Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2007) 4337–4353
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

A data-adaptive methodology for finding an optimal weighted generalized Mann–Whitney–Wilcoxon statistic☆

Majnu John[a,1], Carey E. Priebe[b,*]

[a]*Department of Medicine, Johns Hopkins School of Medicine, Baltimore, MD 21205, USA*
[b]*Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, MD 21218, USA*

Received 15 November 2005; received in revised form 25 May 2006; accepted 2 June 2006
Available online 27 June 2006

Abstract

Xie and Priebe [2002. "Generalizing the Mann–Whitney–Wilcoxon Statistic". J. Nonparametric Statist. 12, 661–682] introduced the class of weighted generalized Mann–Whitney–Wilcoxon (WGMWW) statistics which contained as special cases the classical Mann–Whitney test statistic and many other nonparametric distribution-free test statistics commonly used for the two-sample testing problem. The two-sample test that they proposed was based on any statistic within the class of WGMWW statistics optimal in the Pitman asymptotic efficacy (PAE) sense. In this paper, among other things, we show via simulation studies that for finite samples the PAE-optimal WGMWW test has substantially higher empirical power compared to the classical Mann–Whitney test for various underlying densities (especially for those densities for which Mann–Whitney test is considered a better alternative to parametric tests such as *t*-tests). The PAE-optimal WGMWW test is not a candidate for the practitioner's toolbox since the corresponding test statistic contains parameters which are functions of the underlying null distribution function of the samples. The main thrust of this paper is in introducing a data-adaptive alternative to the PAE-optimal WGMWW test, which has efficacy and power as good as the latter. We provide an estimate $\hat{\psi}$ for the PAE function ψ of a WGMWW statistic, and our test is based on a $\hat{\psi}$-optimal WGMWW statistic. We prove strong consistency of $\hat{\psi}$, thereby showing that our test has approximately the same efficacy as the ψ-optimal WGMWW test for large sample sizes. Via simulation studies we show that for finite samples the empirical power of $\hat{\psi}$-optimal WGMWW test is almost the same as ψ-optimal WGMWW test for various underlying densities. We also analyze magnetic imaging data related to subjects with and without Alzheimer's disease to illustrate our methodology. In summary, we present a strong competitor for the classical Mann–Whitney–Wilcoxon test and many other existing nonparametric distribution-free tests, especially for moderate and large samples.

Keywords: Nonparametric tests; Two sample problem; Mann–Whitney–Wilcoxon; Pitman asymptotic efficacy; Asymptotic power

☆ This research is based on the first author's Ph.D. Dissertation prepared at the Johns Hopkins University.

* Corresponding author. Tel.: +1 410 516 7200; fax: +1 410 516 7459.
E-mail addresses: mjohn@jhmi.edu (M. John), cep@jhu.edu (C.E. Priebe).

[1] The author gratefully thanks his advisor Carey Priebe for his introduction to the problem and his strong encouragement.

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.06.003

1. Introduction

One of the central themes of nonparametric testing theory is the two-sample problem. The famous Mann–Whitney test, equivalent to the Wilcoxon rank sum test, is a solution of the two-sample problem, and is considered nowadays as one of the breakthroughs of 20th century statistics. Xie and Priebe (2002) proposed a new solution of the two-sample problem, and their test has higher efficacy than most other nonparametric two-sample tests existing in the literature, including the Mann–Whitney test. The gain in efficacy for their test compared to the classical Mann–Whitney test was seen to be phenomenal in some cases (e.g., when the underlying density is strongly skewed) and was seen to be substantial in other cases (e.g., when the underlying density is either asymmetric bimodal or heavily kurtotic). But their test was not data-adaptive, since the test statistic that they proposed had parameters which were functions of the unknown underlying distribution function (i.e., the distribution function of both the samples under the null hypothesis). In this paper, we propose a methodology which makes the two-sample solution proposed by Xie and Priebe (2002) data-adaptive. We show via theoretical results and simulation studies that the test statistic proposed in our data-adaptive methodology has approximately the same efficacy and power as the one proposed in Xie and Priebe (2002), especially for large sample sizes.

The paper is organized as follows. In this introductory section we present background and our data-adaptive methodology. In Section 2, we present our main theoretical results which justify our methodology. The third section contains the numerical simulation results, and Section 4 contains results related to analysis of magnetic resonance imaging data from normally aging subjects and patients with dementia of the Alzheimer type.

1.1. The classical Mann–Whitney–Wilcoxon statistic

Consider two i.i.d. samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from two possibly different populations, with underlying (continuous) distribution functions F and G, respectively. In order to test

$$\mathrm{H}_0 : F = G \text{ vs.}$$

$$\mathrm{H}_\mathrm{A} : F(x) \geqslant G(x) \quad \forall x, \text{ with strict inequality for at least one } x$$

(i.e., stochastic ordering), we may use the classical Mann–Whitney–Wilcoxon (MWW) statistic (Wilcoxon, 1945; Mann and Whitney, 1947) which, in its U-statistic form, is given by

$$\frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} I\left(X_i \leqslant Y_j\right). \tag{1}$$

Here $I(\cdot)$ is the indicator function. Let $F_n(x) = (1/n)\sum_{i=1}^{n} I(X_i \leqslant x)$ and $G_m(x) = (1/m)\sum_{i=1}^{m} I(Y_i \leqslant x)$ be the empirical distribution functions corresponding to F and G, respectively. The statistic in (1), which may also be written as

$$\int_{-\infty}^{\infty} F_n(x)\, \mathrm{d}G_m(x)$$

is an empirical estimate of the functional

$$\int_{-\infty}^{\infty} F(x)\, \mathrm{d}G(x).$$

The classical MWW statistic is robust, in the sense that it is distribution free. That is, the null distribution of (1) does not depend on F. But, generally speaking, it has lesser efficacy compared to many other statistics commonly used to test stochastic ordering. (Throughout this paper, by efficacy we mean Pitman asymptotic efficacy (PAE), Lehmann, 1998; Pitman, 1979.) We can identify distribution functions for which the test based on the classical MWW is inferior to the best parametric tests. For example, if the underlying distributions are both normal, the Student t-test is superior to the MWW statistic in the sense of PAE (Lehmann, 1998).

1.2. A class of statistics with higher efficacy

Xie and Priebe (2002) introduced a class $\mathscr{C}_w^{r,s}$ of weighted generalized Mann–Whitney–Wilcoxon (WGMWW) statistics to test stochastic ordering. This class contains an important subclass $\mathscr{C}_g^{r,s}$ of unweighted generalized Mann–Whitney–Wilcoxon (GMWW) statistics presented in Xie and Priebe (2000). In particular, this inclusion implies that $\mathscr{C}_w^{r,s}$ can have higher efficacy than $\mathscr{C}_g^{r,s}$, in the sense of maximum PAE attainable by statistics within each class. The unweighted generalized version, GMWW, included as special cases the classical MWW, the subsample median statistic (Shetty and Govindarajulu (1988); Kumar (1997)), the subsample maxima statistic (Kochar (1978); Deshpande and Kochar (1980); Stephenson and Ghosh (1985); Ahmad (1996); Adams et al. (2000)) and the subsample minima statistic (Priebe and Cowen, 1999). The statistics in the class $\mathscr{C}_w^{r,s}$ are based on the functional approximation

$$\int_{-\infty}^{\infty} u_r(F(x))\,\mathrm{d}v_s(G(x)) \tag{2}$$

to the functional

$$\int_{-\infty}^{\infty} u(F(x))\,\mathrm{d}v(G(x)). \tag{3}$$

Here, u and v are arbitrary increasing, continuous, real-valued functions on $[0, 1]$. u_r and v_s, respectively, denote the Bernstein polynomials of order r and s (Lorentz, 1986) corresponding to u and v, and r and s are fixed positive integers. Expressing the Bernstein polynomials as weighted sums of tail-binomial polynomials, we may write

$$u_r(\cdot) = \sum_{k=1}^{r} \pi_k b_{k:r}(\cdot), \quad v_s(\cdot) = \sum_{l=1}^{s} \mu_l b_{l:s}(\cdot),$$

where

$$\sum_{k=1}^{r} \pi_k = 1 = \sum_{l=1}^{s} \mu_l \quad \text{and} \quad \pi_k \geqslant 0,\ \mu_l \geqslant 0 \ \text{ for } k = 1, \ldots, r,\ l = 1, \ldots, s. \tag{4}$$

In this case, (2) becomes

$$\sum_{k=1}^{r}\sum_{l=1}^{s} \pi_k \mu_l \Pr\left(X_{k:r} < Y_{l:s}\right). \tag{5}$$

