsampling leads to the new mutual neighbor probabilities. Please note that n here is the size of $\mathcal{X}$, rather than the size of the pooled sample $\mathcal{Z}$. Therefore our limit $p_1(r,s)$ ranges from 0 to $\frac{1}{2}$. The rates at which $p_{x,1}$, $p_{xy,1}$ and $p_{y,1}$ approach the limit differ by a factor of $q(n)$. The limit $p_1(r,s)$ plays a key role in the calculation of the asymptotic variance. Note that as $d \to \infty$, $p_1(r,s)$ simplifies to $\begin{pmatrix} r+s-2 \\ r-1 \end{pmatrix} 2^{-(r+s)}$, which does not depend on λ. The general analytical form of $p_1(r,s)$ is rather complex and is given in (4) at the beginning of Section 7.

Proposition 2. *We have the following relationship between the null neighbor-sharing probabilities:*

$$p_{x,2}(r,s) \sim p_{xy,2}(r,s) \sim p_{y,2}(r,s), \qquad \text{as } n \to +\infty,$$

where $A_n \sim B_n$ is defined as $A_n/B_n \to 1$ as $n \to \infty$.

The proof is given in Section 7. As a side note, we can show that $np_{x,2}(r,s)$, $np_{xy,2}(r,s)$, and $np_{y,2}(r,s)$ approach the same limit as n goes to infinity. However the analytical form of this limit is rather complicated and irrelevant to the proof of the main theorems, and therefore is not given in this work.

3.2 The Asymptotic Null Distribution of The Test Statistic

In this subsection, we first give the asymptotic mean and variance of the test statistic $T_{k,n}$ under the null hypothesis H, and then present the null distribution in the main theorem.

Proposition 3. *The expectation of the test statistic $T_{k,n}$ under the null hypothesis is $\frac{1}{2}$ as n goes to infinity. More specifically*

$$\mathbb{E}_H\left(T_{k,n}\right) = \frac{n-1}{2n-1}, \quad \text{and} \quad \mu_k := \lim_{n\to+\infty} \mathbb{E}_H(T_{k,n}) = \frac{1}{2}.$$

The proof is straightforward given $\mathbb{E}_H(I_r(Z_i, \mathcal{X}, \mathcal{S}_i)) = \frac{n-1}{2n-1}$, $\forall\ i = 1, 2, \cdots, m$. Please note that the ratio q is irrelevant in either the finite sample case or the large sample case.

Proposition 4. *The asymptotic variance of the test statistic $T_{k,n}$ satisfies*

$$\sigma_k^2 = \lim_{n\to+\infty} nk\mathbb{V}ar_H(T_{k,n}) = \frac{\lambda+1}{16\lambda} + k\overline{p}_{1,k}\left(\frac{1}{16} + \frac{1}{8\lambda} + \frac{1}{16\lambda^2}\right), \tag{3}$$

where $\overline{p}_{1,k} = k^{-2}\sum_{r=1}^{k}\sum_{s=1}^{k} p_1(r,s)$, with $p_1(r,s)$ defined as in Proposition 1.

The proof is given in Section 7. The asymptotic variance depends explicitly on λ and k, and implicitly on the dimension d through average mutual neighbor probability $\overline{p}_{1,k}$, which also depends on λ and k. We numerically evaluate $\overline{p}_{1,k}$ and σ_k^2 for different combinations of λ, k and d, and observe a similar pattern of dependence. Therefore, we only present the result for σ_k^2 (Table 1). For $\forall d \leq \infty$, σ_k^2 increases as k increases slightly when λ is fixed, and σ_k^2 decreases as λ increases when k is fixed. These relationships will be useful for us to understand the dependence of asymptotic power on λ and k, which will be discussed in the next subsection.

For the case of equal sample sizes ($\lambda = 1$), our Proposition 4 agrees with Theorem 3.1 in Schilling (1986*a*) ($\lambda_1 = \lambda_2 = 1/2$). In fact, in this case our test statistic $T_{k,n}$ defined in (2) is identical to that in Schilling (1986*a*) and therefore their asymptotic variances should coincide. More precisely, we have $\overline{p}_{1,k} = \overline{p}_1'/2$ where $\overline{p}_1'$ is the notation adopted by Schilling (1986*a*, Theorem 3.1) and our σ_k^2 is actually one-half of the variance σ_k^2 defined in Schilling (1986*a*). The factor 1/2 has to do with the notation n, which represents the size of $\mathcal{X}$ in this work, versus representing the size of $\mathcal{X} \cup \mathcal{Y}$ in Schilling (1986*a*). The former is exactly 1/2 of the latter in the case of equal sample sizes.

Theorem 1. *Suppose the distribution F is absolutely continuous with respect to Lebesgue measure. Suppose $q \equiv q(n) \to \lambda \in [1, +\infty]$ as $n \to \infty$ and $q = O(n^\nu)$ for some $\nu \in (0, 1/9)$. Then $(nk)^{1/2}(T_{k,n} - \mu_k)/\sigma_k$ has a limiting standard normal distribution under the null H, where $\mu_k = 1/2$ and σ_k^2 is defined as in Proposition 4.*

This theorem shows the asymptotic normality of the null distribution. The result includes two cases in which the ratio of the sample sizes goes to either a finite constant or infinity as $n \to \infty$.

3.3 Consistency and Asymptotic Power

In Section 2.1, we discussed the problem associated with the original test statistic $S_{k,n}$ (1) in the setting of the imbalanced two-sample test and we demonstrated via simulation that the test has decreasing power with respect to increasing the sample size ratio q (or λ)(see Figure 1). In fact this problem was implied by the theoretical analysis of the test based on $S_{k,n}$ in Schilling (1986*a*), although the imbalanced data was not the focus of his work. In Section 3.2 of his paper, it was shown that $S_{k,n}$ is consistent under the general alternative K. More specifically,

$$\tilde{\Delta}(\lambda) := \liminf_{n\to\infty} (\mathbb{E}_K S_{k,n} - \mathbb{E}_H S_{k,n}) = \frac{2\lambda}{(1+\lambda)^2} \left(1 - \int \frac{f(x)g(x)\mathrm{d}x}{f(x)/(1+\lambda) + g(x)\lambda/(1+\lambda)}\right) > 0.$$

However, we can see that as λ increases, the consistency result becomes very weak. In fact, as $\lambda \to \infty$, we have $\tilde{\Delta}(\lambda) = o\left(\frac{1}{\lambda}\right)$. Moreover the asymptotic power of the test based on $S_{k,n}$ can be measured by the following efficacy coefficient

$$\begin{aligned}\tilde{\eta}(\lambda) &= \frac{\lim_{n\to\infty}\left(\mathbb{E}_K S_{k,n} - \mathbb{E}_H S_{k,n}\right)}{\lim_{n\to\infty}\left[n\mathbb{V}\mathrm{ar}_H(S_{k,n})\right]^{1/2}} \\ &= \left[1 - \int \frac{f(x)g(x)\mathrm{d}x}{f(x)/(1+\lambda) + g(x)\lambda/(1+\lambda)}\right]\left[\frac{1+\lambda}{4\lambda} + k\overline{p}'_{1,k} - k(1-\overline{p}'_{2,k})\frac{(\lambda-1)^2}{4\lambda(1+\lambda)}\right]^{-1/2} k^{1/2},\end{aligned}$$

where $\overline{p}'_{1,k}$ and $\overline{p}'_{2,k}$ are the average mutual neighbor and neighbor sharing probabilities defined in Schilling (1986*a*) (Section 3.1). This expression implies as $\lambda \to \infty$, $\tilde{\eta}(\lambda) \to 0$. Thus the asymptotic power of the test based on $S_{k,n}$ goes to zero when λ goes to infinity.

Our new test statistic $T_{k,n}$ is designed to address the issue of unbalanced sample sizes. Theorem 2 shows that our new testing procedure is consistent, and, more importantly, the consistency result does not depend on the ratio λ. Furthermore the efficacy coefficient of $T_{k,n}$ implies increasing power with respect to λ.

Theorem 2. *The test based on $T_{k,n}$ is consistent against any general alternative hypothesis K. More specifically,*

$$\lim_{n\to\infty} \mathbb{V}ar_K(T_{k,n}) = 0,$$

and

$$\Delta(\lambda) := \liminf_{n\to\infty}\left(\mathbb{E}_K T_{k,n} - \mathbb{E}_H T_{k,n}\right) > 0.$$

Moreover, $\Delta(\lambda)$ can be expressed as follows,

$$\Delta(\lambda) \equiv \frac{1}{2}\left(1 - \int \frac{f(x)g(x)dx}{f(x)/2 + g(x)/2}\right),$$

which is independent of λ.

