Benchmarking report for Test challenge

created by challengeR v0.3.3
Wiesenfarth, Reinke, Landman, Cardoso, Maier-Hein & Kopp-Schneider (2019)

29 June, 2020

This document presents a systematic report on a benchmark study. Input data comprises raw metric values for all algorithms and test cases. Generated plots are:

- Visualization of assessment data: Dot- and boxplots, podium plots and ranking heatmaps
- Visualization of ranking stability: Blob plots, violin plots and significance maps
- Visualization of ranking robustness: Line plots
- Visualization of cross-task insights

Ranking of algorithms within tasks according to the following chosen ranking scheme:

aggregate using function ("mean") then rank

Ranking list for each task:

c_ideal : Analysis based on 50 test cases which included 0 missing values.

	value_FUN	rank
A1	0.9516761	1
A2	0.8449666	2
A3	0.7416888	3
A4	0.6447502	4
A5	0.5423610	5

c_random : Analysis based on 50 test cases which included 0 missing values.

	value_FUN	rank
A1	0.8046800	1
A2	0.7975213	2
A4	0.7884067	3
A5	0.7697414	4
A3	0.7614808	5

c_worstcase : Analysis based on 120 test cases which included 0 missing values.

	value_FUN	rank
A1	0.9	1
A2	0.9	1
A3	0.9	1

	value_FUN	rank
A4	0.9	1
A5	0.9	1

Consensus ranking according to chosen method euclidean:

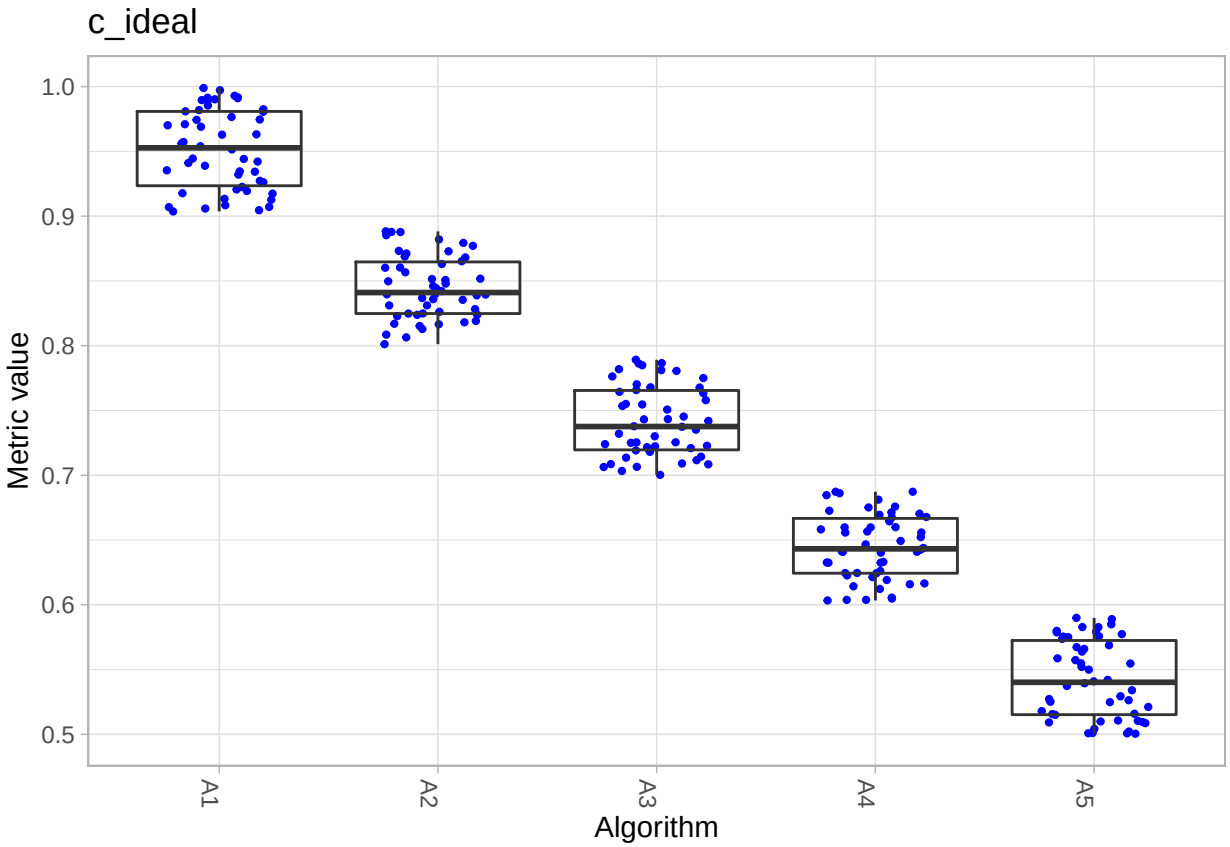
	value	rank
A1	1.667	1
A2	2.333	2
A4	3.333	3
A3	3.667	4
A5	4.000	5

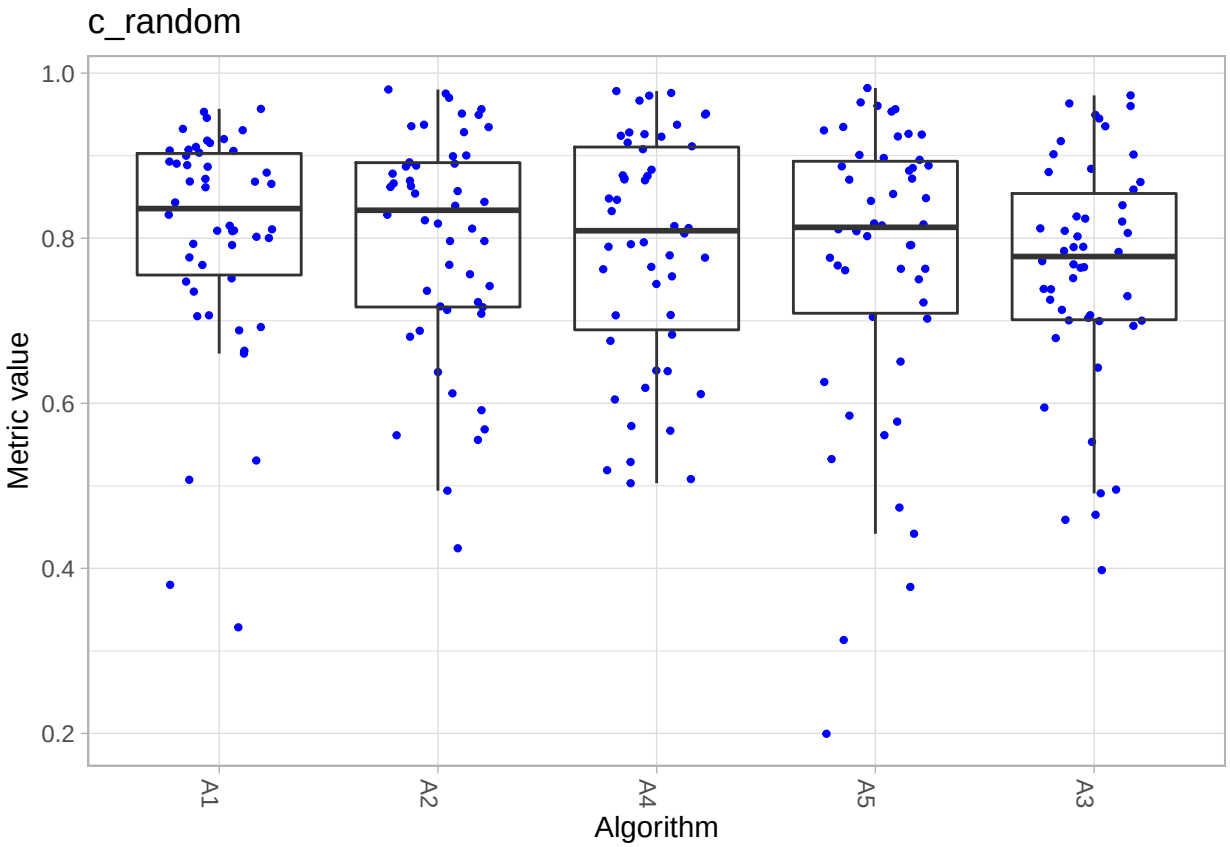
1 Visualization of raw assessment data

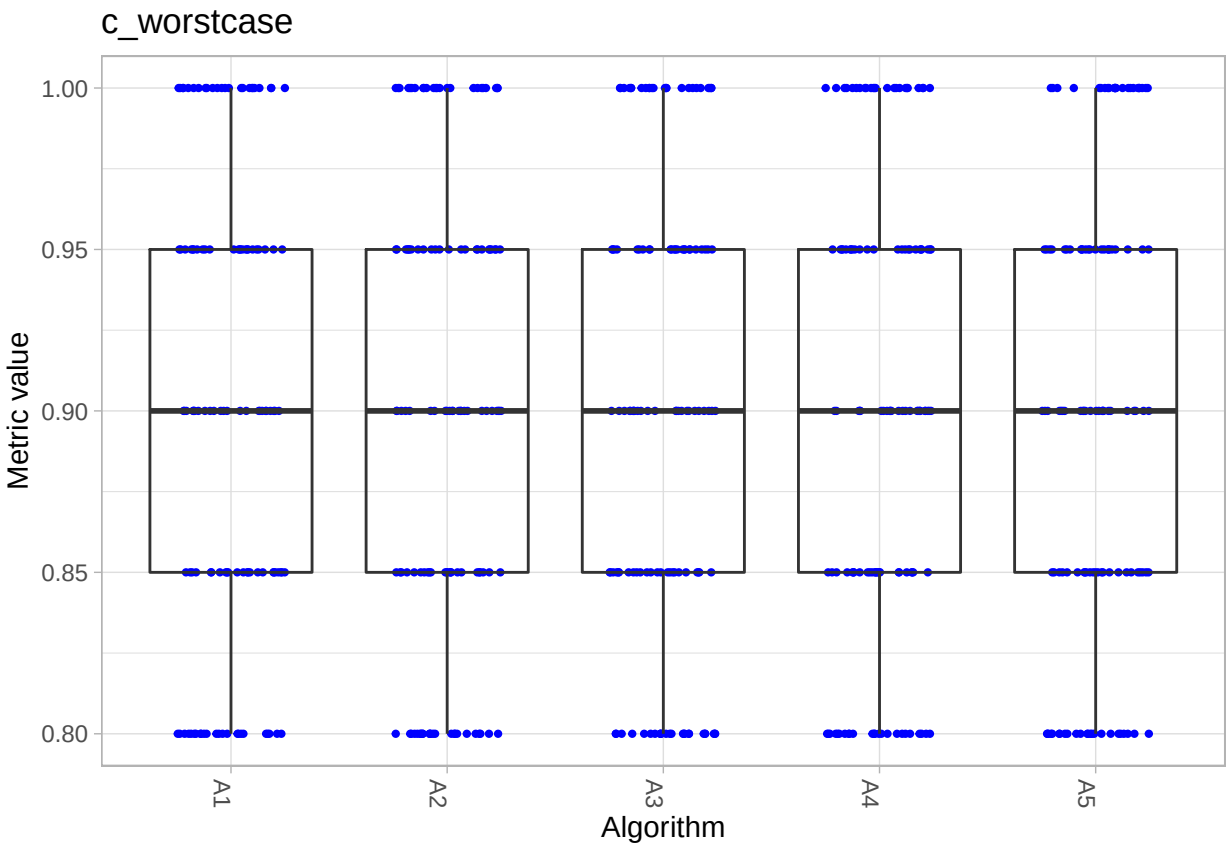
Algorithms are ordered according to chosen ranking scheme for each task.

