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:

- 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.
- Plots of the experimentation with synthetic data: All the plots derived from the experimentation with synthetic data are displayed in many figures.
- Real databases preprocessing: The preprocessings applied to the different real databases.
- 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 $\boldsymbol{D}^2$.
- Plots of the experimentation with real data: All the plots derived from the experimentation with real data are displayed in many figures.

2 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).

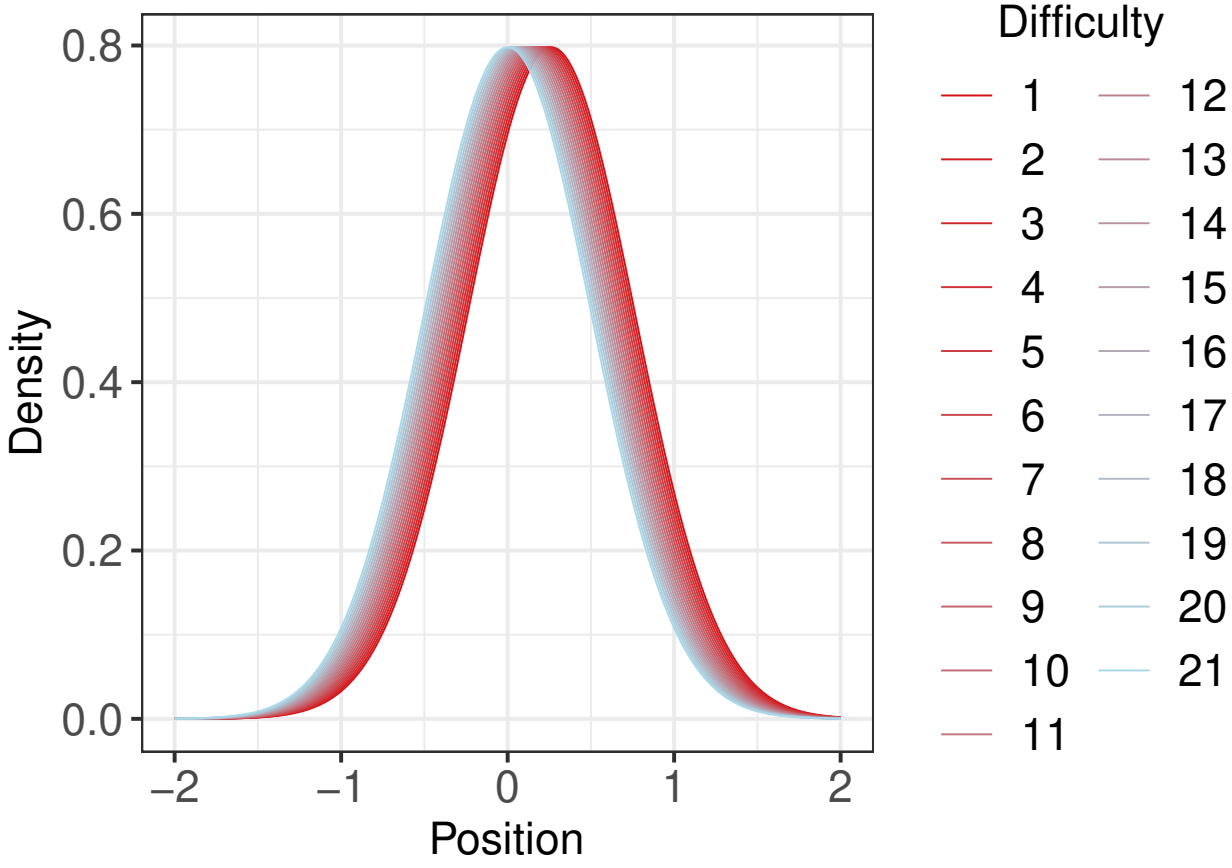
In Table 1 we show the specific values of the parameters of each of the distributions shown in Figure 1. In Table 1 each of those distributions is identified by its associated scenario (differences in location or differences in both location and spread) and difficulty as shown in Figure 1.

3 Plots of the experimentation with synthetic data

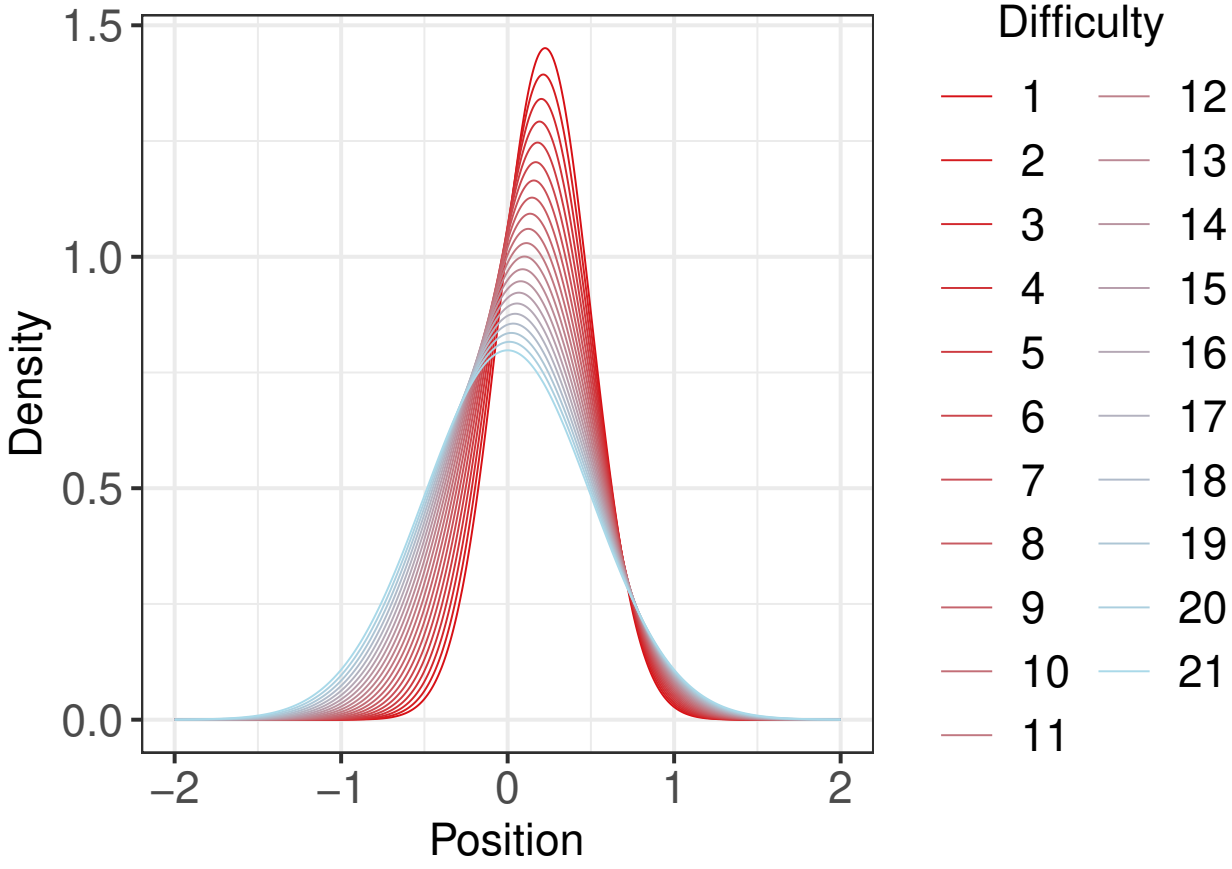
In Figures 2 to 43 the plots of the experimentation with synthetic data can be seen. Specifically, the plots are shown in order, first showing those corresponding to the scenario of differences in location and then showing those corresponding to the scenario of differences in both location and spread.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

4 Real databases preprocessings

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.



(a)

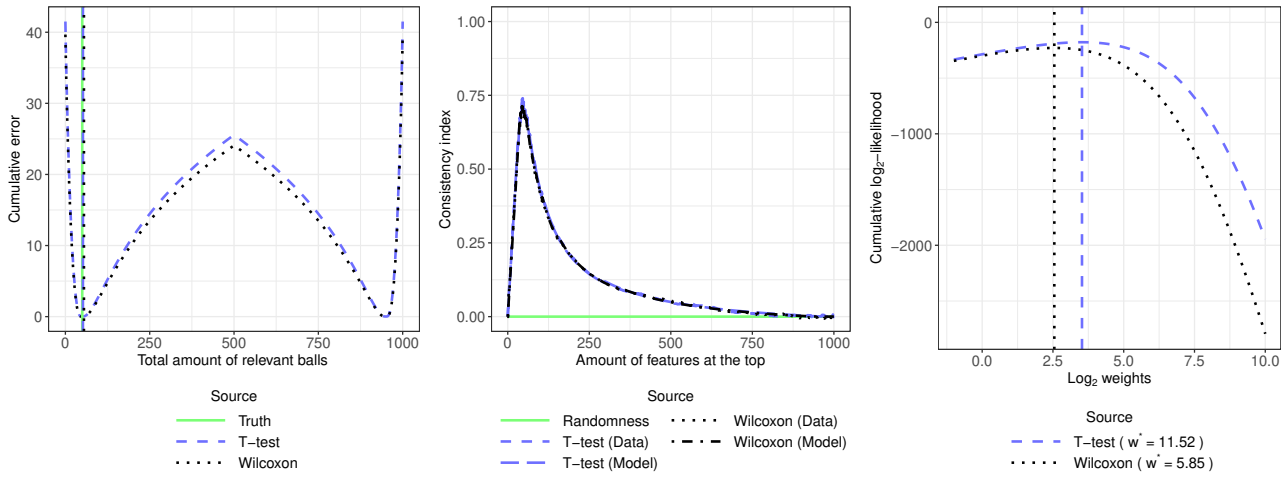


(b)

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

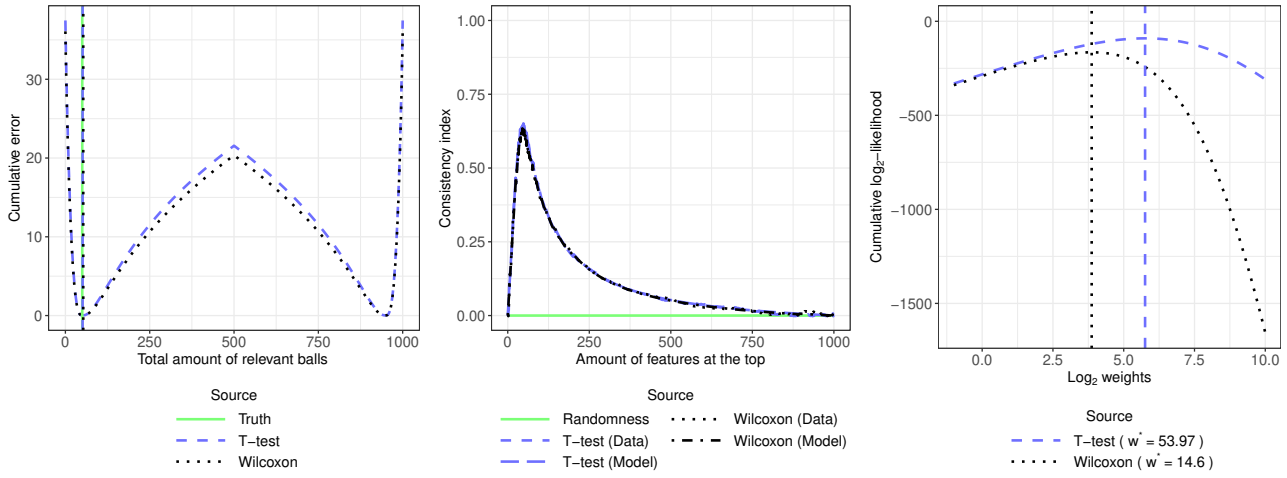
Table 1 Weight ratios for the different combinations of methods, problems and difficulties when dealing with synthetic data