We use the notation $X_{k:r}$ and $X_{k:r}\left(X_{i_1}, \ldots, X_{i_r}\right)$ to denote the kth order statistic in an arbitrary subsample $X_{i_1}, \ldots, X_{i_r}$ of $X_1, \ldots, X_n$ (chosen without replacement). The statistics in the class $\mathscr{C}_w^{r,s}$ are empirical estimates of (5) of the following form:

$$\frac{1}{\binom{n}{r}\binom{m}{s}} \sum_{C} \sum_{k=1}^{r}\sum_{l=1}^{s} \pi_k \mu_l I\left(X_{k:r}\left(X_{i_1}, \ldots, X_{i_r}\right) < Y_{l:s}\left(Y_{j_1}, \ldots, Y_{j_s}\right)\right), \tag{6}$$

where $\sum_C$ extends over all $(r+s)$-tuples of indices $1 \leqslant i_1 < \cdots < i_r \leqslant n$ and $1 \leqslant j_1 < \cdots < j_s \leqslant m$, and $\pi = (\pi_1, \ldots, \pi_r)$, $\mu = \left(\mu_1, \ldots, \mu_s\right)$ are arbitrary weights satisfying (4). Let us denote the statistic in (6) by $\delta_{n,m}^{(\pi,\mu)}$ and the functional in (5) by $\delta^{(\pi,\mu)}$. If we assume that as $n, m \to \infty$, $(n/(n+m)) \to \lambda \in (0, 1)$, then Theorem 2.2 in Xie and Priebe (2002) states that $\sqrt{m+n}\left(\delta_{n,m}^{(\pi,\mu)} - \delta^{(\pi,\mu)}\right)$ is asymptotically normal with mean 0 and variance $\sigma_{\pi,\mu}^2$, where

$$\sigma_{\pi,\mu}^2 = \frac{r^2}{\lambda}\,\mathrm{var}\left\{\int_{-\infty}^{X_1} \sum_{k=1}^{r}\sum_{l=1}^{s} \pi_k \mu_l \left(F_{k:r-1} - F_{k-1:r-1}\right)(x)\,\mathrm{d}G_{l:s}(x)\right\}$$
$$+ \frac{s^2}{(1-\lambda)}\,\mathrm{var}\left\{\int_{-\infty}^{Y_1} \sum_{k=1}^{r}\sum_{l=1}^{s} \pi_k \mu_l \left(G_{l:s-1} - G_{l-1:s-1}\right)(x)\,\mathrm{d}F_{k:r}(x)\right\}.$$

Under H_0, the variance $\sigma^2_{\pi,\mu}$ reduces to $\xi(\pi, \mu, \lambda)/(\lambda(1-\lambda))$, where

$$\begin{aligned}\xi(\pi, \mu, \lambda) = &\left(r^2\lambda + s^2(1-\lambda)\right) \sum_{k=1}^{r}\sum_{l=1}^{s}\sum_{i=1}^{r}\sum_{j=1}^{s} \pi_k\pi_i\mu_l\mu_j \frac{\binom{k+l-2}{l-1}\binom{r+s-k-l}{r-k}}{\binom{r+s-1}{s}} \\ &\times \frac{\binom{i+j-2}{i-1}\binom{r+s-i-j}{r-i}}{\binom{r+s-1}{s}} \sum_{p=k+l-1}^{r+s-1}\sum_{q=i+j-1}^{r+s-1} \binom{r+s-1}{p}\binom{r+s-1}{q} \\ &\times \frac{(p+q)!(2r+2s-p-q-2)!}{(2r+2s-1)!} \\ &- \left\{\frac{\sum_{k=1}^{r}\sum_{l=1}^{s}\pi_k\mu_l\binom{k+l-2}{k-1}\binom{r+s-k-l+1}{r-k}(r-l+1)}{\binom{r+s}{s}}\right\}^2. \end{aligned} \tag{7}$$

Note that $\xi(\pi, \mu, \lambda)$ does not depend on F. If we assume that F has a density function f, the PAE for statistics of the form (6) is given by

$$\psi(\pi, \mu, \lambda) = \frac{\varphi^2(\pi, \mu)}{\xi(\pi, \mu, \lambda)} \tag{8}$$

where

$$\varphi(\pi, \mu) = \sum_{k=1}^{r}\sum_{l=1}^{s} \frac{r!s!\pi_k\mu_l \int_{-\infty}^{\infty} F(x)^{k+l-2}(1-F(x))^{r+s-k-l} f^2(x)\,\mathrm{d}x}{(k-1)!(r-k)!(l-1)!(s-l)!}, \tag{9}$$

and $\xi(\pi, \mu, \lambda)$ is as defined above. As stated in Theorem 3.2 of Xie and Priebe (2002), there exists a PAE-optimal statistic in the class $\mathscr{C}_w^{r,s}$ provided that F is a distribution with finite Fisher information. In other words, for such F there exist π and μ satisfying (4) which maximize the PAE function, $\psi(\pi, \mu)$. Throughout this paper, we assume that F has finite Fisher information.

1.3. A data-adaptive version of the PAE-optimal WGMWW statistic

Although, as noted in the previous subsection, at least one PAE-optimal WGMWW statistic exists, it is in general not possible to find it in practice unless we know the underlying distribution F. This difficulty arises because the PAE-function ψ depends on F. In this paper we propose a strongly consistent estimator $\hat{\psi}(\pi, \mu) = \hat{\varphi}^2(\pi, \mu)/\xi(\pi, \mu, \lambda)$ for $\psi(\pi, \mu)$ for any (fixed) π, μ satisfying (4), where $\hat{\varphi}(\pi, \mu)$ is as given in Theorem 2.1. (Herein, we write $\hat{\psi}$, $\hat{\varphi}$, ψ and, φ for $\hat{\psi}(\pi, \mu)$, $\hat{\varphi}(\pi, \mu)$, $\psi(\pi, \mu)$ and, $\varphi(\pi, \mu)$, respectively, whenever there is no ambiguity.) Our data-adaptive estimator $\hat{\varphi}$ is similar to the estimators based on inverses of spacings of order statistics, presented in Hall (1982) and Grenander (1965).

Xie and Priebe (2002) also introduced the class of WGWSR statistics. This class of one-sample test statistics contained as special cases many of the existing nonparametric distribution-free one sample test statistics, especially the Wilcoxon signed rank statistic. A methodology, similar to that we present in this paper, can be developed to find a PAE-optimal WGWSR statistic data-adaptively. The two sample test that we present in this paper can also be extended into k-sample ($k > 2$) tests for ordered alternatives via Jonckheere-type (Jonckeere, 1954) and Tryon–Hettmansperger-type (Tryon and Hettmansperger, 1973) extensions.

2. Main result

In this section we present our estimator for the PAE function of WGMWW statistics and we show that it is strongly consistent under some mild assumptions. We assume that the underlying distribution function F has a piecewise uniformly continuous density function with nicely behaving tails. The precise statement is given below. As a corollary

to this result, we see that the $\hat{\psi}$-optimal WGMWW statistic, if convergent, is asymptotically equal to the ψ-optimal WGMWW statistic.

Theorem 2.1. *Let $X_{n:1}, \ldots, X_{n:n}$ denote the order statistics of a sample of size n from the distribution F, and let F_n denote the corresponding empirical distribution function. Suppose F has a piecewise uniformly continuous density f, which is ultimately monotonically nonincreasing as $x \to \pm\infty$. If $\varphi = \int_{-\infty}^{\infty} h(F(x)) f^2(x)\,\mathrm{d}x$ and*

$$\hat{\varphi} = \frac{(k-1)}{n^2} \sum_{v=1}^{n-k} h\left(F_n(\rho X_{n:n-v+1} + (1-\rho)X_{n:n-(v+k)+1})\right)\left(X_{n:n-v+1} - X_{n:n-(v+k)+1}\right)^{-1}$$

where

$$h(y) = \sum_{i=1}^{r}\sum_{j=1}^{s} \pi_i \mu_j C(r,s,i,j) y^{i+j-2}(1-y)^{r+s-i-j},$$

$$C(r,s,i,j) = \frac{r!s!}{(i-1)!(r-i)!(j-1)!(s-j)!},$$

then $\hat{\varphi} \to \varphi$ w.p.1. Hence $\hat{\psi} \to \psi$ w.p.1, where $\hat{\psi} = \hat{\varphi}^2/\xi$, $\psi = \varphi^2/\xi$ and ξ is as defined in (7). *Here, we assume that $k > 2$ is an integer and $0 \leqslant \rho \leqslant 1$.*

In Corollary 2.1, we show that the test statistic in our methodology is close to the unknown PAE-optimal WGMWW test statistic. But Theorem 2.1 is our main theoretical justification for our methodology and hence our main result. That is, Theorem 2.1 justifies our test as approximately as efficient as the PAE-optimal WGMWW test. (The proof of Corollary 2.1 is straightforward, and the proof techniques that we use to prove Theorem 2.1 are similar to those used in Hall (1982). See John and Priebe (2005) for details of the proofs.)