1.1 Dot- and boxplots

Dot- and boxplots for visualizing raw assessment data separately for each algorithm. Boxplots representing descriptive statistics over all test cases (median, quartiles and outliers) are combined with horizontally jittered dots representing individual test cases.

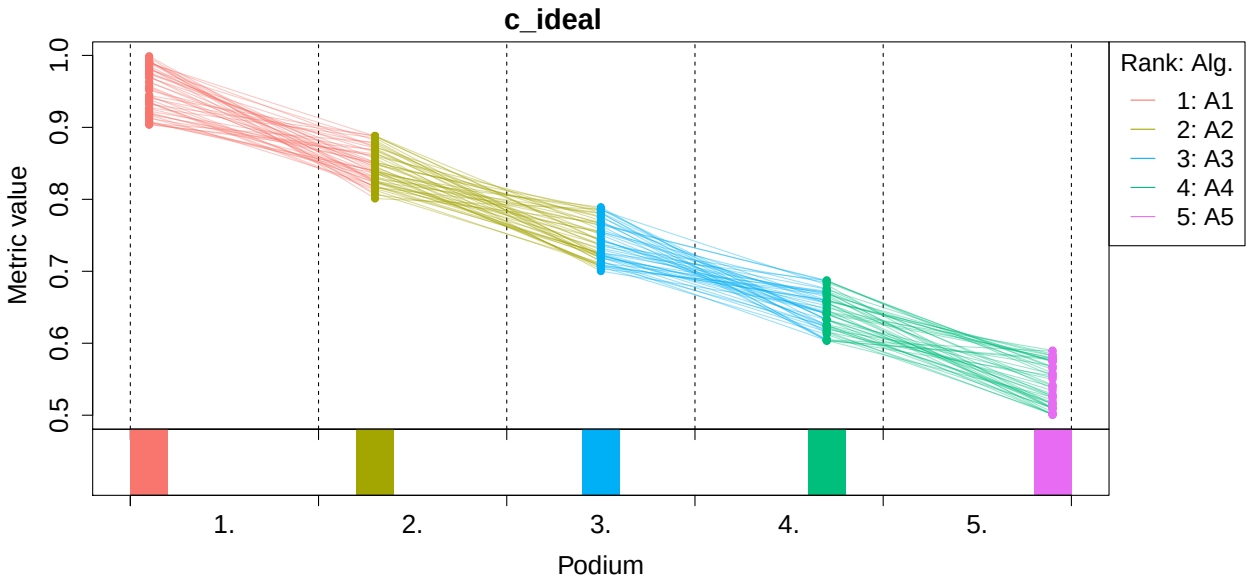

c_ideal
1.0
0.9
0.8
0.7
0.6
0.5
Metric value
A1
A2
A3
A4
A5
Algorithm

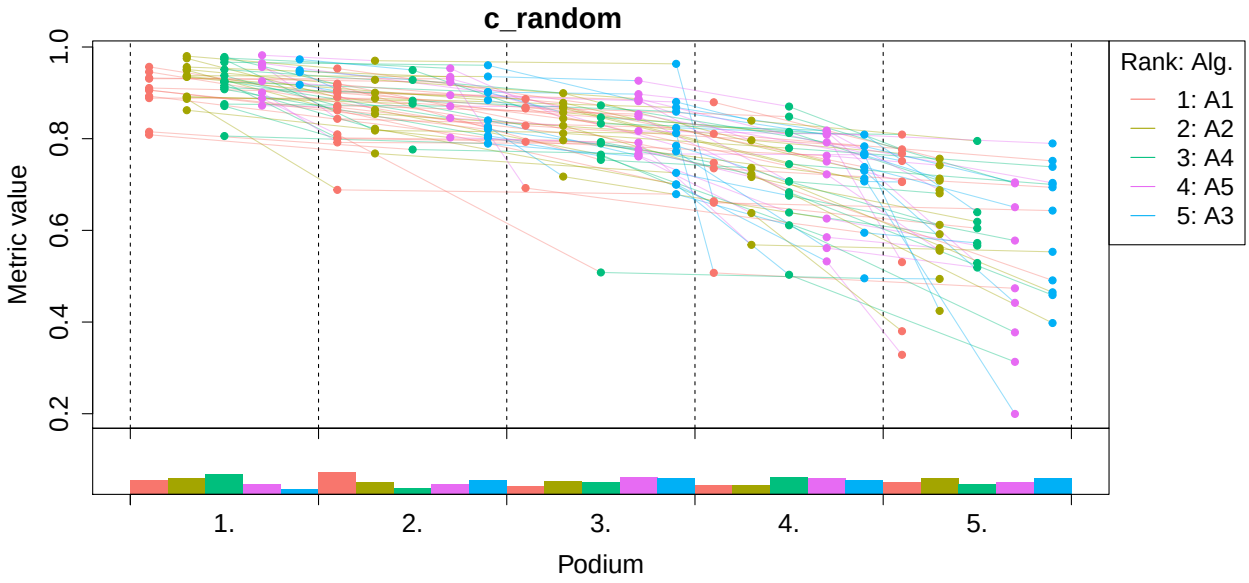

c_random
1.0
0.8
0.6
0.4
0.2
Metric value
A1
A2
A4
A5
A3
Algorithm

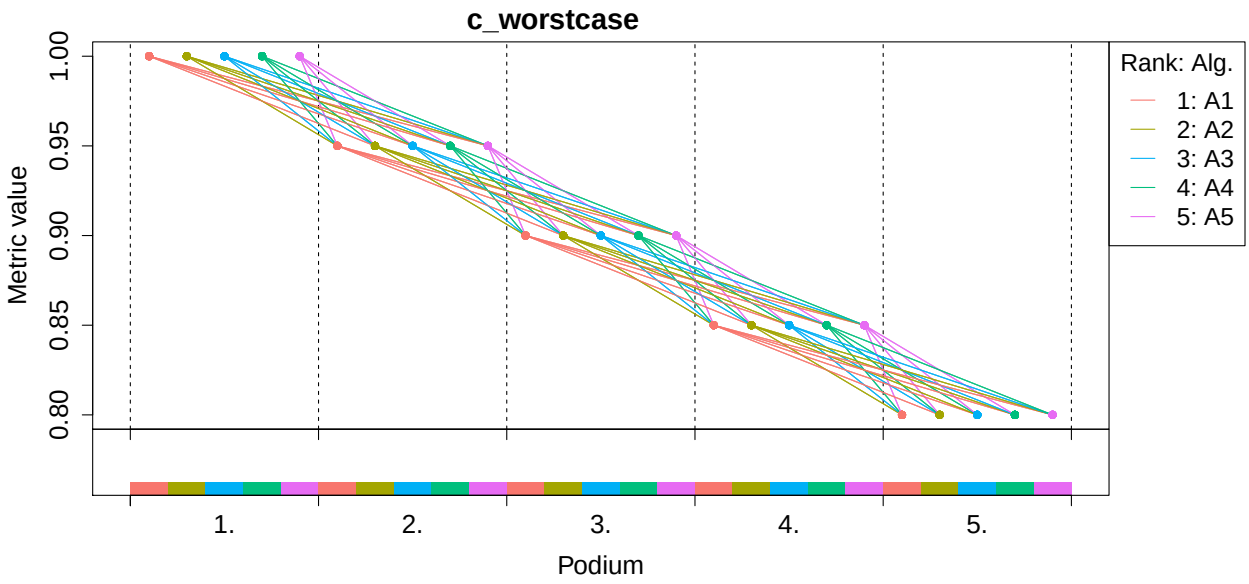


1.2 Podium plots

Podium plots (see also Eugster et al, 2008) for visualizing raw assessment data. Upper part (spaghetti plot): Participating algorithms are color-coded, and each colored dot in the plot represents a metric value achieved with the respective algorithm. The actual metric value is encoded by the y-axis. Each podium (here: p=5) represents one possible rank, ordered from best (1) to last (here: 5). The assignment of metric values (i.e. colored dots) to one of the podiums is based on the rank that the respective algorithm achieved on the corresponding test case. Note that the plot part above each podium place is further subdivided into p "columns", where each column represents one participating algorithm (here: $p = 5$). Dots corresponding to identical test cases are connected by a line, leading to the shown spaghetti structure. Lower part: Bar charts represent the relative frequency for each algorithm to achieve the rank encoded by the podium place.