The proof follows immediately from the results and derivations in Henze (1988, Theorem 4.1), which do not impose the requirements on the differentiability of the density functions of distributions. The details are omitted here. We also provide an alternative detailed proof, similar to Schilling (1986*a*, Theorem 3.4), which requires that the density functions are differentiable, in the supplemental article. Note that the term

$$\frac{1}{2}\int \frac{f(x)g(x)}{f(x)/2 + g(x)/2}\mathrm{d}x$$

is known as Henze-Penrose affinity; see, for example, Neemuchwala et al. (2007). If the Henze-Penrose affinity is higher, $\Delta(\lambda)$ is smaller and hence it becomes harder to test f against g. The efficacy coefficient measuring the asymptotic power of the new test is

$$\begin{aligned}\eta(\lambda) &= \frac{\lim_{n\to\infty} \mathbb{E}_K T_{k,n} - 1/2}{\lim_{n\to\infty}[n\mathbb{V}\mathrm{ar}_H(T_{k,n})]^{1/2}} \\ &= \frac{1}{2}\left(1 - \int \frac{f(x)g(x)\mathrm{d}x}{f(x)/2 + g(x)/2}\right)\left[\frac{\lambda+1}{16\lambda} + k\overline{p}_{1,k}\left(\frac{1}{16} + \frac{1}{8\lambda} + \frac{1}{16\lambda^2}\right)\right]^{-1/2} k^{1/2}.\end{aligned}$$

Note that the denominator contains the asymptotic variance $\sigma_k^2 = \left[\frac{\lambda+1}{16\lambda} + k\overline{p}_{1,k}\left(\frac{1}{16} + \frac{1}{8\lambda} + \frac{1}{16\lambda^2}\right)\right]$, which is a decreasing function of λ. This implies that the asymptotic power increases as λ increases. When λ goes to infinity, we have

$$\lim_{\lambda\to\infty} \eta(\lambda) = 2\left(1 - \int \frac{f(x)g(x)\mathrm{d}x}{f(x)/2 + g(x)/2}\right)\left(1 + k\overline{p}_{1,k}^{\infty}\right)^{-1/2} k^{1/2},$$

where $\overline{p}_{1,k}^{\infty}$ denotes the average of the mutual probabilities $\overline{p}_{1,k}$ defined in Proposition 4 for the $\lambda = \infty$ case. The expression above depends on the underlying distributions f and g, the neighborhood size k and the dimension d. The dependence on k and d is characterized by $k^{1/2}$ in the numerator and by $\left(1 + k\overline{p}_{1,k}^{\infty}\right)^{1/2}$ in the denominator. In Table 2, we give a numerical evaluation of $k\overline{p}_{1,k}^{\infty}$. It is clear that for a fixed d, $k\overline{p}_{1,k}^{\infty}$ increases with k. For a fixed k, $k\overline{p}_{1,k}^{\infty}$ increases with d when $k \geq 2$ and decreases with d when $k = 1$, which implies that the range of $k\overline{p}_{1,k}^{\infty}$ is between $\lim_{d\to\infty} k\overline{p}_{1,1}^{\infty} = 1/4$ and $\lim_{k\to\infty}\lim_{d\to\infty} k\overline{p}_{1,k}^{\infty} = 1/2$. Putting it all together, we conclude that $\left(1 + k\overline{p}_{1,k}^{\infty}\right)^{1/2}$ increases with k much slower than $k^{1/2}$. Hence the efficacy coefficient $\eta(\lambda)$ increases with k, which is consistent with the increasing power with increasing k, as observed in the simulation study (Figure 2, last column).

4. SIMULATION EXAMPLES

We first compare our new testing procedure, the ensemble subsampling based on the nearest neighbor method (**ESS-NN**), with four other testing procedures to illustrate the problem with existing methods and the limitations of a simple treatment of the problem. The first three methods are the **cross-match** method proposed by Rosenbaum (2005); the multivariate **runs** test proposed by Friedman and Rafsky (1979) which is a generalization of the univariate runs test (Wald and Wolfowitz 1940) by using the minimal spanning tree; and the original test based on nearest neighbors (**NN**) by Schilling (1986*a*). These three methods by design are not appropriate for testing

the case of two imbalanced samples. Refer to Section 2 for the detailed discussion on the problem of imbalanced samples. The last method is a simple treatment of the imbalance problem. We select a random subsample from the larger sample of the same size as the smaller sample, and then do the NN test based on the pooled sample. We call this method simple subsampling based on the nearest neighbor method (**SSS-NN**). We examine three simulation models well-studied in the existing literature, considering two sets of parameters for each model.

- Model 1: Multivariate normal with location shift. Both distributions have identity covariance matrix. They are different only in the mean vector for which we choose two sets of simulation parameters $\{d = 1, \mu_x = 0, \mu_y = 0.3\}$ (Model 1.1) and $\{d = 5, \mu_x = 0, \mu_y = 0.75\}$ (Model 1.2).

- Model 2: Multivariate normal with scale difference. The two distributions have zero mean and a scaled identity covariance matrix $\sigma^2 I_d$ for which we choose two sets of parameters, $\{d = 1, \sigma_x = 1, \sigma_y = 1.3\}$ (Model 2.1), and $\{d = 5, \sigma_x = 1, \sigma_y = 1.2\}$ (Model 2.2).

- Model 3: The multivariate random vector $X = (X_1, \ldots, X_d)$ follows the log-normal distribution. That is $\log(X_j) \sim N(\mu, 1)$, where X_j's are independent across $j = 1, \ldots, d$. The two sets of parameters are $\{d = 1, \mu_x = 0, \mu_y = 0.4\}$ (Model 3.1), and $\{d = 5, \mu_x = 0, \mu_y = 0.3\}$ (Model 3.2).

For all simulation settings, the size of the smaller sample is fixed at $n = 100$ and the ratio of the two sample sizes q equals 1, 4, 16, or 64. We conduct each testing procedure to determine whether to reject the null hypothesis at 0.05 significance level. Since the data are indeed generated from two different distributions, a powerful test should reject the null hypothesis with high probability. The critical values of all test statistics are generated using 100 permutations. In each setting, each testing procedure is repeated on 400 independently generated data sets and the proportion of rejections is reported in Table 3 to compare the power of the tests. For the new testing procedure ESS-NN, we also report the empirical type I errors in the parentheses, that is, the proportion of rejections under the null when two samples are generated from the same distributions.

In Table 3, we observed similar patterns in all simulation settings. The overall impression is that the power of runs and NN methods generally decreases with respect to the increase in the

ratio q. The power of the cross-match method does not seem to follow a particular pattern with respect to q, and in particular, with noticeable higher power ($> 60\%$) for $q = 64$ in the three settings of $d = 1$. We checked its type I errors in these settings and found that the false rejection rate to be as high as 58%, which indicates that the observed high power is due to over-rejection, and therefore is not meaningful for comparison. Intuitively the number of cross-matches under the null hypothesis converges to the size of the smaller sample n when the samples become increasingly imbalanced, which makes the test inappropriate. For the simple subsampling method, we expect that on average the power should not be sensitive to q at all because only one subsample of size n of the larger sample is utilized, and we do observe the power to be relatively stable as the ratio q increases. It is clear that only our new test based on ensemble subsampling has overall increasing power as q increases, with type I error being capped at around 0.05.

For the three tests based on nearest neighbor methods, NN, SSS-NN and ESS-NN, we report the results for the neighborhood size $k = 3$ in order to make a fair comparison with the results in Schilling (1986*a*). Both our asymptotic analysis (Section 3.3) and numerical results (Figure 2) indicate that our test is more powerful with a larger k. Our numerical results in Figure 2 suggest the increase in power become marginal after around $k = 11$. It seems wise to choose k around 11 for our new test, considering that computational cost is higher with larger k.

We then compare our method with the state-of-the-art method among two-sample tests, proposed by Gretton et al. (2007). The test statistic is based on Maximum Mean Discrepancy (MMD), namely the maximum difference of the mean function values in the two samples, over a sufficiently rich function class. Larger MMD values indicate a difference between the two underlying distributions. MMD performs strongly compared to many other two-sample tests and is not affected by the imbalance of sample sizes. We compare our method ESS-NN with MMD for Models 1.2, 2.2, and 3.2, and additional three settings for testing the normal mixtures (Table 4). ESS-NN performs as well as MMD for Models 1.2 and 3.2 especially for larger q's, and underperforms MMD for Model 2.2. We further consider the cases in which one or two of the samples are generated from a normal mixture model. In particular we consider the normal mixture consisting of two components with a probability 1/2 from each component. The two components have the same variance and $\mu_1 = -\mu_2$. In the univariate case, each normal component has the following relationship between its mean and

variance, $\sigma^2 = 1 - \mu^2$ with $\mu \in (-1, 1)$. Hence the mixture has mean 0 and variance 1. More generally we define this family of normal mixture in $\mathbb{R}^d$ with the mean vector $\mu 1_d$ and the covariance matrix $(1 - \mu^2)I_d$. We denote this family of the normal mixtures by $NM_d(\mu)$. In the last three settings presented in Table 4, ESS-NN is more powerful. In summary, even though MMD demonstrates strong performance in Models 1.2, 2.2 and 3.2 when the two underlying distributions are different in global parameters such as the mean and the variance, ESS-NN appears more sensitive to local differences in the distributions of the data. In our results of MMD, the kernel parameter is set to the median distance between points in the pooled sample, following suggestions in Gretton et al. (2007). The optimal selection of the parameter is subtle, but can potentially improve the power, and is an area of ongoing research (Gretton et al. 2012).