Corollary 2.1. *For fixed positive integers r and s, consider the following compact subset of $R^r \times R^s$:*

$$S = \left\{ \left(\pi_1, \ldots, \pi_n, \mu_1, \ldots, \mu_s\right) : \pi_i \geqslant 0, \mu_j \geqslant 0, \ \forall i = 1, \ldots, r, \ \ j = 1, \ldots, s, \ \text{and} \ \sum_{i=1}^{r} \pi_i = 1 = \sum_{j=1}^{s} \mu_j \right\}.$$

Define

$$\hat{\psi} : S \to R \ \text{as} \ \hat{\psi} = \frac{\hat{\varphi}^2}{\xi}$$

and

$$\psi : S \to R \ \text{as} \ \psi = \frac{\varphi^2}{\xi},$$

where $\hat{\varphi}$, φ and ξ are as before. Suppose, $y'_n \to y^0$ with some positive probability, where the y'_n and y^0 are in S and the y'_n satisfy

$$\hat{\psi}\left(y'_n\right) \geqslant \hat{\psi}(y), \ \ w.p.1 \quad \forall y \varepsilon S, \quad n \geqslant 1;$$

then $\psi(y^0) \geqslant \psi(y)$, w.p.1, $\forall y \varepsilon S$.

3. Simulations

In this section, we use the notation *M*, *W*, and *D*, respectively, for the two-sample tests based on the classical Mann–Whitney–Wilcoxon statistic, the PAE-optimal WGMWW statistic, and the data-adaptive PAE-optimal WGMWW statistic. The goals of our simulation study are twofold:

(a) to compare the performance of our two-sample test (i.e., D) with other nonparametric tests (namely, M and W), for finite samples and
(b) to study the empirical convergence properties of $\hat{\psi}$-optimal weights.

For our investigation we generate data from five different distributions with densities. Let $g_{(\mu,\sigma)}$ denote the density function for a normal distribution with mean μ and variance σ^2. The densities used for our simulation study are, respectively,

$$
\begin{aligned}
f_1 &: \quad g_{(0,1)} \quad \text{(standard normal density)},\\
f_2 &: \quad \tfrac{1}{5}\, g_{(0,1)} + \tfrac{1}{5}\, g_{\left(\frac{1}{2},\frac{2}{3}\right)} + \tfrac{3}{5}\, g_{\left(\frac{13}{12},\frac{5}{9}\right)} \quad \text{(mildly left skewed)},\\
f_3 &: \quad \sum_{l=0}^{7} \frac{1}{8}\, g_{\left(3\left\{\left(\frac{2}{3}\right)^l - 1\right\},\left(\frac{2}{3}\right)^{2l}\right)} \quad \text{(strongly right skewed)},\\
f_4 &: \quad \tfrac{3}{4}\, g_{(0,1)} + \tfrac{1}{4}\, g_{\left(\frac{3}{2},\left(\frac{1}{3}\right)^2\right)} \quad \text{(asymmetric bimodal)},\\
f_5 &: \quad \tfrac{2}{3}\, g_{(0,1)} + \tfrac{1}{3}\, g_{\left(0,\frac{1}{10}\right)} \quad \text{(heavily kurtotic)}.
\end{aligned}
$$

To be precise, we generate $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from distribution functions $F(\cdot)$ and $G(\cdot) = F(\cdot - \Delta)$, respectively, where F is the distribution associated with the density, say, f and the densities (i.e., f's) that we consider are, namely, f_1 to f_5 listed above. Each of the above densities is a normal mixture. These densities correspond to the first, second, third, eighth and fourth densities considered in Marron and Wand (1992). The plots of these densities are shown in the first column of Fig. 1.

The main focus of our simulation study is to compare the empirical power behavior of the three tests M, W and D for a fixed small shift, $\Delta = 0.1$. That is, the null and alternate hypotheses that we test using the three different tests are

$$\mathrm{H}_0 : \Delta = 0 \quad \text{vs.} \quad \mathrm{H}_\mathrm{A} : \Delta = 0.1.$$

Note that although our test statistic may be used to test for stochastic ordering, we focus on the above location problem (which is an important special case of stochastic ordering), for illustrative purposes.

For a more extensive simulation study see John (2005). We add a disclaimer here: although our simulation study is reasonably extensive, it is not exhaustive. That is, we do not study the performance of our test in all possible scenarios. For example, we do not consider unequal subsample and sample sizes. Our main goal in this section is illustration—to present a few cases where our methodology is substantially better than the existing ones, and also to present other cases where it is slightly (i.e., nonsignificantly) better.

Table 1 gives the PAEs and Table 2 gives the empirical power of M, and of W for various subsamples sizes, for underlying densities f_1 to f_5. Fig. 1 (second and third columns) gives the plots of empirical power vs. α (where $\alpha \in (0, 0.1]$) of the tests M, W and D for sample sizes 10 and 100 for the underlying densities f_1 to f_5. From Table 1, we observe that the W's have higher efficacy (in the PAE sense) than M. In some cases (such as in the cases where the underlying density is f_3, f_4 or f_5), there is substantial gain in efficacy, while as in other cases (such as the cases where the underlying density is f_1 or f_2), the gain is marginal. Unsurprisingly, a similar type of behavior is observed in the empirical power characteristics. As seen in Table 1, the PAEs of the W's are larger than that of M, when the underlying density is f_1 (Normal). In practice, when the population distributions are assumed to differ only in location, M is directly comparable with the Student's t-test which is known to be optimal with PAE of 1 under the assumptions of normality. It is well-known that if the population distributions are normal, the PAE of M is quite high at 0.955. It is no wonder that many statisticians considered the MWW test (i.e., M) the best nonparametric test for the two-sample location problem in the case of F normal. Table 1 shows that this was indeed a false belief.

When the underlying density is f_2 (mildly left skewed), the PAEs of the W's are marginally larger than that of M. It is interesting to note that the gain in empirical power for W and D, gain in efficacy for W, and convergence properties of D are quite similar when underlying densities are f_1 and f_2. This makes some intuitive sense since f_2 is quite 'close' to f_1, in terms of skewness, kurtosis and tailweights (see Fig. 1, first column).

