Simultaneous confidence intervals for ranks using the partitioning principle

Mohamad, D. al; Zwet, E. van; Solari, A.; Goeman, J.

Electronic Journal of Statistics
Vol. 15 (2021) 2608–2646
ISSN: 1935-7524
https://doi.org/10.1214/21-EJS1847

Simultaneous confidence intervals for ranks using the partitioning principle[*][†]

Diaa Al Mohamad and Erik van Zwet

Leiden University Medical Center, Einthovenweg 20. 2333 ZC Leiden, The Nethlerlands
e-mail: diaa.almohamad@gmail.com; e.w.van_zwet@lumc.nl

Aldo Solari

University of Milano-Bicocca, 1 Piazza dell'Ateneo Nuovo. 20126 Milano, Italy
e-mail: aldo.solari@unimib.it

Jelle Goeman

Leiden University Medical Center, Einthovenweg 20. 2333 ZC Leiden, The Nethlerlands
e-mail: j.j.goeman@lumc.nl

Abstract: We consider the problem of constructing simultaneous confidence intervals (CIs) for the ranks of n means based on their estimates together with the (known) standard errors of those estimates. We present a generic method based on the partitioning principle in which the parameter space is partitioned into disjoint subsets and then each one of them is tested at level α. The resulting CIs have then a simultaneous coverage of $1 - \alpha$. We show that any procedure which produces simultaneous CIs for ranks can be written as a partitioning procedure. We present a first example where we test the partitions using the likelihood ratio (LR) test. Then, in a second example we show that a recently proposed method for simultaneous CIs for ranks using Tukey's honest significant difference test has an equivalent procedure based on the partitioning principle. By embedding these two methods inside our generic partitioning procedure, we obtain improved variants. We illustrate the performance of these methods through simulations and real data analysis on hotel ratings. While the novel method that uses the LR test and its variant produce shorter CIs when the number of means is small, the Tukey-based method and its variant produce shorter CIs when the number of means is high.

MSC2020 subject classifications: Primary 62F07; secondary 62F03, 62F30.
Keywords and phrases: Rankings, simple order, likelihood ratio test, Tukey's HSD.

Received March 2020.

1. Introduction

In many applications, we seek to rank objects, entities or institutions on the basis of some numerical characteristic that is measured with uncertainty. One important example is assessing the quality of institutions such as medical centers

[*] This research is supported by VIDI grant number 639.072.412.

[†] This is an original paper.

and universities [19, 3]. Ranking institutions is usually carried out using performance indicators calculated based on samples. However, these indicators are only estimates of the true ones, and they are accompanied by standard errors, so that confidence intervals for their ranks are crucial.

We refer to a collection of CIs for ranks as having pointwise coverage of, say, 95% when the rank of any particular object is covered with 95% probability. We refer to simultaneous coverage if all ranks are covered with 95% probability. The latter is more useful, because it allows us to consider selected centers. For example, it ensures correct coverage for the object with the highest observed rank. Or the second-highest. Or for all objects in the top-5.

In the literature, the ranking problem is considered in several papers mostly focusing on pointwise CIs for the ranks. We mention the parametric bootstrap approach of [19] which is widely used, see [34, 17, 14] among others. Other methods are proposed based on testing pairwise differences between institutions [29, 30, 23, 10].

Methods for simultaneous CIs for ranks are proposed only by [47, 1, 25]. The method of [47] uses the parametric bootstrap to construct CIs for ranks and then Monte-Carlo simulations to estimate the simultaneous coverage. In [1], the authors show that the method of [47] is anticonservative and propose a new method based on Tukey's honest significant difference (HSD) test which ensures that the simultaneous confidence level is at least $1-\alpha$. The method of [25] creates simultaneous CIs for ranks starting from simultaneous CIs for the means which result in less powerful results than the method of [1].

Other papers from the literature considered the ranking problem but not with the objective to calculate CIs for the ranks, see [9, 21, 7, 27, 11, 31, 32, 33, 36, 43, 40]

We introduce in this paper a generic method for simultaneous CIs for ranks of a vector of means. We propose to partition the parameter space by considering all possible orderings of a set of means. Then, using a suitable (local) test, we test all the partitions at level α. The partitioning principle [42, 15, 18] ensures that by doing so, the familywise error rate is controlled at level α which enables us to build simultaneous CIs for ranks at level $1-\alpha$. The properties of the CIs for ranks depend on the local test we use, therefore different choices of the local test will lead to different methods.

Another important property of our generic procedure is that given some procedure that produces simultaneous CIs for ranks, we can construct a local test for the partitions so that the resulting partitioning procedure is equivalent to the original procedure. This shows that all valid procedures for simultaneous CIs are special cases of our approach, motivating the use of our generic procedure when looking for new methods for simultaneous CIs for ranks. Furthermore, in order to improve the partitioning procedure, it suffices to improve its local test which might be easier than improving the original procedure.

We present two examples of local tests. First, we use the likelihood ratio test as a local test and show how the partitioning procedure can be carried out. Although the complexity of the procedure is very high, we show some shortcuts which allow the method to be feasible with up to $30-40$ means with a regular computer. Second, we show that the method of [1] based on Tukey's HSD can be written as a partitioning procedure by giving an explicit local test for the partitions. Using our generic partitioning procedure, we present two variants of the partitioning procedure that uses the LR test and the Tukey-based method which uniformly improve their corresponding methods.

Simulation studies show that the method of [1] based on Tukey's HSD is more powerful than the one based on the LR test especially when the number of means is high while the converse happens when the number of means is small. The improved versions of these methods are shown to be computationally feasible only up to $n = 10$.

The paper is organized as follows. In Section 2, we give a formal definition of the ranking problem and set the objectives. In Section 3, we present the testing problem and show how to use it in order to produce simultaneous confidence intervals for the ranks. In Section 4, we show how, for any procedure for CIs for ranks, we can construct an equivalent one based on the partitioning principle. In Section 5, we present the likelihood ratio (LR) test and use it as a first example of a local test. In Section 6, we revisit the Tukey-based method of [1] and give an equivalent procedure using the partitioning principle. Finally, in Section 7, we compare the results of the LR test and the Tukey-based method on simulated and real data samples. In the Appendix, we provide proofs of main results and further details (algorithms, simulations and data). Software to perform the methods presented in this paper are available in the `ICRanks` package downloadable from CRAN.

2. Context and notation

Let $\mu_{1,T}, \cdots, \mu_{n,T}$ be real valued numbers which represent the unknown true means, that represent for example the true performance of n institutions we want to rank. Denote $\mu_T = (\mu_{1,T}, \cdots, \mu_{n,T})$. For ease of readability, we will use μ_i in place of $\mu_{i,T}$ when reference to the true mean is clear from context. The performance could be the mortality rates of hospitals or the customer rating as in our example in Section 8. We estimate $\mu_{i,T}$ using y_i. We assume that the estimator y_i is calculated based on many independent and identically distributed subjects (e.g. customers, patients), so that it becomes reasonable to assume that y_i is normally distributed with known standard error σ_i. Our starting point, therefore, is a sample $y = (y_1, \cdots, y_n)$ of n independent observations, each drawn from a Gaussian distribution $\mathcal{N}(\mu_{i,T}, \sigma_i^2)$.

Definition 1 (Ranks in the presence of ties)**.** Let $\mu = (\mu_1, \cdots, \mu_n) \in \mathbb{R}^n$. We define the lower-rank of μ_i by

$$l_i = 1 + \sum_{j \neq i} \mathbb{1}_{\mu_j < \mu_i}.$$

We also define the upper-rank of μ_i by

$$u_i = n - \sum_{j \neq i} \mathbb{1}_{\mu_j > \mu_i}.$$

We finally define the set-rank of μ_i as the set of natural numbers $r_i = \{l_i, l_i + 1, \cdots, u_i\}$ denoted here $[l_i, u_i]$. Furthermore, if for all $j \neq i$ we have $\mu_j \neq \mu_i$, then $l_i = u_i$ and the set-rank is the singleton $\{l_i\}$.

When there are ties between the means, then each one of the tied means possesses a set of ranks $r_i = [l_i, u_i]$. For example, suppose that we only have 3 means μ_1, μ_2 and μ_3 such that $\mu_1 = \mu_2 < \mu_3$. Then, the set-rank of μ_1 is $[1, 2]$ which includes both ranks 1 and 2, and the set-rank of μ_2 is also $[1, 2]$, whereas the rank of μ_3 is $[3, 3]$ which is simply rank 3. The rationale of the definition of the set-ranks is that in case of ties, the ranking is arbitrary, and a small perturbation of the true performance may produce any rank in the set of ranks. We call the ranks induced from the observed sample y the empirical ranks. These ranks might be different from the true ranks of the means, and since the sample is assumed to have a continuous distribution, the empirical (set-)ranks are all singletons.

We aim on the basis of the sample $\mathcal{Y}$ to construct (rectangular) simultaneous confidence intervals for the set-ranks of the means $\mu_{1,T}, \cdots, \mu_{n,T}$. In other words, for each i we search for a confidence interval $[L_i, U_i]$ such that:

$$\mathbb{P}\left([l_i, u_i] \subseteq [L_i, U_i], \forall i \in \{1, \cdots, n\}\right) \geq 1 - \alpha$$

for a prespecified confidence level $1 - \alpha$.

3. The testing problem and the partitioning principle

To obtain simultaneous confidence intervals for the set-ranks, we propose to test simultaneously all possible sets of set-ranks and then use the non-rejected ones to construct these CIs. Proposition 1 establishes this result. First, we need to define the hypotheses more formally. Consider the special case of three means A, B and C. The possible set-ranks for A, B, C are

$$\begin{array}{ccc}
 & \{\{1,2,3\},\{1,2,3\},\{1,2,3\}\}; & \\
\{\{1,2\},\{1,2\},\{3\}\}; & \{\{1\},\{2,3\},\{2,3\}\}; & \{\{1,2\},\{3\},\{1,2\}\}; \\
\{\{2,3\},\{1\},\{2,3\}\}; & \{\{2,3\},\{2,3\},\{1\}\}; & \{\{3\},\{1,2\},\{1,2\}\}; \\
\{\{1\},\{2\},\{3\}\}; & \{\{1\},\{3\},\{2\}\}; & \{\{2\},\{3\},\{1\}\}; \\
\{\{2\},\{1\},\{3\}\}; & \{\{3\},\{2\},\{1\}\}; & \{\{3\},\{1\},\{2\}\}.
\end{array} \tag{1}$$

Each set of set-ranks correspond to an ordering of the means. For example, the set-ranks $\{\{1\},\{2\},\{3\}\}$ correspond to $A < B < C$. The set-ranks $\{\{1,2,3\},\{1,2,3\}\ ,\{1,2,3\}\}$ correspond to $A = B = C$. Figure (1) shows all the corresponding cases. Note that these cases partition the space $\mathbb{R}^3$. In order to calculate simultaneous CIs for the ranks of the three means, we propose to test if the vector of means (A, B, C) comply with each of the cases in Figure (1) which is equivalent to testing all sets of set-ranks (1). Note that in this example, if we assume that there no ties, only the third level of hypotheses remain.

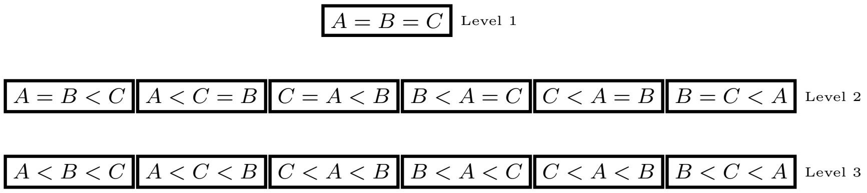

Figure 1: Example of a partitioning scheme with three means A, B and C.

More generally, let $=\!\!\!<$ denote either $=$ or $<$. In order to calculate simultaneous CIs for the ranks of $\mu_{1,T}, \cdots, \mu_{n,T}$, we test the elementary hypotheses

$$H : \mu_{\pi(1)} =\!\!\!< \cdots =\!\!\!< \mu_{\pi(n)}, \tag{2}$$

for all possible configurations of equalities and inequalities among the means and for all permutations $\pi \in \mathcal{S}_n$ (the symmetric group of all permutations with n numbers). Some of these hypotheses are the same because permuting tied means has no effect on the ordering. Therefore, we do not need these redundant copies.

Clearly, the elementary hypotheses (2) become disjoint after we omit the redundant copies. In other words, if $(\mu_1, \cdots, \mu_n) \in H_1 \cap H_2$, then $H_1 = H_2$. Moreover, the union of the parameter spaces implied by these elementary hypotheses is equal to $\mathbb{R}^n$, and hence they form a partitioning of $\mathbb{R}^n$. In order to test all these hypotheses simultaneously at level α, it suffices to test each one of them at level α due to the so-called the partitioning principle [42, 15, 18]. Two examples of tests will be introduced later on, in Sections 5 and 6.

The following result states how the confidence intervals are constructed. This is the general approach for simultaneous CIs for ranks based on testing the elementary hypotheses (2) and using the partitioning principle.

Proposition 1. *Assume we have a statistical test for the elementary hypotheses (2) with significance level equal to α. The union of unrejected elementary hypotheses at level α constitutes simultaneous confidence intervals for the set-ranks of the means $\mu_{1,T}, \cdots, \mu_{n,T}$ at level $1 - \alpha$.*

Although this is useful as a general method, it is not always practical because the number of elementary hypotheses increases rapidly with n. Proposition 2 states the exact number of the elementary hypotheses (2) defining the partitioning of $\mathbb{R}^n$.

Proposition 2. *The total number of elementary hypotheses (2) in the partitioning scheme is*

$$\sum_{l=1}^{n} l! S(n, l)$$

where $S(n, l)$ are the Stirling numbers of the second type.

When there are no ties, the number of hypotheses to test drops to $n!$, therefore, in general the number of hypothesis to test is higher than $n!$, and we can even be more precise and calculate an upper and a lower bound using the result in [37]. In the statistical tool R, function `Stirling2` from package `multicool` calculates these numbers. In any case, the total number of hypotheses is large, and it is very important to find a way to reduce this complexity by finding relations among tested partitions. In the literature, these relations are called shortcuts [12]. We provide two examples where some shortcuts are exploited in order to reduce the number of tested hypotheses.

4. Any procedure generating valid simultaneous CIs for ranks is equivalent to a partitioning procedure

Although the literature on simultaneous CIs for ranks includes the two papers [1, 47], it is always possible to start from pointwise CIs methods such as [23, 10, 29] and correct for multiplicity, for example using Bonferroni's method, so that the resulting CIs become simultaneous, but they tend to be very conservative. In this section, we show that any procedure which produces valid simultaneous CIs for ranks can be written as a partitioning procedure with elementary hypotheses (2) and a suitable statistical test. We note that the class of partitioning procedures is larger than the class on rectangular confidence intervals for ranks, since partitioning may sometimes lead to non-rectangular confidence intervals.