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



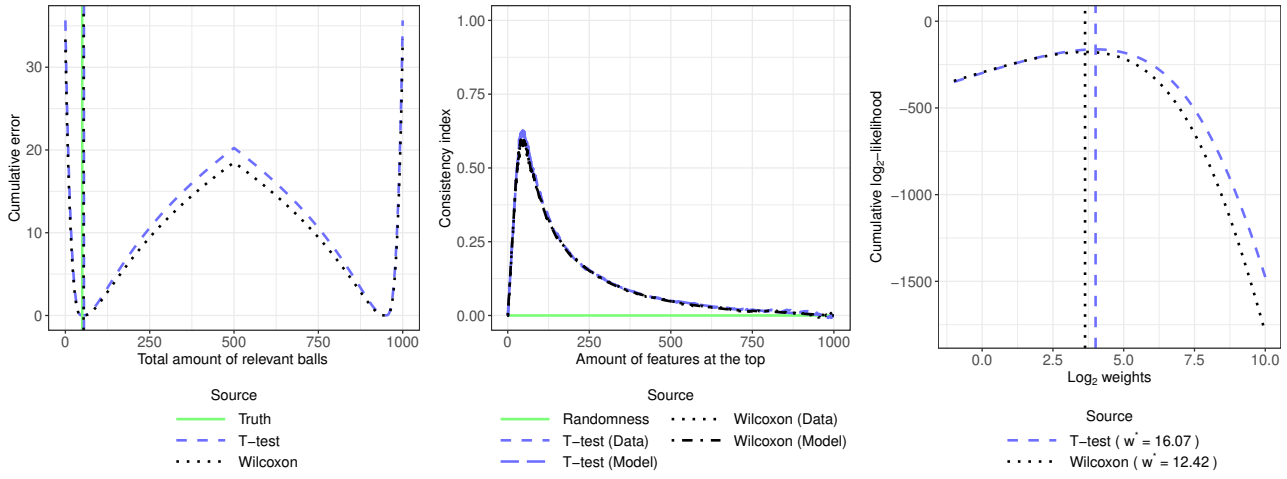
(a) **(b)** **(c)**

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



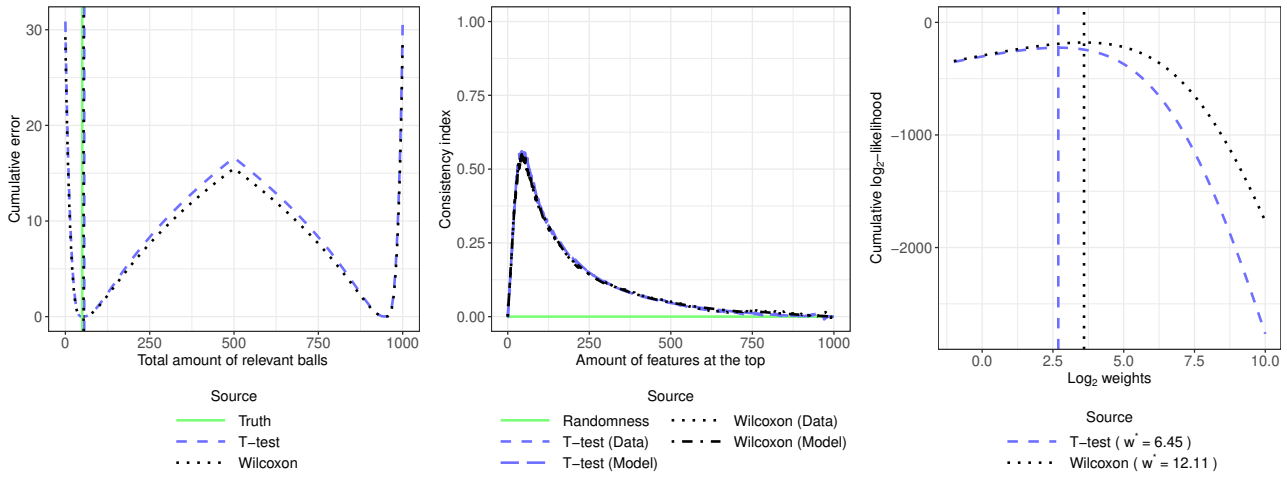
(a) **(b)** **(c)**

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



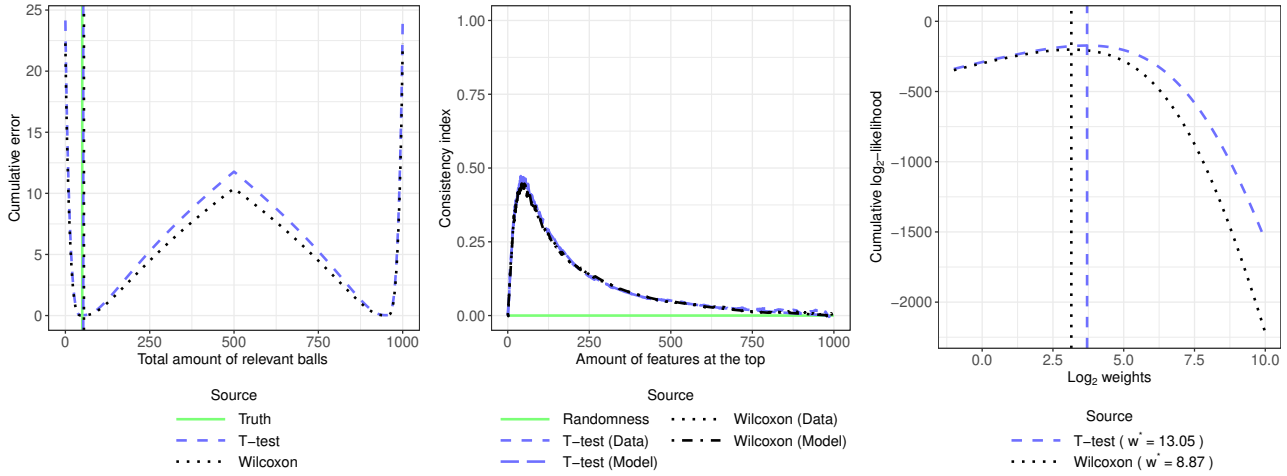
(a) **(b)** **(c)**

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



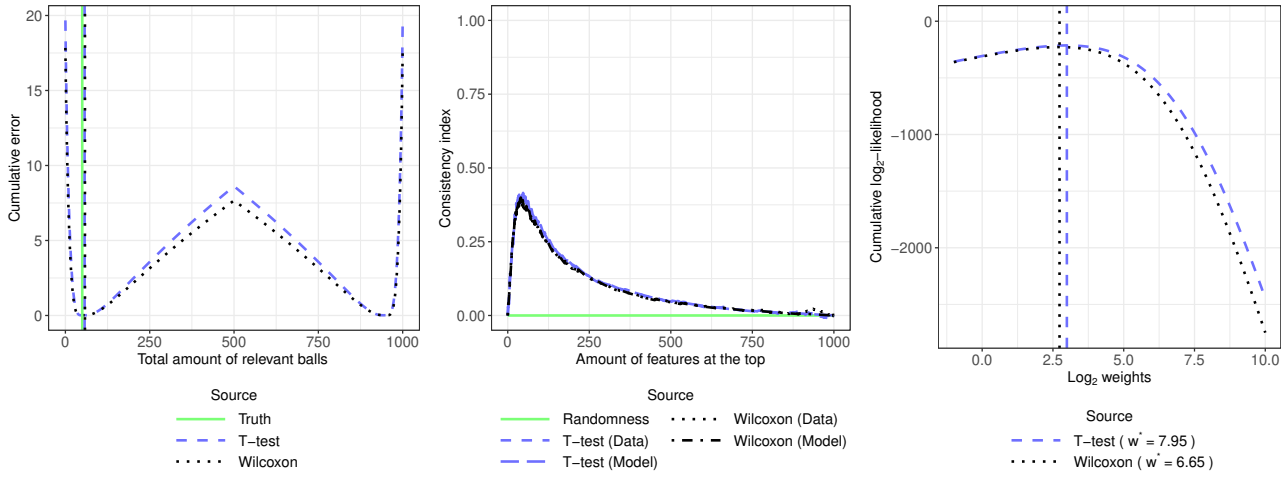
(a) **(b)** **(c)**

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



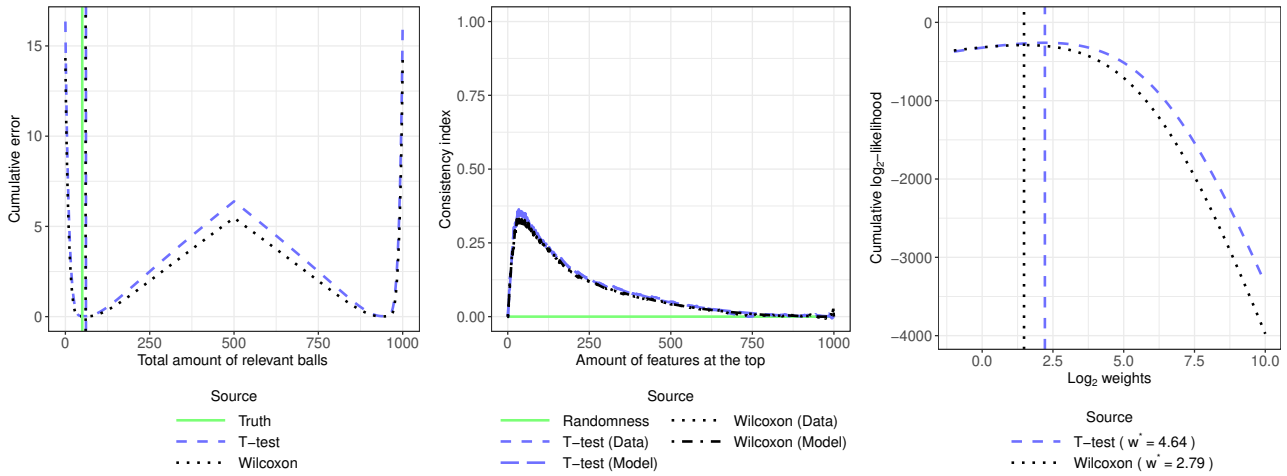
(a) **(b)** **(c)**

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



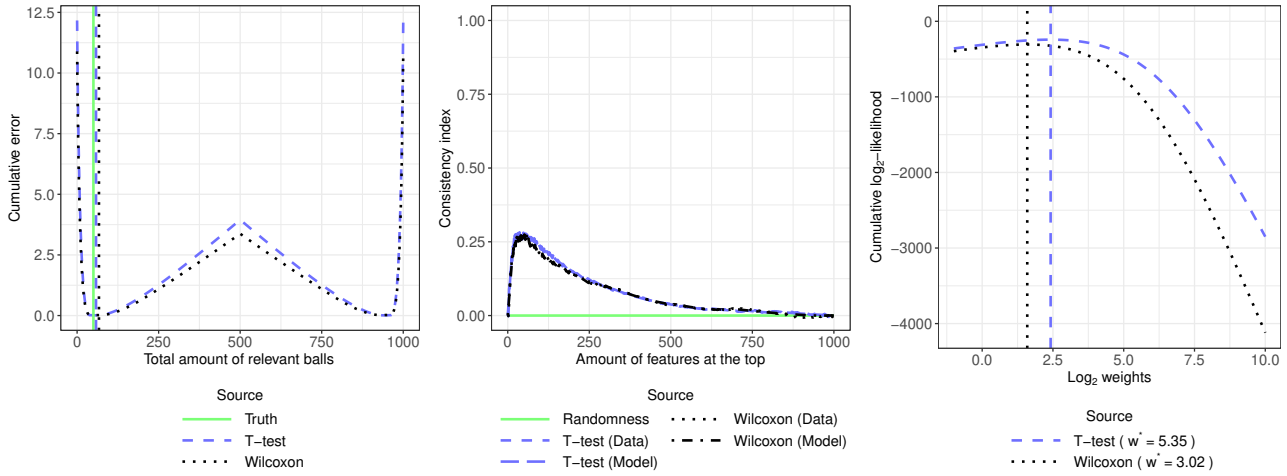
(a) **(b)** **(c)**

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



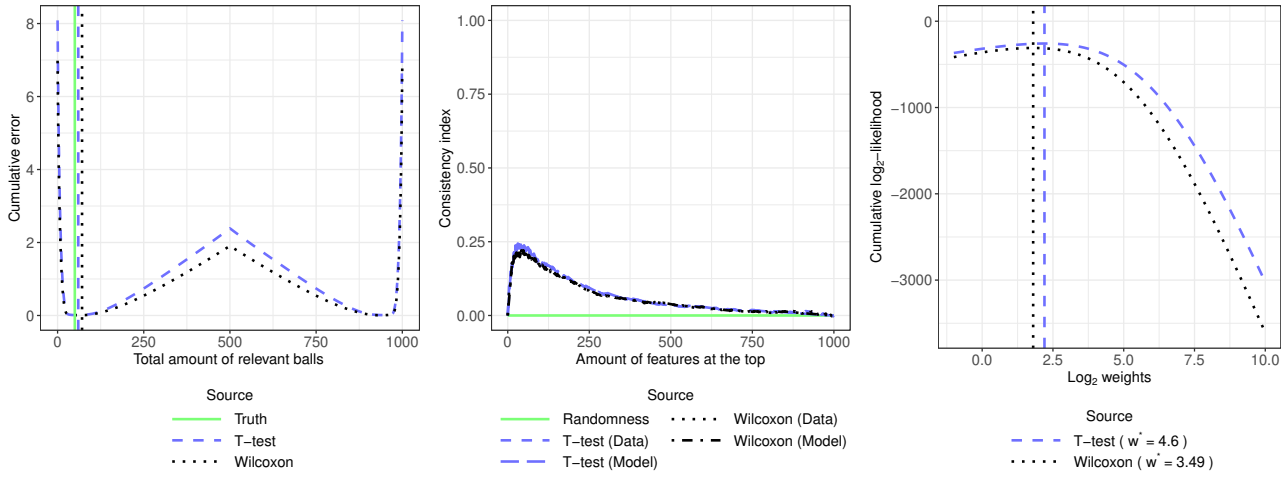
(a) **(b)** **(c)**

