Journal of Statistical Software

November 2020, Volume 96, Issue 6. *doi: 10.18637/jss.v096.i06*

Wilcoxon Rank-Based Tests for Clustered Data with R Package clusrank

Yujing Jiang
Colorado State University

Xin He
University of Maryland

Mei-Ling Ting Lee
University of Maryland

Bernard Rosner
Harvard University

Jun Yan
University of Connecticut

Abstract

Wilcoxon rank-based tests are distribution-free alternatives to the popular two-sample and paired t tests. For independent data, they are available in several R packages such as **stats** and **coin**. For clustered data, in spite of the recent methodological developments, there did not exist an R package that makes them available at one place. We present a package **clusrank** where the latest developments are implemented and wrapped under a unified user-friendly interface. With different methods dispatched based on the inputs, this package offers great flexibility in rank-based tests for various clustered data. Exact tests based on permutations are also provided for some methods. Details of the major schools of different methods are briefly reviewed. Usages of the package **clusrank** are illustrated with simulated data as well as a real dataset from an ophthalmological study. The package also enables convenient comparison between selected methods under settings that have not been studied before and the results are discussed.

Keywords: Wilcoxon rank-sum test, Wilcoxon signed-rank test.

1. Introduction

The Wilcoxon rank-sum and signed-rank tests are important tools for two-group comparisons and paired comparisons, respectively. Unlike their counterparts under the normality assumption, they are attractive because they are rank-based without the need of distributional assumptions. Nonetheless, standard versions of such tests presume independent data, and cannot be applied to clustered data which frequently arise in many fields. Clustered data consist of data obtained from correlated observations from subunits or members in each

cluster, where clusters may be independent but measures from members within each cluster are not. For example, in longitudinal studies or familial studies, measures from observations of the same subject or the same family are not independent but correlated. The effective sample size for clustered data will be different from the number of observations in clusters due to intracluster dependence. Often times, because of the positive intracluster dependence, the variances of the test statistics are underestimated, and as a consequence, the resulting p values are smaller than what they should be. The popular generalized estimating equations (GEE) approach (Liang and Zeger 1986) provides a general regression modeling strategy that accounts for unspecified intracluster dependence, which can be applied to compare two groups by regressing the response to the group indicator and testing whether its coefficient equals zero. Unlike rank-based procedures, however, it is not invariant to monotonic transformations of the data.

Several recent developments have extended the Wilcoxon rank-sum test to allow two-sample comparisons for clustered data. Rosner and Grove (1999) proposed a Mann-Whitney U statistic for clustered data which corrects the variance of the test statistic for four types of intracluster correlation, but did not provide large sample theory. Rosner, Glynn, and Lee (2003) proposed an extended Wilcoxon rank-sum test under the assumptions that all subunit observations (or members) from the same cluster (i.e., subject) belong to the same treatment group, that observations within any cluster are exchangeable, and that the intracluster dependence does not vary across groups. With a test statistic in similar form as the standard Wilcoxon rank-sum test after ranking all the observations combined, they derived the asymptotic mean and variance under the clustered setting that accommodates unequal cluster sizes and possible stratification. Rosner, Glynn, and Lee (2006a) extended their approach in Rosner *et al.* (2003) to accommodate the situation where members of a single cluster may be assigned to different treatment groups, but still assumed exchangeability with the same intracluster dependence across groups. The assumptions of Rosner *et al.* (2003) were relaxed in the approach of Datta and Satten (2005), which is based on within-cluster resampling (Hoffman, Sen, and Weinberg 2001) and remains valid when the cluster sizes are informative. More recently, Dutta and Datta (2016b) extended the idea of within-cluster resampling to further accommodate the case where the number of members in a group within a cluster is informative.

On the other hand, for the one-sample problems or paired comparison problems, Rosner, Glynn, and Lee (2006b) extended the Wilcoxon signed-rank test to clustered data by adjusting the variance of the standard test statistic, assuming a common intracluster correlation across clusters. The cluster sizes are allowed to vary, but the method does not consider the informative cluster sizes where the distribution of paired differences within a cluster depends on the cluster size. Datta and Satten (2008) proposed a signed-rank test based on sampling members within cluster that accounts for informative cluster sizes.

Existing implementations of Wilcoxon rank-sum and signed-rank tests are mostly standard versions where the data are assumed to be independent. They have long been available in standard software, such as `wilcox.test` in the base package **stats** of R (R Core Team 2020), `PROC NPAR1WAY` of SAS (SAS Institute Inc. 2013), and `ranksum` and `signrank` of Stata (StataCorp. 2015). Permutation methods are available in **StatXact** (Cytel Inc. 2013). These implementations cannot, however, handle clustered data in general. One exception is, for instance, multi-center randomized clinical trials, where the centers can be viewed as blocks across the treatment groups. In this case, the R package **coin** (Hothorn, Hornik, van de Wiel, and Zeileis 2008), which provides a powerful toolkit for conditional inferences, can be applied.

The general situation where the sampling unit is cluster, however, is not under the inference framework of the package **coin**.

Despite the popularity of clustered data arising from a wide range of applications such as biomedical and social science studies, the extensions of Wilcoxon rank-sum and signed-rank tests reviewed above have not been implemented in R until very recently. The package **clusrank** (Jiang 2020) that we developed made its first appearance on the Comprehensive R Archive Network (CRAN; https://CRAN.R-project.org/package=clusrank) in December 2015. The package provides implementation of the available rank-sum tests (Rosner *et al.* 2003; Datta and Satten 2005; Rosner *et al.* 2006a) and signed-rank tests (Rosner *et al.* 2006b; Datta and Satten 2008) for clustered data. The methods are grouped into two categories by their authors: RGL for those by Rosner, Glynn, and Lee; and DS for those by Datta and Satten. Note that the RGL methods are available in SAS codes from Dr. Rosner's website (https://sites.google.com/a/channing.harvard.edu/bernardrosner/channing/) and in Stata package **cluswilcox** from Dr. Lee's website (http://cls.umd.edu/mtlee/). R codes for the DS methods are available from Dr. Datta's website (http://www.somnathdatta.org/software/), which were later put into the R package **ClusterRankTest** (Dutta and Datta 2016a) in April, 2016. Modeled after the familiar function `wilcox.test` in the base R package **stats**, the package **clusrank** that we developed unifies both RGL and DS methods under a user-friendly interface that accommodates the specifications from these methods. This makes it very easy for users to compare the performance of different approaches under various settings (see Appendix A).

The rest of the article is organized as follows. The recently available Wilcoxon rank-sum tests and signed-rank tests for clustered data are briefly reviewed in Sections 2 and 3, respectively. The usage of the unified user-level function and major input arguments are described in Section 4. Illustrations of how to access the implemented methods using both simulated data and a real dataset from an ophthalmological study are presented in Section 5. A discussion concludes in Section 6. A comparison study of selected methods that has not been studied previously are reported in Appendix A.

2. Rank-sum test for clustered data

The Wilcoxon rank-sum test is used for two-sample comparison. Let X_{ij} be the j-th observation in the i-th cluster, $1 \leq i \leq N$, $1 \leq j \leq n_i$, where n_i is the size of cluster i. Let δ_{ij} be the group indicator of X_{ij}; $\delta_{ij} = 1$ if X_{ij} is in group 1, and $\delta_{ij} = 0$ if X_{ij} is in group 2. Let R_{ij} be the rank of X_{ij} among all the observations. The observed data consist of $(\mathbf{X}, \boldsymbol{\delta}) = \{X_{ij}, \delta_{ij} : 1 \leq j \leq n_i; 1 \leq i \leq N\}$. Clusters are assumed to be independent, while subunit observations within each cluster are not. The null hypothesis H_0 to be tested is that there is no difference between the two groups; i.e., the distribution of X_{ij} remains the same regardless of the group indicator δ_{ij}. Although the original rank sum test of Wilcoxon (1945) was to test that the two distributions have the same mean versus not, later on the test was formulated with null hypothesis that the two distribution are the same versus not (Fay and Proschan 2010). As we are extending the `wilcox.test` function in R to handle clustered data, we keep in the output the null and alternative hypotheses the same as those in `wilcox.test`; that is, the true difference in locations is equal to zero or not.