Assume that we have a procedure that generates confidence intervals for ranks with joint confidence level $1-\alpha$. Let $[\tilde{L}_1, \tilde{U}_1], \cdots, [\tilde{L}_n, \tilde{U}_n]$ be the corresponding confidence intervals. This means

$$\mathbb{P}_{\mu_T}\left(\forall i, r_i \subseteq [\tilde{L}_i, \tilde{U}_i]\right) \geq 1-\alpha. \tag{3}$$

Let H be some elementary hypothesis from (2). This partition includes all vectors of means having one specific set of set-ranks, say $r_1(H), \cdots, r_n(H)$. Hence, for any $(\mu_1, \cdots, \mu_n)$ under H, we have $r_i(\mu_i) = r_i(H)$ according to Definition 1. We define a local test for H, say φ by

$$\varphi(H) = \begin{cases} 0 & \text{if } r_i(H) \subseteq [\tilde{L}_i, \tilde{U}_i], \forall i; \\ 1 & \text{otherwise .} \end{cases} \tag{4}$$

Proposition 3. *φ is a valid test for H at level α.*

The test does not reject an elementary hypothesis H if the confidence intervals $[\tilde{L}_1, \tilde{U}_1], \cdots, [\tilde{L}_n, \tilde{U}_n]$ cover the set-ranks of any vector of means under H. Since we have a valid local test for the partitions, we can build a partitioning scheme leading to a set of simultaneous CIs with joint confidence level of $1-\alpha$ for the ranks, say $[L_1, U_1], \cdots, [L_n, U_n]$ due to Proposition 1. We show in the following proposition that they are the same as the ones produced by the original procedure (3).

Proposition 4. *The confidence intervals produced by the method (3) are the same as the ones produced by the partitioning procedure using the elementary hypotheses (2) and the local test φ, that is for all i, $L_i = \tilde{L}_i$ and $U_i = \tilde{U}_i$.*

This fundamental result indicates that our partitioning procedure is complete [28, Section 1.8] for constructing simultaneous confidence intervals for ranks: every valid method is a special case of this method. Partitioning should, therefore, be considered as a design principle when thinking about new methods. When designing a method, it suffices to look for a new local test for the elementary hypotheses. In order to improve a procedure uniformly, it suffices to improve the local test. Note that the result holds whether there are ties or not. Two examples are given in the following sections. Note that in these two examples, the resulting confidence intervals are invariant against a translation of the means by a common constant.

5. A first example: The likelihood ratio test

In the literature on ordered hypotheses such as our elementary hypotheses (2), there is not yet a general result about an optimal test. However, as stated by [8] "the method has a strong intuitive appeal and leads to a meaningful test", referring to the likelihood ratio test. Let H be some elementary hypothesis. The likelihood ratio statistic (LR) related to testing H against all alternatives is given by

$$LR = \min_{\mu_1,\cdots,\mu_n \in H} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2}. \tag{5}$$

Note that the term related to the alternative hypothesis is zero since the minimum is attained at $\mu_i = y_i$. In some cases, the minimum in (5) can be calculated directly. For example, when H is the hypothesis under which the means are equal, the minimum in (5) is attained at the average of the observations. In the sequel, we will use the following notation. Let H be the hypothesis $\mu_1 = \mu_2 < \mu_3 < \mu_4 = \mu_5$. We convene the following writing $H : B_1 < B_2 < B_3$ with $B_1 = \{\mu_1, \mu_2\}, B_2 = \{\mu_3\}$ and $B_3 = \{\mu_4, \mu_5\}$. More generally, assume that H can be written as a union of blocks of means $B_1, \cdots, B_l$ where in each block the means are equal under H and such that if μ_i is in block B_t and μ_j is in block B_s such that $t < s$, then $\mu_i < \mu_j$ under H. Using our new notation, we write $H : B_1 < \cdots < B_l$. Let

$$\hat{\mu}_{B_j} = \frac{1}{\sum_{k:\mu_k \in B_j} \frac{1}{\sigma_k^2}} \sum_{s:\mu_s \in B_j} \frac{y_s}{\sigma_s^2}. \tag{6}$$

It can be shown [8] that if $\hat{\mu}_{B_1} \leq \cdots \leq \hat{\mu}_{B_l}$, the minimum in (5) is attained on H and the LR is given by

$$LR = \min_{\mu_1,\cdots,\mu_n \in H} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2} = \sum_{j=1}^{l} \sum_{i:\mu_i \in B_j} \frac{(y_i - \hat{\mu}_{B_j})^2}{\sigma_i^2}, \tag{7}$$

Note that if the block B_j contains only one mean, say μ_k, the argument of the minimum $\hat{\mu}_{B_j}$ is equal to the single observation y_k. Hence, all subsets B_j with a single mean do not appear in the calculation of the LR statistic.

In general, the basic R function `isoreg` or function `activeSet` from package `isotone` [13] can be used to calculate the minimum in (5) using the Pool Adjacent Violators Algorithm known as the PAVA ([5, 45, 8, 6, 13]).

In order to perform a LR test of H against all alternatives, two approaches from the literature are available; an adaptive [46, 24, 2] and a non adaptive one [38]. The adaptive approach compares the LR with a quantile of a chi-square with data-dependent degrees of freedom whereas the non adaptive one compares the LR with the quantile of a mixture of chi-squares. Paragraph B in the Appendix provides further details. We will see later on that the adaptive approach is always more suitable in this context.

5.1. Some shortcuts and practical issues

The complexity of the partitioning scheme is very high (Lemma 2), so that it is important to find a way to simplify calculations as much as possible and preferably reduce the complexity of the algorithm we use.

We start by the most simple and immediate shortcut.

Lemma 1. *If the hypothesis $\mu_1 = \cdots = \mu_n$ is not rejected, then there is no need to test any other hypothesis and the confidence intervals for the ranks are the trivial ones, that is $[1, n]$.*

Without loss of generality, assume that $y_1 < ... < y_n$. The hypotheses in the partitioning scheme are three types according to the relative ordering of the y_i's with respect to the ordering of the μ_i's in the hypothesis. Let $H : B_1 < \cdots < B_l$ be some partition where we group the means which are equal under H in the blocks $B_1, \cdots, B_l$. Let $\hat{\mu}_{B_j}$ be given by (6). We say that

1. H is a correctly ordered hypothesis whenever $\hat{\mu}_{B_1} < \cdots < \hat{\mu}_{B_l}$ and for any $i < j$ then $\mu_i \leq \mu_j$;
2. H is a partially correctly ordered hypothesis whenever $\hat{\mu}_{B_1} < \cdots < \hat{\mu}_{B_l}$ and there exist $i < j$ such that $\mu_i > \mu_j$. This means that the sample means of the blocks respect the empirical ordering (of the y_i's) whereas the means do not;
3. H is an incorrectly ordered hypothesis if there exist $i < j$ such that $\hat{\mu}_{B_i} > \hat{\mu}_{B_j}$, which means that neither the means nor the sample means of the blocks respect the empirical ordering of the y_i's.

Note that in the first two cases, the LR is given by (7) which means that the solution of the PAVA does not pool any adjacent blocks. In the third case, the LR results from the PAVA by pooling all blocks of means $B_i, \cdots, B_j$ violating the empirical ordering into one block, and then using formula (7). The resulting pooled blocks are partially correctly ordered hypotheses. To illustrate the differences among the three types of hypotheses, assume for example $y_1 = 0, y_2 = 2, y_3 = 3$.

First we have $y_1 < y_2 < y_3$. The hypothesis $\mu_1 = \mu_2 < \mu_3$ is a correctly ordered hypothesis since $(y_1 + y_2)/2 < y_3$. The hypothesis $\mu_1 = \mu_3 < \mu_2$ is a partially correctly ordered hypothesis because under the hypothesis $\mu_3 < \mu_2$ and $(y_1 + y_3)/2 < y_2$. Finally, the hypothesis $\mu_2 = \mu_3 < \mu_1$ is an incorrectly ordered hypothesis because $(y_2 + y_3)/2 > y_1$. Note that we only require the knowledge of the observed values of $y_1, \cdots, y_n$ and the relative positions of $\mu_1, \cdots, \mu_n$ under the current hypothesis and not their actual values.

The following propositions state that only the correctly and partially correctly ordered hypotheses are required. When we have a common standard deviation, we show that only a subset of the partially correctly ordered hypotheses is required.

Proposition 5. *Assume that all the standard deviations are the same. If we use the LR to test the elementary hypotheses (2), then in order to obtain simultaneous CIs for ranks at joint level $1 - \alpha$, it suffices to test at level α the following list of hypotheses:*

1. *the correctly ordered hypotheses;*
2. *if the correctly order hypothesis $H : \mu_1 \mathrel{=\!\!<} \cdots \mathrel{=\!\!<} \mu_n$ is not rejected, test all partially correctly ordered hypotheses of the form $\mu_{\pi(1)} \mathrel{=\!\!<} \cdots \mathrel{=\!\!<} \mu_{\pi(n)}$ for all permutations π from the list*

$$\left\{ \begin{array}{ccc} (i, i+1), & & \forall i = 1, \cdots, n-1 \\ (i, i+1, i+2), & (i+2, i+1, i), & \forall i = 1, \cdots, n-2 \\ (i, i+1, i+2, i+3), & (i+3, i+2, i+1, i), & \forall i = 1, \cdots, n-3 \\ \vdots & \vdots & \vdots \\ (1, 2, 3, \cdots, n), & (n, n-1, \cdots, 2, 1) & \end{array} \right. \tag{8}$$

Furthermore, if we apply the list (8) column after column, then for each column it suffices to test until we encounter the first rejected partially correctly ordered hypothesis.

Proposition 5 shows that in order to construct the simultaneous CIs for the ranks, we need to test first all correctly ordered hypotheses. For $n = 3$, these are $\mu_1 = \mu_2 = \mu_3$, $\mu_1 < \mu_2 = \mu_3$, $\mu_1 = \mu_2 < \mu_3$ and $\mu_1 < \mu_2 < \mu_3$. Recall that we assumed $y_1 < y_2 < y_3$. According to Proposition 5, if hypothesis $\mu_1 < \mu_2 = \mu_3$ is not rejected then we need to apply the first column of permutations which are $\pi = (1, 2)$ that results in hypothesis $\mu_2 < \mu_1 = \mu_3$ and $\pi = (1, 2, 3)$ that results in hypothesis $\mu_3 < \mu_2 = \mu_1$. Then we use the second column of permutations which are $\pi = (2, 3)$ that results in $\mu_1 < \mu_3 = \mu_2$ and $(3, 2, 1)$ that results in hypothesis $\mu_2 < \mu_3 = \mu_1$. Otherwise, if the hypothesis $\mu_1 < \mu_2 = \mu_3$ is rejected, we move to the next correctly order hypothesis.

When the standard deviations are not equal, the proof of Proposition 5 is no longer valid. This is because we use (and prove) the fact that the LR does not decrease as the number of permuted means increases. For example, let $H : B_1 < \cdots < B_k$ be some correctly ordered hypotheses. If we permute μ_i with μ_j, then the LR increases. When the standard deviations are not the same, then this result no longer holds in general, especially when we permute two means, one with high standard deviation and one with small standard deviation.

Proposition 6. *Assume that there exist $i \neq j$ such that $\sigma_i \neq \sigma_j$. If we use the LR to test the elementary hypotheses (2), then it suffices to test the correctly ordered and partially correctly ordered hypotheses.*

While this result shows that there is no need to test the incorrectly ordered hypotheses, we do not know how to characterize the set of partially correctly ordered hypotheses in general efficiently. In the Appendix, we provide an algorithm to test all the elementary hypotheses. In that algorithm, we test first the correctly ordered hypotheses, then, we permute the indexes using some $\pi \in \mathcal{S}_n$ and repeat the same procedure while taking into account that some hypotheses become incorrectly ordered because of the permutation and they can be discarded. When $n > 10$, this becomes computationally infeasible. In that case, we may still use the list of permutations (8) and then randomly select another some 10^5 permutations from $\mathcal{S}_n$ and apply them all. Of course, this is an approximation and we might just hope that the resulting CIs get a joint confidence level of $1 - \alpha$, but we have no guarantee that they will. In the Appendix, we provide a few simulations for the case of different standard deviations when $n = 10, 15$ which show that the approximate CIs are still conservative for a variety of vectors of means and vectors of standard deviations.

As we mentioned here above in this section, it is possible to test the LR statistic using either an adaptive test [24, 46, 2] or using a non adaptive test [38]. According to Propositions 5 and 6, we only test the correctly and partially correctly ordered hypotheses for which the LR is given by (7). Therefore, the adaptive quantile is the quantile of a $\chi^2(\ell)$ where ℓ is the number of equalities in H whereas the non adaptive quantile is the quantile of a mixture of the chi-squares $\chi^2(\ell), \cdots, \chi^2(n-1)$. This proves the following corollary.

Corollary 1. *If we compare the LR to the adaptive quantile, the resulting simultaneous CIs for ranks are never longer than the ones obtained if we compare the LR to the non adaptive quantile, that is the quantile of a mixture of chi-squares.*

5.2. An improved variant

We give in this paragraph a way to improve the partitioning procedure when the local test is the LR test using our generic procedure from Section 4. We use the partitioning procedure in order to produce simultaneous CIs for the ranks of $\mu_{1,T}, \cdots, \mu_{n,T}$ and then define function φ_{LR} through (4).

Consider the hypothesis $H : \mu_1 = \cdots = \mu_n$. Using the LR test, it is tested at an exact level α by comparing $LR(H)$ with the quantile q_{n-1} of $\chi^2(n-1)$. However, using the test φ_{LR}, hypothesis H is not rejected not only when $LR(H) \leq q_{n-1}$, but also whenever a set of partitions implying the trivial CIs $[1, n]$ to all the means is not rejected either. For example, the CI for the rank of $\mu_{1,T}$ can be $[1, n]$ if the elementary hypotheses $\mu_1 < \mu_2 = \cdots = \mu_n$ and $\mu_2 < \mu_1 = \cdots = \mu_n$ are both not rejected. This means that the test $\varphi_{\text{LR}}(H)$

does not exhaust the α-level and thus we can estimate the gap between the actual level of the test and α and rescale the test significance level in order to gain power. In Section 7, we show that when $\mu_{1,T} = \cdots = \mu_{n,T}$ with $n = 10$ and $\alpha = 0.1$, then 95.6% of the simulations result in trivial CIs $[1, n]$ for all the means simultaneously. This means that we actually reject $\mu_1 = \cdots = \mu_n$ using function φ_{LR} at the rate of 0.044 instead of the 0.1 used in the simulations.