5. REAL DATA EXAMPLE

We consider a real data example from Corporate Finance, the study of how corporations make their decisions on financing, investment, payout, compensation, and so on. One important question in Corporate Finance is whether macroeconomic conditions and firm profitability affect the financing decisions of corporations. Financing decisions include events like issuing/repurchasing debt and equity. Among the widely accepted proxies for the macroeconomic conditions are term spread, default spread, and real equity return. Conventionally, the firm profitability is measured by the ratio between the operating income before depreciation and total assets for each quarter. Based on these variables, Korajczyk and Lévy (2003) investigated this question using the Kolmogorov-Smirnov two-sample test where the two samples are distinguishable by debt or equity repurchase. Specifically, part of their research concerns financially-unconstrained firms [1] and the firm-event window between the 1st quarter of 1985 (1985Q1) and the 3rd quarter of 1998 (1998Q3). Each observation is a firm quarter pair for which all the variables are available in the firm-event window from the well-known COMPUSTAT and CRSP databases. The data in this analysis are intrinsically imbalanced, in part because stock repurchases (equity repurchase) in the open market usually takes longer time and have a more complex completion procedure compared to the debt repurchases. In

[1] "Unconstrained firms are firms that are not labeled as constrained firms". "Constrained firms do not pay dividends, do not have a net equity or debt purchase (not both) over the event quarter, and have a Tobin's Q greater than one at the end of the event quarter" (Korajczyk and Lévy 2003).

Korajczyk and Lévy (2003), there are $n = 164$ firm quarters corresponding to equity repurchases, while there are $\tilde{n} = 1,769$ firm quarters corresponding to debt repurchases. Using the Kolmogorov-Smirnov two-sample test (KS test), the authors found that the samples are not significantly different in distribution with respect to the three macroeconomic condition indicators, which suggests that no significant association exists between each macroeconomic condition indicator and repurchasing decisions.

In this section, we examine a question similar to one considered by Korajczyk and Lévy (2003) using our new testing procedure. In addition, unlike KS test which is designed for univariate tests, our testing procedure can test multiple variables jointly. We extend the time horizon of the study with firm quarters from 1981Q1 to 2005Q4[2]. There are $n = 305$ firm quarters corresponding to equity repurchases and $\tilde{n} = 4,343$ firm quarters corresponding to debt repurchases. The variables of interest are lagged term spread, lagged credit spread, lagged real stock return, and firm profitability. We use multivariate two-sample tests to explore whether the macroeconomic conditions and profitability are jointly associated with firm repurchase activity.

For the two-sample test on the joint distribution of the four-dimensional variables, the original nearest neighbor method (Schilling 1986*a*) produces a p-value of 0.43 and our method reports a p-value smaller than 0.01, both using $k = 5$. The results are consistent across different k's, from 1 to 30 (Table 5). The significant difference can be confirmed upon visual inspection of the each of the variables separately. In Figure 4, the histograms of the two samples indeed show a difference in the univariate distributions of profitability, with noticeably long tails in the debt repurchases sample. For the univariate test on profitability, both the KS test, which is robust to imbalanced data, and our test produces p-values smaller than 0.01, whereas the p-value for the original nearest neighbor method is 0.82. This shows that our new test improves upon the original nearest neighbor test for imbalanced data. The significance of univariate test also confirms the validity of our test result for the joint distributions, as a difference between marginal distributions implies a difference between joint distributions.

[2]The raw data are from the COMPUSTAT database, the CRSP database, the Board of Governors of Federal Reserve System H.15 Database, and the U.S. Bureau of Labor Statistics CPI database. The cleaned data and R codes are available upon request

6. SUMMARY AND DISCUSSION

We addressed the issue of unbalanced sample sizes in existing nonparametric multivariate two-sample tests. We proposed a new testing procedure which combines the ensemble subsampling with the nearest neighbor method, and demonstrated the superiority of the test by both a simulation study and through real data analysis. In contrast to the original nearest neighbor test, the power of the new test increases as the sample sizes become more imbalanced. Furthermore, we provided asymptotic analysis for our testing procedure, as the ratio of the sample sizes goes to either a finite constant or infinity.

We would like to note that the imbalance in the two samples is not an issue for some existing tests including the Kolmogorov-Smirnov test for the univariate case, the test based on maximum mean discrepancy (MMD) (Gretton et al. 2007), and the Liu-Singh test (Liu and Singh 1993; Zuo and He 2006). We have discussed the test based on MMD in detail in Section 4. The Liu-Singh test uses a multivariate extension of the Wilcoxon rank sum statistic based on depth functions, and is also distribution-free. Zuo and He (2006) derived the explicit asymptotic distribution of the Liu-Singh test under both the null hypothesis and the general alternative hypothesis, as well as the asymptotic power of the test. However there is a practical drawback of the test, that is, the power of the test is sensitive to the depth function and it is difficult to select an "efficient" depth function without knowing what the alternative is.

An interesting topic for future research is to explore the dependence on the distance metric used in the nearest neighbor method. Our current analysis is based on the Euclidean distance, the most commonly used distance metric to define nearest neighbors. A systematic generalization of the Euclidean distance is to define neighborhood using the Mahalanobis metric. This treatment can be viewed as applying a linear transformation of the original sample space before conducting the test based on the Euclidean distances. Intuitively such a linear transformation can be pursued to amplify the distributional difference between the two samples both locally and globally. In this avenue, there has been continuous interest in learning the optimal distance metric for nearest neighbor classification. Hastie and Tibshirani (1996) adapted the idea of linear discriminant analysis in each neighborhood and applied local linear transformation so that the neighborhood is elongated along the most discriminant direction. Weinberger and Saul (2009) proposed a large marginal nearest

neighbor classifier that seeks a linear transformation to make the nearest neighbors share the same class labels as much as possible. In the setting of unsupervised learning, Abou-Moustafa et al. (2011) introduced (semi)-metrics based on convolution kernels for an augmented data space, which is formed by the parameters of the local Gaussian models. The intention was to relax the global Gaussian assumption under which the Euclidean distance is optimal. These ideas can potentially be borrowed to improve the power of the two-sample tests based on nearest neighbors.

Another interesting area of research is related to variation in the test statistic due to subsampling. Subsampling variation introduces another source of randomness to our test statistic. Though this should not be a concern to the effectiveness of our test as both the asymptotic theory and the permutation test have taken this variation into account, more efficient tests can be designed by reducing this variation, for example, by averaging the test statistics from multiple runs of subsampling.

7. SKETCH OF PROOFS

This section provides the sketch of proofs. Readers who are interested in our detailed proofs should refer to the supplemental materials to this paper. We write indicator function of event $\mathcal{A}$ as $\mathbf{1}_{\mathcal{A}}$.

In proposition 1

$$p_1(r,s) = \frac{1}{2}\sum_{i=0}^{h}\sum_{j=0}^{h-i}\sum_{j_1=0}^{h-i-j}\sum_{j_2=0}^{h-i-j-j_1}\binom{r+s-i-j-2}{i,j,j_1,j_2,r-i-j-j_1-1,s-i-j-j_2-1}\mathcal{Q}(\lambda,i,j,j_1,j_2) \tag{4}$$

with $h = \min(r-1, s-1)$, and for all $\lambda \in [1, +\infty]$,

$$\begin{aligned}\mathcal{Q}(\lambda,i,j,j_1,j_2) = {} & 2^{-i-j-j_1-j_2}(\lambda-1)^{j_1+j_2}\lambda^{-(j+j_1+j_2)}(1-C_d)^{i+j+j_1+j_2}C_d^{r+s-2i-2j-j_1-j_2-2} \\ & \times \left(C_d + (1-\lambda^{-1})(1-C_d)/2 + 1\right)^{-(r+s-i-j-1)},\end{aligned}$$

where $0^0 := 1$, $\infty^0 := 1$, and

$$C_d = \frac{2\Gamma(\frac{d}{2}+1)J_d}{\pi^{\frac{1}{2}}\Gamma(\frac{d+1}{2})}, \quad \text{with } J_d = \int_0^{1/2}(1-x^2)^{\frac{d-1}{2}}\mathrm{d}x.$$

Proof of proposition 1

Proof. First, we know that

$$p_{x,1}(r,s) = \frac{1}{2n-1}\mathbb{P}\Big(\{NN_r(Z_1, \mathcal{X}\cup\mathcal{S}_1) = Z_2\}|\{NN_s(Z_2, \mathcal{X}\cup\mathcal{S}_2) = Z_1\}\Big).$$

Define $\mathbf{B}^d[x,\rho]$ as the closed ball in $\mathbb{R}^d$, centered at x, which has radius ρ. We know that the surfaces of the two balls $\mathbf{B}^d[Z_1, ||Z_1 - Z_2||]$ and $\mathbf{B}^d[Z_2, ||Z_1 - Z_2||]$ pass through Z_2 and Z_1, respectively. The two balls have the same volume, which is denoted as $A_d = \pi^{d/2}||Z_1 - Z_2||^d/\Gamma(d/2+1)$. Define B_d to be the volume of the intersection of the two balls, that is, $\mathbf{B}^d[Z_1, ||Z_1 - Z_2||]\cap\mathbf{B}^d[Z_2, ||Z_1 - Z_2||]$. Define $C_d := (A_d - B_d)/A_d$. It is easy to see that $B_d \to 0$ and $C_d \to 1$ as $d \to \infty$, .