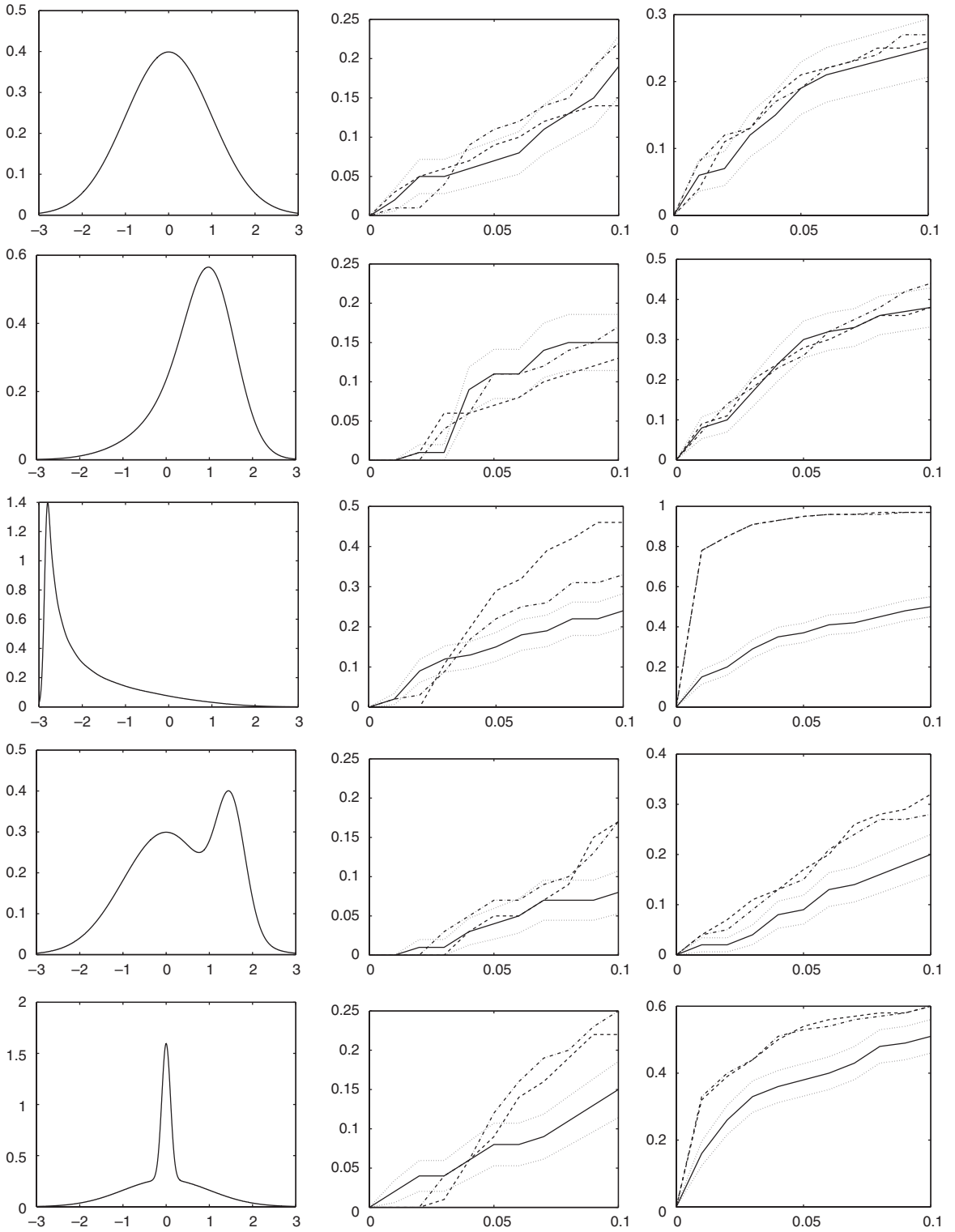

Fig. 1. Plots of the underlying densities f_1, f_2, f_3, f_4, f_5: (in rows 1–5, respectively, of the first column) and the corresponding plots of estimated power vs. α ($\alpha \in (0, 0.1]$) of $M(-)$, $W(--)$, $D(- \cdot -\cdot)$, and confidence intervals $(\cdots)$ for the empirical power of M, for sample sizes $n = m = 10$ (middle column), and $n = m = 100$ (last column). Subsamples size $r = s = 5$ were used for W and D. Monte Carlo sample size, $N = 100$; shift, $\Delta = 0.1$. Parameters k and ρ in $\hat{\psi}$ were chosen to be $k = 3$, $\rho = 0.5$.

Table 1
PAE of M and W's for various underlying densities

Test/density		f_1	f_2	f_3	f_4	f_5
Pitman asymptotic efficacy						
M		0.9549	1.7040	3.7222	0.8309	4.5423
W	$r=s=2$	0.9549	1.8469	9.7806	1.0281	5.5515
	$r=s=3$	0.9889	1.8554	15.0385	1.1927	6.3238
	$r=s=4$	0.9932	1.8650	19.0317	1.3776	6.9395
	$r=s=5$	0.9937	1.8671	27.8753	1.4759	7.4408

Table 2
Empirical power of M, W and D at level 0.05 for the underlying distributions f_1, f_2, f_3, f_4, f_5

Sample sizes	Tests/subsample sizes	$r=s=2$	$r=s=3$	$r=s=4$	$r=s=5$
Underlying density: f_1					
$n=m=10$	M	0.07	0.07	0.07	0.07
	W	0.07	0.10	0.10	0.09
	D	0.09	0.10	0.10	0.11
$n=m=100$	M	0.19	0.19	0.19	0.19
	W	0.19	0.21	0.22	0.21
	D	0.17	0.20	0.20	0.19
Underlying density: f_2					
$n=m=10$	M	0.11	0.11	0.11	0.11
	W	0.07	0.07	0.07	0.07
	D	0.09	0.08	0.10	0.11
$n=m=100$	M	0.30	0.30	0.30	0.30
	W	0.26	0.26	0.27	0.28
	D	0.29	0.29	0.29	0.26
Underlying density: f_3					
$n=m=10$	M	0.15	0.15	0.15	0.15
	W	0.21	0.23	0.24	0.29
	D	0.22	0.23	0.23	0.22
$n=m=100$	M	0.37	0.37	0.37	0.37
	W	0.67	0.85	0.93	0.95
	D	0.67	0.85	0.93	0.95
Underlying density: f_4					
$n=m=10$	M	0.04	0.04	0.04	0.04
	W	0.06	0.07	0.04	0.05
	D	0.04	0.05	0.05	0.07
$n=m=100$	M	0.09	0.09	0.09	0.09
	W	0.11	0.11	0.13	0.17
	D	0.11	0.11	0.15	0.15
Underlying density: f_5					
$n=m=10$	M	0.08	0.08	0.08	0.08
	W	0.00	0.08	0.00	0.09
	D	0.06	0.09	0.11	0.12
$n=m=100$	M 21	0.38	0.38	0.38	0.38
	W	0.38	0.46	0.45	0.54
	D	0.38	0.46	0.51	0.53

Monte Carlo sample size, $N=100$; shift, $\Delta=0.1$. Parameters k and ρ in $\hat{\psi}$ were chosen to be $k=3$, $\rho=0.5$.

Among the five cases considered (corresponding to the five underlying densities f_1 to f_5), the best performance shown by W and D in terms of PAE and empirical power is in the case where the underlying density is f_3 (strongly skewed). As seen in Table 1, there is a phenomenal gain in efficacy for W (nearly an 800% increase for W with $r=s=5$) compared to M. Similar behavior is seen for empirical power of M, W and D, as seen in Table 2.

Table 3
Discrepancy between $\hat{\theta}$'s and θ corresponding to subsample sizes $r = s = 2$ and $r = s = 3$, for the five underlying densities and various sample sizes

Sample sizes/density	f_1	f_2	f_3	f_4	f_5
Subsample size: $r = s = 2$					
$n = m = 25$	1.2137	0.4378	0.0675	0.9436	0.1537
$n = m = 50$	0.6950	0.2244	0.0007	0.6206	0.1308
$n = m = 100$	0.6656	0.1936	0.0000	0.3447	0.0074
$n = m = 250$	0.3933	0.0804	0.0000	0.2821	0.0000
$n = m = 500$	0.2476	0.0607	0.0000	0.2050	0.0000
Subsample size: $r = s = 3$					
$n = m = 25$	0.3292	0.7197	0.0045	0.8408	0.1772
$n = m = 50$	0.3546	0.6514	0.0000	0.5054	0.1436
$n = m = 100$	0.2695	0.5122	0.0000	0.2177	0.0532
$n = m = 250$	0.1143	0.2550	0.0000	0.1474	0.0169
$n = m = 500$	0.0426	0.2495	0.0000	0.0138	0.0152

Monte Carlo sample size, $N = 100$. Parameters k and ρ in $\hat{\psi}$ were chosen to be $k = 3$, $\rho = 0.5$.

When the underlying density is f_4 (asymmetric bimodal), we see that W with $r = s = 5$ has about 78% increase in efficacy compared to M. (See Table 1.) The substantial gain in efficacy for W matches with the substantial gain in empirical power of W and D for large finite samples (e.g., $n = m = 100$) as seen in Tables 1 and 2. The gain in efficacy for W with $r = s = 5$ compared to M is about 64% when the underlying density is f_5, as seen in Table 1. The empirical power of W and D for large finite samples is also substantially higher compared to M. We also note that the PAEs of W's increase with subsample sizes r and s, in all the five cases corresponding to f_1 to f_5. We also have plotted the 95% confidence intervals for the empirical power of M in all the figures.

Importantly, in all the plots, the empirical power behavior of D is almost the same as that of W—the data-adaptive methodology works well, in practice, for reasonable sample sizes.

3.1. Empirical convergence of data-adaptive weights

Consider the class $\mathscr{C}_w^{r,s}$ of WGMMW statistics with fixed subsample sizes r and s. Let $\hat{\theta}^{(i)} = \left(\hat{\pi}^{(i)}, \hat{\mu}^{(i)}\right) = (\hat{\pi}_1{}^{(i)}, \ldots, \hat{\pi}_r^{(i)}, \hat{\mu}_1^{(i)}, \ldots, \hat{\mu}_s^{(i)})$, $i = 1, \ldots, N$, denote the weights obtained by maximizing our estimate of the PAE function for the class $\mathscr{C}_w^{r,s}$ subject to the constraints (4), at the ith Monte Carlo iteration. Let $\theta = (\pi, \mu) = (\pi_1, \ldots, \pi_r, \mu_1, \ldots, \mu_s)$ be the weights obtained by maximizing the PAE function for $\mathscr{C}_w^{r,s}$ subject to (4). The $\hat{\theta}$ depend on the sample sizes. In order to study the empirical convergence of $\hat{\theta}$ to θ as sample sizes increase, we consider the following "discrepancy measure":

$$d\left(\hat{\theta}^{(i)}, i = 1, \ldots, N, \theta\right) = \frac{1}{N} \sum_{k=1}^{N} \left[\left(\sum_{i=1}^{r} \left(\hat{\pi}_i^{(k)} - \pi_i \right)^2 \right) + \left(\sum_{j=1}^{s} \left(\hat{\mu}_j^{(k)} - \mu_j \right)^2 \right) \right],$$

which is the average (over all Monte Carlo iterations) of the squared Euclidean distance between $\hat{\theta}^{(i)}$'s and θ. For illustrative purposes, we restrict our attention to subsample sizes $r = s = 2$ and $r = s = 3$.

The results are presented in Table 3. For both subsample sizes considered, very fast empirical convergence rates are observed for underlying densities f_3 and f_5. The convergence rates are quite good also for $r = s = 2$, when the underling density is f_2, and for $r = s = 3$, when the underlying density is f_1 or f_5. The results in Table 3 support the plots in the second and third column in Fig. 1. That is, in the cases for which the data-adaptive method converges quickly, the empirical power vs. α plots of W and D are nearly identical even for the small sample sizes considered here.

