

# Bootstrapped Edge Count Tests for Nonparametric Two-Sample Inference Under Heterogeneity

Trambak Banerjee

Analytics, Information and Operations Management, University of Kansas

Bhaswar B. Bhattacharya

Department of Statistics and Data Science, University of Pennsylvania  
and

Gourab Mukherjee

Department of Data Sciences and Operations, University of Southern California

April 8, 2023

## Abstract

Nonparametric two-sample testing is a classical problem in inferential statistics. While modern two-sample tests, such as the edge count test and its variants, can handle multivariate and non-Euclidean data, contemporary gargantuan datasets often exhibit heterogeneity due to the presence of latent subpopulations. Direct application of these tests, without regulating for such heterogeneity, may lead to incorrect statistical decisions. We develop a new nonparametric testing procedure that accurately detects differences between the two samples in the presence of unknown heterogeneity in the data generation process. Our framework handles this latent heterogeneity through a composite null that entertains the possibility that the two samples arise from a mixture distribution with identical component distributions but with possibly different mixing weights. In this regime, we study the asymptotic behavior of weighted edge count test statistic and show that it can be effectively re-calibrated to detect arbitrary deviations from the composite null. For practical implementation we propose a Bootstrapped Weighted Edge Count test which involves a bootstrap-based calibration procedure that can be easily implemented across a wide range of heterogeneous regimes. A comprehensive simulation study and an application to detecting aberrant user behaviors in online games demonstrates the excellent non-asymptotic performance of the proposed test.

*Keywords:* composite hypothesis testing; consumer behavior analysis; heterogeneity; two-sample tests.

# 1 Introduction

Nonparametric two-sample testing is a classical problem in inferential statistics. The use of these tests is pervasive across disciplines, such as medicine (Farris and Schopflocher, 1999), consumer research (Folkes et al., 1987), remote sensing (Conradsen et al., 2003) and public policy (Rothman et al., 2006), where detecting distributional differences between the two samples is germane to the ongoing scientific analysis. Nonparametric two-sample tests like the Kolmogorov-Smirnov test, the Wilcoxon rank-sum test, and the Wald-Wolfowitz runs test are extremely popular tools for analyzing univariate data. Multivariate versions of these tests have their origins in the randomization tests of Chung and Fraser (1958) and in the generalized Kolmogorov-Smirnov test of Bickel (1969). Friedman and Rafsky (1979) proposed the first computationally efficient nonparametric two-sample test, the edge count test, for high-dimensional data. Modern versions of the edge count test, such as the weighted and generalized edge count test (Chen and Friedman, 2017; Chen et al., 2018), can handle multivariate data, and can be applied to any data types as long as an informative similarity measure between the data points can be defined. Besides the edge count tests, several tests based on nearest-neighbor distances (Henze, 1984; Schilling, 1986; Chen et al., 2013; Hall and Tajvidi, 2002; Banerjee et al., 2020) and matchings (Rosenbaum, 2005; Mukherjee et al., 2022) have been proposed over the years, and used in variety of applications, such as covariate balancing (Heller et al., 2010a,b), change point detection (Chen and Zhang, 2015; Shi et al., 2017), gene-set analysis (Rahmatallah et al., 2012), microbiome data (Callahan et al., 2016; Holmes and Huber, 2018; Fukuyama, 2020), among others.

Bhattacharya (2019) propose a general framework to study the asymptotic properties of these graph based tests. In addition to the graph based tests, other popular two-sample tests include the energy distance test of Székely (2003) and Székely and Rizzo (2004) (see also Baringhaus and Franz (2004); Aslan and Zech (2005); Székely and Rizzo (2013)), and kernel tests based on the maximum mean discrepancy (see Gretton et al. (2007); Chwialkowski et al. (2015); Ramdas et al. (2015, 2017) and the references therein). Recently, several authors have proposed distribution-free two-sample tests using optimal

transport based multivariate ranks (see [Deb and Sen \(2021\)](#); [Ghosal and Sen \(2019\)](#); [Shi et al. \(2020a\)](#), and [Shi et al. \(2020b\)](#) and the references therein).

While there exists a vast literature on nonparametric two-sample tests, their performance in scenarios where the two samples might contain different heterogeneous structures have not been well-explored before. However, in a host of contemporary data analysis problems (See Ch. 3 of [Holmes and Huber \(2018\)](#), [Rossi et al. \(2012\)](#)) there is a need to conduct inference in presence of unknown heterogeneity in the data generation process. This is particularly important when there are latent subpopulations in the two populations from which the samples were extracted. Detecting distributional differences across the two samples is challenging in the presence of such heterogeneity because the two samples may differ with respect to the rates with which they arise from the underlying subpopulations. In these settings, direct application of existing two-sample tests, without regulating for the latent heterogeneity in the samples, may lead to incorrect statistical decisions and scientific consequences. In this article, we develop a new nonparametric testing procedure that can accurately detect if there are differences between the two samples in the presence of latent heterogeneity in the data generation process. We next present two contemporary data examples to motivate the two-sample testing problem under heterogeneity and discuss how existing tests may lead to incorrect decisions in the presence of heterogeneity.

## 1.1 Motivating examples for testing under heterogeneity

**Example 1: Detecting shift in consumer sentiment and spending pattern** – Detecting shifts in consumer sentiment and their spending pattern in response to exogenous economic shocks, such as a pandemic, war or supply chain constraints, is important from the perspective of public policy and businesses operations across different sectors ([Agarwal et al., 2019](#); [Crouzet et al., 2019](#); [Bruun, 2021](#); [Bartik et al., 2020](#)). Evidence of such changes in spending pattern is used to make policy decisions on the allocation of future resources to tackle the shifting landscape of consumer demand ([Balis, 2021](#); [Liguori and Pittz, 2020](#); [Akpan et al., 2022](#)). Data privacy rules, however, often forbid using personally identifiable information, such as individual consumer demographics and spending patterns over time,

thus ruling out the possibility of acquiring a rich consumer level panel data and utilizing sophisticated tools from causal inference to detect changes in spending patterns in response to the event of interest. A two-sample statistical test of hypothesis is often conducted to determine whether the spending pattern of a sample of consumers before the event of interest is significantly different from the spending pattern of an independent sample of consumers during or after the event of interest. Despite being much less powerful than tests based on comprehensive longitudinal data, the two-sample tests have the advantage of being substantially more privacy-preserving as these non-intrusive tests only need two independent vectors of observations and thus can be implemented across a wide range of contemporary applications. However, there are two main challenges for devising a test of hypothesis that can correctly detect such differences in the spending pattern between these two independent samples:

- (1.) The two samples may exhibit sample size imbalances which presents a challenging setting for nonparametric two-sample testing on multivariate data ([Chen et al., 2018](#)).
- (2.) The consumer base may consist of several heterogeneous subpopulations with respect to their consumption behavior. In business and economic modeling it is now commonplace to model such consumption behavior as a mixture of several distinct consumption patterns ([Fahey et al., 2007](#); [Labeeuw and Deconinck, 2013](#)).

Due to an exogenous shock we can have one of the following three situations regarding the behavior of consumers after the shock:

- I. Consumer behavior does not change and maintains the pre-shock levels.
- II. Consumption behavior changes but there are no new consumption patterns. The change in the consumption distribution is due to changes in the proportions of existing modes of consumption. This is the case where there is switching between the different consumption modes but no new consumption patterns evolve due to the external shock.
- III. Consumption behavior changes and there are new consumption patterns that were previously non-existent.

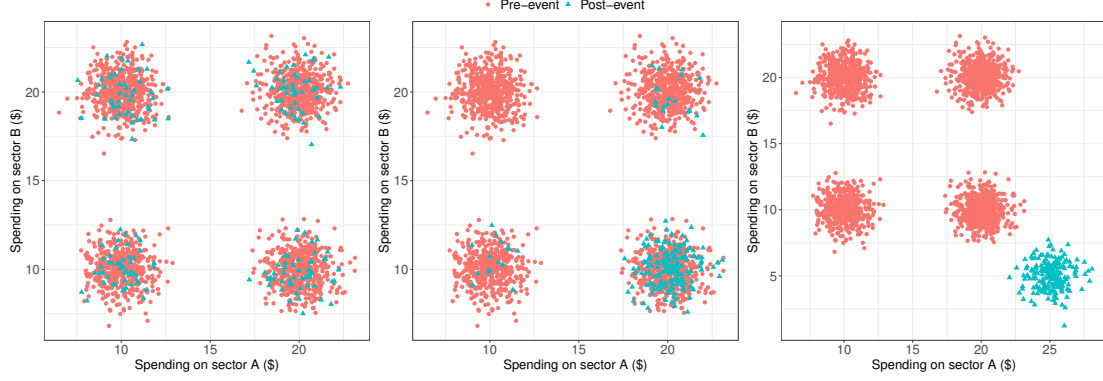


Figure 1: Testing shift in consumer spending pattern due to external economic shock. Green triangles represent a sample of consumers **post-event**, while the red dots represent an independent sample of consumers **pre-event**. The **pre-event** sample exhibits four distinct subpopulations. **Left**– consumption pattern of **post-event** consumers originate from these four subpopulations with equal probabilities. **Center**– majority of the **post-event** consumers arise from one of the four subpopulations without a shift in the levels of their consumption for that subpopulation. **Right**– **post-event** consumers originate from a new subpopulation that exhibits a change in the levels of consumption after the shock. The simulation scheme for these plots is described in Section B of the supplement.

In this paper, we concentrate on the detection of the third case, Case III, where new consumption patterns emerge post-shock. Accurate and timely identification of Case III is important for it allows researchers to understand if deviant forms of consumption (Koskeniemi, 2021) arise after a critical event, which can then be subsequently studied and analyzed with higher granularity data. Statistically, the main challenge here lies in correctly detecting Case III while not misidentifying Cases I and II as type I error. Two-sample testing procedures in the existing literature (discussed at the beginning of Section 1) are designed to distinguish between Cases I and II. We will show that without modifications their direct usage do not produce correct inference for this exercise.

We further elucidate this problem with the help of a simple simulation example. Consider a bivariate consumption problem with consumption on sector A denoted by  $X_1$  and on sector B by  $X_2$ . We have bivariate observations  $(X_1, X_2)$  for two random samples of customers before and after the shock (event). In Figure 1, the red dots represent the sam-

ple of consumers **pre-event**, while the green triangles represent an independent sample of consumers **post-event**. The three plots in Figure 1 present the distribution of spending pattern with respect to  $(X_1, X_2)$  and reveals that the **pre-event** sample exhibits four distinct subpopulations. The leftmost panel represents the setting where the consumption pattern of **post-event** consumers originate from these four subpopulations with equal probabilities. The center panel presents the setting of Case II where a majority of the **post-event** consumers arise from one of the four subpopulations without a shift in the levels of their consumption for that subpopulation. This represents normal consumption with only the proportional representation of the latent states of normal consumption being changed. The rightmost panel, on the other hand, reveals that the **post-event** consumers originate from a new subpopulation that exhibits a change in the levels of consumption after the event (Case III).

