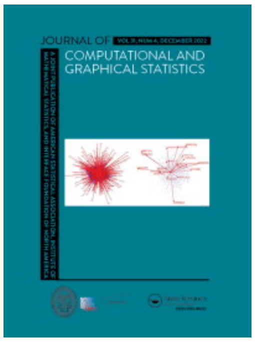

Journal of Computational and Graphical Statistics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/ucgs20

Log-Rank-Type Tests for Equality of Distributions in High-Dimensional Spaces

Linxi Liu, Yang Meng, Xiaoru Wu, Zhiliang Ying & Tian Zheng

To cite this article: Linxi Liu, Yang Meng, Xiaoru Wu, Zhiliang Ying & Tian Zheng (2022) Log-Rank-Type Tests for Equality of Distributions in High-Dimensional Spaces, Journal of Computational and Graphical Statistics, 31:4, 1384-1396, DOI: 10.1080/10618600.2022.2051530

To link to this article: https://doi.org/10.1080/10618600.2022.2051530

JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2022, VOL. 31, NO. 4, 1384–1396
https://doi.org/10.1080/10618600.2022.2051530

Taylor & Francis
Taylor & Francis Group

Log-Rank-Type Tests for Equality of Distributions in High-Dimensional Spaces

Linxi Liu[a], Yang Meng[b], Xiaoru Wu[c], Zhiliang Ying[b], and Tian Zheng[b]

[a]Department of Statistics, University of Pittsburgh, Pittsburgh, PA; [b]Department of Statistics, Columbia University, New York, NY; [c]Facebook, Inc., Menlo Park, CA

ABSTRACT
Motivated by applications in high-dimensional settings, we propose a novel approach for testing the equality of two or more populations by constructing a class of intensity centered score processes. The resulting tests are analogous in spirit to the well-known class of weighted log-rank statistics that are widely used in survival analysis. The test statistics are nonparametric, computationally simple, and applicable to high-dimensional data. We establish the usual large sample properties by showing that the underlying log-rank score process converges weakly to a Gaussian random field with zero mean under the null hypothesis and with a drift under the contiguous alternatives. For the Kolmogorov–Smirnov-type and the Cramer-von Mises-type statistics, we also establish the consistency result for any fixed alternative. Cutoff points for the test statistics are obtained by permutations or a simulation-based resampling method. The new approach is applied to a study of brain activation measured by functional magnetic resonance imaging when performing two linguistic tasks and also to a prostate cancer DNA microarray dataset. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received October 2020
Revised September 2021

KEYWORDS
Consistency;
Distribution-free method;
Gaussian random field;
High-dimensional data;
Nonparametric tests;
Random weighting;
Rank-based tests

1. Introduction

Testing for differences between two populations is one of the classical problems in nonparametric inference. For univariate data, there exists an extensive literature (Hájek, Šidák, and Sen 1999; Hollander, Wolfe, and Chicken 2013; Lehmann and Romano 2005). A common feature in most approaches is that the test statistics are based on the ranks of the observations in the pooled sample, which are invariant under monotone transformations. As a result, the test statistics are distribution free under the null hypothesis. An intrinsic difficulty in extending these tests to the multivariate case is the fact that there is no natural order on the p-dimensional ($p \geq 2$) space and monotone transformation of each coordinate does not necessarily lead to the uniform distribution on the unit p-cube. Consequently, the test statistics are no longer distribution free under the null hypothesis in the multivariate case.

Since the multivariate case has been studied far less thoroughly, there has always been interest in finding nonparametric approaches. As a generalization of the univariate ordering of the pooled sample, Friedman and Rafsky (1979) proposed a multivariate run test based on the minimal spanning tree of the observations. Rosenbaum (2005), on the other hand, introduced the concept of optimal nonbipartite matching and proposed what he called the cross-match test. Based on the number of nearest neighbor type coincidences, Henze (1988) developed a procedure which can be implemented using an approximate permutation test. By further incorporating rank information, Hall and Tajvidi (2002) developed a permutation-based approach for testing group differences. Such two-sample testing methods often suffer from low power when the dimension is moderate to high, even for the basic location or scale alternatives. More recently, Chen and Friedman (2017) and Chen, Chen, and Su (2018) have extended tests that are based on the number of coincidences in nearest neighbors to adapt to high-dimensional settings. In low-dimensional cases, the number of coincidences is expected to be larger under the alternative hypothesis, an evidence against the null hypothesis that motivated their proposed test. However, the number may exhibit an inverse trend in high-dimensional cases due to the high-dimensional geometry, which is especially the case for scale alternatives. Based on this insightful observation, the authors defined a new test statistic which is able to capture these two types of changes simultaneously, thus, the new test is able to retain its power as the dimension increases compared to comparison methods. On the other hand, the test is constructed based on the number of coincidences only, without considering relative ordering of distances (i.e., rank) in nearest neighbors. There are also several methods based on distance between two samples. Aslan and Zech (2005) introduced a test statistic corresponding to the difference in potential energy between two systems. Gretton et al. (2012) proposed to use test functions from a reproducing kernel Hilbert space, and the distance between two samples is defined to be the expectation of kernelized one.

Survival analysis is one of the most active areas where nonparametric tests are the gold standard. The properties of tests in this area are well understood. Most notably, the log-rank test and its weighted versions are commonly used in testing for treatment effects when survival time is the outcome variable. We refer to the work by Mantel and Haenszel (1959), Mantel (1966),

CONTACT Zhiliang Ying ✉ zy9@columbia.edu; Tian Zheng ✉ tzheng@stat.columbia.edu Department of Statistics, Columbia University, New York, NY.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JCGS.

Gehan (1965), Peto and Peto (1972), and Prentice (1978) for some of the initial developments and the work by Gill (1980) for the counting process and associated martingale representations of the weighted log-rank test statistics.

The nonparametric nature, the flexibility of the weight functions, the well-understood theoretical properties and the widely available software tools suggest the utility of using the class of weighted log-rank statistics to connect the problem of testing for differences in high-dimensional data to survival analysis. To that end, we propose a survival analysis approach to the multivariate K-sample problem, by converting multivariate data into survival data. Using this conversion, we are able to make use of the powerful weighted log-rank tests to develop a class of nonparametric tests. This approach is simple to implement and flexible with easy adaptation to different space configurations.

We develop large sample properties for the proposed statistics by studying the underlying score processes. By centering the scores at their marginal intensities, we establish their weak convergence to Gaussian random fields under the null and contiguous alternative hypotheses. For the Kolmogorov–Smirnov-type and Cramer-von Mises-type tests, we also establish the consistency against any fixed alternative. As a practical means to obtain approximate cutoff points, we propose a simulation based resampling method that is easy to implement and has rigorous theoretical justification.

The rest of the article is organized as follows. The proposed tests and their theoretical properties are given in the next section. Section 3 is devoted to simulation studies. In Section 4, the method is applied to two real datasets. Some concluding remarks are given in Section 5.

2. A Class of Weighted Log-Rank Based Test Statistics

We introduce the weighted log-rank score process in this section, and define two types of test statistics based on it. The null distribution for two-sample problems is derived in Theorem 2.2 and Corollary 2.3, while the extension to K-sample problems is summarized in Theorem 2.6. A more computationally feasible approach for approximating the null distribution is proposed based on the results of Theorem 2.5 (two-sample problems) and Theorem 2.7 (K-sample problems). In Section 2.3, we study asymptotic power of the proposed approach, and obtain the consistency in Theorem 2.8 and local properties in Theorem 2.10, respectively. In particular, for two examples about location and scale alternatives, under contiguous alternatives our calculations show explicitly what factors may affect the power.

2.1. Weighted Log-Rank Score Process and Test Statistics

We consider first the case of two populations. Let $X_1, \dots, X_{n_1}$ be a p-dimensional random sample from a population with distribution function F_1 and $X_{n_1+1}, \dots, X_{n_1+n_2}$ be a second random sample from distribution F_2. As sample size $n = n_1+n_2$ goes to infinity, we assume that $n_2/n \to r \in (0,1)$. Suppose that the null hypothesis of interest is

$$H_0 : F_1 = F_2, \tag{1}$$

and that the alternative hypothesis is its complement. For any fixed point x in $\mathbb{R}^p$, let $T_i(x) = d(X_i, x)$, where d is a certain distance metric on $\mathbb{R}^p$. Here, d could be any measure of (dis)similarity. Typically d is taken to be the Euclidean distance, $d(X_i, x) = ||X_i - x||$. Furthermore, define $T_i(x)$-induced counting process

$$N_i(x;t) = I(T_i(x) \leq t), \quad t \geq 0, \quad i = 1, \dots, n,$$

where $I(\cdot)$ denotes the indicator function.

By converting the original observations $X_1, \dots, X_n$ into $T_1(x), \dots, T_n(x)$, we may regard T_i as survival times and connect the two-sample problem with the well-studied weighted log-rank tests for survival data. Define the weighted log-rank score process

$$U_w(x;t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t W_x(s) \left(Z_i - \frac{\sum_{j=1}^n Z_j I(T_j(x) \geq s)}{\sum_{j=1}^n I(T_j(x) \geq s)} \right) dN_i(x;s), \tag{2}$$