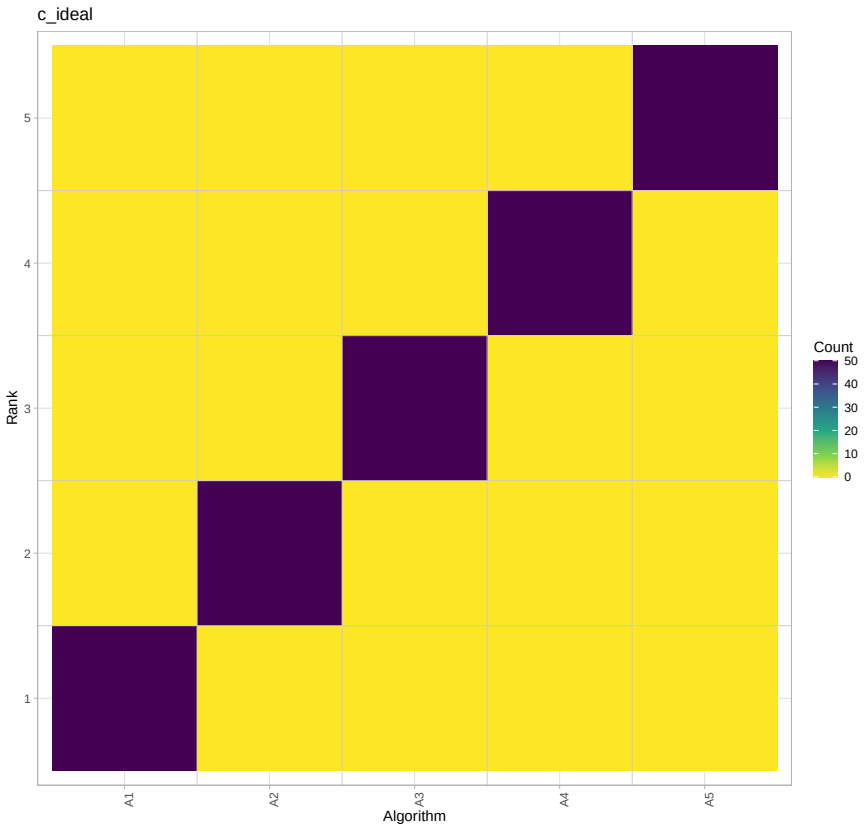

c_ideal
Metric value
1.0
0.9
0.8
0.7
0.6
0.5
1.
2.
3.
4.
5.
Podium
Rank: Alg.
1: A1
2: A2
3: A3
4: A4
5: A5

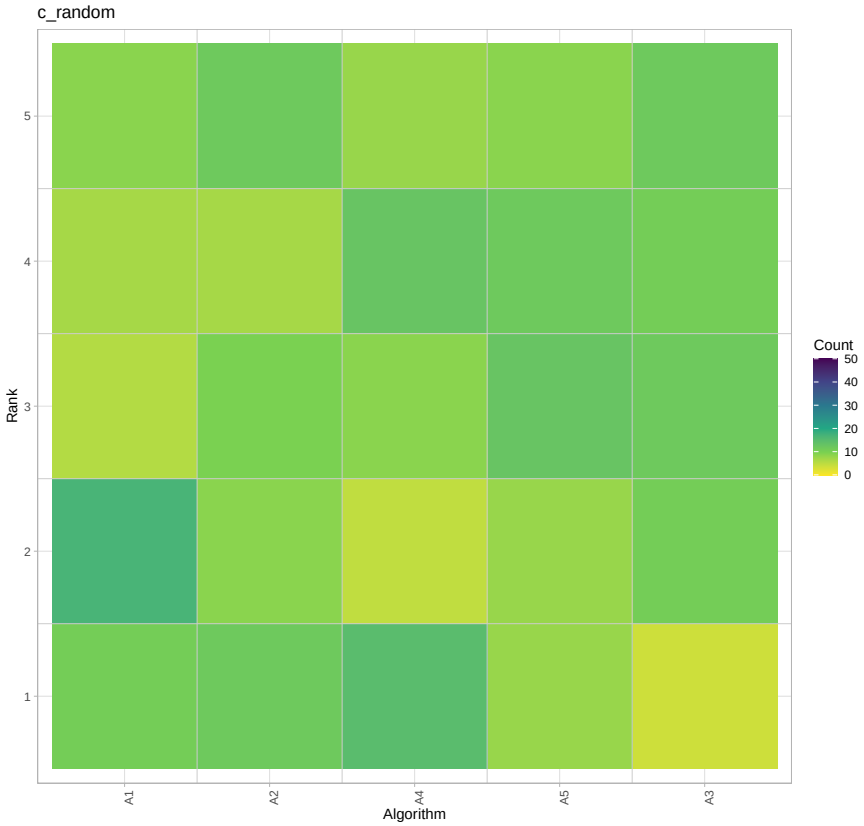

c_random
Metric value
1.0
0.8
0.6
0.4
0.2
1.
2.
3.
4.
5.
Podium
Rank: Alg.
1: A1
2: A2
3: A4
4: A5
5: A3

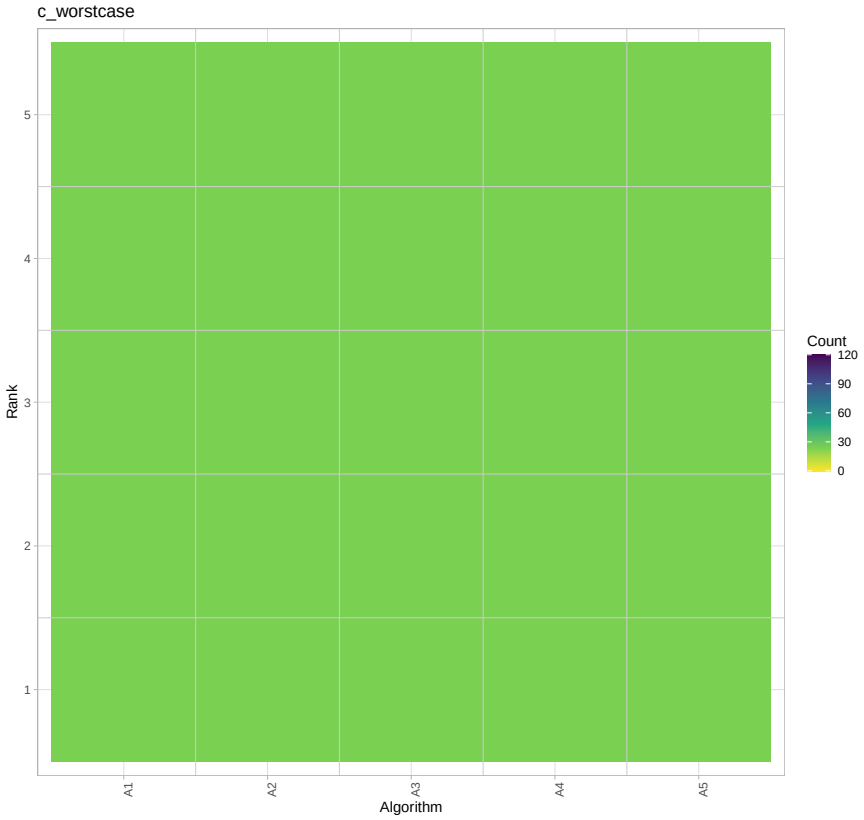


1.3 Ranking heatmaps

Ranking heatmaps for visualizing raw assessment data. Each cell (i, A_j) shows the absolute frequency of test cases in which algorithm A_j achieved rank i.