Distinction between the two settings presented in the center (Case II) and right (Case III) panels of Figure 1 is critical for policy makers and marketers. In this paper, we develop a consistent two-sample testing framework that can detect Case III from Cases I and II in the presence of sample size imbalance as well as heterogeneity. We next discuss another motivating example regarding consumption of digital entertainment, which will be further pursued in Section 5.

**Example 2: Detecting differences in player behavior in online games** – Online gaming is an important component of modern recreational and socialization media (Banerjee et al., 2022). For monetization of these digital products, managers often have to deliver personalized promotions to users based on their product usage. Additionally, through promotional intervention the portal needs to regulate addiction, violence and other deviant consumption patterns (Hull et al., 2014) than can cause long term societal harms. As it is difficult to manually track every game, the portals use automated decision rules to monitor the game sessions and rely on features extracted at regular intervals of time, such as hourly or half-hourly, to constantly check for deviant consumption within each game session. When evidence of preponderance of deviant consumption is available, managerial interventions are made.

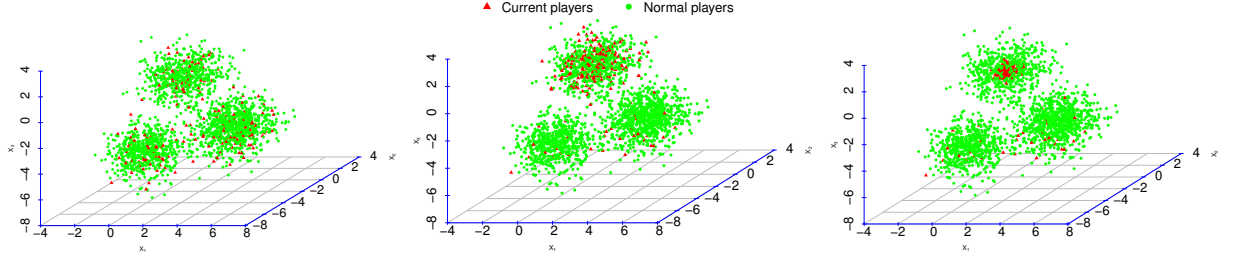


Figure 2: Differences in playing behavior in online video games. Red triangles represent the sample of current players, while the green dots are the sample of normal players from historical logs. The green sample exhibits three distinct subpopulations. **Left**– Case I: Current sample originates from three subpopulations present in the green sample with equal probabilities. **Center**– Case II: most members in the current sample originate from only one of the three subpopulations. **Right**– Case III: current sample originates from one of the three subpopulations but exhibits a different scale. The simulation scheme for these plots is described in Section B of the supplement.

Statistically, the problem here again reduces to the correct detection of scenarios pertaining to Case III described in Figure 1 above. To see this, note that based on historical data a multivariate sample of gaming features from players corresponding to normal gaming characteristics is available. The size of such a sample is huge as it is based on historical logs. The goal is to compare the instantaneous gaming features of a subset (gated by region, age, etc) of currently logged-on players with respect to this sample. The sample size of this subset of players is much smaller than the benchmark normal gaming consumption sample and so, we encounter the issue of sample size imbalance. Additionally, while comparing these two samples we are interested in detecting if there is a sub-population of players with deviant gaming characteristics which would need regulating promotional incentives related to purchase of gaming artifacts. This again corresponds to different Case III scenarios described above. Even without any instance of deviant usage, instantaneous gaming characteristics greatly change over time-of-day, for instance early morning players and evening players have varying characteristics. This corresponds to scenarios in Case II. Thus, the goal will be to detect Case III scenarios while allowing the possibility for scenarios from Cases I and II. We illustrate this through a simulation example in Figure 2.

Consider observing  $d = 3$  dimensional characteristics,  $X_1, X_2$  and  $X_3$  of playing behavior. The green dots in Figure 2 represent the sample of normal players from historical records. It exhibits three distinct subpopulations of equal sizes. The red triangles represent the sample of players from a current session. The leftmost panel depicts a setting where the current sample originates from these three subpopulations with equal probabilities (an instance of Case I). The center panel presents the setting where a majority of the players in the current sample are from one of the three subpopulations (an instance of Case II), and the rightmost panel reveals an instance of Case III where there is a new sub-population represented by red triangles that was non-existent in normal players. Note that, unlike Figure 1 here the deviant sub-population differs from one of the three subpopulations of normal players only in scale and not in locations. Detecting this case is more difficult than Figure 1 and we present a detailed analysis in Section 4. In Section 5 we develop this example and apply our proposed hypothesis testing method to detect addictive behaviors in a real-world gaming dataset.

## 1.2 Two-sample testing under heterogeneity and our contributions

Existing nonparametric two-sample tests cannot distinguish between the scenarios presented in the center and right panels of Figures 1 and 2. For instance, the edge count test (Friedman and Rafsky, 1979), the weighted edge count test (Chen et al., 2018) and the generalized edge count test (Chen and Friedman, 2017) reject the null hypothesis of equality of the two distributions for both Cases II and III in Figures 1 and 2 (see tables 7 and 8 in Section B of the supplement for more details). This is not surprising because these nonparametric two-sample tests are designed to test the simple null hypothesis of equality of the two distributions (Case II vs Case I) and directly using them, without any modification, to detect Case III scenarios from Case II can be misleading. To resolve this conundrum, we develop a hypothesis testing framework based on an appropriately constructed composite null hypothesis (Equations (5)–(7)). Under this composite null hypothesis, we study the properties of the Weighted Edge Count (WEC) test statistic of Chen



et al. (2018) and show that it can be re-calibrated to produce an asymptotically consistent two-sample test. For finite sample applications, we propose a bootstrap-based calibration procedure for the WEC test statistic that allows it to consistently and efficiently detect differences between the two samples under subpopulation level heterogeneity. The ensuing discussion summarizes our key contributions.

- We study the asymptotic properties of the WEC statistic for the heterogeneous two-sample problem (Section 2). Specifically, we show how one can choose a cut-off that renders the WEC statistic asymptotically powerful for detecting any distribution that significantly deviates from the composite null hypothesis (see Equation (7)), which encompasses the possibility that the two samples arise from the same mixture distribution but with possibly different mixing weights (see Proposition 1). This is in contrast to the TRUH test in Banerjee et al. (2020) which can only detect deviations in location under the concerned composite null hypothesis. This is because, unlike the TRUH statistic which compares the nearest neighbor distances between the two-samples, the WEC statistic is based on the number of within sample edges in a similarity graph of the pooled sample. The combinatorial (count-type) nature of the WEC statistic renders it asymptotically distribution-free under the simple (homogeneous) null, that is, the distribution of the test statistic does not depend on the null distribution of the data. Moreover, the WEC statistic estimates the well-known Henze-Penrose divergence (see Equation (9) for the definition) and, consequently, can detect arbitrary differences between two distributions under the homogeneous null. We leverage these properties to obtain a cut-off for the WEC statistic that can detect deviations from the heterogeneous (composite) null, beyond location problems.
- For non-asymptotic usage, we develop a novel bootstrap-based calibration procedure for the weighted edge count test statistic (Algorithm 1 in Section 3) and use it to test the composite null hypothesis that the two samples arise from the same mixture distribution but with possibly different mixing weights (Equations (5)–(7)). Under sample size imbalance, our calibration procedure first explores the larger sample for heterogeneous subgroups. Then it generates an ensemble of mixing weights for the

different subgroups and subsequently constructs a collection of surrogate two-samples under the composite null hypothesis. The WEC test statistic is computed for each such surrogate two-samples and the ensemble of these WEC test statistics is then used to determine the level  $\alpha$  cutoff for testing the composite null hypothesis (Equation (7)).

- Our numerical experiments (Section 4) reveal that across a wide range of simulation settings, the proposed bootstrap calibration procedure results in a conservative level  $\alpha$  cutoff for the WEC test statistic for consistent nonparametric two sample testing involving the composite null hypothesis of Equation (7). This is in contrast to existing nonparametric two-sample tests that may lead to incorrect inference in this regime. Furthermore, our empirical evidence suggests that the bootstrap calibration procedure renders the WEC test more powerful than the recently introduced TRUH test (Banerjee et al., 2020) for the heterogeneous two sample problem.
- We apply our proposed hypothesis testing procedure to detect addictive behaviors in online gaming (Section 5). On an anonymized data available from a large video game company in Asia, we test whether players who login after midnight and players who login early in the morning exhibit deviant playing behavior when compared to players with normal gaming behavior. Our analysis reveals that the playing behavior of players who login after midnight is statistically different from the normal gaming behavior, while those who login early, crucially enough, do not exhibit such differences in their playing behavior. These results confirm the findings in extant research that indicate higher tendency towards game addiction for those players who login late (Lee and Kim, 2017). Existing nonparametric two-sample tests, on the other hand, erroneously conclude aberrant gaming behavior for both sets of players, those who login late and those who login early.

## 2 Nonparametric two-sample testing under heterogeneity

We begin with a review of the edge count tests in Section 2.1. Thereafter, Section 2.2 introduces the heterogeneous two-sample hypothesis testing problem. In Section 2.3 we study the asymptotic properties of the WEC statistic for the heterogeneous two-sample problem.

We first collect some notations that will be used throughout this article. Denote the two independent samples by  $\mathcal{X}_n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$  and  $\mathcal{Y}_m = \{\mathbf{Y}_1, \dots, \mathbf{Y}_m\}$ . Suppose each  $\mathbf{X}_i \in \mathbb{R}^d$ ,  $i = 1, \dots, n$ , is distributed independently and identically according to a distribution that has cumulative distribution function (cdf)  $F_{\mathbf{X}}$ . Similarly, denote  $F_{\mathbf{Y}}$  to be the cdf of the distribution of  $\mathbf{Y}_j \in \mathbb{R}^d$  for  $j = 1, \dots, m$ . Denote  $N = n + m$  and let  $\mathcal{Z}_N = \{\mathcal{X}_n, \mathcal{Y}_m\}$  be the pooled sample.

### 2.1 Edge count tests – a review

In their seminal paper Friedman and Rafsky (1979) introduced the edge-count (EC) test for the classical two-sample hypothesis testing problem:

$$H_0 : F_{\mathbf{X}} = F_{\mathbf{Y}} \text{ versus } H_1 : F_{\mathbf{X}} \neq F_{\mathbf{Y}}. \quad (1)$$

The EC test can be described as follows:

- Construct a similarity graph  $\mathcal{G}$  (based on the pairwise distances between the observations) of the pooled sample  $\mathcal{Z}_N$ .
- Count the number of edges  $R_0$  in the graph  $\mathcal{G}$  with one end-point in sample 1 and other in sample 2, and reject  $H_0$  in (1) if  $R_0$  is ‘small’.