where $Z_i = I(i > n_1)$ and $W_x(s)$ is a weight function. Equation (2) can be viewed as a statistic examining the arrival pattern of data at x: $\frac{\sum_{j=1}^n Z_j I(T_j(x) \geq s)}{\sum_{j=1}^n I(T_j(x) \geq s)}$ is an estimate of proportion of points from F_2 based on all observations at risk (with a survival time $\geq T_i(x)$), while Z_i denoting the arrival of an individual point from F_2. Let $T_{(1)}(x), \dots, T_{(n)}(x)$ denote the order statistics of $T_1(x), \dots, T_n(x)$. The test statistic $U_\omega(x;t)$ is obtained by contrasting the observed arrivals (Z_i's) to the expected ones at time points $T_{(1)}(x), T_{(2)}(x), \dots,$ up to time t. Note that the relative rank of data points in the neighborhood of x affects estimation of expected arrivals at different time points. It is also further captured by the weight function W_x. A widely used class of weight functions in survival analysis corresponds to the G-ρ class (Harrington and Fleming 1982) of weighted log-rank test statistics, where $W_x(t) = w(\widehat{F}(x;t-))$, $w(u) = (1-u)^\rho$, $1 - \widehat{F}(x;t-)$ is the Kaplan–Meier estimate of the survival function from $T_1(x), \dots, T_n(x)$ and $\rho \geq 0$ is a tuning parameter. Let $\tilde{t}$ be an upper bound on the distances $T_i, i = 1, \dots, n$. For a fixed x, $U_w(x;\tilde{t})$ is the usual log-rank test statistic for $\rho = 0$ and it becomes the Peto–Prentice extension of the Wilcoxon statistic when $\rho = 1$. The choice of ρ depends on the projected alternatives: if the difference in the two corresponding hazards is more pronounced for smaller t values, then a larger ρ is preferred. Thus, if the group difference lies in "local features," the Peto–Prentice statistic achieves higher efficiency. For any fixed x, since Equation (1) implies that the T_i have a common distribution, it follows from the usual counting process-martingale approach to survival analysis (Gill 1980; Harrington and Fleming 1982; Anderson et al. 1993) that $U_w(x;t)$ is a zero-mean martingale in t for a suitable σ-filtration under H_0. Viewed as a two-parameter process (of x and t), Theorem 2.2 below shows that U_w converges weakly to a zero-mean Gaussian random field on $\mathbb{R}^{p+1}$. Moreover, the data generating process can be viewed as first independently sampling $Z_i \sim$ Bernoulli(r), then $X_i | Z_i = 0 \sim F_1$ and $X_i | Z_i = 1 \sim F_2$. Thus, (X_i, Z_i) are iid pairs. Let $\widehat{\Gamma}_k(x;t) = n^{-1} \sum_{i=1}^n Z_i^k I(T_i(x) \geq t)$, $k = 0, 1$. By the uniform law of large numbers (see Pollard 1990, sec. 8; van der Vaart and Wellner 1996, sec. 2.4), $\sup_{x,t} \left| \widehat{\Gamma}_k(x;t) - E\left(Z_1^k I(T_1(x) \geq t) \right) \right| \to 0$ a.e. as $n \to \infty$, where the expectation is taken with respect to (X_1, Z_1). The limit will be denoted as $\Gamma_k(x;t)$ for $k = 0, 1$

throughout. Under the null hypothesis, the limit is simply $1 - F_1(t-)$ and $r(1 - F_1(t-))$ for $k = 0, 1$, respectively.

To avoid tail instability, we restrict our attention to the set $\mathcal{R}$, a compact subset of $\{(x, t) \in \mathbb{R}^p \times [0, \infty) : P(T_1(x) > t) > \gamma\}$, where $\gamma \in (0, 1)$ is a small positive constant. We introduce the following two regularity conditions:

Condition 1. For distribution of X_1, a density exists and is bounded.

Condition 2. For any $(x, t) \in \mathcal{R}$, there exists a nonrandom constant B such that the total variation $|W_x(0)| + \int_0^t |dW_x(s)| \leq B$, where $\int_0^t |dW_x(s)|$ is L_1-type total variation.

Remark 2.1. The first condition ensures that $T_i(x)$ has a density, as well as the existence of the hazard rate function. The condition can be further relaxed to the case when the cdf of X_1 has finite number of discontinuities. Meanwhile, if the weight function is chosen in line with the G-ρ class of weighted log-rank test statistic, then Condition 2 is satisfied.

Theorem 2.2. Assume Conditions 1 and 2 hold. Under H_0, the weighted log-rank score process $U_w(\cdot; \cdot)$ converges weakly to a zero-mean Gaussian random field $G_0(\cdot; \cdot)$ on $\mathcal{R}$ with covariance function $C(x_1, t_1; x_2, t_2)$ as

$$E\left[\int_0^{t_1} W_{x_1}(s)\left(Z_1 - \frac{\Gamma_1(x_1; s)}{\Gamma_0(x_1; s)}\right) dM_1(x_1; s) \times \int_0^{t_2} W_{x_2}(s)\left(Z_1 - \frac{\Gamma_1(x_2; s)}{\Gamma_0(x_2; s)}\right) dM_1(x_2; s)\right],$$

where the expectation is taken with respect to (X_1, Z_1), $M_1(x; t) = N_1(x; t) - \int_0^t I(T_1(x) \geq s) d\Lambda(x; s)$ and $\Lambda(x; s)$ is the cumulative hazard function of $T_1(x)$.

Theorem 2.2, along with other asymptotic properties in subsequent developments, will be proved in the Appendix. Using (2) as the basic vehicle, we can construct a variety of test statistics. The main idea is to combine the weighted log-rank statistics at different x locations and to introduce censoring to the observed survival times so that the resulting test statistic has more robust power. Two common ways to summarize over x correspond to the Kolmogorov–Smirnov (K–S) sup-type and the Cramer-von Mises (C-vM) integral-type approaches. The latter also includes the Anderson–Darling test (Anderson and Darling 1954). To those ends, we propose the following statistics:

$$U_1 = \sup_{x \in D} |U_w(x; \tilde{t})| \quad \text{(K-S sup-type)},$$

and

$$U_2 = \int_{x \in D} U_w^2(x; \tilde{t}) \pi(x, \tilde{t}) dx \quad \text{(C-vM integral-type)},$$

where D is a suitably chosen subset of $\mathcal{R}$ and $\pi(x, t)$ is a weight function. Here $\tilde{t}$ is typically an upper bound, which is commonly used in survival analysis to control possible tail instability. Let $T_{(1)}(x), \ldots, T_{(n)}(x)$ denote the order statistics from $\{T_1(x), \ldots, T_n(x)\}$. We may set $\tilde{t}$ to be the kth order statistic $T_{(k)}(x)$, mimicking the Type II censoring in survival analysis. The use of $T_{(k)}(x)$ localizes the discrepancy between the two populations. For example, if x is surrounded locally by the first group of X's, then, with a small k, $U_w(x; T_{(k)}(x))$ tends to take a negative value. Taking a small k is in the spirit of k-nearest neighbor method in nonparametric regression. This idea can be generalized to other censoring schemes to accommodate different data patterns.

The weak convergence of U_w can be used to derive limiting distributions of U_1 and U_2, which are functionals of U_w.

Corollary 2.3. Let G_0 be the Gaussian random field defined in Theorem 2.2. Then, under H_0 with the same assumptions, U_1 and U_2 converge in distribution to $\sup_{x \in D} |G_0(x; \tilde{t})|$ and $\int_{x \in D} G_0^2(x; \tilde{t}) \pi(x, \tilde{t}) dx$, respectively.

Remark 2.4. The choice of test set D essentially reflects a bias-variance tradeoff. For the K–S sup-type statistic, the larger the set D is, the higher the variance is and the closer the mean of test statistic is to the theoretical value. While for the C-vM integral-type statistic, a larger D generally implies a lower variance. At the same time, the strong evidence against the null observed at certain points may be attenuated by other points at which the arrival pattern of the two samples does not differ too much. In Example 2.2, we see that for two normal distributions only different in scales, by setting D to be the mean of the two distributions, the test achieves highest power. However, in practice such an optimal choice is typically not available. We provide three strategies for choosing D. One possible choice is the set of observed data. Given that maximizing over the support of the multivariate distribution is computationally infeasible, this choice enables us to avoid it while providing a reasonable approximation. Setting D to be equally spaced grid (see the application in Section 4.1) is another practical solution for reducing computational cost, especially for low dimensional cases. We can also choose a smaller D which only contains finite number of points. In particular, a clustering based approach is introduced for simulation studies.

In summary, one would prefer to use the *grid* option in low dimensions. For high dimensions, if the dataset is moderate in size, having D coincide with the observed data points offers a good approximation. For larger datasets, the *clustering* option provides a practice tradeoff between performance and computability. We choose D for our simulations and real data results accordingly.

Because of their intractable forms, the limiting distributions in Corollary 2.3 does not immediately lead to the cutoff points for the corresponding tests. An alternative way is to simulate replicates of processes U_w^*, which have asymptotically the same limiting distribution as that of U_w, thereby constructing respective functionals of U_w^* numerically approximates the distribution of U_1 and U_2. Such an approximation is theoretically justified by the following result.

Theorem 2.5. Let $U_w^*(x; t) = n^{-1/2} \sum_{i=1}^n V_i \int_0^t W_x(s)\Big(Z_i - \frac{\widehat{\Gamma}_1(x;s)}{\widehat{\Gamma}_0(x;s)}\Big) d\widehat{M}_i(x; s)$, where $\widehat{M}_i(x; t)$ is the same as $M_i(x; t)$ except with $\Lambda(x; t)$ being replaced by the Nelson–Aalen estimator (Anderson et al. 1993) and V_i are independent standard normal random variables that are independent of $\{(X_i, Z_i)\}_{i=1}^n$, the observed data. Then, when Conditions 1 and 2 hold the conditional distribution of U_w^* given data $\{(X_i, Z_i)\}_{i=1}^n$ converges to the same limiting Gaussian random field G_0 on $\mathcal{R}$ as that of U_w.

Note that the conditional distribution of U_w^* given $\{(X_i, Z_i)\}_{i=1}^n$ can be simulated by repeatedly generating random sequences $\{V_i\}_{i=1}^n$. Therefore, we can approximate the distribution of any functional of U_w by that of the corresponding functional of U_w^*, which can be obtained via simulations. This random weighting method is computationally efficient because $\int_0^t W_x(s)\Big(Z_i - \frac{\widehat{\Gamma}_1(x;s)}{\widehat{\Gamma}_0(x;s)}\Big)d\widehat{M}_i(x;s), i = 1, \dots, n$, are fixed at their observed values for each sample of U_w^*. We refer to Lin, Wei, and Ying (2002) for a comprehensive discussion of such an approach and related methods. Specifically, to obtain the cutoff value for U_1 or U_2, we can either use a permutation test by randomly dividing n observations into two groups of sizes n_1 and n_2 or utilize the simulation based resampling by repeatedly generating $U_1^* = \sup_{x\in D}|U_w^*(x;\tilde{t})|$ or $U_2^* = \int_{x\in D}[U_w^*(x;\tilde{t})]^2\pi(x,\tilde{t})dx$ when the sample size is relatively large.