c_ideal
Rank
5
4
3
2
1
A1
A2
A3
A4
A5
Algorithm
Count
50
40
30
20
10
0

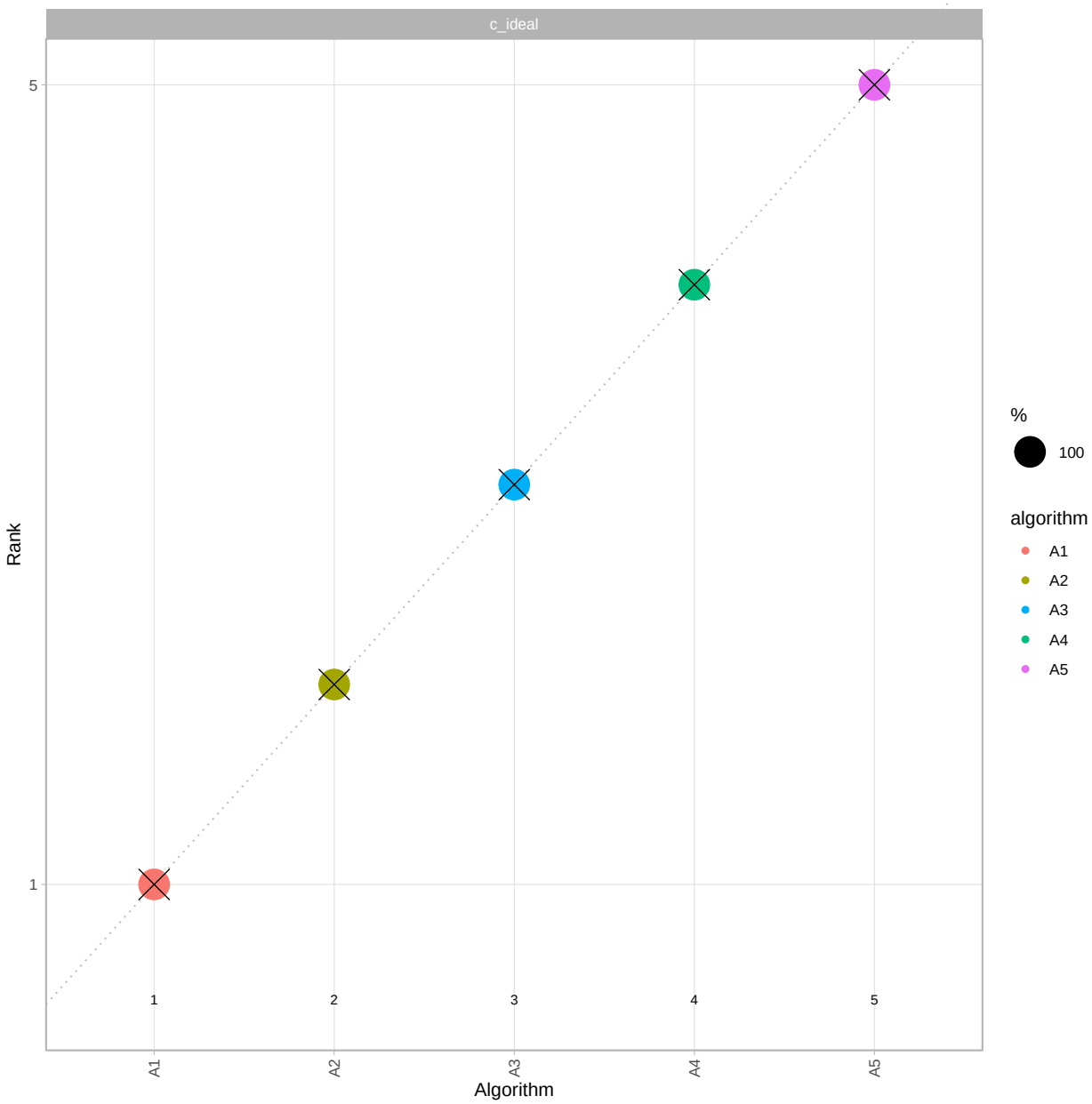

c_random
Rank
5
4
3
2
1
A1
A2
A4
A5
A3
Algorithm
Count
50
40
30
20
10
0

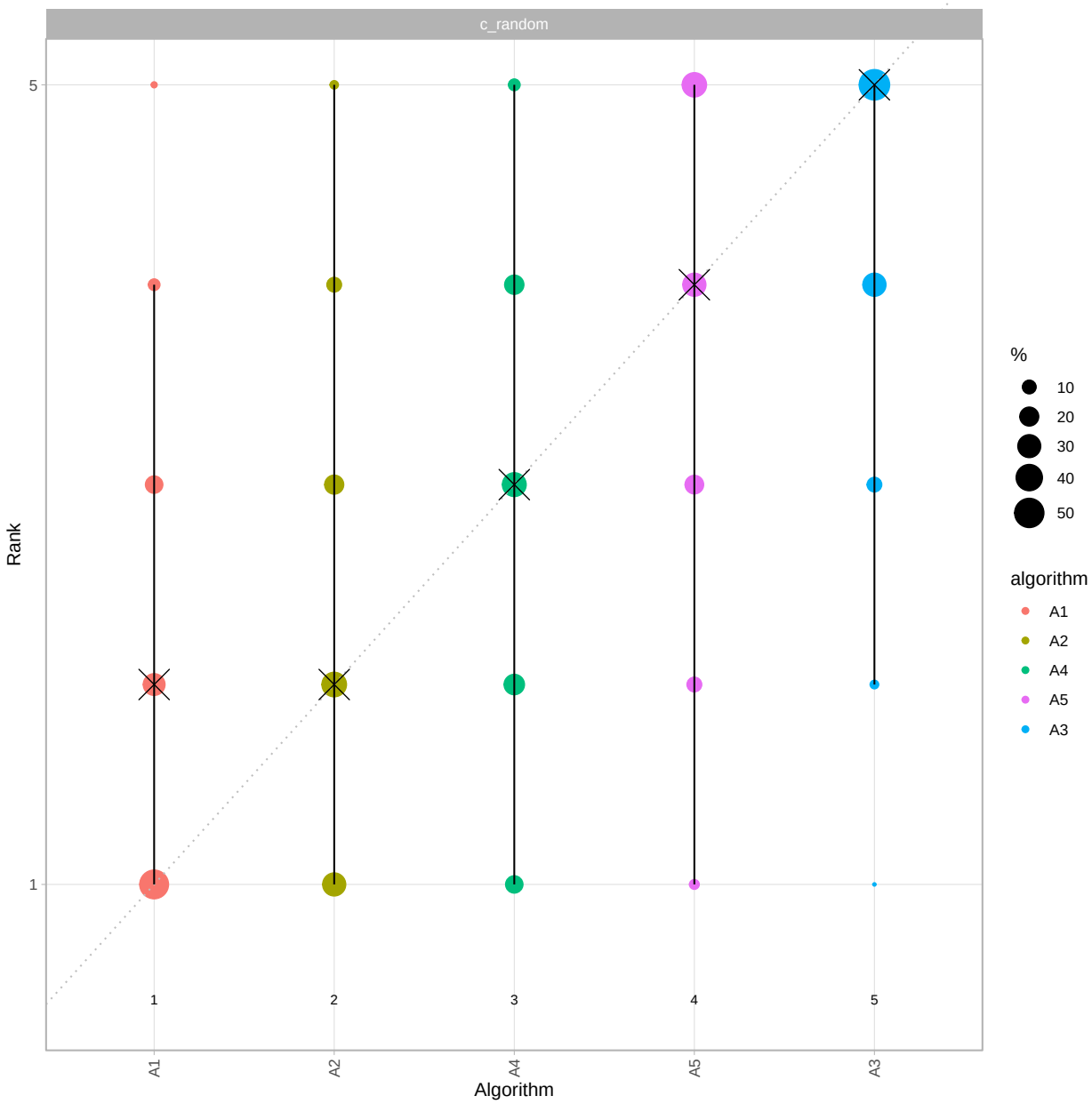


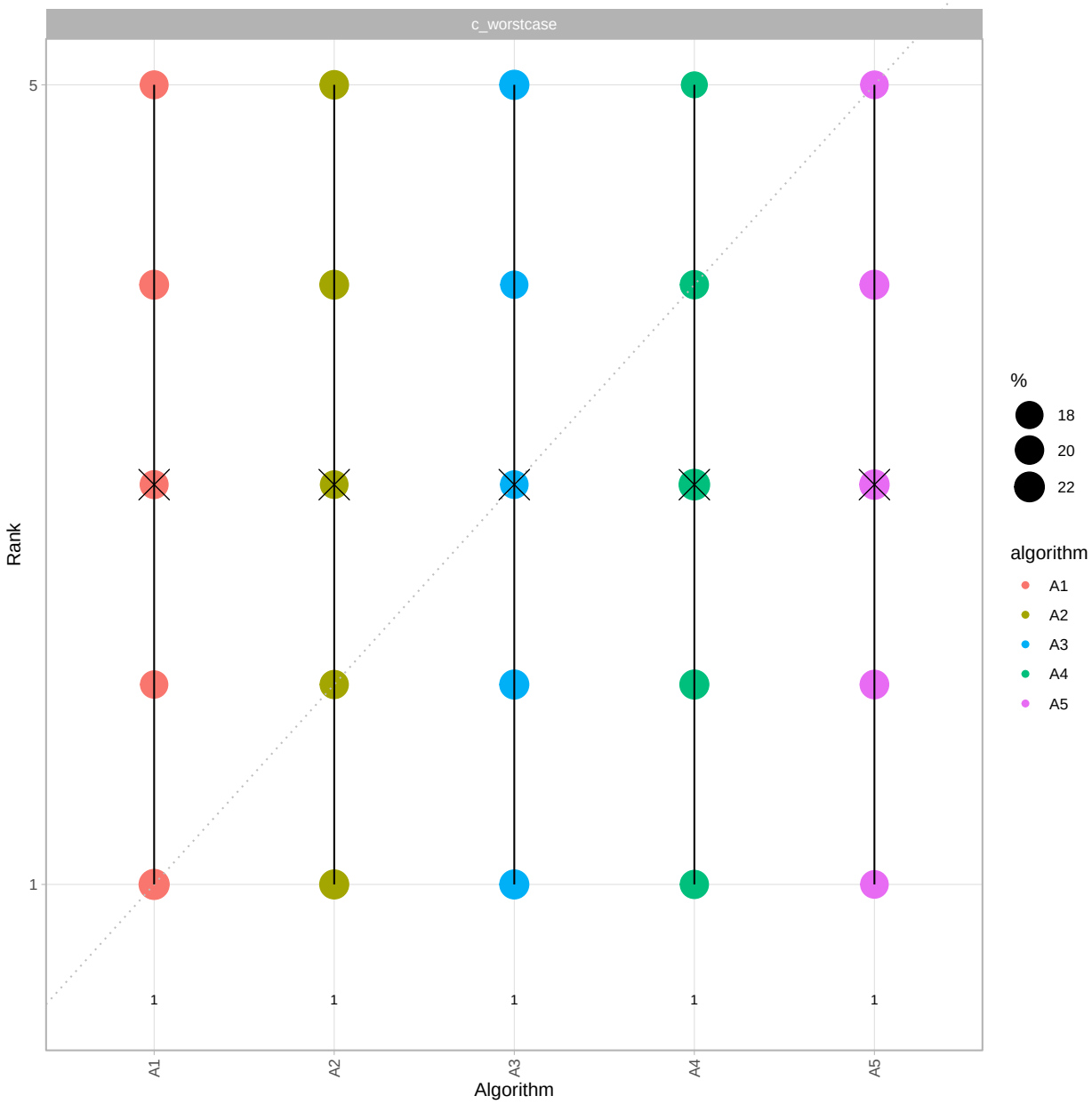
2 Visualization of ranking stability

2.1 *Blob plot* for visualizing ranking stability based on bootstrap sampling

Algorithms are color-coded, and the area of each blob at position $(A_i, \text{rank } j)$ is proportional to the relative frequency A_i achieved rank j across $b = 1000$ bootstrap samples. The median rank for each algorithm is indicated by a black cross. 95% bootstrap intervals across bootstrap samples are indicated by black lines.


c_ideal
5
1
Rank
1
2
3
4
5
A1
A2
A3
A4
A5
Algorithm
%
100
algorithm
A1
A2
A3
A4
A5


c_random
Rank
5
1
1
2
3
4
5
A1
A2
A4
A5
A3
Algorithm
%
10
20
30
40
50
algorithm
A1
A2
A4
A5
A3