2.1. RGL method with cluster-level grouping

The RGL Wilcoxon rank-sum test (Rosner *et al.* 2003) was designed for the scenario where the treatment group is assigned at the cluster-level: i.e., $\delta_{ij} = \delta_i$ for all $1 \leq j \leq n_i$. Define $R_{i+} = \sum_{j=1}^{n_i} R_{ij}$, the sum of the observed ranks of the subunits in the i-th cluster. The Wilcoxon rank-sum statistic is

$$W = \sum_{i=1}^{N} \delta_i R_{i+}. \tag{1}$$

The rationale of the RGL test procedure is random permutation conditioning on the observed R_{i+}, $i = 1, \ldots, N$. Like all permutation-based approaches, the RGL method assumes that subunit observations within each cluster are exchangeable and that the intracluster dependence remains the same across groups. To derive the sampling distribution of W given R_{i+}'s, Rosner *et al.* (2003) stratified on the cluster sizes and investigated R_{i+} between two groups for each cluster size. Let G be the maximum cluster size; i.e., $G = \max_{1 \leq i \leq N} n_i$. Then W in Equation (1) can be written as

$$W = \sum_{g=1}^{G} \sum_{i \in I_g} \delta_i R_{i+} = \sum_{g=1}^{G} W_g, \tag{2}$$

where I_g is the set of indices of clusters whose size is g and $W_g = \sum_{i \in I_g} \delta_i R_{i+}$. The null distribution of W conditioning on R_{i+}'s is obtained by combining all possible permutations of W_g for each cluster size $g \in \{1, \ldots, G\}$. Let N_g be the number of clusters of size g, among which m_g are in group 1 and n_g are in group 2. The total number of permutations is $\prod_{g=1}^{G} \binom{N_g}{m_g}$.

When N is large exhaustive permutation is infeasible and Rosner *et al.* (2003) proposed an asymptotic test statistic $Z = (W - \mathsf{E}(W))/\sqrt{\mathsf{VAR}(W)}$, where $\mathsf{E}(W) = \sum_{g=1}^{G} \mathsf{E}(W_g)$, and $\mathsf{VAR}(W) = \sum_{g=1}^{G} \mathsf{VAR}(W_g)$. Under H_0, for clusters with size g, the distribution of δ_i is Bernoulli with probability m_g/N_g, and we have $\mathsf{E}(\delta_i) = m_g/N_g$, $\mathsf{VAR}(\delta_i) = m_g n_g/N_g^2$, and $\mathsf{COV}(\delta_i, \delta_j) = -m_g n_g/[N_g^2(N_g - 1)]$. The specific expressions needed are shown to be:

$$\mathsf{E}(W_g) = m_g R_{++,g}/N_g, \qquad \mathsf{VAR}(W_g) = \frac{m_g n_g}{N_g(N_g - 1)} \sum_{i \in I_g} \left(R_{i+} - \frac{R_{++,g}}{N_g} \right)^2,$$

where $R_{++,g} = \sum_{i \in I_g} R_{i+}$. Rosner *et al.* (2003) showed that under mild conditions, the asymptotic distribution of the test statistic Z is standard normal. The method can be extended to the case of stratified data.

While Rosner *et al.* (2003) compare groups at each cluster size, the imbalance of sample sizes between two groups across cluster size strata may result in inefficiency (Datta and Satten 2005, p. 911). If only one group shows up at a certain cluster size, the corresponding data will be ignored as no comparison at this cluster size can be made. Furthermore, the rank-sum statistic scores clusters by the sum of ranks of the cluster members, which is expected to perform best if intracluster dependence is weak; otherwise, when the effective number of independent observations per cluster becomes smaller, it overweights larger clusters, and, hence, may have lower efficiency.

2.2. DS method with subunit-level grouping

Unlike the RGL method, the DS method allows subunit observations within the same cluster to have different group memberships. The rationale is rooted in the within-cluster resampling principle of Hoffman *et al.* (2001). The test statistic is constructed by randomly picking one observation from each cluster to form a pseudo-sample and averaging the standard Wilcoxon rank-sum statistic over all possible pseudo-samples. Let X_i^* be a random pick from the i-th cluster in the pseudo-sample, and δ_i^* be its group membership. The Wilcoxon rank-sum statistic for the pseudo-sample is

$$W^* = \frac{1}{N+1} \sum_{i=1}^{N} \delta_i^* R_i^*,$$

where R_i^* is the rank of X_i^* among the pseudo-sample. The test statistic of the DS method is

$$Z = \frac{S - \mathsf{E}(S)}{\sqrt{\widehat{\mathsf{VAR}}(S)}},$$

where $S = \mathsf{E}(W^* | \mathbf{X}, \boldsymbol{\delta})$.

Datta and Satten (2005) derived the quantities needed to calculate the test statistic, using mid-ranks to allow for ties in the data. Let $F_i(x) = n_i^{-1} \sum_{k=1}^{n_i} I(X_{ik} \leq x)$ be the empirical distribution function of the observations from cluster i and define $F_i(x-) = n_i^{-1} \sum_{k=1}^{n_i} I(X_{ik} < x)$. It can be shown that

$$S = \frac{1}{N+1} \sum_{i=1}^{N} \sum_{k=1}^{n_i} \frac{\delta_{ik}}{n_i} \left[1 + \frac{1}{2} \sum_{j \neq i} \left\{ F_j(X_{ik}) + F_j(X_{ik}-) \right\} \right].$$

The expectation turns out to be

$$\mathsf{E}(S) = \mathsf{E}(W^*) = \frac{1}{2} \sum_{i=1}^{N} \frac{n_i}{N}.$$

The variance term $\mathsf{VAR}(S)$ can be estimated by $\widehat{\mathsf{VAR}}(S) = \sum_{i=1}^{N} \{\hat{W}_i - \mathsf{E}(W_i)\}^2$, where

$$\hat{W}_i = \frac{1}{2n_i(N+1)} \sum_{k=1}^{n_i} \left[(N-1)\delta_{ik} - \sum_{j \neq i}^{N} \frac{n_{j1}}{n_j} \right] \left[\hat{F}(X_{ik}) + \hat{F}(X_{ik}-) \right],$$

$$\mathsf{E}(W_i) = \frac{N}{N+1} \left(\frac{n_{i1}}{n_i} - \frac{1}{N} \sum_{j=1} \frac{n_{j1}}{n_j} \right),$$

with n_{j1} being the number of subunits in group 1 from cluster j, and $\hat{F} = \sum_{i=1}^{N} n_i F_i / \sum_{i=1}^{N} n_i$, the pooled empirical distribution function of the observations. The asymptotic distribution of Z is standard normal under mild conditions (Datta and Satten 2005, p. 910).

The DS method can be generalized to the comparison of location among m treatment groups, $m \geq 3$, with the test statistic constructed from a quadratic form of group-wise rank-sum vector. This method allows arbitrary intracluster dependence structure (not necessarily exchangeable as assumed in the RGL method) within each cluster and remains valid when

treatment affects the correlation structure. However, this test cannot be applied to strictly contralateral data (e.g., when each subject in an eye study has exactly one eye under each treatment) due to violation of the assumptions required by the asymptotic theory.

2.3. RGL method with subunit-level grouping

Rosner *et al.* (2006a) extended the RGL method to allow subunit-level grouping; i.e., for each cluster i, treatment group indicator δ_{ij} may take different values for $j = 1, \ldots, n_i$. The idea of this method can be easily explained with balanced data, where all cluster sizes are equal; i.e., $n_i = g$ for all i. The rank-sum statistic is

$$W_{g,N} = \sum_{i=1}^{N} \sum_{j=1}^{g} \delta_{ij} R_{ij},$$

where R_{ij} is the rank of X_{ij} among all the observed data. A cluster may have $q \in \{0, 1, \ldots, g\}$ subunits in group 1. The sampling distribution of $W_{g,N}$ is derived from a two-stage randomization: first, each cluster i is randomly assigned to a random number Q_i according to the observed grouping distribution Q; then, within cluster i, a random Q_i out of g subunits are assigned to group 1 while all the rest are assigned to group 2. Essentially, the first stage determines how many subunits are in group 1 in a cluster and the second stage determines which they are. The two-stage randomization process can be used to devise a random permutation test to exhaust all possibilities for small N and g.

It can be shown that

$$\mathsf{E}(W_{g,N}) = \frac{gN+1}{2} \sum_{q=0}^{g} q N_q,$$

where N_q is the number of clusters with q members in group 1. The variance of $W_{g,N}$ can be estimated by

$$\widehat{\mathsf{VAR}}(W_{g,N}) = \frac{N^2}{(N-1)g^2} \widehat{\mathsf{VAR}}(Q) s_B^2 + N \hat{\mathsf{E}}[Q(g-Q)] s_W^2 / g,$$

where $s_B^2 = \sum_{i=1}^{N} \{R_{i+} - g(gN+1)/2\}^2/N$, $s_W^2 = \sum_{i=1}^{N} \sum_{j=1}^{g} (R_{ij} - R_{i+}/g)^2/\{N(g-1)\}$, and $\widehat{\mathsf{VAR}}$ and $\hat{\mathsf{E}}$ are operated on the empirical distribution of Q. Rosner *et al.* (2006a) showed that

$$Z_{g,N} = \frac{W_{g,N} - \mathsf{E}(W_{g,N})}{\sqrt{\widehat{\mathsf{VAR}}(W_{g,N})}}$$

converges to a standard normal distribution $N \to \infty$ provided that $\lim_{N\to\infty} N_q/N = \xi_q$, where $0 \le \xi_q \le 1$, if $0 < q < g$; or, $0 \le \xi_q < 1$, if $q \in \{0, g\}$. This test is equivalent to the RGL test for balanced data when the treatment is assigned at the cluster-level.