2.2. Extension to the K-Sample Case

Assume we first generate Z_i from a discrete distribution on $\{1, \dots, K\}$ with parameters $(r_1, \dots, r_k)$, where $0 < r_k < 1$ and $\sum_{k=1}^K r_k = 1$. Given $Z_i = k$, $X_i \sim F_k$. Denote data as $\{(X_i, Z_i)\}_{i=1}^n$. After rearrange of orders, for each k, $\{X_{ki}\}_{i=1}^{n_k}$ can be viewed as a random sample from a population with distribution function F_k on $\mathbb{R}^p$. We wish to test the null hypothesis that $F_1 = \cdots = F_K$ against the complementary alternative. Similarly to the two-sample case ($K = 2$), for each fixed point $x \in \mathbb{R}^p$, let $T_i(x) = d(X_i, x)$ and $N_i(x;t) = I(T_i(x) \le t)$. In the K-sample case, the weighted log-rank score process $U_w = (U_{w1}, \dots, U_{w(K-1)})^T$, where

$$U_{wk}(x;t) = n^{-1/2}\sum_{i=1}^n \int_0^t W_x(s)\left(\delta_{ik} - \frac{\widehat{\Gamma}_k(x;s)}{\widehat{\Gamma}(x;s)}\right)dN_i(x;s),$$

where $\delta_{ik} = I(Z_i = k)$, $\widehat{\Gamma}_k(x;t) = \sum_{i=1}^n \delta_{ik}I(T_i \ge t)$ and $\widehat{\Gamma}(x;t) = \sum_{i=1}^n I(T_i \ge t)$. Again, by strong law of large numbers, $\widehat{\Gamma}_k(x;t)/n \to r_k(1 - F_k(t-))$ a.e. and $\widehat{\Gamma}(x;t)/n \to \sum_{k=1}^K r_k(1 - F_k(t-))$ a.e. as $n \to \infty$. The two limits under the null hypothesis are denoted as $\Gamma_k(x;t)$ and $\Gamma(x;t)$, respectively. Under some mild regularity conditions, we have the following result:

Theorem 2.6. Assume Conditions 1 and 2 hold. Under the null hypothesis, the weighted log-rank score process U_w converges weakly to a zero-mean multivariate Gaussian random field on $\mathcal{R}$ with covariance matrix $C(x_1, t_1; x_2, t_2)$, whose (l, j)th element is

$$E\left[\int_0^{t_1} W_{x_1}(s)\left(\delta_{1l} - \frac{\Gamma_l(x_1;s)}{\Gamma(x_1;s)}\right)dM_1(x_1;s) \right. \\ \left. \times \int_0^{t_2} W_{x_2}(s)\left(\delta_{1j} - \frac{\Gamma_j(x_2;s)}{\Gamma(x_2;s)}\right)dM_1(x_2;s)\right],$$

where the expectation is taken with respect to (X_1, Z_1), $M_1(x;t) = N_1(x;t) - \int_0^t I(T_1(x) > s)d\Lambda(x;s)$ and $\Lambda(x;t)$ is the cumulative hazard function of $T_1(x)$.

For statistics defined as functionals of U_w, it is again difficult to obtain analytic forms of their limiting distributions. We therefore, propose to use a simulation based random weighting method to approximate the limiting distributions and to obtain the cutoff points. For theoretical justification, we need the following generalization of Theorem 2.5.

Theorem 2.7. Let

$$U_{wl}^*(x;t) = n^{-1/2}\sum_{k=1}^K \sum_{i=1}^{n_k} V_{ki}\int_0^t W_x(s)\left(\delta_{kl} - \frac{\Gamma_l(x;s)}{\Gamma_.(x;s)}\right)d\widehat{M}_{ki}(x;s),$$

where $\widehat{M}_{ki}(x;t)$ is the same as $M_{ki}(x;t)$ with $\Lambda(x;t)$ replaced by the Nelson–Aalen estimator, V_{ki} are independent standard normals and are independent of the data. Then, under Conditions 1 and 2 the conditional distribution of $U_w^* = (U_{w1}^*, \dots, U_{w(K-1)}^*)^T$ given the data converges to the same limiting multivariate Gaussian random field on $\mathcal{R}$ as in Theorem 2.6.

2.3. Asymptotic Properties Under Alternative Hypotheses

In this section, we establish asymptotic properties for the proposed test statistics under alternative hypothesis. We first consider the case of a fixed alternative and show that the Kolmogorov–Smirnov-type and Cramer-von Mises-type tests are consistent against any such alternative.

Theorem 2.8. Under any fixed alternative $F_2 \ne F_1$, there exists a compact set $\mathcal{R} \subset \{(x,t) : P(T_1(x) > t) > \gamma\}$, such that for any sequence of random variables $\{c_n : n \ge 1\}$ converging to a constant $c > 0$, we have

$$\lim_{n\to\infty} P\Big(\sup_{(x,t)\in\mathcal{R}} |U_w(x;t)| > c_n\Big) = 1$$

and if, $\pi(x,t) > 0$ for all $(x,t) \in \mathcal{R}$,

$$\lim_{n\to\infty} P\Big(\int_{(x,t)\in\mathcal{R}} U_w^2(x;t)\pi(x,t)dtdx > c_n\Big) = 1.$$

Remark 2.9. Because U_w is standardized and converges weakly, the critical values for these tests, for example those obtained via the random weighting method discussed in Section 2.1, should converge in probability to constants under the null and under the alternative. Therefore, the above theorem shows that the two types of tests are consistent when summarized over all x and t in a suitable compact set. However, for a single pair of x and t, it is not hard to see that $U_w(x,t)$ itself may not have any power. For instance, if the first population follows $\mathcal{N}_p(\mu, \Sigma)$ and the second follows $\mathcal{N}_p(-\mu, \Sigma)$ and x is fixed at the origin, then $U_w(x,t)$ will have no power due to the symmetry.

It is customary to study asymptotic power of test statistics by considering contiguous alternatives. We next derive the limiting distributions of the proposed tests under such contiguous alternatives. Following Hájek, Šidák, and Sen (1999), let P_n be the sequence probability measures representing the null and Q_n its contiguous alternatives. Denote by $\frac{dQ_n}{dP_n}$ the Radon–Nikodym derivative between the two measures.

Theorem 2.10. Suppose that, for any x and t, the random vector $\left(U_w(x;t), \log \frac{dQ_n}{dP_n}\right)$ converges in distribution to $\mathcal{N}\left(\binom{0}{-0.5\sigma^2}, \begin{pmatrix} C(x;t) & \tau(x;t) \\ \tau(x;t) & \sigma^2 \end{pmatrix}\right)$ under P_n. Then, under Q_n, $U_w(x;t)$ converges weakly to a Gaussian random field $G_Q(x;t)$ with mean $\tau(x;t)$ and covariance function $C(x;t)$.

Corollary 2.11. Under the assumption in Theorem 2.10, we have for any $\tilde{t}$
(i) $\sup_{x\in D}|U_w(x;\tilde{t})|$ converges in distribution to $\sup_{x\in D}|G_Q(x;\tilde{t})|$.
(ii) $\int_{x\in D} U_w^2(x;\tilde{t})\pi(x,\tilde{t})dx$ converges in distribution to $\int_{x\in D} G_Q^2(x;\tilde{t})\pi(x,\tilde{t})dx$.

Comparing the limiting distribution in Theorem 2.10 with that in Theorem 2.2, we know that the power of the respective tests is governed by the mean drift $\tau(x;t)$ of the Gaussian random field. To gain some insight into the power properties, we calibrate $\tau(x;t)$ in two concrete examples. Recall we assume the following data generating model: $Z \sim \text{Bernoulli}(r)$, $X \sim F_1$ if $Z = 0$, and $X \sim F_2$ otherwise. Let $\delta(x;t) = I(\|X - x\| \leq t)$, and H_x be the distribution function of $\|X - x\| \wedge t$.

Example 2.1. Consider two populations differing in location:

$$H_1 : X \sim f(\cdot - n^{-1/2}\mathbf{c}Z),$$

where f is the density function of X under the null and $\mathbf{c}$ is a p-vector of constants. The alternative is clearly contiguous to H_0 (Hájek, Šidák, and Sen 1999). It can be shown that the drift function has expression

$$\begin{aligned}\tau(x;t) = r(1-r)E[\{\delta\psi \circ H_x(\|X-x\|) \\ + (1-\delta)\Psi \circ H_x(t)\}\mathbf{c}^T f'(X)/f(X)],\end{aligned}$$

where $\psi(u) = w(u) - \int_0^u w(v)/(1-v)dv$ and $\Psi(u) = \phi(u) - w(u)$. In particular, if f is the density for $\mathcal{N}_p(\mu, I)$, where I is the identity matrix, and $t = +\infty$, then $\tau(x;t) = -r(1-r)E\{\psi \circ H_x(\|X-x\|)\mathbf{c}^T(X-\mu)\}$. When $\mu = 0$ and $D = \{0\}$, that is, if under the null hypothesis, X follows a p-dimensional normal distribution with mean zero and if we construct our test statistics using $D = \{0\}$, it is easily seen that $\tau(x;t) = 0$. Thus, the corresponding tests have no power. This is somewhat surprising since the two populations do differ by a contiguous location shift.

Example 2.2. We next consider two populations differing in scale:

$$H_2 : X \sim \exp(-n^{-1/2}\mathbf{c}^T\mathbf{1}_p Z)f(\exp(-n^{-1/2}Z\mathbf{c}) * \cdot),$$

where $\mathbf{1}_p$ is a p-vector consisting of all ones and $\mathbf{a} * \mathbf{b} = (a_1b_1, \ldots, a_pb_p)$ for p-dimensional vectors $\mathbf{a} = (a_1, \ldots, a_p)^T$ and $\mathbf{b} = (b_1, \ldots, b_p)^T$. We can get

$$\begin{aligned}\tau(x;t) = 2r(1-r)E[\{\delta\psi \circ H_x(\|X-x\|) \\ + (1-\delta)\Psi \circ H_x(t)\}(\mathbf{c} * X)^T f'(X)/f(X)].\end{aligned}$$

If we let $t = +\infty$, f be the density of $\mathcal{N}_p(0, \sigma_0^2 I)$ and $\mathbf{c} = c_0\mathbf{1}_p$, then it simplifies to

$$\tau(x;t) = -2r(1-r)c_0\sigma_0^{-2}E\{\psi \circ H_x(\|X-x\|)\,\|X\|^2\}.$$

In this case, if we let D contain only one point x, then $x = 0$ gives the highest power and the farther x moves away from 0, the lower the power is.

3. Simulation

In this section, we demonstrate the Type I error rate control and power gain of the proposed method on a range of synthetic datasets, covering location alternatives, scale alternatives, and a mixture of both. We focus on the two-sample testing problems and consider both the K–S sup-type tests and the C-vM integral-type tests. We use two approaches to choose a set of data dependent test points D: simply choosing the observed data points and a clustering based approach, which will be explained later.

In Tables 1–3, the definition of ρ is the same as that in the G-ρ class of weighted log-rank tests discussed in Section 2.1. For each simulation setting we estimate the Type I error[1] and power based on 400 repetitions, and for each repetition the cutoff value is obtained by 200 permutations.

