Additional file 1: Supplementary material

Ari Urkullu Aritz Pérez Borja Calvo

1 Introduction

In this additional file 1, we gather additional documentation which is not presented in the manuscript. Briefly, this additional documentation consists of:

- Section 2, detailed description of the Algorithm to find $\boldsymbol{a}^*$: An extensive and detailed description of the Algorithm that enables $\boldsymbol{a}^*$ to be found is exposed.
- Section 3, parameters of the synthetic experimentation: The specific values of the parameters of all the distributions used during the experimentation with synthetic data are provided in a table.
- Section 4, plots and tables of the experimentation with synthetic data: All the plots and tables derived from the experimentation with synthetic data are displayed in many figures and tables.
- Section 5, descriptions of the real databases: A description for each real database is presented.
- Section 6, preprocessings of the real databases: The preprocessings applied to the different real databases are exposed.
- Section 7, ovarian cancer database stratification: The stratification process applied after the preprocessing has been carried out when the sampling of the data is tackled in order to derive $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$.
- Section 8, plots and tables of the experimentation with real data: All the plots and tables derived from the experimentation with real data are displayed in many figures and tables.

2 Detailed description of the Algorithm to find $\boldsymbol{a}^*$

First, Algorithm 1, which is the algorithm that enables $\boldsymbol{a}^*$ to be found, is presented. Secondly, the detailed description of Algorithm 1 is exposed.

Algorithm 1 The pseudo-code of the algorithm used for computing the sequence of the amounts of relevant balls $\boldsymbol{a}^*$ of minimum error.

1: **procedure** COMPUTING $\boldsymbol{a}^*$
Input: Estimated expected reproducibility curve ($\hat{\boldsymbol{\rho}}$).
Output: Sequence of the amounts of relevant balls that minimizes the cumulative error function ($\boldsymbol{a}^*$).
2: $n = \text{length}(\hat{\boldsymbol{\rho}})$
3: $\boldsymbol{S} = \text{Zeros}(n, \lfloor n/2 \rfloor + 1)$
4: $\boldsymbol{e} = \text{Zeros}(\lfloor n/2 \rfloor + 1)$
5: **for** $m = 0$ to $\lfloor n/2 \rfloor$ **do**
6: $\boldsymbol{E}_m = \text{Infinites}(n+1, m+1)$
7: $\boldsymbol{E}_m[0,0] = 0$
8: $\boldsymbol{P}_m = \text{Zeros}(n+1, m+1)$
9: **for** $i = 1$ to n **do**
10: **for** $j = 0$ to $min(i,m)$ **do**
11: **if** $j = 0$ **then**
12: $\boldsymbol{E}_m[i,j] = \boldsymbol{E}_m[i-1,j] + e_i(\hat{\rho}_i, j, m)$
13: $\boldsymbol{P}_m[i,j] = j$
14: **end if**
15: **if** $j = i$ **then**
16: $\boldsymbol{E}_m[i,j] = \boldsymbol{E}_m[i-1,j-1] + e_i(\hat{\rho}_i, j, m)$
17: $\boldsymbol{P}_m[i,j] = j-1$
18: **end if**
19: **if** $j \neq 0$ and $j \neq i$ **then**
20: **if** $\boldsymbol{E}_m[i-1,j] < \boldsymbol{E}_m[i-1,j-1]$ **then**
21: $\boldsymbol{E}_m[i,j] = \boldsymbol{E}_m[i-1,j] + e_i(\hat{\rho}_i, j, m)$
22: $\boldsymbol{P}_m[i,j] = j$
23: **else**
24: $\boldsymbol{E}_m[i,j] = \boldsymbol{E}_m[i-1,j-1] + e_i(\hat{\rho}_i, j, m)$
25: $\boldsymbol{P}_m[i,j] = j-1$
26: **end if**
27: **end if**
28: **end for**
29: **end for**
30: $\boldsymbol{e}[m] = \boldsymbol{E}_m[m,n]$
31: $\boldsymbol{S}[.,m] = \text{get\_subproblem\_best\_solution}(\boldsymbol{P}_m)$
32: **end for**
33: $\boldsymbol{a}^* = \text{get\_problem\_best\_solution}(\boldsymbol{S}, \boldsymbol{e})$
34: **return** $\boldsymbol{a}^*$
35: **end procedure**

- Line 2: Sets n to be the amount of balls, which is equal to the amount of features of the estimated expected reproducibility curve $\hat{\boldsymbol{\rho}}$.
- Line 3: Creates the matrix $\boldsymbol{S}$ of n rows and $\lfloor n/2 \rfloor + 1$ columns filled with zeros. This matrix is dedicated to storing in each column the best solution (a sequence $\boldsymbol{a}$) of a different subproblem of the $\lfloor n/2 \rfloor + 1$ subproblems.
- Line 4: Creates the vector $\boldsymbol{e}$ of length $\lfloor n/2 \rfloor + 1$ filled with zeros. This vector is dedicated to storing the $\lfloor n/2 \rfloor + 1$ errors associated to the $\lfloor n/2 \rfloor + 1$ best solutions of the $\lfloor n/2 \rfloor + 1$ subproblems.
- Line 5: Starts the outer loop in which each iteration is dedicated to a different subproblem of the $\lfloor n/2 \rfloor + 1$ subproblems.
- Line 6: Creates the matrix $\boldsymbol{E}_m$ of $n+1$ rows and $\lfloor n/2 \rfloor + 1$ columns filled with infinites. This matrix is dedicated to storing the cumulative errors described in Equation 1 (Equation 13 in the manuscript):

$$E^i_{a_i}(\hat{\boldsymbol{\rho}}) = e_i(\hat{\rho}_i, a_i, a_n) + \min(E^{i-1}_{a_i}(\hat{\boldsymbol{\rho}}), E^{i-1}_{a_i-1}(\hat{\boldsymbol{\rho}})) \tag{1}$$

- Line 7: Initializes the trivial case of $\boldsymbol{E}_m$, in which the cumulative error is always 0 by definition.
- Line 8: Creates the matrix $\boldsymbol{P}_m$ of $n+1$ rows and $\lfloor n/2 \rfloor + 1$ columns filled with zeros. This matrix is dedicated to storing the paths that enable the retrieval of the best solution $\boldsymbol{a}$ for the m-th subproblem. Specifically, for a given

cell in row i and in column j, it stores information regarding the best solution $\boldsymbol{a}$ for the m-th subproblem belonging to the subset of solutions that fulfill $a_i = j$. Briefly, the value of that cell, $\boldsymbol{P}_m[i,j]$, specifies both the value of a_{i-1} for that solution and the column of the previous row of $\boldsymbol{P}_m$ in which the value of a_{i-2} can be located.

- Line 9: Starts the middle loop in which each iteration is dedicated to a different row of the m-th subproblem, given that each row is associated to a different position of the sequences of the amounts of relevant balls.
- Line 10: Starts the inner loop in which each iteration is dedicated to a different column of the m-th subproblem, given that each column is associated to a different amount of relevant balls.
- Lines 11 to 14: Given that $\boldsymbol{E}_m[i-1,j-1]$ is not an option because $j = 0$ (line 11), the only possibility is to set the cumulative error $\boldsymbol{E}_m[i,j]$ to be $e_i(\hat{\rho}_i, j, m) + \boldsymbol{E}_m[i-1,j]$, following the essence of Equation 1. $\boldsymbol{P}_m[i,j]$ is updated in consonance to be j.
- Lines 15 to 18: Given that $\boldsymbol{E}_m[i-1,j]$ is not an option because $j = i$ (line 15), the only possibility is to set the cumulative error $\boldsymbol{E}_m[i,j]$ to be $e_i(\hat{\rho}_i, j, m) + \boldsymbol{E}_m[i-1,j-1]$, following the essence of Equation 1. $\boldsymbol{P}_m[i,j]$ is updated in consonance to be $j-1$.
- Lines 19 to 27: Given that $j \neq 0$ and that $j \neq i$ (line 19) there are two possibilities (either $\boldsymbol{E}_m[i-1,j] < \boldsymbol{E}_m[i-1,j-1]$ is satisfied or not). Each possibility corresponds to a different option in Equation 1. Consequently, each possibility implies a different update of $\boldsymbol{E}_m[i,j]$ and $\boldsymbol{P}_m[i,j]$, according to Equation 1.
- Line 30: The best error achieved in the m-th subproblem is stored in the m-th position of $\boldsymbol{e}$.
- Line 31: Given the matrix of paths $\boldsymbol{P}_m$, the sequence $\boldsymbol{a}$ that is the best solution for subproblem m is derived. First, for this sequence $\boldsymbol{a}$, it is already known that $a_n = m$. Secondly, the rest of the solution $\boldsymbol{a}$ is derived starting from $\boldsymbol{P}_m[m,n]$. Let us recall that the value $\boldsymbol{P}_m[m,n]$ specifies both the value a_{n-1} of the solution $\boldsymbol{a}$ and the column of the previous row of $\boldsymbol{P}_m$ in which the value of a_{n-2} can be located. In conclusion, departing from $\boldsymbol{P}_m[m,n]$ and proceeding recursively, the sequence $\boldsymbol{a}$ is derived. Finally, the retrieved sequence $\boldsymbol{a}$ is stored in the m-th column of $\boldsymbol{S}$.
- Line 33: First, the position in $\boldsymbol{e}$ that stores the minimum value is located. Secondly, the sequence $\boldsymbol{a}$ stored in corresponding row of $\boldsymbol{S}$ is stored in $\boldsymbol{a}^*$.

3 Parameters of the synthetic experimentation

First of all, for the sake of clarity, let us show again in Figure 1 the distributions from which the data are sampled in the experimentation with synthetic data (this figure is also shown in the manuscript).

In Table 1 we show the specific values of the parameters of each of the distributions shown in Figure 1. In Table 1 each of those distributions is identified by its associated scenario (differences in location or differences in both location and spread) and difficulty as shown in Figure 1.

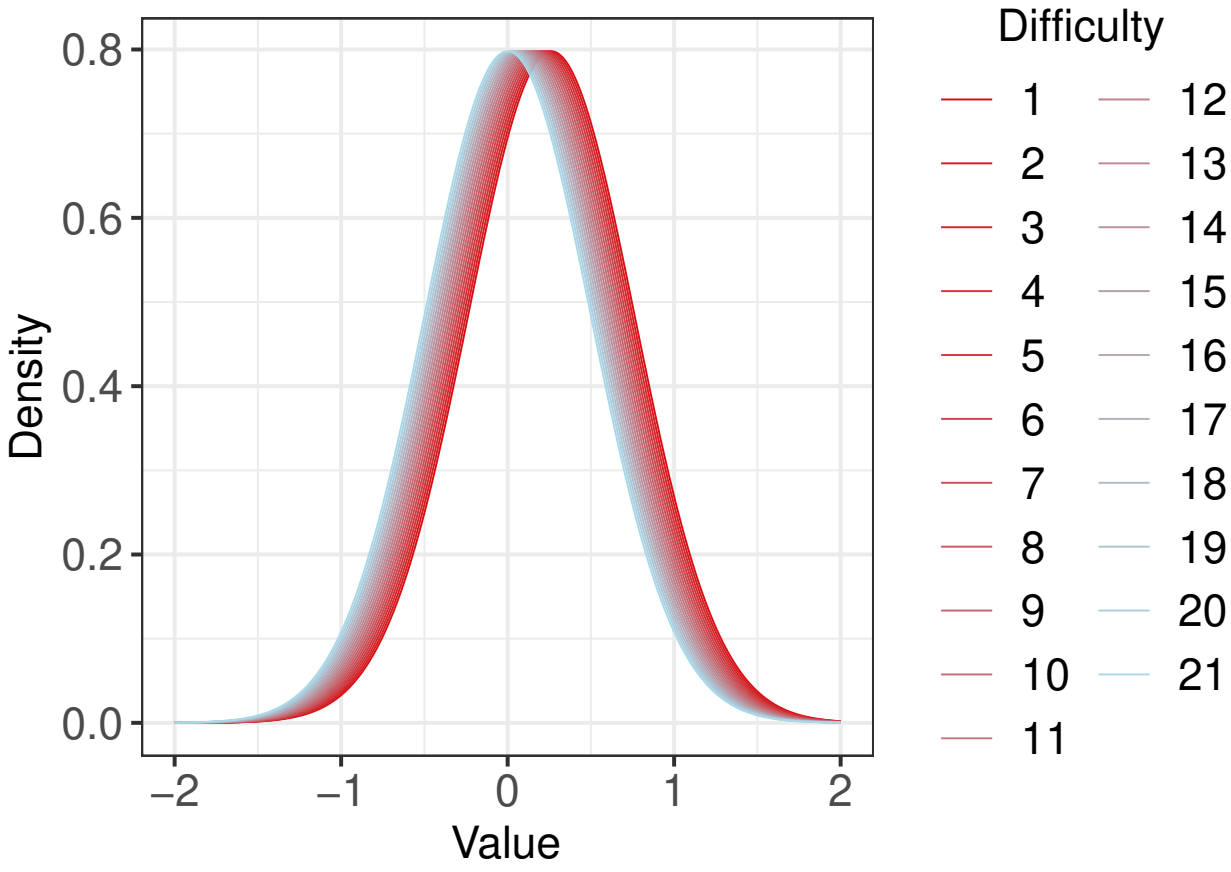
4 Plots and tables of the experimentation with synthetic data

In Figures 2 to 43 the plots of the experimentation with synthetic data can be seen. Specifically, the plots are shown in order, first showing those corresponding to the scenario of differences in location and then showing those corresponding to the scenario of differences in both location and spread. Additionally, the plots belonging to the same scenario are shown in order, first showing those corresponding to difficulty 1 and lastly showing those corresponding to difficulty 21.

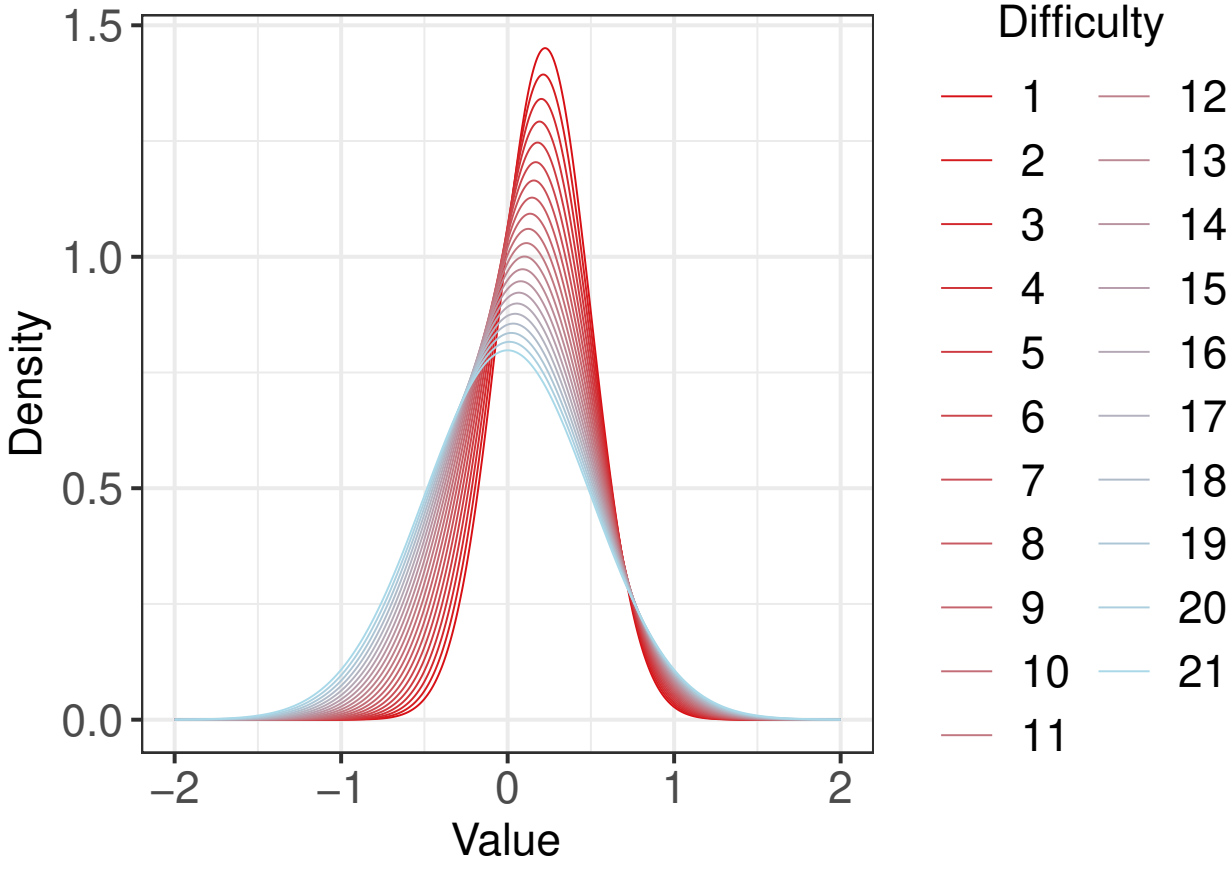
In Tables 2 and 3 the weights and AUC values of the experimentation with synthetic data can be seen.

5 Descriptions of the real databases

- Breast database [3, 5]: This dataset includes 30 features and 569 instances. Specifically, the features consist of visual characteristics of the cell nuclei present in digitized images from patients with breast cancer. The instances of this dataset are breast masses of women with breast tumors, 357 women with benign tumors and 212 women with malignant tumors.



(a)

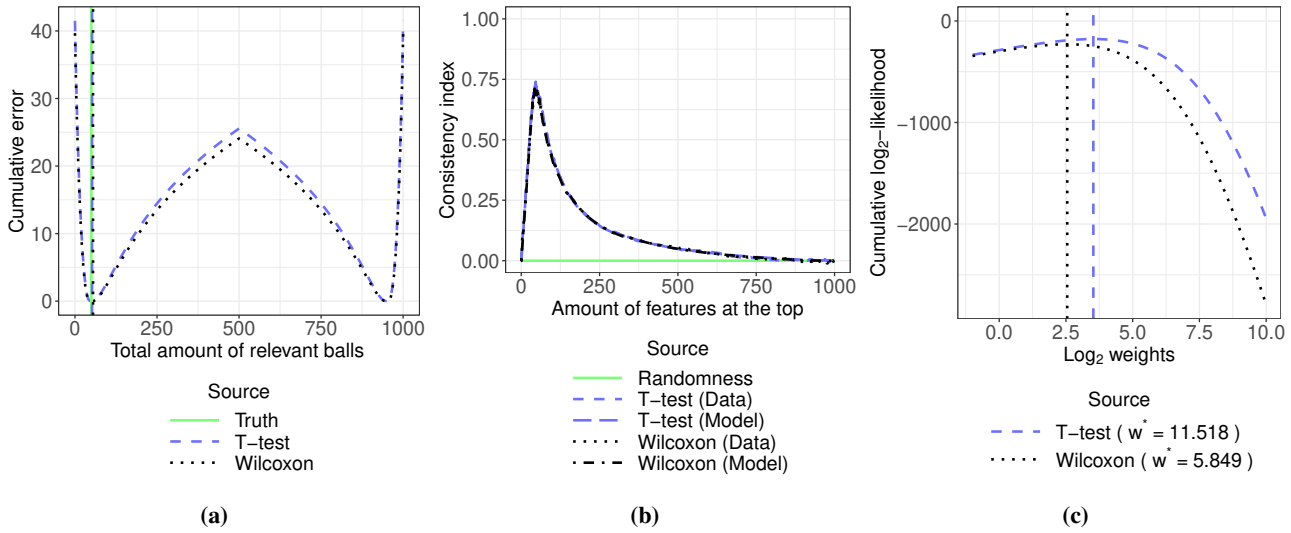


(b)

Figure 1 Distributions used in the scenario of differences in location (1a) and in the scenario of differences in both location and spread (1b)