For unbalanced data, let $N^{(g)}$ be the number of clusters of size g such that $N = \sum_{g=1}^{G} N^{(g)}$, where $G = \max_{1\le i\le N} n_i$ is the maximum cluster size. A test procedure can be constructed by efficiently combining $W_{g,N^{(g)}}$ across all $g \in \{1, \ldots, G\}$. Rosner *et al.* (2006a) proposed to base the test on a combined estimator $\hat{\theta}_N$ of θ, the probability that an observation in group 1 is greater than that of an observation in group 2, which is 1/2 under the null hypothesis. Standardized by a variance estimator $\widehat{\mathsf{VAR}}(\hat{\theta}_N)$, the test statistic $(\hat{\theta}_N - 1/2)/\widehat{\mathsf{VAR}}^{1/2}(\hat{\theta}_N)$ converges to a standard normal distribution as $N \to \infty$ under mild conditions.

3. Signed-rank test for clustered data

The Wilcoxon signed-rank test is often used for paired data comparisons. Let X_{ij} be the paired-difference score for the j-th pair in the i-th cluster, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. The null hypothesis H_0 is that the marginal distribution of X_{ij} is symmetric around 0. Note that unlike the rank-sum test, all subjects belong to a single group in the paired comparison setting. Let R_{ij} be the rank of $|X_{ij}|$ among $\{|X_{ij}|, i = 1, \ldots, N, j = 1 \ldots, n_i\}$. Let $S_{ij} = V_{ij}R_{ij}$ be the signed-rank, where $V_{ij} = \text{sign}(X_{ij})$.

3.1. RGL method: Uninformative cluster size

The RGL method for the rank-sum test (Rosner *et al.* 2003) was adapted to the signed-rank test in Rosner *et al.* (2006b). For balanced data where $n_i = g$ for all i, the clustered Wilcoxon signed-rank statistic is

$$T = \sum_{i=1}^{N} S_{i+} = \sum_{i=1}^{N} \sum_{j=1}^{g} R_{ij}V_{ij},$$

where $S_{i+} = \sum_{j=1}^{g} S_{ij}$ is the rank sum within the i-th cluster and only nonzero X_{ij} are considered in the computation of signed-ranks. The null sampling distribution of T can be obtained from a randomization at the cluster-level conditional on S_{i+}'s. Let δ_i, $i = 1, \ldots, N$, be independent and identically distributed random variables with equal probability being 1 and -1; this distribution has variance 1. The conditional distribution of T given S_{i+}'s is the same as that of $T_p = \sum_{i=1}^{N} \delta_i S_{i+}$. For small N, it is possible to assess the significance of T by enumerating all 2^N possibilities and computing the tail probability. A large sample test can be constructed with $\mathsf{E}(T) = 0$ and $\mathsf{VAR}(T) = \sum_{i=1}^{n} S_{i+}^2$. As $N \to \infty$, $T/\mathsf{VAR}^{1/2}(T)$ converges to a standard normal distribution provided $g < \infty$.

For unbalanced data, Rosner *et al.* (2006b) considered a stratified statistic

$$T = \sum_{i=1}^{N} w_i \bar{S}_i,$$

where $\bar{S}_i = S_{i+}/n_i$ and $w_i = 1/\mathsf{VAR}(\bar{S}_i)$ under H_0. The variance estimator $\widehat{\mathsf{VAR}}(\bar{S}_i)$ is obtained assuming a shared intracluster correlation coefficient. The randomization distribution of T given $\bar{S}_i$'s is that of $T_p = \sum_{i=1}^{N} \delta_i w_i \bar{S}_i$, which facilitates a random permutation test for small N. The large sample test statistic is $T/(\sum_{i=1}^{N} \hat{w}_i^2 \bar{S}_i^2)^{1/2}$, where $\hat{w}_i = 1/\widehat{\mathsf{VAR}}(\bar{S}_i)$. It converges to a standard normal distribution as $N \to \infty$ provided that $G < \infty$ and $\lim_{N\to\infty} N_g/N \to \xi_g$ where $0 \leq \xi_g \leq 1$ for all $g \in \{1, \ldots, G\}$. This method assumes that the cluster size distribution is uninformative, and, hence, is not valid when the distribution of paired differences within a cluster depends on the cluster size.

3.2. DS method: Informative cluster size

Datta and Satten (2008) followed the same principle of within-cluster resampling as in Datta and Satten (2005) in the context of clustered paired data to develop a signed-rank test. This method allows informative cluster sizes as long as the marginal distributions of X_{ij}'s are identical for all i and j. Suppose that from the i-th cluster, paired difference X_{ij}, denoted by X_i^*, is randomly picked and pooled to form a pseudo-sample. Let R_i^* be the mid-rank of $|X_i^*|$,

$i = 1, \ldots, N$ to allow ties in the data, and let $V_i^* = \text{sign}(X_i^*)$. A standard Wilcoxon signed-rank test statistic for the pseudo-sample is $\sum_{i=1}^N S_i^*$, where $S_i = V_i^* R_i^*$. The DS method is based on $T = \mathsf{E}(\sum_{i=1}^N S_i^* | \mathbf{X})$, where $\mathbf{X} = \{X_{ij} : 1 \le i \le N; 1 \le j \le n_i\}$.

To compute T, let

$$\hat{H}_i(x) = \frac{1}{2}\{F_i(x) + F_i(x-)\},$$

where $F_i(x)$ is the empirical distribution function of the observations in cluster i at x and $F_i(x-)$ is the left limit of $F_i(x)$ at x as in the DS method for the rank-sum test. It turns out that

$$T = \sum_{i=1}^N \frac{n_i^+ - n_i^-}{n_i} + \sum_{i=1}^N \frac{1}{n_i} \sum_{k=1}^{n_i} V_{ik} \sum_{j \ne i} H_j(|X_{ij}|),$$

where $n_i^+ = \sum_{j=1}^{n_i} I(X_{ij} > 0)$ and $n_i^- = \sum_{j=1}^{n_i} I(X_{ij} < 0)$. The variance can be estimated by $\widehat{\mathsf{VAR}}(T) = \sum_{i=1}^N \hat{S}_i^2$, where

$$\hat{S}_i = \frac{n_i^+ - n_i^-}{n_i} + \frac{N-1}{n_i} \sum_{k=1}^{n_i} V_{ik} \hat{H}(|X_{ik}|),$$

with $\hat{H}(x) = \sum_{i=1}^N n_i \hat{H}_i(x) / \sum_{i=1}^N n_i$. The standardized test statistic $Z = T / \sqrt{\widehat{\mathsf{VAR}}(T)}$ converges to a standard normal distribution under mild conditions.

Note that the null distribution being tested using the DS method is the distribution of the paired difference of a randomly selected pair from a randomly selected cluster, regardless of the cluster size. Whereas the null distribution of most signed-rank tests is the common distribution of a randomly selected paired difference conditional on the size of the cluster it belongs to. Therefore, the latter framework is a special case of the former. Also, the DS method accounts for the cluster size by assigning equal weight to each cluster instead of each paired difference, e.g., a paired difference from a larger cluster will be assigned a smaller weight than a paired difference from a smaller cluster.

4. Usage

Package **clusrank** provides a unified interface to all the methods reviewed in Sections 2 and 3 through function `clusWilcox.test` which has arguments:

```
clusWilcox.test(x, ...)
```

Argument `x` can be either a numeric vector or a formula. The default interface is called if `x` is a numeric vector, in which case the interface is designed to mimic that of the function `wilcox.test`, i.e., the default S3 method of `clusWilcox.test` has arguments:

```
clusWilcox.test(x, y = NULL, cluster = NULL, group = NULL, stratum = NULL,
  data = NULL, alternative = c("two.sided", "less", "greater"),
  mu = 0, paired = FALSE, exact = FALSE, B = 2000,
  method = c("rgl", "ds", "dd"), ...)
```

The arguments `x`, `y`, `alternative`, `mu`, `paired`, and `exact` have the same meaning as those in the familiar default interface of `wilcox.test`. Clustered rank-sum test is requested if `paired = FALSE`; clustered signed-rank test is requested if `paired = TRUE`. For both tests, the RGL and DS methods are requested with `method` set to be `"rgl"` and `"ds"`, respectively.

