Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ubes20

Detecting Weak Distribution Shifts via Displacement Interpolation

YoonHaeng Hur & Tengyuan Liang

To cite this article: YoonHaeng Hur & Tengyuan Liang (07 May 2024): Detecting Weak Distribution Shifts via Displacement Interpolation, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2024.2335957

To link to this article: https://doi.org/10.1080/07350015.2024.2335957

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2024, VOL. 00, NO. 0, 1–13
https://doi.org/10.1080/07350015.2024.2335957

Detecting Weak Distribution Shifts via Displacement Interpolation

YoonHaeng Hur[a] and Tengyuan Liang[b]

[a]Department of Statistics, University of Chicago, Chicago, IL; [b]Booth School of Business, University of Chicago, Chicago, IL

ABSTRACT

Detecting weak, systematic distribution shifts and quantitatively modeling individual, heterogeneous responses to policies or incentives have found increasing empirical applications in social and economic sciences. Given two probability distributions P (null) and Q (alternative), we study the problem of detecting weak distribution shift deviating from the null P toward the alternative Q, where the level of deviation vanishes as a function of n, the sample size. We propose a model for weak distribution shifts via displacement interpolation between P and Q, drawing from the optimal transport theory. We study a hypothesis testing procedure based on the Wasserstein distance, derive sharp conditions under which detection is possible, and provide the exact characterization of the asymptotic Type I and Type II errors at the detection boundary using empirical processes. We demonstrate how the proposed testing procedure works in modeling and detecting weak distribution shifts in real datasets using two empirical examples: distribution shifts in consumer spending after COVID-19, and heterogeneity in the published p-values of statistical tests in journals across different disciplines.

ARTICLE HISTORY
Received May 2023
Accepted March 2024

KEYWORDS
Displacement interpolation; Hypothesis testing; Optimal transport; Phase transition; Weak distribution shifts

1. Introduction

Classic detection problem aims to distinguish a shifted distribution Q from the null distribution P based on data, formulated as a nonparametric goodness-of-fit test:

$$H_0 : X_1, \ldots, X_n \overset{\text{iid}}{\sim} P \quad \text{versus} \quad H_1 : X_1, \ldots, X_n \overset{\text{iid}}{\sim} Q. \tag{1.1}$$

It is understood that the power of certain test statistics, such as Kolmogorov-Smirnov (Kolmogorov 1933; Smirnoff 1939) or Anderson-Darling (Anderson and Darling 1952), is asymptotically 1, implying that detection is possible if the sample size is large.

In many applications, one confronts situations where the signal in the alternative distribution is weaker. One natural formulation is to replace Q in H_1 (1.1) by a suitable interpolation scheme between P and Q. A common choice is the linear interpolation $(1-\epsilon)P+\epsilon Q$, or the so-called Huber's ϵ-contamination model (Huber 1964), where one parameterizes $\epsilon = \epsilon_n \to 0$ as $n \to \infty$ to represent weak signals. This model represents that only a small ϵ-fraction of the data deviates from P, often seen in applications such as large-scale inference with high-throughput measurements (Efron 2010). For instance, microarray data—widely used in genetics and genomics—measure the expression level of thousands of genes, where only some portions are relevant to detecting certain diseases. See Donoho and Jin (2015) for a comprehensive overview of such applications. In those applications, weak signals are modeled as small perturbations in the frequencies of the histogram of the data, that is, deviations along the y-axis.

Detecting the presence of distribution shifts is also essential in social and economic applications. There, the main interest is often to quantify individual, heterogeneous responses to policies or other incentives. For instance, consider the economic impacts of the coronavirus pandemic (COVID-19) and government policies (Chetty et al. 2020); when analyzing how the economy—measured by weekly statistics such as consumer spending—recovers from the shock caused by the pandemic, it is natural to consider individual, heterogeneous shifts to reflect that each individual adjusts the spending differently according to the income level and other characteristics. In this context, the individual responses can be modeled as perturbations along the x-axis of the data histogram, whereas the aforementioned perturbation along the y-axis (frequencies) rooted in engineering applications is arguably less informative. Another example is the study of the effect of financial or nonfinancial incentives given to students or parents to improve educational performance, measured by test scores (Fryer et al. 2015; Levitt et al. 2016). There, the main interest is measuring the shifts in test score distributions in response to the incentives. It is worth noting that the shifts are arguably heterogeneous, depending on gender, race/ethnicity, and other demographic information (Levitt et al. 2016).

In the above examples, perturbations along the y-axis—the fraction of individuals deemed responsive to a given policy or incentive—may be of interest but are insufficient to model individual, heterogeneous responses. Unlike the earlier genetics example, where the identification of a handful of non-null genes showing distinct expression levels provides crucial scientific evi-

CONTACT YoonHaeng Hur *yoonhaenghur@uchicago.edu* Department of Statistics, University of Chicago, Chicago, IL.

dence of certain diseases, the goal of socioeconomic research is beyond simply finding a proportion of shifted observations; the fundamental interest is to quantify individual, heterogeneous responses to policy, which is more relevant to the shift along the x-axis of histograms.

This article proposes a different interpolation scheme for detecting weak distribution shifts motivated by the above. We study a natural test statistic motivated by optimal transport and conduct an exact study of the asymptotic power of the test. We represent the signal strength as the level of deviation, rather than the proportion of deviated units. In a nutshell, we are interested in how much the data histogram shifts along the x-axis instead of the y-axis. To this end, we model weak signals using displacement interpolation (McCann 1997) with a viewpoint from optimal transport theory (Villani 2003), where the interpolation parameter ϵ represents a certain transport distance of data from the null P along the geodesics, thus, characterizing the level of distribution shifts from P.

1.1. Displacement Interpolation and Optimal Transport

Let P and Q be Borel probability measures on $\mathbb{R}$ whose cumulative distribution functions are F and G, respectively; also, let F^{-1} and G^{-1} be their quantile functions, respectively. Displacement interpolation between P and Q is defined as

$$P_\epsilon := ((1-\epsilon)\mathrm{Id} + \epsilon T)_\# P \quad \forall \epsilon \in [0,1],$$

where $\mathrm{Id}\colon \mathbb{R} \to \mathbb{R}$ is the identity map and $T = G^{-1} \circ F$, namely, the composition of G^{-1} and F. Here, $S_\# P$ denotes the pushforward measure of P by a map $S\colon \mathbb{R} \to \mathbb{R}$ defined by $S_\# P(B) = P\{x \in \mathbb{R} : S(x) \in B\}$ for any Borel set $B \subset \mathbb{R}$. If P is absolutely continuous with respect to the Lebesgue measure, it is known that $T = G^{-1} \circ F$ is the unique monotone map such that $T_\# P = Q$. More importantly, it is also the unique solution to the following optimal transport problem (Brenier 1991; Villani 2003) provided P and Q have finite second moments, namely,

$$T = \underset{\substack{S\colon \mathbb{R}\to\mathbb{R} \\ S_\# P = Q}}{\operatorname{argmin}} \int_{\mathbb{R}} |x - S(x)|^2 \mathrm{d}P(x). \tag{1.2}$$

Intuitively, the constraint $S_\# P = Q$ means that a map S transports mass from the source distribution P to the target distribution Q by moving the infinitesimal mass at x in the support of P to $S(x)$ such that $\mathrm{d}P(x) = \mathrm{d}Q(S(x))$. Accordingly, (1.2) tells that the unique monotone map $T = G^{-1} \circ F$ provides the optimal way of transporting mass from P to Q, minimizing the squared transport distance $|x - T(x)|^2$ on average.

Now, for each x, let us view the segment $\{(1-\epsilon)x + \epsilon T(x) : \epsilon \in [0,1]\}$ as the transport path from x to the destination $T(x)$ along its displacement $T(x) - x$. It turns out that transporting mass from the source distribution P along such a path gives rise to a geodesic connecting P and Q in the Wasserstein space (Ambrosio, Gigli, and Savaré 2005), namely, $\epsilon = \frac{W_p(P,P_\epsilon)}{W_p(P,Q)}$ for all $\epsilon \in [0,1]$, where for $p \geq 1$, we denote by W_p the Wasserstein-p distance defined by

$$W_p(\mu,\nu) := \left(\int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^p \mathrm{d}u \right)^{1/p}$$

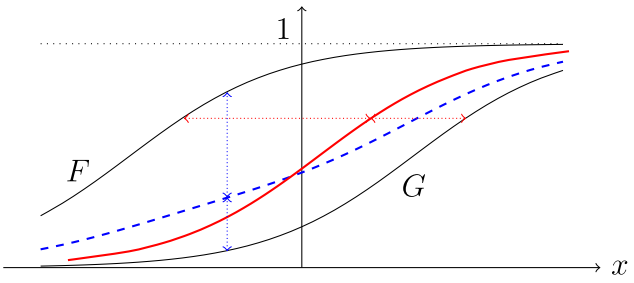


Figure 1. Illustration of two interpolation schemes. Linear interpolation $(1-\epsilon)P + \epsilon Q$ vertically combines the cumulative distribution functions F and G; namely, its cumulative distribution function (blue, dashed) is $(1-\epsilon)F + \epsilon G$. Meanwhile, displacement interpolation (red, solid) horizontally combines F and G, or equivalently, vertically combines the quantile functions F^{-1} and G^{-1}. In other words, the quantile function of $((1-\epsilon)\mathrm{Id} + \epsilon G^{-1} \circ F)_\# P$ is $(1-\epsilon)F^{-1} + \epsilon G^{-1}$.

for any probability measures μ, ν whose quantile functions are F_μ^{-1}, F_ν^{-1}, respectively. To see this, it suffices to observe that the quantile function of P_ϵ is $(1-\epsilon)F^{-1} + \epsilon G^{-1}$ as $P = (F^{-1})_\# U$ implies $P_\epsilon = ((1-\epsilon)\mathrm{Id} + \epsilon G^{-1} \circ F)_\#((F^{-1})_\# U) = ((1-\epsilon)F^{-1} + \epsilon G^{-1})_\# U$, where U is the Lebesgue measure on $[0,1]$; see Figure 1 for details. Therefore, displacement interpolation represents the optimal path—shortest path under the distance W_p—for transporting mass from P to Q, where the parameter ϵ naturally characterizes the deviation from P via the relative distance $\epsilon = \frac{W_p(P,P_\epsilon)}{W_p(P,Q)}$.[1]

We finish this section with a concrete example to contrast two interpolation schemes. Consider $P = N(0,1)$ and $Q = N(\tau, \sigma^2)$ for some $\tau, \sigma > 0$. Linear interpolation $(1-\epsilon)P + \epsilon Q$ is the mixture of two Gaussian distributions P, Q, where the parameter ϵ is essentially the frequency of signals from Q. Displacement interpolation is the optimal transport path from P to Q, where $T(x) = \sigma x + \tau$ for all $x \in \mathbb{R}$ and $P_\epsilon := ((1-\epsilon)\mathrm{Id} + \epsilon T)_\# P = N(\epsilon\tau, (1 - \epsilon + \epsilon\sigma)^2)$; in words, the optimal path is simply to move each x to $\sigma x + \tau$. The resulting interpolation $N(\epsilon\tau, (1 - \epsilon + \epsilon\sigma)^2)$ suggests that the parameter ϵ provides an intuitive characterization of the signal strength which controls the level of distribution shift. Lastly, it is worth noting that displacement interpolation preserves the unimodality as the interpolation is always Gaussian, whereas linear interpolation may result in two modes.