According to Schilling (1986*b*, Theorem 2.1) and Henze (1987, Theorem 1.1 and Lemmas in its proof), we know that to analyze the asymptotic conditional probability of the mutual neighbors,

$$\mathbb{P}\Big(\{NN_r(Z_1, \mathcal{X}\cup\mathcal{S}_1) = Z_2\}|\{NN_s(Z_2, \mathcal{X}\cup\mathcal{S}_2) = Z_1\}\Big)$$

, as n approaches infinity, $Z_1, \cdots, Z_m$ can be viewed as samples from a homogeneous Poisson process with intensity τ. The exact value of τ is not important here because under the null hypothesis the two distributions are equal and hence the effect of τ will be canceled out.

> **Remark.** The problem of computing the mutual neighbor probabilities has been studied extensively in the literature. Clark and Evans (1955), Clark (1955), Cox (1981), Pickard (1982), and Henze (1986), among others, analyzed this problem in the case of homogeneous Poisson processes. Schilling (1986*b*) found the limits of the mutual neighbor probabilities for i.i.d. case as the sample size goes to infinity. However, the author did not rigorously bridge the gap between the homogeneous-Poisson-process case and the i.i.d.-sample case, and assumed that they are equivalent in limit for this particular local problem. Henze (1987) rigorously established the asymptotic equivalence result in weak convergence. Without repeating the exact steps in the proofs to Theorem 1.1, Lemma 2.1, and Lemma 2.2 in Henze (1987), we can directly use the asymptotic equivalence results developed in that paper.

According to (Cox 1981, Page 368), it follows that given that Z_1 is the s-th nearest neighbor to Z_2 in $\mathcal{X}\cap\mathcal{S}_2$, A_d has the distribution with the following density:

$$f(A;s) = (\frac{2\tau}{1+\lambda})^s A^{s-1}\exp(-\tau 2A/(1+\lambda))/(s-1)!, \quad A > 0.$$

Now consider three sub-Poisson processes $\mathcal{B}_1 \equiv \mathcal{S}_1 - \mathcal{S}_2, \mathcal{B}_2 \equiv \mathcal{S}_2 - \mathcal{S}_1, \mathcal{C} = \mathcal{S}_1\cap\mathcal{S}_2$. The intensities of Poisson processes are $\tau_{\mathcal{B}_1} = \tau_{\mathcal{B}_2} = \frac{\tau}{1+\lambda}\left(1 - \frac{1}{\lambda}\right)$ and $\tau_{\mathcal{C}} = \frac{\tau}{1+\lambda}\frac{1}{\lambda}$. Given that the volume is A

and there are i points of $\mathcal{X}$ and j_2 points of $\mathcal{B}_2$ and j points of $\mathcal{C}$ falling in the intersection of the two balls, the conditional probability that Z_2 is the r-th nearest neighbor to Z_1 is given by

$$g(i,j,j_2;A) = \sum_{j_1=0}^{h-i-j-j_2} \frac{1}{(r-i-j-j_1-1)!}\left(\frac{2\tau C_d A}{1+\lambda}\right)^{r-i-j-j_1-1} e^{-\frac{2\tau C_d A}{1+\lambda}} \frac{1}{j_1!}\left(\frac{\lambda-1}{\lambda}\frac{\tau(1-C_d)A}{1+\lambda}\right)^{j_1} e^{-\frac{\lambda-1}{\lambda}\frac{\tau(1-C_d)A}{1+\lambda}},$$

where $\frac{1}{(r-i-j-j_1-1)!}\left(\frac{2\tau C_d A}{1+\lambda}\right)^{r-i-j-j_1-1}\exp\left(-\frac{2\tau C_d A}{1+\lambda}\right)$ is the probability that the Poisson process $\mathcal{X}\cup\mathcal{S}_1$ with intensity $\frac{2\tau}{1+\lambda}$ has $r-i-j-j_1-1$ points lying in the region $\mathbf{B}^d[Z_1,||Z_1-Z_2||]\setminus \mathbf{B}^d[Z_2,||Z_1-Z_2||]$, and $\frac{1}{j_1!}\left(\frac{\lambda-1}{\lambda}\frac{\tau(1-C_d)A}{1+\lambda}\right)^{j_1}\exp\left(-\frac{\lambda-1}{\lambda}\frac{\tau(1-C_d)A}{1+\lambda}\right)$ is the probability that the Poisson process $\mathcal{B}_1$ has j_1 points lying in the region $\mathbf{B}^d[Z_1,||Z_1-Z_2||]\cap\mathbf{B}^d[Z_2,||Z_1-Z_2||]$.

Hence the (conditional) probability, $P_n(r,s)$, that Z_2 is the r-th nearest neighbor to its own s-th nearest neighbor Z_1 is given by

$$P_n(r,s) = \int_0^\infty \left\{ \sum_{i=0}^{h}\sum_{j=0}^{h-i}\sum_{j_2=0}^{h-i-j} \frac{(s-1)!}{i!j!j_2!(s-1-i-j-j_2)!}\left(\frac{1-C_d}{2}\right)^i\left(\frac{1-C_d}{2\lambda}\right)^j \left(\frac{1-C_d}{2}\left(1-\frac{1}{\lambda}\right)\right)^{j_2} C_d^{s-i-j-j_2-1} g(i,j,j_2;A)\right\} f(A;s)\mathrm{d}A,$$

where $h := \min(r-1, s-1)$. So, we get

$$\begin{aligned} P_n(r,s) =& \sum_{i=0}^{h}\sum_{j=0}^{h-i}\sum_{j_2=0}^{h-i-j}\sum_{j_1=0}^{h-i-j-j_2} \binom{r+s-i-j-2}{i,j,j_1,j_2,r-i-j-j_1-1,s-i-j-j_2-1} 2^{-(i+j+j_1+j_2)} \\ &\times (C_d+(1-C_d)(1-1/\lambda)/2+1)^{-(r+s-i-j-1)}(\lambda-1)^{j_1+j_2}\lambda^{-(j+j_1+j_2)} \\ &\times (1-C_d)^{i+j+j_1+j_2} C_d^{r+s-2i-2j-j_1-j_2-2}. \end{aligned}$$

Therefore, $\lim_{n\to+\infty} np_{x,1}(r,s) = \lim_{n\to\infty}\frac{n}{2n-1}P_n(r,s) = p_1(r,s)$.

Note that

$$\begin{aligned} p_{y,1}(r,s) =& \frac{(n-1)^2}{(2n-1)(qn-1)^2} \\ &\times \mathbb{P}\Big(\{\{NN_r(Z_{n+1},\mathcal{X}\cup\mathcal{S}_{n+1}) = Z_{n+2}\}|\{NN_s(Z_{n+2},\mathcal{X}\cup\mathcal{S}_{n+2}) = Z_{n+1}, Z_{n+2}\in\mathcal{S}_{n+1}\}\}\Big), \end{aligned}$$

and

$$p_{xy,1}(r,s) = \frac{n-1}{(2n-1)(qn-1)} \times \mathbb{P}\Big(\{NN_r(Z_1, \mathcal{X} \cup \mathcal{S}_1) = Z_{n+1}\} | \{NN_s(Z_{n+1}, \mathcal{X} \cup \mathcal{S}_{n+1}) = Z_1, Z_{n+1} \in \mathcal{S}_1\}\Big).$$

Using similar arguments, we can analyze the asymptotic behavior of the conditional probability above, and then, show that $\lim_{n\to+\infty} nq^2 p_{y,1}(r,s) = p_1(r,s)$ and $\lim_{n\to+\infty} nqp_{xy,1}(r,s) = p_1(r,s)$.