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



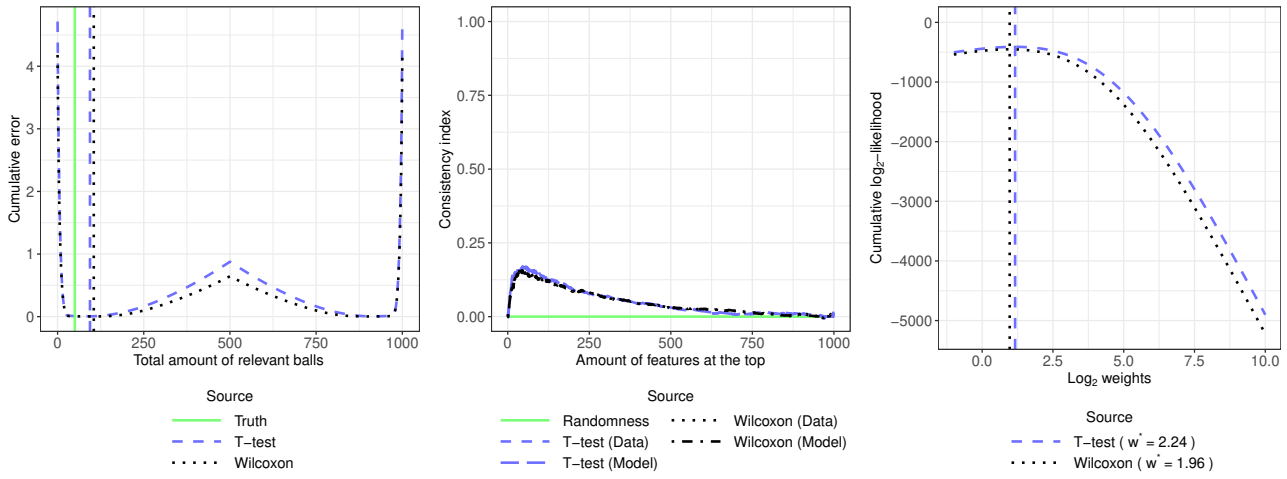
(a) **(b)** **(c)**

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



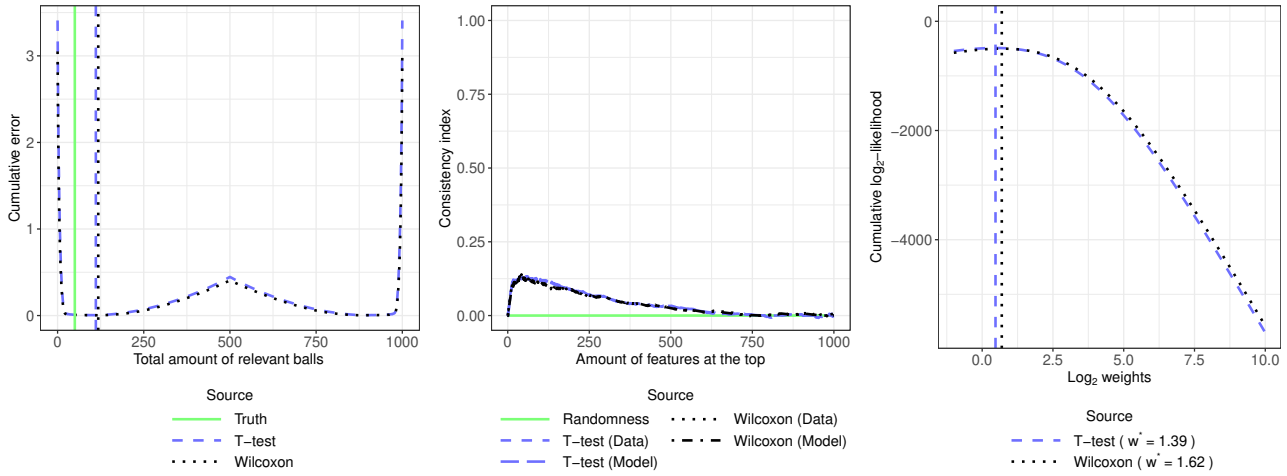
(a) **(b)** **(c)**

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



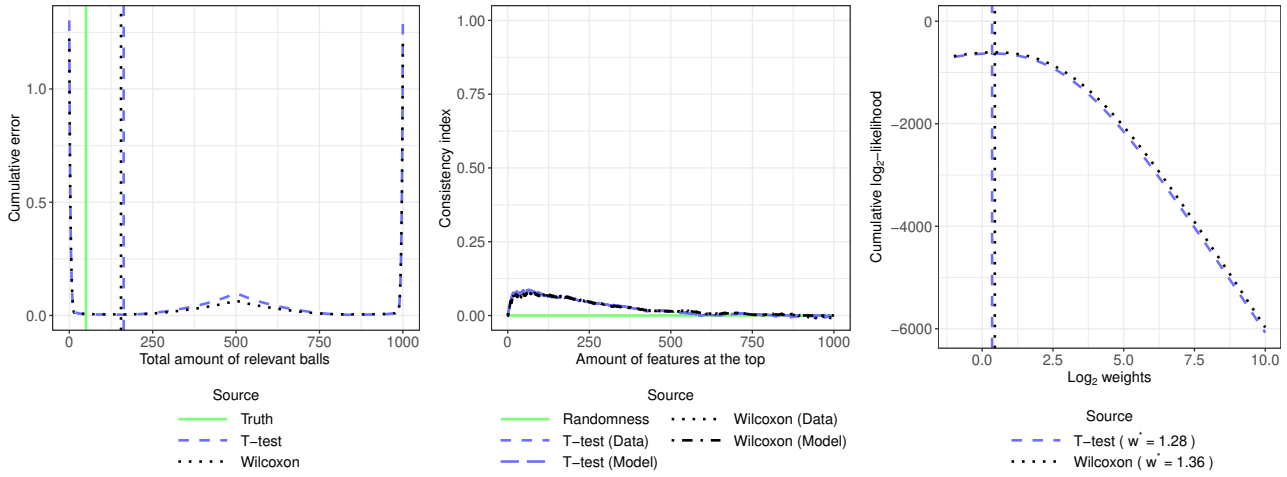
(a) **(b)** **(c)**

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



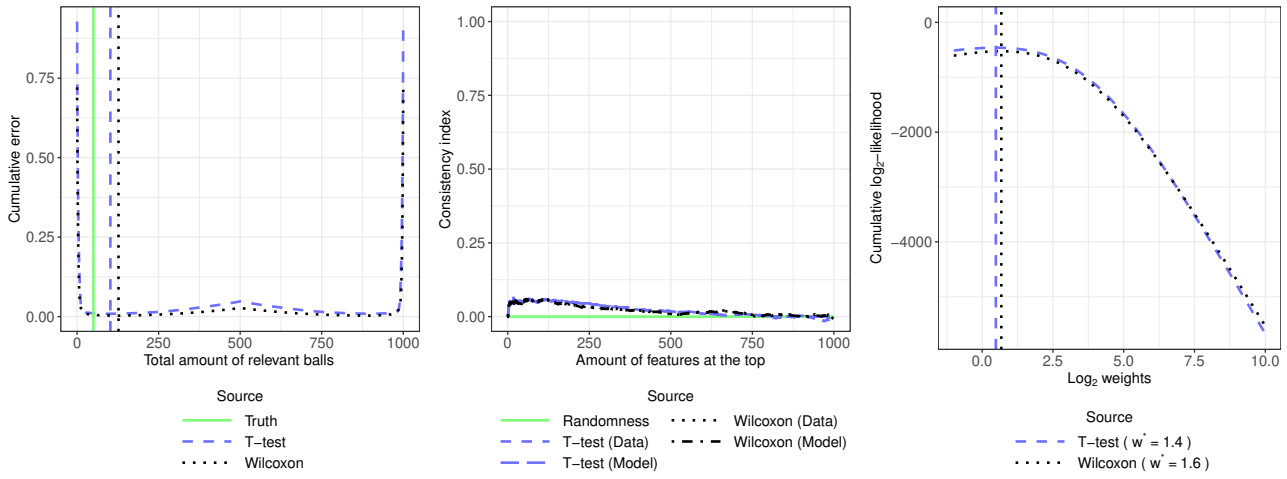
(a) **(b)** **(c)**

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



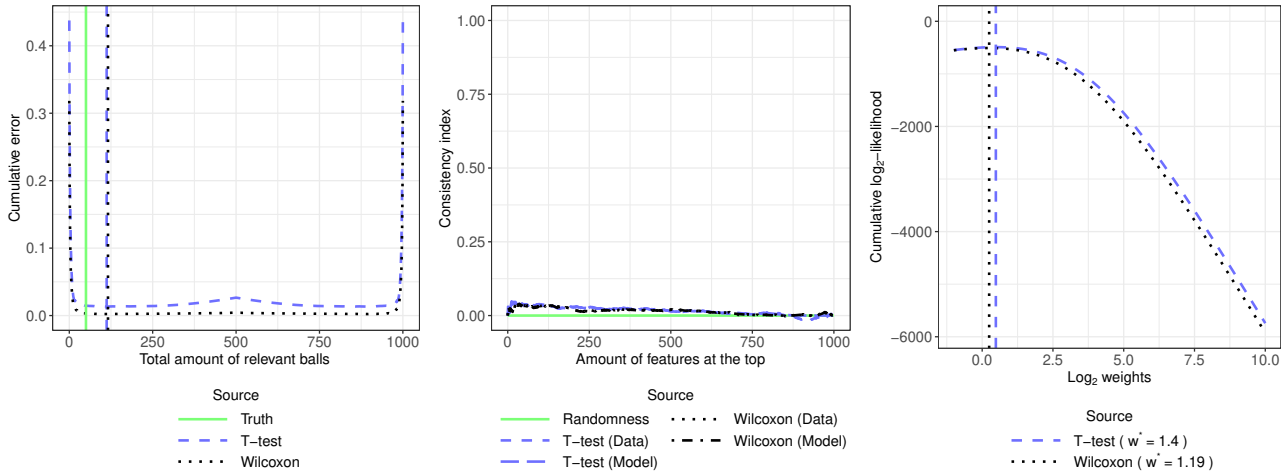
(a) **(b)** **(c)**

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



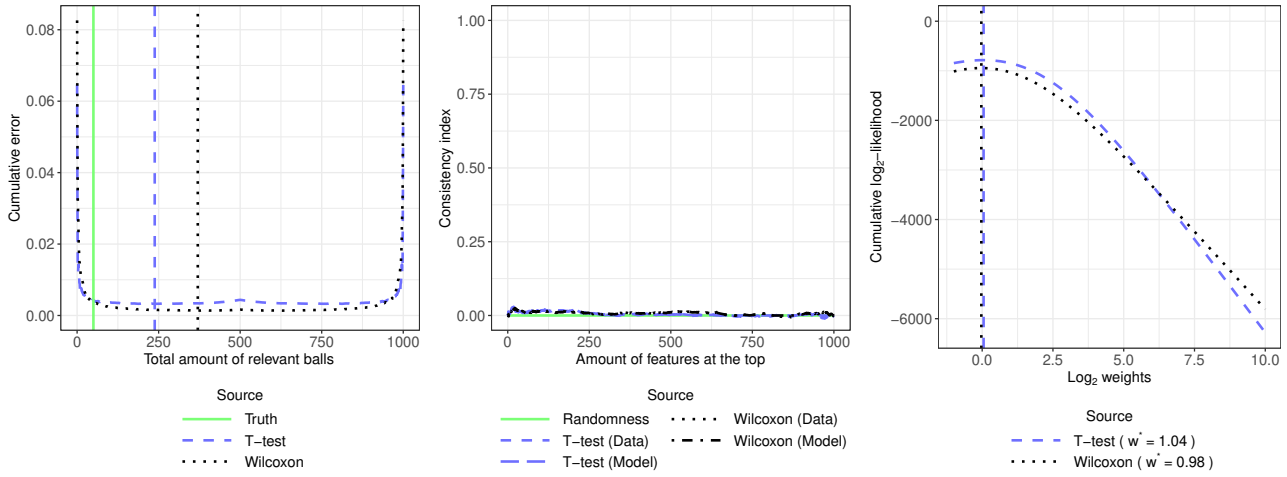
(a) **(b)** **(c)**

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



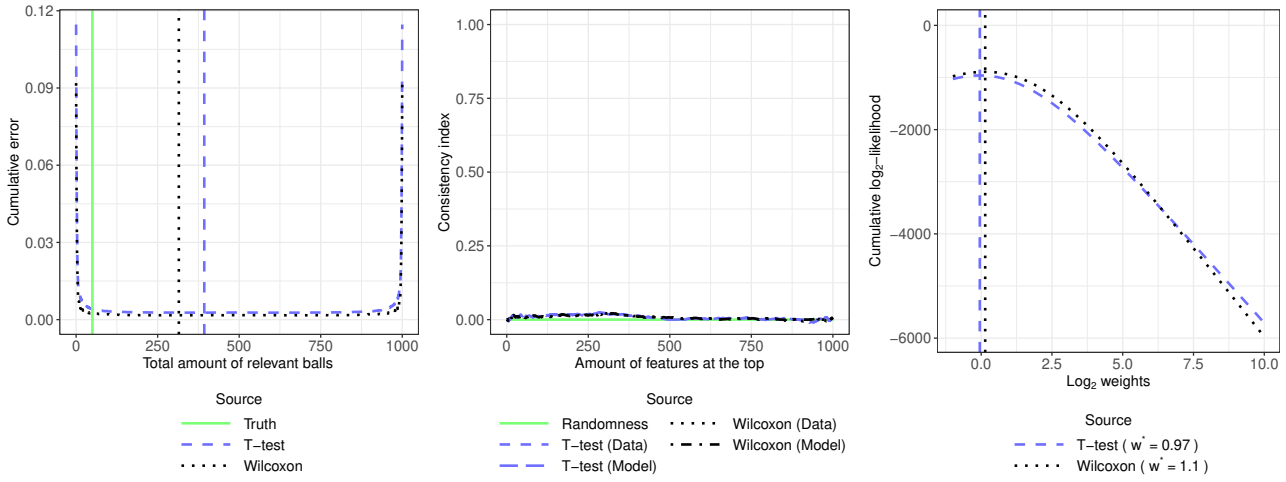
(a) **(b)** **(c)**