Friedman and Rafsky (1979) chose  $\mathcal{G}$  to be the  $\ell$ -minimum spanning tree (MST) of the pooled sample  $\mathcal{Z}_N$  using the  $L_2$  distance<sup>1</sup>. Thereafter, tests based on other similarity

---

<sup>1</sup>A *spanning tree* of a finite set  $S \subset \mathbb{R}^d$  is a connected graph with vertex-set  $S$  and no cycles. A *1-minimum spanning tree* (MST), or simply a MST, of  $S$  is a spanning tree which minimizes the sum of

graphs have been proposed. In particular, [Schilling \(1986\)](#) and [Henze \(1988\)](#) considered tests where  $\mathcal{G}$  is the nearest-neighbor graph and [Rosenbaum \(2005\)](#) proposed a test where  $\mathcal{G}$  is the minimum non-bipartite matching. The aforementioned tests are all asymptotically distribution-free (the asymptotic distribution of  $R_0$  under  $H_0$  in (1) does not depend on the distribution of the data), universally consistent (the test has asymptotic power 1 for all alternatives in (1)), and computationally efficient (running time is polynomial in both the number of data points and dimension), making them readily usable in applications.

Although the EC test is universally consistent (see ([Henze and Penrose, 1999](#), Theorem 2)), [Chen and Friedman \(2017\)](#) and [Chen et al. \(2018\)](#) observed that it has two major limitations: First, empirical evidence suggest that the EC test has low or no power for scale alternatives even though asymptotically the test is consistent for both location and scale alternatives ([Henze and Penrose, 1999](#)). Second, under sample size imbalance the EC test statistic has a relatively large variance and exhibits low power for detecting departures from the null hypothesis in Equation (1). To mitigate these issues, [Chen and Friedman \(2017\)](#) and [Chen et al. \(2018\)](#) suggested new tests based on the within sample edges in the similarly graph  $\mathcal{G}$ . Towards this suppose  $\mathcal{E}_{\mathcal{G}}$  denotes the edge set of  $\mathcal{G}$ . For an edge  $e = (i, j) \in \mathcal{E}_{\mathcal{G}}$ , let

$$J_e = \begin{cases} 2 & \text{if observations } i \text{ and } j \text{ are from sample } \mathcal{Y}_m, \\ 1 & \text{if observations } i \text{ and } j \text{ are from sample } \mathcal{X}_n, \\ 0 & \text{if observations } i \text{ and } j \text{ are from different samples.} \end{cases}$$

For  $k \in \{0, 1, 2\}$ , define

$$R_k = \sum_{e \in \mathcal{E}_{\mathcal{G}}} \mathbb{I}\{J_e = k\}. \quad (2)$$

Note that  $R_0$ , which counts the number of between-sample edges, is the EC statistic introduced before. Similarly,  $R_1$  is the number of edges with both endpoints in  $\mathcal{X}_n$  and  $R_2$  is the number of edges with both endpoints in  $\mathcal{Y}_m$ .

---

distances across the edges of the tree. A  $\ell$ -MST of  $S$ , for  $\ell \geq 2$ , is the union of the edges in the  $(\ell - 1)$ -MST together with the edges of the spanning tree that minimizes the sum of distances across edges subject to the constraint that this spanning tree does not contain any edge of the  $(\ell - 1)$ -MST.

To address the issue of low power for scale alternatives, [Chen and Friedman \(2017\)](#) proposed the Generalized Edge Count (GEC) test, which rejects the null hypothesis  $h_0$  in Equation (1) for large values of

$$\mathcal{R}_g(\mathcal{X}_n, \mathcal{Y}_m) = (R_1 - \mu_1, R_2 - \mu_2) \Sigma^{-1} (R_1, R_2)^T, \quad (3)$$

where  $\mu_k = \mathbb{E}(R_k)$ , for  $k = 1, 2$ , and  $\Sigma$  is the covariance matrix of  $(R_1, R_2)^T$  under the permutation null distribution (see Lemma 3.1 of [Chen and Friedman \(2017\)](#) for the analytical expression of these quantities). The Weighted Edge Count (WEC) test of [Chen et al. \(2018\)](#) addresses the issue of sample-size imbalance by proposing a new test statistic  $\mathcal{R}_w(\mathcal{X}_n, \mathcal{Y}_m)$  based on the weighted sum of the within sample edges,  $R_1$  and  $R_2$ , as follows:

$$\mathcal{R}_w(\mathcal{X}_n, \mathcal{Y}_m) = \frac{1}{N} \left( \frac{m}{N} R_1 + \frac{n}{N} R_2 \right). \quad (4)$$

The weighting scheme controls the variance of  $\mathcal{R}_w$  as opposed to the EC test statistic based on  $R_0$ . The test rejects  $H_0$  in (1) for large values of  $\mathcal{R}_w(\mathcal{X}_n, \mathcal{Y}_m)$ , where the cut-off is computed from the permutation or the asymptotic null distribution under  $H_0 : F_{\mathbf{X}} = F_{\mathbf{Y}}$ . Evidence from authors' empirical studies reveal that under sample size imbalance and for location alternatives, the WEC test is more powerful than both the EC test and the GEC test.

## 2.2 The heterogeneous two-sample problem

In this section we introduce the heterogeneous two-sample hypothesis testing problem using motivating Example 2 from Section 1.1. For ease of exposition, we will refer the population of players who exhibit normal playing behavior as the baseline population. So in our example,  $\mathcal{X}_n$  is an i.i.d random sample of size  $n$  from the baseline population that has cdf  $F_{\mathbf{X}}$  and  $\mathcal{Y}_m$  is an i.i.d random sample of size  $m$  from the population of current players that has cdf  $F_{\mathbf{Y}}$ . We consider a setting where the heterogeneity in the baseline population is represented by  $K$  different subgroups, each having unimodal distributions with distinct modes and cdfs  $F_1, \dots, F_K$ , such that

$$F_{\mathbf{X}} = \sum_{a=1}^K w_a F_a, \quad \text{where} \quad w_a \in (0, 1) \text{ and } \sum_{a=1}^K w_a = 1. \quad (5)$$

Here the number of components  $K$ , the mixing distributions  $F_1, \dots, F_K$ , and the mixing weights  $w_1, \dots, w_K$  are fixed (non-random) but unknown. Furthermore, since  $F_1, \dots, F_K$  are cdfs from unimodal distributions with distinct modes,  $F_{\mathbf{X}}$  is well-defined with a unique specification i.e,  $F_{\mathbf{X}} \neq \sum_{a=1}^K \tilde{w}_a F_a$  if  $\tilde{w}_a \neq w_a$  for at least one  $a \in \{1, \dots, K\}$ . The population of normal players may exhibit two distinct phenomenon. First, the normal players can have similar playing behavior as the baseline population but a different representation of the  $K$  subpopulations than those reflected by the mixing proportions  $\{w_1, \dots, w_K\}$ . This may imply that a few baseline subpopulations are completely absent in the population of normal players. Thus, if the normal players do not exhibit a different playing behavior then their cdf  $F_{\mathbf{Y}}$  lies in a class of distributions  $\mathcal{F}(F_{\mathbf{X}})$  that contains any convex combination of  $\{F_1, \dots, F_K\}$  including the boundaries, that is,

$$\mathcal{F}(F_{\mathbf{X}}) = \left\{ Q = \sum_{a=1}^K \lambda_a F_a : \lambda_1, \lambda_2, \dots, \lambda_K \in [0, 1] \text{ and } \sum_{a=1}^K \lambda_a = 1 \right\}. \quad (6)$$

Note that the left and center panels of Figure 2 are examples of the setting where  $F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$ . Furthermore,  $F_{\mathbf{X}} \in \mathcal{F}(F_{\mathbf{X}})$  in Equation (6). Second, if the normal players exhibit a different playing behavior than the baseline population, then  $F_{\mathbf{Y}}$  would contain at least one non-trivial subpopulation with distribution substantially different from  $\{F_1, F_2, \dots, F_K\}$  or their linear combinations. Then,  $F_{\mathbf{Y}} \notin \mathcal{F}(F_{\mathbf{X}})$  and the right panel of Figure 2 represents this phenomenon. The heterogeneous two-sample problem that we consider in this article involves testing the following composite null hypothesis:

$$H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}}) \quad \text{versus} \quad H_1 : F_{\mathbf{Y}} \notin \mathcal{F}(F_{\mathbf{X}}). \quad (7)$$

Existing nonparametric graph-based two-sample tests, such as the EC test, are designed to test the simple null hypothesis  $H_0 : F_{\mathbf{X}} = F_{\mathbf{Y}}$ , and are not conservative for testing the composite null hypothesis  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  of Equation (7) (see, for example, Proposition 1 in (Banerjee et al., 2020)).

## 2.3 Asymptotic properties of WEC statistic under heterogeneity

In this section we will discuss how one can calibrate the WEC statistic under heterogeneity to obtain a test that is asymptotically powerful for general practical alternatives. For

this we assume that the baseline cdfs  $F_1, F_2, \dots, F_K$  have unimodal densities  $f_1, f_2, \dots, f_K$  (with respect to Lebesgue measure). Therefore, the baseline population will have density  $f_{\mathbf{X}} = \sum_{a=1}^K w_a f_a$  (recall Equation (5)), and the set of distributions in Equation (6) can be represented in terms of the densities  $f_1, f_2, \dots, f_K$  as:

$$\left\{ g = \sum_{a=1}^K \lambda_a f_a : \lambda_1, \lambda_2, \dots, \lambda_K \in [0, 1] \text{ and } \sum_{a=1}^K \lambda_a = 1 \right\},$$

and will be denoted by  $\mathcal{F}(f_{\mathbf{X}})$ . Then, assuming that the population of current players has density  $f_{\mathbf{Y}}$ , the hypothesis testing problem in Equation (7) can be restated as:

$$H_0 : f_{\mathbf{Y}} \in \mathcal{F}(f_{\mathbf{X}}) \quad \text{versus} \quad H_1 : f_{\mathbf{Y}} \notin \mathcal{F}(f_{\mathbf{X}}).$$

To prove our asymptotic results we assume the following:

**Assumption 1.** *The densities  $f_{\mathbf{X}}$  and  $f_{\mathbf{Y}}$  have a common support  $S \subseteq \mathbb{R}^d$  and there exist constants  $M_1, M_2 > 0$  such that  $M_1 \leq f_{\mathbf{X}}(x) \leq M_2$  and  $f_{\mathbf{Y}}(x) \leq M_2$ , for all  $x \in S$ .*

**Assumption 2.** *The weights of the baseline population are bounded below, that is, there exists a known constant  $L > 0$  such that  $w_a > L$ , for  $1 \leq a \leq K$ .*

Now, recall the definition of the WEC statistic  $\mathcal{R}_w(\mathcal{X}_n, \mathcal{Y}_m)$  from Equation (4). Then, we have the following result in the usual asymptotic regime where  $m, n \rightarrow \infty$  such that  $n/m \rightarrow \rho \in (0, \infty)$ .