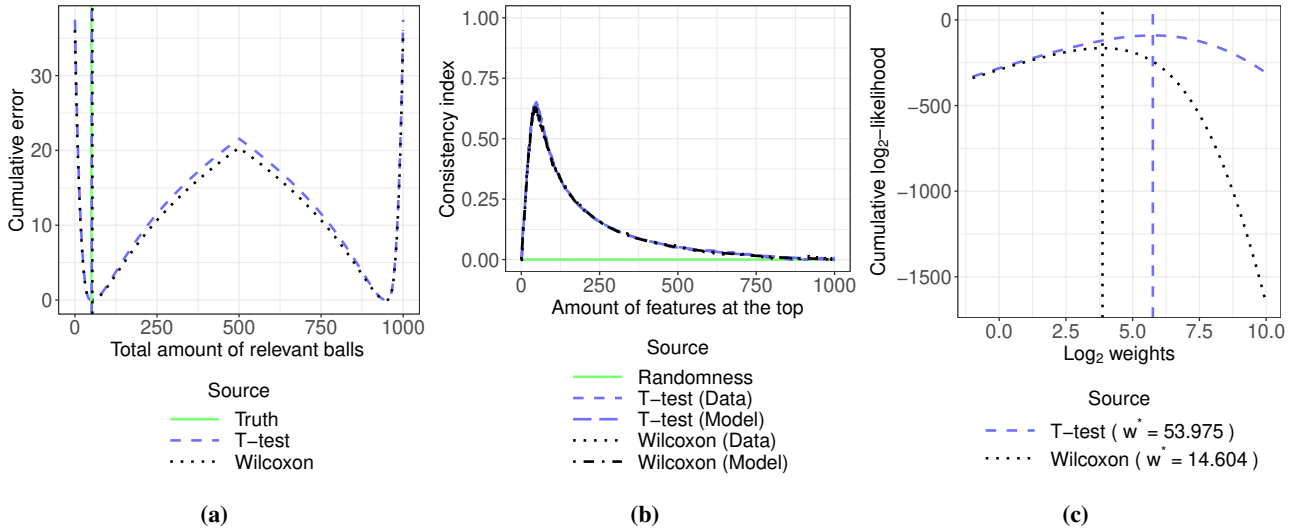
Table 1 Weight ratios for the different combinations of methods, problems and difficulties when dealing with synthetic data

Scenario	Difficulty	Parameters	
		μ	σ^2
Location	1	0.262500	0.25
Location	2	0.249375	0.25
Location	3	0.236250	0.25
Location	4	0.223125	0.25
Location	5	0.210000	0.25
Location	6	0.196875	0.25
Location	7	0.183750	0.25
Location	8	0.170625	0.25
Location	9	0.157500	0.25
Location	10	0.144375	0.25
Location	11	0.131250	0.25
Location	12	0.118125	0.25
Location	13	0.105000	0.25
Location	14	0.091875	0.25
Location	15	0.078750	0.25
Location	16	0.065625	0.25
Location	17	0.052500	0.25
Location	18	0.039375	0.25
Location	19	0.026250	0.25
Location	20	0.013125	0.25
Location	21	0.000000	0.25
Location & spread	1	0.22500	0.0756250000
Location & spread	2	0.21375	0.0819390625
Location & spread	3	0.20250	0.0885062500
Location & spread	4	0.19125	0.0953265625
Location & spread	5	0.18000	0.1024000000
Location & spread	6	0.16875	0.1097265625
Location & spread	7	0.15750	0.1173062500
Location & spread	8	0.14625	0.1251390625
Location & spread	9	0.13500	0.1332250000
Location & spread	10	0.12375	0.1415640625
Location & spread	11	0.11250	0.1501562500
Location & spread	12	0.10125	0.1590015625
Location & spread	13	0.09000	0.1681000000
Location & spread	14	0.07875	0.1774515625
Location & spread	15	0.06750	0.1870562500
Location & spread	16	0.05625	0.1969140625
Location & spread	17	0.04500	0.2070250000
Location & spread	18	0.03375	0.2173890625
Location & spread	19	0.02250	0.2280062500
Location & spread	20	0.01125	0.2388765625
Location & spread	21	0.00000	0.2500000000



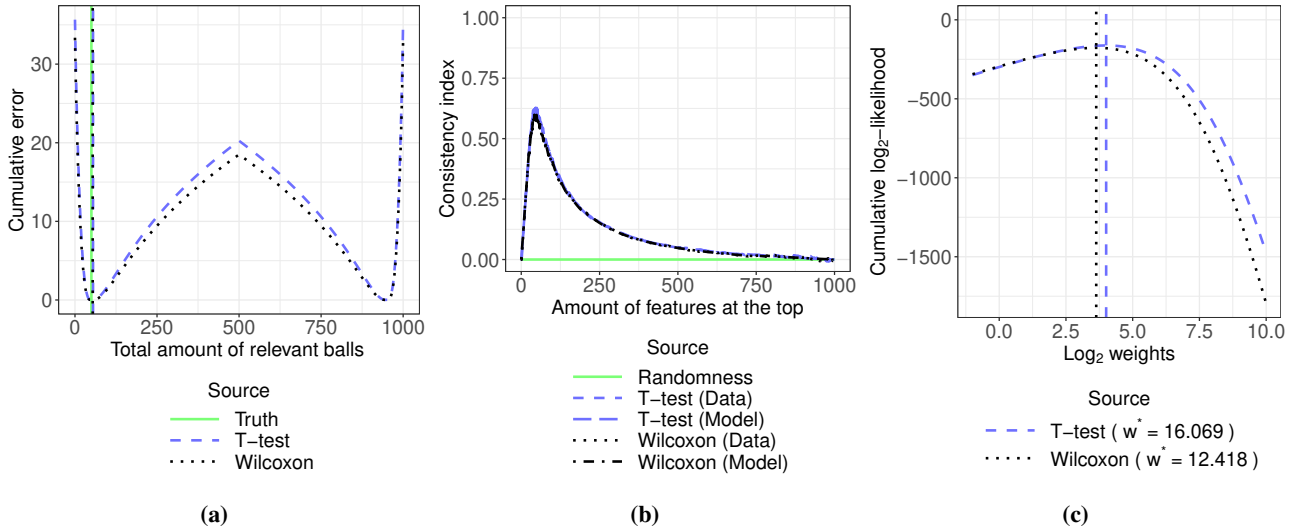
(a) **(b)** **(c)**

Figure 2 Error plot (2a), reproducibility plot (2b) and weight plot (2c) for the difficulty configuration 1, in the differences in location scenario



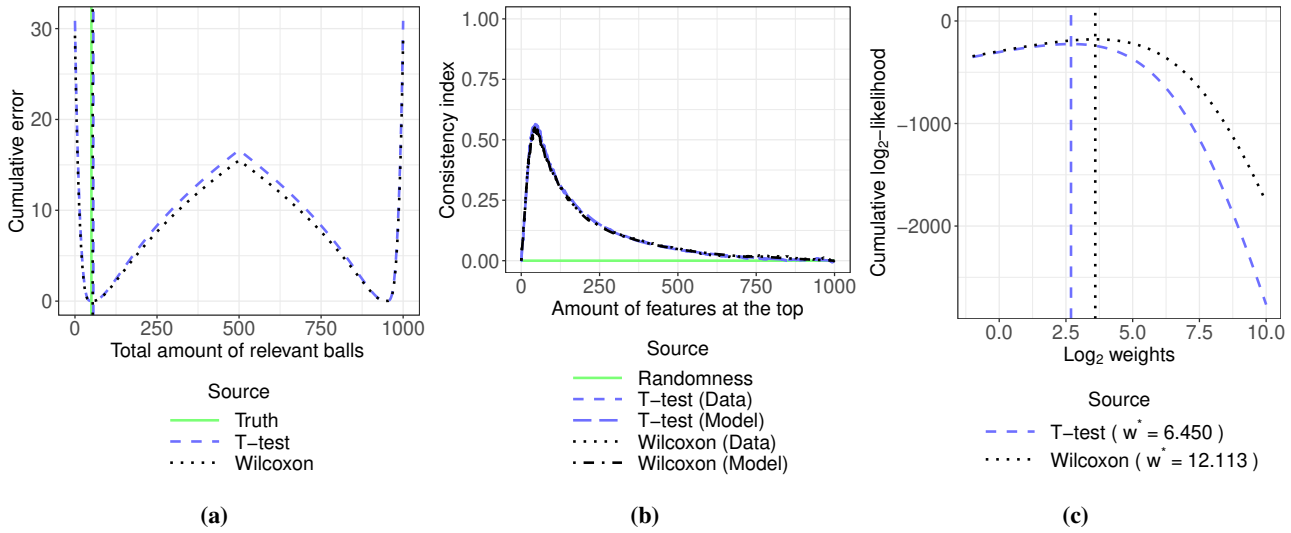
(a) **(b)** **(c)**

Figure 3 Error plot (3a), reproducibility plot (3b) and weight plot (3c) for the difficulty configuration 2, in the differences in location scenario



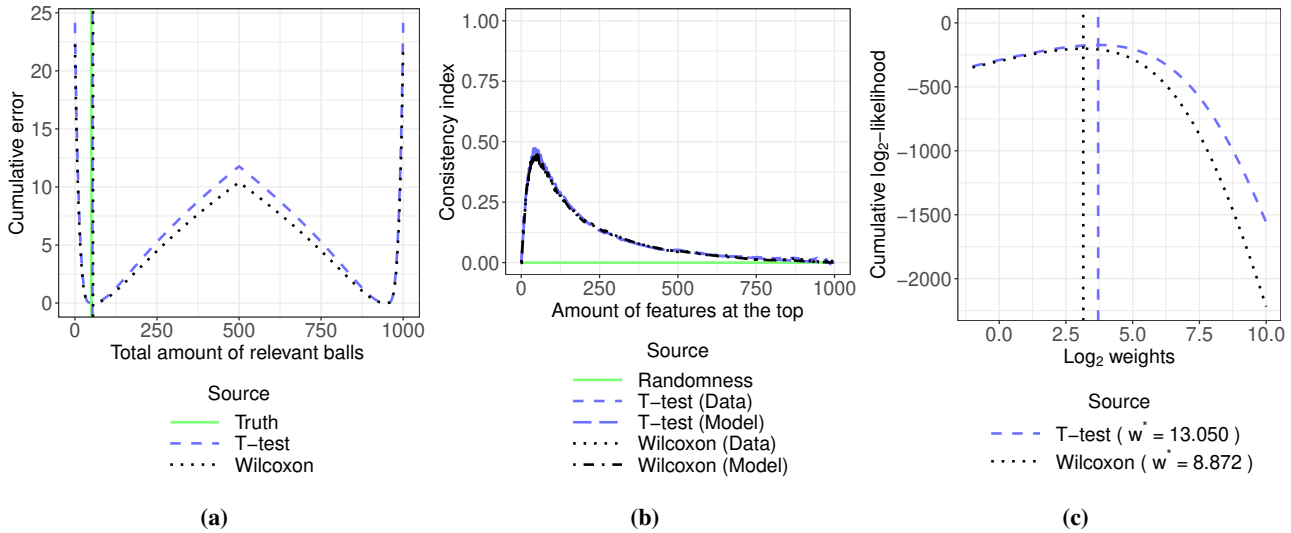
(a) **(b)** **(c)**

Figure 4 Error plot (4a), reproducibility plot (4b) and weight plot (4c) for the difficulty configuration 3, in the differences in location scenario



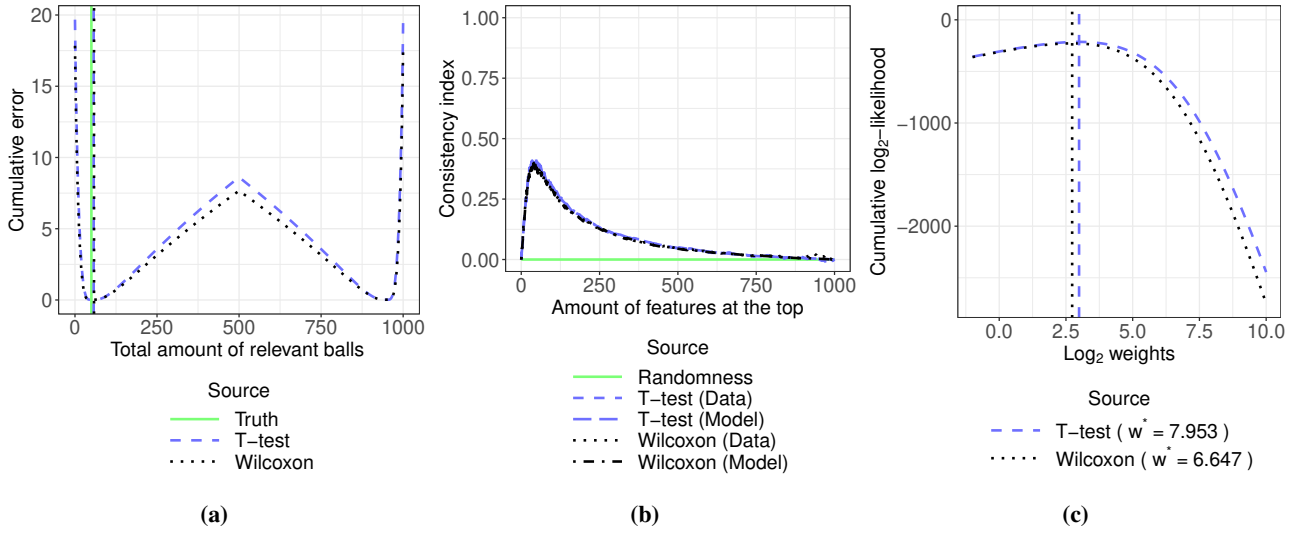
(a) **(b)** **(c)**

Figure 5 Error plot (5a), reproducibility plot (5b) and weight plot (5c) for the difficulty configuration 4, in the differences in location scenario



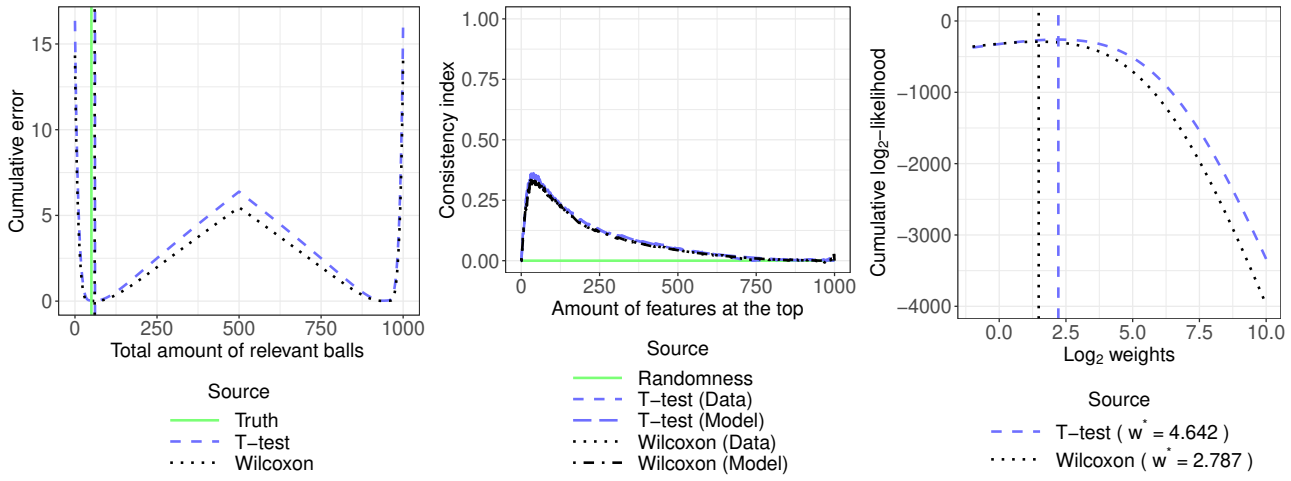
(a) **(b)** **(c)**

Figure 6 Error plot (6a), reproducibility plot (6b) and weight plot (6c) for the difficulty configuration 5, in the differences in location scenario



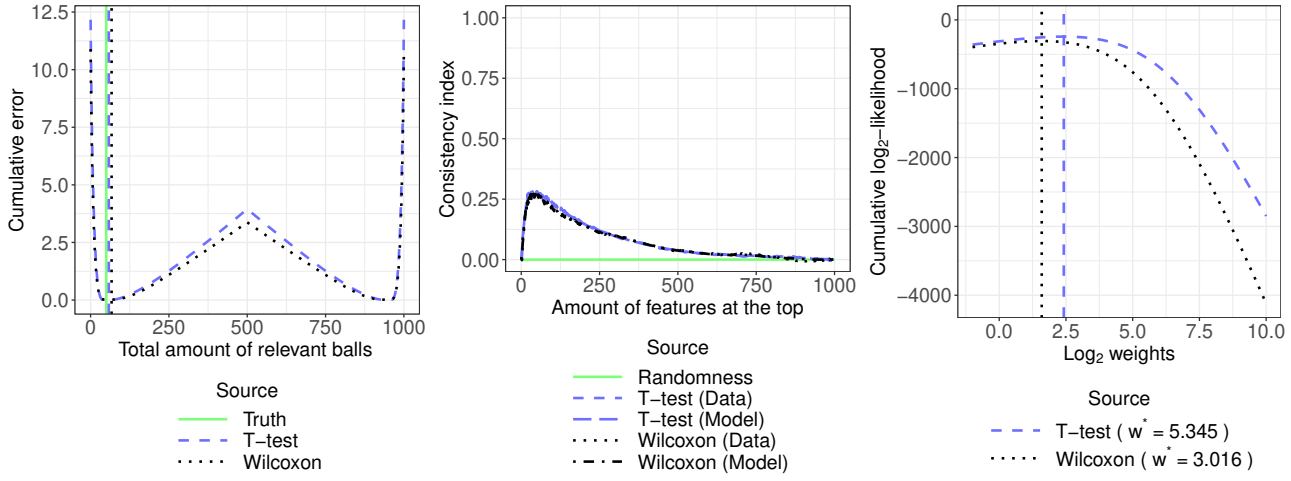
(a) **(b)** **(c)**

Figure 7 Error plot (7a), reproducibility plot (7b) and weight plot (7c) for the difficulty configuration 6, in the differences in location scenario



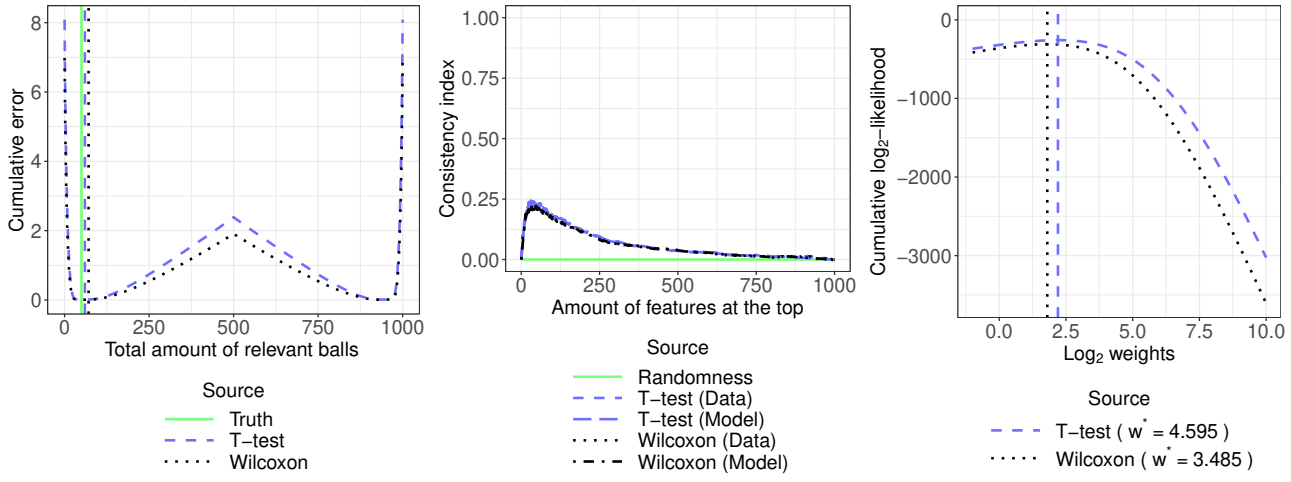
(a) **(b)** **(c)**

Figure 8 Error plot (8a), reproducibility plot (8b) and weight plot (8c) for the difficulty configuration 7, in the differences in location scenario