In order to rescale the local tests φ_{LR}, we need to find a least favorable vector μ_0 with respect to which we can rescale. In other words, μ_0 has to verify for any μ under H,

$$\mathbb{P}_{\mu}(\varphi_{\text{LR}}(H) = 1) \leq \mathbb{P}_{\mu_0}(\varphi_{\text{LR}}(H) = 1).$$

It appears that $\mu_0 = (0, \cdots, 0)$, and the following lemma states this result when the non adaptive critical value is used. The case when we use the adaptive critical value remains unknown.

Lemma 2. *Let H be some elementary hypothesis from (2). Assume that we compare the LR with the non adaptive critical value [38], then for any $\mu \in H$*

$$\mathbb{P}_{\mu}(\varphi_{LR}(H) = 1) \leq \mathbb{P}_{0}(\varphi_{LR}(H) = 1).$$

Let $\varphi_{\text{LR}(\alpha)}$ denote φ_{LR} when the original partitioning procedure is calculated at level α. Rescaling the partitioning procedure defined using $\varphi_{\text{LR}(\alpha)}$ as a local test is done by looking for a zero of the function $\tilde{\alpha} \mapsto \mathbb{P}_0(\varphi_{\text{LR}(\tilde{\alpha})}(H) = 1) - \alpha$ for all the elementary hypotheses H from (2). For any $\tilde{\alpha}$, the probability $\mathbb{P}_0(\varphi_{\text{LR}(\tilde{\alpha})}(H) = 1)$ can be estimated by simulations.

This improved variant is uniformly more powerful than the original partitioning procedure since the rescaled level $\tilde{\alpha}$ is in $[\alpha, 1]$. However, in practice, the variant is computationally feasible only for small number of means. As implemented in our R package `ICRanks`, when the standard deviations are the same, the improvement is computationally feasible up to $n = 10$. When the standard deviations are not the same, then we have to calculate the non adaptive quantile for each one of the elementary hypotheses (2) by simulations or using iterative methods [20] which makes the improvement computationally feasible only for $n \leq 5$.

6. A second example: Tukey's procedure for ranks

Tukey's pairwise comparison procedure [44, 22] well-known as the Honest Significant Difference test (HSD) is an easy way to compare a set of n means based on a Gaussian sample especially in ANOVA models. The interesting point about the procedure is that it provides simultaneous confidence intervals for the differences and controls the FWER at level α. Tukey's HSD is employed by [1] to produce simultaneous CIs for the ranks. The objective here is to review this method and get more insights about it in terms of the partitioning principle.

6.1. The method

Suppose that $y_1, \cdots, y_n$ are generated independently from the Gaussian distributions $\mathcal{N}(\mu_{i,T}, \sigma_i^2)$. In order to produce simultaneous confidence intervals for the ranks of the means $\mu_{1,T}, \cdots, \mu_{n,T}$, we test all hypotheses of the form $H_{i,j} : \mu_i = \mu_j$ using the following rejection region

$$\left\{ \frac{|y_i - y_j|}{\sqrt{\sigma_i^2 + \sigma_j^2}} > q_{1-\alpha} \right\}$$

where $q_{1-\alpha}$ is the quantile of order $1 - \alpha$ of the distribution of the Studentized range

$$\max_{i,j=1,\cdots,n} \frac{|Y_i - Y_j|}{\sqrt{\sigma_i^2 + \sigma_j^2}}, \tag{9}$$

and Y_i and Y_j are two independent Gaussian random variables with mean 0 and standard deviations σ_i and σ_j respectively. The confidence interval for the rank of mean $\mu_{i,T}$, say $[L_i, U_i]$ is calculated by counting how many hypotheses $H_{i,j}$ are rejected and such that $y_j < y_i$ (which yields $L_i - 1$). Then we calculate how many hypotheses $H_{i,j}$ are not rejected and such that $y_j > y_i$ (which yields $n - U_i$).

6.2. A new look at Tukey's pairwise comparison using the partitioning principle

We define a statistical (local) test over the elementary hypotheses (2) which yields the same confidence intervals for the ranks as the method based on Tukey's HSD. Assume that $\sigma_i = \sigma$ for all i. Let $H : B_1 < \cdots < B_l$, where as before B_i is a block of means which are equal under H. For each block, we calculate the maximum and minimum observed values. If the observed maximum of a block B_i (calculated using the y_i's) is larger than the observed minimum of block B_{i+1}, then the two blocks are combined (pooled) into one, say $\tilde{B}_i$. Denote $\tilde{l}$ the number of remaining blocks after pooling. We test the hypotheses H using the following rejection region

$$\left\{ \max_{k=1,\ldots,\tilde{l}} \max_{\mu_j, \mu_s \in \tilde{B}_k} \frac{|y_j - y_s|}{\sqrt{2}\sigma} > q_{1-\alpha} \right\}, \tag{10}$$

where $q_{1-\alpha}$ is the quantile of order $1 - \alpha$ of the Studentized range (9) as in Tukey's HSD procedure. Note that we use the same critical value for all the elementary hypotheses.

Proposition 7. *If we use the Tukey-based method for ranks to construct a new partitioning procedure using the local test φ, then $\{\varphi = 1\}$ is equivalent to the rejection region (10).*

When the standard deviations are not the same, we can show that the partitioning procedure produces slightly shorter CIs for the ranks than the Tukey-based method. Although this would seem as if we obtained an improved procedure through the partitioning procedure, we do not have a proof that the local test (10) is an α-level test and hence the resulting CIs are not guaranteed to have a joint level $1-\alpha$.

6.3. An improved variant based on the partitioning principle

Similarly to the partitioning procedure that uses the LR as a local test, we can define an equivalent partitioning procedure to the Tukey-based method of [1] using the test φ (4). We show using Proposition 3.2 from [1], that $\mu = 0$ is the least favorable case. We consider a new partitioning procedure in which we test the elementary hypotheses (2) using a local test $\varphi = \varphi_{\mathrm{TKY}}$ of the form of (4) that uses the simultaneous CIs for the ranks obtained through the Tukey-based method of [1].

Lemma 3. *Let H be some elementary hypothesis from (2). For any $\mu \in H$*

$$\mathbb{P}_\mu\left(\varphi_{TKY}(H) = 1\right) \leq \mathbb{P}_0(\varphi_{TKY}(H) = 1).$$

Rescaling the partitioning procedure defined using function φ_{TKY} as a local test is done by looking for a zero of the function $\tilde{\alpha} \mapsto \mathbb{P}_0(\varphi_{\mathrm{TKY}(\tilde{\alpha})}(H) = 1) - \alpha$ for all the elementary hypotheses H from (2). The probability $\mathbb{P}_0(\varphi_{\mathrm{TKY}(\tilde{\alpha})}(H) = 1)$ can be estimated through simulations.

Similarly to the case of the partitioning procedure that uses the LR test, in practice, this improvement is computationally feasible on ordinary computers for $n \leq 10$. In contrast to the LR case, when the standard deviations are not the same, the procedure does not imply any further complications and is computationally feasible up to $n = 10$.

7. Simulation study: A comparison of simultaneous coverage and efficiency

The goal of this section is to compare the performance of the novel LR-based method with the Tukey-based method of [1] which is the only method available in the literature which provides valid simultaneous CIs for ranks. Note that [1] show that the method of [47] does not control the joint confidence level of the CIs, therefore, it is unfair to include it in the comparison. We also illustrate the performance of the method of [25] that uses the Sidak correction.

The simulation setup is the following. We estimate the simultaneous coverage of both methods, the LR-based and the Tukey-based method of [1] for vectors of n means with $n \in \{5, 10, 20\}$. We generate 1000 means $\mu_{1,T}, \cdots, \mu_{n,T}$ independently from the Gaussian distribution $\mathcal{N}(0, \tau^2)$ for $\tau \in \{0, 1, 3, 5\}$. For each vector of means, we generate independently a Gaussian sample $y_1, \cdots, y_n$ such that $y_i \sim \mathcal{N}(\mu_{i,T}, 1)$. For each τ, the coverage is estimated as the proportion of vectors of means which are being covered simultaneously by the CIs calculated based on the corresponding samples $y_1, \cdots, y_n$. The results are presented in Table 1. We also calculate the average length of the confidence intervals

$$1 - \hat{R}_n(\alpha) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (U_i - L_i),$$

where $\hat{R}_n(\alpha)$ is the rankability measure defined in [1]. The quantity $1 - \hat{R}_n(\alpha)$ is a measure of efficiency of a method producing CIs for ranks. A better method has shorter CIs and therefore a smaller $1 - \hat{R}_n(\alpha)$. We provide in Appendix D simulations when the standard deviations are not the same, cases when more ties are present among the true means and when the normality assumption is violated.

The results of Table 1 show that on average the Tukey-based method produces shorter confidence intervals than the LR-based one especially as the number of means increases to 20. When the number of means is smaller than 10, our LR-based method produces shorter CIs. Both variants produce shorter CIs than their corresponding methods.

It is not surprising that when $\tau = 0$ (all the means are tied and their true set-ranks are all $[1, n]$), the Tukey-based method delivers CIs for ranks with joint level equal to $1 - \alpha$ because this method is exact when $\mu_{1,T} = \cdots = \mu_{n,T}$ [1, Proposition 3.2]. On the other hand, our novel method based on the LR test does not seem to share this property empirically except for $n = 5$. The method of [25] is the least performing method. We recall Proposition 3.2 from [1] that states that when the standard deviations are the same, the Tukey-based method produces shorter simultaneous confidence intervals for the ranks than the method of [25].

The rescaled version of the Tukey-based method does not improve as much as the rescaled version of the partitioning procedure that uses the likelihood ratio test. When the standard deviations are equal, we show in Lemma 3 in the Appendix that in order to perform the partitioning procedure that uses the local test (10), it suffices to test the correctly ordered hypotheses. We can see the implication of such result on the example of testing the hypothesis $\mu_1 = \cdots = \mu_n$. Indeed, we obtain trivial CIs for the ranks only when that hypothesis is not rejected, because if $\mu_1 = \cdots = \mu_n$ is rejected, then there is no correctly ordered hypothesis that has μ_1 in the nth position except for $\mu_1 = \cdots = \mu_n$. This means that it is not possible to improve the local test for this hypothesis. In the case of the partitioning procedure that uses the LR test, it is possible to improve the level at which we test the hypothesis $\mu_1 = \cdots = \mu_n$. When the standard deviations are not the same, Lemma 3 from the Appendix no longer holds, and the improved procedure may be more efficient.

8. Data analysis

Ratings of hotels is one of the tools that booking websites use to show the quality of these hotels and guide new customers choose a suitable one. Booking.com is

Table 1

Coverage probability and efficiency for different values of τ. Nominal simultaneous level is 90%.

		$n = 5$		$n = 10$		$n = 20$	
		Coverage	Average length of CIs	Coverage	Average length of CIs	Coverage	Average length of CIs
$\tau = 0$	Tukey	0.912	0.790	0.896	0.896	0.911	0.949
	Rescaled Tukey	0.912	0.789	0.895	0.896	-	-
	LR	0.912	0.789	0.942	0.899	0.999	0.950
	Rescaled LR	0.900	0.787	0.928	0.897	-	-
	Klein	0.989	0.799	0.990	0.899	0.991	0.950
$\tau = 1$	Tukey	0.989	0.735	0.997	0.866	0.999	0.931
	Rescaled Tukey	0.988	0.729	0.993	0.861	-	-
	LR	0.984	0.721	0.998	0.874	1.000	0.947
	Rescaled LR	0.974	0.709	0.991	0.864	-	-
	Klein	1.000	0.784	1.000	0.891	1.000	0.945
$\tau = 3$	Tukey	0.993	0.446	1.000	0.585	1.000	0.669
	Rescaled Tukey	0.992	0.425	0.997	0.557	-	-
	LR	0.980	0.393	0.995	0.572	1.000	0.734
	Rescaled LR	0.954	0.348	0.970	0.512	-	-
	Klein	1.000	0.559	1.000	0.670	1.000	0.748
$\tau = 5$	Tukey	0.993	0.298	0.997	0.391	1.000	0.461
	Rescaled Tukey	0.989	0.276	0.991	0.360	-	-
	LR	0.980	0.247	0.993	0.358	1.000	0.499
	Rescaled LR	0.950	0.209	0.954	0.293	-	-
	Klein	1.000	0.375	1.000	0.471	1.000	0.531

one of the world leading websites for booking hotels. A hotel is rated by some of its customers for different criteria such as cleanness, breakfast, etc. An overall rating between 1 and 5 stars is also attributed to the hotel by the customer. We used the data publicly available on the website www.booking.com for a room reservation in the city of Leiden (The Netherlands) to rent a room for one night on the 2^{nd} of May 2019. The query was made on the 15^{th} of April 2019. We restricted our search for hotels with free Wifi, free cancellation and within 1 Km from the city center. We obtained a list of 9 hotels (see raw data in the Appendix). For each hotel, we have the number of customers who rated the hotels for 1, 2, 3, 4 or 5 stars. We compute the average rating for each hotel and its standard error in the following way. Let X be a random variable taking values in the set $\{1,2,3,4,5\}$ which represents the rating of a customer. We calculate

$$y_i = \sum_{j=1}^{5} j \frac{n_{i,j}}{n_i};$$

$$\sigma_i^2 = \sum_{j=1}^{5} j^2 \frac{n_{i,j}}{n_i^2} - \frac{1}{n_i} y_i^2$$

where n_i is the number of customers reviews for the i^{th} hotel and $n_{i,j}$ is the number of customers reviews of j stars in the i^{th} hotel. The result is in table 2.

We apply both the Tukey-based method of [1] and our new LR-based method on this data and calculate simultaneous CIs for the ranks of these hotels at joint

TABLE 2
Average and standard error of ratings of 9 hotels in the city of Leiden (The Netherlands). Simultaneous CIs of joint level 90% for their ranks are calculated using our novel LR-based method and the Tukey-based method.