3.2. Discussion about the choice of r, s, k and ρ

By Theorem 6.3.1 in Xie (1999), whenever $r_1 \leqslant r_2$ and $s_1 \leqslant s_2$ the maximum PAE of WGMWW statistics in $\mathscr{C}_w^{r_2,s_2}$ is not less than that in $\mathscr{C}_w^{r_1,s_1}$ and, moreover, at least one member in $\mathscr{C}_w^{r_2,s_2}$ has strictly larger PAE than does any member in $\mathscr{C}_w^{r_1,s_1}$. So for fixed n, W (and hence D) is more powerful in the PAE sense for larger r and s. The choice of r, s—model selection—becomes an issue of bias–variance tradeoff. For larger r, s, the model bias is smaller, but the estimation variance is larger. Another relevant point is the fact that the weight vectors π and μ arise from the polynomial approximations u_r and v_s (see (2)) to the unknown increasing continuous functions u and v given in (3). Asymptotically, achieving true optimality requires letting r and s go to infinity at some rate, e.g. $\mathrm{O}\left(n^{\alpha}\right)$, $\alpha \in (0, 1)$. How the tests W and D behave in such a scenario is a topic for further research.

The estimator $\hat{\psi}$ in Theorem 2.1 is consistent for any choice of $k > 2$ and $0 \leqslant \rho \leqslant 1$. That is, for large n, D has approximately the same efficacy as W in the PAE sense, for any choice of k and ρ. But the choice of k and ρ do play a role in the properties of $\hat{\psi}$. In this regard, developing the asymptotic distributional properties of $\hat{\psi}$ is a relevant topic for further research.

4. Analysis of magnetic resonance brain imaging data

Changes in the structure of the cerebral cortex have been associated with both dementia of the Alzheimer's type (DAT) and healthy aging. Miller et al. (2003) showed that there exists sharp distinctions in cortical thickness of the cingulate between normal aging subjects and patients with mild DAT. In their study, the thickness of the cortical mantle was quantified using cortical mantle distance maps (CMDMs). The cortical mantle has a thin laminar structure consisting of layers of cerebrospinal fluid (CSF) at the top, layers of gray matter (GM) in the middle, and layers of white matter (WM) at the bottom. Imaging data from magnetic resonance scans of the cingulate gyrus may be transformed and interpolated into $0.5 \times 0.5 \times 0.5\,\mathrm{mm}^3$ voxels. CMDMs measure the distance of each voxel to the GM/WM interface. The data were obtained from elder subjects with mild DAT ($n = 9$) and healthy elders with no evidence of dementia ($n = 10$). The clinical dementia rating scale (CDR) was used to assign the subjects into the above two groups. Subjects with no discernible evidence of dementia on CDR were designated as healthy subjects with CDR0. A score of CDR1 indicates mild DAT.

4.1. Analysis of the data

CMDMs corresponding to anatomically defined left anterior (LA), left posterior (LP), right anterior (RA), right posterior (RP) regions of the cingulate gyrus of all the subjects were available for the study. Samples of size 80 were chosen without replacement from the CMDMs of subjects in the mild DAT (i.e., CDR1) group and healthy aging (i.e., CDR0) group. Let F and G denote the underlying distribution functions of the CMDMs from the CDR1 group and CDR0 group, respectively. The null and alternate hypotheses are

$$\mathrm{H}_0 : F = G \quad \text{vs.}$$

$$\mathrm{H}_A : F(x) \geqslant G(x) \quad \forall x, \text{ with strict inequality for at least one } x.$$

The classical Mann–Whitney–Wilcoxon test (i.e., M), the data-adaptive WGMWW test (i.e., D) with subsample sizes $r = s = 2$, and D with $r = s = 3$ were used to test the above hypotheses. The medians of p-values over 15 trials obtained using each of the above three tests are shown in Table 4. As seen from Table 4, the CMDMs of the CDR1 sample are stochastically larger than those of the CDR0 sample at a significance level 0.05 (using the data-adaptive statistic) for each of the four regions LA, LP, RA, RP. Miller et al. (2003) demonstrate this as well, using M, with larger sample sizes. (We make note of the fact that the data analysis performed here (separately for each of the four regions) parallels that which was done in Miller et al., 2003.) Our point here is that the data-adaptive WGMWW tests are more powerful for this real application and data—again, the data-adaptive methodology works well, in practice, for reasonable sample sizes.

Table 4
Medians of p-values over 15 trials for various tests

Regions/tests	M	$D\ (r=s=2)$	$D\ (r=s=3)$
Medians of p-values over 15 *trials*			
Left anterior	0.0444	0.0202	0.0104
Left posterior	0.0664	0.0285	0.0256
Right anterior	0.0546	0.0313	0.0170
Right posterior	0.0569	0.0360	0.0311

Samples of sizes $n=m=80$ were chosen without replacement from each population.

Acknowledgments

The MR cingulate gyrus data were provided by Dr. John G. Csernansky of Washington University School of Medicine, and the Labeled Cortical Mantle Distance Maps (LCMDM) of the cingulate gyrus were provided by Dr. Michael I. Miller of the Johns Hopkins University. Research support by NIH grants (P20-MH071616 and R01-MH064838) is gratefully acknowledged.

Appendix A. Proofs

In this appendix we present a sketch of the proofs of Theorem 2.1 and Corollary 2.1. The proof is lengthy but the proof techniques are fairly close to that in Hall (1982) and hence we omit most of the details. A reader interested in the details is referred to John and Priebe (2005). We need Lemma A.3 to prove Theorem 2.1. Lemmas A.1 and A.2 are needed to prove Lemma A.3.

Lemma A.1. *Let* $U_{n:1}\leqslant\cdots\leqslant U_{n:n}$ *be order statistics from the* $\mathcal{U}(0,1)$ *distribution. Let* $0\leqslant\theta\leqslant1$. *Then* $U_{n:n-\lfloor n\theta\rfloor+1}\longrightarrow(1-\theta)$, *w.p.*1. *In particular, if we define* $V_{n:\lfloor n\theta\rfloor}=-\log\left(U_{n:n-\lfloor n\theta\rfloor+1}\right)$, *then*

$$V_{n:\lfloor n\theta\rfloor}\longrightarrow-\log(1-\theta)\quad w.p.1$$

and

$$V_{n:\lfloor n\theta\rfloor}+\lambda\left(V_{n:\lfloor n\theta\rfloor+k}-V_{n:\lfloor n\theta\rfloor}\right)\longrightarrow-\log(1-\theta)\quad w.p.1,$$

where $k>0$ *is an integer, and* $0\leqslant\lambda\leqslant1$.

Proof. See John and Priebe (2005). □

Lemma A.2. *Let* $V_{n:r}=-\log F\left(X_{n:n-r+1}\right)$, $1\leqslant r\leqslant n$ (*here* $X_{n:1}<\cdots<X_{n:n}$ *denote the order statistics from some distribution* F, *so that* $V_{n:r}$ *are order statistics from an* $Exponential(1)$ *distribution*); *let* $0\leqslant\pi_1<\pi_2<1$; $k>0$ *be an integer, and* $\mu=1/(k-1)=\Gamma(k-1)/\Gamma(k)$. *Then*

$$\lim_{n\to\infty}E\left\{\frac{1}{n^2}\sum_{r=\lfloor n\pi_1\rfloor}^{\lfloor n\pi_2\rfloor}\left(V_{n:r+k}-V_{n:r}\right)^{-1}\right\}=\frac{\mu}{2}\left\{(1-\pi_1)^2-(1-\pi_2)^2\right\}=\mu\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x.$$

Also,

$$\lim_{n\to\infty}E\left\{\frac{1}{n^2}\sum_{r=\lfloor n\pi_1/k\rfloor}^{\lfloor n\pi_2/k\rfloor}\left(V_{n:rk+k}-V_{n:rk}\right)^{-1}\right\}=\frac{\mu}{2k}\left\{(1-\pi_1)^2-(1-\pi_2)^2\right\}=\frac{\mu}{k}\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x.$$

Proof. See John and Priebe (2005). □

Lemma A.3. *Let* $V_{n:r}$, π_1, π_2, k *and* μ *be as in Lemma* A.2; *then*

$$\frac{1}{n^2}\sum_{r=\lfloor n\pi_1\rfloor}^{\lfloor n\pi_2\rfloor}(V_{n:r+k}-V_{n:r})^{-1}\longrightarrow\mu\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x\quad w.p.1. \tag{A.1}$$

Proof. See John and Priebe (2005). □

Proof of Theorem 2.1. Note that F_n^m is a step function for any positive integer m, and hence piecewise uniformly continuous. So, there exists (extended) real numbers $-\infty=x_0<x_1<\cdots<x_w=\infty$ such that f and F_n^m are both uniformly continuous on each (x_{i-1},x_i) for $1\leqslant i\leqslant w$. First, we shall prove the following two claims:

Claim A.1. *For some* $0<\alpha<\beta<1$ *and* $0\leqslant\rho\leqslant1$,

$$\frac{1}{n^2}\sum_{v=\lfloor n\alpha\rfloor}^{\lfloor n\beta\rfloor}F_n^m\left(\rho X_{n:n-v+1}+(1-\rho)X_{n:n-(v+k)+1}\right)\left(X_{n:n-v+1}-X_{n:n-(v+k)+1}\right)^{-1}$$
$$\longrightarrow\frac{1}{(k-1)}\int_{F^{-1}(1-\alpha)}^{F^{-1}(1-\beta)}F^m(x)f^2(x)\,\mathrm{d}x\quad w.p.1.\ as\ n\to\infty.$$

Claim A.2. *With probability one*,

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{v=1}^{\lfloor n\alpha\rfloor}F_n^m\left(\rho X_{n:n-v+1}+(1-\rho)X_{n:n-(v+k)+1}\right)\left(X_{n:n-v+1}-X_{n:n-(v+k)+1}\right)^{-1}$$

can be made arbitrary small by choosing α *small enough. Similarly*, *with probability one*,

$$\lim_{n\to\infty}\frac{1}{n^2}\sum_{v=\lfloor n\beta\rfloor}^{n-k}F_n^m\left(\rho X_{n:n-v+1}+(1-\rho)\,X_{n:n-(v+k)+1}\right)\left(X_{n:n-v+1}-X_{n:n-(v+k)+1}\right)^{-1}$$

is arbitrary small, *when* β *is sufficiently close to* 1.

These two claims would imply that

$$\frac{1}{n^2}\sum_{v=1}^{n-k}F_n^m\left(\rho X_{n:n-v+1}+(1-\rho)X_{n:n-(v+k)+1}\right)\left(X_{n:n-v+1}-X_{n:n-(v+k)+1}\right)^{-1}$$
$$\longrightarrow\frac{1}{(k-1)}\int_{-\infty}^{\infty}F^m(x)f^2(x)\,\mathrm{d}x\quad w.p.1\ \text{as}\ n\to\infty. \tag{A.2}$$

It is easy to see that (A.2) is sufficient to show that $\hat{\varphi}_n\to\varphi$, *w.p.*1.

Proof of Claim A.1. Define $H(x)=F^{-1}(\mathrm{e}^{-x})$, $x>0$, so that $H'(x)=-\mathrm{e}^{-x}/f\{H(x)\}$. Let

$$V_{n:v}=-\log\{F\left(X_{n:n-v+1}\right)\},\quad 1\leqslant v\leqslant n.$$

By the Mean Value Theorem, for some ϕ (which depends on n and v),

$$H\left(V_{n:v+k}\right)-H\left(V_{n:v}\right)=H'\left\{\phi V_{n:v+k}-(1-\phi)V_{n:v}\right\}\left(V_{n:v+k}-V_{n:v}\right).$$

That is,

$$X_{n:n-(v+k)+1}-X_{n:n-v+1}=H'\left\{V_{n:v}+\phi\left(V_{n:v+k}-V_{n:v}\right)\right\}\left(V_{n:v+k}-V_{n:v}\right).$$

Now choose $0 < \Delta_1 < \Delta_2 < 1$, so that $\left(F^{-1}(1-\Delta_2), F^{-1}(1-\Delta_1)\right)$is any one of the following open intervals:

(1) (x_{i-1}, x_i), for some $i \in \{2, \ldots, w-1\}$,
(2) (a, x_1), for $-\infty = x_0 < a < x_1$,
(3) (x_{w-1}, b), for $x_{w-1} < b < x_w = \infty$.

This assures that both f and F_n^m are uniformly continuous in $\left(F^{-1}(1-\Delta_2), F^{-1}(1-\Delta_1)\right)$. We will need the following two subclaims to prove Claim A.1.

Subclaim A.1. *Let $\Delta_1 \leqslant \pi_1 < \pi_2 \leqslant \Delta_2$, and let $\left(\pi^{(l)}\right)$ be an arbitrary sequence in (Δ_1, Δ_2), converging to π_1. Then we can define the function $\varepsilon_{1,\pi_1}(\cdot) > 0$ on (Δ_1, Δ_2) such that $\varepsilon_{1,\pi_1}\left(\pi^{(l)}\right) \to 0$ uniformly as $l \to \infty$ and*

$$\lim_{n\to\infty} \sup_{\lfloor n\pi_1 \rfloor \leqslant v \leqslant \lfloor n\pi^{(l)} \rfloor} \left| \left[H'\left(V_{n:v} + \phi\left(V_{n:v+k} - V_{n:v}\right)\right)\right]^{-1} - \left[H'\left(-\log\left(1-\pi_1\right)\right)\right]^{-1} \right| \leqslant \varepsilon_{1,\pi_1}\left(\pi^{(l)}\right) \tag{A.3}$$

*w.p.*1. *In other words, there exists a function $\varepsilon_{1,\pi_1}(\cdot) > 0$ on (Δ_1, Δ_2)* such that $\varepsilon_{1,\pi_1}(\pi_2) \to 0$ *uniformly as $\pi_2 \downarrow \pi_1$ and*

$$\lim_{n\to\infty} \sup_{\lfloor n\pi_1 \rfloor \leqslant v \leqslant \lfloor n\pi_2 \rfloor} \left| \left[H'\left(V_{n:v} + \phi\left(V_{n:v+k} - V_{n:v}\right)\right)\right]^{-1} - \left[H'\left(-\log\left(1-\pi_1\right)\right)\right]^{-1} \right| \leqslant \varepsilon_{1,\pi_1}(\pi_2)$$

*w.p.*1. *Since* $\left[H'\left(-\log\left(1-\pi_1\right)\right)\right]^{-1}$ is $-f\left(F^{-1}(1-\pi_1)\right)/1-\pi_1$, *the above statement may be rewritten as*

$$\lim_{n\to\infty} \sup_{\lfloor n\pi_1 \rfloor \leqslant v \leqslant \lfloor n\pi_2 \rfloor} \left| \left[H'\left(V_{n:v} + \phi\left(V_{n:v+k} - V_{n:v}\right)\right)\right]^{-1} - \left[-\frac{f\left(F^{-1}(1-\pi_1)\right)}{1-\pi_1}\right] \right| \leqslant \varepsilon_{1,\pi_1}(\pi_2)$$

*w.p.*1.

Subclaim A.2. *Let π_1, π_2 be as in Subclaim* A.1. *Then there exists a function $\varepsilon_{2,\pi_1}(\cdot) > 0$ on (Δ_1, Δ_2) such that $\varepsilon_{2,\pi_1}(\pi_2) \to 0$ uniformly as $\pi_2 \downarrow \pi_1$ and*

$$\lim_{n\to\infty} \sup_{\lfloor n\pi_1 \rfloor \leqslant v \leqslant \lfloor n\pi_2 \rfloor \rfloor} \left| F_n^m\left(\rho X_{n:n-v+1} + (1-\rho) X_{n:n-(v+k)+1}\right) - F_n^m\left(F^{-1}(1-\pi_1)\right) \right| \leqslant \varepsilon_{2,\pi_1}(\pi_2).$$

See John and Priebe (2005) for the proofs of Subclaims A.1 and A.2.

As a first step towards proving Claim A.1, we observe that the following is true for $\Delta_1 \leqslant \pi_1 < \pi_2 \leqslant \Delta_2$, using Subclaims A.1, A.2, Lemma A.3:

$$\begin{aligned}
&\left\{F^m\left(F^{-1}(1-\pi_1)\right) - \varepsilon_{2,\pi_1}(\pi_2)\right\} \left\{\frac{-f\left(F^{-1}(1-\pi_1)\right)}{1-\pi_1} - \varepsilon_{1,\pi_1}(\pi_2)\right\} \left\{-\mu \int_{\pi_1}^{\pi_2} (1-x)\,\mathrm{d}x\right\} \\
&\quad - \mu F^m\left(F^{-1}(1-\pi_1)\right) f\left(F^{-1}(1-\pi_1)\right)(\pi_2 - \pi_1) \\
&\quad \leqslant \lim_{n\to\infty} \left\{\frac{1}{n^2} \sum_{v=[n\pi_1]}^{[n\pi_2]} F_n^m\left(\rho X_{n:n-v+1} + (1-\rho) X_{n:n-(v+k)+1}\right)\left(X_{n:n-v+1} - X_{n-(v+k)+1}\right)^{-1}\right\} \\
&\qquad - \mu F^m\left(F^{-1}(1-\pi_1)\right) f\left(F^{-1}(1-\pi_1)\right)(\pi_2 - \pi_1) \\
&\quad \leqslant \left\{F^m\left(F^{-1}(1-\pi_1)\right) + \varepsilon_{2,\pi_1}(\pi_2)\right\} \left\{\frac{-f\left(F^{-1}(1-\pi_1)\right)}{1-\pi_1} + \varepsilon_{1,\pi_1}(\pi_2)\right\} \left\{-\mu \int_{\pi_1}^{\pi_2} (1-x)\,\mathrm{d}x\right\} \\
&\qquad - \mu F^m\left(F^{-1}(1-\pi_1)\right) f\left(F^{-1}(1-\pi_1)\right)(\pi_2 - \pi_1) \quad w.p.1.
\end{aligned}$$