**Proposition 1.** *Suppose the similarity graph  $\mathcal{G}$  is the  $\ell$ -MST, for some finite  $\ell \geq 1$ . Then under Assumptions 1 and 2 the following hold:*

(1) *For any  $f_{\mathbf{Y}} \in \mathcal{F}(f_{\mathbf{X}})$ ,*

$$\lim_{m, n \rightarrow \infty} \mathbb{P}_{f_{\mathbf{X}}, f_{\mathbf{Y}}} \left\{ \mathcal{R}_w(\mathcal{X}_n, \mathcal{Y}_m) > \gamma \right\} = 0,$$

$$\text{where } \gamma := \frac{\ell \rho}{(1+\rho)^2} \left\{ 1 + \frac{(1+\rho)K^2}{L\rho^2} \right\}.$$

(2) *Suppose  $\varepsilon := \left\{ \frac{2M_2^3(1+\rho)^3 K^2}{M_1^2 \rho^3 L} \right\}^{\frac{1}{2}}$ . Then*

$$\lim_{m, n \rightarrow \infty} \mathbb{P}_{f_{\mathbf{X}}, f_{\mathbf{Y}}} \left\{ \mathcal{R}_w(\mathcal{X}_n, \mathcal{Y}_m) > \gamma \right\} = 1,$$

*whenever  $f_{\mathbf{Y}} \notin \mathcal{F}(f_{\mathbf{X}})$  is such that  $\int_S \{f_{\mathbf{Y}}(x) - f_{\mathbf{X}}(x)\}^2 dx > \varepsilon^2$ .*

This result shows that if one chooses the cut-off of the WEC statistic based on  $\gamma$ , then the probability of Type I error is asymptotic zero for all  $f_{\mathbf{Y}} \in \mathcal{F}(f_{\mathbf{X}})$ . Moreover, the power of the test is asymptotically 1, whenever  $f_{\mathbf{Y}}$  is  $\varepsilon$ -far (in the  $L_2$ -distance) from the baseline density  $f_{\mathbf{X}}$ . In other words, the WEC statistic calibrated as above is consistent whenever the signal strength (measured in terms of the  $L_2$ -distance) is not too small. The proof of Proposition 1 relies on the following asymptotic result (see (Chen et al., 2018, Theorem 4) and (Henze and Penrose, 1999, Theorem 2)): Suppose  $\mathcal{X}_n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$  and  $\mathcal{Y}_m = \{\mathbf{Y}_1, \dots, \mathbf{Y}_m\}$  are i.i.d. samples from  $d$ -dimensional distributions with absolutely continuous densities  $f_{\mathbf{X}}$  and  $f_{\mathbf{Y}}$ , respectively. Then as  $m, n \rightarrow \infty$  such that  $n/m \rightarrow \rho \in (0, \infty)$ ,

$$\mathcal{R}_w(\mathcal{X}_n, \mathcal{Y}_m) \xrightarrow{P} \frac{\ell\rho}{(1+\rho)^2} \cdot \delta_\rho(f_{\mathbf{X}}, f_{\mathbf{Y}}), \quad (8)$$

where

$$\delta_\rho(f_{\mathbf{X}}, f_{\mathbf{Y}}) = \int_S \frac{\rho f_{\mathbf{X}}^2(x) + f_{\mathbf{Y}}^2(x)}{(\rho f_{\mathbf{X}}(x) + f_{\mathbf{Y}}(x))} dx. \quad (9)$$

The quantity  $\delta_\rho(f_{\mathbf{X}}, f_{\mathbf{Y}})$  is known as the *Henze-Penrose divergence* between the densities  $f_{\mathbf{X}}$  and  $f_{\mathbf{Y}}$  and plays a central role in the consistency of edge-count type tests. The proof of Proposition 1 entails showing that  $\delta_\rho(f_{\mathbf{X}}, f_{\mathbf{Y}})$  is uniformly bounded above under the composite null  $H_0 : f_{\mathbf{Y}} \in \mathcal{F}(f_{\mathbf{X}})$  and is ‘large’ under the alternative (details are given in Section A of the Supplement).

**Remark 1.** *Note that Assumption 1 requires that the baseline density  $f_{\mathbf{X}}$  is bounded below in its support. This is a well-known assumption that is often required in deriving asymptotic properties of methods based on geometric graphs and nearest-neighbors (see, for example, Biau and Devroye (2015); Penrose and Yukich (2003); Zhao and Lai (2022) and the references therein). Although this assumption technically rules out some natural distributions, from a practical standpoint, there is no real concern because one can approximate the univariate density by truncating it to a large interval, on which the above result applies. Moreover, it is possible to remove this assumption by replacing the  $L_2$  separation condition in Proposition 1 with a truncated  $L_2$  separation condition (see Remark 2 in the supplement for details).*



### 3 Bootstrap based calibration of the WEC test statistic

Here we develop a bootstrap based re-calibration procedure for practical implementation of the WEC test under heterogeneity. Under sample size imbalance, our proposed calibration procedure for the WEC test statistic first explores the larger sample  $\mathcal{X}_n$  for heterogeneous subgroups. To determine the number of such subgroups  $K$  in  $\mathcal{X}_n$ , we use the prediction strength approach of [Tibshirani and Walther \(2005\)](#), which gives an estimate  $\hat{K}$  of  $K$ . The class membership of the baseline samples  $\mathcal{X}_n$  is then determined using a  $\hat{K}$ -means algorithm. Denote by  $\hat{J}_a \subseteq \{1, 2, \dots, n\}$  to be the subset of indices estimated to be in class  $a$  by the  $\hat{K}$ -means algorithm where  $1 \leq a \leq \hat{K}$ . Let  $n_a = |\hat{J}_a|$  to be the cardinality of class  $a$  and denote  $\mathcal{X}_{\hat{J}_a} = \{\mathbf{X}_i : i \in \hat{J}_a\}$  to be the corresponding subset of the baseline samples estimated to be in class  $a$ . Note that  $\mathcal{X}_n = \{\mathcal{X}_{\hat{J}_a} : a = 1, 2, \dots, \hat{K}\}$  and  $\sum_{a=1}^{\hat{K}} n_a = n$ . Our calibration procedure then repeats the following three steps a large number of times:

1. Mixing proportions are randomly sampled from the  $\hat{K}$ -dimensional simplex  $\mathcal{S}_{\hat{K}}$  where  $\mathcal{S}_{\hat{K}} = \{(\lambda_1, \dots, \lambda_{\hat{K}}) \in \mathbb{R}^{\hat{K}} : \lambda_a \in [0, 1], \text{ for } 1 \leq a \leq \hat{K}, \text{ and } \sum_{a=1}^{\hat{K}} \lambda_a = 1\}$ .
2. For a given realization of the mixing proportions from  $\mathcal{S}_{\hat{K}}$ , surrogate baseline and normal players samples are generated from  $\mathcal{F}(F_{\mathbf{X}})$ .
3. The WEC test statistic is computed using the two samples obtained from step (2).

We now discuss these three steps below.

For step (1), and for each  $b = 1, \dots, B$ , denote  $(\lambda_1^{(b)}, \dots, \lambda_{\hat{K}}^{(b)})$  to be a random sample from  $\mathcal{S}_{\hat{K}}$ . Given the mixing weights  $\{\lambda_1^{(b)}, \dots, \lambda_{\hat{K}}^{(b)}\}$ , step (2) involves constructing surrogate baseline and normal players samples from  $\mathcal{F}(F_{\mathbf{X}})$  as follows: for each  $a \in \{1, \dots, \hat{K}\}$ , randomly sample  $\lceil m\lambda_a^{(b)} \rceil$  elements without replacement from  $\mathcal{X}_{\hat{J}_a}$ . Then the chosen elements constitute the surrogate normal players sample in class  $a$  and are denoted by  $\mathbb{Y}_a^{(b)} = \{\mathbf{X}_1^{(b)}, \dots, \mathbf{X}_{\lceil m\lambda_a^{(b)} \rceil}^{(b)}\}$ . The remaining  $n_a - \lceil m\lambda_a^{(b)} \rceil$  elements in  $\mathcal{X}_{\hat{J}_a}$  form the residual baseline sample  $\mathbb{X}_a^{(b)}$  in class  $a$ . We combine these samples over the  $\hat{K}$  classes to get the surrogate normal players sample as  $\mathcal{Y}_m^{(b)} = \{\mathbb{Y}_a^{(b)} : a = 1, \dots, \hat{K}\}$  and the corresponding

baseline sample as  $\mathcal{X}_{\tilde{n}}^{(b)} = \{\mathbb{X}_a^{(b)} : a = 1, \dots, \hat{K}\}$ , where  $\tilde{n} = \sum_{a=1}^{\hat{K}} (n_a - \lceil m\lambda_a^{(b)} \rceil)$ . Finally, for step (3) the bootstrapped samples in the  $b^{th}$  round,  $\mathcal{X}_{\tilde{n}}^{(b)}$  and  $\mathcal{Y}_m^{(b)}$  (which are surrogates for  $\mathcal{X}_n$  and  $\mathcal{Y}_m$ , respectively), can be used to compute the WEC test statistic  $\mathcal{R}_w^{(b)} := \mathcal{R}_w(\mathcal{X}_{\tilde{n}}^{(b)}, \mathcal{Y}_m^{(b)})$ .

The bootstrap calibration procedure described above is summarized in Algorithm 1. The computational complexity of our calibration procedure depends on two key steps: (i)

---

**Algorithm 1:** Bootstrap cut-off for a level  $\alpha$  test using  $\mathcal{R}_w(\mathcal{X}_n, \mathcal{Y}_m)$

---

**Input:** The parameters  $n$  and  $\alpha$ . The baseline samples  $\mathcal{X}_n$ , and the estimates  $\hat{K}$  and  $\{\hat{J}_a : a = 1, \dots, \hat{K}\}$  from the  $K$ -means algorithm with  $n_a = |\hat{J}_a|$ .