Hotel name	Average rating	Standard error	CI - LR	CI - rescaled LR	CI - (rescaled) Tukey	Klein
Hotel Mayflower	3.825	0.0258	[8,9]	[9,9]	[8,9]	[8,9]
Best Western City Hotel Leiden	3.888	0.0169	[8,9]	[8,8]	[8,9]	[8,9]
City Resort Hotel Leiden	3.996	0.0197	[7,7]	[7,7]	[7,7]	[7,7]
City Hotel Rembrandt	4.110	0.0191	[5,6]	[5,6]	[5,6]	[5,6]
Ibis Leiden Centraal	4.149	0.0131	[5,6]	[5,6]	[5,6]	[5,6]
Tulip Inn Leiden	4.254	0.0183	[3,4]	[3,4]	[3,4]	[3,4]
Golden Tulip Leiden	4.277	0.0182	[3,4]	[3,4]	[3,4]	[3,4]
Boutique Hotel d'Oude Mors	4.717	0.0154	[2,2]	[2,2]	[2,2]	[2,2]
Boutique Hotel Steenhof Suites	4.839	0.0137	[1,1]	[1,1]	[1,1]	[1,1]

level 90%. We also apply the rescaled versions of these methods presented in paragraphs 5.2 and 6.3. Since the standard errors of the means are not the same, Proposition 5 does not hold so that we have to test all elementary hypotheses (2). Furthermore, the rescaled version of the partitioning procedure that uses the LR as a local test is not computationally feasible, therefore, we use the maximum standard error of all the hotels ratings as the common standard error for all the hotels ratings. The resulting simultaneous CIs are upper-bounds of the CIs that the procedure will produce in case applied. We illustrate the result of the method of [25] that uses the Sidak correction.

The method of [25], the partitioning procedure that uses the LR, the Tukey-based procedure and its rescaled version gave all the same result. The rescaled version of the partitioning procedure that uses the LR delivered the best result. Furthermore, all the methods single out the best and second best hotels. The rescaled version of the partitioning procedure that uses the LR singles out the worst two hotels.

9. Discussion

We presented in this paper a generic method for simultaneous CIs for ranks where we partitioned the parameter space $\mathbb{R}^n$ into sets defined through possible orderings of a set of means $\mu_1, \cdots, \mu_n$. The Partitioning principle allowed to control the FWER below α by testing each set at level α which was used to construct simultaneous CIs for the ranks at level $1-\alpha$. We showed that any procedure producing simultaneous CIs for ranks could be written as a partitioning procedure with a suitable local test for the partitions.

We presented an example of our procedure using the likelihood ratio test and also showed that a recently developed method based on Tukey's HSD could be written as a partitioning procedure. We proposed rescaled versions of these two methods by embedding them inside a new partitioning procedure. Although the rescaled version uniformly improve these methods, they are computationally feasible only up to 10 means. Recall that the procedure that uses the LRT is feasible up to $n = 40$ when the standard deviations are equal and only up to $n = 10$ when they are not equal. The Tukey-based approach has a polynomial complexity and is feasible for large n.

In [1], the authors propose a rescaling method based on empirical evidence in order to reduce the conservativeness of the Tukey-based method. The idea is to rescale the Tukey-based method with respect to a worst-case which is different from our rescaling idea in this paper where the rescaling is done for each partition separately. We believe that a similar method can be developed for our LR-based method which could lead to a procedure that is more computationally feasible.

We assumed the standard errors to be known, which is a standard assumption in most papers considering confidence intervals for ranks, see [34, 40, 25] among others. This assumption becomes challenging when the standard errors are estimated with a few measurements (patients, rating, etc.). In Appendix D.4, a simulation example shows that using estimated standard errors still results in conservative CIs for the ranks with close results to when we used the true standard errors except for the case when there are only three measurements for each sample mean. More extensive simulations are needed and developing rigorous approach under the assumption of unknown standard errors remains an open question.

For a different objective, it is possible to look for the rank of only one prespecified institution that we are interested in. [16] use the partitioning principle to make multiple comparisons to the best or to a control. Combining their work with ours could be the objective of a future work.

We provide in this appendix proofs of the main results in the paper and a detailed algorithm of how to perform the partitioning scheme when we use the likelihood ratio (LR) test. It also includes further simulations and the raw data that we collected for the data analysis section of the paper.

Appendix A: Proofs

A.1. Proof of Proposition 1

Proof. Since the partitioning principle ensures that the FWER is below α, we may write

$$\mathbb{P}\left(\text{Number of type I errors} \geq 1\right) \leq \alpha$$

which is equivalent to

$$1 - \mathbb{P}\left(\text{Number of type I errors} = 0\right) \leq \alpha.$$

Denote $\cup_{i\in I} P_i$ the set of rejected elementary hypotheses at level α and μ_T the true vector of means. We can write

$$\mathbb{P}\left(\mu_T \notin \cup_{i\in I} P_i\right) \geq 1 - \alpha.$$

Since the P_i's partition the parameter space $\mathbb{R}^n$, then

$$\mathbb{P}\left(\mu_T \in \cup_{i\notin I} P_i\right) \geq 1 - \alpha.$$

Finally, recall that each partition represents a single set of set-ranks of the means. Thus, the union of unrejected partitions implies a set of simultaneous confidence intervals for the ranks of the means, this set has a confidence level of at least $1 - \alpha$. □

A.2. Proof of Proposition 2

Following the example in figure 1, we arrange the set of elementary hypotheses by levels according to the number of ties between the means. The 1^{st} level corresponds to the hypothesis where all means are tied. The second level corresponds to hypotheses with $n-1$ ties and so on. The n^{th} level corresponds to hypotheses without any ties. We calculate the number of hypotheses in each level and then sum them up, that is the hypotheses having the same number of inequalities between the means.

At level $n - i$, for $i \in \{0, \cdots, n-1\}$, with i equalities, we have i equalities and $n - i - 1$ inequalities. Any partition H from level $n - i$ can be written as a set of $n - i - 1$ blocks $H : B_1 < \cdots < B_{n-i-1}$ where each block includes means which are related to each others by an equality. Given a set of blocks, the number of different orderings of these blocks is equal to $(n - i - 1)!$. It remains then to calculate the number of possible partitions for a given ordering of the means. This number is the same for all possible orderings. Assume then that $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_n$. The indexes are the set $\{1, \cdots, n\}$ and the blocks are mere ordered subsets (or partitions) of indexes which are disjoint and whose union is equal to the whole set $\{1, \cdots, n\}$. This is an ordered partition of the set $\{1, \cdots, n\}$ [41, Lemma 1.4.11] and the number of partitions of a set of n numbers into $n-i-1$ blocks is equal to the Sterling number of the second kind $S(n, n-i-1)$, see Stanley [41, Section 1.9, page 81]. Finally, the overall number of configurations in level $n - i - 1$ is equal to $(n - i - 1)!S(n, n - i - 1)$.

A.3. Proof of Proposition 3

Due to equation (3), it is straightforward that φ is a valid test for H at level α. Indeed,

$$\begin{aligned}\mathbb{P}(\varphi(H) = 1) &= 1 - \mathbb{P}(\varphi(H) = 0)\\ &= 1 - \mathbb{P}(r_i(H) \subset [\tilde{L}_i, \tilde{U}_i], \forall i)\\ &\leq 1 - (1 - \alpha)\\ &\leq \alpha.\end{aligned}$$

A.4. Proof of Proposition 4

We first show that $[L_i, U_i] \subset [\tilde{L}_i, \tilde{U}_i]$. This is straightforward because by construction of the test φ, a partition (a set of set-ranks) is not rejected only if it induces set-ranks in the CIs $[\tilde{L}_i, \tilde{U}_i]$. Moreover, the confidence intervals for the ranks based on the partitioning scheme are built based on only the unrejected partitions. Thus, the inclusion holds.

We need to show the converse. We show first that there exist permutations $\pi_1, \cdots, \pi_J \in \mathcal{S}_n$ such that for some $J \in \mathbb{N}$

$$[\tilde{L}_1, \tilde{U}_1] \times \cdots \times [\tilde{L}_n, \tilde{U}_n] = \bigcup_{(r_1,\cdots,r_n)=\pi_j(1,\cdots,n), j=1,\cdots,J} \{r_1\} \times \cdots \times \{r_n\}. \quad (11)$$

Indeed, the ranks $1, \cdots, n$ are simultaneously included in the CIs. In other words,

$$\begin{aligned} &\exists j_1, \text{ s.t. } 1 \in [\tilde{L}_{j_1}, \tilde{U}_{j_1}]; \\ &\qquad \vdots \\ &\exists j_n, \text{ s.t. } n \in [\tilde{L}_{j_n}, \tilde{U}_{j_n}]. \end{aligned} \quad (12)$$

In order to show (11), let $r_{1,1} \in [\tilde{L}_1, \tilde{U}_1]$. We start with $r_{1,1} = \tilde{L}_1$. Due to (12), there exist $r_{1,2} \neq r_{1,1}, \cdots, r_{1,n} \neq r_{1,1}$ such that $r_{1,t} \in [\tilde{L}_j, \tilde{U}_j]$ for all $j = 1, \cdots, n$ and such that $(r_{1,1}, \cdots, r_{1,n}) = \pi_1(1, \cdots, n)$ for some permutation $\pi_1 \in \mathcal{S}_n$. We then take $r_{2,1} = \tilde{L}_1 + 1$ (in case $\tilde{U}_1 > \tilde{L}_1$), and do the same as with the first step to obtain $r_{2,t} \in [\tilde{L}_t, \tilde{U}_t]$ for all $t = 1, \cdots, n$ and such that $(r_{2,1}, \cdots, r_{2,n}) = \pi_2(1, \cdots, n)$. We repeat this until we scan all the ranks in the interval $[\tilde{L}_1, \tilde{U}_1]$. This way we scan all the ranks in the first CI through sets of the form $\{r_1\} \times \cdots \times \{r_n\}$. Set

$$A_1 = \bigcup_{s=1}^{\tilde{U}_1 - \tilde{L}_1 + 1} \{r_{s,1}\} \times \cdots \times \{r_{s,n}\};$$

$$\Pi_1 = \bigcup_{s=1}^{\tilde{U}_1 - \tilde{L}_1 + 1} \{\pi_s\}.$$

If $\{r_{1,2}, \cdots, r_{\tilde{U}_1 - \tilde{L}_1 + 1, 2}\} \subsetneq [\tilde{L}_2, \tilde{U}_2]$, we then scan the remaining elements of $[\tilde{L}_2, \tilde{U}_2]$ (that is $[\tilde{L}_2, \tilde{U}_2] \setminus \{r_{1,2}, \cdots, r_{\tilde{U}_1 - \tilde{L}_1 + 1, 2}\}$) in a similar way to the elements of the interval $[\tilde{L}_1, \tilde{U}_1]$ and we set the union of the resulting sets as A_2. Finally, we obtain the sets $A_1, \cdots, A_n$ (some of them may be empty). By construction, we have

$$[\tilde{L}_1, \tilde{U}_1] \times \cdots \times [\tilde{L}_n, \tilde{U}_n] = \bigcup_{i=1}^{n} A_i.$$

Equation (11) shows that the simultaneous CIs for the ranks $[\tilde{L}_i, \tilde{U}_i]$ can be represented by sets of single ranks. Let $\{r_1\} \times \cdots \times \{r_n\}$ be one of these sets. Let $H : \mu_{i_1} < \cdots < \mu_{i_n}$ be a partition such that $i_{r_1} = 1, \cdots, i_{r_n} = \mu_n$. In other words, μ_i is in the r_ith position. This partition is not rejected because the ranks of any vector of means $\mu_1, \cdots, \mu_n \in H$ are actually $r_1, \cdots, r_n$, thus they are in the confidence intervals $[\tilde{L}_1, \tilde{U}_1], \cdots, [\tilde{L}_n, \tilde{U}_n]$ respectively. Therefore, the partition H is not rejected. Hence, for any set of ranks $\{r_1\} \times \cdots \times \{r_n\}$, we can find a partition H which is not rejected using our test φ. Therefore,

$$[\tilde{L}_1, \tilde{U}_1] \times \cdots \times [\tilde{L}_n, \tilde{U}_n] \subset [L_1, U_1] \times \cdots \times [L_n, U_n].$$

A.5. Proof of Proposition 5

The proof requires the following Lemma.

Lemma 4. *Let $B_1, \cdots, B_l$ be subsets that partition the set $\{\mu_1, \cdots, \mu_n\}$ so that $B_i \cap B_j = \emptyset$ and $\cup_i B_i = \{\mu_1, \cdots, \mu_n\}$. Assume that we obtain $\tilde{B}_1, \cdots, \tilde{B}_l$ by swapping μ_{j_1} with μ_{j_2} such that $y_{j_1} < y_{j_2}$ (so that all subsets remain the same except for two). Let $\hat{\mu}_j$ denote the sample mean over block B_j whereas $\tilde{\mu}_j$ denote the sample mean over block $\tilde{B}_j$. Then*

$$\sum_{j=1}^{l} \sum_{\mu_i \in B_j} (y_i - \hat{\mu}_j)^2 \leq \sum_{j=1}^{l} \sum_{\mu_i \in \tilde{B}_j} (y_i - \tilde{\mu}_j)^2.$$

In particular, if $H : B_1 < \cdots < B_l$ and $\tilde{H} : \tilde{B}_1 < \cdots < \tilde{B}_l$ be two (partially) correctly ordered hypotheses, then

$$LR(H) \geq LR(\tilde{H}).$$

Proof. Let $\tilde{B}_{i_1}$ and $\tilde{B}_{i_2}$ be the two subsets that have changed due to swapping μ_{j_1} with μ_{j_2}. Let also B_{i_1} and B_{i_2} be the corresponding subsets before swapping. When both H and $\tilde{H}$ are partially correctly ordered hypotheses, then

$$\begin{aligned} \hat{\mu}_{i_1-1} <& \hat{\mu}_{i_1} < \hat{\mu}_{i_1+1} < \cdots < \hat{\mu}_{i_2-1} < \hat{\mu}_{i_2} < \hat{\mu}_{i_2+1}; \\ \hat{\mu}_{i_1-1} <& \tilde{\mu}_{i_1} < \hat{\mu}_{i_1+1} < \cdots < \hat{\mu}_{i_2-1} < \tilde{\mu}_{i_2} < \hat{\mu}_{i_2+1}. \end{aligned}$$

Thus, we can write easily the LR of both hypotheses H and $\tilde{H}$ as

$$\begin{aligned} LR(H) =& \sum_{s \in \{1,\cdots,l\} \setminus \{i_1, i_2\}} \sum_{\mu_i \in B_s} (y_i - \hat{\mu}_s)^2 + \sum_{\mu_i \in B_{i_1}} (y_i - \hat{\mu}_{i_1})^2 + \sum_{\mu_i \in B_{i_2}} (y_i - \hat{\mu}_{i_2})^2; \\ LR(\tilde{H}) =& \sum_{s \in \{1,\cdots,l\} \setminus \{i_1, i_2\}} \sum_{\mu_i \in B_s} (y_i - \hat{\mu}_s)^2 + \sum_{\mu_i \in \tilde{B}_{i_1}} (y_i - \tilde{\mu}_{i_1})^2 + \sum_{\mu_i \in \tilde{B}_{i_2}} (y_i - \tilde{\mu}_{i_2})^2. \end{aligned}$$