For clustered rank-sum tests, `x`, `cluster`, and `group` are required, which are of the same length; `cluster` and `group` specify the cluster membership and group membership, respectively. Argument `y` is not used for clustered rank-sum tests. The group assignment can be at either the cluster or the subunit level. When using RGL method for data with treatment group assigned at the cluster-level, an optional argument `stratum` can be specified to account for the stratification and therefore provide a more powerful test. The variables `x`, `cluster`, `group` and `stratum` can be found from a data frame specified by argument `data`.

For clustered signed-rank tests, `x` and `cluster` are required while `group` is not needed because the data are paired differences. Argument `x` can be the paired difference between the pre- and post-treatment observations; alternatively, `x` and `y` can specify the pre- and post-treatment observations, respectively. This interface is also similar to that of `wilcox.test`.

Besides the asymptotic tests, exact tests for RGL clustered rank-sum test with cluster-level grouping and the RGL clustered signed-rank test can be requested by setting `exact = TRUE`. However, the exact tests can be very computationally intensive even for a moderate sample size, so it is recommended only for small samples. The `wilcox.test` function by default computes exact p values for sample size less than 50. For clustered data, we recommend 50 observations (instead of clusters) as the threshold for the same reason. The package also provides a remedy by allowing a random permutation test to approximate the exact test with the argument `B`, which is the number of permutations with default value 2000. The random permutation test is also provided for other RGL tests and DS tests, for both rank-sum test and signed-rank test, even when exact permutation test is not available.

The formula interface mimics that of the `wilcox.test` too with arguments `formula`, `subset` and `na.action`, i.e., the S3 method of `clusWilcox.test` where the first argument is of class '`formula`' is given by:

```
clusWilcox.test(formula, data = parent.frame(), subset = NULL,
  na.action = na.omit, alternative = c("two.sided", "less", "greater"),
  mu = 0, paired = FALSE, exact = FALSE, B = 2000,
  method = c("rgl", "ds", "dd"), ...)
```

For clustered rank-sum tests, the left hand side of `formula` should be the data vector to be tested, and the right hand side of `formula` should contain the variables indicating group, cluster and possibly stratum. Except for the group variable, other variables on the right need to be indicated with special terms; for example, in formula `z ~ group + cluster(cid) + stratum(sid)`, `group` identifies the grouping variable, `cid` identifies the clusters, and `sid` identifies the stratum. See Section 5 for detailed illustrations. For clustered signed-rank test, the left hand side of the formula is the paired differences and the right hand side comprises the cluster id. Neither `group` or `stratum` is applicable for signed-rank tests. Other arguments are identical to those in the default interface.

5. Illustrations

5.1. Rank-sum test for clustered data

We use a scheme that is similar to the simulation study in Datta and Satten (2005) to generate data for clustered rank-sum test with both balanced and unbalanced data. For group $g \in \{0, 1\}$, the observations in a cluster are generated as

$$X = \exp(Z_g) + \delta g,$$

where Z_g is a standard multivariate normal random vector with mean zero and an exchangeable or autoregressive of order 1 (AR1) correlation matrix with correlation parameter ρ_g, and δ is the group difference. The AR1 correlation structure is common in longitudinal studies, and it can be used to investigate the robustness of the methods when the intracluster correlation is not exchangeable. A correlation matrix of dimension `dim` with correlation parameter `rho` can be generated as follows, with `ex` for exchangeable and `ar1` for AR1.

```
R> ex  <- function(dim, rho) {
+    diag(1 - rho, dim) + matrix(rho, dim, dim)
+  }
R> ar1 <- function(dim, rho) {
+    rho ^ outer(1:dim, 1:dim, function(x, y) abs(x - y))
+  }
```

Below is a simple implementation that allows different levels of grouping (cluster-level and subunit-level) and unequal cluster sizes. Package **mvtnorm** (Genz *et al.* 2016; Genz and Bretz 2009) is used to generate multivariate normal random vectors.

```
R> library("mvtnorm")
R> datgen.sum <- function(nclus, maxclsize, delta = 0, rho = c(0.1, 0.1),
+    corr = ex, misrate = 0, clusgrp = TRUE) {
+      nn <- nclus * maxclsize
+      Sigma1 <- corr(maxclsize, rho[1])
+      Sigma2 <- corr(maxclsize, rho[2])
+      y1 <- c(t(rmvnorm(nclus, sigma = Sigma1)))
+      y2 <- c(t(rmvnorm(nclus, sigma = Sigma2)))
+      group <- rep(c(0, 1), each = nn)
+      if (!clusgrp) group  <- sample(group, nn, FALSE)
+      cid <- rep(1:(2 * nclus), each = maxclsize)
+      x <- exp(c(y1, y2)) + delta * group
+      dat <- data.frame(x = x, grp = group, cid = cid)
+      drop <-  sort(sample(1:(2 * nn), size = misrate * (2 * nn), FALSE))
+      if (misrate == 0) dat else dat[-drop, ]
+  }
```

There are two required inputs: `nclus` for the number of clusters in each group and `maxclsize` for the maximum cluster size. The difference between groups is specified by `delta`. The group specific intracluster correlation parameter ρ_g is set by `rho`, a numeric vector of length 2, one

for each group. The `corr` argument specifies the function to construct correlation matrix with correlation coefficients in `rho`. To allow unequal cluster sizes, `misrate` specifies the missing rate at which an subunit in a cluster to be excluded at random; when `misrate = 0`, balanced data with equal cluster size are generated. The grouping level is controlled by the logical argument `clusgrp`: `TRUE` for cluster-level and `FALSE` for subunit-level. The function returns a data set with three columns: observation `x`, grouping id `grp`, and cluster id `cid`.

For the replicability of the following demonstration, a random seed is set.

```
R> set.seed(1234)
```

To illustrate the clustered rank-sum test, a data set with 10 clusters of size 3 in each group is generated with $\delta = 0$, exchangeable correlation structure, $\rho_0 = \rho_1 = 0.9$, and cluster-level treatment assignment.

```
R> dat.cl <- datgen.sum(10, 3, 0, c(.9, .9), ex, 0, TRUE)
```

The first and last 6 rows of the data frame look like this:

```
R> cbind(head(dat.cl, 6), "head / tail" = "", tail(dat.cl, 6))
```

```
          x grp cid head / tail         x grp cid
1 0.7322161   0   1             2.3997628   1  19
2 1.1708839   0   1             3.0184715   1  19
3 1.5112823   0   1             4.2536708   1  19
4 0.2516183   0   2             0.6759148   1  20
5 0.6050972   0   2             1.4343341   1  20
6 0.6199984   0   2             0.5985389   1  20
```

We first use the formula interface to perform the RGL asymptotic test:

```
R> library("clusrank")
R> clusWilcox.test(x ~ grp + cluster(cid), dat.cl, method = "rgl")
```

```
        Clustered Wilcoxon rank sum test using Rosner-Glynn-Lee method

data:  x; group: grp; cluster: cid; (from dat.cl)
number of observations: 60;  number of clusters: 20
Z = -1.3613, p-value = 0.1734
alternative hypothesis: true difference in locations is not equal to 0
```

The exact test of the RGL method can be done for data with a small number of clusters when treatment is assigned at cluster-level:

```
R> clusWilcox.test(x ~ grp + cluster(cid), dat.cl, method = "rgl",
+    exact = TRUE, B = 0)
```

```
        Clustered Wilcoxon rank sum test using Rosner-Glynn-Lee method
        (exact permutation)

data:  x; group: grp; cluster: cid; (from dat.cl)
number of observations: 60;  number of clusters: 20
W = 757, p-value = 0.1789
alternative hypothesis: true location is not equal to 0
```

Note that in our experiment, the speed of the exact test is fast when each group contains 10 or 11 clusters, but more clusters can dramatically increase the computing time. Therefore, we recommend the bootstrap option to approximate the exact test when needed.

The numerical interface is illustrated with the DS method by setting `method = "ds"`:

```
R> clusWilcox.test(x, group = grp, cluster = cid, data = dat.cl,
+    method = "ds")

        Clustered Wilcoxon rank sum test using Datta-Satten method

data:  x; group: grp; cluster: cid; (from dat.cl)
number of observations: 60;  number of clusters: 20
Z = 1.3967, p-value = 0.1625
alternative hypothesis: true difference in locations is not equal to 0
```

The test statistics from the two methods (using the asymptotic theory) are very close to each other. Note that, when using the code provided online by the authors of Rosner *et al.* (2003) and Datta and Satten (2005), statistics with different signs may occur. This is a result of using the opposite group in calculating the statistic. To get the matching results, one just needs to switch the group ids.