We also compare our method with the edge-count test by Chen and Friedman (2017), kernel-based maximum mean discrepancy (MMD) test by Gretton et al. (2012), and energy test by Aslan and Zech (2005). For the edge-count test, assume G is a similarity graph constructed on the pooled samples, which has no multi-edge. In other words, any pair of nodes is connected by at most one edge. The k-minimum spanning tree (k-MST) is an example of such graphs. Let R_0 be the number of between-sample edges, R_1 be the number of edges connecting a pair of points both from F_1, and R_2 be the number of edges connecting a pair of points both from F_2. The authors define a new test

Table 1. Performance of the comparison methods under the location shifts scenario.

		Log-rank-type test (K–S)		Log-rank-type test (C-vM)	
		$\rho = 0, 0.5, 1$		$\rho = 0, 0.5, 1$	
p	Δ	Type I error	Power	Type I error	Power
2	0.6	0.05, 0.05, 0.05	0.55, 0.58, 0.57	0.05, 0.06, 0.06	0.54, 0.55, 0.54
10	0.8	0.05, 0.06, 0.05	0.62, 0.65, 0.64	0.05, 0.05, 0.05	0.64, 0.67, 0.69
30	1.1	0.06, 0.06, 0.04	0.53, 0.60, 0.59	0.05, 0.05, 0.05	0.46, 0.56, 0.59
50	1.4	0.05, 0.04, 0.06	0.59, 0.65, 0.64	0.05, 0.06, 0.05	0.60, 0.65, 0.66
70	1.7	0.06, 0.05, 0.04	0.59, 0.65, 0.65	0.05, 0.05, 0.06	0.57, 0.59, 0.60
90	2	0.05, 0.06, 0.04	0.53, 0.58, 0.60	0.04, 0.05, 0.04	0.50, 0.57, 0.59
100	2	0.06, 0.05, 0.06	0.56, 0.63, 0.66	0.05, 0.06, 0.04	0.57, 0.62, 0.64
		Edge-count test		Energy test	
		1-, 3-, 5-MST			
p	Δ	Type I error	Power	Type I error	Power
2	0.6	0.04, 0.04, 0.03	0.04, 0.13, 0.22	0.05	0.64
10	0.8	0.06, 0.04, 0.06	0.06, 0.14, 0.23	0.03	0.77
30	1.1	0.06, 0.04, 0.04	0.15, 0.23, 0.34	0.04	0.89
50	1.4	0.06, 0.05, 0.04	0.19, 0.33, 0.48	0.04	0.97
70	1.7	0.06, 0.06, 0.04	0.35, 0.57, 0.68	0.04	0.98
90	2	0.04, 0.03, 0.03	0.49, 0.73, 0.84	0.06	0.98
100	2	0.05, 0.04, 0.05	0.33, 0.62, 0.80	0.06	0.99
		MMD test		Hotelling's T^2 test	
p	Δ	Type I error	Power	Type I error	Power
2	0.6	0.04	0.60	0.05	0.66
10	0.8	0.04	0.70	0.04	0.72
30	1.1	0.05	0.76	0.04	0.81
50	1.4	0.03	0.84	0.03	0.82
70	1.7	0.03	0.96	0.04	0.78
90	2	0.04	0.98	0.05	0.33
100	2	0.04	0.98	-	-

[1] In all simulations, Type I error is evaluated under the null hypothesis, that is, $\Delta = 0$ for location shifts and $\sigma = 1$ for scale alternatives.

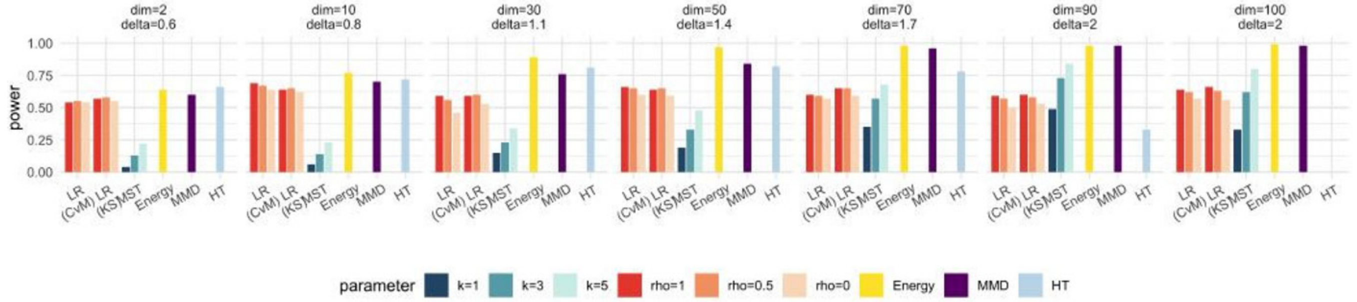

Figure 1. Power analysis under the location shifts scenario. Log-rank (LR) type tests using the Cramer-von Mises (C-vM) tests and the Kolmogorov–Smirnov (K–S) sup-type tests at three ρ values are compared with the edge-count test (MST) by Chen and Friedman (2017) at three k values, as well as the MMD test by Gretton et al. (2012), the energy test by Aslan and Zech (2005), and Hotelling's T-squared test (HT), under seven simulation settings with different numbers of dimension and location shifts. Here, power is estimated based on 400 repetitions, and for each repetition the cutoff value is obtained by 200 permutations.

statistic as

$$S = (R_1 - \mu_1, R_2 - \mu_2)\Sigma^{-1}\begin{pmatrix} R_1 - \mu_1 \\ R_2 - \mu_2 \end{pmatrix},$$

where $(\mu_1, \mu_2)^T$ and Σ are the mean and covariance matrix of the vector $(R_1, R_2)^T$ under the permutation null distribution, respectively. The authors show that under the permutation null and other mild conditions, the test statistic asymptotically follows a χ^2-distribution with degree of freedom 2. In the following simulations, this method is implemented using the `gTests` package by Chen and Friedman (2017). For the kernal based MMD test, the test statistic is defined to be the largest difference in expectations over functions within in the unit ball of a reproducing kernel Hilbert space. The method is implemented using the `kernlab` package. For the energy test, a permutation method is applied to calibrate the distribution of the test statistic $\phi_{n_1 n_2}$ under the null, while the test statistic is

$$\begin{aligned}\phi_{n_1 n_2} &= \frac{1}{2n_1(n_1-1)}\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1} R(|x_i - x_j|) \\ &+ \frac{1}{2n_2(n_2-1)}\sum_{i=1}^{n_2}\sum_{j\neq i}^{n_2} R(|y_i - y_j|) \\ &- \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} R(|x_i - y_j|)\end{aligned}$$

where $R(r) = -\ln r$.

3.1. Scenario I: Location Shifts

Under this simulation scenario, we generate two samples from the p-dimensional multivariate normal distributions $\mathcal{N}(\mu_1, I_p)$ and $\mathcal{N}(\mu_2, I_p)$, respectively, and consider the number of dimension $p = 2, 10, 30, 50, 70, 90$, and 100, and the location shift $\Delta = ||\mu_1 - \mu_2||_2$ takes the value 0.6, 0.8, 1.1, 1.4, 1.7, 2, and 2, respectively. Sample sizes are set to be $n_1 = n_2 = 50$. For the log-rank-type test, the set of test points D is chosen to be the observations themselves. We set $k = 100$, as a relatively large k is generally needed for estimation of hazard functions, and $\rho = 0, 0.5$, and 1. For the edge-count test, we construct 1-, 3-, and 5-MSTs, where 5-MST is recommended by Chen and Friedman (2017). We also provide the results of MMD test, energy test, and Hotelling's T-squared test. The numerical results are summarized in Table 1 and shown in Figure 1.

From the results, we see that both methods can control the Type I error well. For low-dimensional cases ($p \leq 50$), the log-rank-type test, especially the Cramer-von Mises integral-type test, achieves a higher power. As the dimension increases, the power of the edge count test becomes higher.

3.2. Scenario II: Scale Alternatives

For scale alternatives, we generate one sample from the p-dimensional multivariate normal distribution $\mathcal{N}(0, I_p)$, and the other one from $\mathcal{N}(0, \sigma^2 I_p)$. For $p = 2, 5, 10$, and 20, σ takes the value 1.4, 1.25, 1.2, and 1.15, respectively. Again, we set $n_1 = n_2 = 50$. For the log-rank-type test, the choice of D, k, and ρ remain the same as those for the previous simulation. For the edge-count test, we still construct 1-, 3-, and 5-MST. The results of MMD test and energy test are also attached. The numerical results are summarized in Table 2 and shown in Figure 2.

Based on the numerical results, we see that the Type I error is under control for both methods. While the power of the log-rank-type test is higher than that of the edge-count test in almost all cases. This is an expected result as log-rank-type tests are known to be the most powerful for scale alternatives (Hollander, Wolfe, and Chicken 2013).

3.3. Scenario III: Mixed Alternatives

In practice, data often come from complex generating mechanisms where there are multiple subpopulations. Deviation from the null hypothesis may take a different form within each of these subpopulations. Such scenarios are seen in a wide range of applications, such as image analysis, molecular biology, and statistical genetics. In this section, we consider the simulation scenario where the samples are generated under mixed alternatives. In other words, the difference between the two samples can be either location shift or scale difference for different subpopulations. Furthermore, we assume the distributional differences between the two samples only lies in the first five dimensions while the remaining dimensions are noises independently generated from the standard normal distribution for both samples.

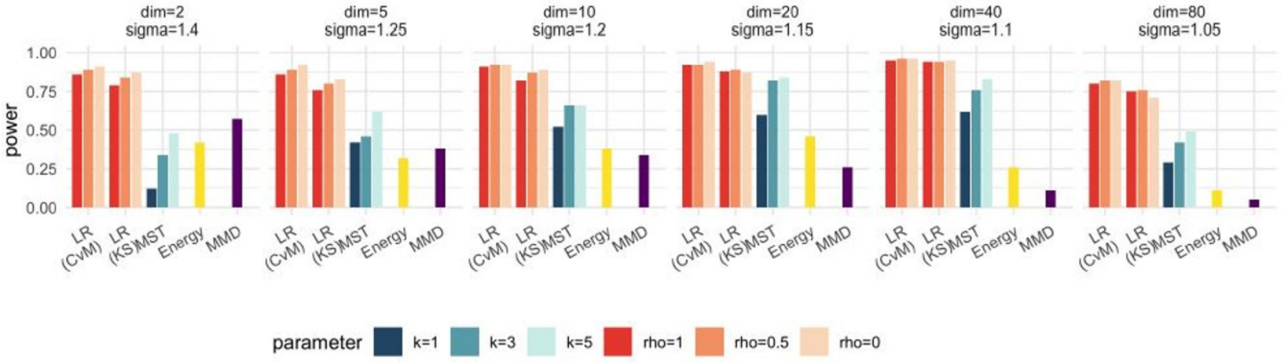

Figure 2. Power analysis under the scale alternatives scenario. Log-rank (LR) type tests using the Cramer-von Mises (C-vM) tests and the Kolmogorov–Smirnov (K–S) sup-type tests at three ρ values, are compared with the edge-count test (MST) by Chen and Friedman (2017) at three k values, the MMD test by Gretton et al. (2012) and the energy test by Aslan and Zech (2005), under four simulation settings with different numbers of dimension and scale differences. Here, power is estimated based on 400 repetitions, and for each repetition the cutoff value is obtained by 200 permutations.