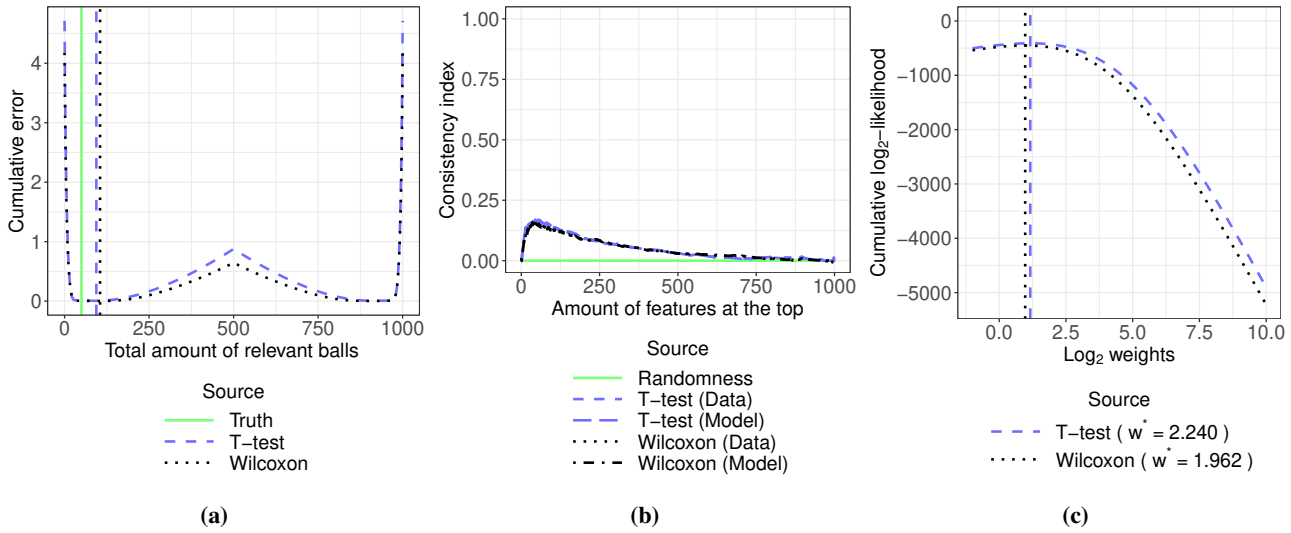
(a) **(b)** **(c)**

Figure 9 Error plot (9a), reproducibility plot (9b) and weight plot (9c) for the difficulty configuration 8, in the differences in location scenario



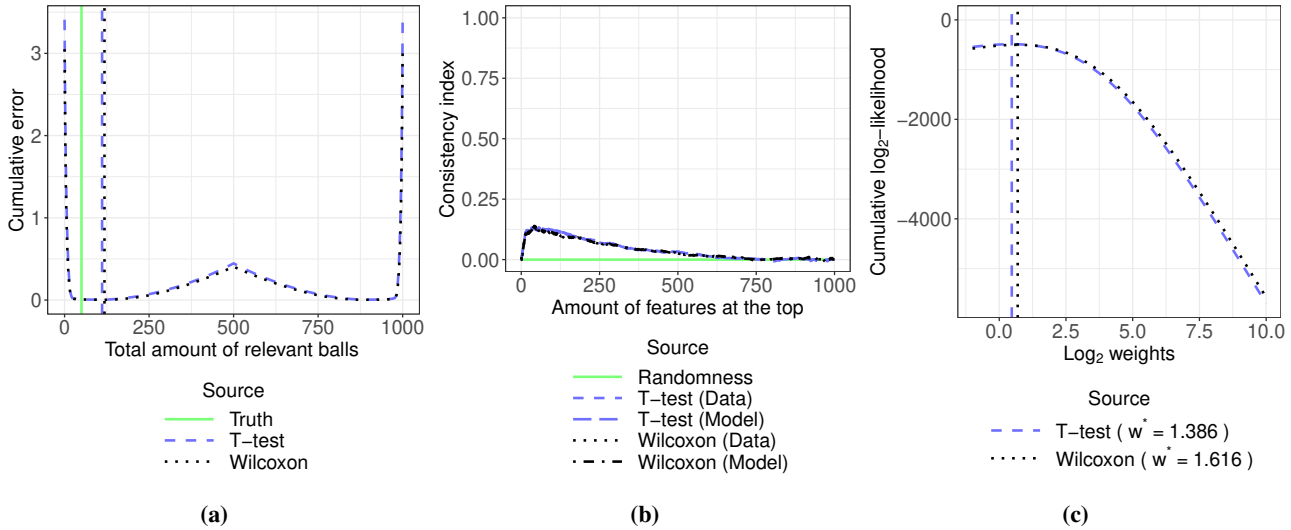
(a) **(b)** **(c)**

Figure 10 Error plot (10a), reproducibility plot (10b) and weight plot (10c) for the difficulty configuration 9, in the differences in location scenario



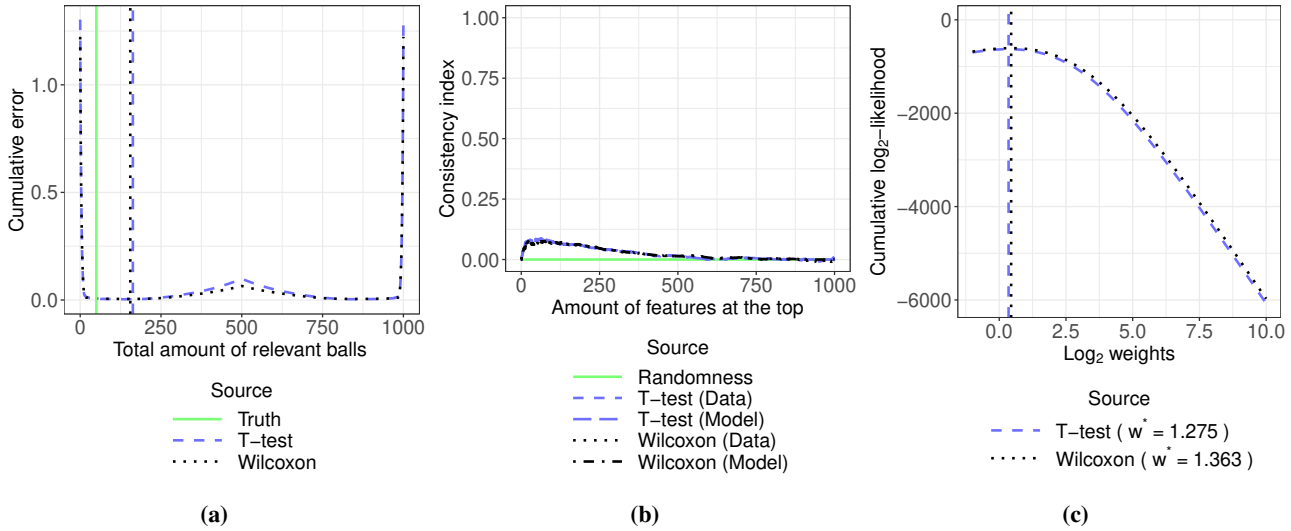
(a) **(b)** **(c)**

Figure 11 Error plot (11a), reproducibility plot (11b) and weight plot (11c) for the difficulty configuration 10, in the differences in location scenario



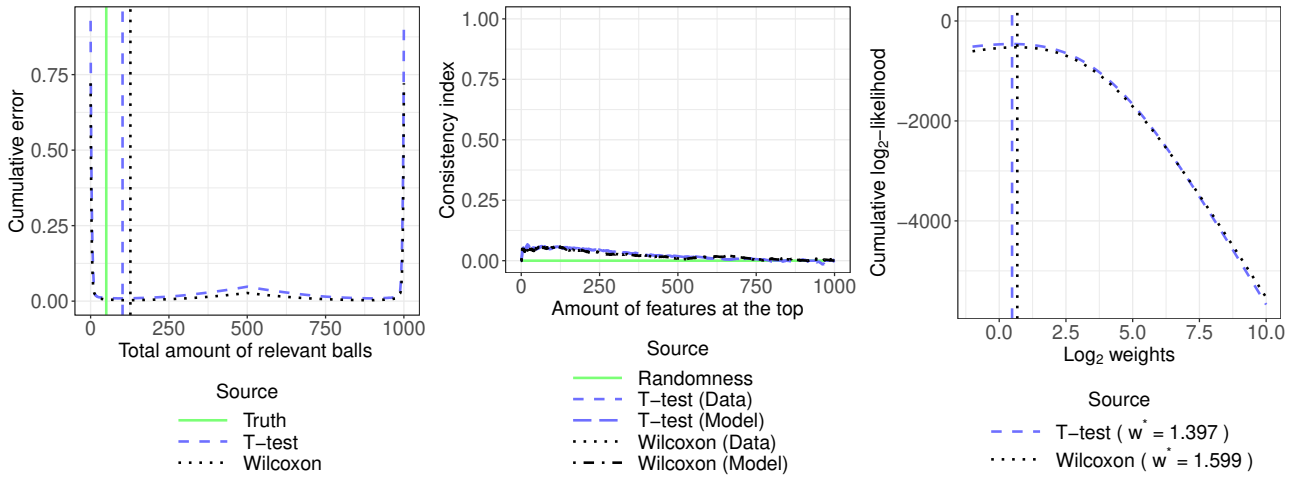
(a) **(b)** **(c)**

Figure 12 Error plot (12a), reproducibility plot (12b) and weight plot (12c) for the difficulty configuration 11, in the differences in location scenario



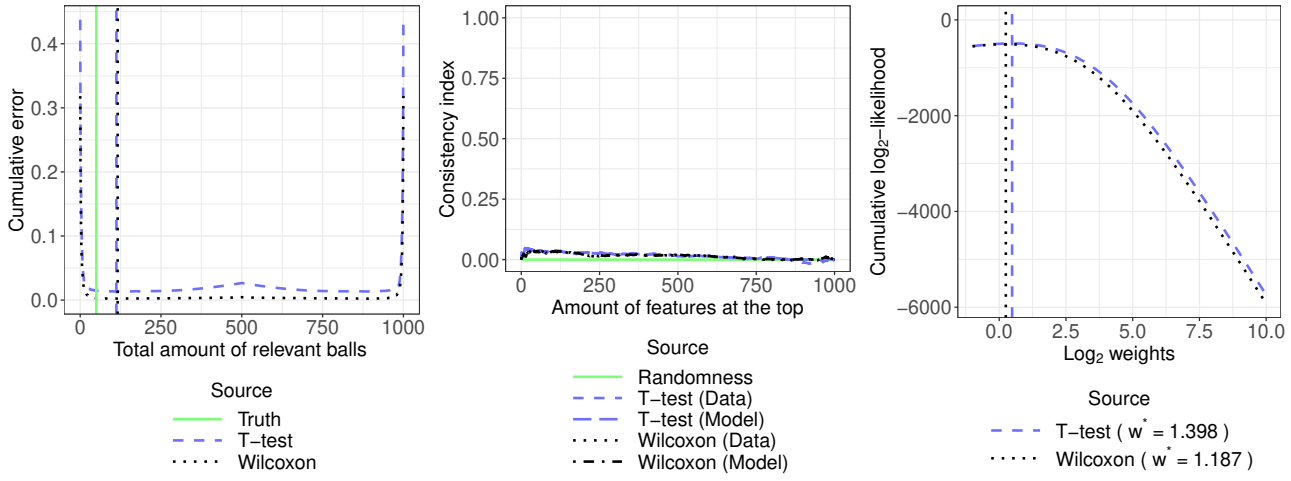
(a) **(b)** **(c)**

Figure 13 Error plot (13a), reproducibility plot (13b) and weight plot (13c) for the difficulty configuration 12, in the differences in location scenario



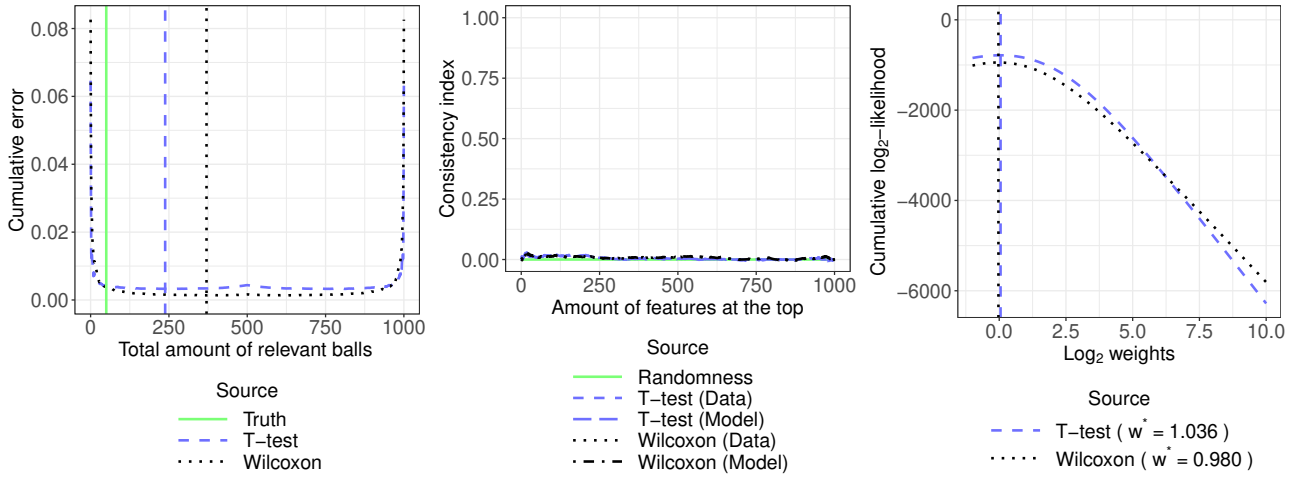
(a) **(b)** **(c)**

Figure 14 Error plot (14a), reproducibility plot (14b) and weight plot (14c) for the difficulty configuration 13, in the differences in location scenario



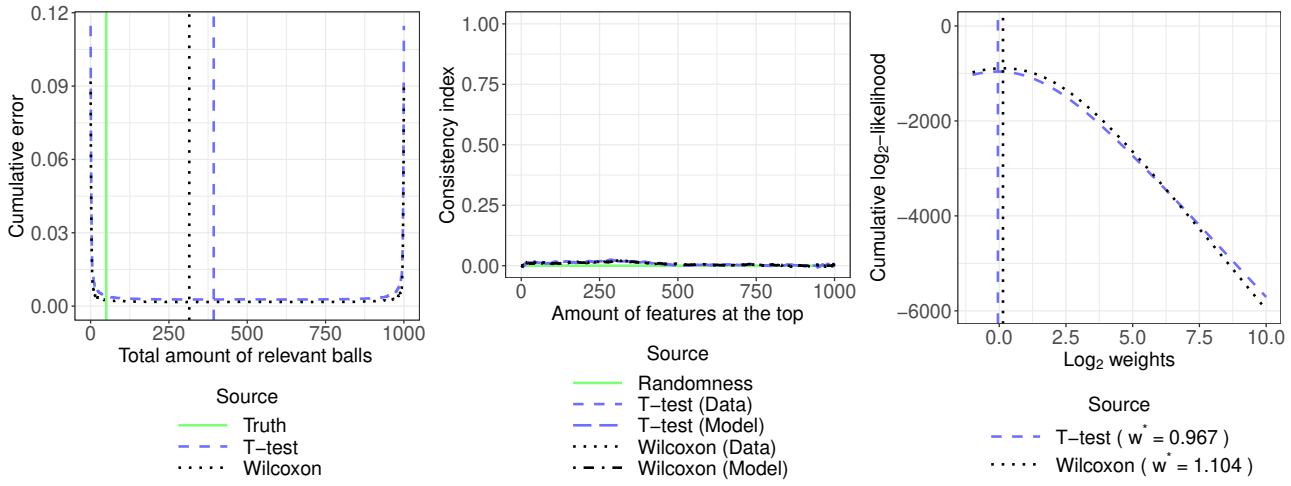
(a) **(b)** **(c)**

Figure 15 Error plot (15a), reproducibility plot (15b) and weight plot (15c) for the difficulty configuration 14, in the differences in location scenario



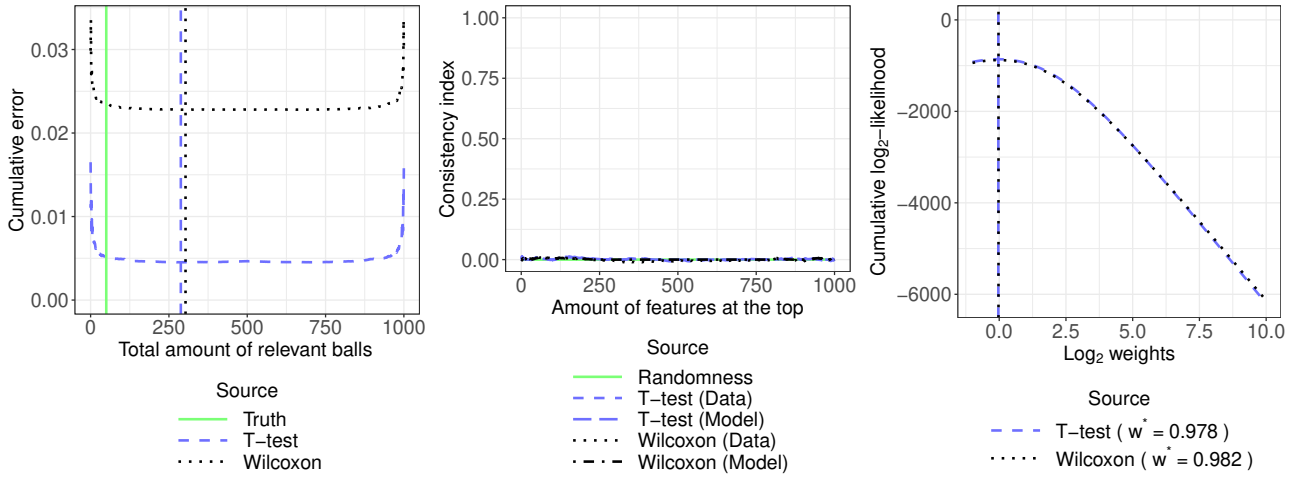
(a) **(b)** **(c)**

Figure 16 Error plot (16a), reproducibility plot (16b) and weight plot (16c) for the difficulty configuration 15, in the differences in location scenario



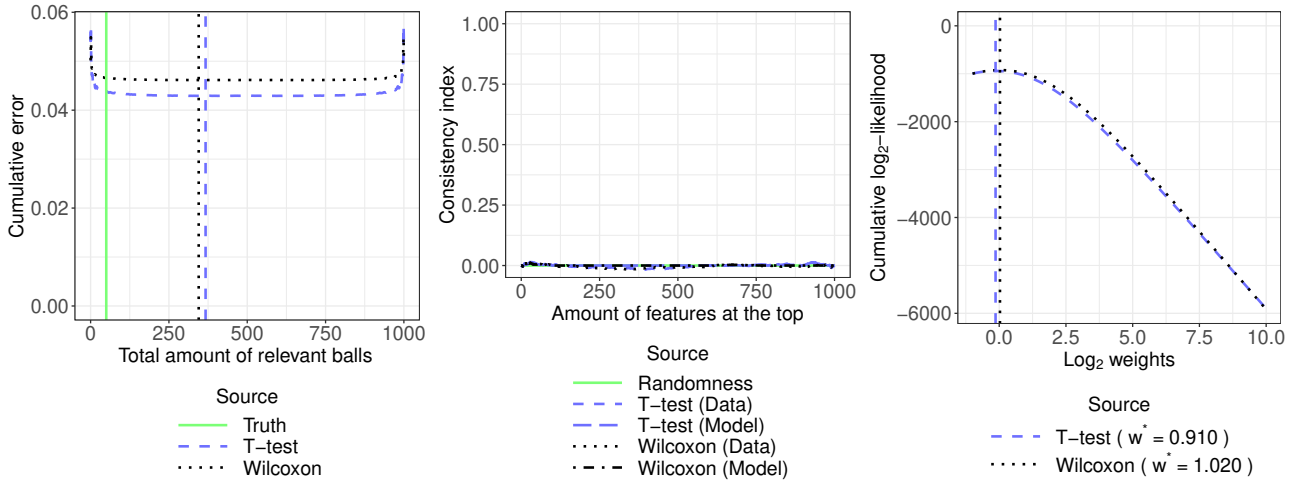
(a) **(b)** **(c)**

Figure 17 Error plot (17a), reproducibility plot (17b) and weight plot (17c) for the difficulty configuration 16, in the differences in location scenario