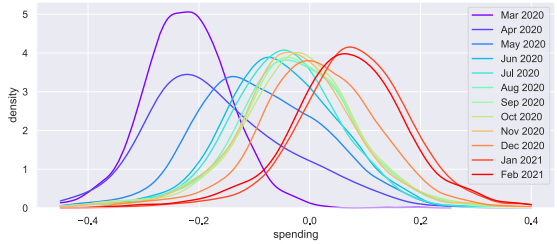
1.2. Distribution Shift as Displacement Interpolation

As discussed in the previous section, displacement interpolation constructs an optimal path transporting mass from P to Q, providing a natural notion of distribution shift from P to Q. Such a viewpoint has recently found several applications in machine learning and computer vision, where displacement interpolation serves as a method to optimally synthesize two objects, such as textures (Rabin et al. 2012), colors (Kolouri et al. 2017), and styles (Mroueh 2020).

Before diving into theory and methodology, in this section, we adopt this viewpoint to model distribution shifts using real-world data; we see that displacement interpolation is better

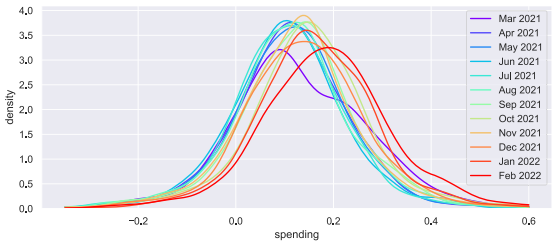
[1] It is possible to generalize the aforementioned concepts—displacement interpolation, optimal transport, and the Wasserstein-p distance—to $\mathbb{R}^d$ with $d > 1$; see (Villani 2003).

(a) Daily average consumer spending



(b) Monthly spending

12 months from March 2020



(c) Monthly spending

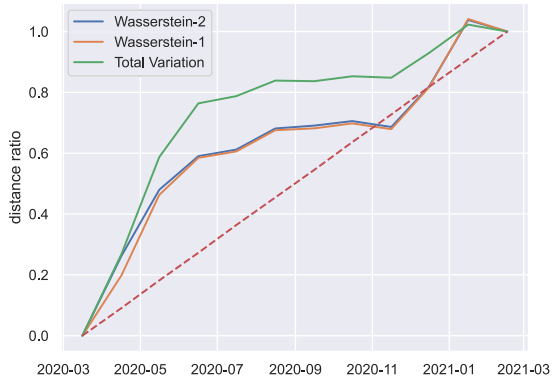
12 months from March 2021

Figure 2. (a) plots a time series of consumer spending from January 13, 2020 to June 5, 2022, where some noticeable events are marked: March 13, 2020 (national emergency declared) and April 15, 2020, January 4, 2021, and March 17, 2021 (first, second, and third stimulus payments start, respectively). (b) and (c) show the smoothed histograms of monthly average consumer spending by county from March 16, 2020 to March 15, 2021 and from March 16, 2021 to March 15, 2022, respectively.

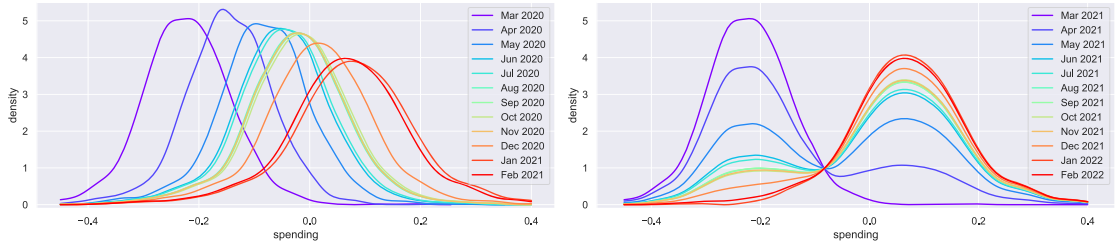
suited to model such distribution shifts than linear interpolation. Later, after laying out the theory, we revisit this empirical example to study the power of our testing procedure. To this end, we use the data from (Chetty et al. 2020), which studies the economic impacts of the coronavirus pandemic (COVID-19) and policy responses in the United States using a wide range of statistics, such as consumer spending, business revenues, employment rates, and so on.[2] Here, we focus on the consumer spending data—recorded as seasonally adjusted percent changes based on anonymized card transactions data—and analyze how consumer expenditures have recovered from the steep plunge caused by COVID-19. Figure 2(a) shows the average consumer spending between January 2021 and June 2022. More specifically, we look at the distribution of monthly average consumer spending over 1655 counties, where the spending is disaggregated based on the ZIP code where the cardholder lives. The monthly average is calculated by averaging the daily expenditures from the 16th of each month to the 15th of the following month. Figure 2(b) shows the smoothed histograms of monthly average spending for the first 12 months since March 2020, which clearly shows how the distribution has shifted in the increasing spending direction, namely, the spending is recovering from the shock of the pandemic. Meanwhile, Figure 2(c) shows the next 12 months from March 2021, where we can no longer observe such an evident distribution shift, suggesting the consumer spending has stabilized after the recovery; see also the corresponding period in Figure 2(a).

[2]Data are publicly available at *https://tracktherecovery.org/*.

Empirically, we illustrate that displacement interpolation serves as a reasonable model for the distribution shift during the recovery period shown in Figure 2(b). To this end, we generate two interpolation paths, displacement interpolation and linear interpolation, and contrast them with the real data. First, let $\{P_{i/11}\}_{i=0}^{11}$ be the distributions of monthly spending by county during that period, namely, they correspond to the 12 histograms of Figure 2(b); here, P_0 and P_1 are the start and end of that period (from March 16, 2020 to April 15, 2020 and from February 16, 2021 to March 15, 2021, respectively). Then, we compute the relative Wasserstein-2 distances $\epsilon_t = \frac{W_2(P_0,P_t)}{W_2(P_0,P_1)}$ and the relative Total Variation (TV) distances $\gamma_t = \frac{TV(P_0,P_t)}{TV(P_0,P_1)}$ for $t \in \{0, 1/11, \ldots, 10/11, 1\}$, as visualized in Figure 3(a). From these, we generate displacement interpolation $Q_t^{\text{dis}} = ((1-\epsilon_t)\text{Id} + \epsilon_t T)_{\#}P_0$, where T is the composition of the quantile function of P_1 and the cumulative distribution function of P_0; similarly, we generate linear interpolation $Q_t^{\text{lin}} = (1-\gamma_t)P_0 + \gamma_t P_1$. Essentially, Q_t^{dis} amounts to displacement interpolation between P_0 and P_1 that shifts from P_0 at the same rate as P_t under the Wasserstein-2 distance, namely, $W_2(P_0, Q_t^{\text{dis}}) = W_2(P_0, P_t)$; analogously, Q_t^{lin} corresponds to linear interpolation such that $TV(P_0, Q_t^{\text{lin}}) = TV(P_0, P_t)$. Figure 3(b) and (c) show Q_t^{dis} and Q_t^{lin}, respectively. Comparing Figure 3(b) and (c) with the real distribution shift in Figure 2(b), we can see that displacement interpolation provides a better approximation. Particularly, displacement interpolation preserves the unimodality of distributions as in Figure 2(b), while linear interpolation creates two



(a) Relative distances ϵ_t and γ_t



(b) Displacement interpolation Q_t^{dis} (c) Linear interpolation Q_t^{lin}

Figure 3. (a) plots the relative Wasserstein-2 distances $\epsilon_t = \frac{W_2(P_0,P_t)}{W_2(P_0,P_1)}$ and the relative TV distances $\gamma_t = \frac{TV(P_0,P_t)}{TV(P_0,P_1)}$ for $t \in \{0, 1/11, \ldots, 1\}$, with the 45 degree line shown as a dashed line; relative Wasserstein-1 distances $\frac{W_1(P_0,P_t)}{W_1(P_0,P_1)}$ are plotted for reference as well, which almost coincide with ϵ_t. (b) and (c) show displacement interpolation $Q_t^{\text{dis}} = ((1-\epsilon_t)\text{Id} + \epsilon_t T)_\# P_0$ and linear interpolation $Q_t^{\text{lin}} = (1-\gamma_t)P_0 + \gamma_t P_1$, respectively.

modes, in reminiscence of the Gaussian example in the previous section.

Together with theoretical foundations in Section 1.1, the above example provides the empirical underpinnings of the displacement interpolation model for distribution shifts. Returning to the detection problem (1.1), in what follows, we will consider a detection problem where Q in H_1 is replaced by displacement interpolation $((1-\epsilon)\text{Id}+\epsilon T)_\# P$ and propose a testing procedure based on the Wasserstein-2 distance. Later, we will apply the proposed testing procedure to revisit the above empirical example and analyze the power under the strong distribution shift during the recovery period.

1.3. Problem Description

Motivated by the discussion above, we study the problem of detecting weak distribution shifts represented as displacement interpolation: given two distributions P and Q on $\mathbb{R}$ whose cumulative distribution functions are F and G, respectively,

$$H_0 : X_1, \ldots, X_n \overset{\text{iid}}{\sim} P \quad \text{versus}$$
$$H_1 : X_1, \ldots, X_n \overset{\text{iid}}{\sim} ((1-\epsilon)\text{Id} + \epsilon G^{-1} \circ F)_\# P. \tag{1.3}$$

Here, F^{-1} and G^{-1} are the quantile functions of P and Q, respectively.

We propose a testing procedure based on the weighted Wasserstein distance between the empirical measure P_n—constructed by the observations $X_1, \ldots, X_n$—and the null distribution P. Assuming a weak signal in a sense $\epsilon = \epsilon_n \to 0$ as $n \to \infty$, we derive sharp conditions under which detection is possible: (a) when $n^{1/2}\epsilon_n \to 0$, detection is impossible; (b) when $n^{1/2}\epsilon_n \to \infty$, the testing procedure has asymptotic power 1; (c) at the detection boundary $n^{1/2}\epsilon_n \to$ constant $\in (0, \infty)$, sharp asymptotic Type I and Type II errors are analyzed using Gaussian processes.