**Output:** The bootstrapped level  $\alpha$  cutoff  $r_{n,m,\alpha}$ .

**for**  $b = 1, \dots, B$  **do**

    STEP 1: Random sample  $\{\lambda_1^{(b)}, \dots, \lambda_{\hat{K}}^{(b)}\}$  from the  $\hat{K}$ -dimensional simplex;

**for**  $a = 1, \dots, \hat{K}$  **do**

**if**  $\lceil m\lambda_a^{(b)} \rceil \leq n_a$  **then**

            STEP 2: Draw a simple random sample  $\mathbb{Y}_a^{(b)} = \{\mathbf{X}_1^{(b)}, \dots, \mathbf{X}_{\lceil m\lambda_a^{(b)} \rceil}^{(b)}\}$  without replacement from  $\mathcal{X}_{\hat{J}_a}$ ;

            STEP 3:  $\mathbb{X}_a^{(b)} = \mathcal{X}_{\hat{J}_a} \setminus \mathbb{Y}_a^{(b)}$  the baseline residual sample in class  $a$ ;

**else**

            Stop: Go to STEP 1;

    Surrogate normal players sample:  $\mathcal{Y}_m^{(b)} = \{\mathbb{Y}_a^{(b)} : a = 1, \dots, \hat{K}\}$ ;

    Baseline sample:  $\mathcal{X}_{\tilde{n}}^{(b)} = \{\mathbb{X}_a^{(b)} : a = 1, \dots, \hat{K}\}$ ;

    STEP 4: Calculate  $\mathcal{R}_w^{(b)} := \mathcal{R}_w(\mathcal{X}_{\tilde{n}}^{(b)}, \mathcal{Y}_m^{(b)})$ ;

STEP 5: Return  $r_{n,m,\alpha} = \min\{\mathcal{R}_w^{(b)} : \frac{1}{B} \sum_{r=1}^B \mathbf{1}\{\mathcal{R}_w^{(r)} \geq \mathcal{R}_w^{(b)}\} \leq \alpha\}$ .

---

computation of the estimated number of clusters  $\hat{K}$ , and (ii) computation of the WEC test statistic over  $B$  bootstrap samples. To estimate  $K$ , we use prediction strength along with a  $K$ -means algorithm where the target number of clusters and the maximum number of iterations over which the  $K$ -means algorithm runs before stopping are both fixed. Thus step (i) has  $O(nd)$  complexity. The calculations in step (ii) can be distributed across the  $B$  bootstrap samples but for each  $b \in \{1, \dots, B\}$  computation of the WEC test statistic

requires calculating the  $N \times N$  distance matrix and constructing the MST on the pooled sample  $\mathcal{Z}_N$ . The former has  $O(n^2d)$  complexity while the latter has  $O((n-1)\log n)$  complexity. Therefore, the overall computational complexity of Algorithm 1 is  $O(n^2d)$ .

### 3.1 Implementation

Here we discuss several aspects related to the implementation of the proposed calibration procedure presented in Algorithm 1. First, for the numerical experiments and real data analysis (Sections 4 and 5, respectively), we set  $B = 200$  and implement a version of Algorithm 1 where we sample the mixing proportions  $\{\lambda_1, \dots, \lambda_{\hat{K}}\}$  only from the corners of  $\mathcal{S}_{\hat{K}}$ , thus placing most weight on the corners of  $\mathcal{S}_{\hat{K}}$ . Second, while Algorithm 1 provides a calibration approach for the WEC test statistic, it can also be used to obtain the level  $\alpha$  cutoff for the GEC test statistic for testing the composite null hypothesis of Equation (7). To do that, we replace  $\mathcal{R}_w(\mathcal{X}_n^{(b)}, \mathcal{Y}_m^{(b)})$  in Algorithm 1 with  $\mathcal{R}_g(\mathcal{X}_n^{(b)}, \mathcal{Y}_m^{(b)})$  where GEC test statistic is defined as  $\mathcal{R}_g(\mathcal{X}_n, \mathcal{Y}_m) = (R_1 - \mu_1, R_2 - \mu_2)\Sigma^{-1}(R_1, R_2)^T$ , with  $\mu_k = \mathbb{E}(R_k)$ ,  $k = 1, 2$  and  $\Sigma$  being the covariance matrix of  $(R_1, R_2)^T$  under the permutation null distribution (See Lemma 3.1 of [Chen and Friedman \(2017\)](#) for the analytical expression of these quantities). In Sections 4 and 5 we report the performance of both WEC and GEC test statistics using the bootstrap calibration procedure presented in Algorithm 1. Third, we rely on the R-package `gTests` to calculate the WEC and GEC statistics using 5–MST on the pooled sample, which is a recommended practical choice ([Chen and Friedman, 2017](#)).

## 4 Numerical experiments

In this section we assess the numerical performance of the bootstrap calibrated WEC (BWEC) and GEC (BGEC) tests against the following four competing testing procedures across a wide range of simulation settings: (i) Edgecount (EC) test, (ii) Generalized edgecount (GEC) test, (iii) Weighted edgecount (WEC) test, and (iv) TRUH test of [Banerjee et al. \(2020\)](#). For implementing the three edge count tests, we rely on the R package `gtests` and while for TRUH, we use the code available at [Banerjee et al. \(2020\)](#) with the default

specification of  $\tau_{fc} = 1$ . Note that amongst the aforementioned four competing tests, only the TRUH statistic is designed to test the composite null hypothesis of Equation (7), while the three variants of the edge count test were developed to test the simple null hypothesis  $H_0 : F_{\mathbf{X}} = F_{\mathbf{Y}}$  against the alternative  $H_1 : F_{\mathbf{X}} \neq F_{\mathbf{Y}}$ .

In our numerical experiments, we simulate  $\mathcal{X}_n$  and  $\mathcal{Y}_m$  from  $F_{\mathbf{X}}$  and  $F_{\mathbf{Y}}$ , respectively, and for each testing procedure, we report the proportion of rejections across 500 repetitions of the composite null hypothesis test  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  vs  $H_1 : F_{\mathbf{Y}} \notin \mathcal{F}(F_{\mathbf{X}})$  at 5% level of significance. The R code that reproduces our simulation results is available at <https://www.dropbox.com/sh/fstb0yzv7yjnfgc/AAA9SbY9trfsTpEmgGEiA6Yva?dl=0>.

## 4.1 Experiment 1

We consider a setting where  $F_{\mathbf{X}}$  is the cdf of a  $d$ -dimensional Gaussian mixture distribution with three components:  $F_{\mathbf{X}} = 0.3\mathcal{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.3\mathcal{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.4\mathcal{N}_d(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ . Here  $\boldsymbol{\mu}_1 = \mathbf{0}_d$ ,  $\boldsymbol{\mu}_2 = -3\mathbf{1}_d$ ,  $\boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2$ , and  $\boldsymbol{\Sigma}_K$ , for  $K = 1, 2, 3$ , are  $d \times d$  symmetric positive definite matrices with eigenvalues randomly generated from the interval  $[1, 10]$ . We consider two scenarios for simulating  $\mathcal{Y}_m$  from  $F_{\mathbf{Y}}$ . In Scenario I we let  $F_{\mathbf{Y}} = 0.1\mathcal{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.1\mathcal{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.8\mathcal{N}_d(\boldsymbol{\mu}_3, \boldsymbol{\Sigma}_3)$ . Thus,  $F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  since  $F_{\mathbf{Y}}$  has all the subpopulations present in  $F_{\mathbf{X}}$  but at different proportions. For Scenario II we consider  $F_{\mathbf{Y}} = 0.1\mathcal{N}_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + 0.1\mathcal{N}_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) + 0.8\mathcal{N}_d(\boldsymbol{\mu}_3, 0.25\boldsymbol{\Sigma}_3)$ . The third component in the above mixture differs with respect to its scale when compared to the third component of  $F_{\mathbf{X}}$ . So this setting presents a scenario where  $F_{\mathbf{Y}} \notin \mathcal{F}(F_{\mathbf{X}})$  and the composite null  $H_0$  is not true.

Table 1: Rejection rates at 5% level of significance: Experiment 1 and Scenario 1 wherein  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  is true.

| Method    | $n = 500, m = 50$ |          |          | $n = 2000, m = 200$ |          |          |
|-----------|-------------------|----------|----------|---------------------|----------|----------|
|           | $d = 5$           | $d = 15$ | $d = 30$ | $d = 5$             | $d = 15$ | $d = 30$ |
| EC test   | 0.224             | 0.144    | 0.134    | 0.710               | 0.364    | 0.270    |
| GEC test  | 0.568             | 0.592    | 0.618    | 0.984               | 0.986    | 0.992    |
| WEC test  | 0.724             | 0.738    | 0.750    | 0.998               | 0.994    | 0.998    |
| TRUH test | 0.028             | 0.030    | 0.020    | 0.020               | 0.018    | 0.008    |
| BGEC test | 0.000             | 0.000    | 0.038    | 0.000               | 0.000    | 0.010    |
| BWEC test | 0.000             | 0.000    | 0.040    | 0.000               | 0.000    | 0.010    |

Table 2: Rejection rates at 5% level of significance: Experiment 1 and Scenario 2 wherein  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  is false.

| Method    | $n = 500, m = 50$ |          |          | $n = 2000, m = 200$ |          |          |
|-----------|-------------------|----------|----------|---------------------|----------|----------|
|           | $d = 5$           | $d = 15$ | $d = 30$ | $d = 5$             | $d = 15$ | $d = 30$ |
| EC test   | 0.310             | 0.000    | 0.000    | 1.000               | 0.000    | 0.000    |
| GEC test  | 1.000             | 1.000    | 1.000    | 1.000               | 1.000    | 1.000    |
| WEC test  | 1.000             | 1.000    | 1.000    | 1.000               | 1.000    | 1.000    |
| TRUH test | 0.002             | 0.000    | 0.000    | 0.000               | 0.000    | 0.000    |
| BGEC test | 0.966             | 1.000    | 0.996    | 0.998               | 1.000    | 1.000    |
| BWEC test | 0.918             | 1.000    | 0.996    | 0.998               | 1.000    | 1.000    |

Table 1 reports the rejection rates for Scenario I for varying  $(n, m, d)$ . We note that TRUH, BWEC and BGEC return rejection rates that are below the prespecified 0.05 level and are conservative across the six testing scenarios considered in Table 1. The three edge count tests, on the other hand, have substantially higher rejection rates. This is not surprising as the three edge count statistics are designed to test the simple null hypothesis  $F_{\mathbf{X}} = F_{\mathbf{Y}}$  as opposed to the composite null hypothesis of Equation (7). For Scenario II, the rejection rates are reported in Table 2 and they reveal that with the exception of TRUH and the EC test, all other competing testing procedures are powerful in detecting departures from  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$ . When  $d$  is moderately high, the EC test, in particular, is known to have low power under sample size imbalance and the presence of scale alternatives exacerbates this problem. The WEC and GEC tests are designed to address these weaknesses of the EC test, and from Table 2 we observe substantially higher power for these two tests compared to the original EC test. The BWEC and BGEC tests, while conservative, continue to be powerful in detecting departures from the composite null hypothesis of Equation (7).

The performance of TRUH test under scenarios I and II is worth noting. From tables 1 and 2, we observe that while TRUH is conservative for testing  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  versus  $H_1 : F_{\mathbf{Y}} \notin \mathcal{F}(F_{\mathbf{X}})$ , it is considerably less powerful than BGEC and BWEC tests for detecting departures from  $H_0$ . In fact, across all our simulation settings TRUH, while conservative, is relatively less powerful than BGEC and BWEC tests. Such a behavior of TRUH is potentially due to its inability to detect departures from  $H_0$  when the components of  $F_{\mathbf{Y}}$  and  $F_{\mathbf{X}}$  differ only with respect to their scales.