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



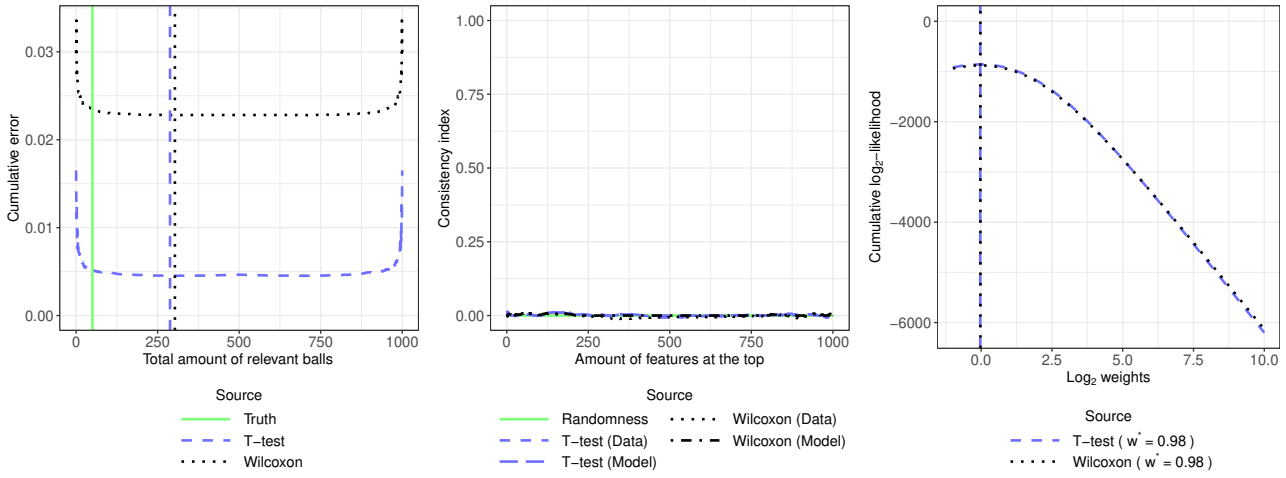
(a) **(b)** **(c)**

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



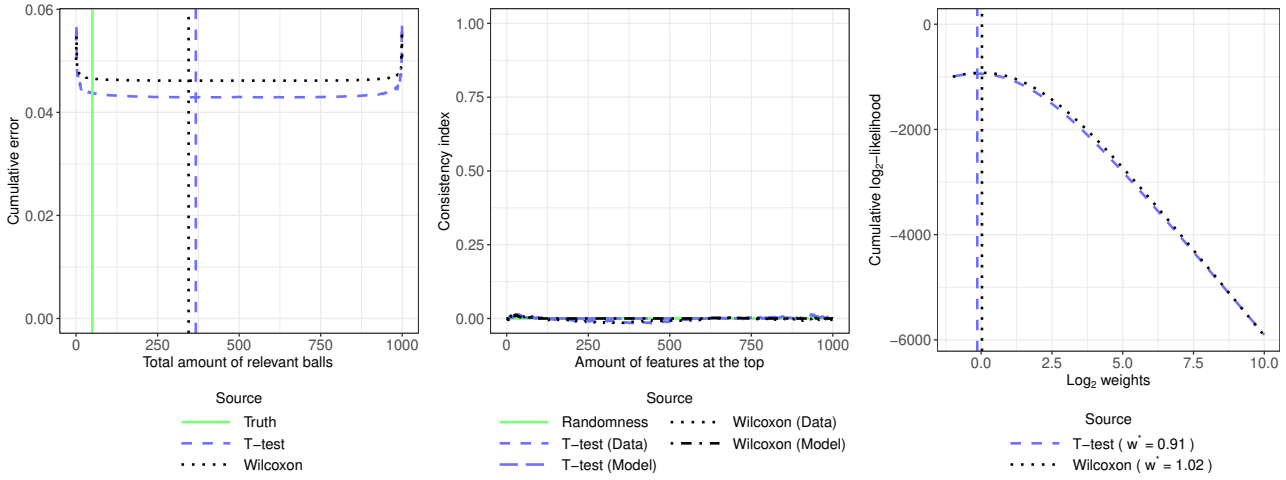
(a) **(b)** **(c)**

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



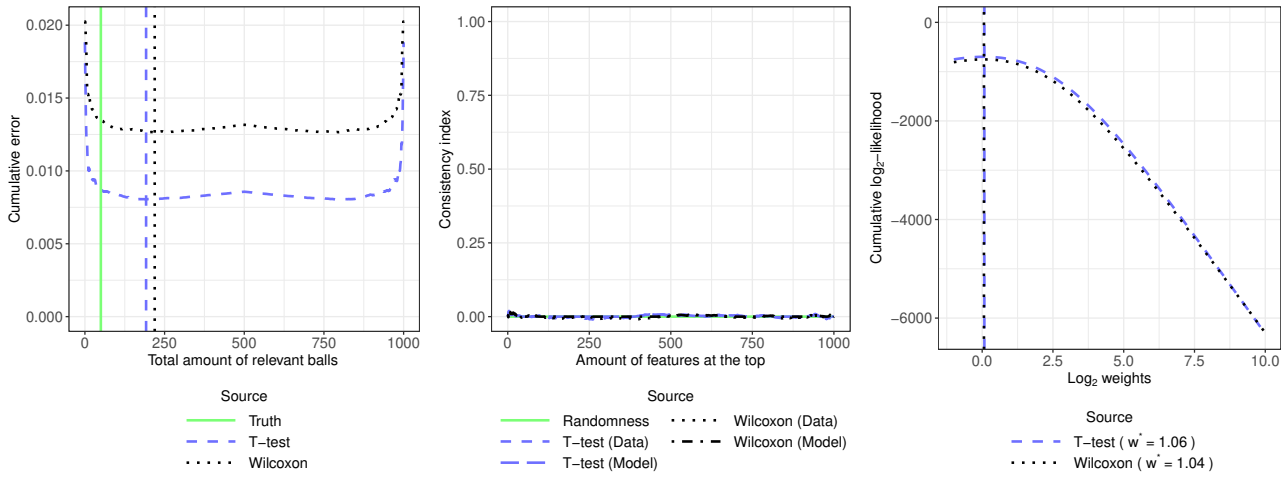
(a) **(b)** **(c)**

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



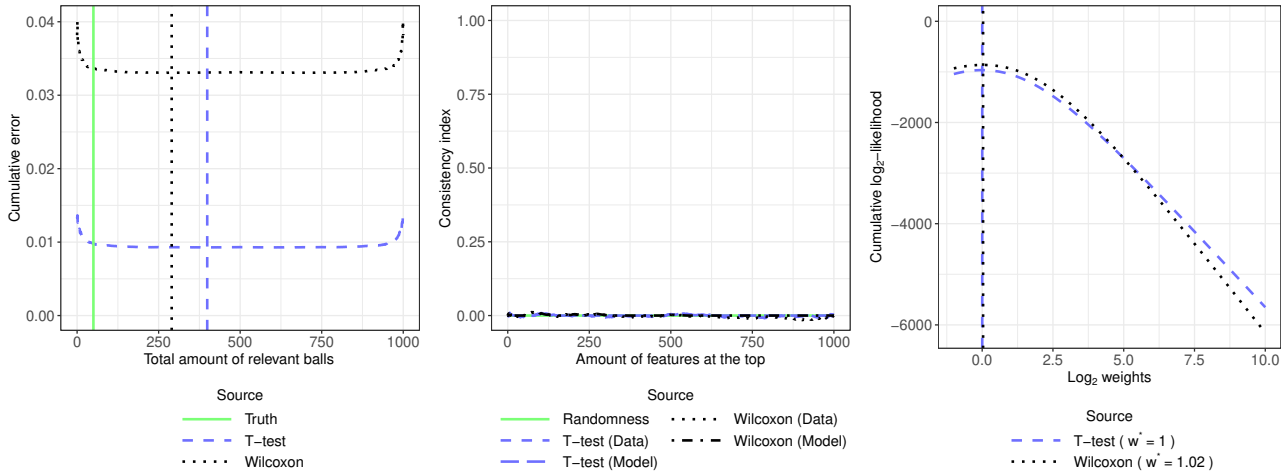
(a) **(b)** **(c)**

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



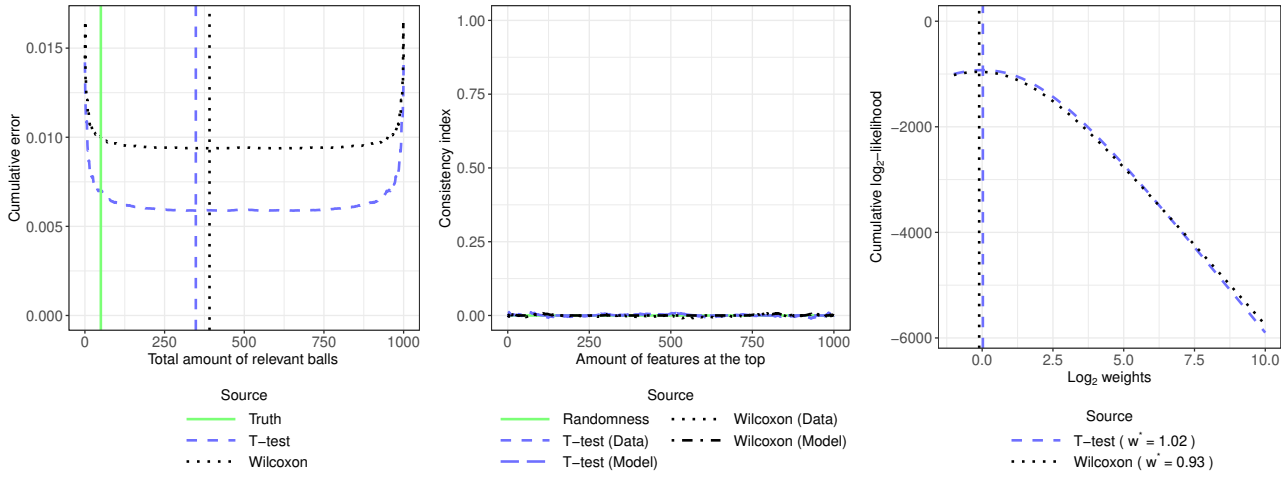
(a) **(b)** **(c)**

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



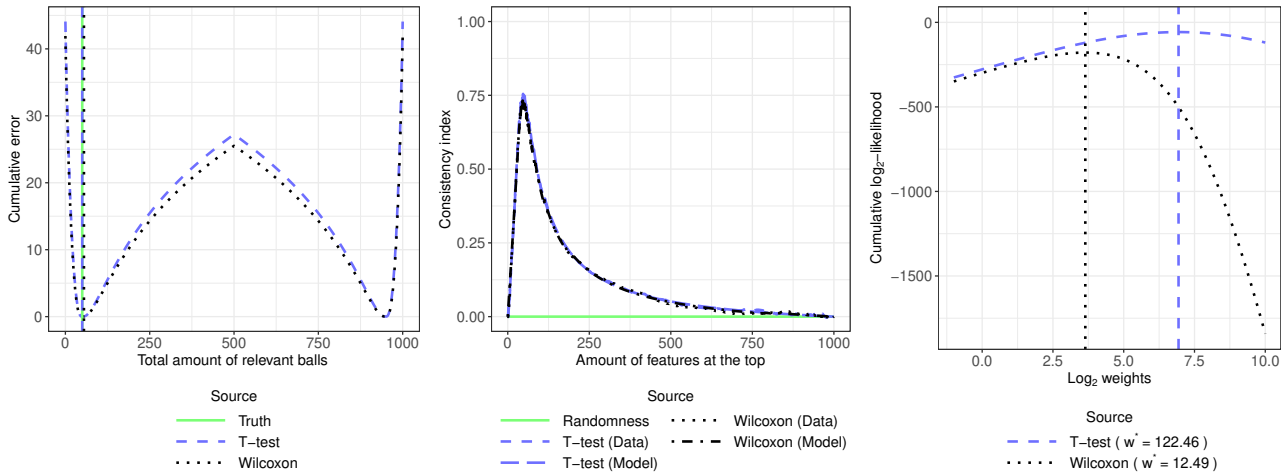
(a) **(b)** **(c)**

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



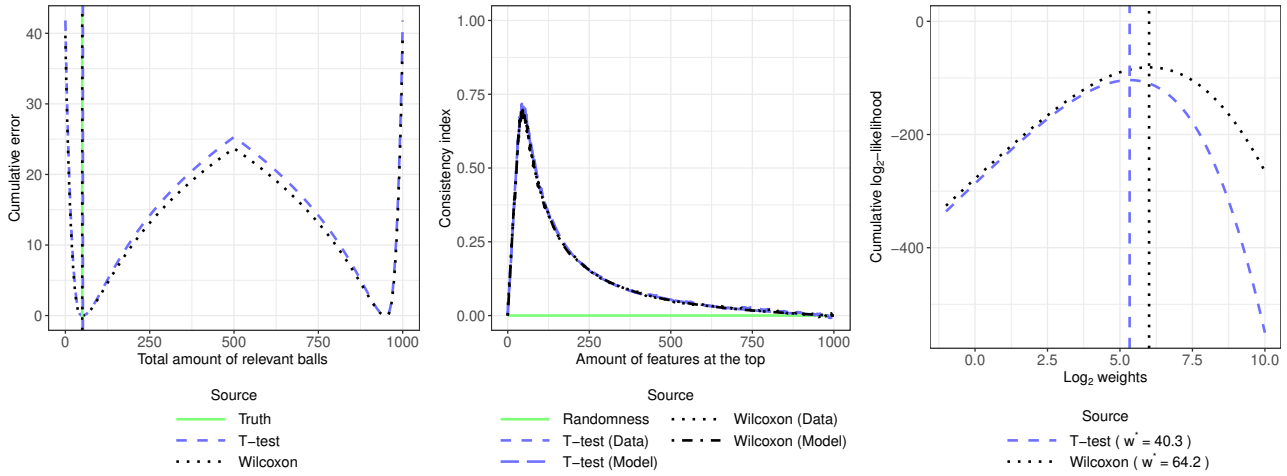
(a) **(b)** **(c)**

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