We study the contribution of the blocks that have changed. Let

$$\begin{aligned} LR(B_{i_1}, B_{i_2}) &= \sum_{\mu_i \in B_{i_1}} (y_i - \hat{\mu}_{i_1})^2 + \sum_{\mu_i \in B_{i_2}} (y_i - \hat{\mu}_{i_2})^2, \\ LR(\tilde{B}_{i_1}, \tilde{B}_{i_2}) &= \sum_{\mu_i \in \tilde{B}_{i_1}} (y_i - \tilde{\mu}_{i_1})^2 + \sum_{\mu_i \in \tilde{B}_{i_2}} (y_i - \tilde{\mu}_{i_2})^2. \end{aligned}$$

Note that the two likelihood ratios $LR(H)$ and $LR(\tilde{H})$ have the same first term. Therefore, it suffices to prove that

$$LR(B_{i_1}, B_{i_2}) \leq LR(\tilde{B}_{i_1}, \tilde{B}_{i_2}). \tag{13}$$

Note that $\#B_{i_1} = \#\tilde{B}_{i_1} = n_{i_1}$ and $\#B_{i_2} = \#\tilde{B}_{i_2} = n_{i_2}$. Moreover, $B_{i_1} \cup B_{i_2} = \tilde{B}_{i_1} \cup \tilde{B}_{i_2} := B_{i_1 i_2}$. Denote $\hat{\mu}_{i_1 i_2}$ the sample mean over $B_{i_1 i_2}$. We have [see 4, Theorem 5]

$$\begin{aligned}\sum_{\mu_i \in B_{i_1 i_2}} (y_i - \hat{\mu}_{i_1 i_2})^2 &= LR(B_{i_1}, B_{i_2}) + \frac{n_{i_1} n_{i_2}}{n_{i_1} + n_{i_2}} (\hat{\mu}_{i_2} - \hat{\mu}_{i_1})^2 \\ &= LR(\tilde{B}_{i_1}, \tilde{B}_{i_2}) + \frac{n_{i_1} n_{i_2}}{n_{i_1} + n_{i_2}} (\tilde{\mu}_{i_2} - \tilde{\mu}_{i_1})^2 .\end{aligned}$$

Finally, since all these terms are non negative, then in order to prove the lemma, it suffices to compare $\hat{\mu}_{i_2} - \hat{\mu}_{i_1}$ with $\tilde{\mu}_{i_2} - \tilde{\mu}_{i_1}$. It is straightforward to see that

$$\hat{\mu}_{i_2} - \hat{\mu}_{i_1} - \tilde{\mu}_{i_2} - \tilde{\mu}_{i_1} = \left(\frac{1}{n_{i_1}} + \frac{1}{n_{i_2}}\right)(y_{i_2} - y_{i_1}) > 0$$

and

$$LR(H) \leq LR(\tilde{H}). \qquad \square$$

Without loss of generality, we assume that $\sigma_i = 1$ for all $i = 1, \cdots, n$. The proof consists of two main parts. We prove in the first part that it suffices to test only hypotheses corresponding to cases 1 and 2. In other words, there is no need to test incorrectly ordered hypotheses (case 3). We show in the second part that not all the hypotheses corresponding to case 2 need to be tested and give only the relevant list.

We prove the first part. Consider a hypothesis H_l from the l^{th} level, that is it contains $l-1$ inequalities. Write this hypothesis as a union of blocks where each block contains all means which are equal under H_l, that is $H_l = A_1 < \cdots < A_l$. Suppose that this hypothesis is incorrectly ordered. According to Proposition 1, we are interested in H_l only if it is not rejected. Suppose then that the hypothesis H_l is not rejected. When we calculate the maximum likelihood under this hypothesis by the pool adjacent violators algorithm (PAVA), adjacent blocks which violate the ordering $\hat{\mu}_{A_1} < \cdots < \hat{\mu}_{A_l}$ will be pooled together. By merging the pooled blocks of hypothesis H_l, we can construct a partially correctly ordered hypothesis $\bar{H}_s = \{\tilde{A}_1, \cdots, \tilde{A}_s\}$ with $s < l$ such that $\hat{\mu}_{\tilde{A}_1} < \cdots < \hat{\mu}_{\tilde{A}_s}$. Note that $LR(\bar{H}_s) = LR(H_l)$ due to the PAVA. Moreover, the adaptive critical value is also the same since it depends on the PAVA solution. Thus, the non rejection of H_l will imply the non rejection of the hypothesis $\bar{H}_s$. The set-ranks induced by H_l are subsets of the set-ranks induced by $\bar{H}_s$ since in the later the pooled blocks become one so that their means are equal under $\bar{H}_s$ whereas they where not under H_l. Thus, testing the partially correctly ordered hypothesis $\bar{H}_s$

We prove the second part. The partially correctly ordered hypotheses result from the correctly ordered hypotheses by switching at least a pair of means in a way that the switching does not result in a modification of the ordering of the observed means inside the blocks. Moreover, the switching only influences the position of the means and not the size of the blocks defining the hypothesis. We need to show two things.

1. If a partially correctly ordered hypothesis is not rejected then the corresponding correctly ordered hypothesis is not rejected either. This allows to conclude that we need to look at switches only if we find a correctly ordered hypothesis which is not rejected. As long as we are rejecting the correctly ordered hypothesis, we do not need to care about partially correctly ordered ones because they are automatically rejected.
2. If a correctly ordered hypothesis is not rejected, then we need to consider permutations of indexes only from the list (8).

We prove the first claim. Let H be any hypothesis (correctly ordered or partially correctly ordered) that consists of l blocks such that $\hat{\mu}_{B_1} < \cdots < \hat{\mu}_{B_l}$. Assume that we switch between two means μ_{j_1} from block B_{i_1} with mean μ_{j_2} from block B_{i_2} such that $j_1 < j_2$. Assume also that this permutation does not result in changing the hypothesis from being (partially) correctly ordered into incorrectly ordered hypothesis. Due to Lemma 4, we have

$$LR(H) \leq LR(\tilde{H}).$$

Now, if the hypothesis $\tilde{H}$ is not rejected, then so does H since they are tested against the same adaptive quantile, that is a quantile of $\chi^2(l)$. Conversely, if the hypothesis H is rejected, then so does $\tilde{H}$.

Last but not least, assume that a partially correctly ordered hypothesis $\tilde{H}$ results from a correctly ordered hypothesis H by permuting s means following some permutation p. It is possible to write p as the composition of a finite set of transpositions, that is there exist $m \leq s$ transpositions τ_i such that $p = \tau_m\tau_2 ... \tau_1$. Applying the permutation p on the set of means indexes is equivalent to applying successively the transpositions on the set of means. In other words, the hypothesis $\tilde{H}$ is the result of m single switches applied successively on the indexes of means considered in H. Denote $\tau(H)$ the hypothesis which results from H by applying the transposition τ on the means indexes. Then

$$\tilde{H} = \tau_1\tau_2 \cdots \tau_m(H)$$

In order to apply Lemma 4, the transpositions must change the positions of two means $\mu_{j_1} < \mu_{j_2}$ (under $\tilde{H}$) only if $y_{j_1} > y_{j_2}$. In order to do so, we start by picking the mean which corresponds to y_1 (the smallest observation), that is μ_1. If it is already in position 1 in $\tilde{H}$, we do nothing, otherwise, we switch it with the mean in position 1 in $\tilde{H}$. We thus set $\tau_1 = (1, i_{y_1})$. More generally, let i_{y_j} be the position of μ_j in $\tilde{H}$. Then, we have

$$\tau_j = (j, i_{y_j}).$$

Some of these transpositions may be the identity function so that only $m \leq s$ transpositions remain. Thus, by recurrence and using Lemma 4, we have

$$LR(\tilde{H}) \geq LR(\tau_1(\tilde{H})) \geq \cdots LR(\tau_{m-1} \cdots \tau_1(\tilde{H})) \geq LR(H)$$

This reads as follows. Any supplementary switch between two means in a (partially) correctly ordered hypothesis results in increasing the LR.

We prove now our second claim. Since we need to consider a partially correctly ordered hypothesis only when the corresponding correctly ordered hypothesis is not rejected, let H be a correctly ordered hypothesis which is not reject. Let $\tilde{H}$ be some partially correctly ordered hypothesis which results from H by permuting the means indexes using a permutation p such that $\tilde{H}$ is not rejected. We show that if $\tilde{H}$ induces wider CI for the rank of $\mu_{i,T}$ than H, then there exist permutations $p_1, \cdots, p_k$ from the list (8) such that the partially correctly ordered hypotheses resulting from applying these permutations on the indexes of the means through H, denoted as before $p_1(H), \cdots, p_k(H)$ are not rejected. Furthermore, the unrejection of those hypotheses result in the same CI for the rank of $\mu_{i,T}$ as $\tilde{H}$. This suffices to conclude that only permutations from the list (8) are needed.

Any permutation has a disjoint decomposition of cycles. Two cycles in this decomposition have disjoint orbits. Two disjoint cycles modify the set-ranks of two disjoint groups of means. Therefore, it is possible to treat each cycle separately. For this reason and without loss of generality, we assume that $p = (i_1, \cdots, i_k)$ is a permutation with one cycle. Note that if the orbit is smaller than n, that is $k < n$, then the permutation p leaves some of the means in their own position. Otherwise, all the means move from their original positions in H to new ones in $\tilde{H}$.

Let $s \in \{1, \cdots, n\}$. Suppose that the original position of mean μ_{i_s} in H is i_{s-1}, then its new position in $\tilde{H}$ is i_s with the convention $i_0 = i_k$. If $i_s > i_{s-1}$, then $\mu_{i_{s-1}}$ moves forward in $\tilde{H}$ (with respect to H). Otherwise, it moves backward in $\tilde{H}$. The proof slightly differs according to whether $\mu_{i_{s-1}}$ moves forward or backward.

We assume first that $\mu_{i_{s-1}}$ moves forward in $\tilde{H}$. It is possible to reorder all the means which have new positions different from i_s by composing p successively with suitable transpositions. The reordering will be done based on the corresponding observed values. We will prove that this reordering results in a decrease of the LR or at least does not increase it. Indeed, we choose the mean with the maximum observed value among the means with new positions different from i_s. If its new position is different from n, say $i_{\max_1}$, then there is some mean whose new position is n and whose observed value is inferior to the maximum. We switch these two by composing p with the transposition $(i_{\max_1}, n)$. This single reordering puts a mean with a small observed value back before another mean with a larger observed value. Therefore, this single reordering does not make the LR increase similarly to (13). Now, we consider again the set of means whose new positions in $(i_{\max_1}, n)p(H) = (i_{\max_1}, n)\tilde{H}$ are different from i_s except for the one who is at position n, that is the set $\{1, \cdots, n-1\} \setminus \{i_s\}$. We choose the mean with maximum observed value. If its new position, say $i_{\max_2}$, is inferior to $n-1$, then we switch it with the one whose new position is $n-1$ by composing $(i_{\max_1}, n)\tilde{H}$ with the transposition $(i_{\max_2}, n-1)$. Similarly to the previous switch, this one also makes the LR decrease (or at least does not increase). We iterate this procedure t times until we reorder all the means whose new positions are different from i_s. The result of this reordering is denoted $\tilde{H}_t$ and is given by

$$\tilde{H}_t = (i_{\max_t}, n-t+1)\cdots(i_{\max_1}, n)\tilde{H}.$$

This can also be written as

$$\tilde{H} = (i_{\max_t}, n-t+1)\cdots(i_{\max_1}, n)\tilde{H}_t.$$

so that using Lemma 4, we have

$$LR(\tilde{H}_t) \leq LR(\tilde{H}). \tag{14}$$

Moreover, we can write $\tilde{H}_t$ explicitly as

$$\begin{aligned}\tilde{H}_t : \mu_1 \mathrel{=\!\!\cdot\!\!<} \cdots \mathrel{=\!\!\cdot\!\!<} \mu_{i_{s-1}-1} \mathrel{=\!\!\cdot\!\!<} \mu_{i_{s-1}+1} \mathrel{=\!\!\cdot\!\!<} \cdots \mathrel{=\!\!\cdot\!\!<} \mu_{i_s} \mathrel{=\!\!\cdot\!\!<} \mu_{i_{s-1}} \mathrel{=\!\!\cdot\!\!<} \mu_{i_s+1} \\ \mathrel{=\!\!\cdot\!\!<} \cdots \mathrel{=\!\!\cdot\!\!<} \mu_n\end{aligned}$$

In other words,

$$\begin{aligned}\tilde{H}_t =& (i_s, \cdots, i_{s-1})H \\ =& (i_s, i_{s-1})\cdots(i_{s-1}+1, i_{s-1})H.\end{aligned}$$

Thus using Lemma 4, we have $LR(\tilde{H}_t) \leq LR(H)$ which together with (14) implies

$$LR(H) \leq LR(\tilde{H}_t) \leq LR(\tilde{H}).$$

We conclude that if $\tilde{H}$ is not rejected, then any mean whose position in H moves forward in $\tilde{H}$ does not get a wider CI for its rank than the CI that it gets from testing the partially correctly ordered hypotheses resulting from applying the list (8) on H.

Last but not least, if μ_{i_s} moves backward in $\tilde{H}$ with respect to H to position i_{s-1}, then similar steps to the previous case allows to reorder the means whose new positions are different from i_{s-1}. Denote the resulting hypothesis $\bar{H}_t$, we have

$$\bar{H}_t = (i_{\max_t}, n-t+1)\cdots(i_{\max_1}, n)\tilde{H},$$

which can be written as

$$\tilde{H} = (i_{\max_t}, n-t+1)\cdots(i_{\max_1}, n)\bar{H}_t,$$

so that

$$LR(\bar{H}_t) \leq LR(\tilde{H}). \tag{15}$$

We can write $\bar{H}_t$ explicitly as

$$\begin{aligned}\bar{H}_t : \mu_1 \mathrel{=\!\!\cdot\!\!<} \cdots \mathrel{=\!\!\cdot\!\!<} \mu_{i_{s-1}-1} \mathrel{=\!\!\cdot\!\!<} \mu_{i_s} \mathrel{=\!\!\cdot\!\!<} \mu_{i_{s-1}} \mathrel{=\!\!\cdot\!\!<} \cdots \mathrel{=\!\!\cdot\!\!<} \mu_{i_s-1} \mathrel{=\!\!\cdot\!\!<} \mu_{i_s+1} \\ \mathrel{=\!\!\cdot\!\!<} \cdots \mathrel{=\!\!\cdot\!\!<} \mu_n\end{aligned}$$

In other words,

$$\begin{aligned}\bar{H} =&(i_{s-1}, \cdots, i_s)H\\ =&(i_{s-1}, i_s)\cdots(i_s-1, i_s)H.\end{aligned}$$

Using Lemma 4, we get $LR(\bar{H}_t) \leq LR(H)$ which together with (15) implies

$$LR(H) \leq LR(\bar{H}_t) \leq LR(\tilde{H}).$$

We conclude that if $\tilde{H}$ is not rejected, then any mean whose position in H moves backward in $\tilde{H}$ does not get a wider CI for its rank than the CI that it gets from testing the partially correctly ordered hypotheses resulting from applying the list (8) on H.