2.2 *Violin plot* for visualizing ranking stability based on bootstrapping

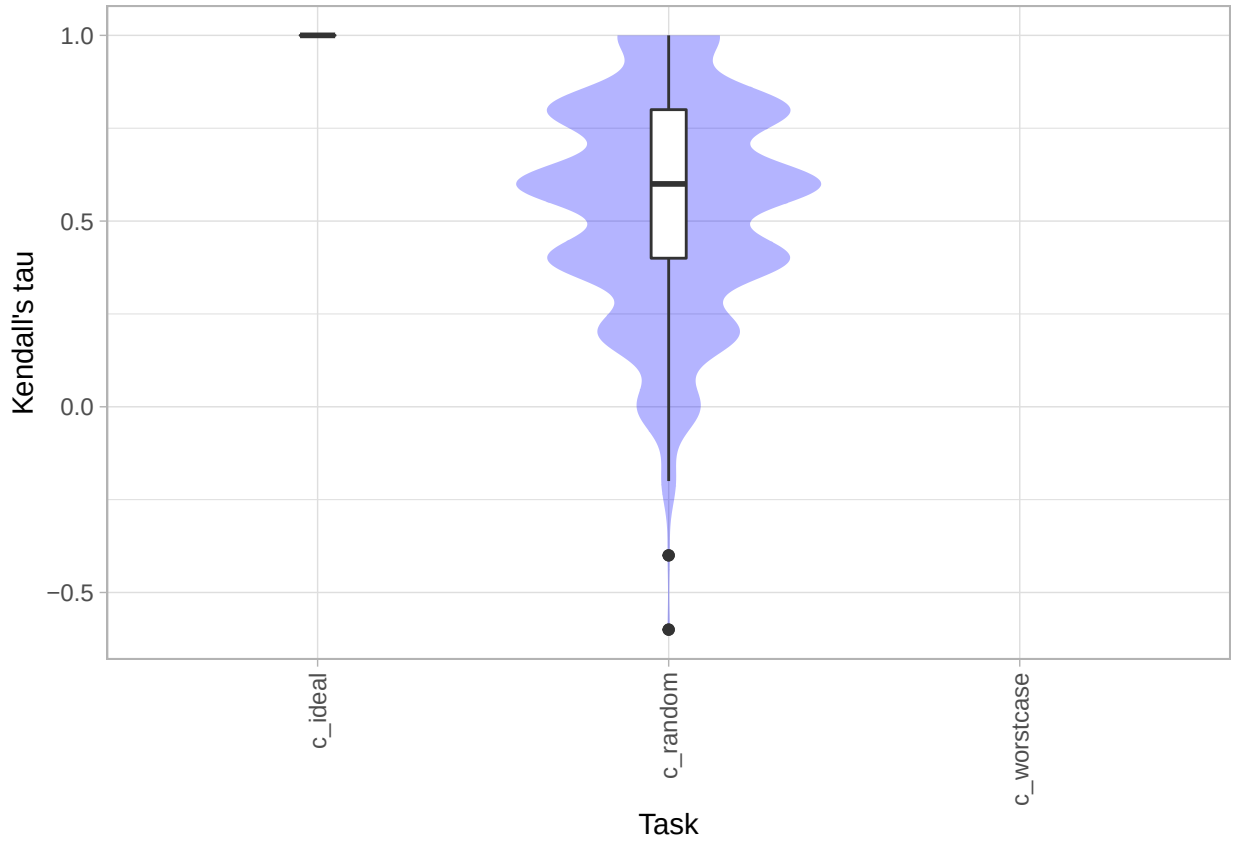
The ranking list based on the full assessment data is pairwisely compared with the ranking lists based on the individual bootstrap samples (here $b = 1000$ samples). For each pair of rankings, Kendall's τ correlation is computed. Kendall's τ is a scaled index determining the correlation between the lists. It is computed by evaluating the number of pairwise concordances and discordances between ranking lists and produces values between -1 (for inverted order) and 1 (for identical order). A violin plot, which simultaneously depicts a boxplot and a density plot, is generated from the results.

```
##
##
## Summary Kendall's tau
```

```
##          Task   mean median q25 q75
## 1     c_ideal 1.0000    1.0 1.0 1.0
## 2    c_random 0.5368    0.6 0.4 0.8
## 3 c_worstcase    NaN     NA  NA  NA
```

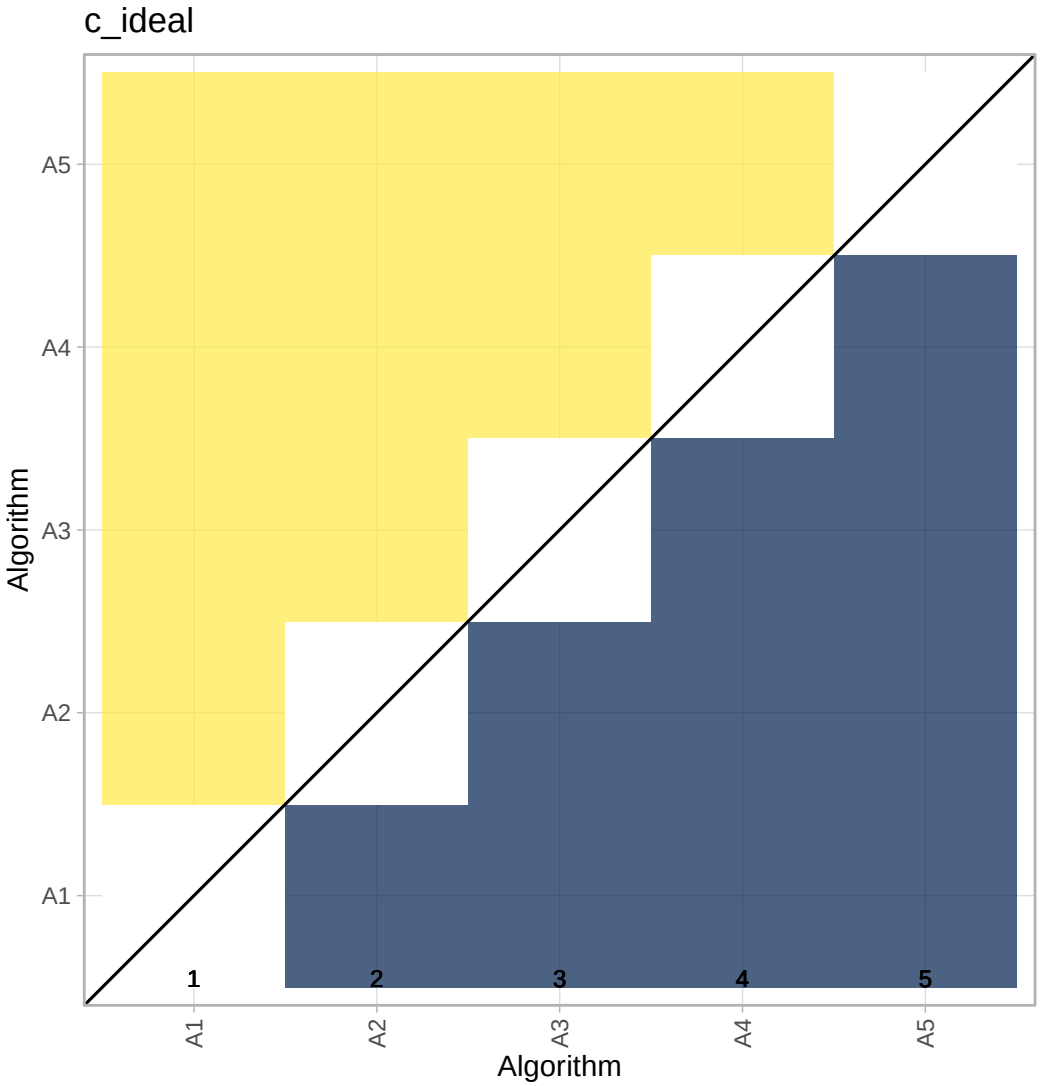
```
## Warning: Removed 1000 rows containing non-finite values (stat_ydensity).
```

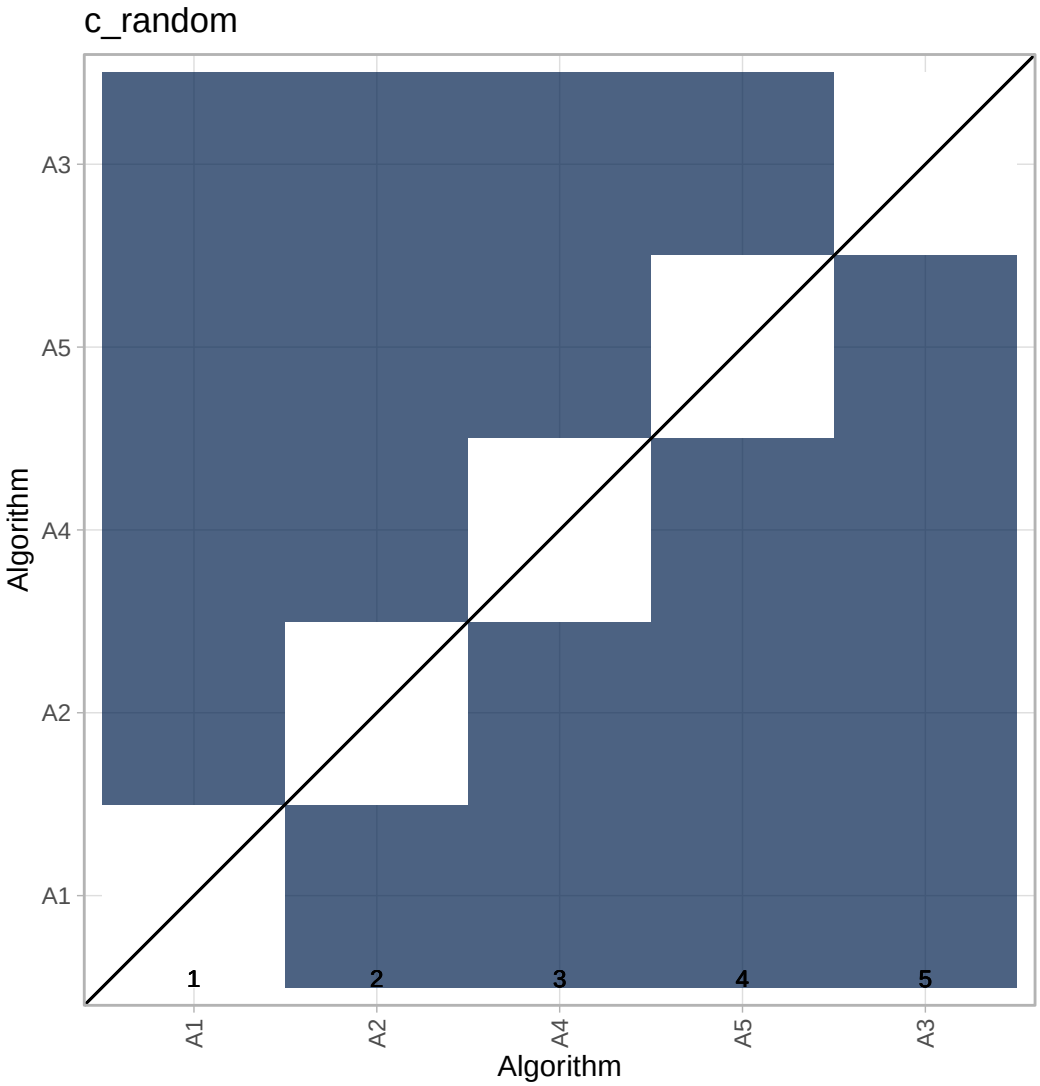
```
## Warning: Removed 1000 rows containing non-finite values (stat_boxplot).
```