Table 2. Performance of the comparison methods under the scale alternatives scenario.

		Log-rank-type test (K–S)		Log-rank-type test (C-vM)	
		$\rho = 0, 0.5, 1$		$\rho = 0, 0.5, 1$	
p	σ	Type I error	Power	Type I error	Power
2	1.4	0.05, 0.05, 0.05	0.87, 0.84, 0.79	0.06, 0.05, 0.04	0.91, 0.89, 0.86
5	1.25	0.06, 0.06, 0.04	0.83, 0.80, 0.76	0.06, 0.06, 0.06	0.92, 0.89, 0.86
10	1.2	0.06, 0.05, 0.06	0.89, 0.87, 0.82	0.05, 0.06, 0.05	0.92, 0.92, 0.91
20	1.15	0.06, 0.05, 0.06	0.87, 0.89, 0.88	0.05, 0.05, 0.04	0.94, 0.92, 0.92
40	1.1	0.06, 0.04, 0.04	0.95, 0.94, 0.94	0.05, 0.05, 0.04	0.96, 0.96, 0.95
80	1.05	0.06, 0.05, 0.03	0.71, 0.76, 0.75	0.06, 0.05, 0.04	0.82, 0.92, 0.80
		Edge-count test		Energy test	
		1-, 3-, 5-MST			
p	σ	Type I error	Power	Type I error	Power
2	1.4	0.06, 0.04, 0.06	0.12, 0.34, 0.48	0.03	0.42
5	1.25	0.05, 0.06, 0.06	0.42, 0.46, 0.62	0.06	0.32
10	1.2	0.04, 0.03, 0.04	0.52, 0.66, 0.66	0.05	0.38
20	1.15	0.06, 0.05, 0.04	0.60, 0.82, 0.84	0.04	0.46
40	1.1	0.05, 0.04, 0.06	0.62, 0.76, 0.83	0.05	0.26
80	1.05	0.05, 0.06, 0.06	0.29, 0.42, 0.49	0.05	0.11
		MMD test			
p	σ	Type I error	Power		
2	1.4	0.05	0.57		
5	1.25	0.06	0.38		
10	1.2	0.06	0.34		
20	1.15	0.04	0.26		
40	1.1	0.03	0.11		
80	1.05	0.04	0.05		

Table 3. Numerical results for the mixture distributions.

	Log-rank-type test (K–S)					
	$\rho = 0, 0.5$		$\rho = 0, 0.5$		$\rho = 0, 0.5$	
	3 clusters		5 clusters		7 clusters	
p	Type I error	Power	Type I error	Power	Type I error	Power
25	0.06, 0.06	0.88, 0.86	0.06, 0.05	0.88, 0.89	0.05, 0.05	0.90, 0.91
50	0.04, 0.06	0.85, 0.85	0.04, 0.06	0.85, 0.86	0.05, 0.05	0.86, 0.84
75	0.05, 0.06	0.82, 0.83	0.05, 0.06	0.80, 0.81	0.05, 0.06	0.84, 0.85
100	0.04, 0.04	0.79, 0.77	0.05, 0.06	0.79, 0.78	0.06, 0.05	0.83, 0.82
	Log-rank-type test (C-vM)					
	$\rho = 0, 0.5$		$\rho = 0, 0.5$		$\rho = 0, 0.5$	
	3 clusters		5 clusters		7 clusters	
p	Type I error	Power	Type I error	Power	Type I error	Power
25	0.05, 0.04	0.87, 0.86	0.07, 0.06	0.89, 0.88	0.06, 0.07	0.89, 0.86
50	0.06, 0.06	0.84, 0.82	0.06, 0.06	0.87, 0.85	0.05, 0.05	0.83, 0.83
75	0.05, 0.07	0.83, 0.82	0.05, 0.07	0.84, 0.81	0.06, 0.07	0.83, 0.82
100	0.04, 0.05	0.81, 0.79	0.05, 0.06	0.83, 0.79	0.07, 0.08	0.82, 0.80
	Edge-count test					
	5-, 25-, 50-, 75-, 100-MST					
p	Type I error	Power				
25	0.05, 0.04, 0.03, 0.05, 0.06	0.76, 0.85, 0.85, 0.86, 0.83				
50	0.06, 0.07, 0.06, 0.06, 0.07	0.71, 0.81, 0.81, 0.80, 0.77				
75	0.05, 0.06, 0.05, 0.04, 0.04	0.63, 0.77, 0.75, 0.76, 0.73				
100	0.06, 0.04, 0.04, 0.06, 0.07	0.57, 0.68, 0.69, 0.68, 0.64				
	Energy test		MMD test			
p	Type I error	Power	Type I error	Power		
25	0.05	0.84	0.06	0.81		
50	0.04	0.61	0.04	0.50		
75	0.06	0.46	0.04	0.33		
100	0.05	0.38	0.04	0.22		

The distributions of the two samples under the first five dimensions are specified as follows:

$$\pi_l(x) = 0.15\phi(x; \mu_l^{(1)}, I) + 0.65\phi(x; \mu^{(2)}, \Sigma_l^{(2)}) \\ +0.1\phi(x; \mu^{(3)}, I) + 0.1\phi(x; \mu^{(4)}, I), \quad \text{for } l = 1, 2,$$

where $\phi(\cdot; \mu, \Sigma)$ denotes the density of a normal distribution with mean μ and variance Σ, $\Sigma_1^{(2)} = I$, $\Sigma_2^{(2)} = 2.85I$, and $\mu_1^{(1)}, \mu_2^{(1)}, \mu^{(2)}, \mu^{(3)}$ and $\mu^{(4)}$ are independently generated from $\mathcal{N}(0, 2I)$. Here, we assume there are a total of four subpopulations, where the two samples differ within the first two subpopulations, a location shift for the first one and a scale difference for the second one. We generate 100 samples from each distribution. The total number of dimension increases from 25 to rev100. For the log-rank-type test, we set $k = 200$, and ρ to be 0 and 0.5. For the edge-count test, we construct 5-, 25-, 50-, 75-, and 100-MSTs. The numerical results are summarized in Table 3 and shown in Figure 3.

Here we use an alternative method to choose the set of test points D: we apply the k-means clustering on the merged samples with different choices of k. The centroids obtained after this step will be the set D at the testing step. As shown in Figure 3 and Table 3, the misspecification of the number of clusters k does not seem to affect the power of the log-rank-type test.

For some of settings under this scenario, the Type I error is slightly higher than nominal value of 0.05 for both methods, for example, 0.07. This is due to the fact that, as the dimension

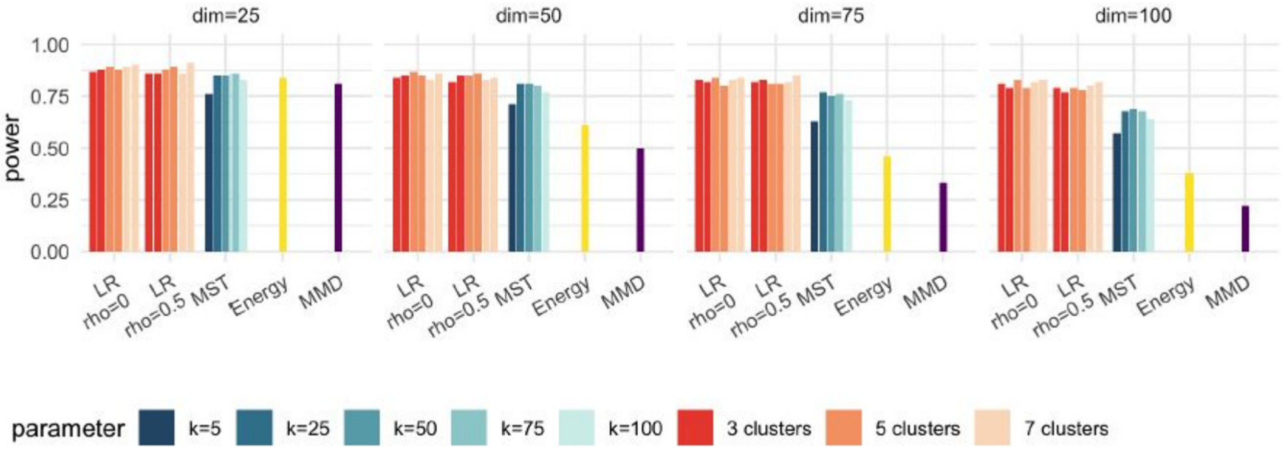

Figure 3. Power analysis under the mixed alternatives scenario. Log-rank (LR) type tests using the Cramer-von Mises (C-vM) tests and the Kolmogorov–Smirnov (K–S) sup-type tests at two ρ values and three clustering specifications are compared with the edge-count test (MST) by Chen and Friedman (2017) at four k values, the MMD test by Gretton et al. (2012), and the energy test by Aslan and Zech (2005), under four simulation settings with different numbers of noisy dimension. Here, power is estimated based on 400 repetitions, and for each repetition the cutoff value is obtained by 200 permutations.

increases, a pooled sample with 200 data points is still relatively sparse for the proposed distance based tests. The power of our method is consistently higher across all simulation settings under this scenario of mixed alternative. The advantage of the proposed methods becomes more pronounced as the number of noisy dimensions increases. A closer look at the edge-count test reveals that the number of edges connecting any two points from F_2 (denoted as R_2) varies in different ways for location and scale alternatives in high dimensions. In particular, R_2 would increase for the location alternative, while decrease for the scale alternative. When we consider a scenario that contains a mixture of both location and scale differences, the changes in R_2 can cancel with each other for the edge-count test. The proposed log-rank-type tests are able to capture the difference by consolidating evidence against the null hypothesis from multiple starting points in D. This feature of the log-rank-type test is particularly useful in practice, as the difference between two samples may exhibit a complicated pattern.

