Results

May 8, 2024

1 Tables of Friedman, Bonferroni-Dunn, Holm, Hochberg and Hommel Tests

Table 1: Average Rankings of the algorithms

Algorithm	Ranking
LWRK	2.36
MS	3.68
SPT	1.41
WINQ	4.00
—(6) PPO—PPO-model-0	4.34
—(0) NN—Baseline-model-0	5.22

Friedman statistic considering reduction performance (distributed according to chi-square with 5 degrees of freedom: 823.9542857142534.
P-value computed by Friedman Test: 2.7445967720751696E-10.

Iman and Davenport statistic considering reduction performance (distributed according to F-distribution with 5 and 1495 degrees of freedom: 364.41667511914875.
P-value computed by Iman and Daveport Test: 2.220446049250313E-16.

Table 2: Holm / Hochberg Table for $\alpha = 0.05$

i	algorithm	$z = (R_0 - R_i)/SE$	p	Holm/Hochberg/Hommel
5	—(0) NN—Baseline-model-0	24.94230485397384	2.5880749051374236E-137	0.01
4	—(6) PPO—PPO-model-0	19.15953076272007	8.059654018230552E-82	0.0125
3	WINQ	16.933708282312896	2.5387017706388603E-64	0.016666666666666666
2	MS	14.860638325070983	5.93569157227532E-50	0.025
1	LWRK	6.197388082702176	5.740783105650818E-10	0.05

Bonferroni-Dunn's procedure rejects those hypotheses that have a p-value ≤ 0.01.
Hochberg's procedure rejects those hypotheses that have a p-value ≤ 0.05.
Hommel's procedure rejects all hypotheses.

Table 3: Holm / Hochberg Table for $\alpha = 0.10$

i	algorithm	$z = (R_0 - R_i)/SE$	p	Holm/Hochberg/Hommel
5	—(0) NN—Baseline-model-0	24.94230485397384	2.5880749051374236E-137	0.02
4	—(6) PPO—PPO-model-0	19.15953076272007	8.059654018230552E-82	0.025
3	WINQ	16.933708282312896	2.5387017706388603E-64	0.03333333333333333
2	MS	14.860638325070983	5.93569157227532E-50	0.05
1	LWRK	6.197388082702176	5.740783105650818E-10	0.1

Bonferroni-Dunn's procedure rejects those hypotheses that have a p-value ≤ 0.02.
Hochberg's procedure rejects those hypotheses that have a p-value ≤ 0.1.

Table 4: Adjusted p-values

i	algorithm	unadjusted p	p_{Bonf}	p_{Holm}	p_{Hoch}	p_{Homm}
1	—(0) NN—Baseline-model-0	2.5880749051374236E-137	1.2940374525687117E-136	1.2940374525687117E-136	1.2940374525687117E-136	1.2940374525687117E-136
2	—(6) PPO—PPO-model-0	8.059654018230552E-82	4.029827009115276E-81	3.223861607292221E-81	3.223861607292221E-81	3.223861607292221E-81
3	WINQ	2.5387017706388603E-64	1.2693508853194302E-63	7.616105311916581E-64	7.616105311916581E-64	7.616105311916581E-64
4	MS	5.93569157227532E-50	2.96784578613766E-49	1.187138314455064E-49	1.187138314455064E-49	1.187138314455064E-49
5	LWRK	5.74078310565081 8E-10	2.870391552825409E-9	5.74078310565081 8E-10	5.74078310565081 8E-10	5.74078310565081 8E-10