1.4. Related Literature

Sparse mixture detection. A popular approach to formulating the weak signal detection problem is to replace the alternative hypothesis of (1.1) with Huber's ϵ-contamination model $(1-\epsilon)P + \epsilon Q$, also known as sparse mixtures, where $\epsilon = \epsilon_n \to 0$ as $n \to \infty$. Donoho and Jin (2004) proposes Tukey's higher criticism as a test statistic and analyzes the asymptotic phase transition depending on the rate $\epsilon = \epsilon_n \to 0$ and the signal strength. A recent endeavor extending the higher criticism test to compare two large frequency tables is given in Donoho and Kipnis (2022). See also Cai, Jessie Jeng, and Jin (2011) and Cai and Wu (2014) on optimal detection of general sparse mixtures.

Wasserstein distances for testing. Our testing procedure is based on the Wasserstein distance, which has been studied extensively in the testing literature. For example, Munk and Czado (1998) and Del Barrio, Giné, and Utzet (2005) study Goodness-of-Fit (GoF) testing using the Wasserstein distance between the null hypothesis and the empirical measure based on the observations; see Hallin, Mordant, and Segers (2021) for

an extension to the multivariate case. As mentioned earlier, the standard GoF testing is related to the classic detection problem (1.1), where the alternative distribution is fixed to some signal Q or a collection of distributions from a specific parametric family (de Wet 2002; Csörgő 2003).

Notation. For $a, b \in \mathbb{R}$, denote $a \wedge b = \min\{a, b\}$. For positive sequences $\{a_n\}$ and $\{b_n\}$, denote $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. We write $Y_n \rightsquigarrow Y$ if a sequence $\{Y_n\}$ of random variables converges weakly to some random variable Y.

2. Main Results

Testing procedure. We propose a distance-based test statistic for the testing problem (1.3), motivated by optimal transport and displacement interpolation. More specifically, we compare P_n—the empirical measure based on the observations $X_1, \ldots, X_n$—with P under the following distance, called the weighted Wasserstein distance, and reject H_0 if it is larger than a specific critical value.

Definition 1. Let ω be a finite Borel measure on $(0, 1)$. For two distributions μ and ν on $\mathbb{R}$, we define the ω-weighted Wasserstein distance by

$$W_{2,\omega}(\mu, \nu) := \left(\int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^2 \mathrm{d}\omega(u) \right)^{1/2},$$

where F_μ^{-1}, F_ν^{-1} denote the quantile functions of μ, ν, respectively.

Remark 1. Note in Definition 1 that $W_{2,\omega} = W_2$ if ω is the Lebesgue measure. Weighted versions of the Wasserstein distance are introduced in (de Wet 2002; Csörgő 2003). Here, we introduce the weighted base measure as in the Anderson-Darling test because, for certain detection problems, the signal may hide unevenly among quantiles. For example, the signal is contained in the extreme quantiles in higher criticism (Donoho and Jin 2004).

Asymptotic phase transition. We analyze the testing error in the asymptotic regime where $\epsilon = \epsilon_n$ vanishes as $n \to \infty$. More precisely, we rewrite (1.3) as follows specifying dependency on the sample size n:

$$H_0^{(n)} : X_1, \ldots, X_n \overset{\text{iid}}{\sim} P,$$
$$H_1^{(n)} : X_1, \ldots, X_n \overset{\text{iid}}{\sim} ((1 - \epsilon_n)\mathrm{Id} + \epsilon_n G^{-1} \circ F)_\# P,$$

where $\lim_{n\to\infty} \epsilon_n = 0$, namely, $\epsilon_n = o(1)$. Our main result shows that the asymptotic testing error behaves differently depending on the vanishing rate of ϵ_n, highlighting a phase transition phenomenon. In particular, we study a testing procedure such that the asymptotic Type I error is a given level $\alpha \in (0, 1)$ by deriving the limit of the test statistic $W_{2,\omega}(P_n, P)$ under the null hypothesis, based on the fundamental limit theorem of the empirical quantile process: assuming F admits a positive density f,

$$\sqrt{n}(F_n^{-1}(u) - F^{-1}(u)) \rightsquigarrow \frac{\mathbf{B}_u}{f(F^{-1}(u))}, \tag{2.1}$$

where F_n^{-1} is the empirical quantile function based on $X_1, \ldots, X_n$ from $H_0^{(n)}$ and $(\mathbf{B}_u)_{u\in[0,1]}$ is the standard Brownian bridge, namely, a mean-zero Gaussian process whose covariance satisfies $\mathbb{E}[\mathbf{B}_{u_1}\mathbf{B}_{u_2}] = u_1 \wedge u_2 - u_1 u_2$; rigorous asymptotic analysis is provided in the subsequent section. For such a testing procedure, we can characterize the asymptotic Type II error based on the three phases of ϵ_n determined by

$$\lim_{n\to\infty} n^{1/2}\epsilon_n = \begin{cases} 0, \\ \infty, \\ \gamma \in (0, \infty). \end{cases}$$

We formally state the main result as follows.

Theorem 1. Suppose F has a density f that is continuous and bounded away from 0 on some compact interval I_F and is 0 on $\mathbb{R}\backslash I_F$, G^{-1} is bounded on $(0, 1)$, and $G^{-1} \circ F$ is Lipschitz. Let ω be a finite Borel measure on $(0, 1)$ that is absolutely continuous with respect to the Lebesgue measure such that $W_{2,\omega}(P, Q) \neq 0$. Also, let Ψ be the cumulative distribution function of

$$\int_0^1 \left| \frac{\mathbf{B}_u}{f(F^{-1}(u))} \right|^2 \mathrm{d}\omega(u), \tag{2.2}$$

where $(\mathbf{B}_u)_{u\in[0,1]}$ is the standard Brownian bridge. Fix $\alpha \in (0, 1)$ and let C_α be the $(1-\alpha)$th quantile of Ψ, that is, $\Psi(C_\alpha) = 1-\alpha$. Consider the following testing procedure:

$$\text{reject } H_0^{(n)} \text{ if and only if } nW_{2,\omega}^2(P_n, P) > C_\alpha.$$

Then, the asymptotic Type I error is α, and the asymptotic Type II error is as follows.

(i) If $n^{1/2}\epsilon_n \to 0$, the asymptotic Type II error is $1 - \alpha$.
(ii) If $n^{1/2}\epsilon_n \to \infty$, the asymptotic Type II error is 0.
(iii) If $n^{1/2}\epsilon_n$ is a constant, say $n^{1/2}\epsilon_n = \gamma > 0$, the asymptotic Type II error is

$$\Psi_\gamma(C_\alpha - \gamma^2 W_{2,\omega}^2(P, Q)),$$

where Ψ_γ is the cumulative distribution function of

$$\int_0^1 \left| \frac{\mathbf{B}_u}{f(F^{-1}(u))} \right|^2 \mathrm{d}\omega(u) + 2\gamma \int_0^1 \frac{\mathbf{B}_u}{f(F^{-1}(u))} \cdot \left(G^{-1}(u) - F^{-1}(u)\right) \mathrm{d}\omega(u). \tag{2.3}$$

Remark 2. Theorem 1 implies that the asymptotic testing error, namely, the sum of the asymptotic Type I and II errors, is 1 (undetectable) if $n^{1/2}\epsilon_n \to 0$ and α (detectable) if $n^{1/2}\epsilon_n \to \infty$. The phase transition occurs at the boundary if $n^{1/2}\epsilon_n$ is a constant, where the testing error is determined by the constant $\gamma := n^{1/2}\epsilon_n$ and $\Delta := W_{2,\omega}(P, Q)$, which denotes the signal strength. The detection boundary (iii) still holds if we replace $n^{1/2}\epsilon_n = \gamma$ with $\lim_{n\to\infty} n^{1/2}\epsilon_n = \gamma > 0$; see the formal proof of Theorem 1 which is deferred to Appendix A. The main ideas of the analysis will be presented in the next section.

3. Asymptotic Analysis

In this section, we rigorously derive the asymptotic limit of the test statistic $W_{2,\omega}(P_n, P)$, under the null $H_0^{(n)}$ and the alternative $H_1^{(n)}$, respectively.

3.1. Preliminaries

Let $\ell^\infty(0,1)$ denote the set of all bounded functions defined on $(0,1)$, which is a Banach space under the uniform norm given by

$$\|h\|_\infty = \sup_{u\in(0,1)} |h(u)| \quad \forall h \in \ell^\infty(0,1).$$

We first provide the exact statement of (2.1) based on weak convergence in $\ell^\infty(0,1)$, see van der Vaart and Wellner (1996); in what follows, for any random element Z in $\ell^\infty(0,1)$, let $Z(u)$ denote the random variable at coordinate $u \in (0,1)$.

Assumption 1. F has a density f that is continuous and bounded away from 0 on some compact interval I_F and is 0 on $\mathbb{R}\backslash I_F$.

Lemma 1. Let F_n^{-1} be the empirical quantile function based on $X_1,\dots,X_n$ that are iid from a cumulative distribution function F. Under Assumption 1, view $(\sqrt{n}(F_n^{-1}-F^{-1}))_{n\in\mathbb{N}}$ as a sequence of random elements in $\ell^\infty(0,1)$, then it converges weakly to some tight measurable random element $\mathbf{H}$ in $\ell^\infty(0,1)$, which we denote as

$$\sqrt{n}(F_n^{-1} - F^{-1}) \rightsquigarrow \mathbf{H} \text{ in } \ell^\infty(0,1). \tag{3.1}$$

The limit $\mathbf{H}$ satisfies the following.

(i) $\{\mathbf{H}(u) : u \in (0,1)\}$ is a mean-zero Gaussian process with covariance function

$$\mathbb{E}[\mathbf{H}(u_1)\mathbf{H}(u_2)] = \frac{u_1 \wedge u_2 - u_1u_2}{f(F^{-1}(u_1))f(F^{-1}(u_2))} \quad \forall u_1,u_2 \in (0,1).$$

(ii) The sample path $u \mapsto \mathbf{H}(u)$ is continuous.

Remark 3. Lemma 1 is from Lemma 3.9.23 of (van der Vaart and Wellner 1996). Assumption 1 ensures that F^{-1} is bounded on $(0,1)$, thereby viewing $\sqrt{n}(F_n^{-1} - F^{-1})$ as a random element in $\ell^\infty(0,1)$. Though this assumption rules out distributions supported on the whole real line, such as normal distributions, we can still apply Assumption 1 to such distributions by restricting their support to a sufficiently large yet bounded interval. Alternatively, one may modify Lemma 1 by considering weak convergence in $\ell^\infty[\delta, 1-\delta]$ with a suitable $\delta > 0$. Such a modification provides limit theorems of the integration of $|F_n^{-1} - F^{-1}|^2$ on $[\delta, 1-\delta]$, often called the trimmed Wasserstein distance (Munk and Czado 1998). It is also possible to modify Lemma 1 by considering weak convergence in $L^2(0,1)$ under suitable assumptions on the behavior of F^{-1} near the endpoints 0 and 1 (Del Barrio, Giné, and Utzet 2005).