3.4. *Summary of Simulation Results*

From these three simulation studies, a few trends emerge. For the location alternatives (Figure 1 and Table 1), a larger value of ρ generally leads to a higher power. This is especially the case for the Cramer-von Mises type test. As ρ increases, observations close to points in D gain more weight in the test statistics. Therefore, it implies that in this case the inference is mainly driven by the local discrepancy. For the scale alternatives (Figure 2 and Table 2), a smaller ρ translates to more power, and the corresponding test statistic is designed to capture the global discrepancy. But overall, the performance of the proposed method is relatively stable across differ ρ values. The edge-count test, on the other hand, is more sensitive to the choice of k.

When the size of test points D is relatively large, such as the case for location and scale alternatives, the power of the Cramer-von Mises intergral-type test is, in general, slightly higher than that of the Kolmogorov–Smirnov sup-type test; while for smaller sizes of D, as under the mixed alternatives scenario, the power of the Kolmogorov–Smirnov sup-type test becomes slightly higher. A possible explanation is that when D contains more points, the variance of the test statistic for the K–S type test becomes larger, resulting in a lower power.

4. Applications

4.1. *Application to Functional Magnetic Resonance Imaging of Brain Activity Data*

The functional magnetic resonance imaging (fMRI) of brain activity dataset, previously analyzed by Rosenbaum (2005) using a minimum distance pairing approach, consists of two measure-

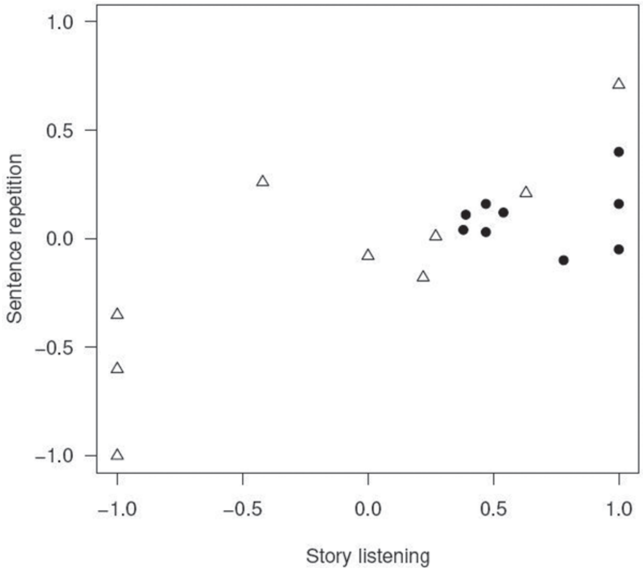

Figure 4. A scatterplot showing the laterality index for story listening and sentence repetition for patients (△) and controls (●) in the fMRI example.

Table 4. *p*-values for FMRI data.

	Cramer-von Mises (C-vM)			Kolmogorov–Smirnov (K–S)		
	$\rho = 0$	$\rho = 1$	$\rho = 2$	$\rho = 0$	$\rho = 1$	$\rho = 2$
$k = 9$	0.0196	0.0230	0.0258	0.0292	0.0292	0.0384
$k = 18$	0.0164	0.0172	0.0204	0.0050	0.0102	0.0266

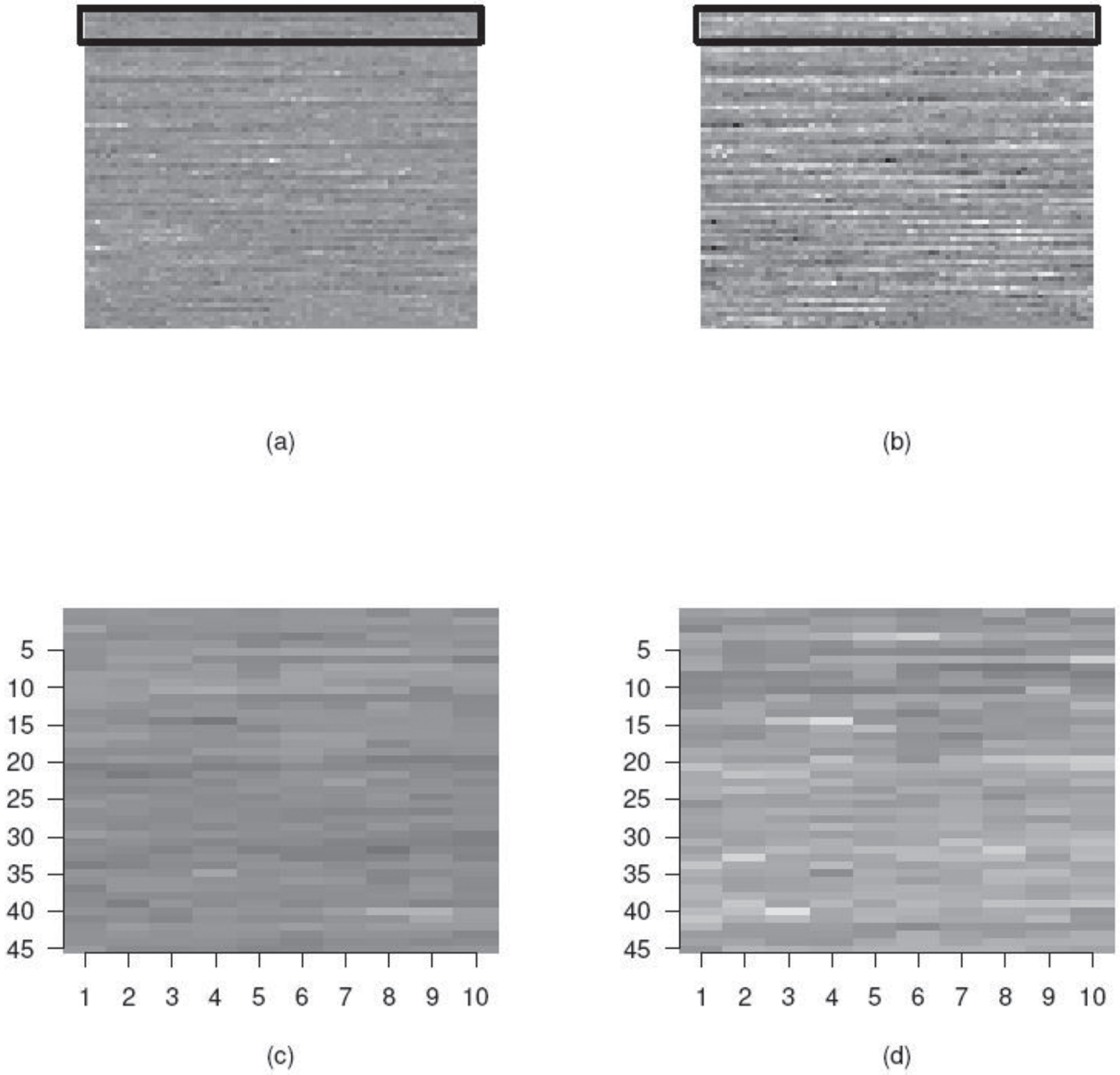

Figure 5. Graphs showing grayscale maps of mean gene expressions for (a) 5153 genes in cancer population, (b) 5153 genes in control population, (c) 450 genes in cancer population, (d) 450 genes in control population.

ments on each of the 18 subjects. All eighteen subjects are right handed. Half of them have arteriovenous malformations in the left hemisphere and the other half are normal controls. The two measurements are laterality indices calculated for listening to a story and repeating a sentence mentally after listening to it, respectively. When the subject is listening to a story or repeating a sentence while undergoing fMRI, the number of activated pixels in both the left and the right hemisphere's temporal lobe, L and R, are recorded. For each task, the laterality index is calculated as $(L-R)/(L+R)$; see Lehéricy et al. (2002) for more details about this dataset. The laterality index is a continuous variable measuring the relative activation of the left and the right hemispheres during the language tasks. In some extreme cases, the laterality index is 1 if all the increased activation is on the left, -1 if all is on the right and 0 if the activation on both sides is the same.

Figure 4 is a scatterplot of the laterality index for sentence repetition against the laterality index for story listening, with patients indicated by a triangle and controls indicated by a bullet. The primary goal of the study is to examine whether the impaired left hemisphere affects the performance of language tasks. To this end, we implement the weighted log-rank approach to compare two groups in the two-dimensional space. Due to the small sample size, we use a fixed D constructed as $\{(-1+0.4i, -1+0.4j), i, j = 1, \dots, 5\}$ and 5000 permutations to evaluate the p-values.

Table 4 displays the p-values corresponding to the weighted log-rank approach with variable combinations of the censoring time k and weight ρ. All p-values are less than 0.05, so the null hypothesis that the distributions of the laterality indices for two groups are the same, is implausible. This conclusion agrees with that in Rosenbaum (2005), where the p-value was found to be 0.0259.