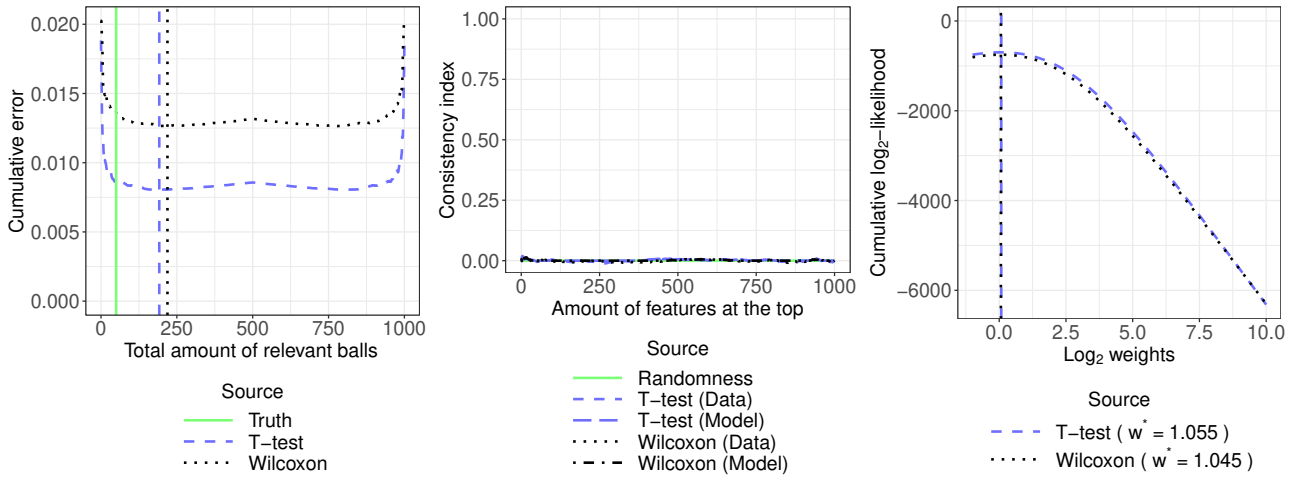
(a) **(b)** **(c)**

Figure 18 Error plot (18a), reproducibility plot (18b) and weight plot (18c) for the difficulty configuration 17, in the differences in location scenario



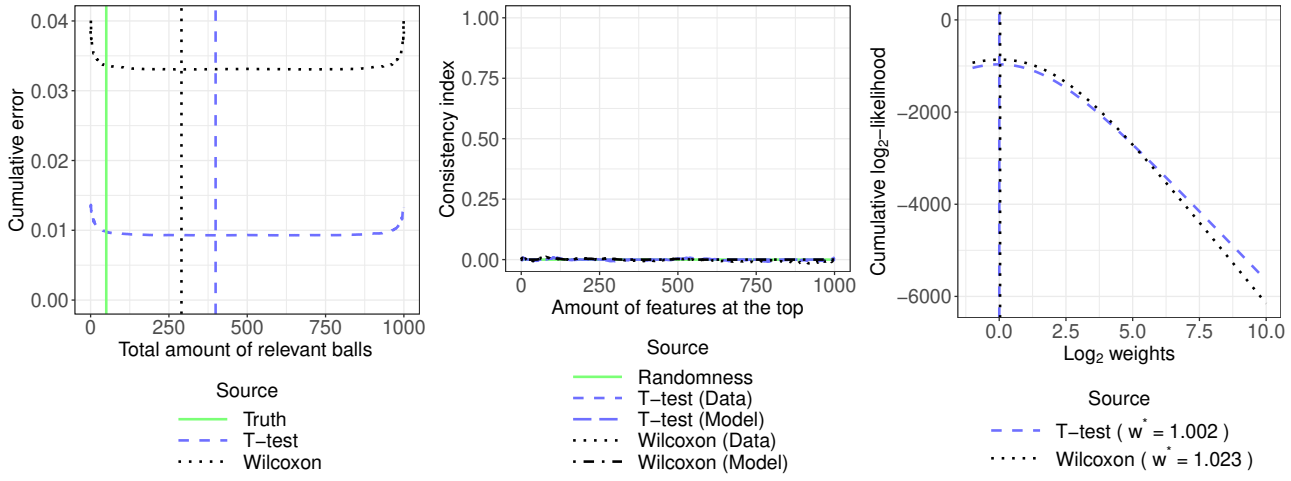
(a) **(b)** **(c)**

Figure 19 Error plot (19a), reproducibility plot (19b) and weight plot (19c) for the difficulty configuration 18, in the differences in location scenario



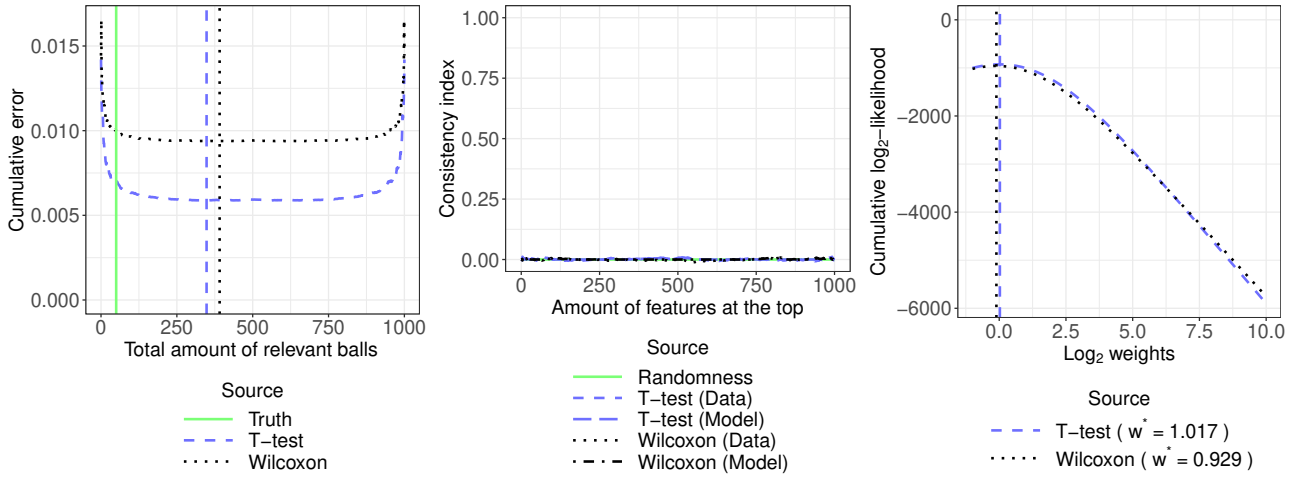
(a) **(b)** **(c)**

Figure 20 Error plot (20a), reproducibility plot (20b) and weight plot (20c) for the difficulty configuration 19, in the differences in location scenario



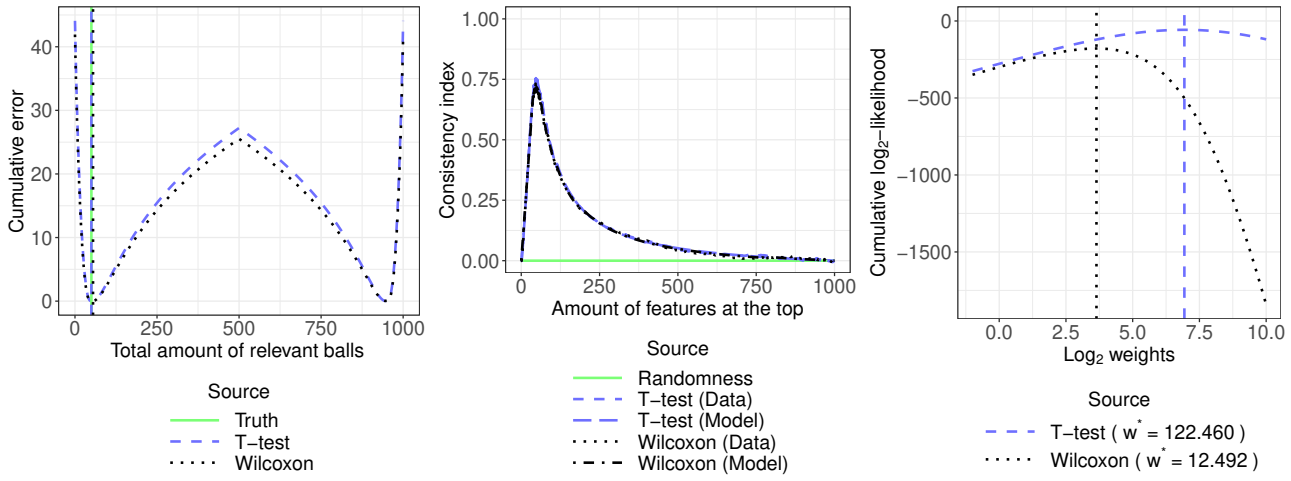
(a) **(b)** **(c)**

Figure 21 Error plot (21a), reproducibility plot (21b) and weight plot (21c) for the difficulty configuration 20, in the differences in location scenario



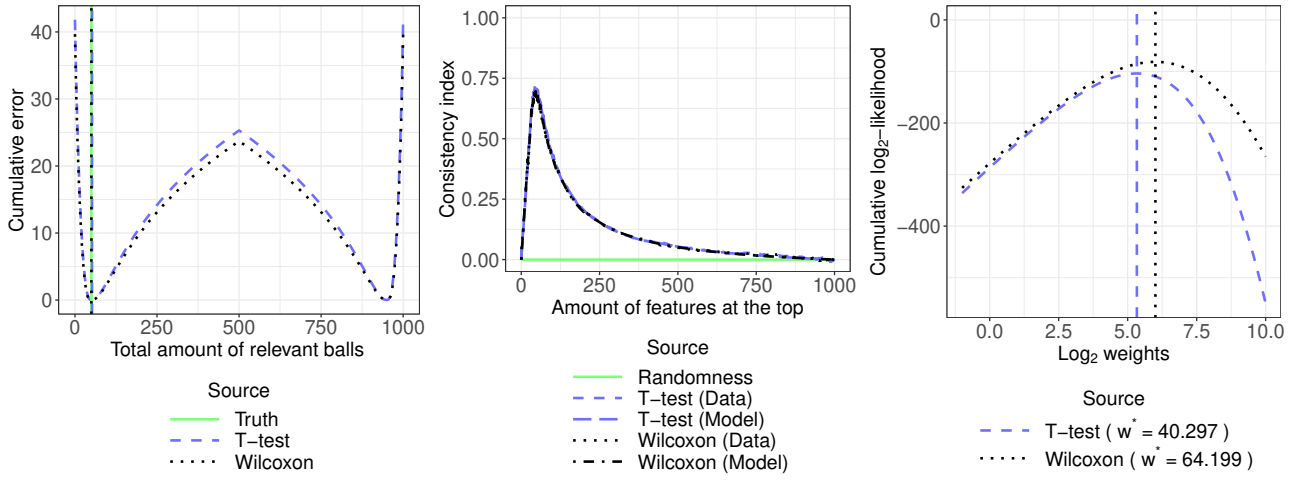
(a) **(b)** **(c)**

Figure 22 Error plot (22a), reproducibility plot (22b) and weight plot (22c) for the difficulty configuration 21, in the differences in location scenario



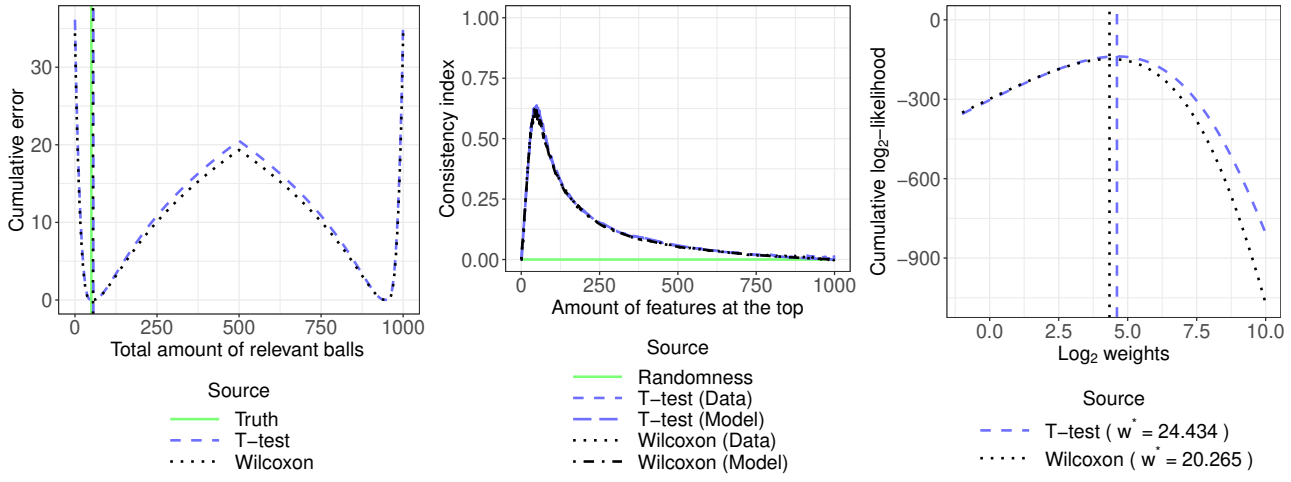
(a) **(b)** **(c)**

Figure 23 Error plot (23a), reproducibility plot (23b) and weight plot (23c) for the difficulty configuration 1, in the differences in both location and spread scenario



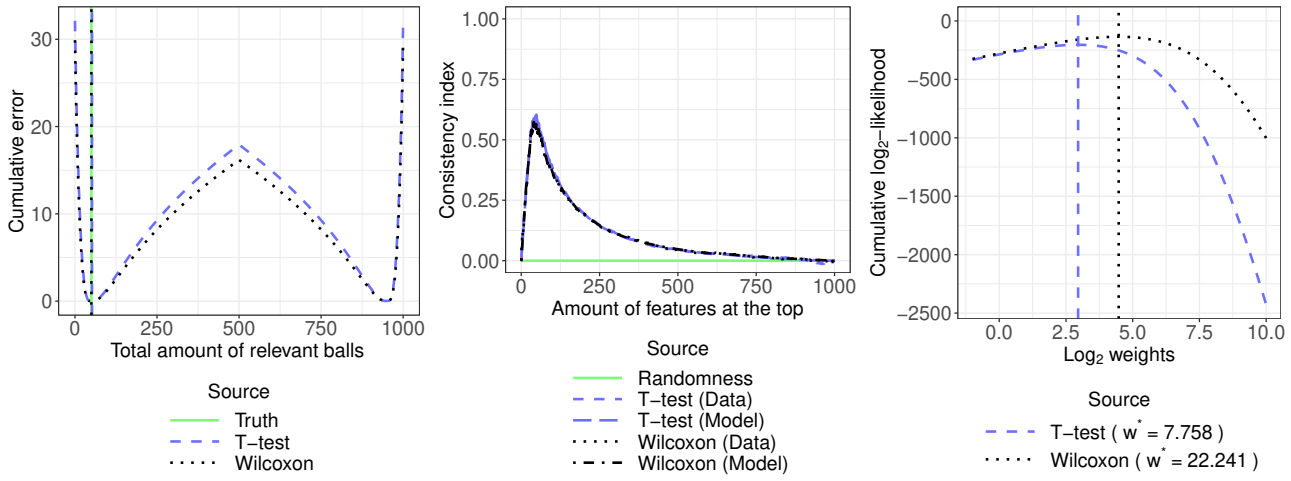
(a) **(b)** **(c)**

Figure 24 Error plot (24a), reproducibility plot (24b) and weight plot (24c) for the difficulty configuration 2, in the differences in both location and spread scenario



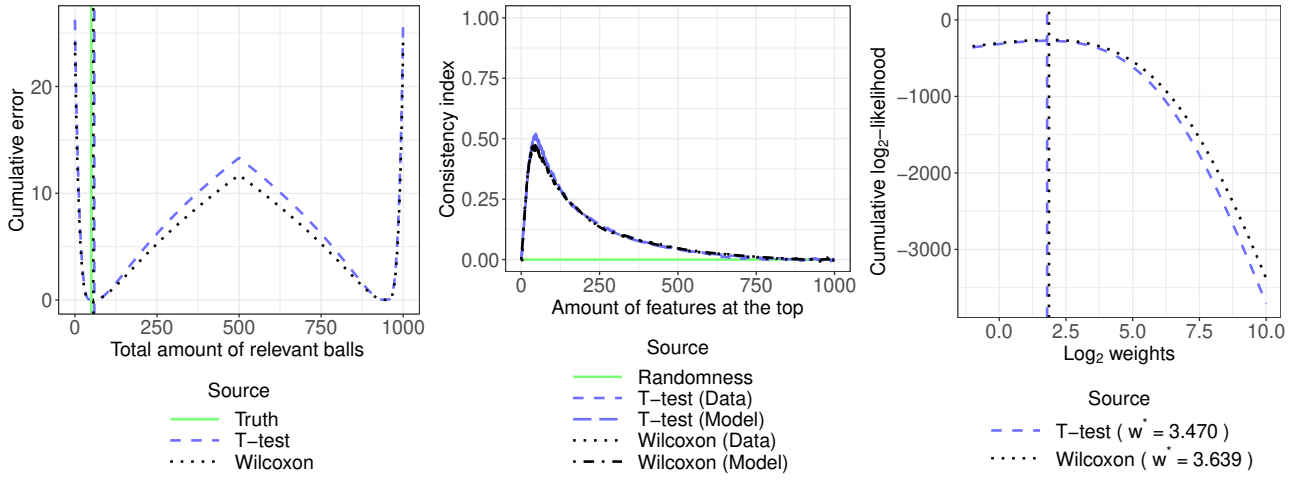
(a) **(b)** **(c)**

Figure 25 Error plot (25a), reproducibility plot (25b) and weight plot (25c) for the difficulty configuration 3, in the differences in both location and spread scenario



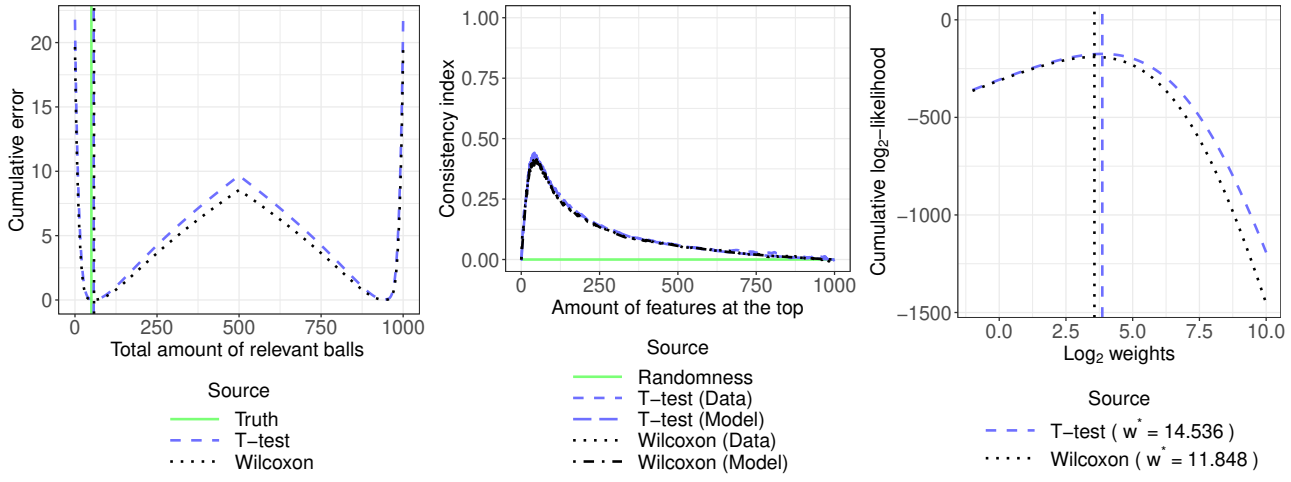
(a) **(b)** **(c)**

Figure 26 Error plot (26a), reproducibility plot (26b) and weight plot (26c) for the difficulty configuration 4, in the differences in both location and spread scenario



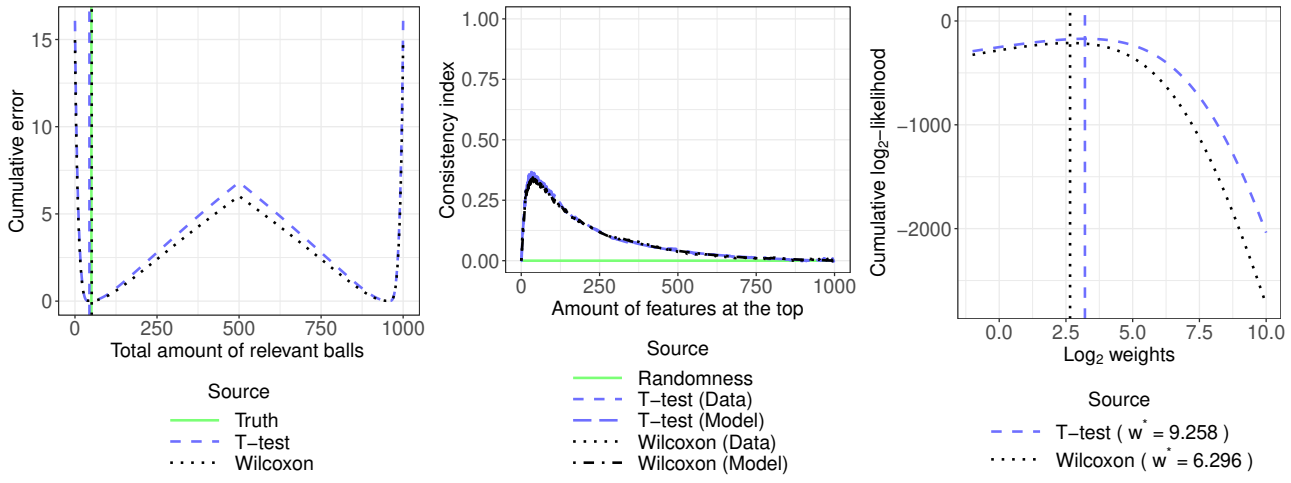
(a) **(b)** **(c)**

Figure 27 Error plot (27a), reproducibility plot (27b) and weight plot (27c) for the difficulty configuration 5, in the differences in both location and spread scenario