For data with cluster-level groups, the RGL method allows an extra stratum variable to accommodate stratification in the data. For illustration, we simply add an extra column `strat` to the dataset and perform the RGL test:

```
R> dat.cl$strat <- rep(rep(1:2, each = 15), 2)
R> cbind(head(dat.cl, 6), "head / tail" = "", tail(dat.cl, 6))

          x grp cid strat head / tail         x grp cid strat
1 0.7322161   0   1     1             2.3997628   1  19     2
2 1.1708839   0   1     1             3.0184715   1  19     2
3 1.5112823   0   1     1             4.2536708   1  19     2
4 0.2516183   0   2     1             0.6759148   1  20     2
5 0.6050972   0   2     1             1.4343341   1  20     2
6 0.6199984   0   2     1             0.5985389   1  20     2

R> clusWilcox.test(x ~ grp + cluster(cid) + stratum(strat), dat = dat.cl,
+    method = "rgl")
```

```
        Clustered Wilcoxon rank sum test using Rosner-Glynn-Lee method

data:  x; group: grp; cluster: cid; stratum: strat; (from dat.cl)
number of observations: 60;  number of clusters: 20
Z = -1.3271, p-value = 0.1845
alternative hypothesis: true difference in locations is not equal to 0
```

The DS method can compare more than two groups. If user set `method = "rgl"` while the data contains more than 2 groups, a warning will show up and the test will be switched to use the DS method. We illustrate this by assigning 4 groups to this data:

```
R> dat.cl$grp <- rep(1:4, each = 15)
R> cbind(head(dat.cl, 6), "head / tail" = "", tail(dat.cl, 6))

          x grp cid strat head / tail         x grp cid strat
1 0.7322161   1   1     1             2.3997628   4  19     2
2 1.1708839   1   1     1             3.0184715   4  19     2
3 1.5112823   1   1     1             4.2536708   4  19     2
4 0.2516183   1   2     1             0.6759148   4  20     2
5 0.6050972   1   2     1             1.4343341   4  20     2
6 0.6199984   1   2     1             0.5985389   4  20     2

R> clusWilcox.test(x ~ grp + cluster(cid), dat = dat.cl, method = "ds")

        Clustered Wilcoxon rank sum test using Datta-Satten method using
        Chi-square test

data:  x; group: grp; cluster: cid; (from dat.cl)
number of observations: 60;  number of clusters: 20
number of groups: 4
chi-square test statistic = 2.0471, p-value = 0.5627
```

5.2. Signed-rank test for clustered data

To illustrate clustered signed-rank tests, we generate data from a scheme slightly modified from simulation scenario 1 in Datta and Satten (2008). The paired differences in a cluster are generated as

$$X = \text{sign}(Z)\exp(|Z|),$$

where Z is a multivariate normal random vector with mean δ and an exchangeable or AR1 correlation structure with parameter ρ. This is implemented by the following function:

```
R> datgen.sgn <- function(nclus, maxclsize, delta = 0, rho = 0.1,
+    corr = ex, misrate = 0) {
+      nn <- nclus * maxclsize
+      Sigma <- corr(maxclsize, rho)
+      z <- delta + c(t(rmvnorm(nclus, sigma = Sigma)))
```

```
+      x <- sign(z) * exp(abs(z))
+      cid <- rep(1:nclus, each = maxclsize)
+      dat <- data.frame(x = x, cid = cid)
+      drop <- sort(sample(1:nn, size = misrate * nn, FALSE))
+      if (misrate == 0) dat else dat[-drop, ]
+  }
```

The arguments of `datgen.sgn` match those of `datgen.sum`, except that it does not need a `clusgrp` argument because the data are already differences between two groups.

For illustration, we generate a dataset that consists of 10 clusters of size 3, with exchangeable correlation parameter $\rho = 0.5$.

```
R> dat.sgn <- datgen.sgn(10, 3, cor = ex, rho = 0.5)
R> cbind(head(dat.cl, 6), "head / tail" = "", tail(dat.cl, 6))
```

```
          x grp cid strat head / tail         x grp cid strat
1 0.7322161   1   1     1             2.3997628   4  19     2
2 1.1708839   1   1     1             3.0184715   4  19     2
3 1.5112823   1   1     1             4.2536708   4  19     2
4 0.2516183   1   2     1             0.6759148   4  20     2
5 0.6050972   1   2     1             1.4343341   4  20     2
6 0.6199984   1   2     1             0.5985389   4  20     2
```

The RGL signed-rank test is performed with:

```
R> clusWilcox.test(x ~ cluster(cid), dat.sgn, paired = TRUE, method = "rgl")
```

```
        Clustered Wilcoxon signed rank test using Rosner-Glynn-Lee method

data:  x; cluster: cid; (from dat.sgn)
number of observations: 30;  number of clusters: 10
Z = 0.47709, p-value = 0.6333
alternative hypothesis: true shift in location is not equal to 0
```

The DS signed-rank test is performed with

```
R> clusWilcox.test(x ~ cluster(cid), dat.sgn, paired = TRUE, method = "ds")
```

```
        Clustered Wilcoxon signed rank test using Datta-Satten method

data:  x; cluster: cid; (from dat.sgn)
number of observations: 30;  number of clusters: 10
Z = 0.45109, p-value = 0.6519
alternative hypothesis: true shift in location is not equal to 0
```

5.3. A real data example

The package contains a real dataset named `amd` from a retrospective observational study (Ferrara and Seddon 2015) that aims to characterize the phenotype associated with a rare variant in the complement factor H (CFH) R1210C, a protein involved in the age-related macular degeneration (AMD). The data contains measures of 283 eyes from 143 patients, among whom 62 had the rare variant and 81 did not. The clusters are the patients and the subunits are the eyes. The outcome variable was the AMD grading score based on the clinical age-related maculopathy staging (CARMS), which has 5 levels. The first 3 levels correspond to the size of drusen which is an intermediate marker of AMD, while level 4 and 5 correspond to different types of AMD, with level 4 indicating the presence of geographic atrophy (GA) and level 5 indicating the presence of choroidal neovascularization (CNV). The Wilcoxon rank-sum test is to be carried out on two subsets: 1) the subset of observations with CARMS grade 1, 2, 3 or 4; and 2) the subset of observations with CARMS grade 1, 2, 3 or 5. In the `amd` data, the outcome variable is `CARMS`; the patients ids indicating the clusters are stored in variable `ID`; the rare variant grouping at the cluster (patient) level is indicated by variable `Variant`.

We apply both the RGL and DS method to the first subset:

```
R> data("amd", package = "clusrank")
R> clusWilcox.test(CARMS ~ Variant + cluster(ID), data = amd,
+    subset = CARMS %in% c(1, 2, 3, 4), method = "rgl")

        Clustered Wilcoxon rank sum test using Rosner-Glynn-Lee method

data:  CARMS; group: Variant; cluster: ID; (from amd)
number of observations: 196;  number of clusters: 112
Z = -4.3993, p-value = 1.086e-05
alternative hypothesis: true difference in locations is not equal to 0

R> clusWilcox.test(CARMS ~ Variant + cluster(ID), data = amd,
+    subset = CARMS %in% c(1, 2, 3, 4), method = "ds")

        Clustered Wilcoxon rank sum test using Datta-Satten method

data:  CARMS; group: Variant; cluster: ID; (from amd)
number of observations: 196;  number of clusters: 112
Z = -4.4823, p-value = 7.384e-06
alternative hypothesis: true difference in locations is not equal to 0
```

The p values from both tests are close and less than 0.001, which implies strong evidence of an association between the presence of CFH R1210C rare variant and the CARMS grade with GA as the advanced stage.