To end the proof, since any transposition (i, j) is the composition of transpositions $(i, i+1), \cdots, (j-1, j)$, we conclude that for any partially correctly ordered hypothesis that we do not reject, we may construct partially correctly ordered hypotheses using the list (8) which are not rejected either and which produce the same CIs for the ranks of $\mu_{1,T}, \cdots, \mu_{n,T}$.

Finally, if we test the list (8) column after column, then for each column it suffices to test until one of the permutations gets rejected then the remaining permutations with a larger orbit (the set of indexes to permute) will automatically be rejected. Indeed, by Lemma 4, as the orbit of the permutation contains more means, the LR increases.

A.6. Proof of Proposition 6

See the first part of the proof of Proposition 5.

A.7. Proof of Lemma 2

Proof. We characterize the event $\{\varphi_{\text{LR}}(H) = 1\}$ when $\mu \in H$. We abbreviate PP for the partitioning procedure that uses φ_{LR} as a local test, and PLR for the partitioning procedure which uses the LR as a local test. Let $[L_i, U_i]$ for $i = 1, \cdots, n$ be the set of simultaneous CIs produced by PLR. Note that according to Proposition 4, the simultaneous CIs produced by PP are the same as the ones produced by PLR, which are $[L_i, U_i]$ for $i = 1, \cdots, n$. For $\mu = (\mu_1, \cdots, \mu_n)$, let $r_i(H)$ be the set-rank of μ_i when $\mu \in H$. According to the definition of φ_{LR}, we reject H in PP ($\varphi_{\text{LR}} = 1$) if for any $\mu \in H$, $r_i(H) \nsubseteq [L_i, U_i]$ for some $i \in \{1, \cdots, n\}$. In other words,

$$\{\varphi_{\text{LR}}(H) = 1\} = \bigcup_{i=1}^{n} \{r_i(H) \nsubseteq [L_i, U_i]\}.$$

The event $\{r_i(H) \nsubseteq [L_i, U_i]\}$ occurs if we reject all the elementary hypotheses $\bar{H}$ in PLR that have μ_i in one of the positions $j \in r_i(H)$. We may now write

$$\{\varphi_{\text{LR}}(H) = 1\} = \bigcup_{i=1}^{n} \bigcap_{\bar{H}: r_i(\bar{H}) \subset r_i(H)} \{LR(Y, \bar{H}) > q(\bar{H})\}.$$

Recall that [20] if $Y = (y_1, \cdots, y_n)$ and $V = \text{diag}(\sigma_1^2, \cdots, \sigma_n^2)$, then

$$LR(Y, \bar{H}) = \|Y - \bar{H}\|_V^2.$$

Using Proposition 3.12.1 from [39], if $\mu \in \bar{H}$, then

$$\|Y - \bar{H}\|_V^2 \leq \|Y - \mu - \bar{H}\|_V^2$$

so that

$$\{LR(Y, \bar{H}) > q(\bar{H})\} \subset \{LR(Y - \mu, \bar{H}) > q(\bar{H})\}.$$

Since Y has a mean μ under H, then we prove the lemma. □

Further results for the Tukey-based method

Lemma 5. *For the partitioning procedure defined for the Tukey-based method using the local test* φ_{TKY}*, it suffices to test only the correctly ordered hypotheses, that is the hypotheses whose ordering does not violate the empirical one.*

Let H be an elementary hypothesis. Without loss of generality, suppose that it has only three blocks $H : B_1 < B_2 < B_3$. Suppose that the empirical ordering is such that $\max_{\mu_i \in B_1} y_i > \min_{\mu_i \in B_2} y_i$, then our testing procedure will pool B_1 and B_2 into $\tilde{B}_1$. In the same spirit of the proof of Proposition 5 and according to Proposition 1, if H is rejected, this changes nothing in terms of the confidence intervals and we only need to look at the unrejected hypotheses.

Suppose now, that H is not rejected, then

$$\max_{\mu_j \in \tilde{B}_1} \frac{|y_j - y_{i_1}|}{\sqrt{\sigma_{i_1}^2 + \sigma_j^2}} \leq q_{1-\alpha}, \text{ and } \max_{\mu_j \in H_{i,3}} \frac{|y_j - y_{i_3}|}{\sqrt{\sigma_{i_3}^2 + \sigma_j^2}} \leq q_{1-\alpha} \tag{16}$$

where y_{i_1} and y_{i_3} correspond to the smallest observed values related to the means in blocks $\tilde{B}_1$ and B_3 respectively. The hypothesis $\tilde{H} : \tilde{B}_1 < B_3$ is also an elementary hypothesis whose ordering coincides with the empirical one so that it is a correctly ordered one. Besides, this hypothesis is not rejected due to (16) because on the one hand, it has the same test statistic as H_i and on the other hand, it is tested against the same common critical value $q_{1-\alpha}$. Thus, for any hypothesis H with incorrect ordering, there exists a correctly ordered hypothesis $\tilde{H}$ which has the same test statistic so that whenever one of them is not rejected the other one is not, either.

Proposition 8. *Assume that we have a common standard deviation* σ*. In terms of ranks, the partitioning procedure defined using the rejection region (10) is equivalent to the Tukey-based method of [1]. In other words, they produce the same simultaneous confidence intervals for the ranks of the means* $\mu_{1,T}, \cdots, \mu_{n,T}$ *at level* $1 - \alpha$*.*

Due to Lemma 5, we only need to test the correctly ordered hypotheses. The rejection region for these hypotheses turns out to be a calculus of the maximum of the maximal differences inside the blocks composing the hypothesis.

Take mean $\mu_{i,T}$. Suppose that with the Tukey-based procedure, we determine a confidence interval for the rank of $\mu_{i,T}$ to be $[L_i, U_i]$. This means that we could not reject all hypotheses $\mu_i = \mu_j$ for $j \in [L_i, U_i]$. In other words, we have:

$$\frac{|y_i - y_j|}{\sqrt{2\sigma^2}} \leq q_{1-\alpha}, \quad \forall j \in [L_i, U_i].$$

Besides, we reject all hypotheses $\mu_i = \mu_l$ for $l \leq L_i - 1$ and $l \geq U_i + 1$. In other words

$$\frac{|y_l - y_i|}{\sqrt{2\sigma^2}} > q_{1-\alpha}, \quad \forall l \in \{1, \cdots, L_i - 1\} \cup \{U_i + 1, \cdots, n\}.$$

Let us check what is the confidence interval that we can get using the partitioning with (10) from these rejections and non rejections. First of all, we have

$$\frac{y_{U_i+1} - y_i}{\sqrt{2\sigma^2}} > q_{1-\alpha}, \qquad \frac{y_i - y_{L_i-1}}{\sqrt{2\sigma^2}} > q_{1-\alpha}$$

Thus any partition containing the block $\mu_i = \cdots = \mu_{U_i+1}$ or the block $\mu_{L_i-1} = \cdots = \mu_i$ (or larger ones) is rejected using the rejection region (10). This also entails that any hypothesis producing a larger confidence interval (more equalities) will also be rejected. Therefore, we can conclude that the confidence interval for $\mu_{i,T}$ produced by the partitioning procedure is at most the one produced by the Tukey-based method, that is $[L_i, U_i]$.

Suppose now that with the partitioning procedure, we get a confidence interval for $\mu_{i,T}$ equal to $[L_P, U_P]$. We are then sure that any hypothesis containing the block $\mu_i = \cdots = \mu_{U_P+1}$ or the block $\mu_{L_P-1} = \cdots = \mu_i$ is also rejected. In particular, the hypotheses $\{\mu_1 < \cdots < \mu_i = \cdots = \mu_{U_P+1} < \cdots < \mu_n\}$ and $\{\mu_1 < \cdots < \mu_{L_P-1} = \cdots = \mu_i < \cdots < \mu_n\}$ are rejected. This means that

$$\max_{j=i,\cdots,U_P+1} \frac{|y_i - y_j|}{\sqrt{2\sigma^2}} = \frac{y_{j_1} - y_i}{\sqrt{2\sigma^2}} > q_{1-\alpha}, \max_{j=L_P-1,\cdots,i} \frac{|y_i - y_j|}{\sqrt{2\sigma^2}} = \frac{y_i - y_{j_0}}{\sqrt{2\sigma^2}} > q_{1-\alpha}.$$

for some $j_0 \in \{L_P - 1, \cdots, i\}$ and $j_1 \in \{i, \cdots, U_P + 1\}$ verifying

$$\begin{aligned} &\forall j \in \{i, \cdots, U_P + 1\}, \qquad \frac{y_{j_1} - y_i}{\sqrt{2\sigma^2}} > \frac{|y_i - y_j|}{\sqrt{2\sigma^2}} \\ &\forall j \in \{L_P - 1, \cdots, i\}, \qquad \frac{y_i - y_{j_0}}{\sqrt{2\sigma^2}} > \frac{|y_i - y_j|}{\sqrt{2\sigma^2}}. \end{aligned}$$

This entails that with the Tukey-based procedure, we must reject hypotheses $\mu_i = \mu_{j_1}$ and $\mu_{j_0} = \mu_i$. Thus, the confidence interval provided by Tukey's procedure is at most the confidence interval produced by the partitioning, that is $[L_P, U_P]$.

We proved that the Tukey-based procedure cannot produce larger confidence intervals than the partitioning procedure using (10), and that the latter cannot produce larger confidence intervals than the former. Hence, Both methods are equivalent in terms of ranks, that is they produce the same simultaneous confidence intervals for the ranks.

A.8. Proof of Proposition 7

Proof. Assume $y_1 < \cdots < y_n$. Let H be a correctly ordered hypothesis that consists of l blocks, that is $H : B_1 < \cdots < B_l$. Assume that H is not rejected. Let μ_{i_s} (μ_{i_t}, resp.) denote the mean with the smallest (highest, resp.) observed value in block B_i. Since H is not rejected, then for all $j \in \{i_s, \cdots, i_t\}$, the rank CI of μ_j includes the ranks $\{i_s, \cdots, i_t\}$. Since the standard deviations are the same, then it implies that Tukey's procedure does not reject the hypothesis $\mu_{i_s} = \mu_{i_t}$ and any hypothesis $\mu_k = \mu_r$ for $k, r \in \{i_s, \cdots, i_t\}$. This means that if H is not rejected, then

$$\max_{k,r\in\{i_s,\cdots,i_t\}} |y_k - y_r| < q_{1-\alpha}. \tag{17}$$

Similarly, if for all blocks of means in H (17) holds, then $\mu_k = \mu_r$ is not rejected for $\mu_k, \mu_r \in B_i$ for $i = 1, \cdots, l$. Thus, not rejecting H is equivalent to

$$\max_{i=1,\cdots,l} \max_{\mu_k,\mu_r\in B_i} |y_k - y_r| < q_{1-\alpha}.$$

Let $\tilde{H}$ be a hypothesis that results from H by switching μ_i with μ_j. Assume also that $\mu_i \in B_s$ and $\mu_j \in B_t$ and denote $\tilde{B}_s$ and $\tilde{B}_t$ the new blocks after switching μ_i with μ_j. We only need to take care of the blocks $\tilde{B}_s, B_{s+1}$, If $\varphi(\tilde{H}) = 1$, then $\tilde{H}$ is not rejected and μ_i gets rank j whereas μ_j gets rank i. On the other hand, since the empirical ranks are never rejected, μ_i has already rank i in its rank CI. Since the standard deviations are assumed equal, then $\mu_i = \mu_j$ is not rejected by the Tukey procedure. Moreover, for all $k \in \{i+1, \cdots, j-1\}$, Tukey's procedure does not reject $\mu_i = \mu_k$.

Since $\tilde{H}$ is not rejected, then all means in block $\tilde{B}_s$ get the same set-rank. If μ_{i_s} corresponds to the mean with the lowest observed value in block $\tilde{B}_s$, then μ_j gets also rank i_s. Similarly, if μ_{i_t} corresponds to the mean with the highest observed value in block $\tilde{B}_t$, then μ_i gets also rank i_t. Since $y_i - y_{i_s} < y_j - y_{i_s}$, then Tukey's procedure does not reject any of the hypotheses $\mu_i = \mu_k$ for any $k \in \{i_s, \cdots, i_j\}$. This is equivalent to pooling all the blocks $B_s, \cdots, B_t$ into one block. Moreover, not rejecting H is equivalent to $y_{i_t} - y_{i_s} < q$ where q is the Studentized range quantile.

More generally, any conflict of ordering between the empirical ranks and the ranks that the elementary hypothesis imply leads to pooling all the blocks of means in between and all means in these blocks share the same set-ranks. □

A.9. Proof of Lemma 3

Proof. Let $[L_i, U_i]$ for $i = 1, \cdots, n$ be the set of simultaneous CIs produced by the Tukey-based method. Note that according to Proposition 4, the simultaneous CIs produced by partitioning procedure defined on the elementary hypotheses (2) through function φ_{TKY} are also $[L_i, U_i]$ for $i = 1, \cdots, n$. For $\mu = (\mu_1, \cdots, \mu_n)$, let $r_i(H)$ be the set-rank of μ_i when $\mu \in H$.

Using Proposition 3.2 from [1], we have

$$\mathbb{P}_{\mu=0}(r_i \subset [L_i, U_i]) = \alpha.$$

On the other hand,

$$\begin{aligned}\mathbb{P}_{\mu}\left(\varphi_{\mathrm{TKY}}(H) = 1\right) &\leq \alpha \\ &\leq \mathbb{P}_{\mu=0}(r_i \subset [L_i, U_i]) \\ &\leq \mathbb{P}_{\mu=0}\left(\varphi_{\mathrm{TKY}}(H) = 1\right). \end{aligned}$$ □

Appendix B: Testing a simple order

Let $Y_1, \cdots, Y_p$ be random variables distributed independently as $\mathcal{N}(\mu_{i,T}, \sigma_i^2)$ for $i = 1, \cdots, p$. We test the null hypothesis $H : \mu_1 \leq \cdots \leq \mu_p$ against all alternatives based on the observation $(y_1, \ldots, y_n)$. The likelihood ratio can be calculated using the pool adjacent violators algorithm known as the PAVA (Bartholomew [8], van Eeden C. [45]). Function `isoreg` in the statistical program R does the job. Note that the maximum likelihood estimator results from the vector $y = (y_1, \ldots, y_n)$ by pooling certain adjacent observations so that the maximum likelihood estimator has ℓ distinct coordinates at most equal to n. From the literature, [38] proposed to compare the LR statistic with the quantile of a mixture of chi-squares with degrees of freedom ranging from 1 to n. In our paper, this refers to the nonadaptive test since the critical value does not adapt to the form of the maximum likelihood estimator. The nonadaptive test is defined by

$$\mathbb{P}(LR > \gamma) \leq \mathbb{P}_{\mu_i=0,\forall i}(LR > \gamma) = \sum_{j=1}^{n-1} w_{j,n} q_{n-j}$$

where

$$w_{1,n} = \frac{1}{n}, w_{n,n} = \frac{1}{n!}, w_{j,n} = \frac{1}{n} w_{j-1,n-1} + \frac{n-1}{n} w_{j,n-1}$$

These weights can be calculated using Stirling numbers of the second kind, see [35].