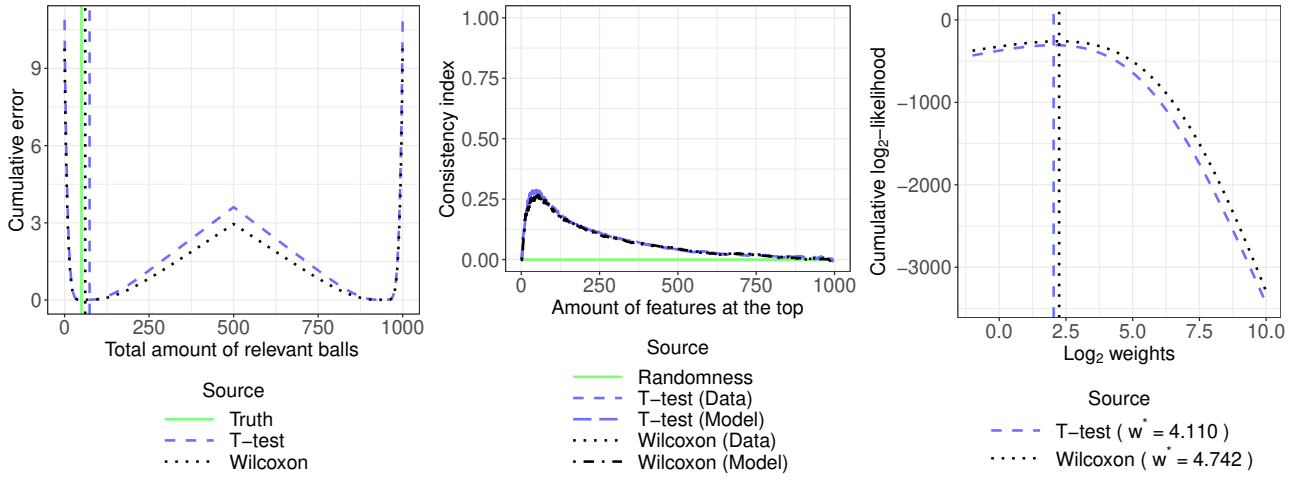
(a) **(b)** **(c)**

Figure 28 Error plot (28a), reproducibility plot (28b) and weight plot (28c) for the difficulty configuration 6, in the differences in both location and spread scenario



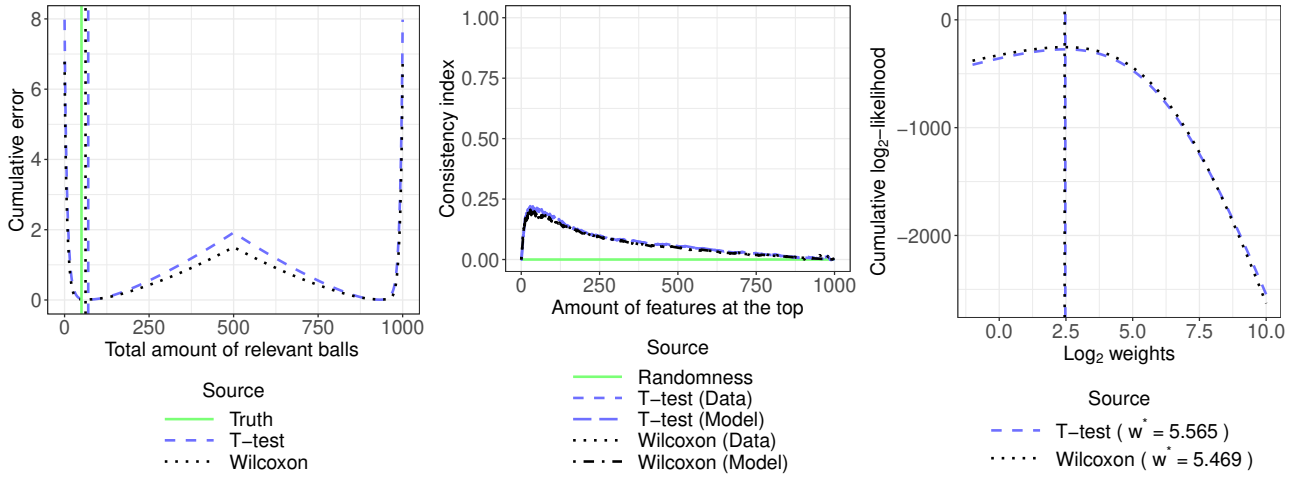
(a) **(b)** **(c)**

Figure 29 Error plot (29a), reproducibility plot (29b) and weight plot (29c) for the difficulty configuration 7, in the differences in both location and spread scenario



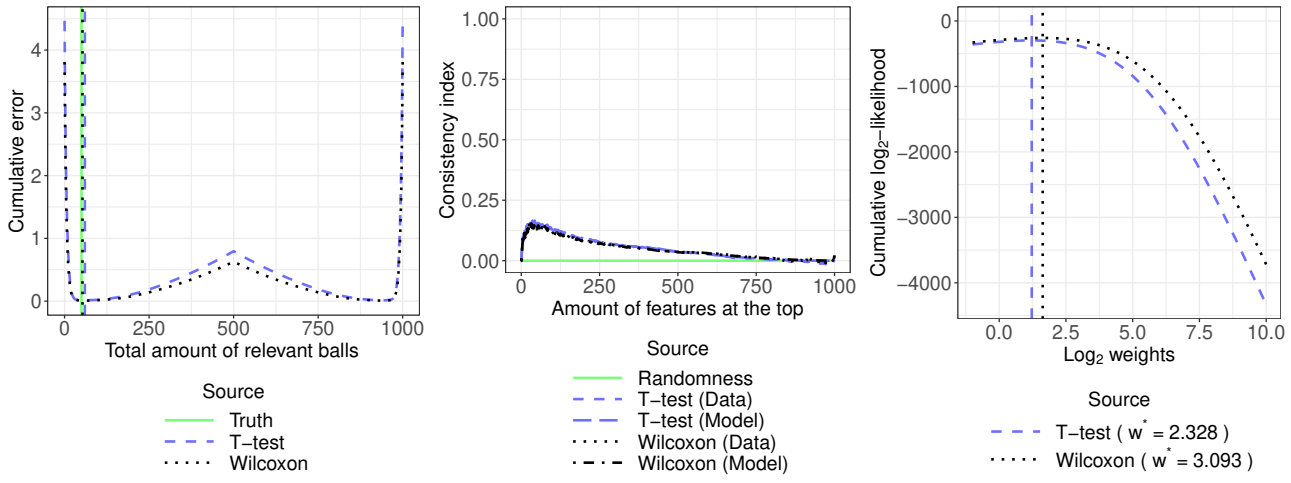
(a) **(b)** **(c)**

Figure 30 Error plot (30a), reproducibility plot (30b) and weight plot (30c) for the difficulty configuration 8, in the differences in both location and spread scenario



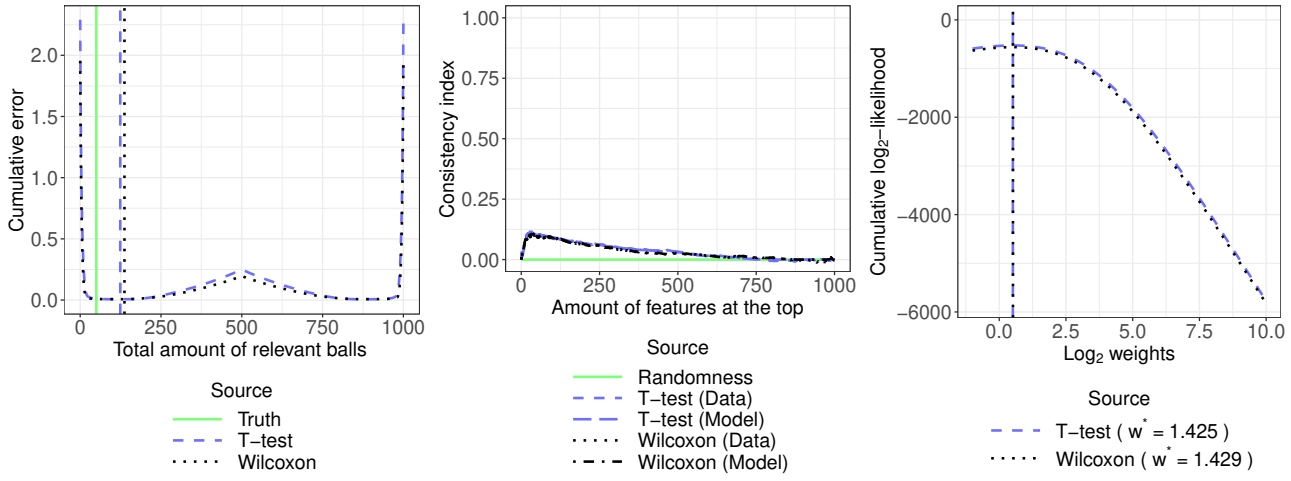
(a) **(b)** **(c)**

Figure 31 Error plot (31a), reproducibility plot (31b) and weight plot (31c) for the difficulty configuration 9, in the differences in both location and spread scenario



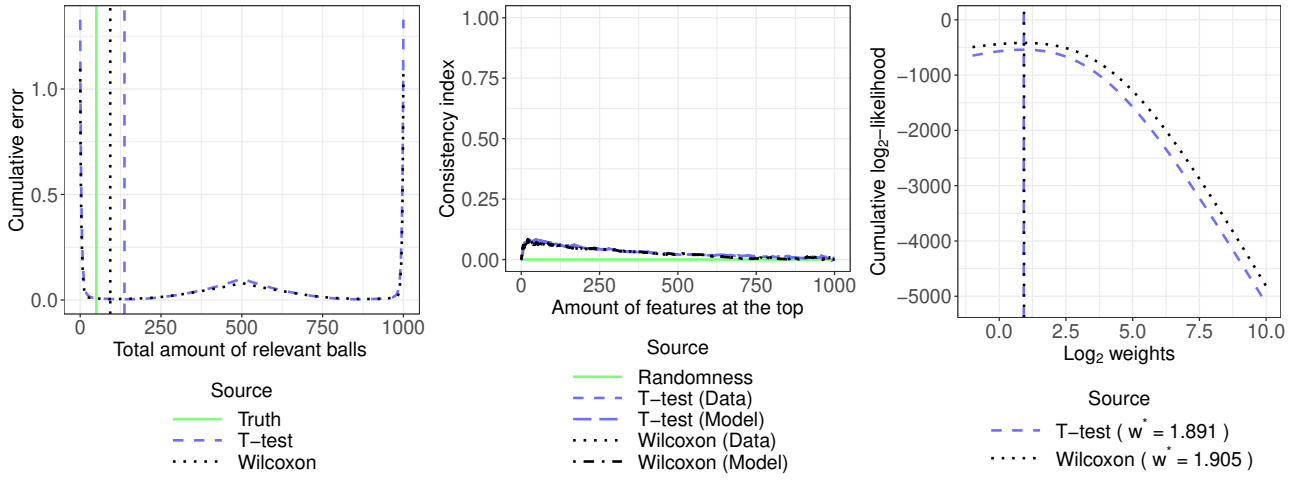
(a) **(b)** **(c)**

Figure 32 Error plot (32a), reproducibility plot (32b) and weight plot (32c) for the difficulty configuration 10, in the differences in both location and spread scenario



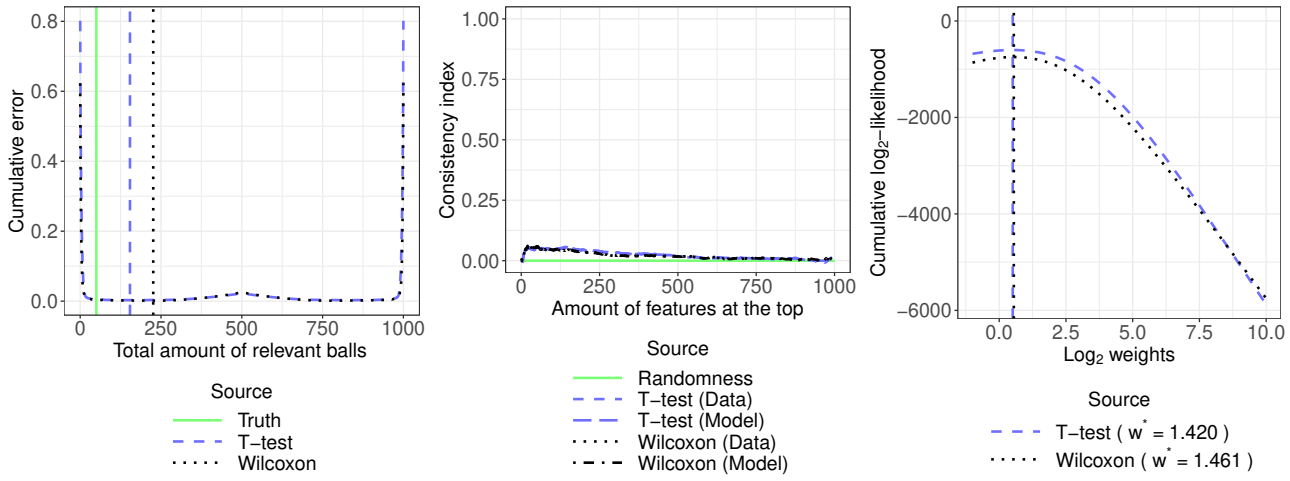
(a) **(b)** **(c)**

Figure 33 Error plot (33a), reproducibility plot (33b) and weight plot (33c) for the difficulty configuration 11, in the differences in both location and spread scenario



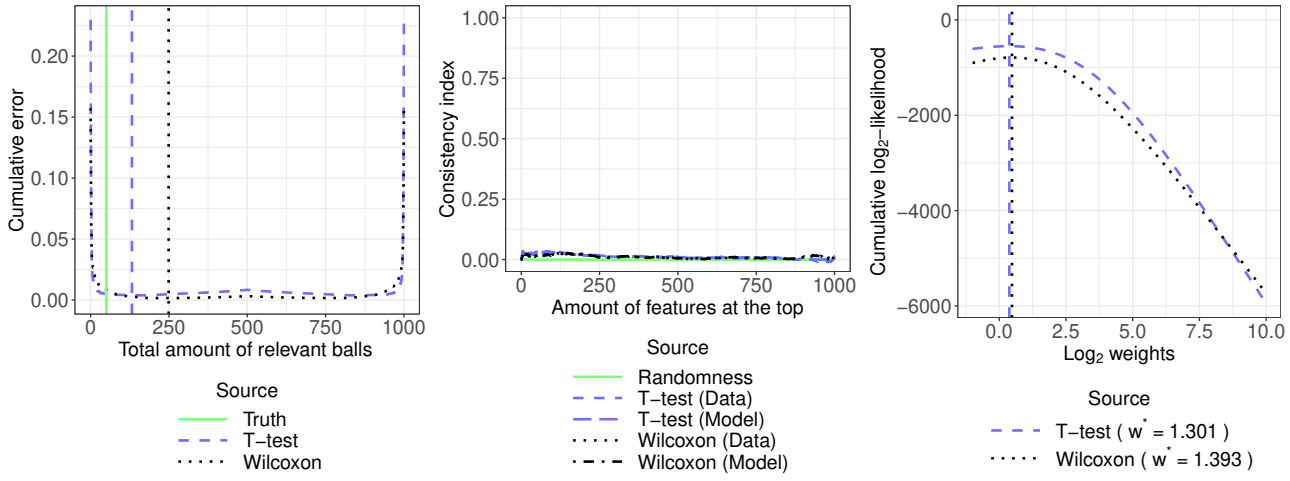
(a) **(b)** **(c)**

Figure 34 Error plot (34a), reproducibility plot (34b) and weight plot (34c) for the difficulty configuration 12, in the differences in both location and spread scenario



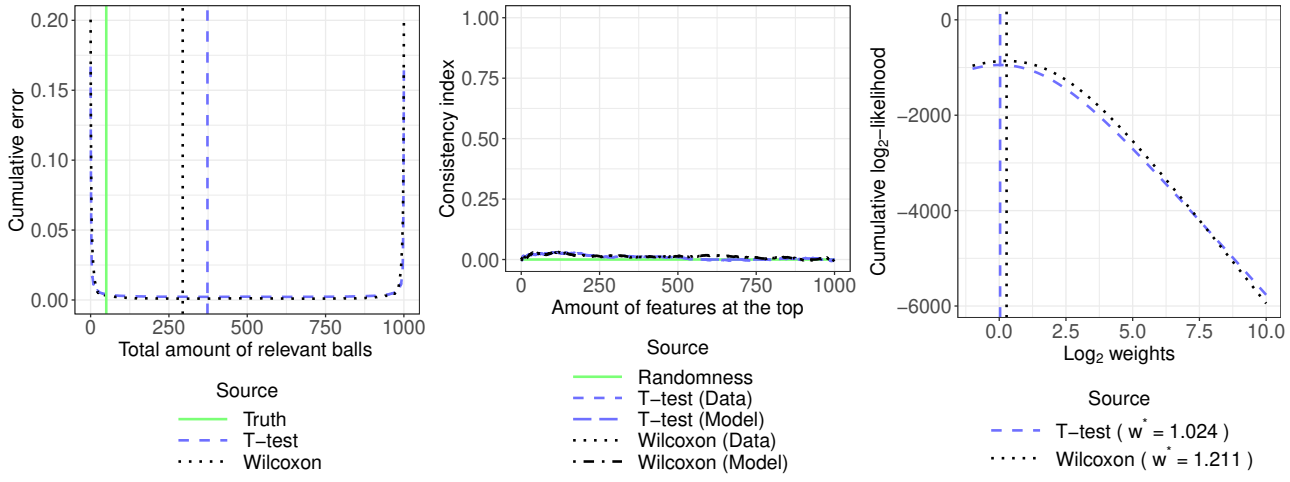
(a) **(b)** **(c)**

Figure 35 Error plot (35a), reproducibility plot (35b) and weight plot (35c) for the difficulty configuration 13, in the differences in both location and spread scenario



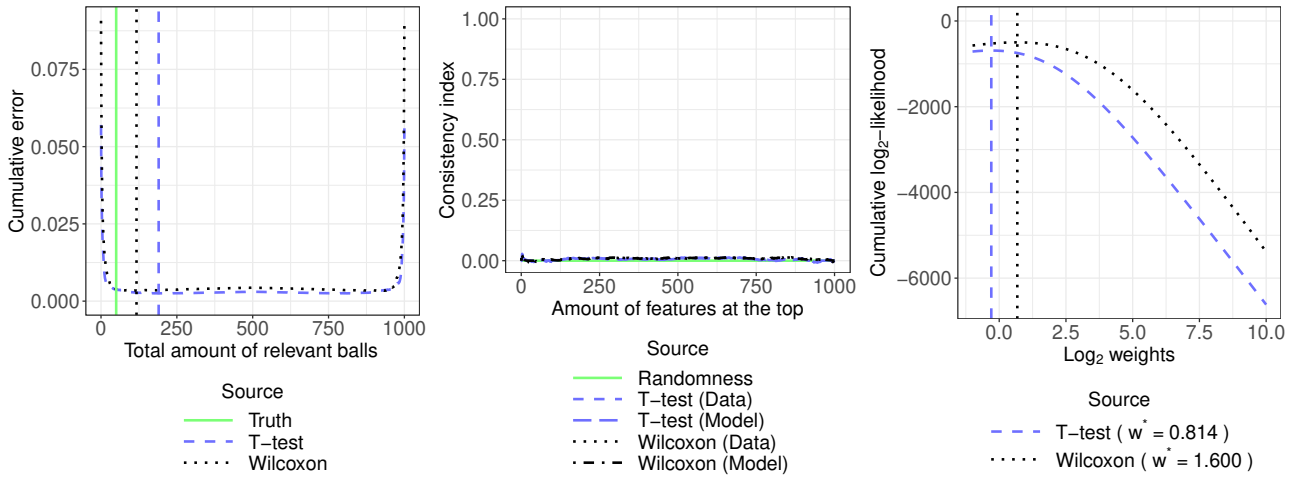
(a) **(b)** **(c)**

Figure 36 Error plot (36a), reproducibility plot (36b) and weight plot (36c) for the difficulty configuration 14, in the differences in both location and spread scenario