Remark 4. In Lemma 1, the limit $\mathbf{H}$ is tight, meaning that for any $\epsilon > 0$, we can find a compact subset K of $\ell^\infty(0,1)$ such that $\mathbb{P}(\mathbf{H} \in K) \geq 1-\epsilon$. Though this technicality is not used explicitly in the main analysis, it is required to apply the extended continuous mapping theorem, which we adapt from Theorem 1.11.1 of van der Vaart and Wellner (1996) and rewrite as Theorem A.1.

Next, we consider the following integrated processes: letting $\mathbf{H}_n := \sqrt{n}(F_n^{-1} - F^{-1})$, define

$$A_n := \int_0^1 |\mathbf{H}_n(u)|^2 \mathrm{d}\omega(u), \tag{3.2}$$

$$B_n := \int_0^1 |(G^{-1}\circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - F^{-1}(u)|^2 \mathrm{d}\omega(u), \tag{3.3}$$

$$C_n := \int_0^1 \mathbf{H}_n(u)\cdot\big((G^{-1}\circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - F^{-1}(u)\big)\,\mathrm{d}\omega(u). \tag{3.4}$$

Under the assumptions of Lemma 1, the limit of A_n is essentially the limit of the test statistic $nW_{2,\omega}^2(P_n,P)$ under the null. Later, we will use B_n and C_n to derive the limit of the test statistic under the alternatives. The following lemma derives the limits of the above processes, which we prove in Appendix A.

Lemma 2. Let F_n^{-1} be the empirical quantile function based on $X_1,\dots,X_n$ that are iid from a cumulative distribution function F. Let ω be a finite Borel measure on $(0,1)$ that is absolutely continuous with respect to the Lebesgue measure. Under Assumption 1 and assuming G^{-1} is bounded on $(0,1)$, let $\mathbf{H}_n = \sqrt{n}(F_n^{-1} - F^{-1})$ and $\mathbf{H}$ be the random element mentioned in Lemma 1, then

$$A_n \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \mathrm{d}\omega(u), \tag{3.5}$$

$$B_n \rightsquigarrow \int_0^1 |G^{-1}(u) - F^{-1}(u)|^2 \mathrm{d}\omega(u), \tag{3.6}$$

$$C_n \rightsquigarrow \int_0^1 \mathbf{H}(u)\cdot(G^{-1}(u) - F^{-1}(u))\mathrm{d}\omega(u), \tag{3.7}$$

where A_n, B_n, C_n are as in (3.2), (3.3), (3.4), respectively.

3.2. Asymptotic Distributions

Now, we analyze the limit of the test statistic $W_{2,\omega}(P_n,P)$ under $H_0^{(n)}$. By (3.5) of Lemma 2, the following holds.

Proposition 1 (Limit under the null). For each $n \in \mathbb{N}$, let P_n be the empirical measure based on $X_1,\dots,X_n$ following $H_0^{(n)}$. Let ω be a finite Borel measure on $(0,1)$ that is absolutely continuous with respect to the Lebesgue measure. Under Assumption 1,

$$nW_{2,\omega}^2(P_n,P) \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \mathrm{d}\omega(u),$$

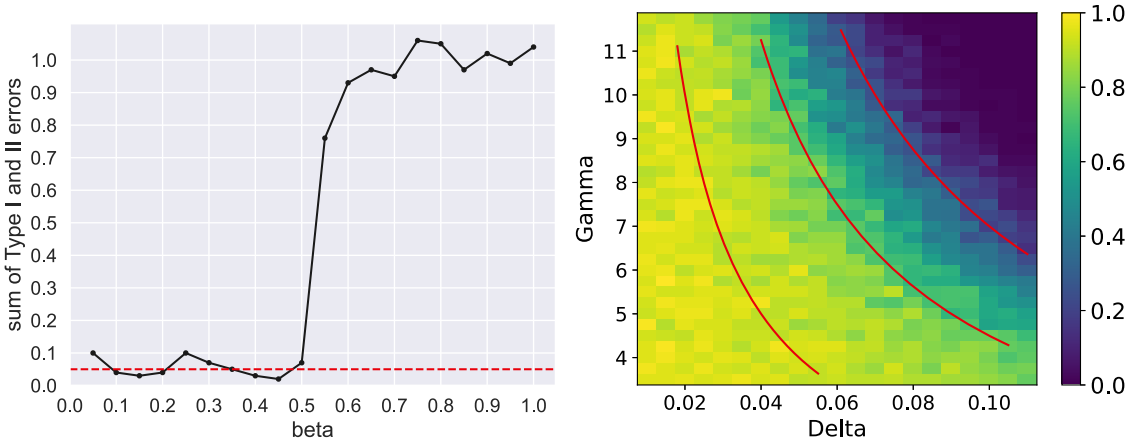
where $\mathbf{H}$ is the random element mentioned in Lemma 1.

Next, we analyze the limit of the test statistic under $H_1^{(n)}$.

Theorem 2 (Limit under the alternatives). For each $n \in \mathbb{N}$, let P_n be the empirical measure based on $X_1,\dots,X_n$ following $H_1^{(n)}$. Let ω be a finite Borel measure on $(0,1)$ that is absolutely continuous with respect to the Lebesgue measure. Under Assumption 1 and assuming G^{-1} is bounded on $(0,1)$ and $G^{-1}\circ F$ is Lipschitz, we can characterize the limit of $nW_{2,\omega}^2(P_n,P)$ as follows.

(i) If $n^{1/2}\epsilon_n \to 0$,

$$nW_{2,\omega}^2(P_n,P) \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \mathrm{d}\omega(u). \tag{3.8}$$



(a) Sum of Type I and Type II errors (b) Type II errors

Figure 4. (a) visualizes the sum of Type I and Type II errors of Experiment 1; the red dashed line represents the level $\alpha = 0.05$. (b) plots Type II errors of Experiment 2 as a color map, where the red solid curves represent $\gamma \cdot \Delta =$ constant $\in \{0.2, 0.45, 0.7\}$

(ii) If $n^{1/2}\epsilon_n \to \infty$,

$$n^{1/2}\epsilon_n^{-1}\left(W_{2,\omega}^2(P_n, P) - \epsilon_n^2 W_{2,\omega}^2(P, Q)\right) \rightsquigarrow 2\int_0^1 \mathbf{H}(u)\cdot\left(G^{-1}(u) - F^{-1}(u)\right)\mathrm{d}\omega(u). \quad (3.9)$$

(iii) If $n^{1/2}\epsilon_n$ is a constant, say $n^{1/2}\epsilon_n = \gamma > 0$,

$$nW_{2,\omega}^2(P_n, P) - \gamma^2 W_{2,\omega}^2(P, Q) \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \mathrm{d}\omega(u) + 2\gamma\int_0^1 \mathbf{H}(u) \cdot \left(G^{-1}(u) - F^{-1}(u)\right)\mathrm{d}\omega(u). \quad (3.10)$$

Remark 5. One can also show that the limit distribution in (3.9) is a mean-zero Gaussian distribution with variance

$$\int_0^1\int_0^1 \frac{4(u_1 \wedge u_2 - u_1u_2)}{f(F^{-1}(u_1))f(F^{-1}(u_2))}\cdot\left(G^{-1}(u_1) - F^{-1}(u_1)\right) \cdot\left(G^{-1}(u_2) - F^{-1}(u_2)\right)\mathrm{d}\omega(u_1)\mathrm{d}\omega(u_2).$$

Remark 6. The Lipschitzness assumption on $G^{-1} \circ F$ is used for (3.9), but not for (3.8) and (3.10). Such an assumption is well studied in the literature; see Appendix of Bobkov and Ledoux (2019).

4. Simulations

4.1. Phase Transition and Power Analysis

This section delivers results using two experiments to numerically verify Theorem 1. For both experiments, we fix $P =$ Unif$[0, 1]$ and let ω be the Lebesgue measure so that Ψ is the cumulative distribution function of

$$\int_0^1 |\mathbf{B}_u|^2\mathrm{d}u.$$

Such a distribution is well studied, and the quantile of Ψ is available, see Table 1 of Anderson and Darling (1952); we choose $\alpha = 0.05$, then $C_\alpha \approx 0.46136$.

Experiment 1: phase transition. In this experiment, we verify that the phase transition occurs at $\beta = 0.5$ for the scaling $\epsilon_n = n^{-\beta}$. To this end, we fix $Q = N(0, 1)$ and let $\epsilon_n = n^{-\beta}$ with $n = 10^6$, and a range of $\beta \in \{0.05, 0.1, \ldots, 0.95, 1\}$. For each β, we compute $nW_2^2(P_n, P)$ using samples from $H_0^{(n)}$ and $H_1^{(n)}$, which we repeat 100 times, then compute the Type I and Type II errors using those 100 realizations, respectively. We plot the sum of Type I and Type II errors against β. Figure 4(a) confirms that the phase transition occurs at $\beta = 0.5$.