4.2. Application to Prostate Cancer DNA Microarray Data

Worldwide, prostate cancer is the third most common cancer and the cause of 6% of cancer deaths in men (Parkin, Bray, and Devesa 2001). In the United States, it is the most frequently diagnosed and the second leading cause of cancer death in men (Jemal et al. 2003). A first step in better understanding the prostate cancer is to test whether genes are expressed differ-

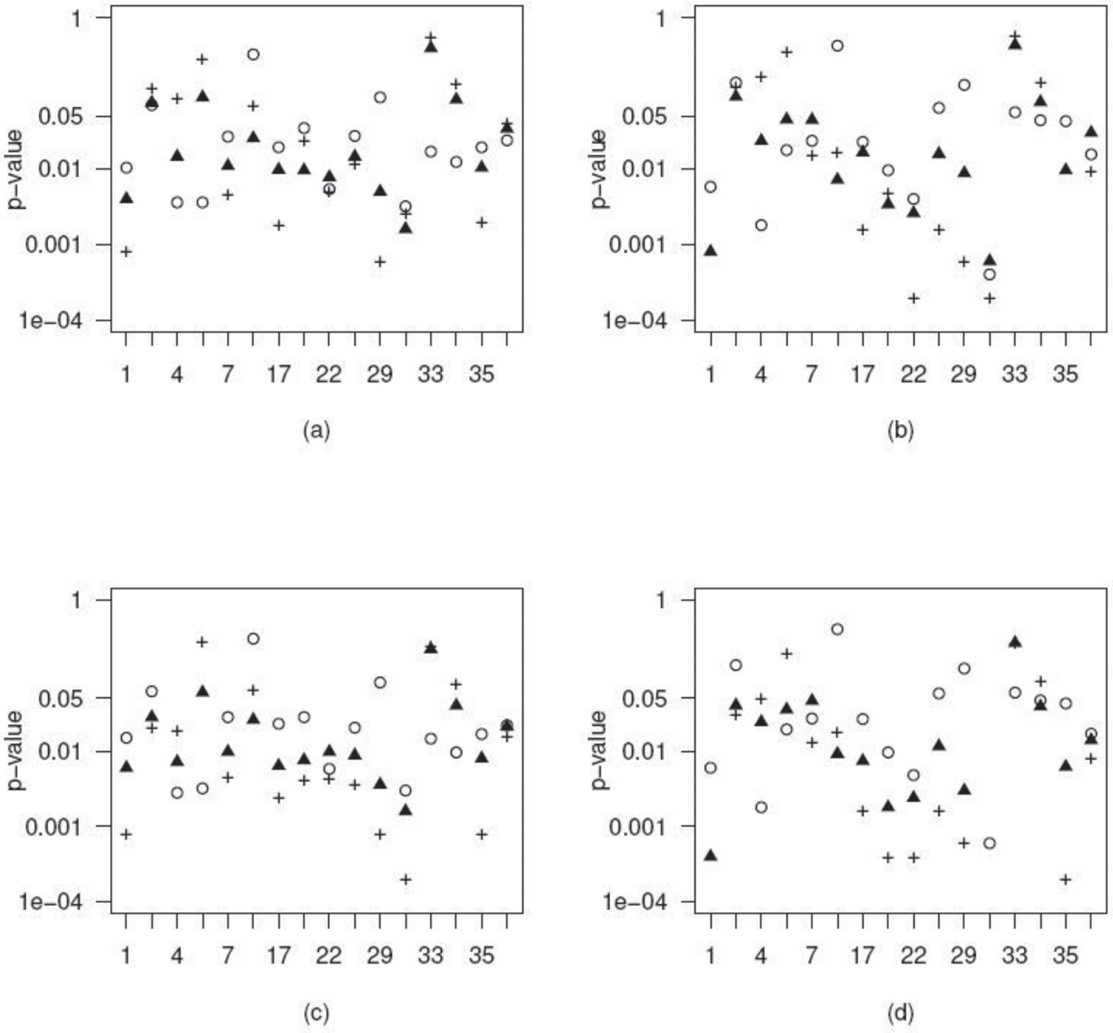

Figure 6. Graphs showing p-values for (a) the Cramer-von Mises (C-vM) test when $\rho = 0$, (b) the Kolmogorov–Smirnov (K–S)) test when $\rho = 0$, (c) the Cramer-von Mises (C-vM) test when $\rho = 1$ and (d) the Kolmogorov–Smirnov (K–S) test when $\rho = 1$ (○ : $k = 20$, ▲ : $k = 50$, + : $k = 103$). The p-values are plotted after a $\log_{10}$ transformation.

entially in tumor compared to nontumor samples. To explore potential molecular variation, DNA microarray profiling studies are conducted. As a new multiplex technology, DNA microarray can contain tens of thousands of probes, leading to a high dimensional problem. To illustrate the proposed method, we study a case–control prostate cancer DNA microarray dataset (Lapointe et al. 2004), which is available at *http://microarray-pubs-stanford.edu/prostateCA/*. The prostate cancer dataset consists of 62 primary prostate tumors (61 adenocarcinomas and one adenoid cystic tumor) and 41 matched normal prostate tissues (from the noncancerous region of the prostate). For each tissue, the expressions of 5153 genes were measured. So we have $62 + 41 = 103$ samples in a 5153-dimensional space. The upper part of Figure 5 displays the grayscale maps of mean expression value of each gene for tumor and nontumor groups, respectively. We realign the 5153 genes in a 72×72 matrix by row according to the percentage of missing data, from low to high, that is, each row from left to right, the percentage of missing data is nondecreasing. The last 31 elements in the last row of the matrix are set to zero just to make up a square matrix. Throughout the following analysis, we use the permutation tests to obtain the cutoff values.

First we take out the 450 genes without any missing data and apply the weighted log-rank approach in a 450-dimensional space with D composed of two 450-vectors, each being the mean expression values within the respective group. The null hypothesis of no difference in distribution is rejected at levels well below 10^{-3} for all reasonable combinations of ρ and k. Then we narrow down the attention to shorter segments of genes. Specifically, we reshape the 450 genes into a 45×10 matrix (see the bottom part of Figure 5) and each time we test for group difference for one row, that is, in a 10-dimensional space. Among the 45 rows, there are 29 rows with extremely small p-values and thus in Figure 6, we only display the p-value for the remaining 16 rows (1, 2, 4, 5, 7, 15, 17, 18, 22, 24, 29, 32, 33, 34, 35, 37).

Note that the patterns of the four graphs are quite similar, which shows the consistency of the procedure. For some rows, the p-value can drop tremendously with more samples. For row 29, censoring at the 20th observation from the mean expression value cannot reject the null hypothesis at 5% level while including 50 or 103 samples can detect the difference. This might suggest the distributions for those rows are quite similar around the mean expression values but differ from each other over the places far away from the means in a 10-dimensional Euclidean space. For some other rows, for example, row 4, censoring at the 20th observation may give the smallest p-value, which might suggest a discrepancy in a local neighborhood of the mean expression values. In contrast, the three k values give similar p-values for some rows. For instance, row 37. This may imply a more uniform distribution over the 10-dimensional space.

5. Discussion

The powerful class of nonparametric weighted log-rank tests is used to test the difference between multivariate distributions in high dimensions. This is done by converting the original observations, possibly in a high dimensional space, into “survival times.” Such an approach reduces greatly the complexity of the original problem. It also allows the application of available tools, including software packages for implementation and large sample properties for theoretical justification. Because the choice of the weight function is intuitive and well understood in survival analysis, it may be possible to use the intuition from survival analysis to gain insights into which weight function or functions should be used in the original multivariate data. Additionally, different weight functions may be combined to achieve optimal testing efficiency and robustness, as in the work by Gastwirth (1985).

As a simple example, we take Type II censoring to illustrate the flexibility of the proposed method. This type of censoring may be viewed as a special kind of weight functions. By varying the number of failure times to be included, this type of censoring magnifies “local” group differences as opposed to more global differences. In this connection, general weight functions provide even greater flexibility.

The usual asymptotic properties are established for the proposed tests. In particular, we establish the weak convergence of the basic processes under the null and contiguous alternative hypotheses. The weak convergence is then used to derive limiting distributions for the test statistics which are functionals of the basic processes. The limiting distributions under contiguous alternatives may shed some light on the asymptotic efficiency of the proposed tests in some simple cases.

Converting spatial points to distances is a crucial component of the proposed method. Instead of Euclidean space, one may consider a general metric space. In particular, the method is readily applicable to functional data when a suitable metric can be introduced on the corresponding function space. Modern empirical process theory is also applicable to deriving asymptotic properties.

The use of the proposed approach for high dimensional problems is of particular importance. We believe that asymptotic properties can be extended to very high dimensional problem, including those with p being larger than n. We further note that the approach may also be extended to other kinds of high dimensional learning problems. We are currently investigating the use of the survival analysis approach in classification problems in high dimensional spaces.

Appendix: Proofs

Proofs of the theorems given in Section 2 are outlined here.

Proof of Theorem 2.2. Under the null hypothesis, for each fixed x, $T_i(x), i = 1, \dots, n$, are iid random variables. We can write

$$U_w(x;t) = n^{-1/2} \sum_{i=1}^{n} \int_0^t W_x(s) \left(Z_i - \frac{\widehat{\Gamma}_1(x;s)}{\widehat{\Gamma}_0(x;s)} \right) dM_i(x;s),$$

where $\Lambda(x;s)$ is the common cumulative hazard function of $T_i(x)$, by noting that

$$\sum_{i=1}^{n} \left(Z_i - \frac{\sum_{l=1}^{n} Z_l I(T_l(x) \geq s)}{\sum_{l=1}^{n} I(T_l(x) \geq s)} \right) I(T_i(x) \geq s) = 0,$$

for all x and t.

Next, we can apply the functional central limit theorem (Pollard 1990, Theorem 10.6) to show that $U_n^1(x;t) = n^{-1/2} \sum_{i=1}^{n} M_i(x;t)$ and $U_n^2(x;t) = n^{-1/2} \sum_{i=1}^{n} Z_i M_i(x;t)$ converge weakly to their respective limiting Gaussian processes. In particular, conditions (ii) and (v) thereof are satisfied as (X_i, Z_i)’s can be viewed as iid random variables. Conditions (iii) and (iv) hold because envelopes can be chosen to be $B^*/\sqrt{n}$ for some constant B^*. The processes are also manageable (condition (i)) as $M_i(x,t)$ involves the indicator function and a smooth distance function.

Let

$$R(x;t) = \int_0^t W_x(s) \left(\frac{\widehat{\Gamma}_1(x;s)}{\widehat{\Gamma}_0(x;s)} - \frac{\Gamma_1(x;s)}{\Gamma_0(x;s)} \right) \left(n^{-1/2} \sum_{i=1}^{n} dM_i(x,s) \right).$$

Since

$$\sup_{(x,t)\in\mathcal{R}} \left| \frac{\widehat{\Gamma}_1(x;s)}{\widehat{\Gamma}_0(x;s)} - \frac{\Gamma_1(x;s)}{\Gamma_0(x;s)} \right| \to 0 \quad \text{a.e.},$$

we can apply Lemma A.3 and the argument for eq. (2.7) in Bilias, Gu, and Ying (1997) to show $\sup_{(x,t)\in\mathcal{R}} |R(x;t)| = o_p(1)$.

Now we have

$$\begin{aligned} U_w(x;t) - R(x;t) &\\ &= n^{-1/2} \sum_{i=1}^{n} \int_0^t W_x(s) \left(Z_i - \frac{\Gamma_1(x;s)}{\Gamma_0(x;s)} \right) dM_i(x;s) \\ &= \underbrace{\int_0^t W_x(s) \left(n^{-1/2} \sum_{i=1}^{n} Z_i dM_i(x;s) \right)}_{\text{Term 1}} \\ &\quad + \underbrace{\int_0^t W_x(s) \cdot \frac{\Gamma_1(x;s)}{\Gamma_0(x;s)} \left(n^{-1/2} \sum_{i=1}^{n} dM_i(x;s) \right)}_{\text{Term 2}}. \end{aligned} \tag{3}$$

As $E(M_i(x;s)|Z_i) = 0$, the summands of (3) are mean zero iid random variables under the null hypothesis. Thus, by the classical multivariate central limit theorem, $U_w(\cdot;\cdot)$ converges in finite dimensional distributions to a Gaussian random field, whose covariance function is given by $C(x_1, t_1; x_2, t_2)$ in Theorem 2.2.