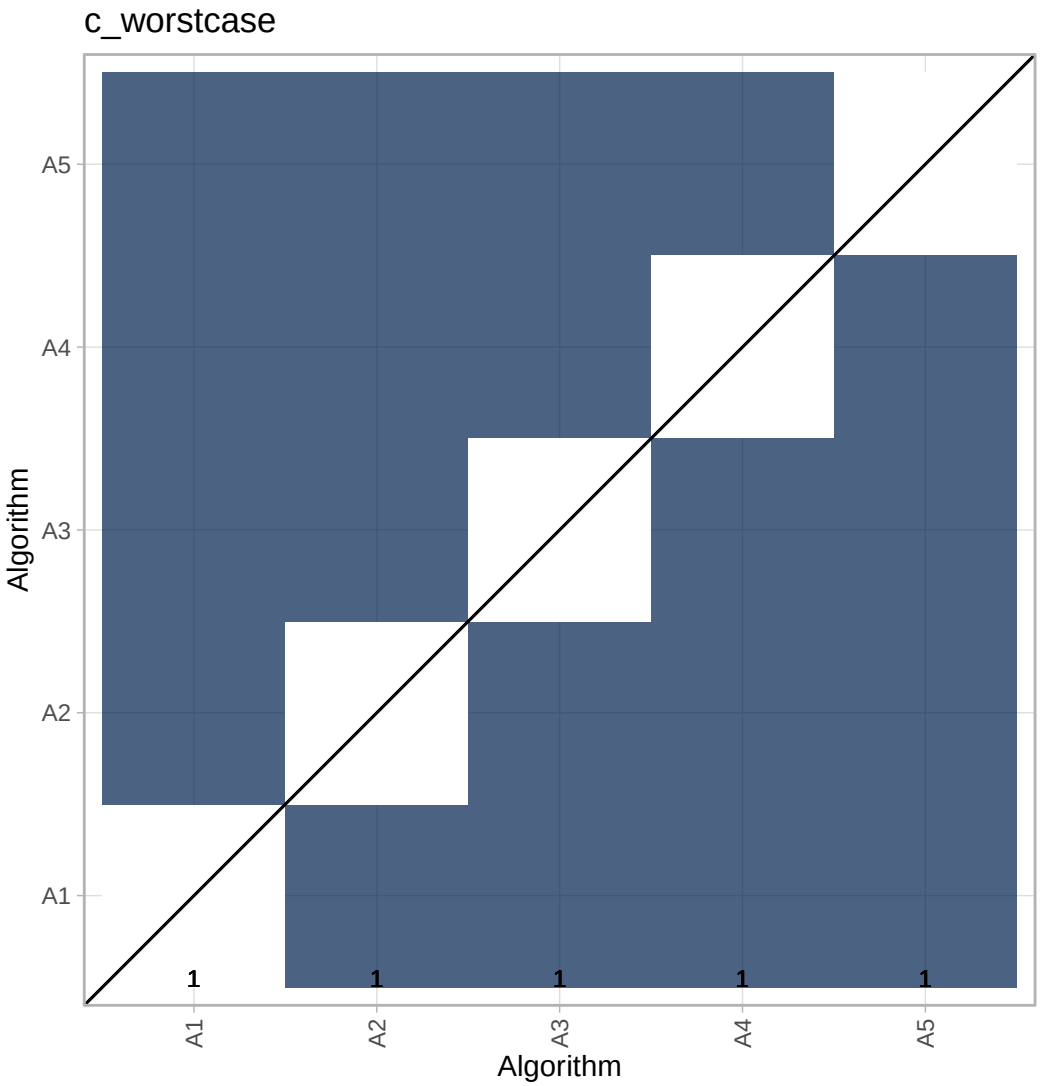


2.3 *Significance maps* for visualizing ranking stability based on statistical significance

Significance maps depict incidence matrices of pairwise significant test results for the one-sided Wilcoxon signed rank test at a 5% significance level with adjustment for multiple testing according to Holm. Yellow shading indicates that metric values of the algorithm on the x-axis were significantly superior to those from the algorithm on the y-axis, blue color indicates no significant difference.

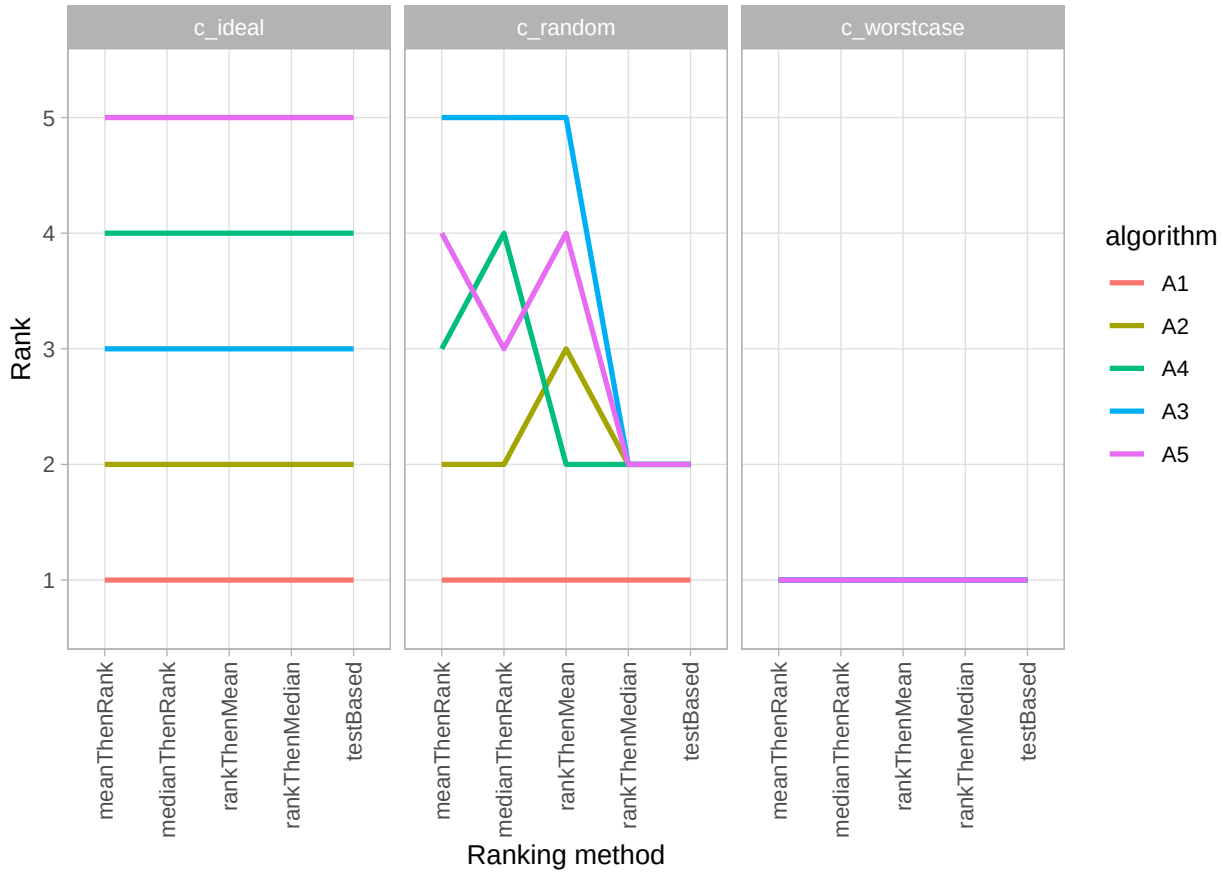

c_ideal
A5
A4
A3
A2
A1
Algorithm
1
2
3
4
5
A1
A2
A3
A4
A5
Algorithm


c_random
A3
A5
A4
A2
A1
Algorithm
1
2
3
4
5
A1
A2
A4
A5
A3
Algorithm



2.4 Ranking robustness to ranking methods

Line plots for visualizing rankings robustness across different ranking methods. Each algorithm is represented by one colored line. For each ranking method encoded on the x-axis, the height of the line represents the corresponding rank. Horizontal lines indicate identical ranks for all methods.



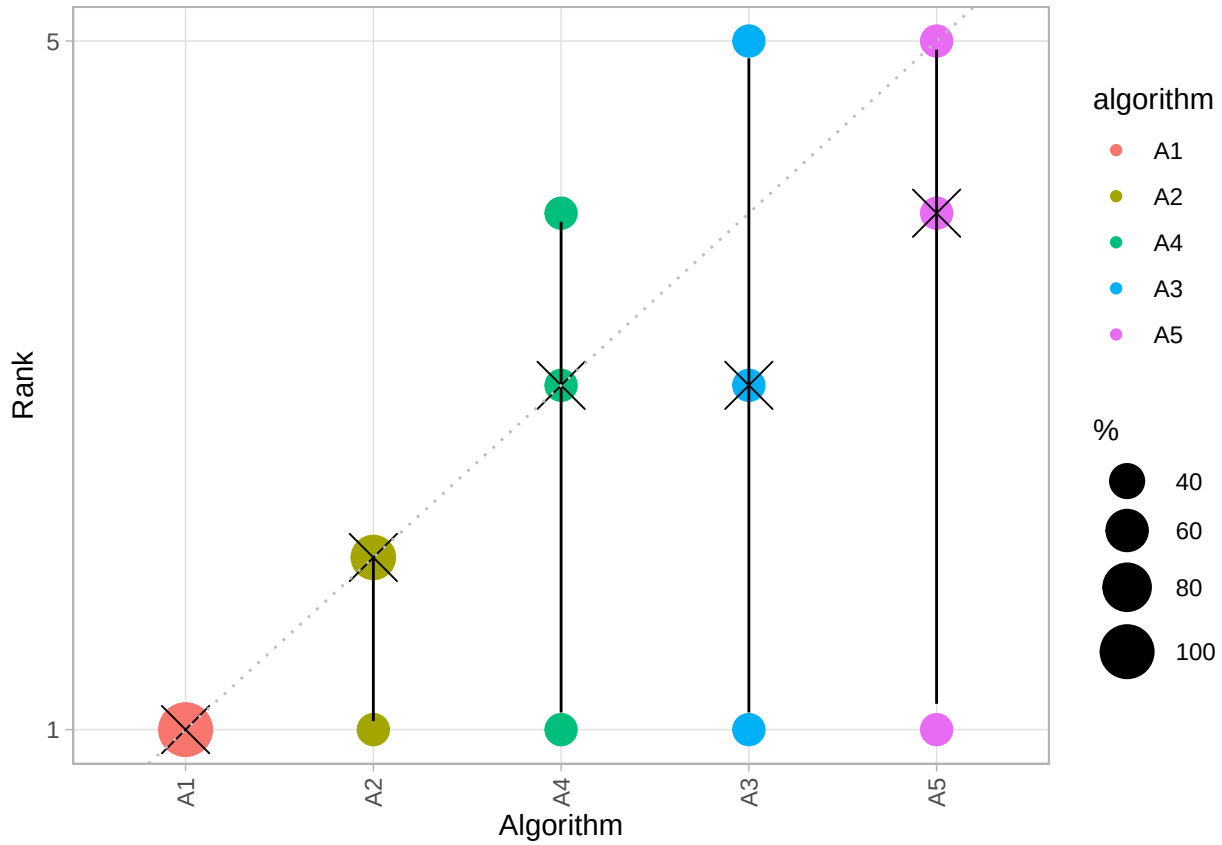
3 Visualization of cross-task insights

Algorithms are ordered according to consensus ranking.