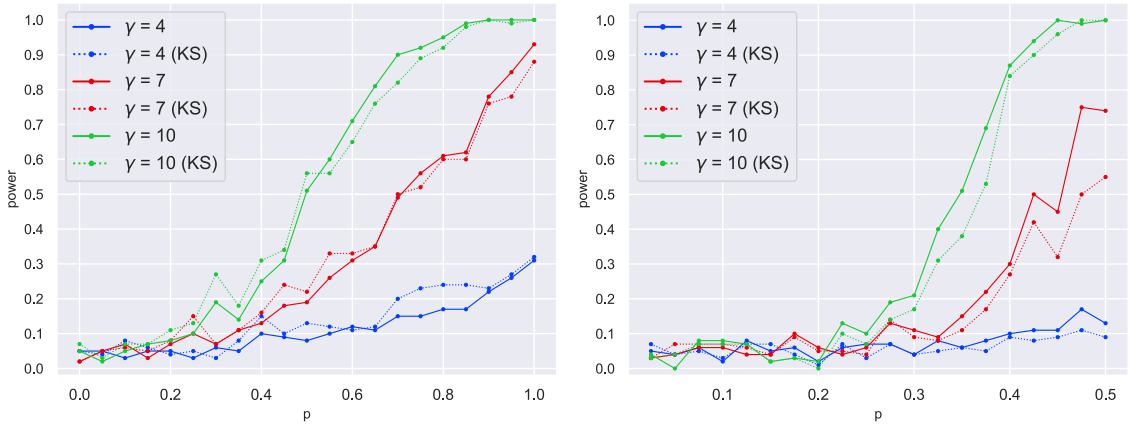
Experiment 2: sharp power analysis. We focus on the power behavior at the detection boundary, namely, $\epsilon_n = \gamma n^{-0.5}$ with $n = 10^6$. Specifically, we analyze the Type II error based on two parameters representing the signal strength: γ and $\Delta := W_2(P, Q)$. By Theorem 1, the Type II error should be approximately $\Psi_\gamma(C_\alpha - \gamma^2\Delta^2)$. To vary Δ, we parameterized Q as follows: the quantile function of Q is $u \mapsto u + \frac{p}{2\pi}\sin(2\pi u)$, where we vary $p \in (0, 1)$; note that this is a valid quantile function as it is monotonically increasing. Then,

$$W_2^2(P, Q) = \frac{p^2}{4\pi^2}\int_0^1 |\sin(2\pi u)|^2\,\mathrm{d}u = \frac{p^2}{8\pi^2} = \Delta^2,$$

which results in $\Delta^2 \in (0, (8\pi^2)^{-1})$. We can easily generate Q's with desired signal strength Δ using this parameterization. For each pair (Δ, γ), where $\Delta \in \{0.01, 0.015, \ldots, 0.105, 0.11\}$ and $\gamma \in \{3.5, 3.75, \ldots, 11.5, 11.75\}$, we compute the Type II error and visualize it as a color map. Figure 4(b) visualizes Type II errors on the Δ-γ plane using colors; this essentially plots $\Psi_\gamma(C_\alpha - \gamma^2\Delta^2)$. Notice that the level set of Type II errors and the curve $\gamma \cdot \Delta =$ constant do not perfectly coincide because Type II errors can vary on the curve $\gamma \cdot \Delta =$ constant, namely, $\Psi_\gamma(C_\alpha -$ constant$^2)$ can change as γ varies.

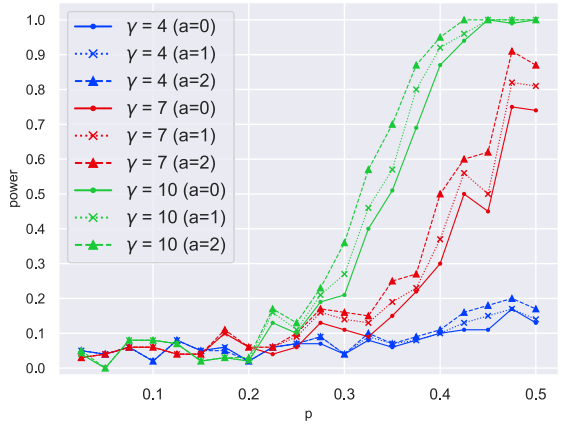
4.2. Comparison to Other Methods and Robustness Checks

This section first compares the power of the proposed procedure and the Kolmogorov-Smirnov (KS) test. Later, we carry out preliminary robustness checks on how the choice of the weight



(a) Powers (Example I)

(b) Powers (Example II)



(c) Powers (Example II) of different weight measures

Figure 5. (a) and (b) show the powers of the proposed testing procedure and the KS test in Example I and Example II, respectively; solid lines correspond to the proposed testing procedure, whereas dotted lines represent the KS test. (c) shows the powers of the proposed testing procedures in the setting of Example II, using the weight measure given by (4.2). By design, the solid lines of (b) and (c) coincide; both correspond to the Lebesgue measure, namely, $a = 0$ in (4.2).

measure ω affects the power. Throughout this section, we again fix $P = \text{Unif}[0,1]$ and $n = 10^6$.

Example I. First, we repeat the previous setting used for the boundary case: $\epsilon_n = \gamma n^{-0.5}$ and the quantile function of Q is $u \mapsto u + \frac{p}{2\pi}\sin(2\pi u)$. This time, we parameterize $p \in \{0, 0.05, \ldots, 0.95, 1\}$ instead of $\Delta = W_2(P,Q)$ and restrict our interest to $\gamma \in \{4, 7, 10\}$. For each pair (p,γ), we compute the power of the proposed testing procedure—with ω being the Lebesgue measure—and the KS test; to ensure both are asymptotically level α tests with $\alpha = 0.05$, we use the critical value $C_\alpha \approx 0.46136$ for the proposed testing procedure as before and the critical value 1.36 for the KS test statistic $\sqrt{n} \cdot \text{KS}(P_n, P) = \sqrt{n} \cdot \sup_{x\in\mathbb{R}} |F_n(x) - F(x)|$ based on Table 1 of (Shorack and Wellner 2009). Figure 5(a) shows the results. First, observe that for $\gamma \in \{7, 10\}$, the powers of both tests exceed 0.5 once the parameter p is above certain values, which means that both are better than a trivial test that randomly rejects the null with probability 0.5. In this case, the power of the proposed procedure is slightly larger than that of the KS test, as the solid lines are above the dotted lines. For $\gamma = 4$, the power of the KS test is mostly larger than that of the proposed procedure, as shown in the blue lines; in this case, however, the powers of both tests are far less than 0.5, implying that both tests practically failed. In other words, the signal strength $\gamma = 4$ is too weak for both tests to detect.

Example II. Next, we consider a different setting where the quantile function of Q differs with that of $P = \text{Unif}[0,1]$ at tails. To this end, suppose the quantile function of Q is as follows:

$$\begin{cases} u + 0.45 \cdot \frac{2p}{\pi} \cos\left(\frac{\pi u}{2p}\right) & 0 \le u \le p, \\ u & p \le u \le 1-p, \\ u - 0.45 \cdot \frac{2p}{\pi} \cos\left(\frac{\pi(1-u)}{2p}\right) & 1-p \le u \le 1, \end{cases} \tag{4.1}$$

where we parameterize $p \in \{0.025, 0.05, \ldots, 0.5\}$. Note that the quantile function of Q deviates from the identity only at tails, namely, $[0,p] \cup [1-p, 1]$. As in Example I, we keep using $\epsilon_n = \gamma n^{-0.5}$ for $\gamma \in \{4, 7, 10\}$ and compute the powers for each pair (p,γ) which are shown in Figure 5(b). For $\gamma \in \{7, 10\}$, the results are similar to Example I; for sufficiently large p, both tests have enough powers. Now, for $\gamma = 4$, the power of the proposed procedure is slightly larger than that of the KS test; however, as

in Example I, both tests are practically useless as their powers are too small; namely, $\gamma = 4$ is again too weak for them to detect.

The role of the weight measure. Lastly, we compare the powers by differing the weight measure ω. Let us keep considering the setting of Example II. As the quantile function Q mainly deviates from that of P at tails, it is reasonable to put more weights at tails for better performance when computing the test statistic. To verify this, consider a simple weight measure whose density with respect to the Lebesgue measure is a quadratic function: for $a \geq 0$,

$$\frac{\mathrm{d}\omega(u)}{\mathrm{d}u} = a\left(u - \frac{1}{2}\right)^2 + 1 - \frac{a}{12}, \tag{4.2}$$

which satisfies $\int_0^1 \mathrm{d}\omega = 1$. Clearly, $a = 0$ corresponds to the Lebesgue measure; large a means more weights at both tails. We compute the powers of the proposed procedure for $a \in \{0, 1, 2\}$. Here, we again set critical values to ensure that all of them are asymptotically level $\alpha = 0.05$; for $a = 0$, we can reuse the previous critical value as before. For $a \in \{1, 2\}$, we estimate the quantile of the asymptotic distribution $\int_0^1 |\mathbf{B}_u|^2 \mathrm{d}\omega(u)$ by Monte Carlo simulations. Figure 5(c) shows the results. As expected, we obtain larger powers with the quadratic weight measure, namely, both $a = 1$ and $a = 2$ achieve better performance than the unweighted procedure $a = 0$; particularly, assigning more weights at tails ($a = 2$) yields the largest power. For $\gamma = 4$, though we have increased powers for the weighted procedures, they are still practically too weak to detect the signal, as discussed in the previous examples.

4.3. Application I: Distribution Shifts in Consumer Spending

We revisit the data example in Section 1.2 and apply the proposed testing procedure to study the power. Recall that $\{P_{i/11}\}_{i=0}^{11}$ are the distributions of monthly average spending by county during the recovery period between March 16, 2020 and March 15, 2021. For $t \in \{0, 1/11, \ldots, 10/11, 1\}$ and $n \in \{10, 50, 100, 500\}$, we construct the empirical measure P_t^n using n points that are iid from P_t and compute $W_2(P_0, P_t^n)$. Essentially, for each t and sample size n, we want to distinguish P_t from P_0 using a finite sample. To this end, we first estimate the quantile of $W_2(P_0, P_0^n)$ via Monte Carlo simulations to define a level $\alpha = 0.05$ test; this will give us a critical value, say, C_α^n. Then, for each $t > 0$, we compute $W_2(P_0, P_t^n)$ and reject the null—that is, data are from P_0—if it exceeds C_α^n; by repeating this for 100 times, we can evaluate the power of the test using simulations. The results are shown in the first 4 rows (Recovery Period) of Table 1. In this case, for $t \geq 3/11$, namely, after three months from March, 2020, we can detect the distribution shift from P_0 for any sample size n. In other words, we can tell there has been a significant recovery after three months; indeed, the shift is significant enough so we can tell the difference from P_0 by estimating P_t using only 10 randomly chosen counties instead of the total 1655 counties. The first month after P_0, that is, for $t = 1/11$, the shift is relatively weaker compared to the subsequent months, so $n = 10$ yields power 0.56, which is not enough to detect the shift; in other words, for $t = 1/11$, we need at least $n = 50$ to distinguish it from P_0.

Table 1. Powers for the recovery and stable periods.

	$11t$											
n	1	2	3	4	5	6	7	8	9	10	11	
10	0.56	0.98	1	1	1	1	1	1	1	1	1	Recovery period
50	0.98	1	1	1	1	1	1	1	1	1	1	
100	1	1	1	1	1	1	1	1	1	1	1	
500	1	1	1	1	1	1	1	1	1	1	1	
10	0.04	0.05	0.09	0.1	0.05	0.07	0.1	0.05	0.07	0.11	0.23	Stable period
50	0.2	0.23	0.36	0.38	0.22	0.18	0.1	0.14	0.09	0.4	0.75	
100	0.57	0.45	0.66	0.66	0.45	0.29	0.43	0.33	0.16	0.76	0.96	
500	1	0.99	1	1	1	0.94	1	0.99	0.88	1	1	

How about the power for the stable period? We repeat the above procedure by taking $\{P_{i/11}\}_{i=0}^{11}$ as the distributions during the stable period, namely, they amount to the histograms in Figure 2(c); again, P_0 and P_1 correspond to the start and end of this period (from March 16, 2021 to April 15, 2021 and from February 16, 2022 to March 15, 2022). Recall from Figure 2(c) that we no longer see a significant distribution shift as in the recovery period. The resulting powers are shown in the last four rows (Stable Period) of Table 1. Unlike the recovery period, we can observe the powers are much smaller for $n \leq 100$; particularly, most of the powers—except for a few periods with $n = 100$—are far below 0.5, meaning that we cannot detect a meaningful shift from P_0 using finite samples. For detection to be possible, we can see that we need a sufficiently large sample size $n = 500$.