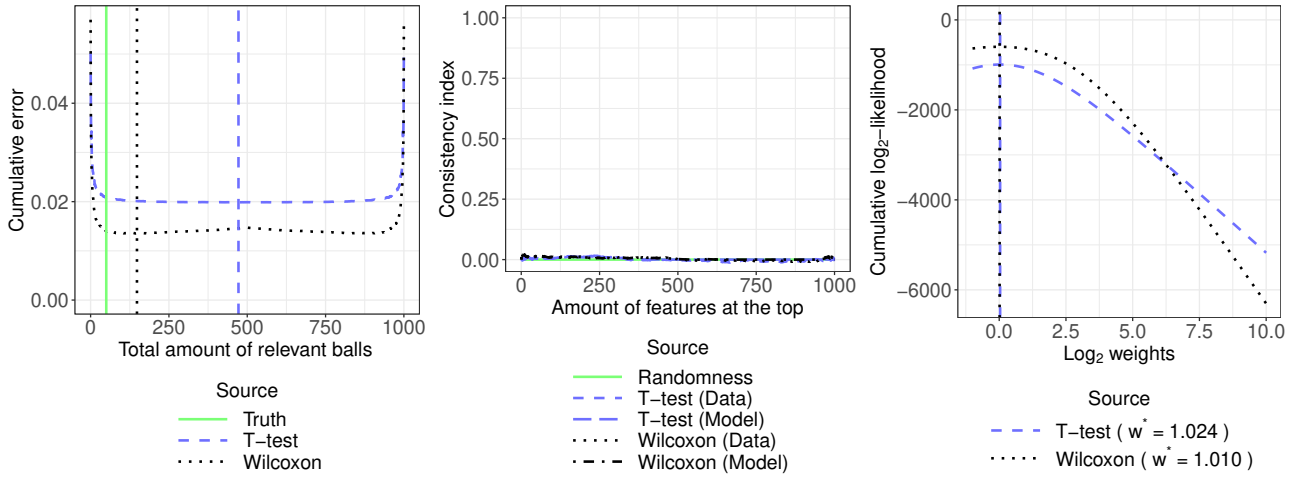
(a) **(b)** **(c)**

Figure 37 Error plot (37a), reproducibility plot (37b) and weight plot (37c) for the difficulty configuration 15, in the differences in both location and spread scenario



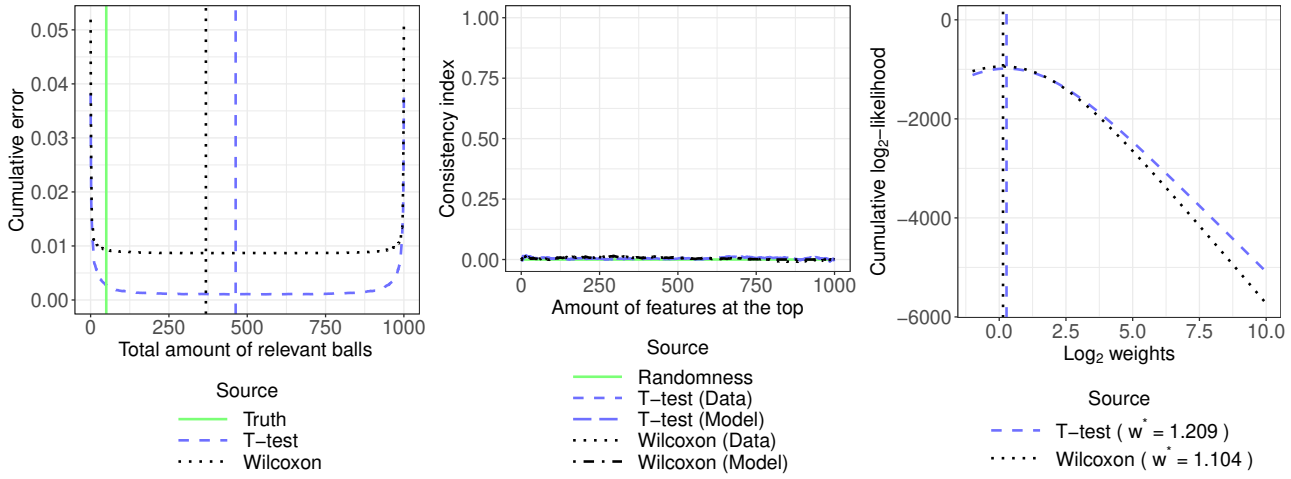
(a) **(b)** **(c)**

Figure 38 Error plot (38a), reproducibility plot (38b) and weight plot (38c) for the difficulty configuration 16, in the differences in both location and spread scenario



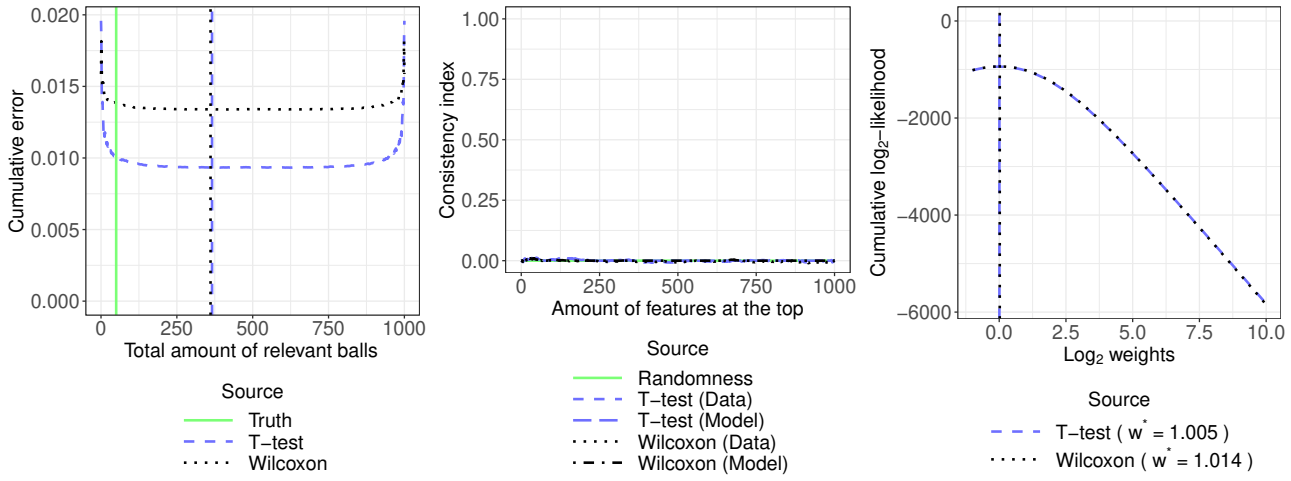
(a) **(b)** **(c)**

Figure 39 Error plot (39a), reproducibility plot (39b) and weight plot (39c) for the difficulty configuration 17, in the differences in both location and spread scenario



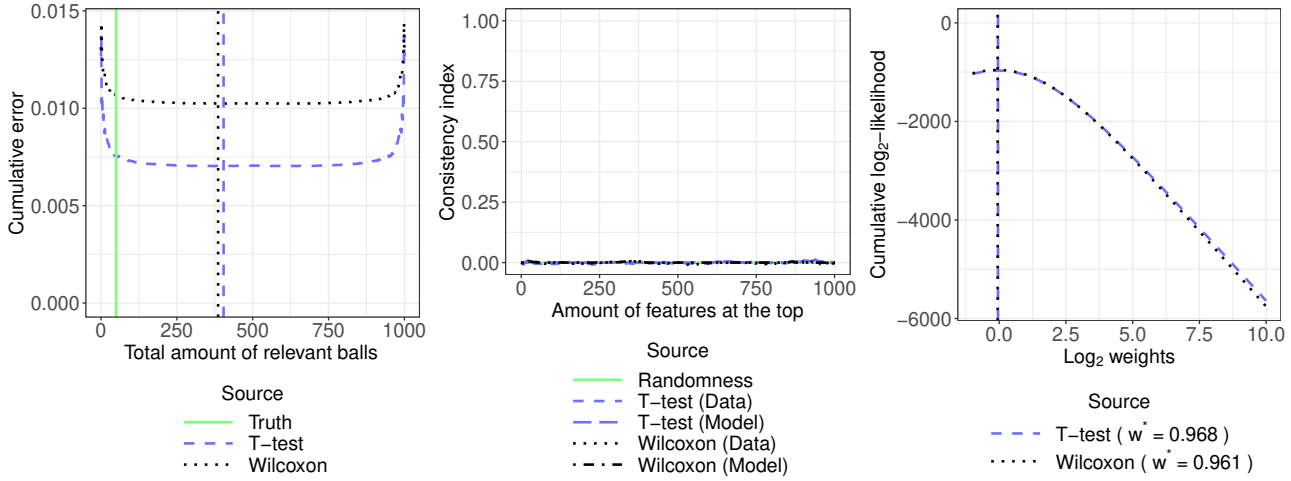
(a) **(b)** **(c)**

Figure 40 Error plot (40a), reproducibility plot (40b) and weight plot (40c) for the difficulty configuration 18, in the differences in both location and spread scenario



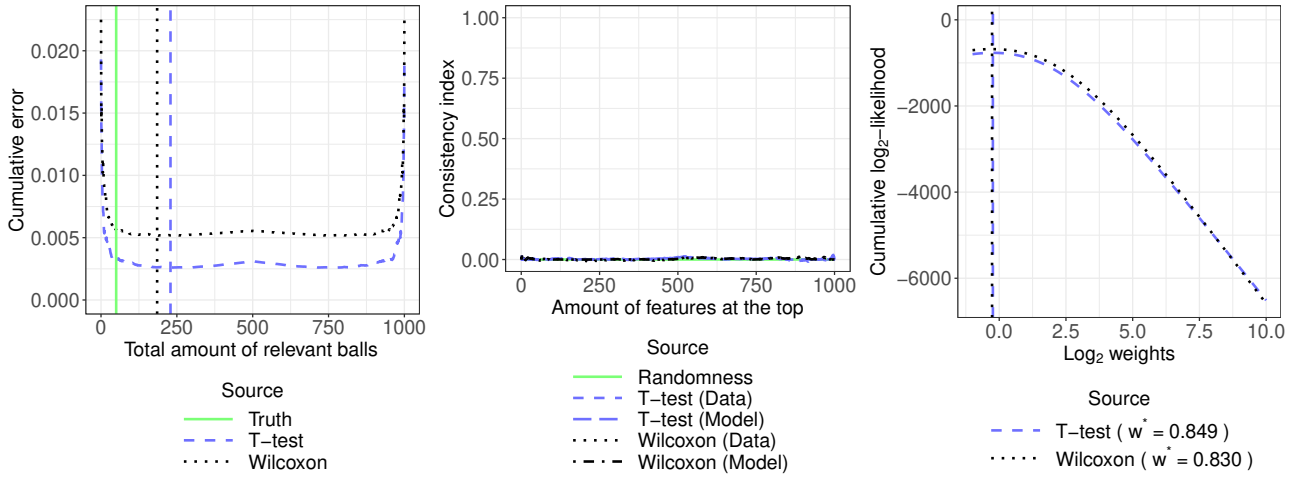
(a) **(b)** **(c)**

Figure 41 Error plot (41a), reproducibility plot (41b) and weight plot (41c) for the difficulty configuration 19, in the differences in both location and spread scenario



(a) **(b)** **(c)**

Figure 42 Error plot (42a), reproducibility plot (42b) and weight plot (42c) for the difficulty configuration 20, in the differences in both location and spread scenario



(a) **(b)** **(c)**

Figure 43 Error plot (43a), reproducibility plot (43b) and weight plot (43c) for the difficulty configuration 21, in the differences in both location and spread scenario

Table 2 w^* and AUC values when the relevant features show differences only in location

Difficulty	Method	w^*	Average data AUC	Model AUC
1	T-test	11.518	0.99148	0.97250
1	Wilcoxon test	5.849	0.98946	0.95259
2	T-test	53.975	0.98500	0.98882
2	Wilcoxon test	14.604	0.98267	0.97205
3	T-test	16.069	0.98233	0.96627
3	Wilcoxon test	12.418	0.97970	0.95985
4	T-test	6.450	0.97515	0.93749
4	Wilcoxon test	12.113	0.97090	0.95435
5	T-test	13.050	0.96332	0.95136
5	Wilcoxon test	8.872	0.95915	0.93693
6	T-test	7.953	0.94897	0.92738
6	Wilcoxon test	6.647	0.94338	0.91295
7	T-test	4.642	0.93835	0.89594
7	Wilcoxon test	2.787	0.93154	0.87626
8	T-test	5.345	0.91565	0.88733
8	Wilcoxon test	3.016	0.90990	0.85423
9	T-test	4.595	0.89733	0.85790
9	Wilcoxon test	3.485	0.88775	0.82226
10	T-test	2.240	0.85830	0.76064
10	Wilcoxon test	1.962	0.85010	0.74880
11	T-test	1.386	0.83564	0.69693
11	Wilcoxon test	1.616	0.82877	0.70229
12	T-test	1.275	0.79122	0.63219
12	Wilcoxon test	1.363	0.78047	0.64459
13	T-test	1.397	0.74516	0.66799
13	Wilcoxon test	1.599	0.73620	0.65572
14	T-test	1.398	0.71663	0.65058
14	Wilcoxon test	1.187	0.70986	0.62836
15	T-test	1.036	0.65179	0.54314
15	Wilcoxon test	0.980	0.64477	0.54778
16	T-test	0.967	0.61097	0.52834
16	Wilcoxon test	1.104	0.60878	0.56503
17	T-test	0.978	0.58113	0.51393
17	Wilcoxon test	0.982	0.57768	0.51062
18	T-test	0.910	0.54310	0.50799
18	Wilcoxon test	1.020	0.54333	0.50883
19	T-test	1.055	0.51729	0.52850
19	Wilcoxon test	1.045	0.51828	0.51656
20	T-test	1.002	0.50171	0.50972
20	Wilcoxon test	1.023	0.50197	0.51516
21	T-test	1.017	0.50194	0.52346
21	Wilcoxon test	0.929	0.50303	0.50471

Table 3 w^* and AUC values when the relevant features show differences both in location and spread

Difficulty	Method	w^*	Average data AUC	Model AUC
1	T-test	122.460	0.99407	0.99536
1	Wilcoxon test	12.492	0.99164	96449
2	T-test	40.297	0.99137	0.98715
2	Wilcoxon test	64.199	0.98905	0.99124
3	T-test	24.434	0.98295	0.97447
3	Wilcoxon test	20.265	0.97917	0.97005
4	T-test	7.758	0.97871	0.95781
4	Wilcoxon test	22.241	0.97396	0.97343
5	T-test	3.470	0.96880	0.90336
5	Wilcoxon test	3.639	0.96373	0.9211
6	T-test	14.536	0.95099	0.95008
6	Wilcoxon test	11.848	0.94507	0.93907
7	T-test	9.258	0.93612	0.94695
7	Wilcoxon test	6.296	0.92903	0.91514
8	T-test	4.110	0.91651	0.84912
8	Wilcoxon test	4.742	0.90731	0.87484
9	T-test	5.565	0.88543	0.86763
9	Wilcoxon test	5.469	0.87457	0.86828
10	T-test	2.328	0.85350	0.81743
10	Wilcoxon test	3.093	0.84282	0.84080
11	T-test	1.425	0.81481	0.6842
11	Wilcoxon test	1.429	0.80419	0.67806
12	T-test	1.891	0.76569	0.68487
12	Wilcoxon test	1.905	0.75853	0.69762
13	T-test	1.420	0.73233	0.65252
13	Wilcoxon test	1.461	0.72517	0.62383
14	T-test	1.301	0.68675	0.61855
14	Wilcoxon test	1.393	0.67817	0.58700
15	T-test	1.024	0.64637	0.54806
15	Wilcoxon test	1.211	0.64123	0.59103
16	T-test	0.814	0.60370	0.57824
16	Wilcoxon test	1.600	0.59955	0.61535
17	T-test	1.024	0.57310	0.52380
17	Wilcoxon test	1.010	0.56957	0.55216
18	T-test	1.209	0.53356	0.55400
18	Wilcoxon test	1.104	0.53449	0.55425
19	T-test	1.005	0.51847	0.50532
19	Wilcoxon test	1.014	0.51820	0.50659
20	T-test	0.968	0.50659	0.50427
20	Wilcoxon test	0.961	0.50810	0.50814
21	T-test	0.849	0.49810	0.54628
21	Wilcoxon test	0.830	0.49794	0.53650

- Mice database [2]: This dataset contains 77 features and 1080 instances. The features consist of protein expression levels of 77 proteins, while the instances correspond to different measurements in 38 control mice and 34 mice with Down's syndrome (multiple measurements of each protein were carried out for each mouse).
- SECOM database [4]: This dataset has 591 features and 1567 instances. Briefly, this dataset is a semiconductor manufacturing process dataset in which the features correspond to process signals. Regarding the instances, each of them corresponds to a different production entity, the instances being divided into 1463 correct productions and 104 faulty productions.
- Arcene database [1]: This dataset includes 10000 features and 900 instances. Specifically, this dataset consists of mass-spectrometric data of biological samples. In particular, 7000 features are measurements derived from the biological samples, while the remaining 3000 features are non-real features added as a distracting factor. Besides, the 900 instances consists of individuals that can be divided into two groups, 398 patients with cancer and 502 healthy individuals.
- Ovarian database [6]: This dataset contains 27578 features and 540 instances. The features consists of β-values that denote the methylation level of different CpG-sites on the peripheral whole blood of 540 postmenopausal women. The instances of this dataset correspond to 274 healthy women, 131 women with ovarian cancer yet to be treated and 135 women with ovarian cancer already treated, i.e., 274 controls and 266 cases. The groups are age-matched and the range of ages covered by the database is from 49 to 91 years.

6 Preprocessing of the real databases

We divide this section into two subsections. The first one is dedicated to the preprocessing applied to the databases extracted from the UCI repository. The second one is dedicated to the preprocessing applied to the ovarian cancer database.

6.1 UCI repository databases preprocessing

The preprocessing applied to the selected databases from the UCI repository consists of the following step:

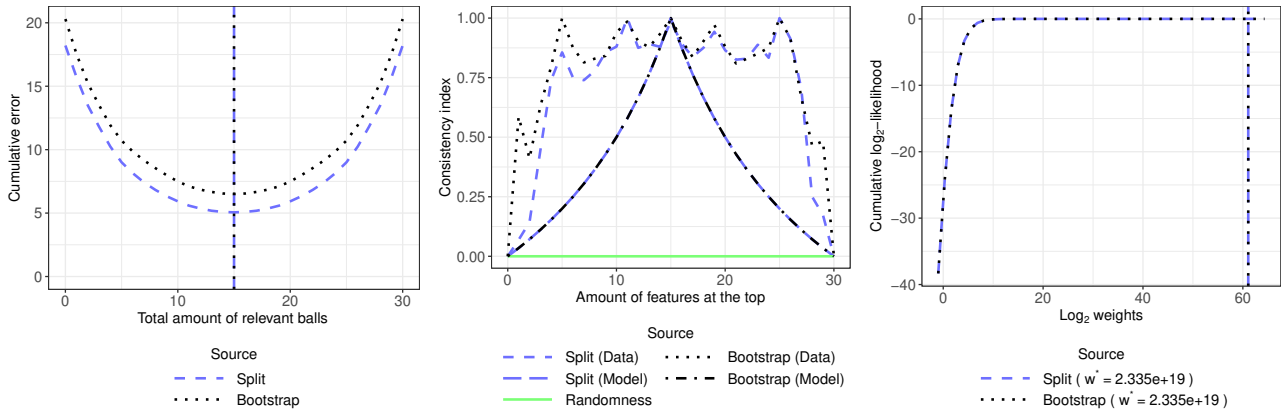
1. The CpG sites that have missing values for more than 50% of the individuals are removed[1].