3.1 Characterization of algorithms

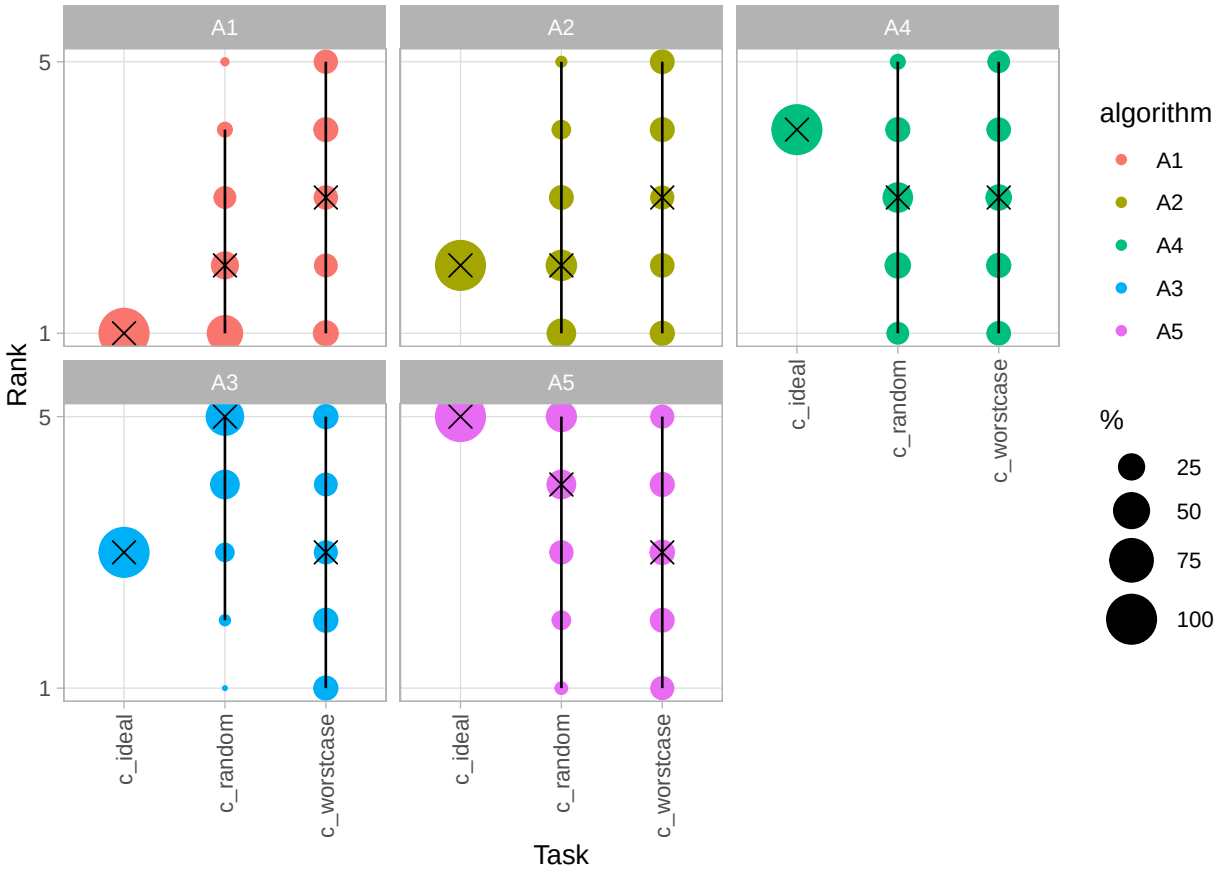
3.1.1 Ranking stability: Variability of achieved rankings across tasks

Blob plot similar to the one shown in Section 2.1 substituting rankings based on bootstrap samples with the rankings corresponding to multiple tasks. This way, the distribution of ranks across tasks can be intuitively visualized.

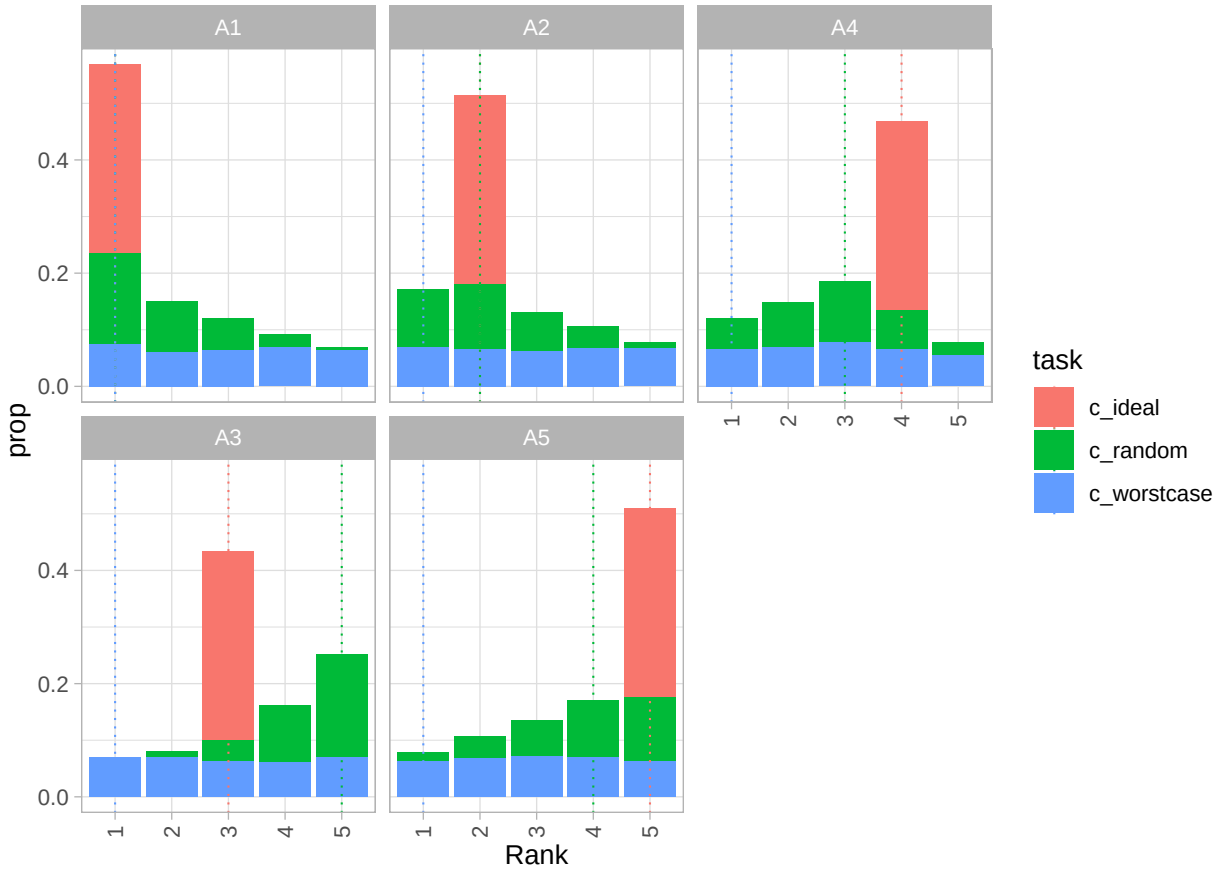


3.1.2 Ranking stability: Ranking variability via bootstrap approach

Blob plot of bootstrap results over the different tasks separated by algorithm allows another perspective on the assessment data. This gives deeper insights into the characteristics of tasks and the ranking uncertainty of the algorithms in each task.



An alternative representation is provided by a stacked frequency plot of the observed ranks, separated by algorithm. Observed ranks across bootstrap samples are displayed with colouring according to task. For algorithms that achieve the same rank in different tasks for the full assessment data set, vertical lines are on top of each other. Vertical lines allow to compare the achieved rank of each algorithm over different tasks.