## 4.2 Experiment 2

In the setting of Experiment 2, we let  $F_{\mathbf{X}}$  and  $F_{\mathbf{Y}}$  to include non-Gaussian components. So we let  $F_{\mathbf{X}} = 0.5 \text{ Gam}_d(\text{shape} = 5\mathbf{1}_d, \text{rate} = \mathbf{1}_d, \Sigma_1) + 0.5 \text{ Exp}_d(\text{rate} = \mathbf{1}_d, \Sigma_2)$ , where  $\text{Gam}_d$  and  $\text{Exp}_d$  are  $d$ -dimensional Gamma and Exponential distributions. The multivariate Gamma and Exponential distributions are constructed using a Gaussian copula and the R-package `lcmix` (Dvorkin, 2012; Xue-Kun Song, 2000) allows sampling from these distributions. The correlation matrices  $\Sigma_1$  and  $\Sigma_2$  are tapering matrices with positive and negative autocorrelations as follows:  $(\Sigma_1)_{ij} = 0.7^{|i-j|}$  and  $(\Sigma_2)_{ij} = -0.9^{|i-j|}$  for  $1 \leq i, j \leq d$ . For simulating  $\mathcal{Y}_n$  from  $F_{\mathbf{Y}}$ , we consider two scenarios. In Scenario I,  $F_{\mathbf{Y}} = 0.05 \text{ Gam}_d(\text{shape} = 5\mathbf{1}_d, \text{rate} = \mathbf{1}_d, \Sigma_1) + 0.95 \text{ Exp}_d(\text{rate} = \mathbf{1}_d, \Sigma_2)$ . In this setting  $F_{\mathbf{Y}}$  has both the components of  $F_{\mathbf{X}}$  but with different mixing proportions. So  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  is true, however this is a particularly challenging setting as a majority of the samples from  $F_{\mathbf{Y}}$  arise from only one of the components of  $F_{\mathbf{X}}$ . In Scenario II,  $F_{\mathbf{Y}} = 0.8 \text{ Gam}_d(\text{shape} = 5\mathbf{1}_d, \text{rate} = \mathbf{1}_d, \Sigma_1) + 0.2 \text{ Exp}_d(\text{rate} = 1.5\mathbf{1}_d, 0.25\Sigma_3)$  where  $(\Sigma_3)_{ij} = 0.9^{|i-j|}$ . In this setting, the second component of  $F_{\mathbf{Y}}$  differs from the second component of  $F_{\mathbf{X}}$  with respect to their correlation matrices, and so  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  is not true.

Table 3: Rejection rates at 5% level of significance: Experiment 2 and Scenario 1 wherein  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  is true.

| Method    | $n = 500, m = 50$ |          |          | $n = 2000, m = 200$ |          |          |
|-----------|-------------------|----------|----------|---------------------|----------|----------|
|           | $d = 5$           | $d = 15$ | $d = 30$ | $d = 5$             | $d = 15$ | $d = 30$ |
| EC test   | 0.418             | 0.312    | 0.258    | 0.944               | 0.798    | 0.684    |
| GEC test  | 0.812             | 0.810    | 0.774    | 1.000               | 1.000    | 1.000    |
| WEC test  | 0.928             | 0.908    | 0.912    | 1.000               | 1.000    | 1.000    |
| TRUH test | 0.000             | 0.000    | 0.000    | 0.000               | 0.000    | 0.000    |
| BGEC test | 0.002             | 0.004    | 0.000    | 0.000               | 0.000    | 0.000    |
| BWEC test | 0.000             | 0.004    | 0.000    | 0.000               | 0.000    | 0.000    |

Table 3 reports the rejection rates for Scenario I. We find that both TRUH and the bootstrapped WEC and GEC tests support  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  while the three edge count tests overwhelmingly reject  $H_0$ , which is an incorrect decision under Scenario I. Table 4 reports the rejection rates for Scenario II when the sample size imbalance is 0.05 as opposed

Table 4: Rejection rates at 5% level of significance: Experiment 2 and Scenario 2 wherein  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  is false.

| Method    | $n = 500, m = 25$ |          |          | $n = 2000, m = 100$ |          |          |
|-----------|-------------------|----------|----------|---------------------|----------|----------|
|           | $d = 5$           | $d = 15$ | $d = 30$ | $d = 5$             | $d = 15$ | $d = 30$ |
| EC test   | 0.424             | 0.250    | 0.096    | 0.978               | 0.926    | 0.806    |
| GEC test  | 0.724             | 0.822    | 0.838    | 1.000               | 1.000    | 1.000    |
| WEC test  | 0.824             | 0.886    | 0.904    | 1.000               | 1.000    | 1.000    |
| TRUH test | 0.050             | 0.040    | 0.036    | 0.026               | 0.050    | 0.024    |
| BGEC test | 0.188             | 0.332    | 0.398    | 0.190               | 0.990    | 1.000    |
| BWEC test | 0.194             | 0.340    | 0.378    | 0.188               | 0.990    | 1.000    |

to 0.1 in all of the earlier settings. Detecting departures from  $H_0$  under such imbalance is a difficult task as a majority of the samples from  $F_{\mathbf{Y}}$  arise from the first component of  $F_{\mathbf{X}}$  and the competing testing procedures must rely on a few observations to reject  $H_0$  in Scenario II. Table 4 reveals that the rejection rates of the three edge count tests and the proposed BWEC, BGEC tests are substantially higher than TRUH. While the edge count tests have the highest rejection rates across both scenarios I and II, the rejection rates of BWEC, BGEC tests improve as the sample size increases in Table 4. The two scenarios under Experiment 2 reveal that the bootstrapped calibrated WEC and GEC tests are conservative than edge count tests for testing the composite null hypothesis  $H_0 : F_{\mathbf{Y}} \in \mathcal{F}(F_{\mathbf{X}})$  and more powerful than TRUH for detecting departures from  $H_0$ .

In Section C of the supplement we report the performance of BWEC, BGEC tests under an additional setting where the samples  $\mathcal{X}_n$  and  $\mathcal{Y}_m$  exhibit zero inflation across the  $d$  dimensions.

## 5 Detecting player addiction in online video games

Online video game addiction is a phenomenon wherein a small subgroup of players are involved in excessive and compulsive use of these games that may ultimately result in social and/or emotional problems (Lemmens et al., 2009). Infact, game addiction was included as a disorder in the Diagnostic and Statistical Manual for Mental Disorders (see DSM-5-TR from the American Psychiatric Association<sup>2</sup>) because of an increased risk of clinically sig-

<sup>2</sup><https://psychiatry.org/psychiatrists/practice/dsm>

nificant problems associated with online gaming (Petry et al., 2014). Therefore, for game managers identifying and regulating addicted players is critical because incorrectly rewarding addiction via promotions may lead to high reputation risk for the gaming platform.

Extant research finds that players who login late at night exhibit a higher tendency towards game addiction (Lee and Kim, 2017) and until recently South Korea had prohibited young players from playing online video games between midnight and 6:00 AM. In this application, we rely on an anonymized data available from a large video game company in Asia to test whether players who login after midnight exhibit deviant playing behavior when compared to players with baseline gaming behavior. Our data hold player level information for  $d = 16$  playing characteristics, such as player’s game level, number of friends that they have, number of strategic missions that they completed in the game, etc across 7 days (see Table 11 in Supplementary Section D for a description of these characteristics). For each day, we have access to the following three samples; a sample of players who login post midnight (**Late**)  $\mathcal{Y}_1 = \{\mathbf{Y}_{11}, \dots, \mathbf{Y}_{1m_1}\}$ , an independent sample of players who login between 8 AM - 9 AM local time (**Early**),  $\mathcal{Y}_2 = \{\mathbf{Y}_{21}, \dots, \mathbf{Y}_{2m_2}\}$  and an independent sample of players who exhibit normal playing behavior (**Baseline**)  $\mathcal{X}_n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ . Suppose  $\mathcal{Y}_1$  are a random sample from a distribution with CDF  $F_{\mathbf{Y}_1}^{(1)}$  and  $\mathcal{Y}_2$  are a random sample from a distribution with CDF  $F_{\mathbf{Y}_2}^{(2)}$ . We consider the following two hypothesis testing problems: (i) whether the playing behavior of **Late** players is different from **Baseline** players,  $H_{01} : F_{\mathbf{Y}_1}^{(1)} \in \mathcal{F}(F_{\mathbf{X}})$  vs  $H_{11} : F_{\mathbf{Y}_1}^{(1)} \notin \mathcal{F}(F_{\mathbf{X}})$ , and (ii) whether the playing behavior of **Early** players is different from **Baseline** players,  $H_{02} : F_{\mathbf{Y}_2}^{(2)} \in \mathcal{F}(F_{\mathbf{X}})$  vs  $H_{12} : F_{\mathbf{Y}_2}^{(2)} \notin \mathcal{F}(F_{\mathbf{X}})$ .

Figure 3 provides a box-plot of the 16 playing characteristics on day 5. It reveals that **Late** players seem to play relatively larger number of games (num\_games), spend more time playing with the game environment (pve\_requests and pve\_time) and are relatively more engaged with their friends and social connections within the game (numfrnd\_games, numfrnd\_played, numguild\_played) than their counterparts in **Baseline**. The **Early** players, on the other hand, do not exhibit such stark contrasts in their playing behavior when compared to the **Baseline**. A t-SNE plot (Van der Maaten and Hinton, 2008) of the  $d = 16$  dimensional



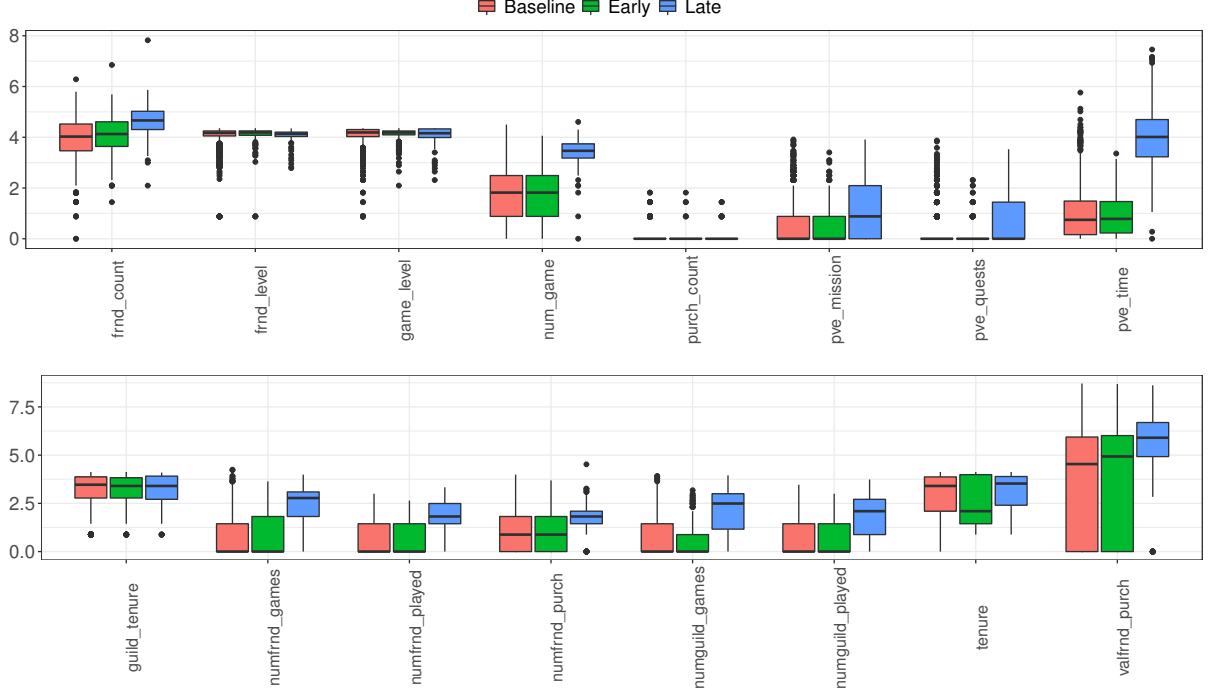


Figure 3: Box plot of the 16 playing characteristics on day 5. Data are  $\text{arcsinh}$  transformed and the variable `pve_time` is reported in hours. Here  $m_1 = 143$ ,  $m_2 = 216$  and  $n = 2,340$ . See Table 11 in Supplementary Section D for a description of these characteristics.