□

Proof of Proposition 2

Proof. We have

$$\begin{aligned} p_{y,2}(r,s) &\equiv \mathbb{P}\left(\{NN_r(Z_{n+1}, \mathcal{X} \cup \mathcal{S}_{n+1}) = NN_s(Z_{n+2}, \mathcal{X} \cup \mathcal{S}_{n+2})\}\right) \\ &\sim \mathbb{P}\left(\{NN_r(Z_{n+1}, \mathcal{X} \cup \mathcal{S}_{n+1} \cup \{Z_{n+2}\}) = NN_s(Z_{n+2}, \mathcal{X} \cup \mathcal{S}_{n+2} \cup \{Z_{n+1}\})\}\right) \\ &\sim p_{x,2}(r,s). \end{aligned}$$

Similarly, we have

$$\begin{aligned} p_{xy,2}(r,s) &\equiv \mathbb{P}\left(\{NN_r(Z_1, \mathcal{X} \cup \mathcal{S}_1) = NN_s(Z_{n+1}, \mathcal{X} \cup \mathcal{S}_{n+1})\}\right) \\ &\sim \mathbb{P}\left(\{NN_r(Z_1, \mathcal{X} \cup \mathcal{S}_1 \cup \{Z_{n+1}\}) = NN_s(Z_{n+1}, \mathcal{X} \cup \mathcal{S}_{n+1} \cup \{Z_1\})\}\right) \\ &\sim p_{x,2}(r,s). \end{aligned}$$

□

Proof of Proposition 4

Proof. We denote the index sets of the two samples by $\Omega_x = \{1, \cdots, n\}$ and $\Omega_y = \{n+1, \cdots, m\}$, with $m = n + \tilde{n}$. We know that

$$\mathbb{V}\mathrm{ar}_H(mkT_{k,n}) = \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{r=1}^{k}\sum_{s=1}^{k} w_i w_j \mathbb{P}_H\Big(I_r(Z_i, \mathcal{X}, \mathcal{S}_i) = I_s(Z_j, \mathcal{X}, \mathcal{S}_j) = 1\Big) - \big(mk\mathbb{E}_H(T_{k,n})\big)^2, \quad (5)$$

where $w_i = \frac{1+q}{2}$ for $i \in \Omega_x$ and $\frac{1+q}{2q}$ for $i \in \Omega_y$. For terms in which $i = j$, we know that

$$\mathbb{P}_H\Big(I_r(Z_i, \mathcal{X}, \mathcal{S}_i) = I_s(Z_j, \mathcal{X}, \mathcal{S}_j) = 1\Big) = \mathbf{1}_{\{r=s\}}\left(\frac{1}{2} - \frac{1}{4n}\right) + \mathbf{1}_{\{r\neq s\}}\left(\frac{1}{4} - \frac{3}{8n}\right). \quad (6)$$

For each term in which (1) $i \neq j \in \Omega_x$, or (2) $i \neq j \in \Omega_y$, or (3) $i \in \Omega_x,\ j \in \Omega_y$, there are always five mutually exclusive and exhaustive cases involved:

(i) $NN_r(Z_i, \mathcal{X} \cup \mathcal{S}_i) = Z_j, NN_s(Z_j, \mathcal{X} \cup \mathcal{S}_j) = Z_i$;

(ii) $NN_r(Z_i, \mathcal{X} \cup \mathcal{S}_i) = NN_s(Z_j, \mathcal{X} \cup \mathcal{S}_j)$;

(iii) $NN_r(Z_i, \mathcal{X} \cup \mathcal{S}_i) = Z_j$, but $NN_s(Z_j, \mathcal{X} \cup \mathcal{S}_j) \neq Z_i$;

(iv) $NN_r(Z_i, \mathcal{X} \cup \mathcal{S}_i) \neq Z_j$, but $NN_s(Z_j, \mathcal{X} \cup \mathcal{S}_j) = Z_i$;

(v) $NN_r(Z_i, \mathcal{X} \cup \mathcal{S}_i) \neq Z_j, NN_s(Z_j, \mathcal{X} \cup \mathcal{S}_j) \neq Z_i$, and $NN_r(Z_i, \mathcal{X} \cup \mathcal{S}_i) \neq NN_s(Z_j, \mathcal{X} \cup \mathcal{S}_j)$.

Let the null probabilities of these events be denoted by $p_{x,i}(r,s)$, $p_{y,i}(r,s)$, and $p_{xy,i}(r,s)$, respectively, for the three scenarios, where $i = 1, \cdots, 5$. Therefore, we have for $i \neq j$,

$$
\begin{aligned}
&\mathbb{P}_H(I_r(Z_i, \mathcal{X}, \mathcal{S}_i) \doteq \mathbf{1}_{\{i,j\in\Omega_x\}} p_{x,1}(r,s) + \mathbf{1}_{\{i,j\in\Omega_y\}} p_{y,1}(r,s) \\
&+ \mathbf{1}_{\{i,j\in\Omega_x\}}(1 - \frac{1}{1+q} - \frac{2q}{(1+q)^2 n}) p_{x,2}(r,s) + \mathbf{1}_{\{i,j\in\Omega_y\}}(\frac{1}{1+q} - \frac{2q}{(1+q)^2 n}) p_{y,2}(r,s) \\
&+ \mathbf{1}_{\{i,j\in\Omega_x\}}(\frac{1}{2} - \frac{1}{2n})(\frac{1}{2n} - p_{x,1}(r,s)) + \mathbf{1}_{\{i,j\in\Omega_y\}}(\frac{1}{2} - \frac{1}{4n} - \frac{1}{4qn})(\frac{1}{2qn} - p_{y,1}(r,s)) \\
&+ \mathbf{1}_{\{i,j\in\Omega_x\}}(\frac{1}{2} - \frac{1}{2n})(\frac{1}{2n} - p_{x,1}(r,s)) + \mathbf{1}_{\{i,j\in\Omega_y\}}(\frac{1}{2} - \frac{1}{4n} - \frac{1}{4qn})(\frac{1}{2qn} - p_{y,1}(r,s)) \\
&+ \mathbf{1}_{\{i,j\in\Omega_x\}}(\frac{1}{4} - \frac{11}{16n} + \frac{1}{16qn}) \left(1 - \frac{1}{n} + p_{x,1}(r,s) - p_{x,2}(r,s)\right) \\
&+ \mathbf{1}_{\{i,j\in\Omega_y\}}(\frac{1}{4} - \frac{3}{16n} - \frac{7}{16nq}) \left(1 - \frac{1}{qn} + p_{y,1}(r,s) - p_{y,2}(r,s)\right) \\
&+ 2 \times \mathbf{1}_{\{i\in\Omega_x, j\in\Omega_y\}}(\frac{1}{4} - \frac{1}{16n} + \frac{3}{16nq}) \left(1 - \frac{1}{2n} - \frac{1}{2qn} + p_{xy,1}(r,s) - p_{xy,2}(r,s)\right). \qquad (7)
\end{aligned}
$$

We plug the long equation (7) together with (6) into the formula of the asymptotic variance (5), and then after re-arranging the terms we can achieve the result of the theorem. □

Proof of Theorem 1

Proof. In order to invoke (Chatterjee 2008, Theorem 3.4), we write

$$
f_i(z_1, \cdots, z_m) = \begin{cases} \frac{1}{2k} \sum_{r \leq k} I_r(z_i, \mathcal{X}, \mathcal{S}_i) & \text{if } 1 \leq i \leq n; \\ \frac{1}{2qk} \sum_{r \leq k} I_r(z_i, \mathcal{X}, \mathcal{S}_i) & \text{if } n+1 \leq i \leq m. \end{cases}
$$

Define

$$G_{k,n} = \frac{1}{\sqrt{m}} \sum\nolimits_{i \le m} f_i(Z_1, \cdots, Z_m) = \frac{\sqrt{m}}{1+q} T_{k,n},$$

and

$$W_{k,n} = \frac{G_{k,n} - \mathbb{E}G_{k,n}}{\sigma(G_{k,n})} = \frac{T_{k,n} - \mathbb{E}T_{k,n}}{\sigma(T_{k,n})}.$$

After re-arranging terms we have

$$(nk)^{1/2} \left(T_{k,n} - \mu_k\right) / \sigma_k = \frac{\sigma(T_{k,n})(nk)^{1/2}}{\sigma_k} W_{k,n} + \frac{(nk)^{1/2}(\mathbb{E}(T_{k,n}) - \mu_k)}{\sigma_k}.$$

According to Propositions 3 and 4, we know that

$$\frac{\sigma(T_{k,n})(nk)^{1/2}}{\sigma_k} \to 1 \quad \text{and} \quad \frac{(nk)^{1/2}(\mathbb{E}(T_{k,n}) - \mu_k)}{\sigma_k} \to 0, \quad \text{as } n \to \infty.$$

Thus, it suffices to show that $\mathbb{P}\left(W_{k,n} \le x\right) \to \Phi(x), \quad \forall\ x \in \mathbb{R}$. For a constant $\zeta \in (0,1)$ that is small enough such that $4.5\nu + 4\zeta < 1/2$ and $\nu + 2\zeta < 1$, we define

$$K(n) := k(1+q)n^{\zeta}. \tag{8}$$

We focus on the big probability set $\mathcal{A}_n$ on which for all Z_i, the k nearest neighbors among $\mathcal{X} \cup \mathcal{S}_i$ are in its $K(n)$ nearest neighbors among $\mathcal{X} \cup \mathcal{Y}$, that is, $\mathcal{A}_n = \cap_{i \le n} \mathcal{A}_{n,i}$, where $\mathcal{A}_{n,i} := \left\{\omega \mid \cup_{r \le k} NN_r(Z_i, \mathcal{X} \cup \mathcal{S}_i) \subseteq \cup_{r \le K(n)} NN_r(Z_i, \mathcal{X} \cup \mathcal{Y})\right\}$. Then, we can get