Table 5: Holm / Shaffer Table for $\alpha = 0.05$

i	algorithms	$z = (R_0 - R_i)/SE$	p	Holm	Shaffer
15	SPT vs. —(0) NN—Baseline-model-0	24.94230485397384	2.5880749051374236E-137	0.0033333333333333335	0.0033333333333333335
14	SPT vs. —(6) PPO—PPO-model-0	19.15953076272007	8.059654018230552E-82	0.0035714285714285718	0.005
13	LWRK vs. —(0) NN—Baseline-model-0	18.74491677127167	2.1300488055684953E-78	0.0038461538461538464	0.005
12	SPT vs. WINQ	16.933708282312896	2.5387017706388603E-64	0.004166666666666667	0.005
11	MS vs. SPT	14.860638325070983	5.93569157227532E-50	0.004545454545454546	0.005
10	LWRK vs. —(6) PPO—PPO-model-0	12.962142680017894	2.0056576955139222E-38	0.005	0.005
9	LWRK vs. WINQ	10.736320199610718	6.872948590264443E-27	0.005555555555555556	0.0071428571428571435
8	MS vs. —(0) NN—Baseline-model-0	10.08166652890286	6.658702603608468E-24	0.00625	0.0071428571428571435
7	LWRK vs. MS	8.663250242368807	4.58500405902497E-18	0.007142857142857143	0.0071428571428571435
6	WINQ vs. —(0) NN—Baseline-model-0	8.008596571660949	1.1602483329669463E-15	0.008333333333333333	0.008333333333333333
5	LWRK vs. SPT	6.197388082702176	5.74078310565081 8E-10	0.01	0.0125
4	—(6) PPO—PPO-model-0 vs. —(0) NN—Baseline-model-0	5.7827740912537715	7.34787097868295E-9	0.0125	0.0125
3	MS vs. —(6) PPO—PPO-model-0	4.298892437649088	1.7165374778670143E-5	0.016666666666666666	0.016666666666666666
2	WINQ vs. —(6) PPO—PPO-model-0	2.2258224804071767	0.02602607440858743	0.025	0.025
1	MS vs. WINQ	2.073069957241912	0.038165769458831544	0.05	0.05

Hommel's procedure rejects all hypotheses.
Nemenyi's procedure rejects those hypotheses that have a p-value $\leq 0.0033333333333333335$.
Holm's procedure rejects those hypotheses that have a p-value ≤ 0.025.
Shaffer's procedure rejects those hypotheses that have a p-value $\leq 0.0033333333333333335$.
Bergmann's procedure rejects these hypotheses:

- LWRK vs. MS
- LWRK vs. SPT
- LWRK vs. WINQ
- LWRK vs. —(6) PPO—PPO-model-0
- LWRK vs. —(0) NN—Baseline-model-0
- MS vs. SPT
- MS vs. WINQ
- MS vs. —(6) PPO—PPO-model-0
- MS vs. —(0) NN—Baseline-model-0
- SPT vs. WINQ
- SPT vs. —(6) PPO—PPO-model-0
- SPT vs. —(0) NN—Baseline-model-0
- WINQ vs. —(6) PPO—PPO-model-0
- WINQ vs. —(0) NN—Baseline-model-0
- —(6) PPO—PPO-model-0 vs. —(0) NN—Baseline-model-0

Table 6: Holm / Shaffer Table for $\alpha = 0.10$

i	algorithms	$z = (R_0 - R_i)/SE$	p	Holm	Shaffer
15	SPT vs. —(0) NN—Baseline-model-0	24.94230485397384	2.5880749051374236E-137	0.006666666666666667	0.006666666666666667
14	SPT vs. —(6) PPO—PPO-model-0	19.15953076272007	8.059654018230552E-82	0.0071428571428571435	0.01
13	LWRK vs. —(0) NN—Baseline-model-0	18.74491677127167	2.1300488055684953E-78	0.007692307692307693	0.01
12	SPT vs. WINQ	16.933708282312896	2.5387017706388603E-64	0.008333333333333333	0.01
11	MS vs. SPT	14.860638325070983	5.93569157227532E-50	0.009090909090909092	0.01
10	LWRK vs. —(6) PPO—PPO-model-0	12.962142680017894	2.0056576955139222E-38	0.01	0.01
9	LWRK vs. WINQ	10.736320199610718	6.872948590264443E-27	0.011111111111111112	0.014285714285714287
8	MS vs. —(0) NN—Baseline-model-0	10.08166652890286	6.658702603608468E-24	0.0125	0.014285714285714287
7	LWRK vs. MS	8.663250242368807	4.58500405902497E-18	0.014285714285714287	0.014285714285714287
6	WINQ vs. —(0) NN—Baseline-model-0	8.008596571660949	1.1602483329669463E-15	0.016666666666666666	0.016666666666666666
5	LWRK vs. SPT	6.197388082702176	5.740783105650818E-10	0.02	0.025
4	—(6) PPO—PPO-model-0 vs. —(0) NN—Baseline-model-0	5.7827740912537715	7.34787097868295E-9	0.025	0.025
3	MS vs. —(6) PPO—PPO-model-0	4.298892437649088	1.7165374778670143E-5	0.03333333333333333	0.03333333333333333
2	WINQ vs. —(6) PPO—PPO-model-0	2.2258224804071767	0.02602607440858743	0.05	0.05
1	MS vs. WINQ	2.073069957241912	0.038165769458831544	0.1	0.1