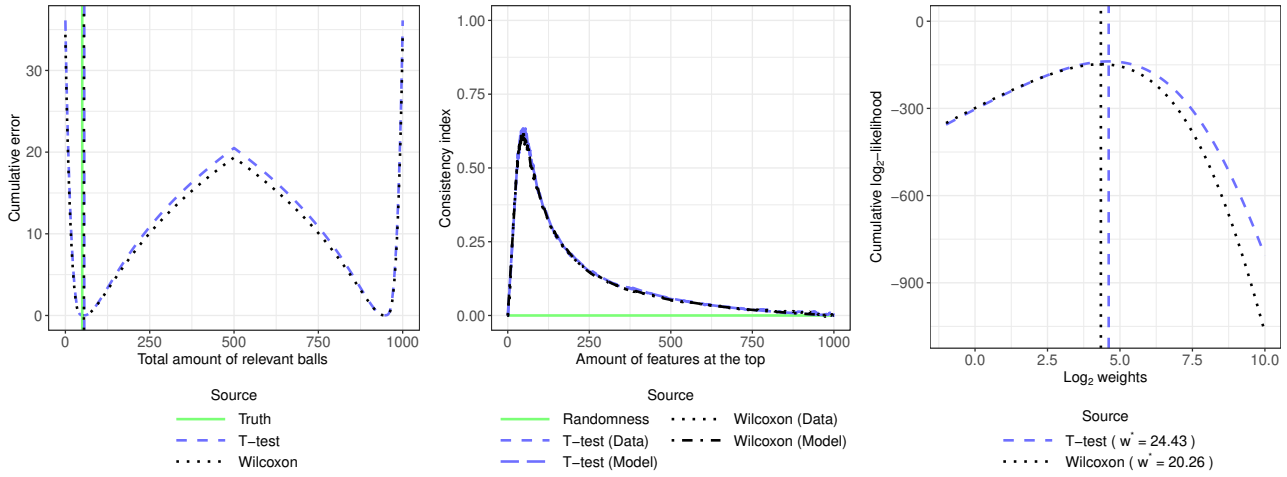
(a) **(b)** **(c)**

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



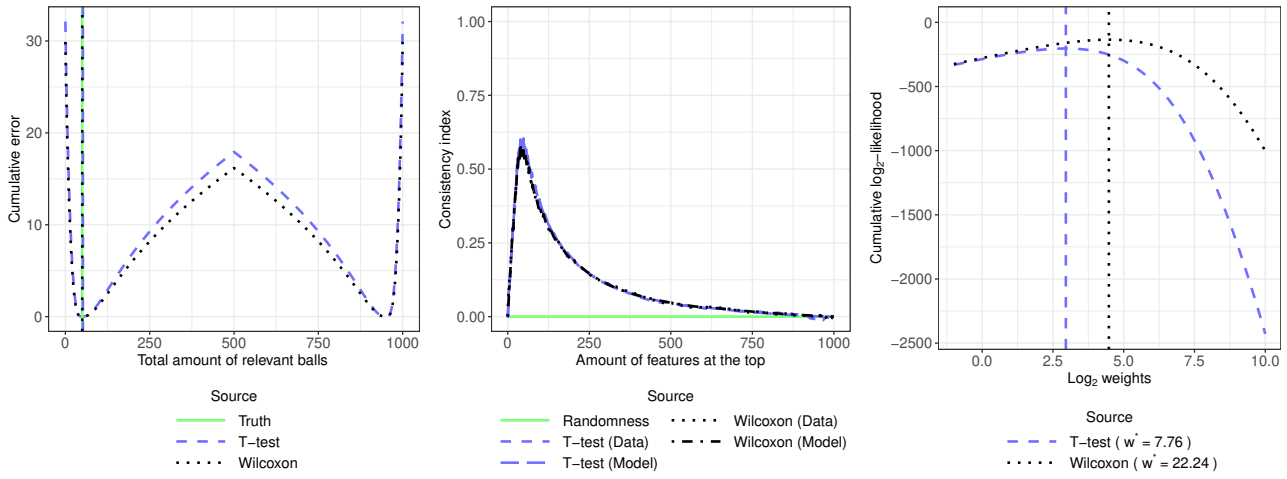
(a) **(b)** **(c)**

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



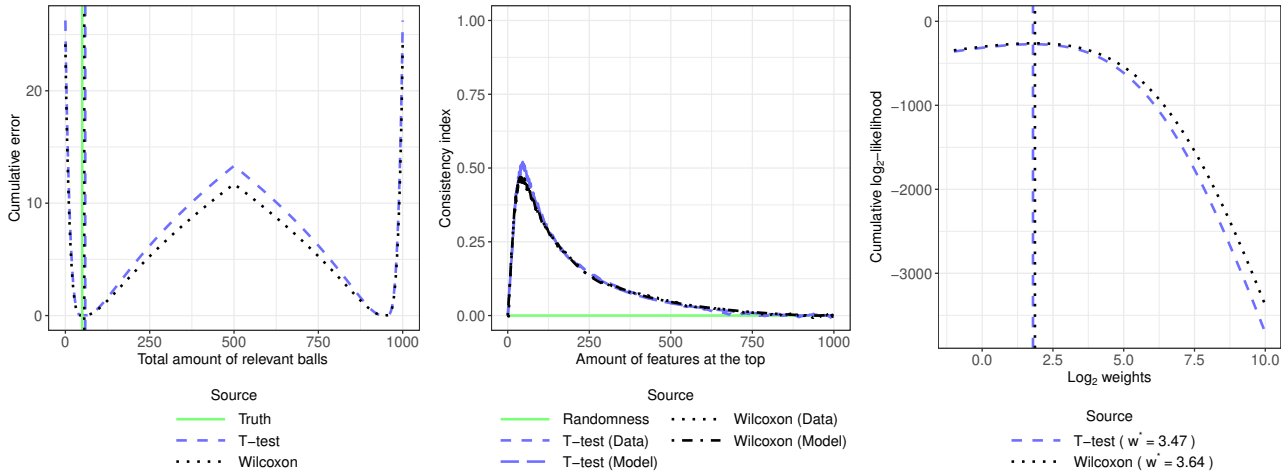
(a) **(b)** **(c)**

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



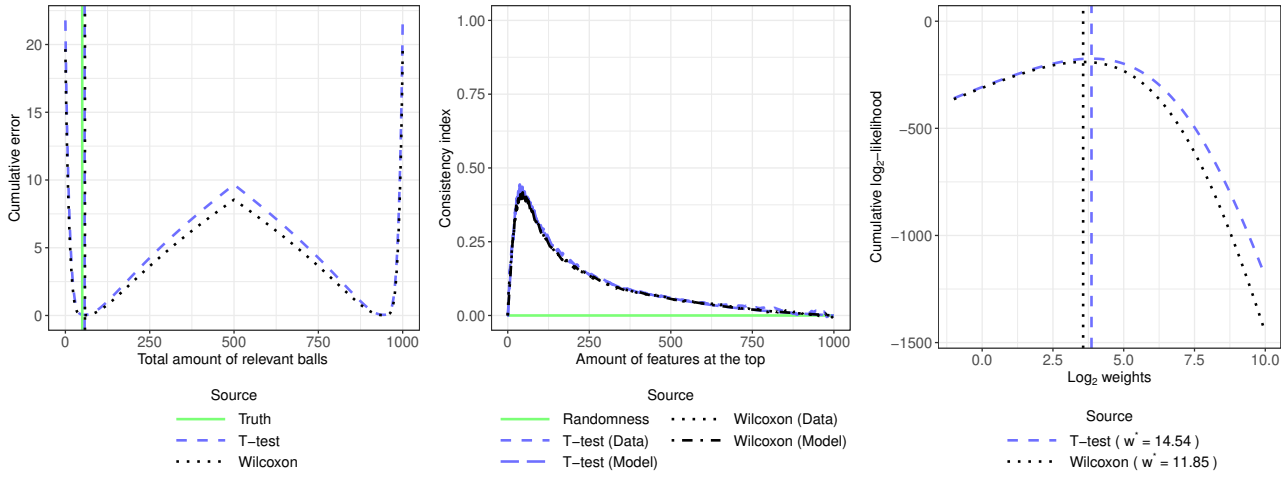
(a) **(b)** **(c)**

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



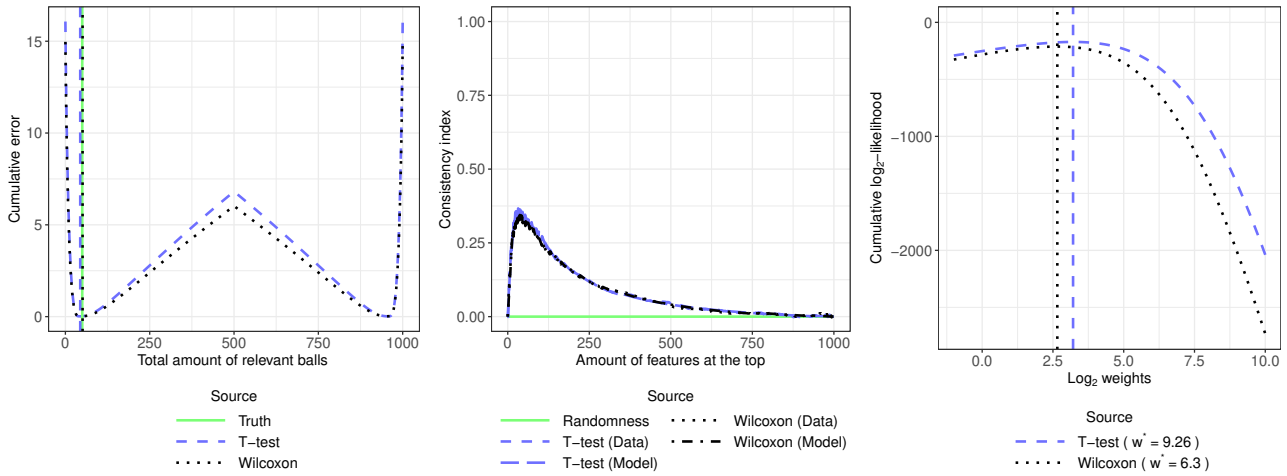
(a) **(b)** **(c)**

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



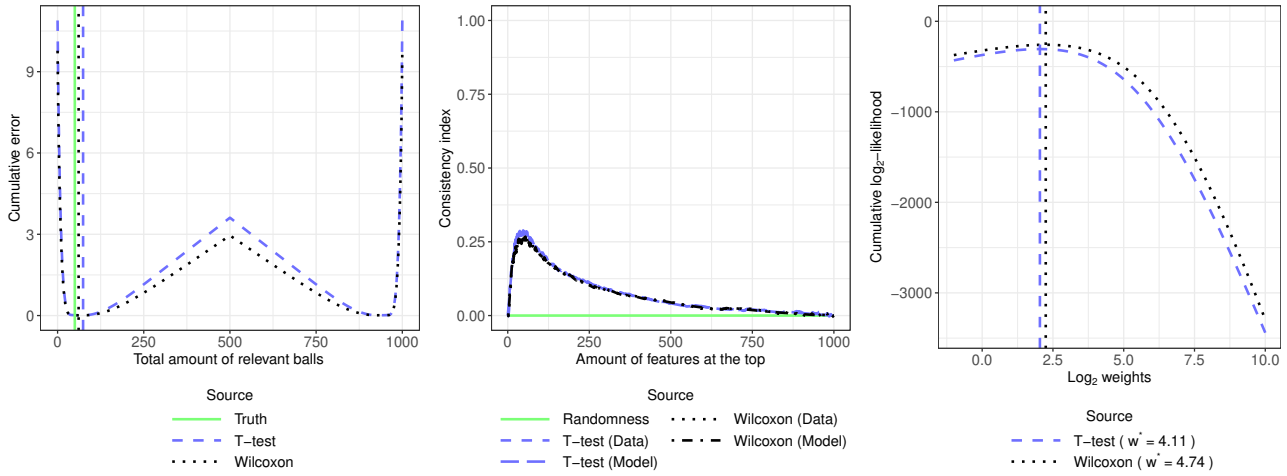
(a) **(b)** **(c)**

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



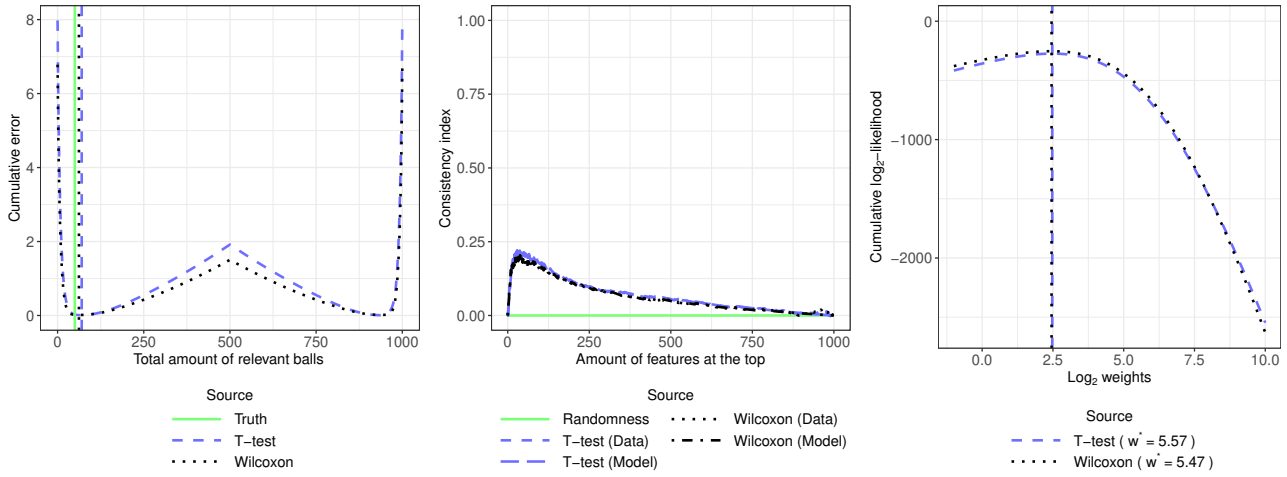
(a) **(b)** **(c)**

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