For the RGL method, a stratifying variable `AgeSex` which categorizes the patients into 6 strata by their age and gender can be used as a control variable:

```
R> clusWilcox.test(CARMS ~ Variant + cluster(ID) + stratum(AgeSex),
+    data = amd, subset = CARMS %in% c(1, 2, 3, 4))
```

```
        Clustered Wilcoxon rank sum test using Rosner-Glynn-Lee method

data:  CARMS; group: Variant; cluster: ID; stratum: AgeSex; (from amd)
number of observations: 196;  number of clusters: 112
Z = -4.0797, p-value = 4.509e-05
alternative hypothesis: true difference in locations is not equal to 0
```

The p value is still less than 0.001 after controlling for age and gender.
We then apply the two tests to the second subset:

```
R> clusWilcox.test(CARMS ~ Variant + cluster(ID), data = amd, method = "rgl",
+    subset = CARMS %in% c(1, 2, 3, 5))

        Clustered Wilcoxon rank sum test using Rosner-Glynn-Lee method

data:  CARMS; group: Variant; cluster: ID; (from amd)
number of observations: 224;  number of clusters: 121
Z = -1.8484, p-value = 0.06455
alternative hypothesis: true difference in locations is not equal to 0

R> clusWilcox.test(CARMS ~ Variant + cluster(ID), data = amd, method = "ds",
+    subset = CARMS %in% c(1, 2, 3, 5))

        Clustered Wilcoxon rank sum test using Datta-Satten method

data:  CARMS; group: Variant; cluster: ID; (from amd)
number of observations: 224;  number of clusters: 121
Z = -2.7311, p-value = 0.006312
alternative hypothesis: true difference in locations is not equal to 0
```

This time, the p values of the two approaches are easily discernable, which is not a surprise because the methods are based on different assumptions. The DS method reports a p value of 0.006312, in contrast to 0.06455 from the RGL method. Using traditional significance level such as 0.01 or 0.05 will lead to different conclusions about the association between the presence of CFH R1210C rare variant and the symptom with CNV as the advanced stage. Again, the RGL method can be applied with age and gender controlled:

```
R> clusWilcox.test(CARMS ~ Variant + cluster(ID) + stratum(AgeSex),
+    data = amd, subset = CARMS %in% c(1, 2, 3, 5), method = "rgl")

        Clustered Wilcoxon rank sum test using Rosner-Glynn-Lee method

data:  CARMS; group: Variant; cluster: ID; stratum: AgeSex; (from amd)
number of observations: 224;  number of clusters: 121
Z = -1.8519, p-value = 0.06404
alternative hypothesis: true difference in locations is not equal to 0
```

The p value from the RGL method remains virtually unchanged after controlling for the age/gender strata.

6. Discussion

Clustered data are frequently encountered in scientific research, and rank-based tests for clustered data are an indispensable tool like their counterparts, the Wilcoxon tests for independent data. The package **clusrank** provides two schools, the RGL method and the DS method, of the recently developed rank-sum tests and signed-rank tests in a unified, higher-level, user-friendly interface. For the DS methods, our implementation gives the same results as those in the **ClusterRankTest** R package. Users need to be aware of the applicability of these tests when using them. For example, the RGL method assumes exchangeability within clusters and does not account for informative cluster sizes; the DS method cannot be applied to contralateral designs with exactly one subunit in each group within a cluster; both asymptotic tests require that the number of clusters to be reasonably large.

Implementation of other rank-based methods for clustered data can be considered in future development of the package **clusrank**. For the rank-sum test, the case of informative group size within a cluster (Dutta and Datta 2016b) has been added to the package, though it is available in the package **ClusterRankTest** (Dutta and Datta 2016a). For the signed-rank test, Larocque (2005) proposes a variance estimator based on certain sums of squares over independent clusters. Sign tests and signed-rank tests for multivariate clustered data (Larocque 2003; Larocque, Nevalainen, and Oja 2007; Haataja, Larocque, Nevalainen, and Oja 2009) and multilevel data (Larocque, Nevalainen, and Oja 2008) have been studied. Some tests allow the distributions in two groups to have different scales and/or shapes under the null hypothesis (Larocque, Haataja, Nevalainen, and Oja 2010). Making these tests available would be of interest to many users.

Acknowledgments

We thank Dr. Johanna Seddon for sharing the data from their phenotypic characterization study of the H R1210C rare variant. Rosner and Lee were supported in part by grant NIH R01EY022445 from the National Eye Institute.

References

Cytel Inc (2013). *StatXact 10: Statistical Software for Exact Nonparametric Interence*. Cambridge. URL http://www.cytel.com/.

Datta S, Satten GA (2005). "Rank-Sum Tests for Clustered Data." *Journal of the American Statistical Association*, **100**(471), 908–915. doi:10.1198/016214504000001583.

Datta S, Satten GA (2008). "A Signed-Rank Test for Clustered Data." *Biometrics*, **64**(2), 501–507. doi:10.1111/j.1541-0420.2007.00923.x.

Dutta S, Datta S (2016a). ***ClusterRankTest**: Rank Tests for Clustered Data*. R package version 1.0, URL https://CRAN.R-project.org/package=ClusterRankTest.

Dutta S, Datta S (2016b). "A Rank-Sum Test for Clustered Data When the Number of Subjects in a Group within a Cluster Is Informative." *Biometrics*, **72**(2), 432–440. doi:10.1111/biom.12447.

Fay MP, Proschan MA (2010). "Wilcoxon-Mann-Whitney or t Test? On Assumptions for Hypothesis Tests and Multiple Interpretations of Decision Rules." *Statistics Surveys*, **4**, 1–39. doi:10.1214/09-ss051.

Ferrara D, Seddon JM (2015). "Phenotypic Characterization of Complement Factor H R1210C Rare Genetic Variant in Age-Related Macular Degeneration." *JAMA Ophthalmology*, **133**(7), 785–791. doi:10.1001/jamaophthalmol.2015.0814.

Genz A, Bretz F (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T (2016). ***mvtnorm**: Multivariate Normal and t Distributions*. R package version 1.0-5, URL http://CRAN.R-project.org/package=mvtnorm.

Haataja R, Larocque D, Nevalainen J, Oja H (2009). "A Weighted Multivariate Signed-Rank Test for Cluster-Correlated Data." *Journal of Multivariate Analysis*, **100**(6), 1107–1119. doi:10.1016/j.jmva.2008.10.009.

Hoffman EB, Sen PK, Weinberg CR (2001). "Within-Cluster Resampling." *Biometrika*, **88**(4), 1121–1134. doi:10.1093/biomet/88.4.1121.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). "Implementing a Class of Permutation Tests: The **coin** Package." *Journal of Statistical Software*, **28**(8), 1–23. doi:10.18637/jss.v028.i08.

Jiang Y (2020). ***clusrank**: Wilcoxon Rank Sum Test for Clustered Data*. R package version 1.0-0, URL https://CRAN.R-project.org/package=clusrank.

Larocque D (2003). "An Affine-Invariant Multivariate Sign Test for Cluster-Correlated Data." *Canadian Journal of Statistics*, **31**, 437–455. doi:10.2307/3315855.

Larocque D (2005). "The Wilcoxon Signed-Rank Test for Cluster Correlated Data." In P Duchesne, B Rémillard (eds.), *Statistical Modeling and Analysis for Complex Data Problems*, pp. 309–323. Springer-Verlag. doi:10.1007/0-387-24555-3_15.

Larocque D, Haataja R, Nevalainen J, Oja H (2010). "Two Sample Tests for the Nonparametric Behrens-Fisher Problem with Clustered Data." *Journal of Nonparametric Statistics*, **22**(6), 755–771. doi:10.1080/10485250903469728.

Larocque D, Nevalainen J, Oja H (2007). "A Weighted Multivariate Sign Test for Cluster-Correlated Data." *Biometrika*, **94**(2), 267–283. doi:10.1093/biomet/asm026.

Larocque D, Nevalainen J, Oja H (2008). "One-Sample Location Tests for Multilevel Data." *Journal of Statistical Planning and Inference*, **138**(8), 2469–2482. doi:10.1016/j.jspi.2007.10.006.

Liang KY, Zeger SL (1986). "Longitudinal Data Analysis Using Generalized Linear Models." *Biometrika*, **73**(1), 13–22. doi:10.1093/biomet/73.1.13.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosner B, Glynn RJ, Lee MLT (2003). "Incorporation of Clustering Effects for the Wilcoxon Rank Sum Test: A Large-Sample Approach." *Biometrics*, **59**(4), 1089–1098. doi:10.1111/j.0006-341x.2003.00125.x.

Rosner B, Glynn RJ, Lee MLT (2006a). "Extension of the Rank Sum Test for Clustered Data: Two-Group Comparisons with Group Membership Defined at the Subunit Level." *Biometrics*, **62**(4), 1251–1259. doi:10.1111/j.1541-0420.2006.00582.x.

Rosner B, Glynn RJ, Lee MLT (2006b). "The Wilcoxon Signed Rank Test for Paired Comparisons of Clustered Data." *Biometrics*, **62**(1), 185–192. doi:10.1111/j.1541-0420.2005.00389.x.

Rosner B, Grove JD (1999). "Use of the Mann-Whitney U-Test for Clustered Data." *Statistics in Medicine*, **18**(11), 1387–1400. doi:10.1002/(sici)1097-0258(19990615)18:11<1387::aid-sim126>3.0.co;2-v.

SAS Institute Inc (2013). *SAS/STAT Software, Version 9.4*. Cary. URL http://www.sas.com/.

StataCorp (2015). *Stata Data Analysis Statistical Software: Release 14*. StataCorp LP, College Station. URL http://www.stata.com/.

Wilcoxon F (1945). "Individual Comparisons by Ranking Methods." *Biometrics Bulletin*, **1**(6), 80–83. doi:10.2307/3001968.