The adaptive LR test compares the likelihood ratio statistics with the quantile of a $\chi^2(p-\ell)$ at order $1-\alpha$. The adaptive critical value is given by

$$q(y, \alpha) = q_{p-\ell}.$$

Theorem 1 from [2] shows that this adaptive LR test has level α.

Similarly, if we want to test $H : \mu_1 = \cdots = \mu_m \leq \mu_{m+1} \leq \cdots \leq \mu_p$, then the PAVA provides a solution where the first m observations are always pooled (possibly together with other ones). The adaptive LR test compares the LR statistic with the quantile of a $\chi^2(p-\ell)$ at order $1-\alpha$ where ℓ is the number of levels in the result of the PAVA. Note that $p - \ell \in \{m-1, \cdots, p-1\}$. The nonadaptive test compares the LR statistic with the quantile of a mixture of chi-squares with degrees of freedom ranging from $m-1, \cdots, p-1$.

Appendix C: Algorithms to calculate the confidence intervals for the ranks based on the partitioning principle

We present an algorithm to produce simultaneous confidence intervals (CIs) for ranks based on the partitioning scheme presented in the paper and using the likelihood ratio (LR) test. The algorithm groups the elementary hypotheses in n levels where each level l contains all hypotheses with $l-1$ inequality for $l = 1, \cdots, n$. Figure (1) is reproduced here as an illustration.

It is important to find a suitable way to represent or code the hypotheses so that the generation of these codes is efficiently carried out with a statistical package. We provide two ways of representing the hypotheses. Other possibilities can exist and finding a simpler way to generate and keep track of the hypotheses may improve significantly the performance of the algorithm.

When the standard deviations are the same, then according to Proposition 5, it suffices to test the correctly ordered hypotheses. Then for each unrejected hypotheses, we apply the list of permutations (8) on the means indexes and test them again. When the standard deviations are not the same, then according to Proposition 6, we need to avoid testing only the incorrectly ordered hypotheses. Since we do not know how to represent the partially correctly ordered hypotheses efficiently, then we need to test all the elementary hypotheses keeping in mind that for each hypothesis the LR statistic is calculated using the PAVA. Still, for the correctly ordered and partially correctly ordered hypotheses, the LR statistic has an explicit formula and we do not need to use the PAVA. Therefore, a `PAVA-Check` procedure is needed which tests if the hypothesis is incorrectly ordered. If so, then the hypothesis is skipped, otherwise, we test it. To go through all the partitions, we may start by the correctly ordered hypotheses. Then, we permute the means indexes using some $\pi \in \mathcal{S}_n$ and test the same hypotheses with an additional step to check if the hypothesis is incorrectly ordered using our `PAVA-Check` procedure. We have to go through all the permutations from $\mathcal{S}_n$ in order to map all the elementary hypotheses. However, when the number of means exceeds 10, it becomes computationally infeasible. Therefore, we sample randomly a set of permutations from $\mathcal{S}_n$, say 10^5 permutations, and then apply them on the means indexes. As pointed out in the paper, the list of permutations from Proposition 5 seemed in practice very efficient, therefore we can use it as well.

In any case, testing the correctly ordered hypotheses is in the core of all these algorithms. Therefore, we will present two algorithms to do so, and then elaborate on them in order to include the partially correctly ordered hypotheses.

C.1. A level-by-level algorithm

In this algorithm, the idea is to use the partitioning scheme presented in figure (1) for three centers. We start by explaining the case of the correctly ordered hypotheses. In practice, it is not possible to code all the correctly hypotheses prior to the testing, because this concerns keeping in hand a matrix of size

$2^{n-1} \times c$ where c is the length (or the lengths) of the representation. Thus, for "normal" computers it becomes easily impossible to generate such matrix (or structure) as n grows. Therefore, it is necessary to be able to generate the configurations (representations) one by one to avoid memory issues.

We propose to represent a hypothesis by keeping track of the positions of the inequalities so that a hypothesis is made into groups of means which are equal under that hypothesis. This is the same representation considered in the paper. Let $H : B_1 < \cdots < B_l$. Since the hypotheses are grouped in levels where the level number is given by the number of inequalities, then H belongs to the $(l+1)$th level. This representation of H also provides an efficient way to calculate the LR. Indeed, since we only test hypotheses with a correct ordering w.r.t the empirical one, the PAVA is not needed and the LR for some partition is only a sum of averages of the blocks of equal centers and our representation tells us directly where are the bounds of each block. Indeed, the LR is given by

$$LR = \min_{\mu_1,\cdots,\mu_n \in H} \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\sigma_i^2} = \sum_{j=1}^{l} \sum_{i:\mu_i \in B_j} \frac{(y_i - \hat{\mu}_{B_j})^2}{\sigma_i^2},$$

where

$$\hat{\mu}_{B_j} = \frac{1}{\sum_{k:\mu_k \in B_j} \frac{1}{\sigma_k^2}} \sum_{s:\mu_s \in B_j} \frac{y_s}{\sigma_s^2}.$$

The first level has only one hypothesis which is $\mu_1 = \cdots = \mu_n$. This hypothesis is tested at the beginning of the procedure. The hypotheses from level 2 to level $n-1$ are coded according to the positions of the inequalities among the means in the following manner. Consider first the case of 3 means A, B and C, the representation of the correctly ordered hypotheses (excluding the 1st level), say $A < B = C$, $A = B < C$ and $A < B < C$, is the set

$$\begin{array}{lcl} A < B = C & \rightarrow & (0) \\ A = B < C & \rightarrow & (1) \\ A < B < C & \rightarrow & (0,1) \end{array}$$

For $n \leq 25$, it is possible (on regular computer) to use function `combn` from the `utils` package in the statistical program `R` in order to generate efficiently the set of configurations for levels 2 to $n-1$. For higher values of n, we need to generate these configurations one by one in order to avoid memory issues.

In the `ICRanks` package, we generate the representations for any n in the same way by considering the following function C well-known in combinatorics as the combinatorial number system, see [26]. Consider level $l+1$ where the hypotheses have l inequality. Let $(c_1, \cdots, c_l)$ be a vector of natural numbers such that $c_1 < \cdots < c_l$. Define function C as follows

$$C(c_1, \cdots, c_l) = \binom{c_1}{1} + \cdots + \binom{c_l}{l}.$$

This is a one-to-one function between the set of configurations

$$\{(c_1, \cdots, c_l) \in \mathbb{N}^k, 0 \leq c_1 < \cdots < c_l \leq n-2,\}$$

which represent the correctly ordered hypotheses from level number $l+1$ and the set of numbers

$$S_l = \left\{1, \cdots, \binom{n-1}{l}\right\}.$$

In order to generate the coding, we go through the numbers from S_l. For each number m, we calculate the inverse of function C using Algorithm 1.

```
Algorithm 1: An iterative algorithm to calculate $C^{-1}$.
  Data: Level number $l$ and a number $m$ between 1 and $\max S_l$.
  Result: A vector $(c_1, \cdots, c_l)$ such that $0 < c_1 < \cdots < c_l < n$.
  Set $m_1 = m$;
  Find the maximum natural number $c_l$ such that $c_l!/(l!(c_l - l)!) \leq m_1$;
  Update $m_1 = m_1 - c_l!/(l!(c_l - l)!)$;
  for $i$ from $l-1$ to 1 do
      Find the maximum number $c_i$ such that $c_i!/(i!(c_i - i)!) \leq m_1$;
      Update $m_1 = m_1 - c_i!/(i!(c_i - i)!)$;
  end
```

Algorithm 2 provides a pseudo-code of the procedure explained here above when the standard deviations are the same. If the standard deviations are not equal, Algorithm 3 provides the corresponding pseudo-code. In both algorithms, the set Π refers to the list of permutations (8). The set S represents a subset of $\mathcal{S}_n$ selected randomly that the user provides. For $n \leq 10$, we can take $S = \mathcal{S}_n$, otherwise it becomes computationally infeasible with a normal laptop.

Appendix D: Extended simulations

D.1. The case of different standard deviations

We generate randomly 1000 vectors of means $(\mu_{1,T}, \cdots, \mu_{n,T})$ with $n = 10, 15$ independently according to a Gaussian distribution $\mathcal{N}(0, \tau)$ for $\tau = 1, 3$ and vectors of standard deviations $(\sigma_1, \cdots, \sigma_n)$ according to a uniform distribution over $[0, 3]$. For each couple, a vector of means and a vector of standard deviations, we generate 1000 Gaussian samples $y_1, \cdots, y_n$ such that $y_i \sim \mathcal{N}(\mu_{i,T}, \sigma_i)$. We calculate the simultaneous coverage for $\alpha = 0.1$ for each vector of means. The result is in figure 2. When $\tau = 3$, the number of random permutations is 10^6 whereas it is 10^3 when $\tau = 1$. The list of permutations (8) was also used together to permute the means indexes.

It appears that when the means are far from each other, then it becomes difficult to find the permutations required to make the CIs conservative.

D.2. What if the normality assumption is not valid

For the unusual situations when the normality assumption is not fulfilled but e.g. is affected by skewness in the distribution we illustrate what happens if the

Algorithm 2: $(1-\alpha)$-Simultaneous CIs when the standard deviations are the same.

```
Data: $y_1, \cdots, y_n, \sigma$. Significance level $\alpha$.
Result: $(1-\alpha)$-simultaneous CIs for the ranks of $\mu_{1,T}, \cdots, \mu_{n,T}$.
if the hypothesis $\mu_1 = \cdots = \mu_n$ is not rejected then
    Set confidence intervals to $[1, n]$; $\forall i, a_i = 1, b_i = n$.
else
    for $l$ from 2 to $n-1$ do
        m = choose$(n-1, l-1)$ ;
        for $i$ from 1 to $m$ do
            Generate a hypothesis $H_{i,l}$ using $C^{-1}(i)$;
            Calculate the LR under $H_{i,l}$ ;
            if $LR(H_{i,l}, y_1, \cdots, y_n) \leq \chi^2_{1-\alpha}(n-l)$ then
                Update the ranks;
                for $\pi \in \Pi$ do
                    $(\tilde{y}_1, \cdots, \tilde{y}_n) = (y_{\pi(1)}, \cdots, y_{\pi(n)})$;
                    Do a PAVA-Check;
                    if $H_{i,l}$ is (partially) correctly ordered then
                        Calculate the LR under $H_{i,l}$ using $(\tilde{y}_1, \cdots, \tilde{y}_n)$;
                        if $LR(H_{i,l}, \tilde{y}_1, \cdots, \tilde{y}_n) \leq \chi^2_{1-\alpha}(n-l)$ then
                            Update the ranks;
                        end
                    end
                end
            end
        end
    end
end
```

true distribution of the data is the Gamma distribution with shape λ and scale $1/\sqrt{\lambda}$ with $\lambda \in \{1, 2, 5\}$. An observation y_i is generated using

$$y_i \sim \mu_{i,T} + \text{Gamma}(\lambda, 1/\sqrt{\lambda})$$

The expectation of y_i is $\mu_{i,T} + \sqrt{\lambda}$ and the standard deviation is 1. Since the ranking problem does not change if all the means are translated by the same fixed quantity, that is $\sqrt{\lambda}$, then we are in the same context of the paper as described in Section 2. Note that as the shape value increases, the Gamma distribution takes closer form to the Gaussian distribution. As in the simulations of the paper, we generate the means $\mu_{i,T}$'s independently from the Gaussian distribution $\mathcal{N}(0, \tau^2)$ for $\tau = 0.5, 1, 2$. For each value of τ, we generate 1000 n-samples of means $\mu_T = (\mu_{1,T}, \cdots, \mu_{n,T})$ for $n = 10$.

The results of Table (3) shows that deviations from the normality assumption has a small effect when ties are present. When there are no ties, the resulting simultaneous CIs are still conservative.

Algorithm 3: $(1-\alpha)$-Simultaneous CIs when the standard deviations are not the same.

```
Data: Sample $y_1, \cdots, y_n$. Standard deviations $\sigma_1, \cdots, \sigma_n$. A significance level $\alpha$.
Result: For each $i, [a_i, b_i]$ such that $\mu_{i,T} \in [a_i, b_i]$ with joint probability greater than
        $1-\alpha$.
if the hypothesis $\mu_1 = \cdots = \mu_n$ is not rejected then
    Set confidence intervals to $[1, n]$; $\forall i, a_i = 1, b_i = n$.
else
    for $\pi \in \Pi \cup S$ do
        $(\tilde{y}_1, \cdots, \tilde{y}_n) = (y_{\pi(1)}, \cdots, y_{\pi(n)})$;
        for $l$ from 2 to $n-1$ do
            m = choose($n-1, l-1$) ;
            for $i$ from 1 to $m$ do
                Generate a hypothesis $H_{i,l}$ using $C^{-1}(i)$;
                Do a PAVA-Check;;
                if $H_{i,l}$ is (partially) correctly ordered then
                    Calculate the LR under $H_{i,l}$ ;
                    if $LR(H_{i,l}, \tilde{y}_1, \cdots, \tilde{y}_n) \leq \chi^2_{1-\alpha}(n-l)$ then
                        Update the ranks;
                    end
                end
            end
        end
    end
end
```

Table 3

Coverage probability and efficiency for different values of τ. *Nominal simultaneous coverage is* 90%