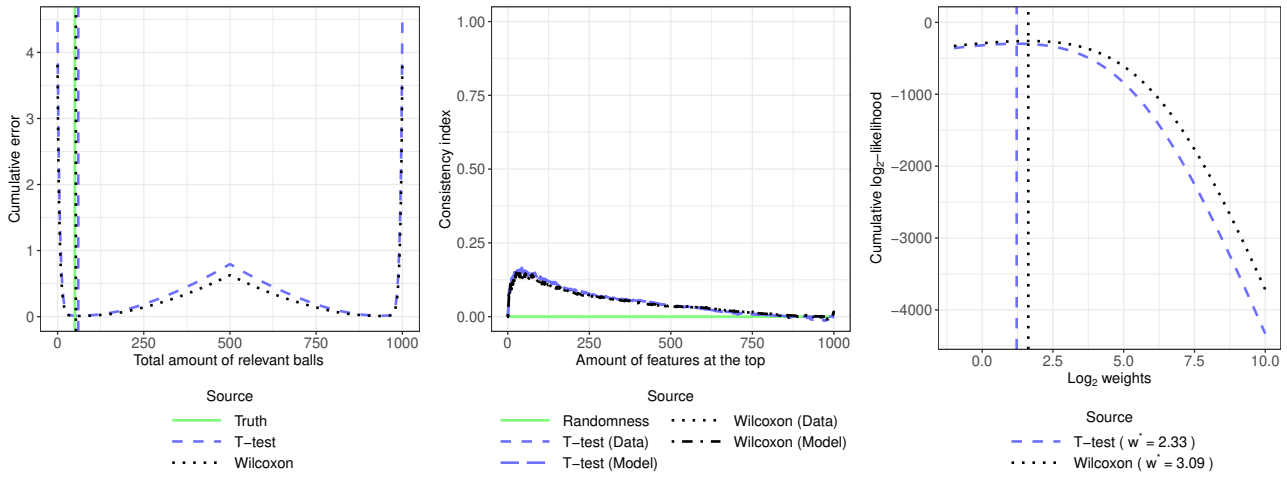
(a) **(b)** **(c)**

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



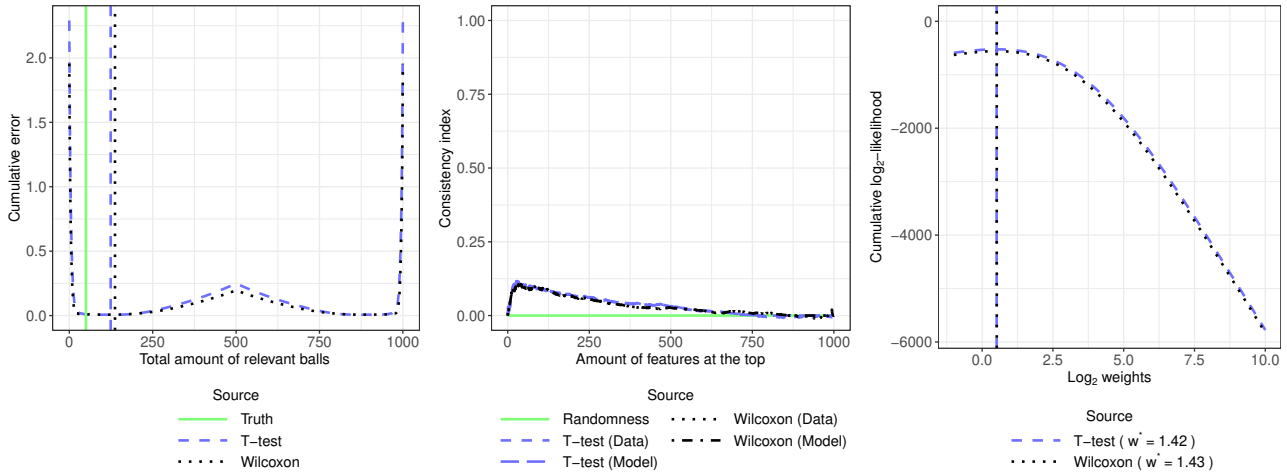
(a) **(b)** **(c)**

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



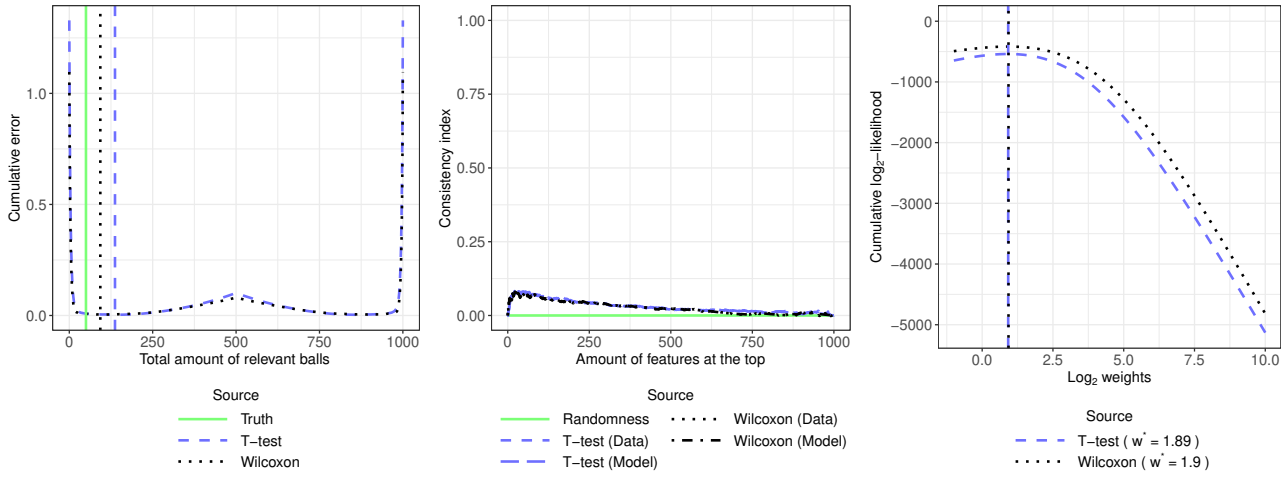
(a) **(b)** **(c)**

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



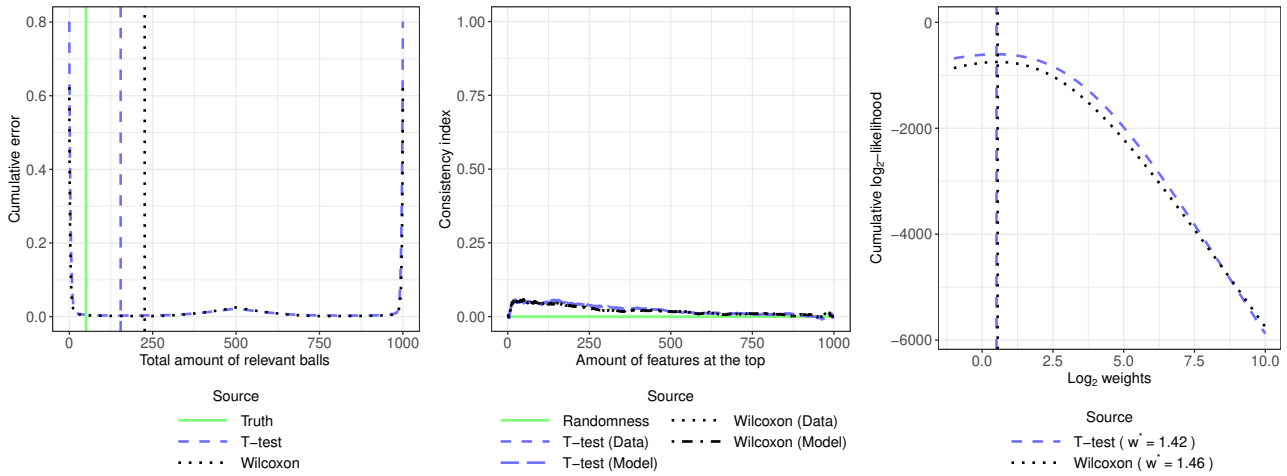
(a) **(b)** **(c)**

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



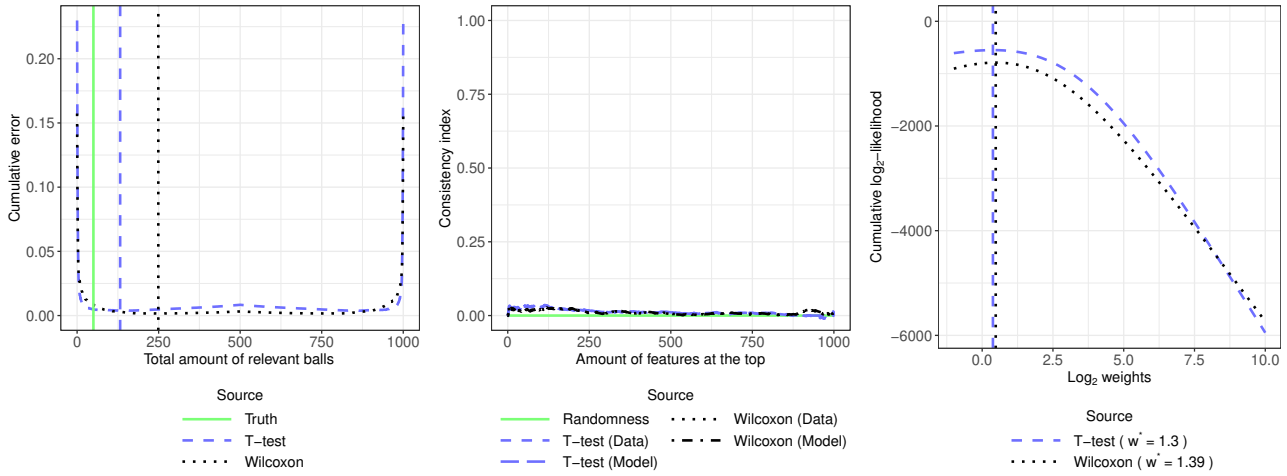
(a) **(b)** **(c)**

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



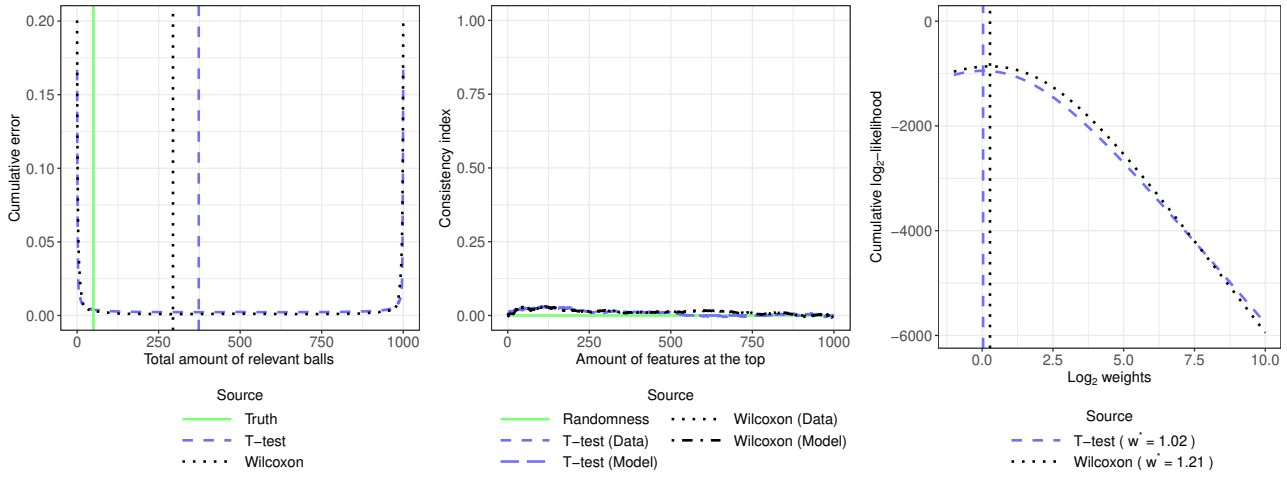
(a) **(b)** **(c)**

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



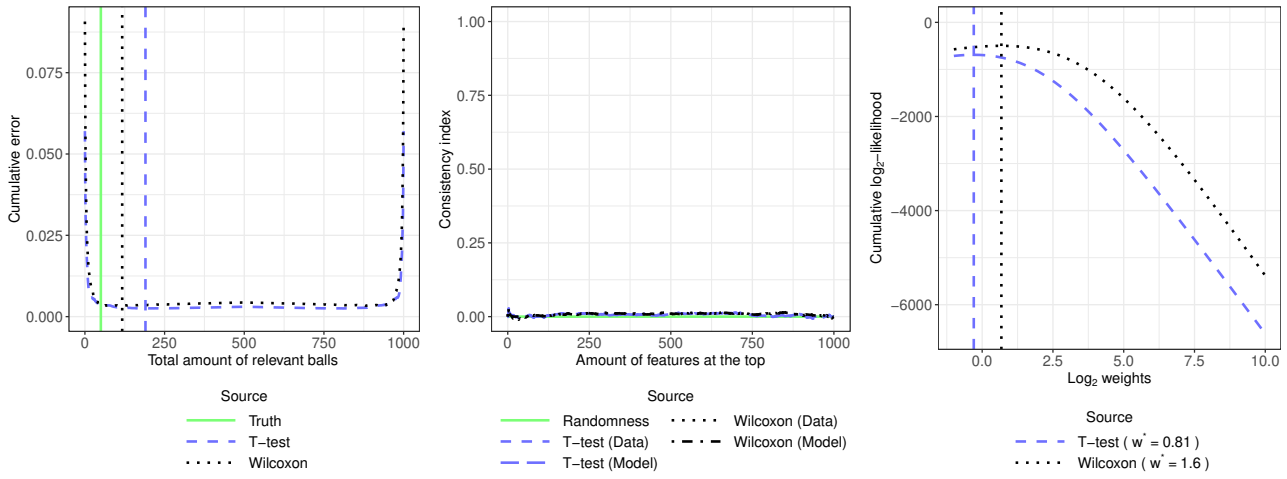
(a) **(b)** **(c)**

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