Nemenyi's procedure rejects those hypotheses that have a p-value $\leq 0.006666666666666667$.
Shaffer's procedure rejects those hypotheses that have a p-value $\leq 0.006666666666666667$.
Bergmann's procedure rejects these hypotheses:

- LWRK vs. MS
- LWRK vs. SPT
- LWRK vs. WINQ
- LWRK vs. —(6) PPO—PPO-model-0
- LWRK vs. —(0) NN—Baseline-model-0
- MS vs. SPT
- MS vs. WINQ
- MS vs. —(6) PPO—PPO-model-0
- MS vs. —(0) NN—Baseline-model-0
- SPT vs. WINQ
- SPT vs. —(6) PPO—PPO-model-0
- SPT vs. —(0) NN—Baseline-model-0
- WINQ vs. —(6) PPO—PPO-model-0
- WINQ vs. —(0) NN—Baseline-model-0
- —(6) PPO—PPO-model-0 vs. —(0) NN—Baseline-model-0

Table 7: Adjusted p-values

i	hypothesis	unadjusted p	p_{Neme}	p_{Holm}	p_{Shaf}	p_{Berq}
1	SPT vs .—(0) NN—Baseline-model-0	2.5880749051374236E-137	3.882112357706135E-136	3.882112357706135E-136	3.882112357706135E-136	3.882112357706135E-136
2	SPT vs .—(6) PPO—PPO-model-0	8.059654018230552E-82	1.2089481027345829E-80	1.1283515625522774E-80	8.059654018230552E-81	8.059654018230552E-81
3	LWRK vs .—(0) NN—Baseline-model-0	2.1300488055684953E-78	3.1950732083527428E-77	2.769063447239044E-77	2.1300488055684953E-77	2.1300488055684953E-77
4	SPT vs .WINQ	2.5387017706388603E-64	3.8080526559582904E-63	3.0464421247666324E-63	2.5387017706388605E-63	1.7770912394472022E-63
5	MS vs .SPT	5.93569157227532E-50	8.90353735841298E-49	6.529260729502852E-49	5.93569157227532E-49	3.561414943365192E-49
6	LWRK vs .—(6) PPO—PPO-model-0	2.0056576955139222E-38	3.0084865432708834E-37	2.005657695513922E-37	2.005657695513922E-37	1.2033946173083534E-37
7	LWRK vs .WINQ	6.872948590264443E-27	1.0309422885396664E-25	6.185653731237999E-26	4.81106401318511E-26	2.749179436105777E-26
8	MS vs .—(0) NN—Baseline-model-0	6.658702603608468E-24	9.988053905412702E-23	5.326962082886774E-23	4.661091822525928E-23	4.661091822525928E-23
9	LWRK vs .MS	4.58500405902497E-18	6.87750608853745 4E-17	3.2095028413174785E-17	3.2095028413174785E-17	1.834001623609988E-17
10	WINQ vs .—(0) NN—Baseline-model-0	1.1602483329669463E-15	1.7403724994504194E-14	6.961489997801678E-15	6.961489997801678E-15	4.640993331867785E-15
11	LWRK vs .SPT	5.740783105650818E-10	8.611174658476226E-9	2.870391552825409E-9	2.296313242260327E-9	2.296313242260327E-9
12	—(6) PPO—PPO-model-0 vs .—(0) NN—Baseline-model-0	7.34787097868295E-9	1.1021806468024425E-7	2.93914839147318E-8	2.93914839147318E-8	1.46957419573659E-8
13	MS vs .—(6) PPO—PPO-model-0	1.716537477867014 3E-5	2.5748062168005215E-4	5.149612433601043E-5	5.149612433601043E-5	5.149612433601043E-5
14	WINQ vs .—(6) PPO—PPO-model-0	0.02602607440858743	0.39039111612881144	0.05205214881717486	0.05205214881717486	0.02602607440858743
15	MS vs .WINQ	0.038165769458831544	0.5724865418824732	0.05205214881717486	0.05205214881717486	0.038165769458831544