In summary, the results in Table 1 essentially demonstrate that the proposed testing procedure can detect the distribution shifts during the recovery period as it is reasonably approximated by displacement interpolation as seen in Figure 2(b), while there are no such shifts to be captured by the proposed procedure during the stable period because the distributions are similar to each other as shown in Figure 2(c).

4.4. Application II: p-value Heterogeneity across Disciplines

We apply our testing procedure to the data set from Head et al. (2015), which collects p-values of statistical tests published in journals across different disciplines. In this example, we use our procedure to determine if the distribution of p-values differs depending on disciplines.

To this end, we first collect p-values across nine disciplines subsampled from total 21 disciplines, then take 80% of them as the null distribution P_{Null}, which consists of 90,950 observations; this is shown as the solid blue curve in Figure 6. Then, from the remaining 20%, we sample two disciplines "medical and health sciences" (P_{MH}) and "multidisciplinary" (P_{Mu}), which consist of 12,928 and 5731 observations, respectively.[3] Figure 6 shows the cumulative distribution functions of the null P_{Null}, P_{MH}, and P_{Mu}, which are close to each other. This suggests that both P_{MH} and P_{Mu} are weak signals that are hard to distinguish from P_{Null} visually. To tackle this problem, we apply our testing procedure to $H_0 : P_{\text{MH}} = P_{\text{Null}}$. First, we compute the test statistic $T_{\text{MH}} := nW_2^2(P_{\text{MH}}, P_{\text{Null}})$, where $n := |P_{\text{MH}}| =$ 12,928 observations. Then, setting $\alpha = 0.05$, we estimate the

[3]They have the largest number of observations among all nine disciplines.

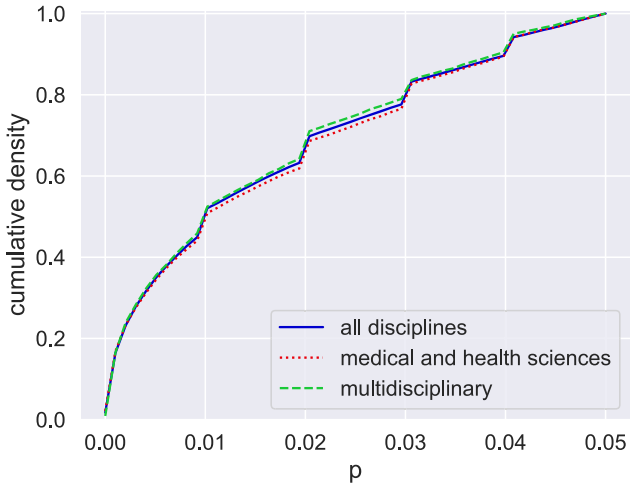


Figure 6. This figure plots the cumulative distribution functions of p-values (below 0.05) across nine disciplines subsampled from (Head et al. 2015), which is shown as the solid blue curve, along with two disciplines: "medical and health sciences" (red, dotted) and "multidisciplinary" (green, dashed).

$(1-\alpha)$th quantile of $nW_2^2(P_n, P_{\text{Null}})$, where P_n is the empirical measure based on $X_1, \dots, X_n$ that are iid from P_{Null}, which can be done by resampling P_n. We obtain $T_{\text{MH}} = 0.004821$ which is larger than the estimated quantile 0.002918; also, we can estimate the p-value of T_{MH}, which is $0.005 \ll \alpha$, suggesting that we reject $H_0 : P_{\text{MH}} = P_{\text{Null}}$. We apply the same procedure to P_{Mu} with $n := |P_{\text{Mu}}| = 5{,}731$ observations. We obtain $T_{\text{Mu}} = 0.002334$, which is slightly below the estimated quantile 0.002412, and the p-value of T_{Mu} is $0.066 > \alpha$, implying that we cannot $H_0 : P_{\text{Mu}} = P_{\text{Null}}$ at level α. Therefore, our testing procedure provides a rigorous framework to detect weak signals that are otherwise indistinguishable, as visualized in Figure 6.

5. Discussion

Comparison with existing methods. Notice that the problem (1.3) can be viewed as an instance of general testing

$$H_0 : X_1, \dots, X_n \overset{\text{iid}}{\sim} P \quad \text{versus} \quad H_1 : X_1, \dots, X_n \overset{\text{iid}}{\sim} Q_n,$$

where $Q_n \to P$ as $n \to \infty$; our problem is the case where Q_n is given as displacement interpolation between P and Q. Existing approaches to such a general problem characterize detection by using the following (related) notions: likelihood ratio $\frac{\mathrm{d}Q_n}{\mathrm{d}P}$ (sec. 13.10.1 of van der Vaart and Wellner 1996) or Hellinger distance $H^2(P, Q_n)$ (sec. 13.1 of Lehmann and Romano 2005). The former approach—often referred to as Le Cam's third lemma—derives limit distributions based on asymptotic normality; the latter characterizes the detection boundary depending on the rate of $nH^2(P, Q_n)$. The first principle behind these methods is clear: quantify a distance/discrepancy between P, Q_n and characterize the limit. Our results also use such a first principle: we use weighted Wasserstein distances. Though our method and the existing methods share a similar high-level idea, the existing methods are unsuitable for our problem for several reasons. First, though Le Cam's third lemma applies to any abstract setting under certain conditions, it does not lead to the exact characterization of testing errors unless there is a suitable parametric assumption. Second, the Hellinger distance is not preferred in analyzing the case where Q_n is given as displacement interpolation as its relationship with the interpolation parameter ϵ_n is not transparent. Moreover, the Hellinger distance-based characterization generally does not calculate testing errors at the detection boundary. On the other hand, our method motivated by weighted Wasserstein distances not only interplays well with displacement interpolation but also provides the exact characterization of testing errors, including the boundary case; moreover, unlike the likelihood ratio test, which requires information on both P and Q_n, our test is implementable as long as P is known.

Lifting technical assumptions. As remarked in Remark 6, the Lipschitzness assumption on $G^{-1} \circ F$ is used in the detectable case ($n^{1/2}\epsilon_n \to \infty$), but not in the other two cases. Removing the Lipschitzness assumption, there is no restriction on Q as long as its quantile function G^{-1} is bounded. Accordingly, our main results hold even when Q is discrete; in particular, there are cases where our method applies, but Le Cam's third lemma cannot because contiguity is not satisfied. As a concrete example, suppose $P = \text{Unif}[0,1]$ and $Q = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$. Then, $Q_n := ((1-\epsilon_n)\text{Id} + \epsilon_n G^{-1} \circ F)_{\#}P$ is a uniform measure supported on the union of two disjoint intervals $U_n := [0, (1-\epsilon_n)/2] \cup [(1+\epsilon_n)/2, 1]$. One can verify that $Q_n^{\otimes n}$—the n tensor product of Q_n—is not contiguous with respect to $P^{\otimes n}$ because $Q_n^{\otimes n}(U_n^n) = 1$ while $P^{\otimes n}(U_n^n) = (1-\epsilon_n)^n \to 0$ at the boundary $\epsilon_n \asymp n^{-1/2}$. Hence, Le Cam's third lemma cannot be applied to derive the limit distribution for such a case, whereas our method can. Lastly, we briefly discuss Assumption 1. Though it is a reasonable assumption to impose from a practical viewpoint as discussed in Remark 3, it is natural to ask—from a technical viewpoint—whether such an assumption can be relaxed. To circumvent Assumption 1, we can use an alternative pivotal test based on the weighted Wasserstein between Unif[0, 1] and the empirical measure based on $F(X_1), \dots, F(X_n)$; then, we can show that similar theory holds.

Optimal testing procedure. Displacement interpolation motivates the weighted Wasserstein distance as a natural test statistic for (1.3). While the main theory of this article focuses on the asymptotic power of the testing procedure based on the weighted Wasserstein distance, we believe there are several interesting theoretical questions to be addressed in future work. Particularly, one may ask whether the proposed procedure is optimal for testing (1.3). The simulation results in Section 4.2 suggest that optimality would depend on P and Q. Hence, it would be interesting to derive a minimax optimal procedure that minimizes the worst case testing error over some class of distributions P, Q, which we leave as important future work.

Extension to two-sample testing. The problem (1.3) itself and the proposed procedure essentially postulate that P is known. In the empirical applications presented in Section 4, we treated finite data points as the null distribution P and applied the proposed procedure. Though the purpose of such treatment was to demonstrate the applicability of the proposed procedure in simplified settings, the fact that the finite data points, say,

$Y_1, \ldots, Y_m$, are usually noisy samples from some distribution P of interest is a crucial limitation in practice. To address this issue, one may consider a two-sample testing framework asking if $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are from the same distribution. We can still consider a local alternative given as displacement interpolation, say, $X_1, \ldots, X_n \overset{\text{iid}}{\sim} ((1-\epsilon)\mathrm{Id} + \epsilon G^{-1} \circ F)_{\#}P$ and $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} P$. The weighted Wasserstein distance between the empirical measures of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ is deemed a plausible test statistic under this framework. We leave the analysis of such a two-sample testing framework as future work.