6.2 Ovarian cancer database preprocessing

The preprocessing we applied to the ovarian cancer database is based on what was done by Wang et al [7]. The preprocessing consists of applying the following steps sequentially to the matrix of β-values available in the GEO database:

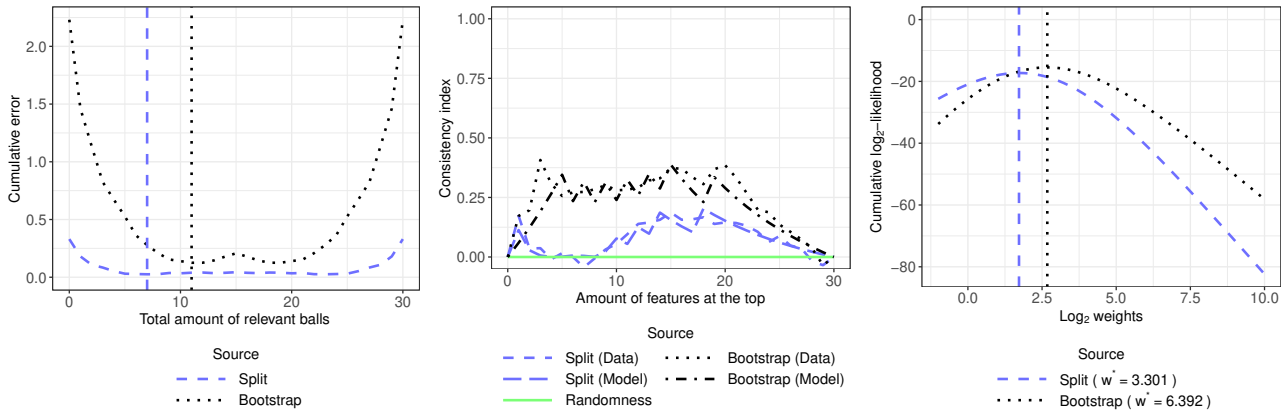
1. Among all the ovarian cancer cases, only those who gave their blood at the time of their diagnosis prior to treatment have been used.
2. Samples whose bisulfite conversion efficiencies are too low (< 4000) have been removed.
3. Data from batches 10-12 have been removed.
4. In order to remove outlier samples, for each sample, we have computed the average of the distances of all the values (using raw total intensities) of its CpG sites regarding their median values across samples. Then, all the generated averages are used to calculate the range $(Q_1 - 1.5 \cdot \text{IQR}, Q_3 + 1.5 \cdot \text{IQR})$. Finally, all those samples whose averages are not within that range are removed.
5. All those individuals that do not cover at least 95% of the CpG sites with a detection p-value smaller than 0.05 are removed.
6. All the CpG sites whose detection p-values are not below 0.05 in all samples are removed.
7. All the CpG sites that do not have numeric values (e.g., NA values) for at least 50 individuals per group are removed.
8. The CpG sites that have missing values for more than 99.55% of the individuals are removed[2].

[1] This preprocessing step was included in order to enable the computation of the ranking method based on the coefficients of a linear SVM.
[2] This preprocessing step was included in order to enable the computation of the ranking method based on the coefficients of a linear SVM.



(a) **(b)** **(c)**

Figure 44 Error plot (44a), reproducibility plot (44b) and weight plot (44c) when the ranking algorithm based on the mutual information is applied to the breast cancer database



(a) **(b)** **(c)**

Figure 45 Error plot (45a), reproducibility plot (45b) and weight plot (45c) when the ranking algorithm based on the linear SVM is applied to the breast cancer database

7 Ovarian cancer database stratification

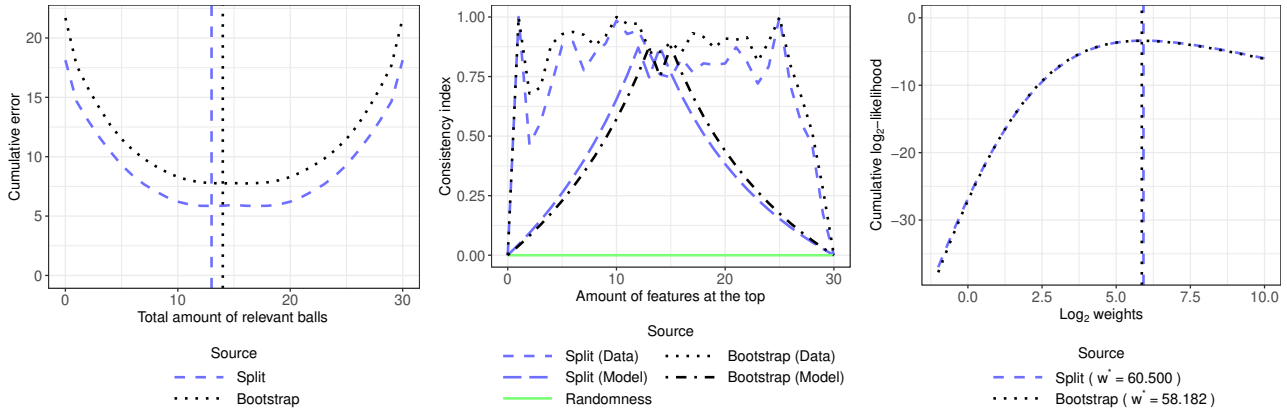
In the following lines, the stratification procedure is explained for the different sampling procedures:

- Splitting in halves: In this sampling procedure, for each group, the individuals are sorted according to their age. Then, for each group, each of its odd individuals is assigned randomly either to belong to $\boldsymbol{D}^{(1)}$ or to belong to $\boldsymbol{D}^{(2)}$. Finally, for each group, each of its even individuals is assigned to $\boldsymbol{D}^{(1)}$ if its previous odd individual was assigned to $\boldsymbol{D}^{(2)}$ or is assigned to $\boldsymbol{D}^{(2)}$ if its previous odd individual was assigned to $\boldsymbol{D}^{(1)}$.
- Bootstrapping: In this sampling procedure, the individuals for both $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$ are directly randomly sampled with replacement from $\boldsymbol{D}$.

8 Plots and tables of the experimentation with real data

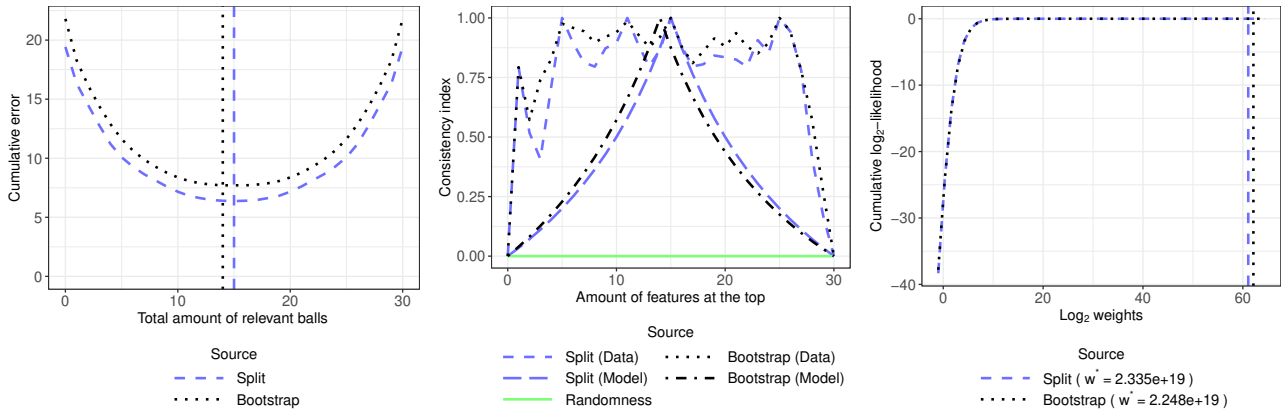
In Figures 44 to 63 the plots and tables of the experimentation with real data can be seen.

In Tables 4 and 5 the weights and AUC values of the experimentation with real data can be seen.



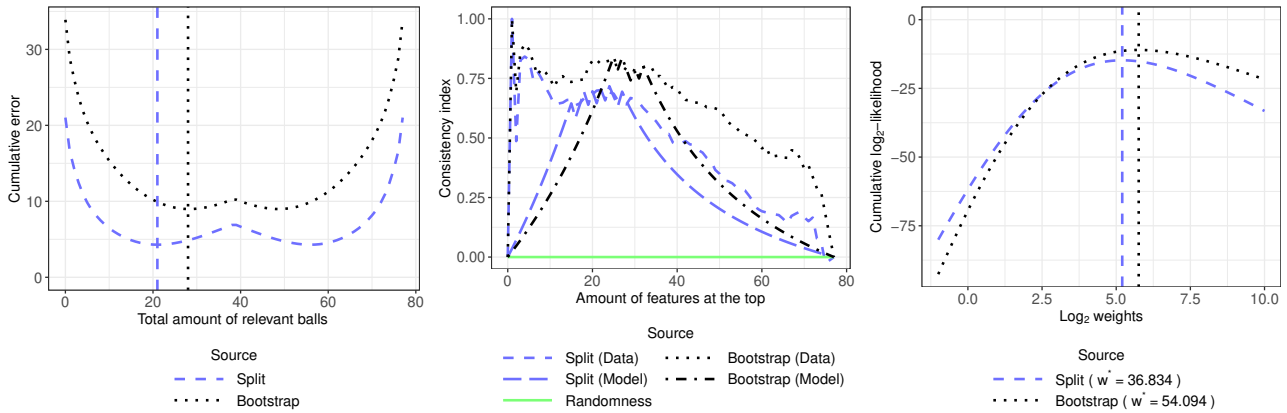
(a) **(b)** **(c)**

Figure 46 Error plot (46a), reproducibility plot (46b) and weight plot (46c) when the ranking algorithm based on the T-test is applied to the breast cancer database



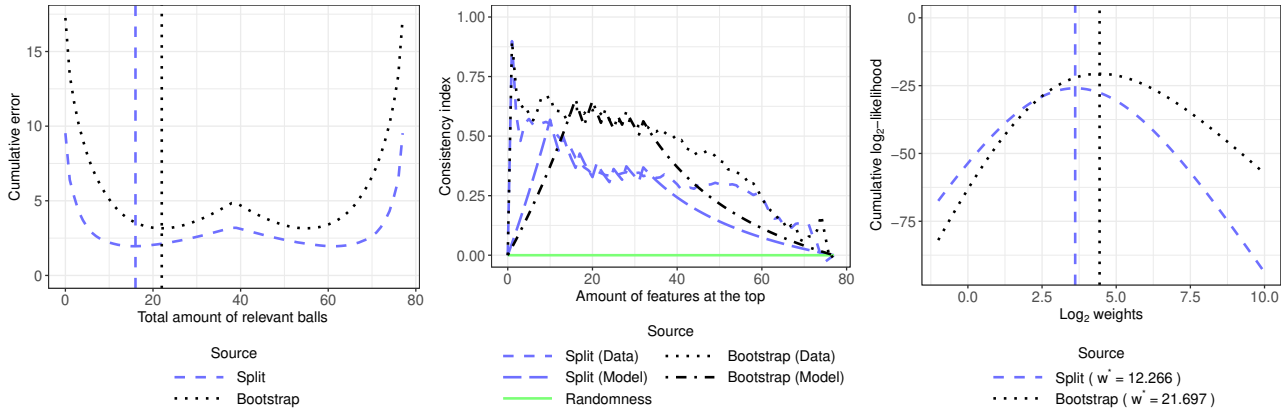
(a) **(b)** **(c)**

Figure 47 Error plot (47a), reproducibility plot (47b) and weight plot (47c) when the ranking algorithm based on the Wilcoxon test is applied to the breast cancer database



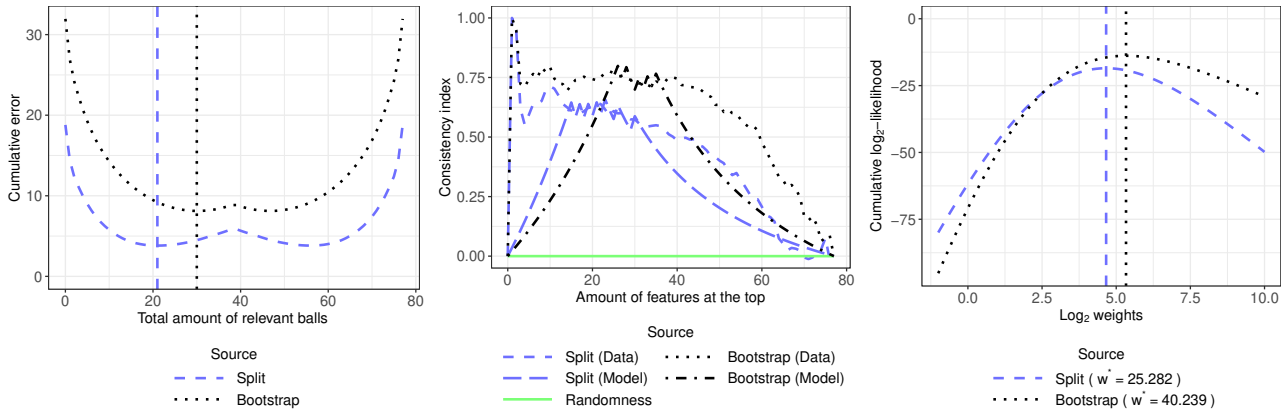
(a) **(b)** **(c)**

Figure 48 Error plot (48a), reproducibility plot (48b) and weight plot (48c) when the ranking algorithm based on the mutual information is applied to the mice database



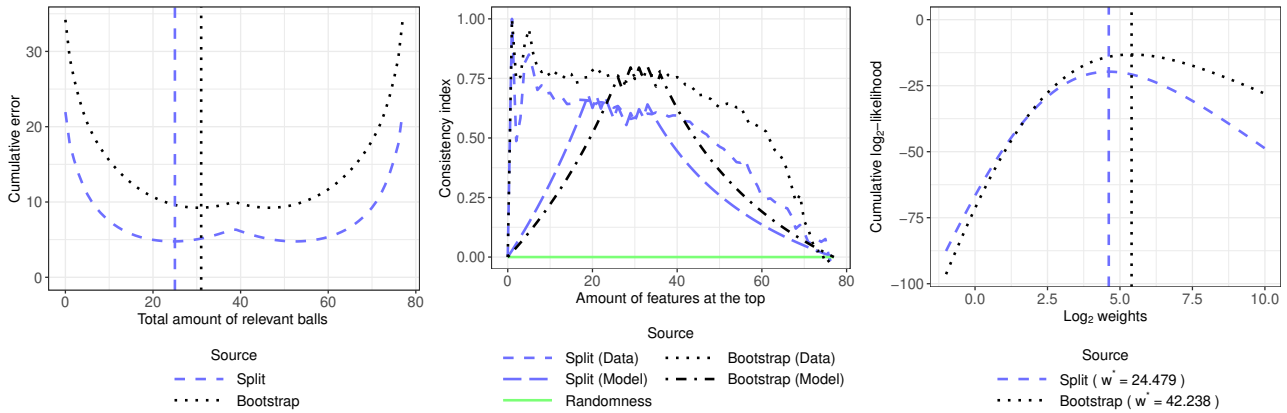
(a) **(b)** **(c)**

Figure 49 Error plot (49a), reproducibility plot (49b) and weight plot (49c) when the ranking algorithm based on the linear SVM is applied to the mice database



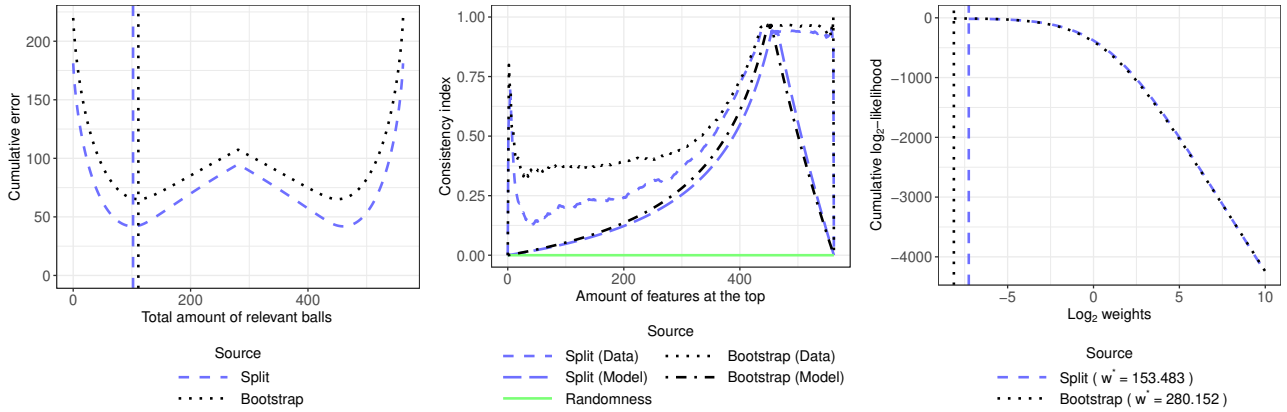
(a) **(b)** **(c)**

Figure 50 Error plot (50a), reproducibility plot (50b) and weight plot (50c) when the ranking algorithm based on the T-test is applied to the mice database



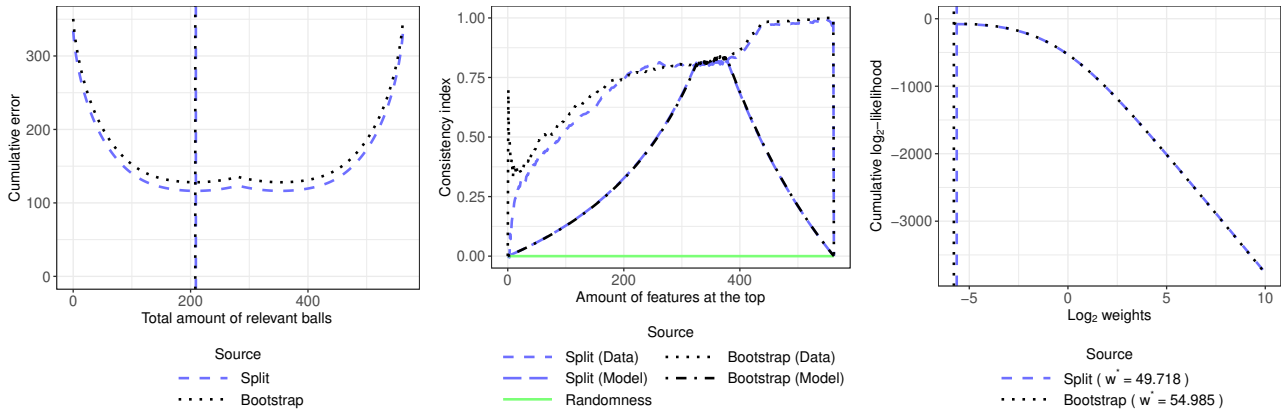
(a) **(b)** **(c)**

Figure 51 Error plot (51a), reproducibility plot (51b) and weight plot (51c) when the ranking algorithm based on the Wilcoxon test is applied to the mice database



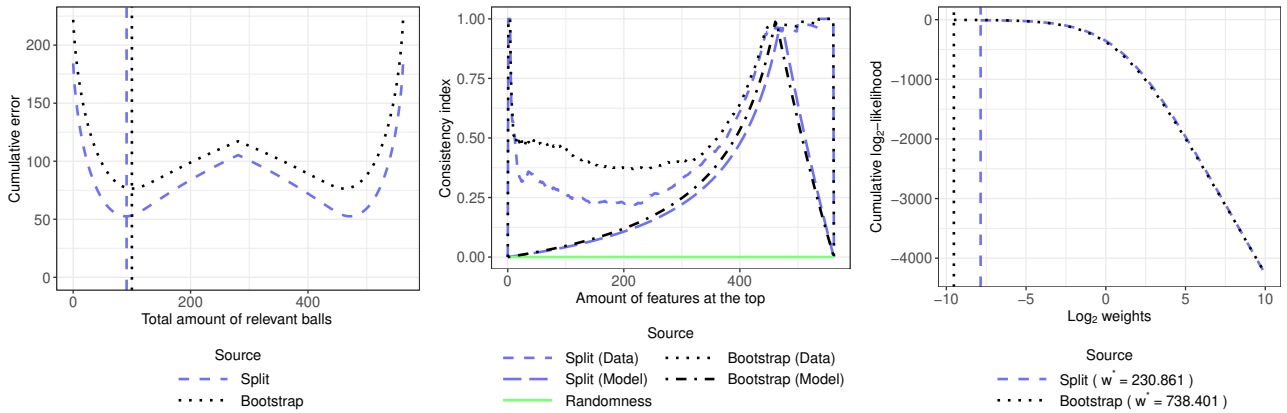
(a) **(b)** **(c)**

Figure 52 Error plot (52a), reproducibility plot (52b) and weight plot (52c) when the ranking algorithm based on the mutual information is applied to the SECOM database