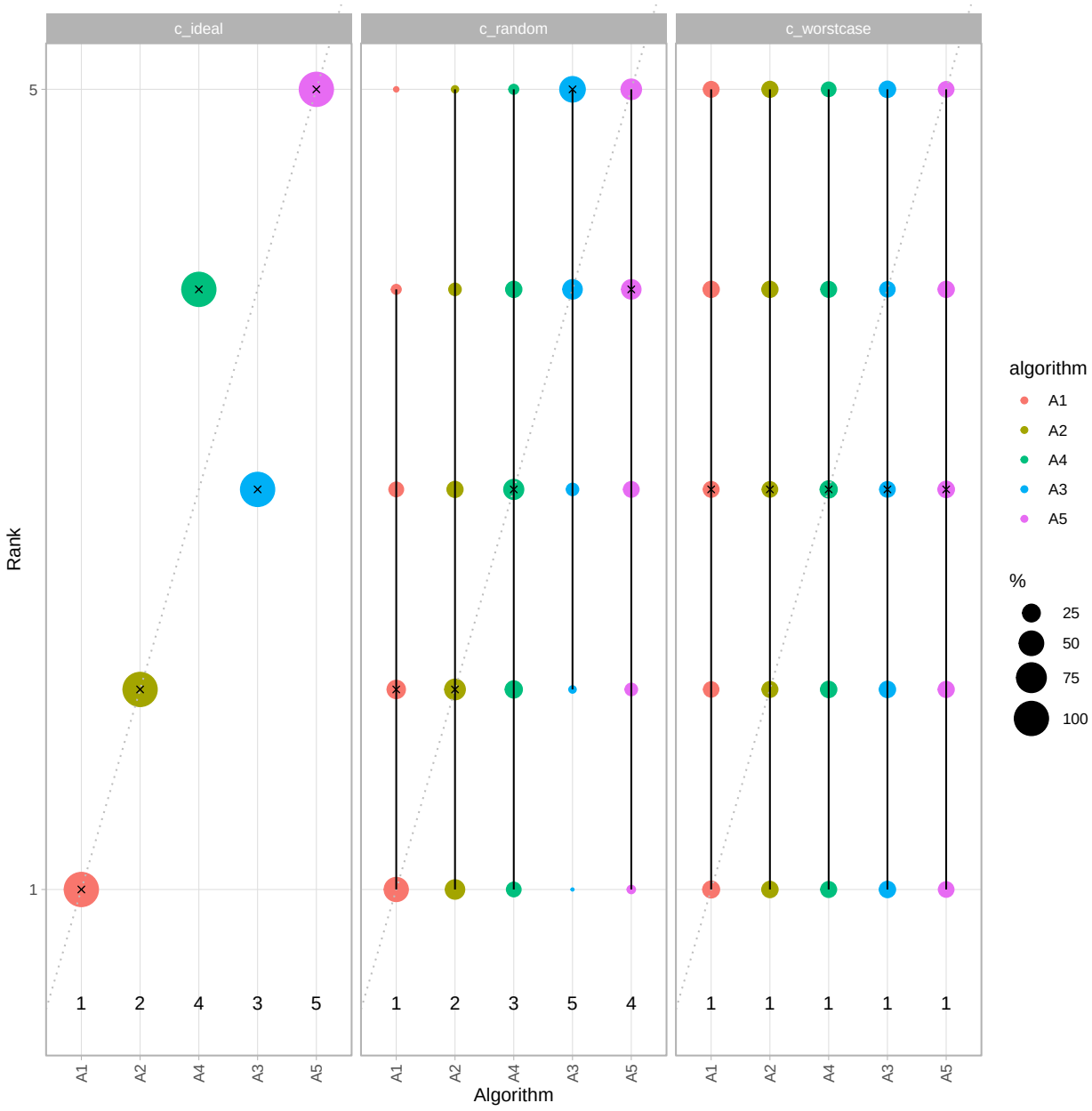
3.2 Characterization of tasks

3.2.1 Visualizing bootstrap results

To investigate which tasks separate algorithms well (i.e., lead to a stable ranking), two visualization methods are recommended.

Bootstrap results can be shown in a blob plot showing one plot for each task. In this view, the spread of the blobs for each algorithm can be compared across tasks. Deviations from the diagonal indicate deviations from the consensus ranking (over tasks). Specifically, if rank distribution of an algorithm is consistently below the diagonal, the algorithm performed better in this task than on average across tasks, while if the rank distribution of an algorithm is consistently above the diagonal, the algorithm performed worse in this task than on average across tasks. At the bottom of each panel, ranks for each algorithm in the tasks is provided.

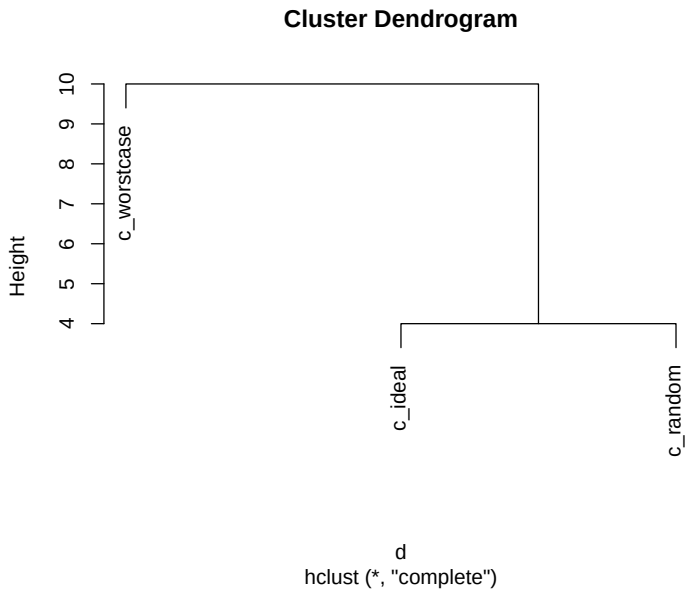
Same as in Section 2.1 but now ordered according to consensus.

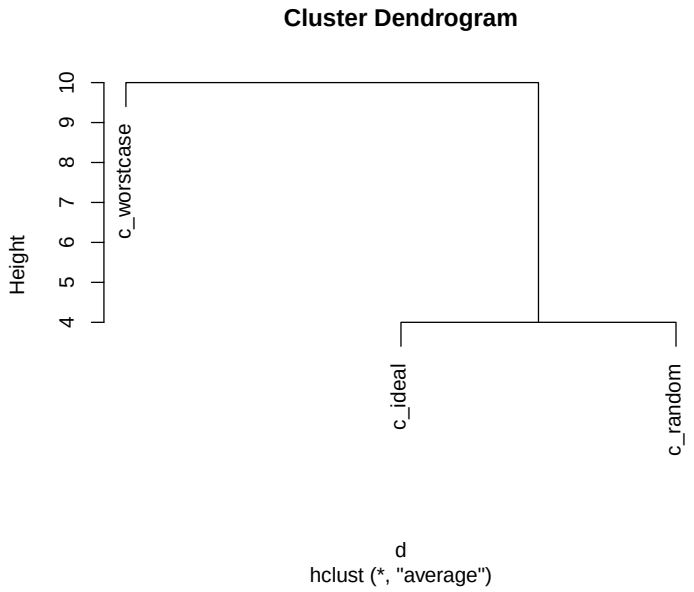


3.2.2 Cluster Analysis

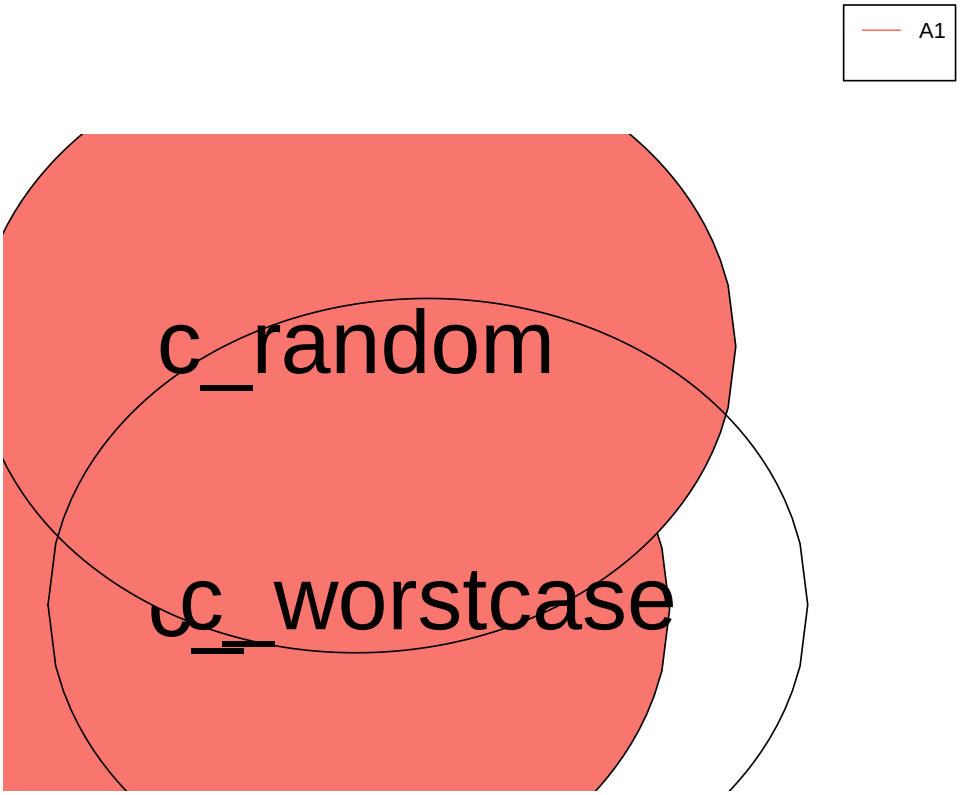
Dendrogram from hierarchical cluster analysis} and *network-type graphs* for assessing the similarity of tasks based on challenge rankings.

A dendrogram is a visualization approach based on hierarchical clustering. It depicts clusters according to a chosen distance measure (here: Spearman's footrule) as well as a chosen agglomeration method (here: complete and average agglomeration).





In network-type graphs (see Eugster et al, 2008), every task is represented by a node and nodes are connected by edges whose length is determined by a chosen distance measure. Here, distances between nodes are chosen to increase exponentially in Spearman's footrule distance with growth rate 0.05 to accentuate large distances. Hence, tasks that are similar with respect to their algorithm ranking appear closer together than those that are dissimilar. Nodes representing tasks with a unique winner are colored-coded by the winning algorithm. In case there are more than one first-ranked algorithms in a task, the corresponding node remains uncolored.



4 References

Wiesenfarth, M., Reinke, A., Landman, B.A., Cardoso, M.J., Maier-Hein, L. and Kopp-Schneider, A. (2019). Methods and open-source toolkit for analyzing and visualizing challenge results. *arXiv preprint arXiv:1910.05121*

M. J. A. Eugster, T. Hothorn, and F. Leisch, "Exploratory and inferential analysis of benchmark experiments," Institut fuer Statistik, Ludwig-Maximilians-Universitaet Muenchen, Germany, Technical Report 30, 2008. [Online]. Available: http://epub.ub.uni-muenchen. de/4134/.