data in figures 4 and 5 provide further insights into the behavior of the **Late** and **Early** players. For both days 1 and 4, these figures exhibit the underlying heterogeneity in the **Baseline** player sample. Moreover, the **Late** sample occupies a distinct position in the two dimensional space that is away from the bulk of the **Baseline** sample.

Tables 5 and 6 report the  $p$ -values for the two testing problems. In Table 5, all competing tests, with the exception of TRUH, reject the null hypothesis  $H_{01} : F_{\mathbf{Y}_1}^{(1)} \in \mathcal{F}(F_{\mathbf{X}})$  across the 7 days and conclude that the playing behavior of the **Late** players is significantly different from **Baseline**. This corroborates the visual evidence found in figures 3, 4 and 5 that indeed the players who login late differ from their counterparts in **Baseline** as far as these 16 playing characteristics are concerned. TRUH, on the other hand, does not provide such a consistent picture across the seven days and fails to reject  $H_{01}$  in four out of the seven days. The numerical experiments of Section 4 reveal that TRUH is relatively less powerful than BGEC and BVEC tests in detecting departures from  $H_{01}$  and the result in Table

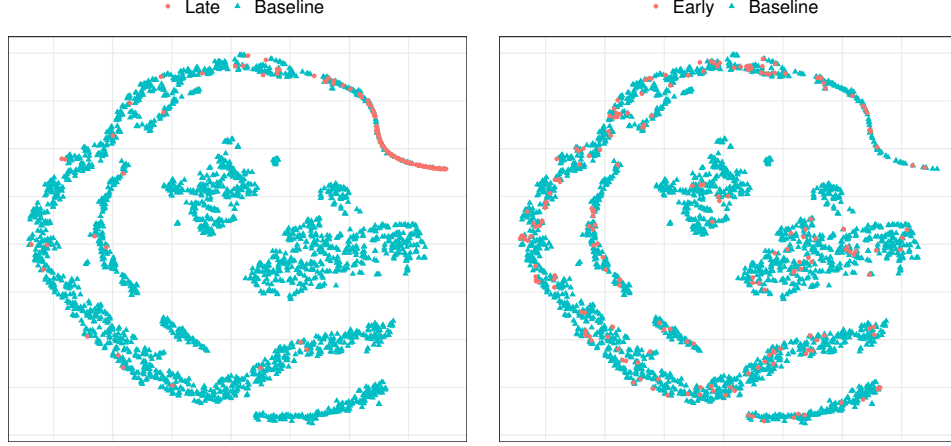


Figure 4: A t-SNE plot of the data for Day 1. The  $d = 16$  playing attributes are projected to a two dimensional space.

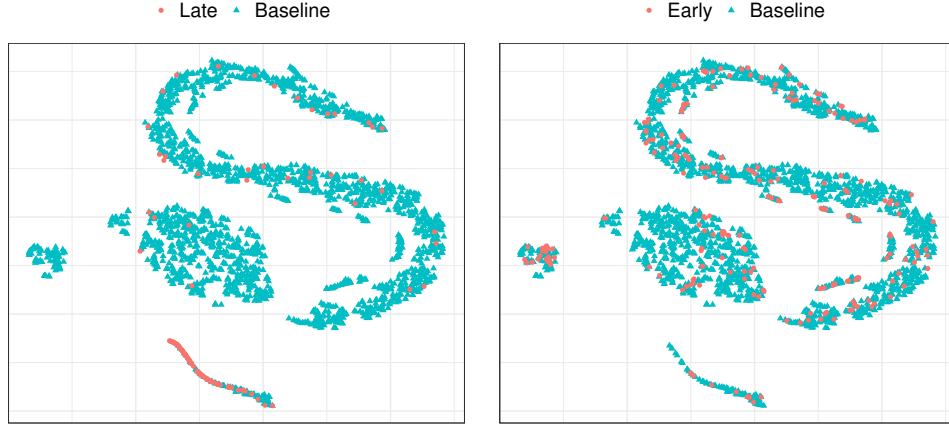


Figure 5: A t-SNE plot of the data for Day 4. The  $d = 16$  playing attributes are projected to a two dimensional space.

| Day | $n$   | $m_1$ | EC test     | GEC test    | WEC test    | TRUH        | BGEC test   | BWEC test   |
|-----|-------|-------|-------------|-------------|-------------|-------------|-------------|-------------|
| 1   | 2,558 | 133   | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | 1.000       | $< 10^{-3}$ | $< 10^{-3}$ |
| 2   | 2,374 | 103   | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ |
| 3   | 2,073 | 163   | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | 1.000       | $< 10^{-3}$ | $< 10^{-3}$ |
| 4   | 2,291 | 126   | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ |
| 5   | 2,340 | 143   | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | 1.000       | $< 10^{-3}$ | $< 10^{-3}$ |
| 6   | 2,560 | 72    | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | 1.000       | $< 10^{-3}$ | $< 10^{-3}$ |
| 7   | 2,268 | 140   | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ |

Table 6:  $p$ -values for testing  $H_{02} : F_{\mathbf{Y}_2}^{(2)} \in \mathcal{F}(F_{\mathbf{X}})$  vs  $H_{12} : F_{\mathbf{Y}_2}^{(2)} \notin \mathcal{F}(F_{\mathbf{X}})$

| Day | $n$   | $m_2$ | EC test     | GEC test    | WEC test    | TRUH  | BGEC test | BWEC test |
|-----|-------|-------|-------------|-------------|-------------|-------|-----------|-----------|
| 1   | 2,558 | 203   | 0.702       | $< 10^{-3}$ | $< 10^{-3}$ | 0.525 | 0.863     | 0.917     |
| 2   | 2,374 | 223   | 0.193       | $< 10^{-3}$ | $< 10^{-3}$ | 0.995 | 1.000     | 1.000     |
| 3   | 2,073 | 171   | 0.864       | 0.001       | 0.002       | 0.715 | 1.000     | 1.000     |
| 4   | 2,291 | 232   | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | 0.795 | 1.000     | 1.000     |
| 5   | 2,340 | 216   | $< 10^{-3}$ | $< 10^{-3}$ | $< 10^{-3}$ | 0.105 | 0.995     | 0.998     |
| 6   | 2,560 | 177   | 0.487       | $< 10^{-3}$ | $< 10^{-3}$ | 0.890 | 1.000     | 1.000     |
| 7   | 2,268 | 199   | 0.708       | $< 10^{-3}$ | $< 10^{-3}$ | 0.855 | 0.978     | 0.988     |

5 is potentially another empirical evidence in that direction. The  $p$ -values reported in Table 6 exhibit an interesting pattern for the testing problem  $H_{02} : F_{\mathbf{Y}_2}^{(2)} \in \mathcal{F}(F_{\mathbf{X}})$  vs  $H_{12} : F_{\mathbf{Y}_2}^{(2)} \notin \mathcal{F}(F_{\mathbf{X}})$ . We note that TRUH, BWEC and BGEC tests fail to reject  $H_{02}$  across the seven days, while the decisions from GEC and WEC tests are exactly the opposite, potentially demonstrating the non-conservativeness of these tests for testing the composite null hypothesis  $H_{02}$ . On five out of the seven days, EC test fails to reject  $H_{02}$  and gives the impression that it is conservative for testing  $H_{02}$ . However, and as observed in Section 4, at moderately high dimensions the EC test statistic suffers from variance boosting under sample size imbalance and demonstrates low power.

## 6 Discussion

The multivariate samples from modern gargantuan datasets often involve heterogeneity and direct application of existing nonparametric two-sample tests, such as the edge count tests, may lead to incorrect scientific decisions if they are not properly calibrated for the underlying latent heterogeneity. In this article, we demonstrate that under a composite null hypothesis, that allows the possibility that two samples may originate from the same mixture distribution but with possibly different mixing weights, the weighted edge count test can be re-calibrated to obtain a test which is asymptotically powerful for general alternatives. For practical implementation of this test, we develop a bootstrap calibrated weighted edge count test and demonstrate its excellent finite sample properties across a wide range of simulation studies. On a real world video game dataset, we use our testing procedure to detect addictive playing behavior and find that in comparison to players

who exhibit normal playing behavior, players who login late to the game exhibit aberrant behavior while players who login early do not, thus confirming some of the findings in existing literature related to video game addiction.

Our future research will be directed towards extending the proposed testing framework in two directions. First, it will be interesting to develop an extension of the kernel two-sample tests of [Gretton et al. \(2007\)](#) to the heterogeneous setting and devise an efficient re-calibration procedure for these tests for consistent two-sample testing under the composite null hypothesis of Equation (7). Second, for dealing with samples that involve high dimensional features, such as single cell RNA-seq data where  $d \sim 10^4$ , or data on consumer preferences for high dimensional product attributes, we will be interested in developing an extension of our testing procedure that allows incorporating such data-types into our heterogeneous two-sample testing framework.