		$n=5$		$n=10$	
		Coverage	Average length of CIs	Coverage	Average length of CIs
$\tau=0$, shape = 5	Tukey	0.898	0.785	0.897	0.810
	Rescaled Tukey	0.898	0.784	0.895	0.809
	LR	0.887	0.7822	0.929	0.813
	Rescaled LR	0.875	0.774	0.907	0.811
$\tau=0$, shape = 2	Tukey	0.878	0.780	0.886	0.807
	Rescaled Tukey	0.876	0.778	0.886	0.806
	LR	0.880	0.777	0.906	0.810
	Rescaled LR	0.865	0.769	0.891	0.808
$\tau=0$, shape = 1	Tukey	0.892	0.777	0.877	0.803
	Rescaled Tukey	0.891	0.776	0.875	0.802
	LR	0.879	0.773	0.887	0.807
	Rescaled LR	0.869	0.767	0.865	0.805
$\tau=1$, shape = 5	Tukey	0.979	0.735	0.988	0.864
	Rescaled Tukey	0.977	0.727	0.978	0.859
	LR	0.974	0.723	0.990	0.873
	Rescaled LR	0.954	0.697	0.978	0.863
$\tau=1$, shape = 2	Tukey	0.976	0.733	0.969	0.863
	Rescaled Tukey	0.971	0.726	0.963	0.859
	LR	0.968	0.719	0.971	0.873
	Rescaled LR	0.950	0.697	0.960	0.863
$\tau=1$, shape = 1	Tukey	0.962	0.737	0.941	0.863
	Rescaled Tukey	0.959	0.732	0.938	0.859
	LR	0.950	0.724	0.952	0.871
	Rescaled LR	0.931	0.702	0.936	0.863

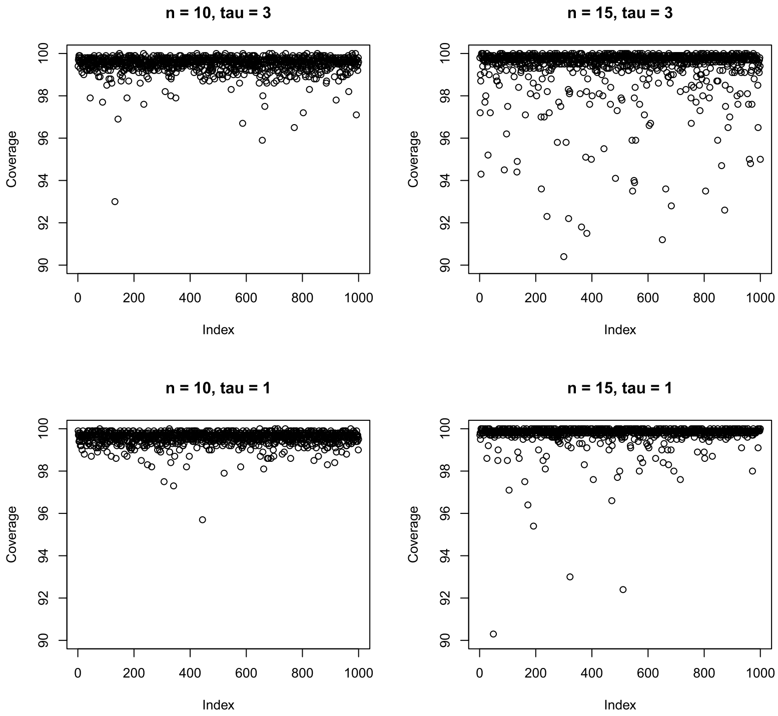

Figure 2: Simulation for the coverage of the partitioning procedure that uses the LR as a local test when the standard deviations are not the same using vectors of means and standard deviations randomly selected.

D.3. Example with more ties

Here is an example of 3 groups of 3 means (so that $n = 9$) and also 2 groups of 4 means ($n = 8$). We follow the same setup as in Section 7, but we only use $\tau = 1$ (recall that we generate the true means from $\mathcal{N}(0, \tau)$). The results of Table (4) show conservative confidence intervals but less than when there are no ties.

TABLE 4
Coverage probability and efficiency when ties are present. Nominal simultaneous coverage is 90%

	Three groups $n = 9$		Two groups $n = 8$	
	Coverage	Average length of CIs	Coverage	Average length of CIs
Tukey	0.964	0.861	0.950	0.853
Rescaled Tukey	0.947	0.857	0.936	0.849
LR	0.976	0.865	0.963	0.853
Rescaled LR	0.952	0.857	0.947	0.844

D.4. Example with estimated standard errors

In this example, we consider 2 groups of 4 means ($n = 8$). We follow the same setup as in paragraph D.3, but we only use $\tau = 1$ (recall that we generate the true means from $\mathcal{N}(0, \tau)$). For each true mean, we generate m observations randomly from the Gaussian distribution $\mathcal{N}(\mu_{i,T}, \sqrt{m})$. Then, we calculate the sample means and sample standard errors. Note that the true standard error is 1 in order to get comparable results to paragraph D.3. The results of Table (5) show very close results to when we used the true standard error for $m = 30$. For $m = 3$, the simultaneous coverage goes slightly below the nominal level.

Table 5

Coverage probability and efficiency when ties are present. Nominal simultaneous coverage is 90%

	Three groups $m = 3$		Three groups $m = 5$		Two groups $m = 30$	
	Coverage	Average length of CIs	Coverage	Average length of CIs	Coverage	Average length of CIs
Tukey	0.880	0.837	0.908	0.844	0.939	0.847
Rescaled Tukey	0.868	0.832	0.902	0.840	0.932	0.843
LR	0.898	0.836	0.928	0.845	0.955	0.848
Rescaled LR	0.857	0.824	0.902	0.834	0.936	0.838

Appendix E: Data for hotels ratings

We collected the following dataset from the website of Booking.com for a room reservation in the city of Leiden (The Netherlands) to rent a room for one night on the 2^{nd} of May 2019. The query was made on the 15^{th} of April 2019. We restricted our search for hotels with free Wifi, free cancellation and within 1 Km from the city center.

Table 6

Dataset of hotels ratings.

Hotel name	Rating	Total number of reviews	1 Star	2 Stars	3 Stars	4 Stars	5 Stars
Ibis Leiden Centeral	8.2	2706	4	31	336	1523	812
Boutique Hotel d'Oude Mors	9.3	1033	0	1	18	253	761
Tulip Inn Leiden	8.4	1308	3	17	103	712	474
Golden Tulip Leiden	8.4	1551	5	14	131	778	619
City Hotel Rembrandt	8.1	1499	6	33	202	807	451
Hotel Mayflower	7.5	1027	8	55	224	552	186
Best Western City Hotel Leiden	7.7	2146	10	89	422	1216	405
City Resort Hotel Leiden	7.8	1695	14	55	310	861	455
Boutique Hotel Steenhof Suites	9.5	473	0	0	9	68	398

References

[1] Diaa Al Mohamad, Jelle J. Goeman, and Erik W. van Zwet. Simultaneous confidence intervals for ranks with application to ranking institutions. *Biometrics*, 2020. Accepted.

[2] Diaa Al Mohamad, Erik W. Van Zwet, Eric Cator, and Jelle J. Goeman. Adaptive critical value for constrained likelihood ratio testing. *Biometrika*, 107(3):677–688, 05 2020. MR4138983

[3] Russell G. Almond, Charles Lewis, John W. Tukey, and Duanli Yan. Displays for comparing a given state to many others. *The American Statistician*, 54(2):89–93, 2000.

[4] Tom M. Apostol and Mamikon A. Mnatsakanian. Sums of squares of distances in m-space. *The American Mathematical Monthly*, 110(6):516–526, 2003. MR1984403

[5] Miriam Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and Edward Silverman. An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, 26(4):641–647, 12 1955. MR0073895

[6] R.E. Barlow. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. J. Wiley, 1972.

[7] Andres F. Barrientos, Deborshee Sen, Garritt L Page, and David B Dunson. Bayesian inferences on uncertain ranks and orderings, 2019.

[8] D. J. Bartholomew. A test of homogeneity for ordered alternatives. ii. *Biometrika*, 46(3/4):328–335, 1959. MR0112204

[9] Robert E. Bechhofer. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Ann. Math. Statist.*, 25(1):16–39, 03 1954. MR0060197

[10] Ting Bie. Confidence intervals for ranks: Theory and applications in binomial data. Master's thesis, Uppsala University, Sweden, 2013. Master thesis under the supervision of R. Larsson.

[11] Hans C. van Houwelingen and Ronald Brand. Empirical bayes methods for monitoring health care quality. In *Bulletin of the Institute of Statistics*, Finland, 1999.

[12] Violeta Calian, Dongmei Li, and Jason C. Hsu. Partitioning to uncover conditions for permutation tests to control multiple testing error rates. *Biometrical Journal*, 50(5):756–766, 2008. MR2542341

[13] Jan de Leeuw, Kurt Hornik, and Patrick Mair. Isotone optimization in R: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009.

[14] Chris Feudtner, Jay G. Berry, Gareth Parry, Paul Hain, Rustin B. Morse, Anthony D. Slonim, Samir S. Shah, and Matt Hall. Statistical uncertainty of mortality rates and rankings for children's hospitals. *Pediatrics*, 128(4):e966–e972, 2011.

[15] H. Finner and K. Strassburger. The partitioning principle: a powerful tool in multiple decision theory. *Ann. Statist.*, 30(4):1194–1213, 08 2002. MR1926174

[16] Helmut Finner and Klaus Strassburger. Step-up related simultaneous confidence intervals for mcc and mcb. *Biometrical Journal*, 49(1):40–51, 2007. MR2339215

[17] Robert B. Gerzoff and G. David Williamson. Who's number one? the impact of variability on rankings based on public health indicators. *Public Health Reports (1974-)*, 116(2):158–164, 2001.

[18] Jelle J. Goeman and Aldo Solari. The sequential rejection principle of familywise error control. *Ann. Statist.*, 38(6):3782–3810, 12 2010. MR2766868

[19] Harvey Goldstein and David J. Spiegelhalter. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 159(3):385–443, 1996.

[20] Ulrike Grömping. Inference with linear equality and inequality constraints using R: The package ic.infer. *Journal of Statistical Software*, 33(10):1–31, 2010.

[21] Nicholas C. Henderson and Michael A. Newton. Making the cut: improved ranking and selection for large-scale inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):781–804, 2016. MR3534350

[22] Y. Hochberg and A. C. Tamhane. *Multiple comparison procedures*. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1987. MR0914493

[23] S. Holm. Confidence intervals for ranks. *Department of Mathematical Statistics*, 2012. Unpublished manucript.

[24] Geoffrey J. Iverson and Steven Alex Harp. A conditional likelihood ratio test for order restrictions in exponential families. *Mathematical Social Sciences*, 14(2):141 – 159, 1987. MR0917841

[25] M. Klein, T. Wright, and J. Wieczorek. A simple joint confidence region for a ranking of k populations. *Research Report Series, U.S. Bureau of the Census*, pages 1–18, 2018. MR4098963

[26] Donald E. Knuth. *The Art of Computer Programming, Volume 4, Fascicle 3: Generating All Combinations and Partitions*. Addison-Wesley Professional, 2005. MR2251472

[27] Nan M. Laird and Thomas A. Louis. Empirical bayes ranking methods. *Journal of Educational Statistics*, 14(1):29–46, 1989. MR3983320

[28] E. L. Lehmann and Joseph P. Romano. *Testing statistical hypotheses*. Springer Texts in Statistics. Springer, New York, third edition, 2005. MR2135927

[29] Oscar Lemmers, Jan A. M. Kremer, and George F. Borm. Incorporating natural variation into IVF clinic league tables. *Human Reproduction*, 22(5):1359–1362, 2007.

[30] Oscar Lemmers, Mireille Broeders, André Verbeek, Gerard Den Heeten, Roland Holland, and George F Borm. League tables of breast cancer screening units: Worst-case and best-case scenario ratings helped in exposing real differences between performance ratings. *Journal of Medical Screening*, 16(2):67–72, 2009. PMID: 19564518.

[31] Rongheng Lin, Thomas A. Louis, Susan M. Paddock, and Greg Ridgeway. Loss function based ranking in two-stage, hierarchical models. *Bayesian Anal.*, 1(4):915–946, 12 2006. MR2282211

[32] Rongheng Lin, Thomas A. Louis, Susan M. Paddock, and Greg Ridgeway. Ranking usrds provider specific smrs from 1998–2001. *Health Services and Outcomes Research Methodology*, 9(1):22–38, 2009.

[33] Hester F. Lingsma, Marinus JC Eijkemans, and Ewout W. Steyerberg. Incorporating natural variation into ivf clinic league tables: The expected rank. *BMC Medical Research Methodology*, 9(1):53, 2009.

[34] E. Clare Marshall and David J. Spiegelhalter. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ: British Medical Journal*, 316:1701–1705, 1998.

[35] R. E. Miles. The complete amalgamation into blocks, by weighted means, of a finite set of real numbers. *Biometrika*, 46(3/4):317–327, 1959. MR0112167

[36] Hisashi Noma, Shigeyuki Matsui, Takashi Omori, and Tosiya Sato. Bayesian ranking and selection methods using hierarchical mixture models in microarray studies. *Biostatistics*, 11(2):281, 2010.

[37] B. C. Rennie and A. J. Dobson. On stirling numbers of the second kind. *Journal of Combinatorial Theory*, 7(2):116 – 121, 1969. MR0241310

[38] Tim Robertson and Edward J. Wegman. Likelihood ratio tests for order restrictions in exponential families. *Ann. Statist.*, 6(3):485–505, 05 1978. MR0471147

[39] M. J. Silvapulle and P. K. Sen. *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. Wiley Series in Probability and Statistics. Wiley, 2004. MR2099529

[40] David J. Spiegelhalter. Funnel plots for comparing institutional performance. *Statistics in Medicine*, 24(8):1185–1202, 2005. MR2134573

[41] Richard P. Stanley. *Enumerative Combinatorics: Volume 1*. Cambridge University Press, New York, NY, USA, 2nd edition, 2011. MR2868112

[42] G. Stefansson, W. Kim, and J. C. Hasu. On confidence sets in multiple comparisons. *Statistical Decision Theory and Related Topics IV(S. S. Gupta and J. O. Berger, eds.)*, 2:89–104, 1988. MR0927125

[43] Paris P. Tekkis, Peter McCulloch, Adrian C. Steger, Irving S. Benjamin, and Jan D. Poloniecki. Mortality control charts for comparing performance of surgical units: validation study using hospital mortality data. *BMJ*, 326(7393):786, 2003.

[44] J. W. Tukey. The problem of multiple comparisons. *The Collected Works of John W. Tukey VIII. Multiple Comparisons: 1948–1983*, pages 1–300, 1953. Unpublished manuscript. MR1263027

[45] C. van Eeden. *Testing and estimating ordered parameters of probability distributions*. PhD thesis, University of Amsterdam, 1958. MR0102874

[46] Peter C. Wollan and Richard L. Dykstra. Conditional tests with an order restriction as a null hypothesis. In Richard Dykstra, Tim Robertson, and Farroll T. Wright, editors, *Advances in Order Restricted Statistical Inference*, pages 279–295, New York, NY, 1986. Springer New York. MR0875659

[47] Shunpu Zhang, Jun Luo, Li Zhu, David G. Stinchcomb, Dave Campbell, Ginger Carter, Scott Gilkeson, and Eric J. Feuer. Confidence intervals for ranks of age-adjusted rates across states or counties. *Statistics in Medicine*, 33(11):1853–1866, 2014. MR3256907