Appendix A: Proofs

A.1. Proof of Lemma 2

Proof. Recall from Lemma 1 that $\mathbf{H}_n \rightsquigarrow \mathbf{H}$ in $\ell^\infty(0,1)$. Now, define

$$\ell_m^\infty(0,1) = \{h \in \ell^\infty(0,1) : h \text{ is measurable}\},$$

then one can verify that $\ell_m^\infty(0,1)$ is a closed subset of the Banach space $(\ell^\infty(0,1), \|\cdot\|_\infty)$. Also, as the sample path of $\mathbf{H}_n$ is always monotone and the sample path of $\mathbf{H}$ is always continuous, $\mathbf{H}_n$ and $\mathbf{H}$ take values in $\ell_m^\infty(0,1)$. Now, define a map $\mathcal{I}\colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\mathcal{I}(h) = \int_0^1 |h(u)|^2 \mathrm{d}\omega(u).$$

We claim that $\mathcal{I}$ is continuous. To this end, consider a sequence $(h_n)_{n\in\mathbb{N}}$ in $\ell_m^\infty(0,1)$ and $h \in \ell_m^\infty(0,1)$ such that $\|h_n - h\|_\infty \to 0$. Then, $\sup_{n\in\mathbb{N}} \|h_n\|_\infty \le M$ for some $M > 0$, hence, the dominated convergence theorem shows that $\mathcal{I}(h_n) \to \mathcal{I}(h)$. Now, applying Theorem A.1 (stated below), we conclude that $\mathcal{I}(\mathbf{H}_n) \rightsquigarrow \mathcal{I}(\mathbf{H})$, which proves (3.5). Similarly, for each $n \in \mathbb{N}$, define a map $\mathcal{E}_n\colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\mathcal{E}_n(h) = \int_0^1 |(G^{-1} \circ F)(F^{-1}(u) + n^{-1/2}h(u)) - F^{-1}(u)|^2 \mathrm{d}\omega(u).$$

For any sequence $(h_n)_{n\in\mathbb{N}}$ in $\ell_m^\infty(0,1)$ and $h \in \ell_m^\infty(0,1)$ such that $\|h_n - h\|_\infty \to 0$, we claim

$$\mathcal{E}_n(h_n) \to \int_0^1 |G^{-1}(u) - F^{-1}(u)|^2 \mathrm{d}\omega(u). \tag{A.1}$$

As G^{-1} is continuous almost everywhere on $(0,1)$, we can see that $G^{-1} \circ F \circ (F^{-1} + n^{-1/2}h_n)$ converges to G^{-1} almost everywhere; as ω is absolutely continuous with respect to the Lebesgue measure, this convergence holds ω-almost everywhere as well. As G^{-1} and F^{-1} are bounded on $(0,1)$, the dominated convergence theorem shows (A.1). By Theorem A.1,

$$\mathcal{E}_n(\mathbf{H}_n) \rightsquigarrow \int_0^1 |G^{-1}(u) - F^{-1}(u)|^2 \mathrm{d}\omega(u),$$

showing (3.6). Lastly, to prove (3.7), for each $n \in \mathbb{N}$, define a map $\mathcal{J}_n\colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\mathcal{J}_n(h) = \int_0^1 h(u) \cdot \Big((G^{-1} \circ F)(F^{-1}(u) + n^{-1/2}h(u)) - F^{-1}(u)\Big) \mathrm{d}\omega(u).$$

Also, define $\mathcal{J}\colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\mathcal{J}(h) = \int_0^1 h(u) \cdot \Big(G^{-1}(u) - F^{-1}(u)\Big) \mathrm{d}\omega(u).$$

Similarly to $\mathcal{I}$, one can verify that $\mathcal{J}$ is continuous. Also, for any sequence $(h_n)_{n\in\mathbb{N}}$ in $\ell_m^\infty(0,1)$ and $h \in \ell_m^\infty(0,1)$ such that $\|h_n - h\|_\infty \to 0$, we can show that $\mathcal{J}_n(h_n) \to \mathcal{J}(h)$ by means of the dominated convergence theorem. Therefore, $\mathcal{J}_n(\mathbf{H}_n) \rightsquigarrow \mathcal{J}(\mathbf{H})$ holds by Theorem A.1, hence, (3.7) holds. □

Theorem A.1 (Extended Continuous Mapping Theorem). Let $\mathcal{H}$ and $\mathcal{K}$ be metric spaces. Let $\mathcal{H}_0$ be a Borel subset of $\mathcal{H}$ and consider a measurable map $\mathcal{I}\colon \mathcal{H}_0 \to \mathcal{K}$. Suppose there is a sequence $(\mathcal{I}_n)_{n\in\mathbb{N}}$ of maps from $\mathcal{H}_0$ to $\mathcal{K}$ satisfying the following: for any sequence $(h_n)_{n\in\mathbb{N}}$ in $\mathcal{H}_0$ converging to some $h \in \mathcal{H}_0$, the sequence $(\mathcal{I}_n(h_n))_{n\in\mathbb{N}}$ converges to $\mathcal{I}(h)$ in $\mathcal{K}$. If a sequence $(H_n)_{n\in\mathbb{N}}$ of random elements in $\mathcal{H}$ converges weakly to a tight measurable random element H in $\mathcal{H}$, where both H_n and H take values in $\mathcal{H}_0$, the sequence $(\mathcal{I}_n(H_n))_{n\in\mathbb{N}}$ of random elements in $\mathcal{K}$ converges weakly to the random element $\mathcal{I}(H)$ in $\mathcal{K}$.

Remark A.1. Theorem A.1 is adapted from Theorem 1.11.1 of van der Vaart and Wellner (1996), where the latter uses separability instead of tightness. As tightness implies separability, we have stated Theorem A.1 with tightness, which is sufficient in our analysis.

A.2. Proof of Theorem 2

Proof. For each $n \in \mathbb{N}$, let $\phi_n = (1-\epsilon_n)\mathrm{Id} + \epsilon_n G^{-1} \circ F$. As we are concerned with the limit distribution of $nW_{2,\omega}^2(P_n, P)$, we may assume $(X_n)_{n\in\mathbb{N}}$ is iid from P and let P_n be the empirical measure based on $\phi_n(X_1), \ldots, \phi_n(X_n)$ as $\phi_n(X_1), \ldots, \phi_n(X_n)$ also follow $H_1^{(n)}$. Now, let F_n^{-1} be the empirical quantile function based on $X_1, \ldots, X_n$, then

$$\begin{aligned}
&\phi_n \circ F_n^{-1} - F^{-1} \\
&\quad = (1-\epsilon_n)F_n^{-1} + \epsilon_n G^{-1} \circ F \circ F_n^{-1} - F^{-1} \\
&\quad = (1-\epsilon_n)n^{-1/2}\mathbf{H}_n + \epsilon_n(G^{-1} \circ F \circ (F^{-1} + n^{-1/2}\mathbf{H}_n) - F^{-1}),
\end{aligned}$$

where $\mathbf{H}_n := \sqrt{n}(F_n^{-1} - F^{-1})$. Hence, using A_n, B_n, C_n defined in (3.2), (3.3), (3.4), respectively, we have

$$W_{2,\omega}^2(P_n, P) = (1-\epsilon_n)^2 n^{-1} A_n + \epsilon_n^2 B_n + 2(1-\epsilon_n)n^{-1/2}\epsilon_n C_n.$$

Case I: Suppose $n^{1/2}\epsilon_n \to 0$. Recall that we have shown in Lemma 2 that A_n, B_n, and C_n are weakly convergent. Note that

$$nW_{2,\omega}^2(P_n, P) = (1-\epsilon_n)^2 A_n + (n^{1/2}\epsilon_n)^2 B_n + 2(1-\epsilon_n)(n^{1/2}\epsilon_n)C_n.$$

As $(n^{1/2}\epsilon_n)^2 B_n, (n^{1/2}\epsilon_n)C_n \rightsquigarrow 0$, by Slutsky's theorem,

$$nW_{2,\omega}^2(P_n, P) \rightsquigarrow \lim_{n\to\infty} A_n = \int_0^1 |\mathbf{H}(u)|^2 \mathrm{d}\omega(u).$$

Case II: Suppose $n^{1/2}\epsilon_n \to \infty$ and note that

$$\begin{aligned}
&n^{1/2}\epsilon_n^{-1}\left(W_{2,\omega}^2(P_n, P) - \epsilon_n^2 W_{2,\omega}^2(P, Q)\right) \\
&\quad = (1-\epsilon_n)^2 (n^{1/2}\epsilon_n)^{-1} A_n \\
&\qquad + n^{1/2}\epsilon_n(B_n - W_{2,\omega}^2(P, Q)) + 2(1-\epsilon_n)C_n.
\end{aligned}$$

First, notice that $(n^{1/2}\epsilon_n)^{-1}A_n \rightsquigarrow 0$. We claim $n^{1/2}\epsilon_n(B_n - W_{2,\omega}^2(P, Q)) \rightsquigarrow 0$. To this end, observe that

$$\begin{aligned}
&B_n - W_{2,\omega}^2(P, Q) \\
&\quad = \int_0^1 |(G^{-1} \circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - F^{-1}(u)|^2 \mathrm{d}\omega(u) \\
&\qquad - \int_0^1 |G^{-1}(u) - F^{-1}(u)|^2 \mathrm{d}\omega(u) \\
&\quad = \int_0^1 |(G^{-1} \circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - G^{-1}(u)|^2 \mathrm{d}\omega(u) \\
&\qquad + 2\int_0^1 \Big((G^{-1} \circ F)(F^{-1}(u) + n^{-1/2}\mathbf{H}_n(u)) - G^{-1}(u)\Big)
\end{aligned}$$

$$\cdot \left(G^{-1}(u) - F^{-1}(u)\right) \mathrm{d}\omega(u)$$
$$=: B_n^o + 2B_n^*.$$

Let L be the Lipschitz constant of $G^{-1} \circ F$, then

$$n^{1/2}\epsilon_n |B_n^*| \le n^{1/2}\epsilon_n \int_0^1 Ln^{-1/2} |\mathbf{H}_n(u)| \cdot |G^{-1}(u) - F^{-1}(u)| \mathrm{d}\omega(u)$$
$$\le L\epsilon_n \sqrt{A_n} W_{2,\omega}^2(P, Q),$$

where $\epsilon_n\sqrt{A_n} \rightsquigarrow 0$ by Lemma 2, hence, $n^{1/2}\epsilon_n B_n^* \rightsquigarrow 0$. Similarly,

$$n^{1/2}\epsilon_n |B_n^0| \le n^{1/2}\epsilon_n \int_0^1 L^2 n^{-1} |\mathbf{H}_n(u)|^2 \mathrm{d}\omega(u)$$
$$= L^2 n^{-1/2}\epsilon_n A_n \rightsquigarrow 0,$$

hence, $n^{1/2}\epsilon_n B_n^o \rightsquigarrow 0$. Therefore, applying Slutsky's theorem, we have

$$n^{1/2}\epsilon_n^{-1}\left(W_{2,\omega}^2(P_n, P) - \epsilon_n^2 W_{2,\omega}^2(P, Q)\right)$$
$$\rightsquigarrow \lim_{n\to\infty} 2(1-\epsilon_n)C_n$$
$$= 2\int_0^1 \mathbf{H}(u) \cdot \left(G^{-1}(u) - F^{-1}(u)\right) \mathrm{d}\omega(u).$$

Case III: Suppose $n^{1/2}\epsilon_n = \gamma$, then

$$nW_{2,\omega}^2(P_n, P) = (1-\epsilon_n)^2 A_n + \gamma^2 B_n + 2(1-\epsilon_n)\gamma C_n$$
$$= (1-\epsilon_n) \cdot ((1-\epsilon_n)A_n + 2\gamma C_n) + \gamma^2 B_n.$$