$$\begin{aligned}
\mathbb{P}\mathcal{A}_n^c &\le m\mathbb{P}\mathcal{A}_{n,1}^c = m(1 - \mathbb{P}\mathcal{A}_{n,1}) \\
&\le m(1 - \mathbb{P}(\text{there are at least } k \text{ points of } \mathcal{S}_1 \text{ lying in} \\
&\qquad \text{the } K(n) \text{ nearest neighbors of } Z_1 \text{ among } \mathcal{Y})) \quad \text{(see below for more explanations)} \\
&= m\mathbb{P}(\text{there are at most } k-1 \text{ points of } \mathcal{S}_1 \text{ lying in} \\
&\qquad \text{the } K(n) \text{ nearest neighbors of } Z_1 \text{ among } \mathcal{Y}) \\
&\le mk \binom{K(n)}{k-1} \binom{nq - K(n)}{n-k+1} \Big/ \binom{nq}{n} \\
&= O\left(nq^{2-k} K(n)^{k-1} a(\lambda)^{K(n)/(1+q)}\right) = o\left(n^{k+\nu} a(\lambda)^{kn^{\zeta}}\right) = o\left(1\right),
\end{aligned} \tag{9}$$

where $a(\lambda) \equiv (1 - 1/(1+\lambda))^{1+\lambda}$ is a constant on $(0,1)$. The second inequality above (9) is due to the fact that $\mathcal{B}_{n,1} := \{\text{at least } k \text{ points of } \mathcal{S}_1 \text{ lie in the } K(n) \text{ nearest neighbors of } Z_1 \text{ among } \mathcal{Y}\}$ and $\mathcal{B}_{n,1} \subseteq \mathcal{A}_{n,1}$. More precisely, suppose event $\mathcal{B}_{n,1}$ holds and consider the $K(n)$ nearest neighbors

of Z_1 among the points of $\mathcal{Y}$. The $K(n)$ balls are colored black. Each of these balls is recolored red (covering the original black color) if it belongs to $\mathcal{S}_1$. Therefore, at least k of these $K(n)$ balls are red (i.e. the event $\mathcal{B}_{n,1}$ holds). Now, let us focus on the $K(n)$ nearest neighbors of Z_1 among the points of the bigger set $\mathcal{X} \cup \mathcal{Y}$, which is a set of balls not necessarily identical to the previously colored $K(n)$ balls, with all other $m+n-K(n)-1$ points eliminated. Each of these balls is colored yellow if it belongs to $\mathcal{X}$ and is kept as red if it belongs to $\mathcal{S}_1 \subset \mathcal{Y}$; otherwise it is colored black as before. Some of the black balls and red balls of the original arrangement may now have been eliminated by being recolored as yellow. The key point is that the number of black and red balls that are eliminated equals to the number of yellow balls that are added. Therefore, the number of eliminated red balls is less than or equal to the number of added yellow balls. Thus, we have at least k yellow and red balls after adding yellow balls and eliminating red/black balls (i.e. $\mathcal{A}_{n,1}$ holds). Therefore, we have proved $\mathcal{B}_{n,1} \subseteq \mathcal{A}_{n,1}$[3].

Denote $F_n(x) := \mathbb{P}(W_{k,n} \leq x | \mathcal{A}_n)$ and denote $\epsilon_n := d_L(F_n, \Phi)$, the Lévy distance between F_n and Φ. By definition of the Lévy distance and the Mean Value Theorem, we have

$$F_n(x) - \Phi(x) \leq \Phi(x + \epsilon_n) + \epsilon_n - \Phi(x) \leq \left(1 + \frac{1}{2\pi}\right) \epsilon_n,$$
$$F_n(x) - \Phi(x) \geq \Phi(x - \epsilon_n) - \epsilon_n - \Phi(x) \geq -\left(1 + \frac{1}{2\pi}\right) \epsilon_n.$$

Thus,

$$|F_n(x) - \Phi(x)| \leq \left(1 + \frac{1}{2\pi}\right) \epsilon_n. \tag{10}$$

From (Huber 1981, Page 33-34), we know that the following relation between the Lévy distance and the Wasserstein (or the Kantorovich) distance,

$$\epsilon_n \leq \sqrt{d_W(F_n, \Phi)}, \tag{11}$$

where $d_W(F_n, \Phi)$ is the Wasserstein (or Kantorovich) distance between F_n and Φ. Given the set $\mathcal{A}_n$, we know that each function f_i only depends on the $K(n)$ nearest neighbors of the point z_i. Moreover, based on Proposition 4, it follows that $\sigma(G_{k,n}) \asymp 1/\sqrt{q}$. By the definition of $K(n)$ in (8) and the assumption on q, we know that $K(n) = O\left(n^{\nu+\zeta}\right)$. For the large constant p such that

[3]This relatively short and conceptual proof is suggested by one of our anonymous referees. An alternative proof which is more explicit can be found in the supplemental materials

$4.5\nu + 4\zeta < (p - 8 - 8\nu)/(2p)$, we invoke Theorem 3.4 in (Chatterjee 2008) directly to get the following bound,

$$\begin{aligned}
|F_n(x) - \Phi(x)| &\leq \left(1 + \frac{1}{2\pi}\right)\epsilon_n \leq \left(1 + \frac{1}{2\pi}\right)\sqrt{d_W(F_n, \Phi)} \\
&\leq C\frac{K(n)^2}{\sigma(G_{k,n})(n(1+q))^{(p-8)/(4p)}} + C\frac{K(n)^{3/2}}{\sigma^{3/2}(G_{k,n})(n(1+q))^{(p-6)/(4p)}} \\
&\leq C'K(n)^2 n^{-(p-8)/(4p)} q^{1/2-(p-8)/(4p)} + C'K(n)^{3/2} n^{-(p-6)/(4p)} q^{3/4-(p-6)/(4p)} \\
&\leq C''n^{2.25\nu+2\zeta-(p-8-8\nu)/(4p)} + C''n^{2.25\nu+1.5\zeta-(p-6)/(4p)} = o(1),
\end{aligned}$$

where C, C', and C'' are universal constants and the first two inequalities result from (10) and (11), respectively. Because $\mathbb{P}(W_{k,n} \leq x) = \mathbb{P}(\mathcal{A}_n)\mathbb{P}(W_{k,n} \leq x|\mathcal{A}_n) + \mathbb{P}(\mathcal{A}_n^c)\mathbb{P}(W_{k,n} \leq x|\mathcal{A}_n^c)$, then we have $\mathbb{P}(W_{k,n} \leq x) \to \Phi(x), \quad \forall\, x \in \mathbb{R}$. □

REFERENCES

Abou-Moustafa, K., Shah, M., De La Torre, F., and Ferrie, F. (2011), "Relaxed Exponential Kernels for Unsupervised Learning," *Pattern Recognition*, pp. 184–195.

Aslan, B., and Zech, G. (2005), "New Test for the Multivariate Two-Sample Problem Based on the Concept of Minimum Energy," *Jour. Statist. Comp. Simul.*, 75(2), 109–119.

Baringhaus, L., and Franz, C. (2004), "On a New Multivariate Two-sample Test," *Journal of Multivariate Analysis*, 88(1), 190–206.

Bickel, P. J. (1969), "A Distribution Free Version of the Smirnov Two Sample Test in the p-variate Case," *Ann. Math. Statist.*, 40, 1–23.

Chatterjee, S. (2008), "A New Method of Normal Approximation," *Ann. Probab.*, 36(4), 1584–1610.

Chung, J., and Fraser, D. (1958), "Randomization Tests for a Multivariate Two-sample Problem," *Journal of the American Statistical Association*, 53(283), 729–735.

Clark, P. J. (1955), "Grouping in Spatial Distributions," *Science*, 123, 373 – 374.

Clark, P. J., and Evans, F. C. (1955), "On Some Aspects of Spatial Pattern in Biological Populations," *Science*, 121(3142), 397 – 398.

Cox, T. F. (1981), "Reflexive Nearest Neighbours," *Biometrics*, 37(2), 367–369.

Friedman, J. H., and Rafsky, L. C. (1979), "Multivariate Generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Ann. Statist.*, 7(4), 697–717.

Gretton, A., Borgwardt, K., Rasch, M., Schlkopf, B., and Smola, A. (2007), "A Kernel Method for the Two Sample Problem," *Advances in Neural Information Processing Systems 19*, pp. 513–520.

Gretton, A., Borgwardt, K., Rasch, M., Scholkopf, B., and Smola, A. (2012), "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, 16, 723–773.

Hall, P., and Tajvidi, N. (2002), "Permutation Tests for Equality of Distributions in High-Dimensional Settings," *Biometrika*, 89(2), 359–374.

Hastie, T., and Tibshirani, R. (1996), "Discriminant Adaptive Nearest Neighbor Classification," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(6), 607–616.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data mining, Inference, and Prediction*, Springer Series in Statistics, 2nd edn, New York: Springer-Verlag.

He, H., and Garcia, E. (2009), "Learning from Imbalanced Data," *Knowledge and Data Engineering, IEEE Transactions on*, 21(9), 1263–1284.

Henze, N. (1984), "On the Number of Random Points with Nearest Neighbour of the Same Type and a Multivariate Two-Sample Test (in German)," *Metrika*, 31, 259–273.

Henze, N. (1986), "On the Probability That a Random Point Is the jth Nearest Neighbour to Its Own kth Nearest Neighbour," *J. Appl. Prob.*, 23(1), 221–226.

Henze, N. (1987), "On the Fraction of Random Points with Specified Nearest-Neighbour Interrelations and Degree of Attraction," *Adv. in Appl. Probab.*, 19(4), 873–895.