A. A simulation study

Comparisons of the RGL and DS methods under some common scenarios have not been studied in the recent literature. Such comparison can be easily done with the package **clusrank**. Using the two data generation functions defined above, we conduct a simulation study comparing their sizes and powers in a few settings. The following function generates replicates of data for a given scenario and returns the empirical power for a given significance level.

```
R> simpower <- function(nrep, level, paired, nclus, maxclsize,
+    delta, rho, corr, misrate, ...) {
+      do1rep <- function() {
+        datgen <- if (paired) datgen.sgn else datgen.sum
+        formula <- if (paired) x ~ cluster(cid)
+          else x ~ cluster(cid) + grp
+        dat <- datgen(nclus, maxclsize, delta, rho, corr, misrate, ...)
+        p.rgl <- clusWilcox.test(formula, paired = paired, data = dat,
+          method = "rgl")$p.value
+        p.ds  <- clusWilcox.test(formula, paired = paired, data = dat,
+          method = "ds" )$p.value
+        c(rgl = p.rgl, ds = p.ds)
+      }
+      sim <- t(replicate(nrep, do1rep()))
+      apply(sim, 2, function(x) mean(x < level))
+  }
```

The first two arguments, `nrep` and `level` specify the number of replications and the desired significance level, respectively. Argument `paired` is a logical scalar to switch between the two methods, `TRUE` for signed-rank tests and `FALSE` for rank-sum tests. Other arguments have the same meanings as those in `datgen.sum` or `datgen.sgn` depending on the value of `paired`. The last argument `...` is used to supply `clusgrp`, which is only needed by `datgen.sum` for the level of groups. The function returns the empirical rejection rates of the RGL and DS methods for the setting defined by the inputs.

As an example, consider the rank-sum tests in a setting with cluster-level grouping, each group containing 20 clusters of size 3, with exchangeable correlation parameter $\rho = 0.5$ in both groups. The group difference is set to be $\delta = 0$. We do this experiment with 1000 replicates at a significance level of 0.05:

```
R> simpower(1000, 0.05, FALSE, 20, 3, 0.0, c(0.5, 0.5), ex, 0,
+    clusgrp = TRUE)
```

```
  rgl    ds
0.052 0.056
```

The empirical sizes of both methods are close to the nominal level 0.05.

Similarly, a comparison for signed-rank tests can be done. This time we use an AR1 correlation setting with $\rho = 0.5$.

```
R> simpower(1000, 0.05, TRUE, 20, 3, 0, 0.5, ar1, 0)

  rgl    ds
0.055 0.057
```

Again, both methods have empirical sizes close to the nominal level. It means that the RGL method is robust to the violation of the exchangeability assumption in this setting.

A.1. Rank-sum tests for clustered data

The code for replicating the simulation study and for producing the tables below is provided in the supplementary R script `simulation.R`.

Comparison between the RGL and DS methods for the rank-sum tests has never been done previously under subunit-level treatment group. Nor has it been done when the intracluster dependence is not exchangeable. Therefore we will present these comparisons in this session. Consider two equal size groups, with 20 or 50 clusters, and with exchangeable or AR1 intracluster correlation structure as specified in `datgen.sum`. The correlation parameters for the two groups were set to be $(0.1, 0.1)$, $(0.5, 0.5)$, or $(-0.1, 0.9)$. The maximum cluster size $G = \max_i n_i$ has three levels: 2, 5, and 10. The `simpower` function makes the comparison very easy, and the `misrate` argument allows cluster sizes to be random, which is an additional scenario of interest that has not been compared before. Two missing rates, 0 and 0.5, were considered, representing balanced data and unbalanced data with missing completely at random, respectively. This is different from the informative cluster size setting studied in Datta and Satten (2005), where cluster size depends on group. The group difference δ was set to be 0, 0.2 and 0.5. For each setting, empirical rejection rates was obtained from 4000 replicates.

The results are summarized in Table 1–2. The empirical size ($\delta = 0$) of both methods are close to the nominal level 0.05 in all the settings considered in this study. The empirical power ($\delta \neq 0$) for both methods increases as the cluster size increases, and as the missing rate decreases. For balanced data, the powers of the two tests are very close regardless of the level of the treatment group assignment and the intracluster configurations. For unbalanced data from maximum cluster size 10 and missing rate 0.5, the DS method has higher power with cluster-level group assignment, while the RGL method has higher power with subunit-level group assignment. As this setting has average cluster size 5, it can be compared with the balanced data setting with cluster size fixed at 5. Both methods have higher power in the random cluster size setting when the treatment group is assigned at the subunit-level. Under cluster-level group assignment, it appears that the RGL method has lower power with random cluster size, while the DS method has lower power with fixed cluster size, though the differences are not big. Both tests seem to be robust to the intracluster correlation structure.

A.2. Signed-rank tests for clustered data

For the signed-rank test, Datta and Satten (2008) did not have settings with completely random cluster size, and their non-exchangeable intracluster dependence is different from our

Group	Missing	Max	Corr	Group	$\delta = 0$		$\delta = 0.2$		$\delta = 0.5$	
level	rate	n_i	ρ	size	RGL	DS	RGL	DS	RGL	DS
cluster	0	2	0.1, 0.1	20	4.6	4.9	17.5	18.0	63.6	64.3
				50	4.9	5.1	37.7	38.1	96.2	96.3
			0.5, 0.5	20	4.7	5.1	13.7	14.0	52.8	54.0
				50	5.2	5.3	28.2	28.5	90.8	91.0
			−0.1, 0.9	20	4.5	4.7	15.7	16.2	58.9	59.6
				50	5.3	5.4	32.0	32.5	94.8	95.0
		5	0.1, 0.1	20	4.8	5.1	29.8	30.6	90.0	90.5
				50	4.8	5.0	64.4	64.8	99.9	99.9
			0.5, 0.5	20	4.8	5.1	17.0	17.6	61.9	62.8
				50	4.5	4.7	36.3	36.6	95.2	95.3
			−0.1, 0.9	20	5.1	5.3	17.3	18.1	79.6	80.4
				50	5.7	5.8	42.2	42.7	99.7	99.7
	0.5	10	0.1, 0.1	20	4.3	4.8	24.7	28.3	83.1	88.5
				50	5.2	4.9	57.9	58.5	99.9	99.9
			0.5, 0.5	20	4.7	4.8	14.1	16.7	51.7	62.1
				50	5.1	5.2	30.9	34.8	92.3	95.1
			−0.1, 0.9	20	4.7	4.8	14.7	18.2	66.4	76.8
				50	4.5	5.1	36.8	40.6	99.3	99.5
subunit	0	2	0.1, 0.1	20	5.0	4.9	19.3	18.4	69.6	68.9
				50	5.2	5.2	40.2	40.3	97.3	97.2
			0.5, 0.5	20	4.9	5.2	19.6	18.9	70.2	68.5
				50	5.4	5.5	39.3	38.8	97.2	96.9
			−0.1, 0.9	20	4.9	4.7	19.6	18.9	69.1	67.0
				50	4.5	4.5	40.9	40.6	97.4	97.1
		5	0.1, 0.1	20	5.1	4.7	39.8	38.1	98.0	97.2
				50	4.7	4.5	76.8	76.3	100.0	100.0
			0.5, 0.5	20	4.4	4.2	42.5	40.0	96.5	95.0
				50	4.3	4.5	78.1	76.4	100.0	100.0
			−0.1, 0.9	20	5.4	5.5	42.8	39.2	97.2	95.5
				50	4.9	4.9	76.2	74.9	100.0	100.0
	0.5	10	0.1, 0.1	20	5.3	5.1	38.4	35.4	96.1	94.4
				50	4.5	4.2	74.7	68.8	100.0	100.0
			0.5, 0.5	20	4.1	4.5	43.0	34.5	97.2	92.8
				50	4.7	5.0	78.3	71.0	100.0	100.0
			−0.1, 0.9	20	4.9	4.9	45.9	34.9	97.6	93.2
				50	5.2	5.4	80.7	70.3	100.0	100.0

Table 1: Empirical rejection percentage of the RGL and the DS methods for rank-sum tests at nominal significance level 0.05 when intracluster correlation is exchangable. The results are based on 4000 datasets. Each group contains same number of clusters.