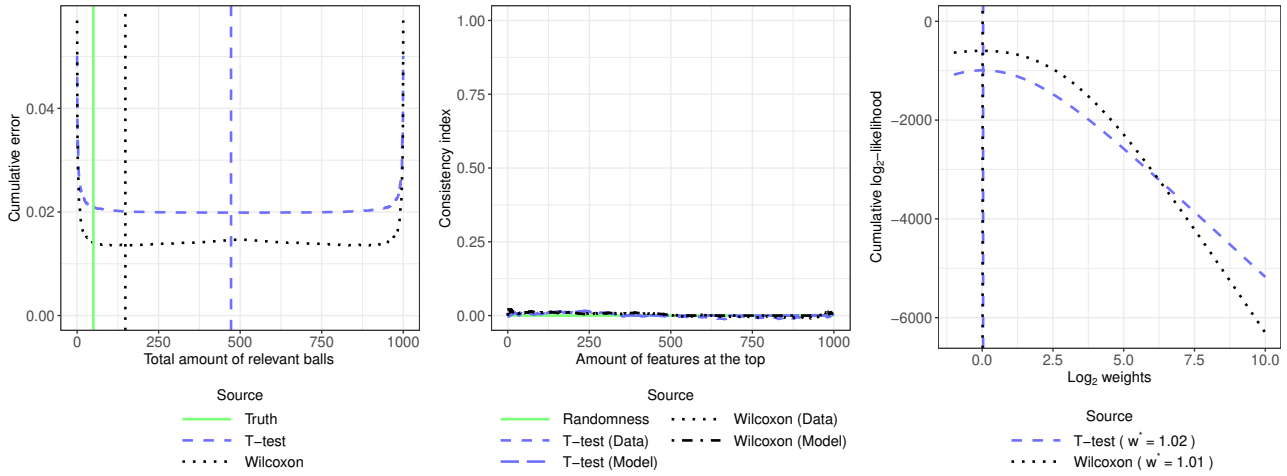
(a) **(b)** **(c)**

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



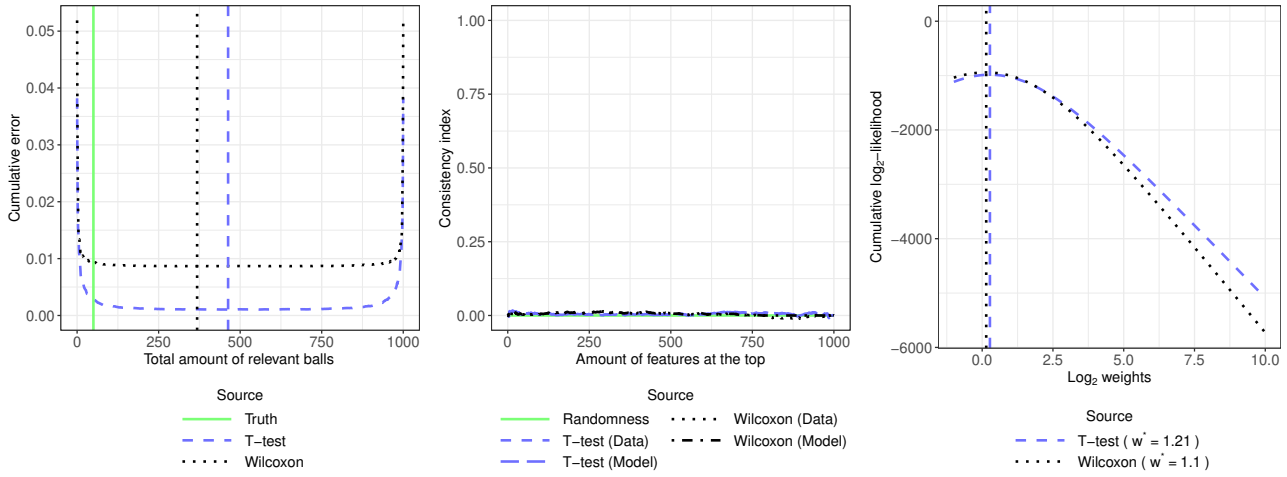
(a) **(b)** **(c)**

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



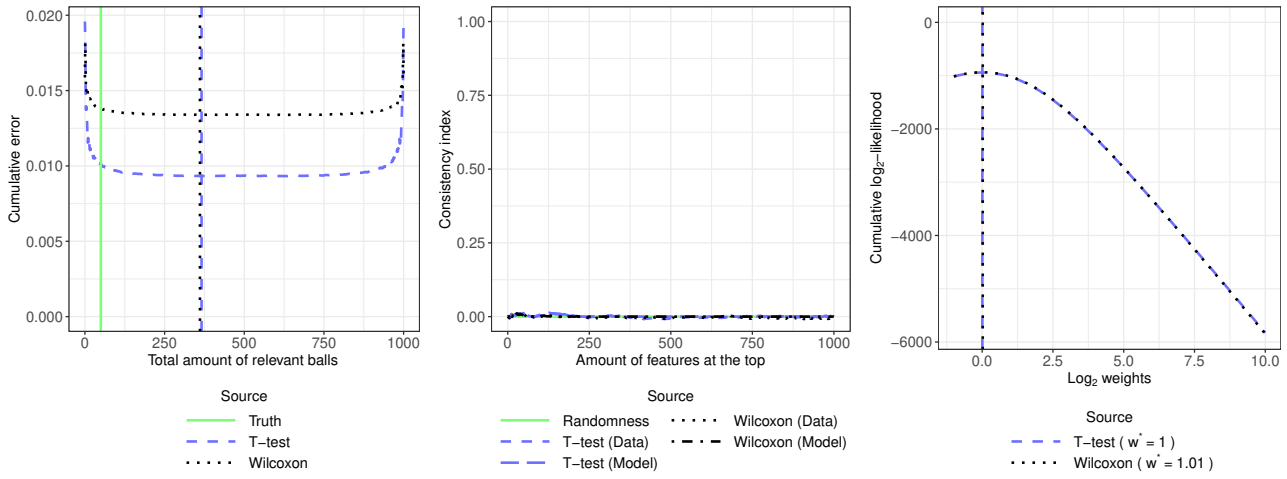
(a) **(b)** **(c)**

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



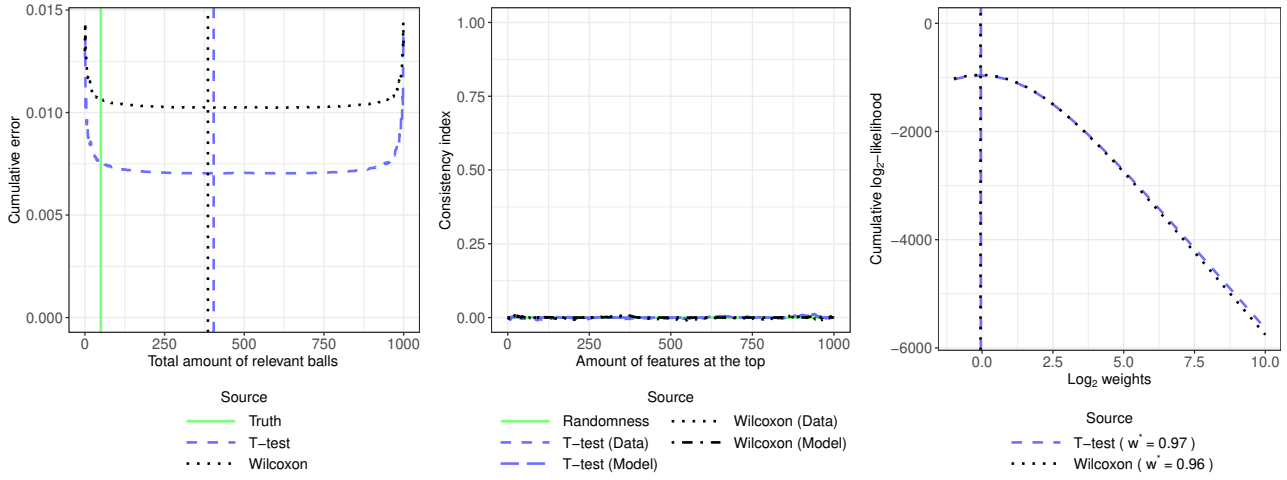
(a) **(b)** **(c)**

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



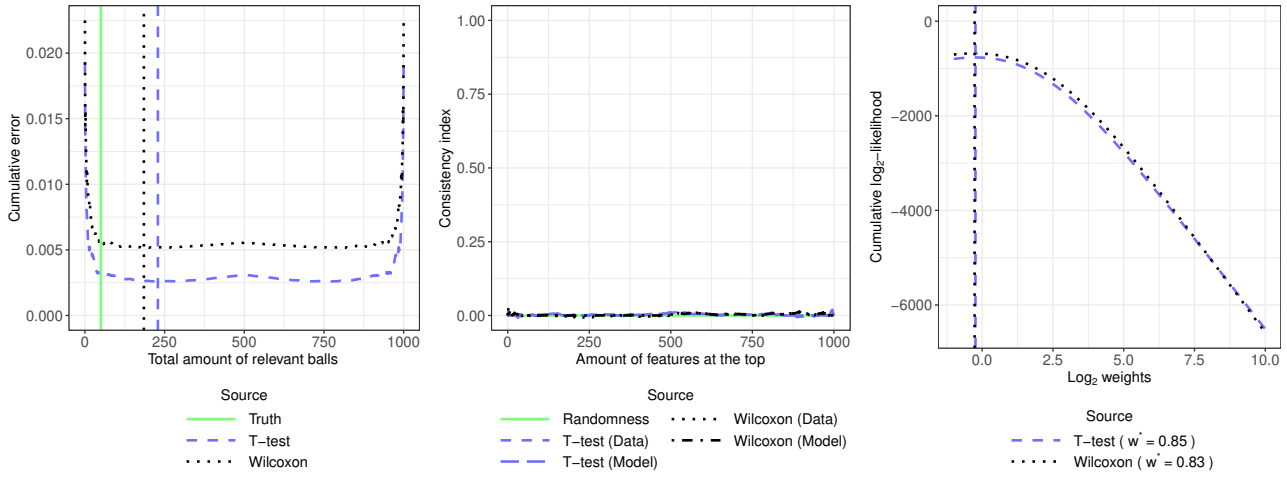
(a) **(b)** **(c)**

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

4.1 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].

4.2 Ovarian cancer database preprocessing

The preprocessing we applied to the ovarian cancer database is based on what was done by Wang et al [1]. 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].