## References

- Sumit Agarwal, Pulak Ghosh, Jing Li, and Tianyue Ruan. Digital payments induce over-spending: Evidence from the 2016 demonetization in india. 2019. URL [https://abfer.org/media/abfer-events-2019/annual-conference/economic-transformation-of-asia/AC19P4028\\_Digital\\_Payments\\_Induce\\_Excessive\\_Spending\\_Evidence\\_from\\_Demonetization\\_in\\_India.pdf](https://abfer.org/media/abfer-events-2019/annual-conference/economic-transformation-of-asia/AC19P4028_Digital_Payments_Induce_Excessive_Spending_Evidence_from_Demonetization_in_India.pdf). 3
- Ikpe Justice Akpan, Elijah Abasifreke Paul Udoh, and Bamidele Adebisi. Small business awareness and adoption of state-of-the-art technologies in emerging and developing markets, and lessons from the covid-19 pandemic. *Journal of Small Business & Entrepreneurship*, 34(2):123–140, 2022. 3
- B Aslan and G Zech. New test for the multivariate two-sample problem based on the concept of minimum energy. *Journal of Statistical Computation and Simulation*, 75(2): 109–119, 2005. 2
- Janet Balis. 10 truths about marketing after the pandemic. 2021. URL <https://hbr.org/2021/03/10-truths-about-marketing-after-the-pandemic>. 3
- Trambak Banerjee, Bhaswar B Bhattacharya, and Gourab Mukherjee. A nearest-neighbor based nonparametric test for viral remodeling in heterogeneous single-cell proteomic data. *Annals of Applied Statistics*, 14(4):1777–1805, 2020. 2, 9, 10, 14, 19
- Trambak Banerjee, Peng Liu, Gourab Mukherjee, Shantanu Dutta, and Hai Che. Joint modeling of playing time and purchase propensity in massively multiplayer online role playing games using crossed random effects. 2022. URL <https://gmukherjee.github.io/pdfs/crejm.pdf>. 6
- Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of multivariate analysis*, 88(1):190–206, 2004. 2

- Alexander W Bartik, Marianne Bertrand, Zoe Cullen, Edward L Glaeser, Michael Luca, and Christopher Stanton. The impact of covid-19 on small business outcomes and expectations. *Proceedings of the national academy of sciences*, 117(30):17656–17666, 2020. [3](#)
- Bhaswar B Bhattacharya. A general asymptotic framework for distribution-free graph-based two-sample tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):575–602, 2019. [2](#)
- Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015. [16](#)
- Peter J Bickel. A distribution free version of the smirnov two sample test in the p-variate case. *The Annals of Mathematical Statistics*, 40(1):1–23, 1969. [2](#)
- Kayla Bruun. Supply chain disruptions limit consumer spending. 2021. URL <https://morningconsult.com/2021/09/27/supply-chain-disruptions-limit-consumer-spending/>. [3](#)
- Ben J Callahan, Kris Sankaran, Julia A Fukuyama, Paul J McMurdie, and Susan P Holmes. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. *F1000Research*, 5, 2016. [2](#)
- Hao Chen and Jerome H Friedman. A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association*, 112(517):397–409, 2017. [2](#), [8](#), [12](#), [13](#), [19](#)
- Hao Chen and Nancy Zhang. Graph-based change-point detection. *The Annals of Statistics*, 43(1):139–176, 2015. [2](#)
- Hao Chen, Xu Chen, and Yi Su. A weighted edge-count two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 113(523):1146–1155, 2018. [2](#), [4](#), [8](#), [12](#), [13](#), [16](#)
- Lisha Chen, Winston Wei Dou, and Zhihua Qiao. Ensemble subsampling for imbalanced multivariate two-sample tests. *Journal of the American Statistical Association*, 108(504):1308–1323, 2013. [2](#)
- James H Chung and Donald AS Fraser. Randomization tests for a multivariate two-sample problem. *Journal of the American Statistical Association*, 53(283):729–735, 1958. [2](#)
- Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28, 2015. [2](#)
- Knut Conradsen, Allan Aasbjerg Nielsen, Jesper Schou, and Henning Skriver. A test statistic in the complex wishart distribution and its application to change detection in polarimetric sar data. *IEEE Transactions on Geoscience and Remote Sensing*, 41(1):4–19, 2003. [2](#)
- Nicolas Crouzet, Apoorv Gupta, and Filippo Mezzanotti. Shocks and technology adoption: Evidence from electronic payment systems. *Techn. rep., Northwestern University Working Paper*, 2019. [3](#)
- Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, pages 1–16, 2021. [3](#)

- Daniel Dvorkin. *lcmix: Layered and chained mixture models*, 2012. URL <https://R-Forge.R-project.org/projects/lcmix/>. R package version 0.3/r5. 22
- Michael T Fahey, Christopher W Thane, Gemma D Bramwell, and W Andy Coward. Conditional gaussian mixture modelling for dietary pattern analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(1):149–166, 2007. 4
- Karen B Farris and Donald P Schopflocher. Between intention and behavior: an application of community pharmacists’ assessment of pharmaceutical care. *Social science & medicine*, 49(1):55–66, 1999. 2
- Valerie S Folkes, Susan Koletsky, and John L Graham. A field study of causal inferences and consumer reaction: the view from the airport. *Journal of consumer research*, 13(4): 534–539, 1987. 2
- Jerome H Friedman and Lawrence C Rafsky. Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *The Annals of Statistics*, pages 697–717, 1979. 2, 8, 11
- Julia Fukuyama. *phyloseqgraphtest: Graph-based permutation tests for microbiome data*. 2020. URL <https://cran.rstudio.com/web/packages/phyloseqGraphTest/index.html>. 2
- Promit Ghosal and Bodhisattva Sen. Multivariate ranks and quantiles using optimal transportation and applications to goodness-of-fit testing. *arXiv preprint arXiv:1905.05340*, 2019. 3
- Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007. 2, 28
- Peter Hall and Nader Tajvidi. Permutation tests for equality of distributions in high-dimensional settings. *Biometrika*, 89(2):359–374, 2002. 2
- Ruth Heller, Shane T Jensen, Paul R Rosenbaum, and Dylan S Small. Sensitivity analysis for the cross-match test, with applications in genomics. *Journal of the American Statistical Association*, 105(491):1005–1013, 2010a. 2
- Ruth Heller, Paul R Rosenbaum, and Dylan S Small. Using the cross-match test to appraise covariate balance in matched pairs. *The American Statistician*, 64(4):299–309, 2010b. 2
- Norbert Henze. On the number of random points with nearest neighbour of the same type and a multivariate two-sample test. *Metrika*, 31:259–273, 1984. 2
- Norbert Henze. A multivariate two-sample test based on the number of nearest neighbor type coincidences. *The Annals of Statistics*, 16(2):772–783, 1988. 12
- Norbert Henze and Mathew Penrose. On the multivariate runs test. *The Annals of Statistics*, 27(1):290–298, 1999. 12, 16
- Susan Holmes and Wolfgang Huber. *Modern statistics for modern biology*. Cambridge University Press, 2018. 2, 3
- Jay G Hull, Timothy J Brunelle, Anna T Prescott, and James D Sargent. A longitudinal study of risk-glorifying video games and behavioral deviance. *Journal of personality and social psychology*, 107(2):300, 2014. 6

- Aino Koskenniemi. Deviant consumption meets consumption-as-usual: The construction of deviance and normality within consumer research. *Journal of Consumer Culture*, 21(4):827–847, 2021. [5](#)
- Wouter Labeeuw and Geert Deconinck. Residential electrical load model based on mixture model clustering and markov models. *IEEE Transactions on Industrial Informatics*, 9(3):1561–1569, 2013. [4](#)
- Changho Lee and Ocktae Kim. Predictors of online game addiction among korean adolescents. *Addiction Research & Theory*, 25(1):58–66, 2017. [10](#), [24](#)
- Jeroen S Lemmens, Patti M Valkenburg, and Jochen Peter. Development and validation of a game addiction scale for adolescents. *Media psychology*, 12(1):77–95, 2009. [23](#)
- Eric W Liguori and Thomas G Pittz. Strategies for small business: Surviving and thriving in the era of covid-19. *Journal of the International Council for Small Business*, 1(2):106–110, 2020. [3](#)
- Somabha Mukherjee, Divyansh Agarwal, Nancy R Zhang, and Bhaswar B Bhattacharya. Distribution-free multisample tests based on optimal matchings with applications to single cell genomics. *Journal of the American Statistical Association*, 117(538):627–638, 2022. [2](#)
- Mathew D Penrose and Joseph E Yukich. Weak laws of large numbers in geometric probability. *The Annals of Applied Probability*, 13(1):277–303, 2003. [16](#)
- Nancy M Petry, Florian Rehbein, Douglas A Gentile, Jeroen S Lemmens, Hans-Jürgen Rumpf, Thomas Mößle, Gallus Bischof, Ran Tao, Daniel SS Fung, Guilherme Borges, et al. An international consensus for assessing internet gaming disorder using the new dsm-5 approach. *Addiction*, 109(9):1399–1406, 2014. [24](#)
- Yasir Rahmatallah, Frank Emmert-Streib, and Galina Glazko. Gene set analysis for self-contained tests: complex null and specific alternative hypotheses. *Bioinformatics*, 28(23):3073–3080, 2012. [2](#)
- Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. [2](#)
- Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017. [2](#)
- Paul R Rosenbaum. An exact distribution-free test comparing two multivariate distributions based on adjacency. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):515–530, 2005. [2](#), [12](#)
- Peter E Rossi, Greg M Allenby, and Rob McCulloch. *Bayesian statistics and marketing*. John Wiley & Sons, 2012. [3](#)
- Russell L Rothman, Ryan Housam, Hilary Weiss, Dianne Davis, Rebecca Gregory, Tebeb Gebretsadik, Ayumi Shintani, and Tom A Elasy. Patient understanding of food labels: the role of literacy and numeracy. *American journal of preventive medicine*, 31(5):391–398, 2006. [2](#)
- Mark F Schilling. Multivariate two-sample tests based on nearest neighbors. *Journal of the American Statistical Association*, 81(395):799–806, 1986. [2](#), [12](#)



- Hongjian Shi, Mathias Drton, and Fang Han. Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, pages 1–16, 2020a. [3](#)
- Hongjian Shi, Marc Hallin, Mathias Drton, and Fang Han. On universally consistent and fully distribution-free rank tests of vector independence. *arXiv preprint arXiv:2007.02186*, 2020b. [3](#)
- Xiaoping Shi, Yuehua Wu, and Calyampudi Radhakrishna Rao. Consistent and powerful graph-based change-point test for high-dimensional data. *Proceedings of the National Academy of Sciences*, 114(15):3873–3878, 2017. [2](#)
- Gábor J Székely. E-statistics: The energy of statistical samples. *Bowling Green State University, Department of Mathematics and Statistics Technical Report*, 3(05):1–18, 2003. [2](#)
- Gábor J Székely and Maria L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 5(16.10):1249–1272, 2004. [2](#)
- Gábor J Székely and Maria L Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013. [2](#)
- Robert Tibshirani and Guenther Walther. Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, 14(3):511–528, 2005. [17](#)
- Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. [24](#)
- Peter Xue-Kun Song. Multivariate dispersion models generated from gaussian copula. *Scandinavian Journal of Statistics*, 27(2):305–320, 2000. [22](#)
- Puning Zhao and Lifeng Lai. Analysis of KNN density estimation. *IEEE Transactions on Information Theory*, 2022. [16](#)