Henze, N. (1988), "A Multivariate Two-Sample Test Based on the Number of Nearest Neighbor Type Coincidences," *Ann. Statist.*, 16(2), 772–783.

Henze, N., and Penrose, M. (1999), "On the Multivariate Run Test," *Ann. Statist.*, 27(1), 290–298.

Huber, P. J. (1981), *Robust statistics*, New York: John Wiley & Sons Inc. Wiley Series in Probability and Mathematical Statistics.

Korajczyk, R. A., and Lévy, A. (2003), "Capital Structure Choice: Macroeconomic Conditions and Financial Constraints," *Journal of Financial Economics*, 68(1), 75–109.

Liu, R., and Singh, K. (1993), "A Quality Index Based on Data Depth and Multivariate Rank Tests," *Journal of the American Statistical Association*, pp. 252–260.

Neemuchwala, H., Hero, A., Zabuawala, S., and Carson, P. (2007), "Image Registration Methods in High-Dimensional Space," *Int. J. of Imaging Syst. and Techn.*, 16, 130145.

Pickard, D. K. (1982), "Isolated Nearest Neighbors," *J. Appl. Probab.*, 19(2), 444–449.

Rosenbaum, P. (2005), "An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency," *Journal of the Royal Statistical Society. Series B*, 67(4), 515–530.

Schilling, M. F. (1986*a*), "Multivariate Two-sample Tests Based on Nearest Neighbors," *J. Amer. Statist. Assoc.*, 81(395), 799–806.

Schilling, M. F. (1986*b*), "Mutual and Shared Neighbor Probabilities: Finite- and Infinite-Dimensional Results," *Adv. in Appl. Probab.*, 18(2), 388–405.

Smirnoff, N. (1939), "On the Estimation of the Discrepancy between Empirical Curves of Distribution for Two Independent Samples," *Bulletin de lUniversite de Moscow, Serie internationale (Mathematiques)*, 2, 3–14.

Wald, A., and Wolfowitz, J. (1940), "On a Test Whether Two Samples are from the Same Population," *The Annals of Mathematical Statistics*, 11(2), 147–162.

Weinberger, K., and Saul, L. (2009), "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *The Journal of Machine Learning Research*, 10, 207–244.

Weiss, L. (1960), "Two-sample Tests For Multivariate Distributions," *The Annals of Mathematical Statistics*, 31(1), 159–164.

Woods, K., Solks, J., Priebe, C., Kegelmeyer, W., Doss, C., and Bowyer, K. (1994), "Comparative Evaluation of Pattern Recognition Techniques for Detection of Microcalcifications in Mammography. State of The Art in Digital Mammographic Image Analysis,".

Zuo, Y., and He, X. (2006), "On the Limiting Distributions of Multivariate Depth-based Rank Sum Statistics and Related Tests," *The Annals of Statistics*, 34(6), 2879–2896.

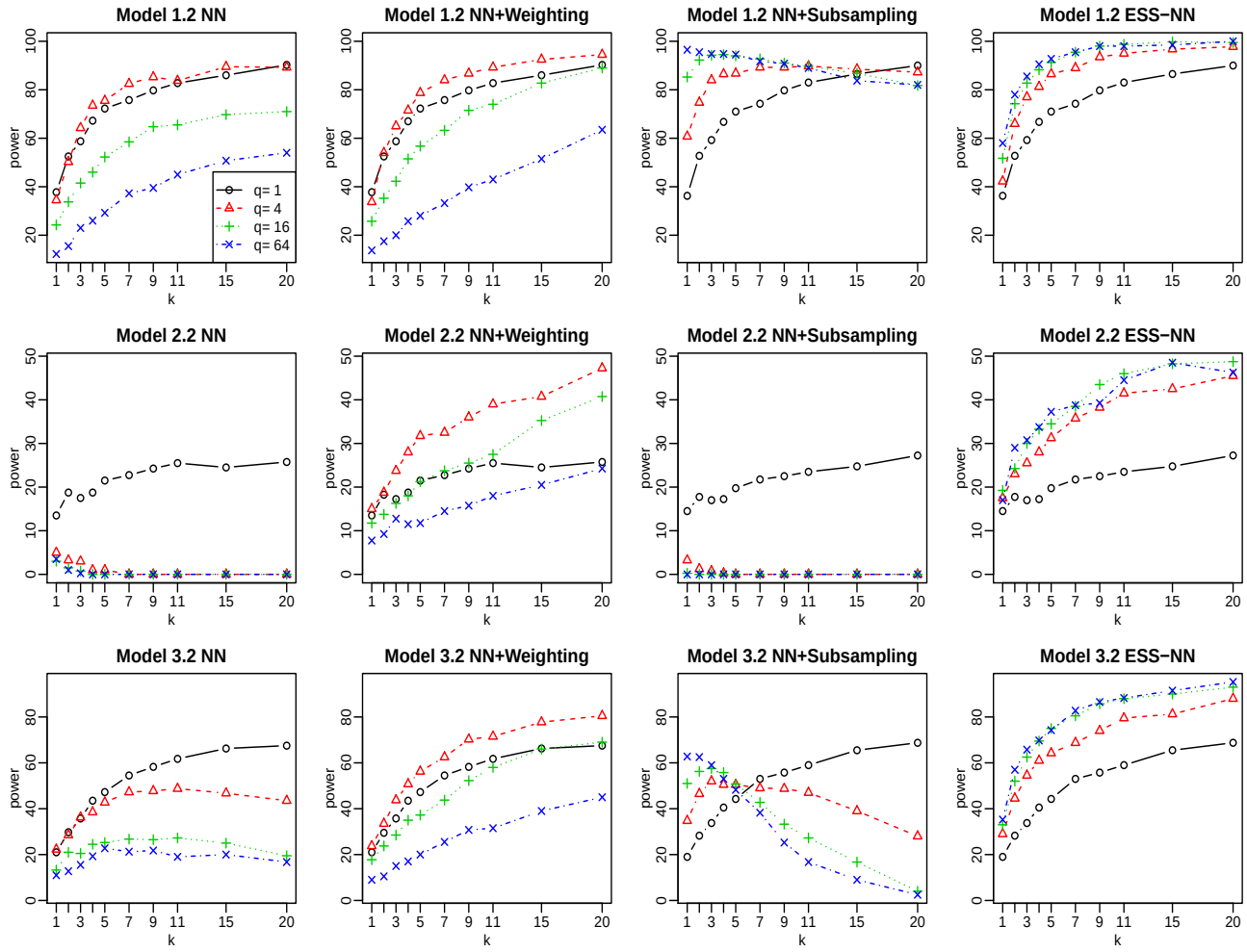

Figure 2: Simulation results comparing the power of original nearest neighbor method (NN), NN+Weighting, the unweighted statistic $T^u_{k,n}$ (NN+Subsampling) and the new weighted statistic $T_{k,n}$ (ESS-NN), for different ratios of the sample sizes $q = 1, 4, 16, 64$. The two samples are generated from the three simulation settings with $d = 5$ in Section 4. Power is approximated by the proportion of rejections over 400 runs of the testing procedures. A sequence of neighborhood sizes k are used.

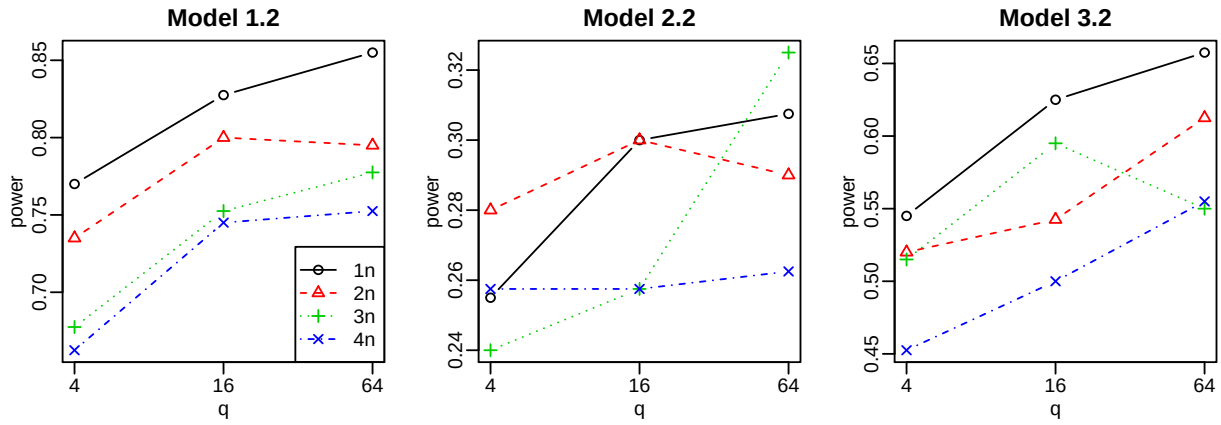

Figure 3: Simulation results comparing the power of the statistic T_{k,n_s} for different subsubsample sizes $n_s = n, 2n, 3n, 4n$, at the different ratios of the sample sizes $q = 4, 16, 64$. The two samples are generated from the three simulation settings with $d = 5$ in Section 4. Power is approximated by the proportion of rejections over 400 runs of the testing procedures.