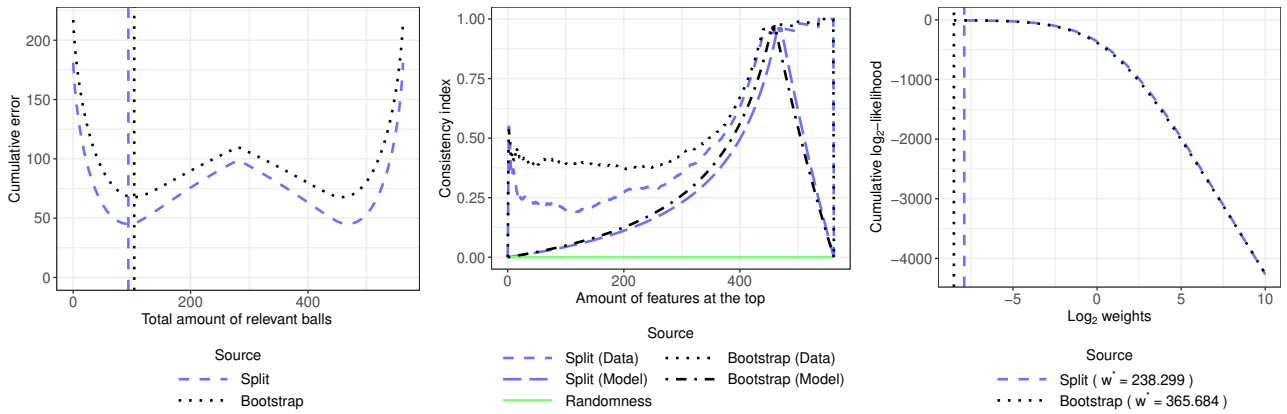
(a) **(b)** **(c)**

Figure 53 Error plot (53a), reproducibility plot (53b) and weight plot (53c) when the ranking algorithm based on the linear SVM is applied to the SECOM database



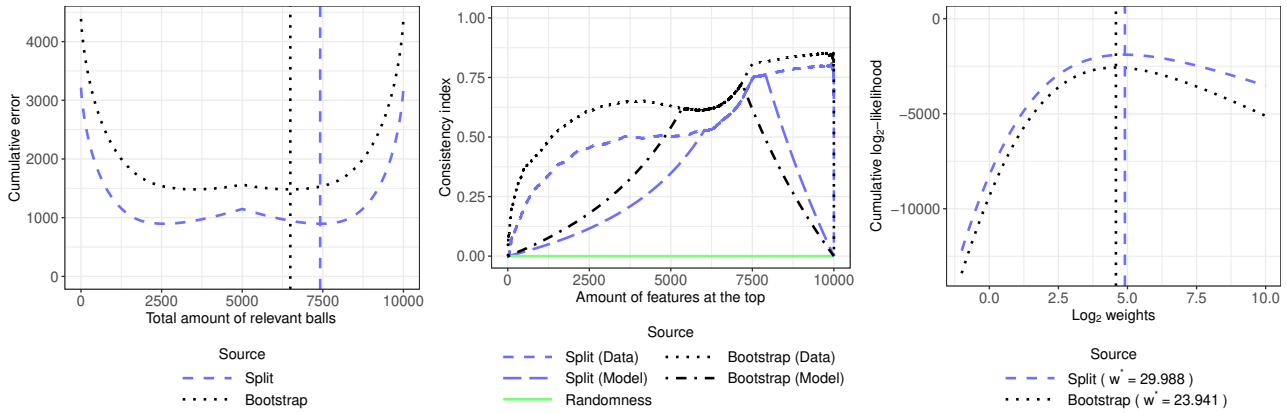
(a) **(b)** **(c)**

Figure 54 Error plot (54a), reproducibility plot (54b) and weight plot (54c) when the ranking algorithm based on the T-test is applied to the SECOM database



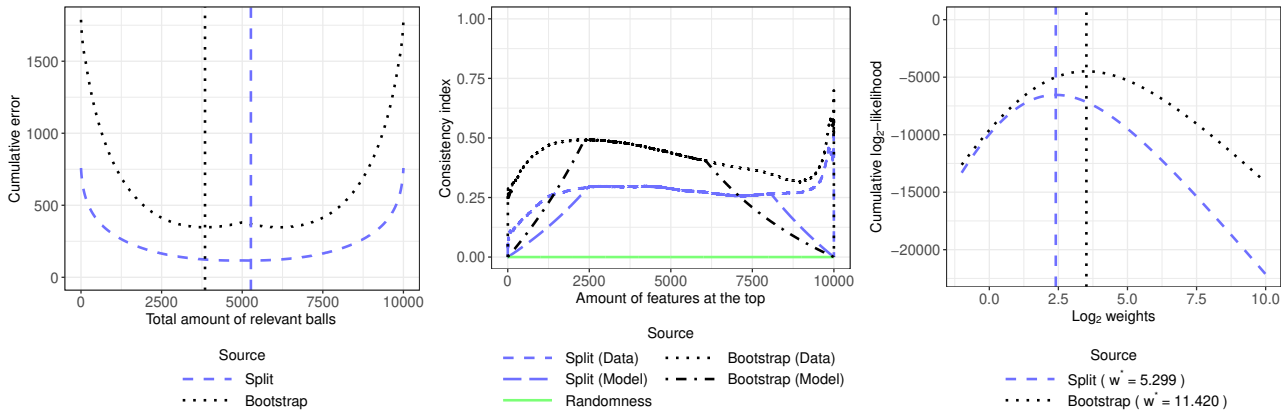
(a) **(b)** **(c)**

Figure 55 Error plot (55a), reproducibility plot (55b) and weight plot (55c) when the ranking algorithm based on the Wilcoxon test is applied to the SECOM database



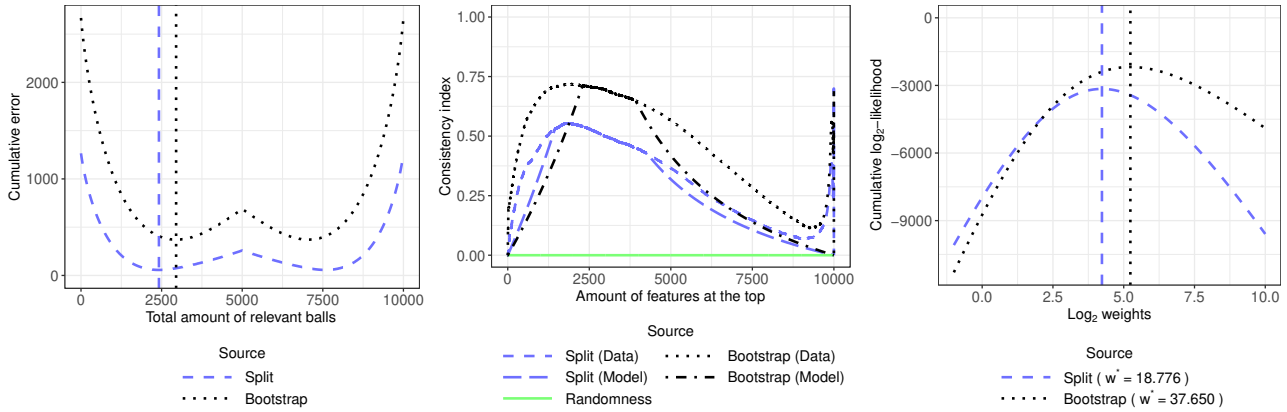
(a) **(b)** **(c)**

Figure 56 Error plot (56a), reproducibility plot (56b) and weight plot (56c) when the ranking algorithm based on the mutual information is applied to the arcene database



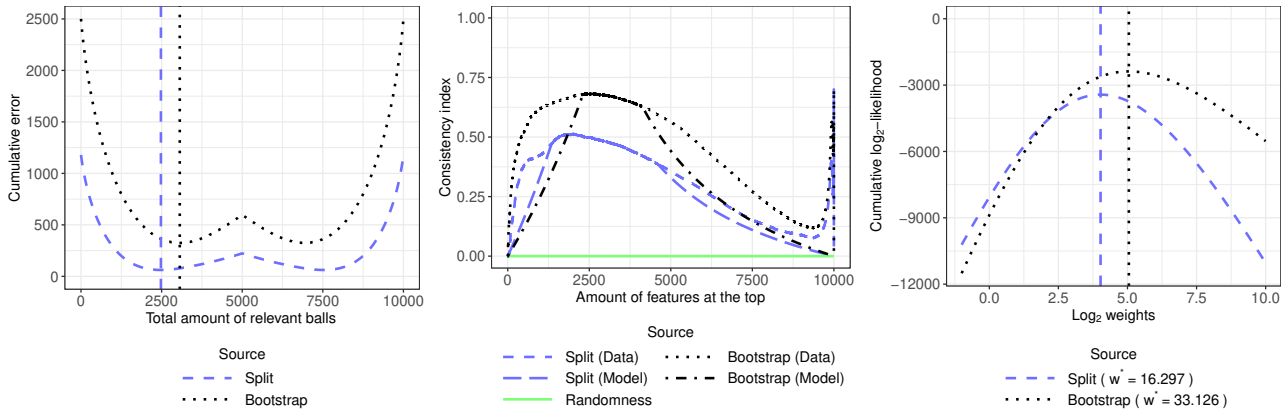
(a) **(b)** **(c)**

Figure 57 Error plot (57a), reproducibility plot (57b) and weight plot (57c) when the ranking algorithm based on the linear SVM is applied to the arcene database



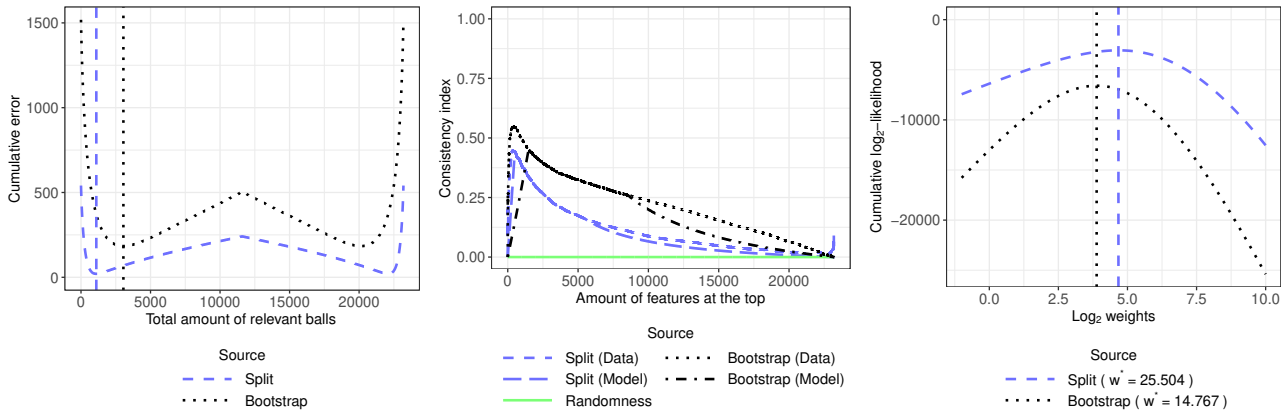
(a) **(b)** **(c)**

Figure 58 Error plot (58a), reproducibility plot (58b) and weight plot (58c) when the ranking algorithm based on the T-test is applied to the arcene database



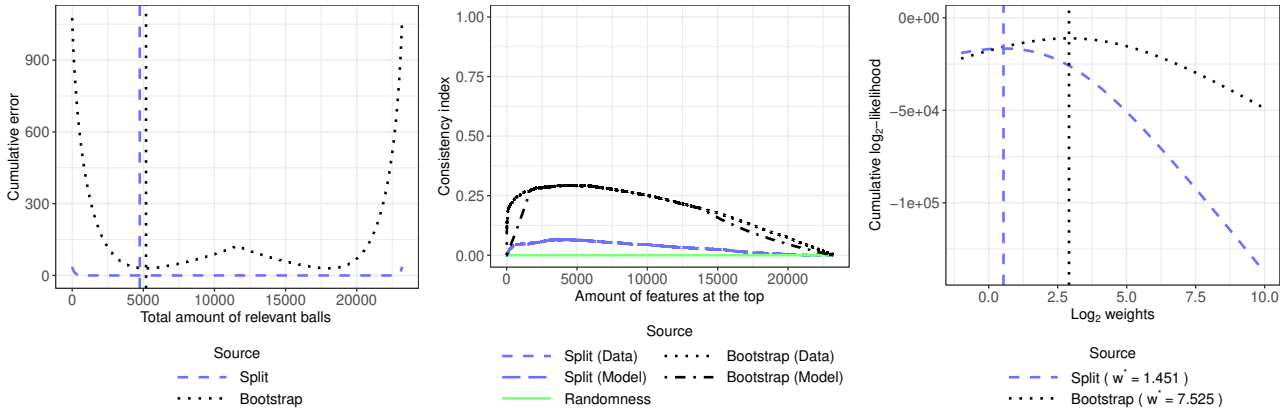
(a) **(b)** **(c)**

Figure 59 Error plot (59a), reproducibility plot (59b) and weight plot (59c) when the ranking algorithm based on the Wilcoxon test is applied to the arcene database



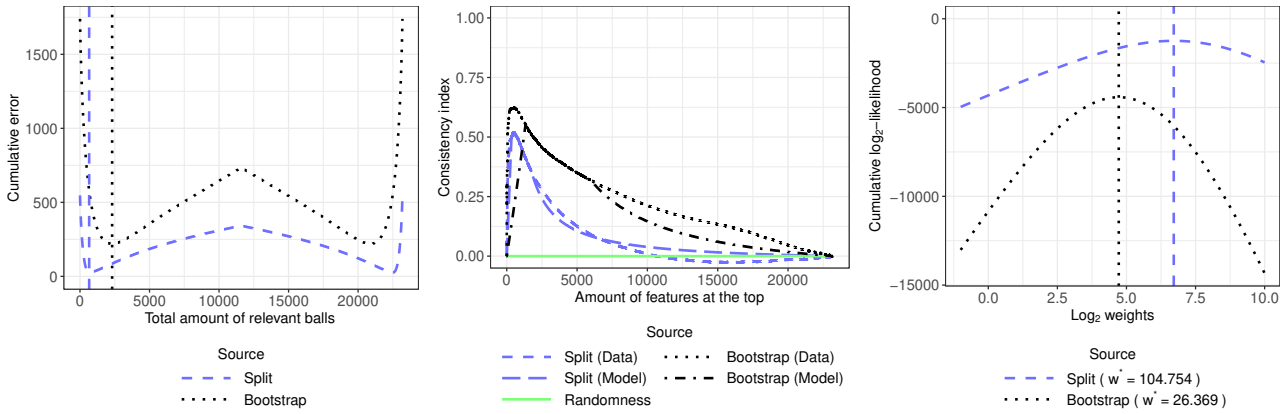
(a) **(b)** **(c)**

Figure 60 Error plot (60a), reproducibility plot (60b) and weight plot (60c) when the ranking algorithm based on the mutual information is applied to the ovarian cancer database



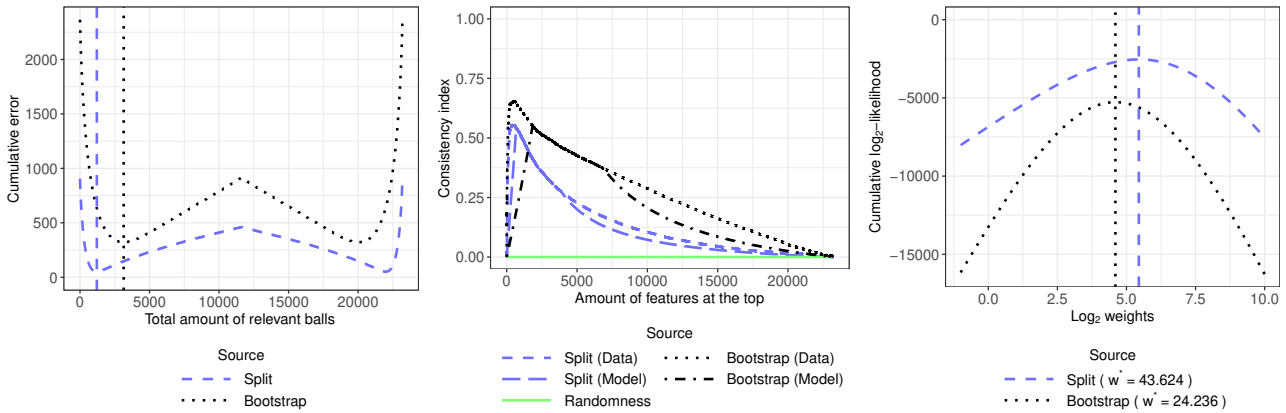
(a) **(b)** **(c)**

Figure 61 Error plot (61a), reproducibility plot (61b) and weight plot (61c) when the ranking algorithm based on the linear SVM is applied to the ovarian cancer database



(a) **(b)** **(c)**

Figure 62 Error plot (62a), reproducibility plot (62b) and weight plot (62c) when the ranking algorithm based on the T-test is applied to the ovarian cancer database



(a) **(b)** **(c)**

Figure 63 Error plot (63a), reproducibility plot (63b) and weight plot (63c) when the ranking algorithm based on the Wilcoxon test is applied to the ovarian cancer database

Table 4 w^* values for the ranking methods when applied to the real datasets

Database	Estimation	Mutual information	SVM	T-test	Wilcoxon test
Breast cancer	Random split	$2.335 \cdot 10^{19}$	3.301	60.500	$2.335 \cdot 10^{19}$
Breast cancer	Bootstrap	$2.335 \cdot 10^{19}$	6.392	58.182	$2.248 \cdot 10^{19}$
Mice	Random split	36.834	12.266	25.282	24.479
Mice	Bootstrap	54.094	21.697	40.239	42.238
SECOM	Random split	153.483	49.718	230.861	238.299
SECOM	Bootstrap	280.152	54.985	738.401	365.684
Arcene	Random split	29.988	5.299	18.776	16.297
Arcene	Bootstrap	23.941	11.420	37.650	33.126
Ovarian cancer	Random split	25.504	1.451	104.754	43.624
Ovarian cancer	Bootstrap	14.767	7.525	26.369	24.236

Table 5 Model AUC values for the ranking methods when applied to the real datasets

Database	Estimation	Mutual information	SVM	T-test	Wilcoxon test
Breast cancer	Random split	1.00000	0.73292	0.99548	1.00000
Breast cancer	Bootstrap	1.00000	0.88995	0.99554	1.00000
Mice	Random split	0.98554	0.94365	0.97874	0.97846
Mice	Bootstrap	0.99417	0.97438	0.99078	0.99158
SECOM	Random split	0.99968	0.99520	0.99986	0.99986
SECOM	Bootstrap	0.99988	0.99575	0.99998	0.99990
Arcene	Random split	0.98496	0.87718	0.96597	0.95932
Arcene	Bootstrap	0.98079	0.94586	0.98705	0.98459
Ovarian cancer	Random split	0.96858	0.66808	0.99278	0.98301
Ovarian cancer	Bootstrap	0.94793	0.89923	0.97239	0.97071

References

[1] Guyon, I., Gunn, S., Ben-Hur, A., Dror, G.: Result analysis of the nips 2003 feature selection challenge. In: Advances in neural information processing systems, pp. 545–552 (2005)

[2] Higuera, C., Gardiner, K.J., Cios, K.J.: Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. PloS one **10**(6), e0129126 (2015)

[3] Mangasarian, O.L., Street, W.N., Wolberg, W.H.: Breast cancer diagnosis and prognosis via linear programming. Operations Research **43**(4), 570–577 (1995)

[4] McCann, M., Li, Y., Maguire, L., Johnston, A.: Causality challenge: benchmarking relevant signal components for effective monitoring and process control. In: Causality: Objectives and Assessment, pp. 277–288 (2010)

[5] Street, W.N., Wolberg, W.H., Mangasarian, O.L.: Nuclear feature extraction for breast tumor diagnosis. In: Biomedical image processing and biomedical visualization, vol. 1905, pp. 861–870. International Society for Optics and Photonics (1993)

[6] Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Weisenberger, D.J., Shen, H., Campan, M., Noushmehr, H., Bell, C.G., Maxwell, A.P., et al.: Age-dependent dna methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome research **20**(4), 440–446 (2010)

[7] Wang, S.: Method to detect differentially methylated loci with case-control designs using illumina arrays. Genetic epidemiology **35**(7), 686–694 (2011)