Rewriting the above inequalities after expanding the terms we get,

$$
\begin{aligned}
&\mu F^m\left(F^{-1}(1-\pi_1)\right)\frac{f\left(F^{-1}(1-\pi_1)\right)}{1-\pi_1}\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x\\
&-\mu F^m\left(F^{-1}(1-\pi_1)\right)f\left(F^{-1}(1-\pi_1)\right)(\pi_2-\pi_1)\\
&+\varepsilon_{1,\pi_1}(\pi_2)\left\{\mu\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x\right\}\left\{F^m\left(F^{-1}(1-\pi_1)\right)\right\}\\
&-\varepsilon_{2,\pi_1}(\pi_2)\left\{\mu\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x\right\}\left\{\frac{f\left(F^{-1}(1-\pi_1)\right)}{1-\pi_1}\right\}\\
&-\varepsilon_{3,\pi_1}(\pi_2)\left\{\mu\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x\right\}\\
&\leqslant\lim_{n\to\infty}\left\{\frac{1}{n^2}\sum_{v=\lfloor n\pi_1\rfloor}^{\lfloor n\pi_2\rfloor}F_n^m\left(\rho X_{n:n-v+1}+(1-\rho)X_{n:n-(v+k)+1}\right)\left(X_{n:n-v+1}-X_{n-(v+k)+1}\right)^{-1}\right\}\\
&\quad-\mu F^m\left(F^{-1}(1-\pi_1)\right)f\left(F^{-1}(1-\pi_1)\right)(\pi_2-\pi_1)\\
&\leqslant\mu F^m\left(F^{-1}(1-\pi_1)\right)\frac{f\left(F^{-1}(1-\pi_1)\right)}{1-\pi_1}\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x\\
&\quad-\mu F^m\left(F^{-1}(1-\pi_1)\right)f\left(F^{-1}(1-\pi_1)\right)(\pi_2-\pi_1)\\
&\quad-\varepsilon_{1,\pi_1}(\pi_2)\left\{\mu\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x\right\}\left\{F^m\left(F^{-1}(1-\pi_1)\right)\right\}\\
&\quad+\varepsilon_{2,\pi_1}(\pi_2)\left\{\mu\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x\right\}\left\{\frac{f\left(F^{-1}(1-\pi_1)\right)}{1-\pi_1}\right\}\\
&\quad-\varepsilon_{3,\pi_1}(\pi_2)\left\{\mu\int_{\pi_1}^{\pi_2}(1-x)\,\mathrm{d}x\right\}\quad\text{w.p.1.}
\end{aligned}
\tag{A.4}
$$

In (A.4), we used $\varepsilon_{3,\pi_1}(\pi_2)$ for $\varepsilon_{1,\pi_1}(\pi_2)\,\varepsilon_{1,\pi_1}(\pi_2)$. Note that $\varepsilon_{3,\pi_1}(\pi_2)$ also converges uniformly to zero, as $\pi_2\downarrow\pi_1$. Partition $[\Delta_1,\Delta_2]$ into t equal intervals:

$$\Delta_1=\pi_0<\pi_1<\cdots<\pi_t=\Delta_2.$$

Then using (A.4) we obtain

$$
\begin{aligned}
&\mu\sum_{i=1}^{t}\left\{F^m\left(F^{-1}(1-\pi_{i-1})\right)\frac{f\left(F^{-1}(1-\pi_{i-1})\right)}{1-\pi_{i-1}}\int_{\pi_{i-1}}^{\pi_i}(1-x)\,\mathrm{d}x\right\}\\
&-\mu\sum_{i=1}^{t}\left\{F^m\left(F^{-1}(1-\pi_{i-1})\right)f\left(F^{-1}(1-\pi_{i-1})\right)(\pi_i-\pi_{i-1})\right\}\\
&+\mu\sum_{i=1}^{t}\left\{\varepsilon_{1,\pi_{i-1}}(\pi_i)\left\{\int_{\pi_{i-1}}^{\pi_i}(1-x)\,\mathrm{d}x\right\}\left\{F^m\left(F^{-1}(1-\pi_{i-1})\right)\right\}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&- \mu \sum_{i=1}^{t} \left\{ \varepsilon_{2,\pi_{i-1}}(\pi_i) \left\{ \int_{\pi_{i-1}}^{\pi_i} (1-x)\,\mathrm{d}x \right\} \left\{ \frac{f\left(F^{-1}(1-\pi_{i-1})\right)}{1-\pi_{i-1}} \right\} \right\} \\
&- \mu \sum_{i=1}^{t} \left\{ \varepsilon_{3,\pi_{i-1}}(\pi_i) \int_{\pi_{i-1}}^{\pi_i} (1-x)\,\mathrm{d}x \right\} \\
&\leqslant \lim_{n\to\infty} \left\{ \frac{1}{n^2} \sum_{r=\lfloor n\Delta_1 \rfloor}^{\lfloor n\Delta_2 \rfloor} F_n^m \left(\rho X_{n:n-v+1} + (1-\rho) X_{n:n-(v+k)+1}\right) \left(X_{n:n-v+1} - X_{n-(v+k)+1}\right)^{-1} \right\} \\
&\quad - \mu \sum_{i=1}^{t} \left\{ F^m\left(F^{-1}(1-\pi_{i-1})\right) f\left(F^{-1}(1-\pi_{i-1})\right)(\pi_i - \pi_{i-1}) \right\} \\
&\leqslant \mu \sum_{i=1}^{t} \left\{ F^m\left(F^{-1}(1-\pi_{i-1})\right) \frac{f\left(F^{-1}(1-\pi_{i-1})\right)}{1-\pi_{i-1}} \int_{\pi_{i-1}}^{\pi_i} (1-x)\,\mathrm{d}x \right\} \\
&\quad - \mu \sum_{i=1}^{t} \left\{ F^m\left(F^{-1}(1-\pi_{i-1})\right) f\left(F^{-1}(1-\pi_{i-1})\right)(\pi_i - \pi_{i-1}) \right\} \\
&\quad - \mu \sum_{i=1}^{t} \left\{ \varepsilon_{1,\pi_{i-1}}(\pi_i) \left\{ \int_{\pi_{i-1}}^{\pi_i} (1-x)\,\mathrm{d}x \right\} \left\{ F^m\left(F^{-1}(1-\pi_{i-1})\right) \right\} \right\} \\
&\quad + \mu \sum_{i=1}^{t} \left\{ \varepsilon_{2,\pi_{i-1}}(\pi_i) \left\{ \int_{\pi_{i-1}}^{\pi_i} (1-x)\,\mathrm{d}x \right\} \left\{ \frac{f\left(F^{-1}(1-\pi_{i-1})\right)}{1-\pi_{i-1}} \right\} \right\} \\
&\quad - \mu \sum_{i=1}^{t} \left\{ \varepsilon_{3,\pi_{i-1}}(\pi_i) \int_{\pi_{i-1}}^{\pi_i} (1-x)\,\mathrm{d}x \right\} \quad \text{w.p.1.}
\end{aligned} \tag{A.5}
$$

In the limit as $t \to \infty$, (A.5) becomes

$$
\begin{aligned}
0 \leqslant \lim_{n\to\infty} &\left\{ \frac{1}{n^2} \sum_{r=\lfloor n\Delta_1 \rfloor}^{\lfloor n\Delta_2 \rfloor} F_n^m \left(\rho X_{n:n-v+1} + (1-\rho) X_{n:n-(v+k)+1}\right) \left(X_{n:n-v+1} - X_{n-(v+k)+1}\right)^{-1} \right\} \\
&- \mu \int_{\Delta_1}^{\Delta_2} F^m\left(F^{-1}(1-x)\right) f\left(F^{-1}(1-x)\right) \mathrm{d}x \leqslant 0 \quad \text{w.p.1.}
\end{aligned}
$$

(See John and Priebe, 2005 for more details.) That is,

$$
\begin{aligned}
&\lim_{n\to\infty} \left\{ \frac{1}{n^2} \sum_{v=\lfloor n\Delta_1 \rfloor}^{\lfloor n\Delta_2 \rfloor} F_n^m \left(\rho X_{n:n-v+1} + (1-\rho) X_{n:n-(v+k)+1}\right) \left(X_{n:n-v+1} - X_{n-(v+k)+1}\right)^{-1} \right\} \\
&= \mu \int_{\Delta_1}^{\Delta_2} F^m\left(F^{-1}(1-x)\right) f\left(F^{-1}(1-x)\right) \mathrm{d}x \\
&= \mu \int_{F^{-1}(1-\Delta_2)}^{F^{-1}(1-\Delta_1)} F^m(x) f^2(x)\,\mathrm{d}x, \quad \text{w.p.1.}
\end{aligned} \tag{A.6}
$$