We claim that

$$(1-\epsilon_n)A_n + 2\gamma C_n \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \mathrm{d}\omega(u)$$
$$+ 2\gamma \int_0^1 \mathbf{H}(u) \cdot \left(G^{-1}(u) - F^{-1}(u)\right) \mathrm{d}\omega(u).$$

We apply the same argument used in the proof of Lemma 2; for each $n \in \mathbb{N}$, define $\mathcal{K}_n \colon \ell_m^\infty(0,1) \to \mathbb{R}$ by

$$\mathcal{K}_n(h) = (1-\epsilon_n)\int_0^1 |h(u)|^2 \mathrm{d}\omega(u) + 2\gamma \int_0^1 h(u)$$
$$\cdot \left((G^{-1} \circ F)(F^{-1}(u) + n^{-1/2}h(u)) - F^{-1}(u)\right) \mathrm{d}\omega(u).$$

In the proof of Lemma 2, we have shown that

$$\mathcal{K}_n(h_n) \to \int_0^1 |h(u)|^2 \mathrm{d}\omega(u) + 2\gamma \int_0^1 h(u) \cdot \left(G^{-1}(u) - F^{-1}(u)\right) \mathrm{d}\omega(u)$$

for any sequence $(h_n)_{n\in\mathbb{N}}$ in $\ell_m^\infty(0,1)$ and $h \in \ell_m^\infty(0,1)$ such that $\|h_n - h\|_\infty \to 0$. Therefore, the extended continuous mapping theorem shows that

$$(1-\epsilon_n)A_n + 2\gamma C_n = \mathcal{K}_n(\mathbf{H}_n) \rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \mathrm{d}\omega(u)$$
$$+ 2\gamma \int_0^1 \mathbf{H}(u) \cdot \left(G^{-1}(u) - F^{-1}(u)\right) \mathrm{d}\omega(u).$$

Hence, by Slutsky's theorem,

$$nW_{2,\omega}^2(P_n, P) = (1-\epsilon_n) \cdot ((1-\epsilon_n)A_n + 2\gamma C_n) + \gamma^2 B_n$$
$$\rightsquigarrow \int_0^1 |\mathbf{H}(u)|^2 \mathrm{d}\omega(u) + 2\gamma \int_0^1 \mathbf{H}(u)$$
$$\cdot \left(G^{-1}(u) - F^{-1}(u)\right) \mathrm{d}\omega(u) + \gamma^2 W_{2,\omega}^2(P, Q).$$

□

A.3. Proof of Theorem 1

Proof. For the asymptotic Type I error, we invoke Proposition 1: letting P_n be the empirical measure based on $X_1, \ldots, X_n$ from $H_0^{(n)}$,

$$\text{the asymptotic Type I error} = \lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) > C_\alpha)$$
$$= \mathbb{P}(A > C_\alpha) = 1 - \Psi(C_\alpha) = \alpha,$$

where A is the limit distribution of $nW_{2,\omega}^2(P_n, P)$ under $H_0^{(n)}$, namely, $A =$ (2.2). Here, weak convergence implies the above limit as the open set (C_α, ∞) is a continuity set of Ψ. Therefore, the asymptotic Type I error is α.

Next, we compute the asymptotic Type II error: assuming P_n be the empirical measure based on $X_1, \ldots, X_n$ from $H_1^{(n)}$,

$$\text{the asymptotic Type II error} = \lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) \le C_\alpha).$$

If $n^{1/2}\epsilon_n \to 0$, we have shown in Theorem 2 that the limit distribution of $nW_{2,\omega}^2(P_n, P)$ is exactly A, hence

$$\lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) \le C_\alpha) = \mathbb{P}(A \le C_\alpha) = \Psi(C_\alpha) = 1 - \alpha.$$

If $n^{1/2}\epsilon_n \to \infty$, we apply Slutsky's theorem by multiplying $n^{-1/2}\epsilon_n^{-1}$ to both sides of (3.9), which yields

$$\epsilon_n^{-2}\left(W_{2,\omega}^2(P_n, P) - \epsilon_n^2 W_{2,\omega}^2(P, Q)\right) \rightsquigarrow 0.$$

In other words, $\epsilon_n^{-2} W_{2,\omega}^2(P_n, P)$ converges to a constant $W_{2,\omega}^2(P, Q) > 0$ under $H_1^{(n)}$, hence,

$$\lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) \le C_\alpha)$$
$$= \lim_{n\to\infty} \mathbb{P}(\epsilon_n^{-2} W_{2,\omega}^2(P_n, P) \le n^{-1}\epsilon_n^{-2} C_\alpha) = 0.$$

If $n^{1/2}\epsilon_n = \gamma > 0$, the limit distribution of $T_n := nW_{2,\omega}^2(P_n, P) - \gamma^2 W_{2,\omega}^2(P, Q)$ under $H_1^{(n)}$ is (2.3) by (3.10), hence

$$\lim_{n\to\infty} \mathbb{P}(nW_{2,\omega}^2(P_n, P) \le C_\alpha) = \lim_{n\to\infty} \mathbb{P}(T_n \le C_\alpha - \gamma^2 W_{2,\omega}^2(P, Q))$$
$$= \Psi_\gamma(C_\alpha - \gamma^2 W_{2,\omega}^2(P, Q)).$$

Here, weak convergence implies the above limit as the set $(-\infty, C_\alpha]$ is a continuity set of Ψ_γ. □

Remark A.2. As noted in Remark 2, we may replace the detection boundary $n^{1/2}\epsilon_n = \gamma > 0$ with $\lim_{n\to\infty} n^{1/2}\epsilon_n = \gamma > 0$ by modifying the proofs of Theorem 2 and 1.

Acknowledgments

TL and YH thank the editors and anonymous referees for the constructive feedback on strengthening the empirical results.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Liang acknowledges the generous support from the NSF Career Award (DMS-2042473), and the William Ladany Faculty Fellowship at the University of Chicago Booth School of Business.

ORCID

YoonHaeng Hur http://orcid.org/0000-0001-6308-8896

References

Ambrosio, L., Gigli, N., and Savaré, G. (2005), *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Basel: Birkhäuser. [2]

Anderson, T. W., and Darling, D. A. (1952), "Asymptotic Theory of Certain "Goodness of Fit" Criteria based on Stochastic Processes," *The Annals of Mathematical Statistics*, 23, 193–212. [1,7]

Bobkov, S., and Ledoux, M. (2019), *One-Dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances*, Providence, RI: American Mathematical Society. [7]

Brenier, Y. (1991), "Polar Factorization and Monotone Rearrangement of Vector-Valued Functions," *Communications on Pure and Applied Mathematics*, 44, 375–417. [2]

Cai, T., Jessie Jeng, X., and Jin, J. (2011), "Optimal Detection of Heterogeneous and Heteroscedastic Mixtures," *Journal of the Royal Statistical Society*, Series B, 73, 629–662. [4]

Cai, T., and Wu, Y. (2014), "Optimal Detection of Sparse Mixtures Against a Given Null Distribution," *IEEE Transactions on Information Theory*, 60, 2217–2232. [4]

Chetty, R., Friedman, J. N., Stepner, M., and the Opportunity Insights Team (2020), "The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data," Technical Report, National Bureau of Economic Research. [1,3]

Csörgő, S. (2003), "Weighted Correlation Tests for Location-Scale Families," *Mathematical and Computer Modelling*, 38, 753–762. [5]

de Wet, T. (2002), "Goodnes-of-Fit Tests for Location and Scale Families based on a Weighted l_2-Wasserstein Distance Measure," *Test*, 11, 89–107. [5]

Del Barrio, E., Giné, E., and Utzet, F. (2005), "Asymptotics for L_2 Functionals of the Empirical Quantile Process, with Applications to Tests of Fit based on Weighted Wasserstein Distances," *Bernoulli*, 11, 131–189. [4,6]

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *The Annals of Statistics*, 32, 962–994. [4,5]

——— (2015), "Higher Criticism for Large-Scale Inference, Especially for Rare and Weak Effects," *Statistical Science*, 30, 1–25. [1]

Donoho, D. L., and Kipnis, A. (2022), "Higher Criticism to Compare Two Large Frequency Tables, with Sensitivity to Possible Rare and Weak Differences," *The Annals of Statistics*, 50, 1447–1472. [4]

Efron, B. (2010), *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Cambridge: Cambridge University Press. [1]

Fryer, R. G., Levitt, S. D., List, J. A., et al. (2015), "Parental Incentives and Early Childhood Achievement: A Field Experiment in Chicago Heights," Technical Report, National Bureau of Economic Research. [1]

Hallin, M., Mordant, G., and Segers, J. (2021), "Multivariate Goodness-of-Fit Tests based on Wasserstein Distance," *Electronic Journal of Statistics*, 15, 1328–1371. [4]

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015), "The Extent and Consequences of p-hacking in Science," *PLoS Biology*, 13, e1002106. [9,10]

Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *The Annals of Mathematical Statistics*, 35, 73–101. [1]

Kolmogorov, A. N. (1933), "Sulla Determinazione Empirica di una legge didistribuzione," *Giorn Dell'inst Ital Degli Att*, 4, 89–91. [1]

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017), "Optimal Mass Transport: Signal Processing and Machine-Learning Applications," *IEEE Signal Processing Magazine*, 34, 43–59. [2]

Lehmann, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses* (Third ed.). Springer. [10]

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2016), "The Behavioralist Goes to School: Leveraging Behavioral Economics to Improve Educational Performance," *American Economic Journal: Economic Policy*, 8, 183–219. [1]

McCann, R. J. (1997), "A Convexity Principle for Interacting Gases," *Advances in Mathematics*, 128, 153–179. [2]

Mroueh, Y. (2020), "Wasserstein Style Transfer," in *International Conference on Artificial Intelligence and Statistics* (Vol. 108), pp. 842–852, PMLR. [2]

Munk, A., and Czado, C. (1998), "Nonparametric Validation of Similar Distributions and Assessment of Goodness of Fit," *Journal of the Royal Statistical Society*, Series B, 60, 223–241. [4,6]

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012), "Wasserstein Barycenter and its Application to Texture Mixing," in *Scale Space and Variational Methods in Computer Vision*, eds. L. Calatroni, M. Donatelli, S. Morigi, M. Prato, and M. Santacesaria, pp. 435–446, Cham: Springer. [2]

Shorack, G. R., and Wellner, J. A. (2009), *Empirical Processes with Applications to Statistics*, Philadelphia: Society for Industrial and Applied Mathematics. [8]

Smirnoff, N. (1939), "Sur les écarts de la courbe de distribution empirique," *Matematicheskii Sbornik*, 48, 3–26. [1]

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer. [6,10,11]

Villani, C. (2003), *Topics in Optimal Transportation*, Providence, RI: American Mathematical Society. [2]