Group	Missing	Max	Corr	Group	$\delta = 0$		$\delta = 0.2$		$\delta = 0.5$	
level	rate	n_i	ρ	size	RGL	DS	RGL	DS	RGL	DS
cluster	0	2	0.1, 0.1	20	5.0	5.3	17.2	17.9	64.3	65.2
				50	4.7	4.7	37.0	37.6	96.4	96.5
			0.5, 0.5	20	4.9	5.2	14.1	14.9	52.1	53.1
				50	4.6	4.7	29.1	29.6	88.6	88.9
			−0.1, 0.9	20	4.8	5.1	14.0	14.7	57.6	58.6
				50	5.0	5.0	32.4	32.9	94.4	94.5
		5	0.1, 0.1	20	4.9	5.2	35.1	36.0	95.2	95.4
				50	4.6	4.7	69.6	69.9	100.0	100.0
			0.5, 0.5	20	4.8	5.1	19.6	20.5	75.6	76.4
				50	5.3	5.4	46.6	47.0	99.0	99.0
			−0.1, 0.9	20	4.9	5.1	18.0	18.6	78.8	79.7
				50	4.5	4.7	42.4	42.9	99.5	99.6
	0.5	10	0.1, 0.1	20	4.3	4.5	31.1	33.9	91.1	93.7
				50	5.6	4.9	71.2	69.4	100.0	100.0
			0.5, 0.5	20	5.1	5.3	20.8	25.2	75.6	82.8
				50	5.1	5.4	50.4	52.8	99.5	99.6
			−0.1, 0.9	20	4.8	5.2	16.0	19.1	67.1	78.4
				50	5.7	5.2	41.2	44.6	99.3	99.6
subunit	0	2	0.1, 0.1	20	5.4	5.1	17.3	17.0	68.8	67.9
				50	5.3	5.3	39.3	39.1	97.4	97.2
			0.5, 0.5	20	5.1	5.2	19.5	18.7	68.5	67.2
				50	5.1	5.1	41.4	41.0	97.1	96.8
			−0.1, 0.9	20	5.5	5.5	19.3	18.8	69.4	67.6
				50	5.3	5.5	40.4	40.2	97.2	96.9
		5	0.1, 0.1	20	4.8	4.5	39.6	38.2	97.3	96.8
				50	4.6	4.8	78.0	77.4	100.0	100.0
			0.5, 0.5	20	5.3	5.5	41.5	40.0	96.9	96.2
				50	5.1	5.1	78.3	77.1	100.0	100.0
			−0.1, 0.9	20	4.9	5.1	42.4	39.8	97.4	96.2
				50	4.9	4.7	78.4	76.9	100.0	100.0
	0.5	10	0.1, 0.1	20	5.1	5.0	38.5	36.1	96.3	94.3
				50	5.3	5.3	75.6	70.2	100.0	100.0
			0.5, 0.5	20	4.8	5.0	39.2	35.5	96.5	93.6
				50	4.6	5.2	76.2	70.5	100.0	100.0
			−0.1, 0.9	20	4.8	4.8	43.6	35.9	97.4	93.4
				50	5.1	5.0	79.5	71.6	100.0	100.0

Table 2: Empirical rejection percentage of the RGL and the DS methods for rank-sum tests at nominal significance level 0.05 when intracluster correlation is AR1. The results are based on 4000 datasets. Each group contains same number of clusters.

Missing	Max	Corr	Group	$\delta = 0$		$\delta = 0.2$		$\delta = 0.5$	
rate	n_i	ρ	size	RGL	DS	RGL	DS	RGL	DS
0	2	0.1	20	5.0	5.2	20.8	20.9	78.3	78.7
			50	4.6	4.6	44.5	44.5	99.7	99.7
		0.5	20	4.6	5.0	16.2	16.6	67.5	68.4
			50	4.7	4.8	36.3	36.6	97.5	97.6
		0.9	20	4.2	4.4	13.9	14.4	57.4	58.5
			50	5.1	5.1	28.3	28.8	93.2	93.2
	10	0.1	20	5.0	5.1	48.5	48.9	99.8	99.8
			50	5.2	5.2	88.1	88.1	100.0	100.0
		0.5	20	4.5	4.8	19.2	19.9	81.0	81.5
			50	4.9	4.9	46.4	46.8	99.6	99.6
		0.9	20	4.9	5.2	14.0	14.6	58.0	59.0
			50	5.3	5.4	28.7	28.9	94.5	94.5
0.5	5	0.1	20	4.9	5.5	21.2	20.8	84.7	80.5
			50	4.5	4.9	49.8	46.2	100.0	99.7
		0.5	20	4.8	5.0	15.5	16.1	66.9	67.0
			50	4.5	4.4	35.0	34.7	97.9	97.7
		0.9	20	4.2	4.8	14.2	14.8	54.7	55.5
			50	4.8	5.0	27.7	27.9	92.6	92.8
	10	0.1	20	4.7	4.7	33.1	33.3	97.2	96.8
			50	5.3	5.0	72.5	69.4	100.0	100.0
		0.5	20	5.0	5.1	18.4	19.1	76.8	77.3
			50	5.3	5.3	40.7	41.0	99.2	99.2
		0.9	20	5.2	5.5	12.8	13.1	57.9	58.6
			50	4.8	4.8	29.3	29.3	94.1	94.2

Table 3: Empirical rejection percentage of the RGL and DS methods for signed-rank tests at nominal significance level 0.05 with exchangable intracluster correlation. The results are based on 4000 datasets. Each group contains same number of clusters.

AR1 setting. We considered settings similar to those for the rank-sum test. The paired differences were generated for 20 or 50 clusters. The intracluster correlation parameter was set to be 0.1, 0.5, and 0.9. The true difference δ was set to be 0, 0.2 and 0.5. For each setting, 4000 replicates were generated.

Selected results are summarized in Table 3–4. In all settings in this study, the empirical sizes of both methods are close to the nominal level, including the cases under of AR1 correlation where the exchangeability assumption for the RGL method is violated The empirical power of both tests increases as the cluster size increases, and as the intracluster dependence decreases. Completely random cluster size reduces the power in comparison to the cases where the cluster sizes are fixed at their means. In all the settings considered here, the two methods performed similarly.

Missing rate	Max n_i	Corr ρ	Group size	$\delta = 0$ RGL	$\delta = 0$ DS	$\delta = 0.2$ RGL	$\delta = 0.2$ DS	$\delta = 0.5$ RGL	$\delta = 0.5$ DS
0	2	0.1	20	5.0	5.1	19.7	20.2	79.7	80.1
			50	4.4	4.5	43.7	43.9	99.4	99.4
		0.5	20	4.8	5.3	16.1	16.4	66.8	67.5
			50	4.2	4.2	34.9	35.2	97.5	97.5
		0.9	20	5.4	5.6	12.5	13.2	56.8	58.1
			50	4.9	4.9	28.5	28.9	93.3	93.5
	10	0.1	20	4.5	4.5	63.9	64.1	100.0	100.0
			50	5.2	5.2	97.2	97.2	100.0	100.0
		0.5	20	4.3	4.3	38.0	38.4	98.0	98.1
			50	4.7	4.7	76.5	76.5	100.0	100.0
		0.9	20	5.5	5.7	16.1	16.7	68.6	69.5
			50	5.0	5.1	34.0	34.1	97.8	97.9
0.5	5	0.1	20	4.3	4.7	24.5	22.4	87.5	82.0
			50	5.0	5.1	53.8	47.7	99.9	99.5
		0.5	20	5.0	5.2	17.9	17.7	73.4	72.0
			50	5.1	5.3	40.2	38.9	99.3	98.9
		0.9	20	4.8	5.0	14.1	14.3	58.3	59.2
			50	5.6	5.8	28.6	28.9	94.1	94.0
	10	0.1	20	4.8	4.6	40.9	37.9	99.2	98.5
			50	5.1	4.7	82.0	78.5	100.0	100.0
		0.5	20	5.2	5.1	27.8	27.9	94.3	93.5
			50	4.9	4.9	63.5	62.5	100.0	100.0
		0.9	20	5.0	5.3	15.8	16.4	65.5	66.5
			50	5.1	5.1	34.7	34.9	97.4	97.3

Table 4: Empirical rejection percentage of the RGL and DS methods for signed-rank tests at nominal significance level 0.05 with AR1 intracluster correlation. The results are based on 4000 datasets. Each group contains same number of clusters.

Affiliation:

Yujing Jiang
Department of Statistics
Colorado State University
851 Oval Dr., Fort Collins, Colorado 80523, United States of America
Email: yujing.jiang@colostate.edu

Xin He, Mei-Ling Ting Lee
Department of Epidemiology and Biostatistics
University of Maryland
SPH Building #255, College Park, MD 20742, United States of America
Email: xinhe@umd.edu, mltlee@umd.edu

Bernard Rosner
Department of Biostatistics
Harvard University
181 Longwood Avenue, Boston, Massachusetts 02115, United States of America
Email: stbar@channing.harvard.edu

Jun Yan
Department of Statistics
Professor of Statistics
University of Connecticut
215 Glenbrook Rd. Unit 4120, Storrs, CT 06269, United States of America
Email: jun.yan@uconn.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
November 2020, Volume 96, Issue 6 *Submitted:* 2016-11-10
doi:10.18637/jss.v096.i06 *Accepted:* 2019-10-20