Now for any α, β satisfying $0 < \alpha < \beta < 1$, we can partition the interval (α, β) as a disjoint union of open intervals:

$$\left(\Delta_1^{(0)}, \Delta_2^{(0)}\right) \cup \left(\Delta_1^{(1)}, \Delta_2^{(1)}\right) \cup \cdots \cup \left(\Delta_1^{(p)}, \Delta_2^{(p)}\right),$$

where $\alpha = \Delta_1^{(0)}$, $\beta = \Delta_2^{(p)}$, $\Delta_2^{(i-1)} = \Delta_1^{(i)}$, for $i = 1, \ldots, p$, and f and F_n^m are uniformly continuous in $\left(F^{-1}\left(1 - \Delta_2^{(i)}\right), F^{-1}\left(1 - \Delta_1^{(i)}\right)\right)$, for $i = 1, \ldots, p$. Hence, (A.6) implies that

$$\lim_{n\to\infty} \left\{ \frac{1}{n^2} \sum_{v=\lfloor n\alpha \rfloor}^{\lfloor n\beta \rfloor} F_n^m \left(\rho X_{n:n-v+1} + (1-\rho) X_{n:n-(v+k)+1}\right) \left(X_{n:n-v+1} - X_{n-(v+k)+1}\right)^{-1} \right\}$$
$$= \mu \int_{F^{-1}(1-\beta)}^{F^{-1}(1-\alpha)} F^m(x) f^2(x)\, \mathrm{d}x \quad \text{w.p.1.}$$

This proves Claim A.1. □

Proof of Claim A.2. $\left|H'(V_{n:v} + \phi(V_{n:v+k} - V_{n:v}))^{-1}\right| = \exp(V_{n:v} + \phi(V_{n:v+k} - V_{n:v}))\, f(H(V_{n:v} + \phi(V_{n:v+k} - V_{n:v})))$. Since $V_{n:v}$ is increasing in v,

$$\exp(V_{n:v} + \phi(V_{n:v+k} - V_{n:v})) \leqslant \exp(V_{n:v+k}).$$

Also, H is a monotone nonincreasing function. Therefore, $H(V_{n:v} + \phi(V_{n:v+k} - V_{n:v})) \geqslant H(V_{v+k})$. But we assume that f is ultimately monotone nonincreasing. Hence, if we choose v sufficiently small, then

$$f(H(V_{n:v} + \phi(V_{n:v+k} - V_{n:v}))) \leqslant f(H(V_{n:v+k})).$$

Combining all these facts, we see that if α is sufficiently small, then for all sufficiently large n, w.p.1,

$$\left|\left\{H'(V_{n:v} + \phi(V_{n:v+k} - V_{n:v}))\right\}^{-1}\right| \leqslant \exp(V_{n:v+k})\, f(H(V_{n:v+k}))$$
$$\leqslant C f\left(X_{n:n-(v+k)+1}\right),$$

uniformly in $1 \leqslant v \leqslant \lfloor n\alpha \rfloor$. Here, C is some positive constant. Then,

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{v=1}^{\lfloor n\alpha \rfloor} F_n^m \left(\rho X_{n:n-v+1} + (1-\rho) X_{n:n-(v+k)+1}\right) \left(X_{n:n-v+1} - X_{n:n-(v+k)+1}\right)^{-1}$$
$$\leqslant C \lim_{n\to\infty} \frac{1}{n^2} \sum_{v=1}^{\lfloor n\alpha \rfloor} \left\{ F_n^m \left(\rho X_{n:n-v+1} + (1-\rho) X_{n:n-(v+k)+1}\right) f\left(X_{n:n-(v+k)+1}\right) (V_{n:v+k} - V_{n:v})^{-1} \right\}$$

w.p.1, when α is small enough. But $F_n^m \leqslant 1$ always, and $f\left(X_{n:n-(v+k)+1}\right) \leqslant K$, for some positive constant K, when v is small enough, since f is ultimately nonincreasing. This implies that, when α is small enough, w.p.1,

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{r=1}^{\lfloor n\alpha \rfloor} F_n^m \left(\rho X_{n:n-v+1} + (1-\rho) X_{n:n-(v+k)+1}\right) \left(X_{n:n-v+1} - X_{n:n-(v+k)+1}\right)^{-1}$$
$$\leqslant C' \lim_{n\to\infty} \frac{1}{n^2} \sum_{r=1}^{\lfloor n\alpha \rfloor} (V_{n:v+k} - V_{n:v})^{-1}$$
$$= C' \int_0^{\alpha} (1-x)\, \mathrm{d}x.$$

Here C' is some positive constant and the last equality follows from Lemma A.3. Clearly,

$$\lim_{\alpha \downarrow 0} \int_0^{\alpha} (1 - x)\, \mathrm{d}x = 0$$

and this proves the first assertion in the claim. The second assertion follows similarly. □

As noted previously, the two claims combined prove the theorem. □

Proof of Corollary 2.1. See John and Priebe (2005). □

References

Adams, B.E., et al., 2000. Ahmad, I.A. (1996). A class of Mann–Whitney–Wilcoxon type statistics. Amer. Statist. 50, 324–327: Comment by Adams, Adams, Chang, Etzel, Kuo, Montemayor, and Schucany; and Reply. Amer. Statist. 54, 160.

Ahmad, I.A., 1996. A class of Mann–Whitney–Wilcoxon type statistics. Amer. Statist. 50, 324–327.

Deshpande, J.V., Kochar, S.C., 1980. Some competitors of tests based on powers of ranks for the two-sample problem. Sankhya Ser. B 42, 236–241.

Grenander, U., 1965. Some direct estimates of the mode. Ann. Math. Statist. 36 (1), 131–138.

Hall, P., 1982. Limit theorems for estimators based on inverses of spacings of order statistics. Ann. Probab. 10 (4), 992–1003.

John, M., 2005. A data-adaptive methodology for finding an optimal weighted generalized Mann–Whitney–Wilcoxon statistic. Ph.D. Dissertation, Department of Applied Mathematics and Statistics, Johns Hopkins University.

John, M., Priebe, C.E., 2005. A data-adaptive methodology for finding an optimal weighted generalized Mann–Whitney–Wilcoxon statistic. Technical Report No. 652, Department of Applied Mathematics and Statistics, Johns Hopkins University.

Jonckeere, A.R., 1954. A Distribution-free *k*-sample test against ordered alternatives. Biometrika 41 (1/2), 133–145.

Kochar, S.C., 1978. A class of distribution-free tests for the two-sample slippage problem. Comm. Statist.—Theory Methods A 7 (13), 1243–1252.

Kumar, N., 1997. A class of two-sample tests for location based on sub-sample medians. Comm. Statist.—Theory Methods 26, 943–951.

Lehmann, E.L., 1998. Nonparametrics: Statistical Methods Based on Ranks, Revised. Prentice-Hall, Englewood Cliffs, NJ.

Lorentz, G.G., 1986. Bernstein Polynomials. second ed. Chelsea, New York.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Statist. 18, 50–60.

Marron, J.S., Wand, M.P., 1992. Exact mean integrated square error. Ann. Statist. 20, 712–736.

Miller, M.I., et al., 2003. Labelled cortical depth maps quantifying cortical change during aging. Proc. Nat. Acad. Sci. 100 (25), 15172–15177.

Pitman, E.J.G., 1979. Some Basic Theory for Statistical Inference. Chapman & Hall, New York.

Priebe, C.E., Cowen, L.J., 1999. A generalized Mann–Whitney–Wilcoxon statistic. Comm. Statist.—Theory Methods 28 (12), 2871–2878.

Shetty, I.D., Govindarajulu, Z., 1988. A two-sample test for location. Comm. Statist.—Theory Methods 17, 2389–2401.

Stephenson, R.W., Ghosh, M., 1985. Two sample nonparametric tests based on subsamples. Comm. Statist.—Theory Methods 14 (7), 1669–1684.

Tryon, P.V., Hettmansperger, T.P., 1973. A class of non-parametric tests for homogeneity against ordered alternatives. Ann. Statist. 1 (6), 1061–1070.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics 1, 80–83.

Xie, J., 1999. Generalizing the Mann–Whitney–Wilcoxon statistic. Ph.D. Dissertation, Department of Applied Mathematics and Statistics, Johns Hopkins University.

Xie, J., Priebe, C.E., 2000. Generalizing the Mann–Whitney–Wilcoxon Statistic. J. Nonparametric Statist. 12, 661–682.

Xie, J., Priebe, C.E., 2002. A weighted generalization of the Mann–Whitney–Wilcoxon statistic. J. Statist. Plann. Inference 102, 441–466.