It remains to show the “tightness” of $U_w(\cdot;\cdot)$, which will be done by showing tightness of both Terms 1 and 2. The weight function $W_x(\cdot)$ has bounded variation under Condition 2. Together with weak convergence of $U_n^1(x;t)$ and $U_n^2(x;t)$, using a similar argument as that for Theorem 2.1 in Bilias, Gu, and Ying (1997), we can show both of them are tight.

Proof of Theorem 2.5. Conditioning on the data $\{(X_i, Z_i)\}_{i=1}^n$, $\{V_i\}_{i=1}^n$ are the only random components in U_w^*. Since V_i are generated to be iid standard normals and independent of the data, it follows from the central limit theorem and a straightforward covariance calculation that U_w^* converges in finite dimensional distributions to a zero-mean Gaussian process with the same covariance function $C(x_1, t_1; x_2, t_2)$ as defined in Theorem 2.2. Thus, it suffices to prove the tightness of U_w^*, which can be obtained in a similar way as that for Theorem 2.2; see Lin et al. (2000) for arguments in a more general case.

Proof of Theorem 2.8. Let $X \sim F_1$ and $Y \sim F_2$. When $F_1 \neq F_2$, for p-dimensional cubes of the form $\{(x_1, \dots x_p) : x_i \in [a_i, b_i] \text{ with } a_i < b_i\}$, we can find at least one cube, denoted as A, such that $P(X \in A) \neq P(Y \in A)$. Otherwise the two distributions will be the same. We may also assume the two probabilities are bounded away from 1. Then $\mathcal{R}$ can be chosen as $A \times [0, \text{diam}(A)] \cap \{(x, t) : P(T_1(x) > t) > \gamma\}$ for some small γ such that for any $x \in A$ there exist a $t > 0$ with $(x, t) \in \mathcal{R}$. Given x, let $t_x = \sup\{t : (x, t) \in \mathcal{R}\}$. For any x at which $d(X, x)$ and $d(Y, x)$ follow different distributions in the sense that there exists $t' \leq t_x$ such that $P(d(X, x) \leq t') \neq P(d(Y, x) \leq t')$, we can find a $t \leq t_x$ making the weighted log-rank test consistent. Therefore, it suffices to show that if X and Y differ in distribution, such an $x \in A$ exists. We prove by contradiction. Suppose such an x does not exist, then for any $\Delta > 0$ small enough,

$$\frac{P(d(X, x) \leq \Delta)}{\Delta^p} = \frac{P(d(Y, x) \leq \Delta)}{\Delta^p}.$$

Letting Δ go to zero on both sides, we know that $f_X(x) = f_Y(x)$ at any $x \in A$, where f_X and f_Y are respective densities. This contradicts with the assumption.

Proof of Theorem 2.10. By Le Cam's third lemma (Hájek, Šidák, and Sen 1999), $U_w(x; t)$ converges in finite dimensional distribution to $G_Q(x; t)$ under Q_n. Moreover, we can show that $\left(U_w(x; t), \log \frac{dQ_n}{dP_n}\right)$ is tight. The tightness of $U_w(x; t)$ has been obtained in the proof of Theorem 2.2, and $\log \frac{dQ_n}{dP_n}$ converges in distribution to a normal distribution, as it is scaled summation of iid random variables.

Theorems 2.6 and 2.7 are direct generalizations of Theorems 2.2 and 2.5, respectively. And Corollaries 2.3 and 2.11 can be shown by applying continuous mapping theorem to Theorems 2.2 and 2.10.

Supplementary Materials

Computational codes for the numerical experiments presented in this article.

ORCID

Tian Zheng http://orcid.org/0000-0003-4889-0391

References

Anderson, P. K., Borgan, Ø., Gill, R. D., and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, Springer Series in Statistics, New York: Springer. [1385,1386]

Anderson, T. W., and Darling, D. A. (1954), "A Test of Goodness of Fit," *Journal of the American Statistical Association*, 49, 765–769. doi: 10.1080/01621459.1954.10501232. [1386]

Aslan, B., and Zech, G. (2005), "New Test for the Multivariate Two-sample Problem Based on the Concept of Minimum Energy," *Journal of Statistical Computation and Simulation*, 75, 109–119. doi: 10.1080/00949650410001661440. [1384,1388,1389,1390,1391]

Bilias, Y., Gu, M., and Ying, Z. (1997), "Towards a General Asymptotic Theory for Cox Model with Staggered Entry," *The Annals of Statistics*, 25, 662–682. doi: 10.1214/aos/1031833668. [1394]

Chen, H., and Friedman, J. H. (2017), "A New Graph-Based Two-Sample Test for Multivariate and Object Data," *Journal of the American Statistical Association*, 112, 397–409. doi: 10.1080/01621459.2016.1147356. [1384,1388,1389,1390,1391]

Chen, H., Chen, X., and Su, Y. (2018), "A Weighted Edge-Count Two-Sample Test for Multivariate and Object Data," *Journal of the American Statistical Association*, 113, 1146–1155. doi: 10.1080/01621459.2017.1307757. [1384]

Friedman, J. H., and Rafsky, L. C. (1979), "Multivariate Generalizations of the Wald-Wolfowitz and Smirnov Two-sample Tests," *The Annals of Statistics*, 7, 697–717. doi: 10.1214/aos/1176344722. [1384]

Gastwirth, J. L. (1985), "The Use of Maximin Efficiency Robust Tests in Combining Contingency Tables and Survival Analysis," *Journal of the American Statistical Association*, 80, 380–384. [1394]

Gehan, E. A. (1965), "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples," *Biometrika*, 52, 203–223. [1385]

Gill, R. D. (1980), *Censoring and Stochastic Integrals*. Mathematical Centre tracts. Amsterdam: Mathematisch Centrum. [1385]

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012), "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, 13, 723–773. [1384,1388,1389,1390,1391]

Hájek, J., Šidák, Z., and Sen, P. K. (1999), *Theory of Rank Tests*, New York: Elsevier. [1384,1387,1388,1395]

Hall, P., and Tajvidi, N. (2002), "Permutation Tests for Equality of Distributions in High-Dimensional Settings," *Biometrika*, 89, 359–374. [1384]

Harrington, D. P., and Fleming, T. R. (1982), "A Class of Rank Test Procedures for Censored Survival Data," *Biometrika*, 69, 553–566. [1385]

Henze, N. (1988), "A Multivariate Two-sample Test Based on the Number of Nearest Neighbor Type Coincidences," *The Annals of Statistics*, 16, 772–783. doi: 10.1214/aos/1176350835. [1384]

Hollander, M., Wolfe, D. A., and Chicken, E. (2013), *Nonparametric Statistical Methods*, Wiley Series in Probability and Statistics, Hoboken, NJ: Wiley. [1384,1389]

Jemal, A., Murray, T., Samuels, A., Ghafoor, A., Ward, E., and Thun, M. J. (2003), "Cancer Statistics, 2003," *CA: A Cancer Journal for Clinicians*, 53, 5–26. doi: 10.3322/canjclin.53.1.5. [1392]

Lapointe, J., Li, C., Higgins, J. P., van de Rijn, M., Bair, E., Montgomery, K., Ferrari, M., Egevad, L., Rayford, W., Bergerheim, U., Ekman, P., DeMarzo, A. M., Tibshirani, R., Botstein, D., Brown, P. O., Brooks, J. D., and Pollack, J. R. (2024), "Gene Expression Profiling Identifies Clinically Relevant Subtypes of Prostate Cancer," *Proceedings of the National Academy of Sciences*, 101, 811–816. doi: 10.1073/pnas.0304146101. [1393]

Lehéricy, S., Biondi, A., Sourour, N., Vlaicu, M., du Montcel, S. T., Cohen, L., Vivas, E., Capelle, L., Faillot, T., Casasco, A., Le Bihan, D., and Marsault, C. (2002), "Arteriovenous Brain Malformations: Is Functional MR Imaging Reliable for Studying Language Reorganization in Patients? Initial Observations," *Radiology*, 223, 672–682. doi: 10.1148/radiol.2233010792. [1392]

Lehmann, E. L., and Romano, J. P. (2005), "Generalizations of the Familywise Error Rate," *The Annals of Statistics*, 33, 1138–1154. doi: 10.1214/009053605000000084. [1384]

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000), "Semiparametric Regression for the Mean and Rate Functions of Recurrent Events," *Journal of the Royal Statistical Society*, Series B, 62, 711–730. [1395]

Lin, D. Y., Wei, L. J., and Ying, Z. (2002), "Model-Checking Techniques Based on Cumulative Residuals," *Biometrics*, 58, 1–12. doi: 10.1111/j.0006-341X.2002.00001.x. [1387]

Mantel, N. (1966), "Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration," *Cancer Chemotherapy Reports*, 50, 163–170. [1384]

Mantel, N., and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease," *JNCI: Journal of the National Cancer Institute*, 22, 719–748. doi: 10.1093/jnci/22.4.719. [1384]

Parkin, D. M., Bray, F. I., and Devesa, S. S. (2001), "Cancer Burden in the Year 2000. The Global Picture," *European Journal of Cancer*, 37, S4–S66. doi: 10.1016/s0959-8049(01)00267-2. [1392]

Peto, R., and Peto, J. (1972), "Asymptotically Efficient Rank Invariant Test Procedures," *Journal of the Royal Statistical Society*, Series A, 135, 185–207. [1385]

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, Conference Board of the Mathematical Science: NSF-CBMS Regional Conference Series in Probability and Statistics, Hayward, CA: Institute of Mathematical Statistics. [1385,1394]

Prentice, R. L. (1978), "Linear Rank Tests with Right Censored Data," *Biometrika*, 65, 167–179. [1385]

Rosenbaum, P. R. (2005), "An Exact Distribution-Free Test Comparing Two Multivariate Distributions Based on Adjacency," *Journal of the Royal Statistical Society*, Series B, 67, 515–530. doi: 10.1111/j.1467-9868.2005.00513.x. [1384,1391,1392]

van der Vaart, A., and Wellner, J. (1996), "*Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics, New York: Springer-Verlag. [1385]