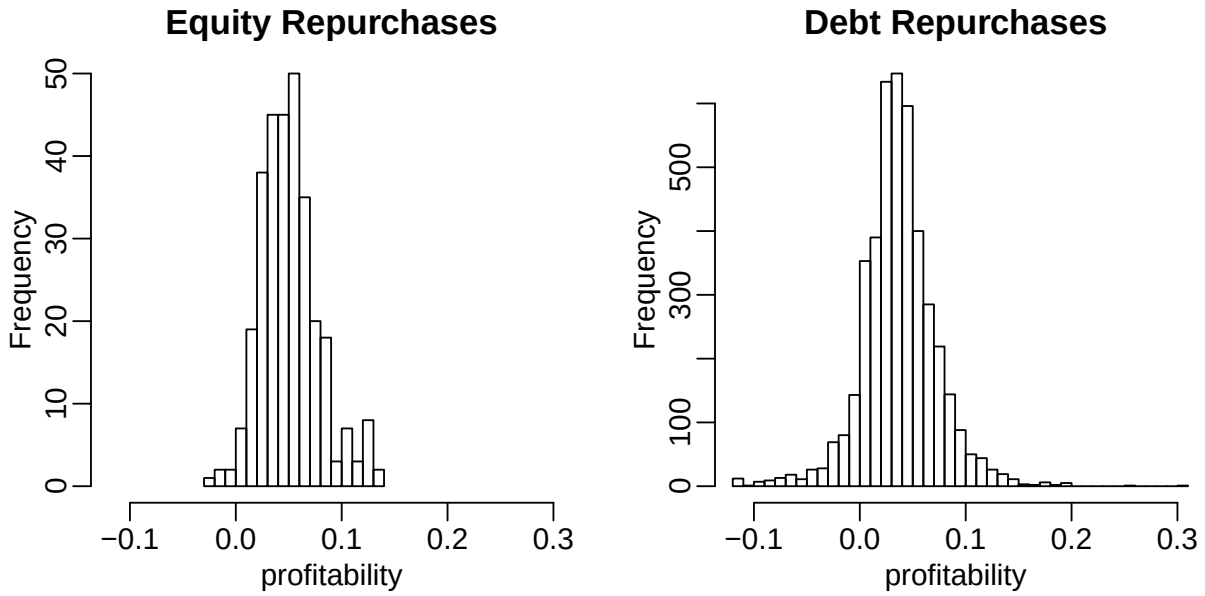

Figure 4: The histograms of profitability comparing the equity repurchases sample and the debt repurchases sample.

Table 1: Numerical evaluation of the asymptotic variance σ_k^2 (3), for different combinations of the dimension $d = 1, 5, \infty$, the neighborhood size $k = 1, 3, 5, 10, 30$, and the ratio of sample sizes $\lambda = 1, 4, 16, 64, \infty$.

		$\lambda = 1$	$\lambda = 4$	$\lambda = 16$	$\lambda = 64$	$\lambda = \infty$
	k=1	0.208	0.107	0.087	0.082	0.080
	k=3	0.218	0.108	0.087	0.082	0.081
$d = 1$	k=5	0.223	0.109	0.087	0.082	0.081
	k=10	0.228	0.109	0.088	0.082	0.081
	k=30	0.234	0.112	0.089	0.083	0.082
	k=1	0.195	0.104	0.085	0.080	0.079
	k=3	0.208	0.109	0.088	0.083	0.082
$d = 5$	k=5	0.215	0.111	0.090	0.085	0.083
	k=10	0.223	0.114	0.092	0.087	0.085
	k=30	0.230	0.118	0.095	0.089	0.087
	k=1	0.188	0.103	0.084	0.080	0.078
	k=3	0.203	0.109	0.088	0.084	0.082
$d = \infty$	k=5	0.211	0.112	0.091	0.086	0.084
	k=10	0.219	0.115	0.093	0.088	0.086
	k=30	0.228	0.118	0.095	0.090	0.088

Table 2: Numerical evaluation of $k\overline{p}_{1,k}$ at $\lambda = \infty$ ($\overline{p}_{1,k}$ is defined in Proposition 4), for different combinations of the dimension $d = 1, 2, 3, 5, 10, \infty$ and the neighborhood size $k = 1, 2, 3, 5, 10, 30, \infty$.

	k=1	k=2	k=3	k=5	k=10	k=30	$k = \infty$
d=1	0.286	0.292	0.291	0.293	0.307	0.365	
d=2	0.277	0.299	0.309	0.324	0.356	0.419	
d=3	0.271	0.303	0.319	0.341	0.379	0.435	
d=5	0.264	0.307	0.330	0.359	0.398	0.444	
d=10	0.255	0.311	0.339	0.372	0.409	0.448	
$d = \infty$	0.250	0.312	0.344	0.377	0.412	0.449	0.5

Table 3: Simulation results comparing the power of cross-match, runs, nearest neighbor method (NN), simple subsampling based on NN (SSS-NN) and ensemble subsampling based on NN (ESS-NN), for the sample size ratio $q = 1, 4, 16, 64$. The simulation settings are detailed in Section 4. Power is approximated by the proportion of rejections over 400 runs of each testing procedure on independently generated data. In the parentheses are empirical type I errors, i.e. the proportions of rejections under the null.

			cross-match	runs	NN	SSS-NN	ESS-NN
Model 1	dim=1	q=1	0.10	0.13	0.12	0.10	0.11 (0.05)
		q=4	0.08	0.11	0.11	0.13	0.12 (0.08)
		q=16	0.07	0.12	0.08	0.11	0.12 (0.04)
		q=64	0.62 (0.58)	0.06	0.05	0.13	0.17 (0.05)
	dim=5	q=1	0.36	0.58	0.59	0.60	0.59 (0.06)
		q=4	0.37	0.57	0.64	0.54	0.77 (0.05)
		q=16	0.26	0.36	0.41	0.53	0.83 (0.04)
		q=64	0.25 (0.13)	0.25	0.23	0.59	0.85 (0.05)
Model 2	dim=1	q=1	0.12	0.15	0.13	0.14	0.15 (0.05)
		q=4	0.13	0.13	0.13	0.14	0.20 (0.08)
		q=16	0.06	0.10	0.09	0.14	0.17 (0.04)
		q=64	0.66 (0.58)	0.06	0.08	0.15	0.23 (0.05)
	dim=5	q=1	0.14	0.22	0.17	0.17	0.17 (0.06)
		q=4	0.15	0.00	0.03	0.15	0.26 (0.05)
		q=16	0.13	0.00	0.01	0.18	0.30 (0.04)
		q=64	0.17 (0.13)	0.00	0.00	0.18	0.31 (0.05)
Model 3	dim=1	q=1	0.18	0.18	0.16	0.17	0.16 (0.04)
		q=4	0.14	0.20	0.18	0.17	0.27 (0.06)
		q=16	0.07	0.12	0.09	0.19	0.30 (0.05)
		q=64	0.65 (0.58)	0.09	0.08	0.19	0.28 (0.05)
	dim=5	q=1	0.24	0.38	0.36	0.34	0.34 (0.07)
		q=4	0.33	0.24	0.36	0.37	0.54 (0.08)
		q=16	0.25	0.15	0.20	0.38	0.62 (0.05)
		q=64	0.26 (0.10)	0.11	0.15	0.38	0.66 (0.06)

Table 4: Simulation results comparing the test based on MMD and the new test ESS-NN, for the sample size ratio $q = 1, 4, 16, 64$. The simulation settings are detailed in Section 4. Power is approximated by the proportion of rejections over 400 runs of each testing procedure on independently generated data.

	MMD				ESS-NN ($k = 15$)			
	q=1	q=4	q=16	q=64	q=1	q=4	q=16	q=64
Model 1 (dim=5)	0.99	1.00	1.00	1.00	0.87	0.97	0.99	0.99
Model 2 (dim=5)	0.61	0.87	0.89	0.92	0.25	0.43	0.48	0.49
Model 3 (dim=5)	0.92	0.98	0.99	1.00	0.66	0.81	0.90	0.92
$N(0,1)$ vs $NM_1(0.9)$	0.60	0.89	0.92	0.92	0.79	0.93	0.96	0.98
$N(0,I_5)$ vs $NM_5(0.4)$	0.12	0.17	0.22	0.24	0.22	0.37	0.37	0.41
$NM(0.7)$ vs $NM_1(0.9)$	0.29	0.50	0.61	0.62	0.59	0.77	0.83	0.81

Table 5: P-values for comparing the joint distributions of the four variables between the firm quarters related to equity repurchases and those related to debt repurchases. The variables are lagged term spread, lagged credit spread, lagged real stock return, and firm profitability. Both the original nearest neighbor method (NN) and the ensemble subsampling based on nearest neighbor method (ESS-NN) are applied. The p-values are obtained using different neighborhood sizes $k = 1, 3, 5, 10, 30$.

	k=1	k=3	k=5	k=10	k=30
NN	0.449	0.367	0.432	0.056	0.54
ESS-NN	0.004	0.006	0	0	0