5 Ovarian cancer database stratification

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

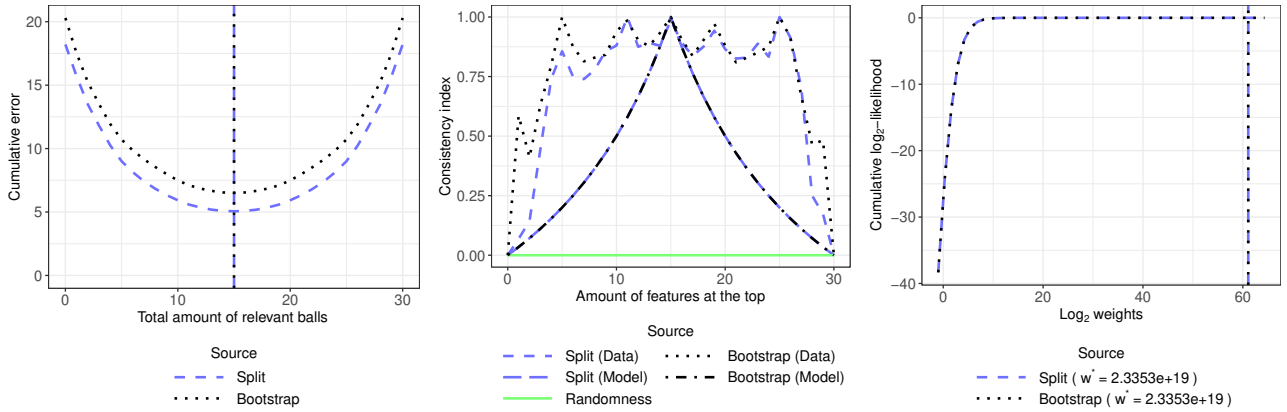
- Splitting in halves: In this sampling procedure, for each group its individuals are sorted according to their age. Then, for each group, each of its odd individuals is assigned randomly either to belong to $\boldsymbol{D}^1$ or to belong to $\boldsymbol{D}^2$. Finally, for each group, each if its even 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 $\boldsymbol{D}$.

6 Plots of the experimentation with real data

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

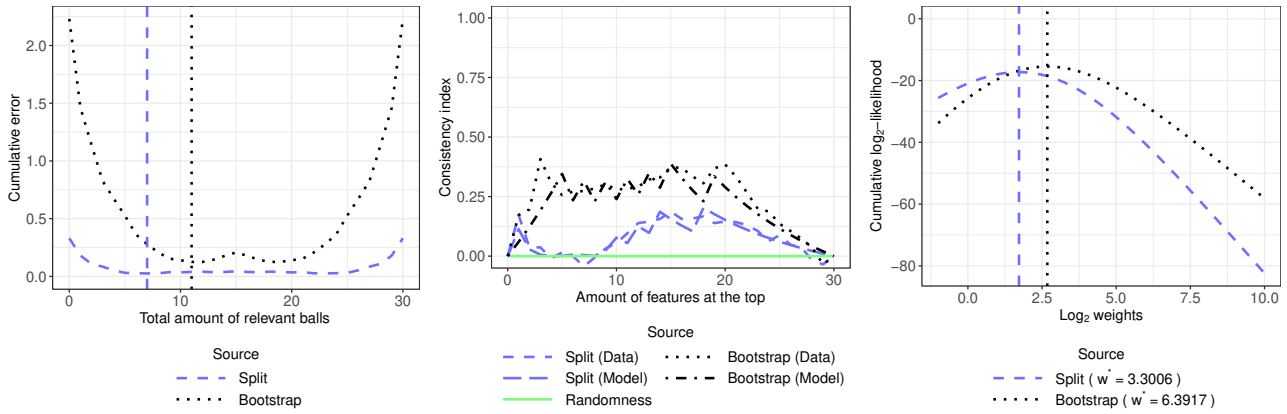
[1]This preprocessing step was included in order to enable the computation of the ranking method based on the coefficients of a linear SVM.

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



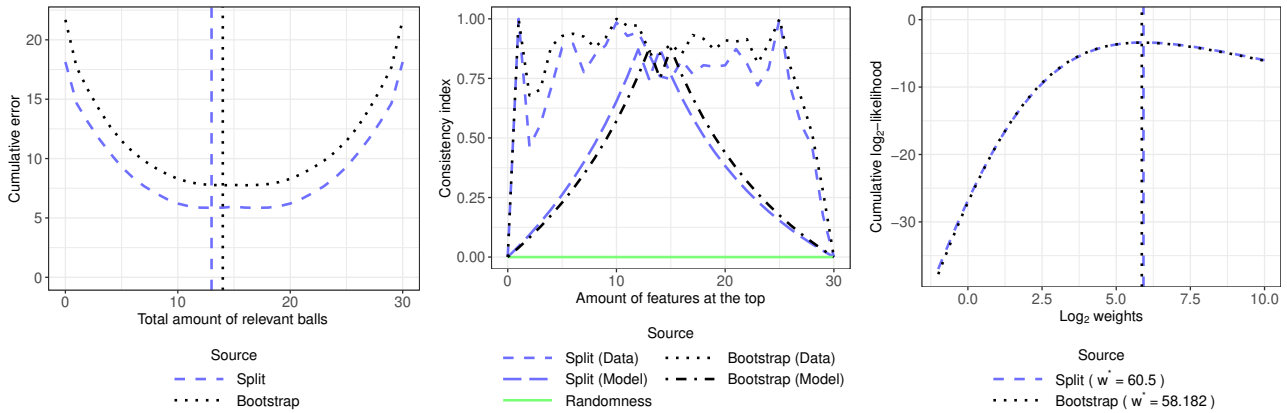
(a) **(b)** **(c)**

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



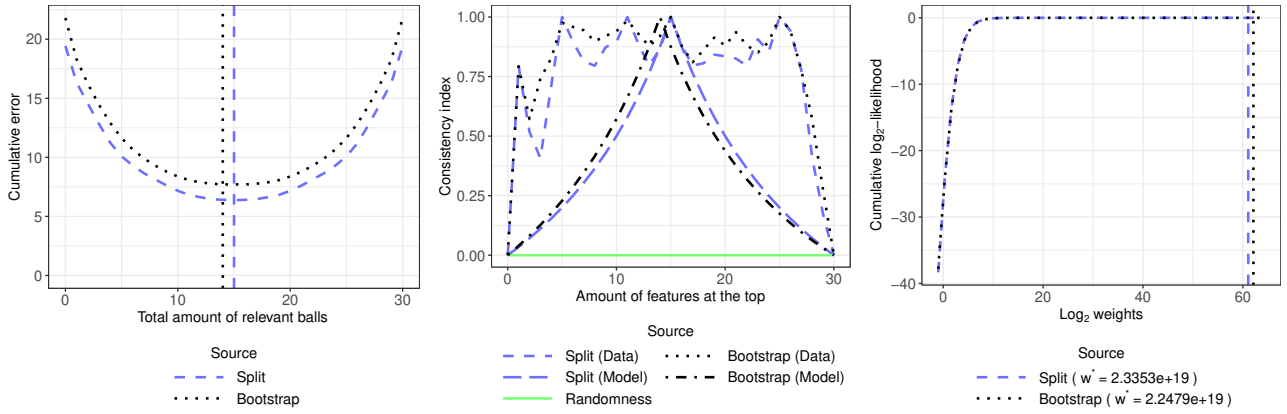
(a) **(b)** **(c)**

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



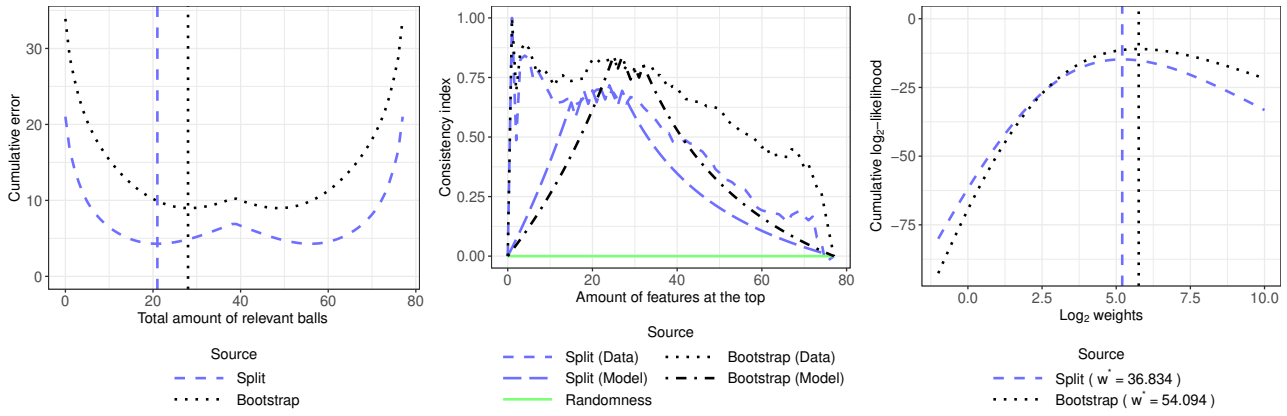
(a) **(b)** **(c)**

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



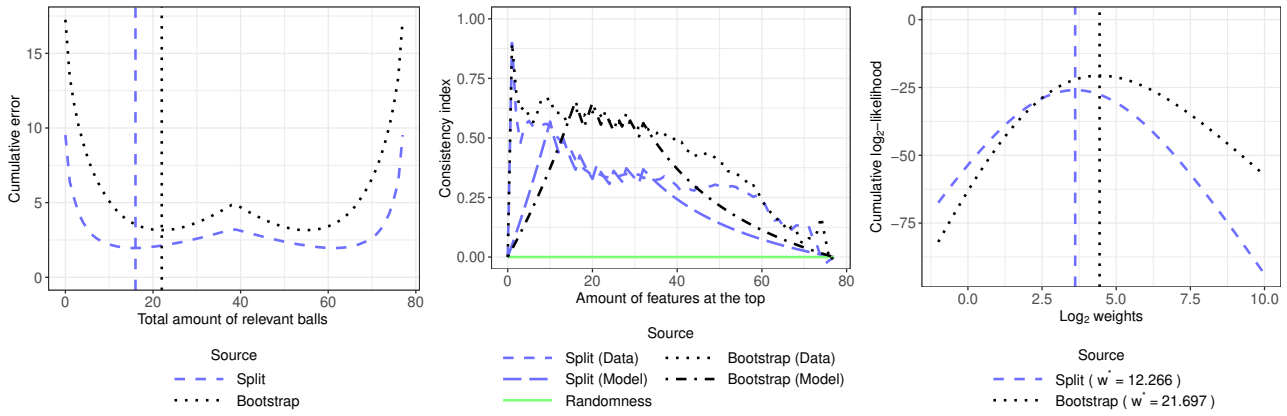
(a) **(b)** **(c)**

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



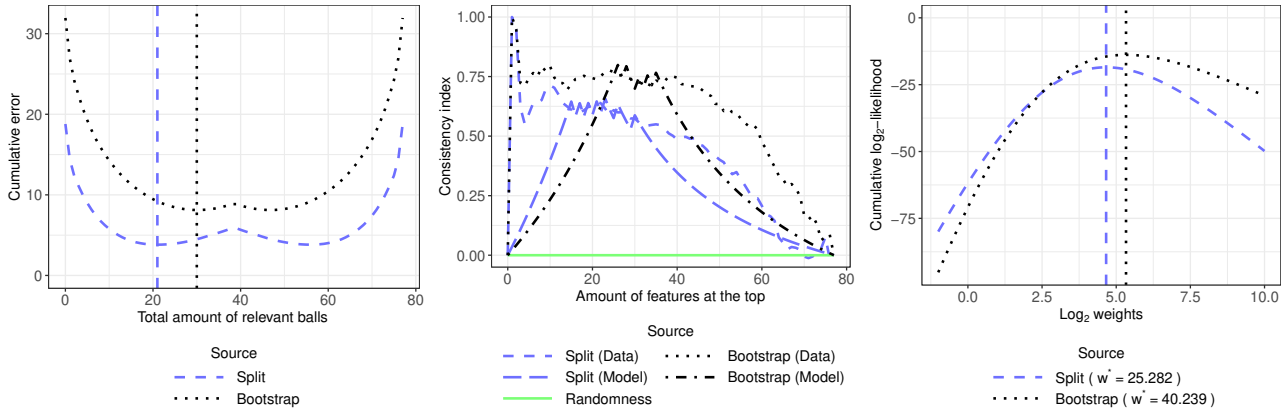
(a) **(b)** **(c)**

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



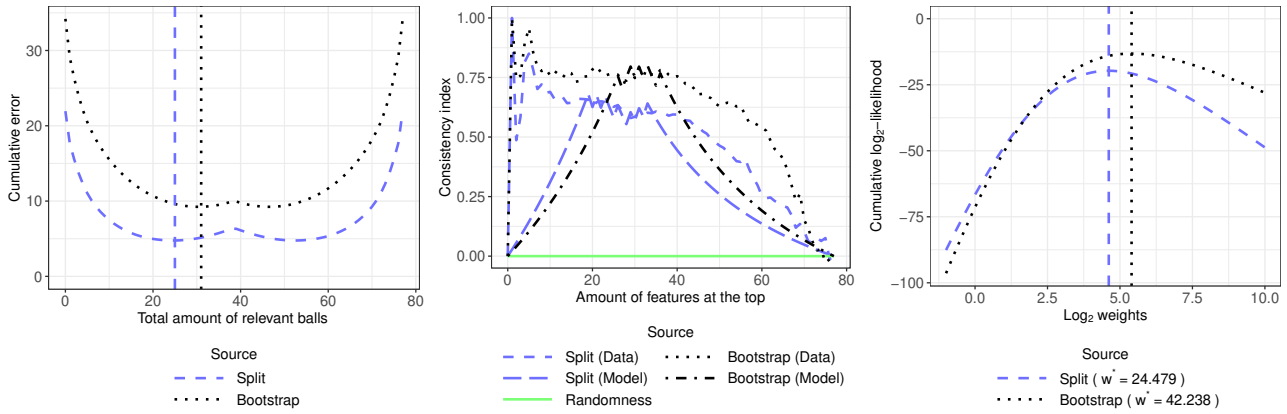
(a) **(b)** **(c)**

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



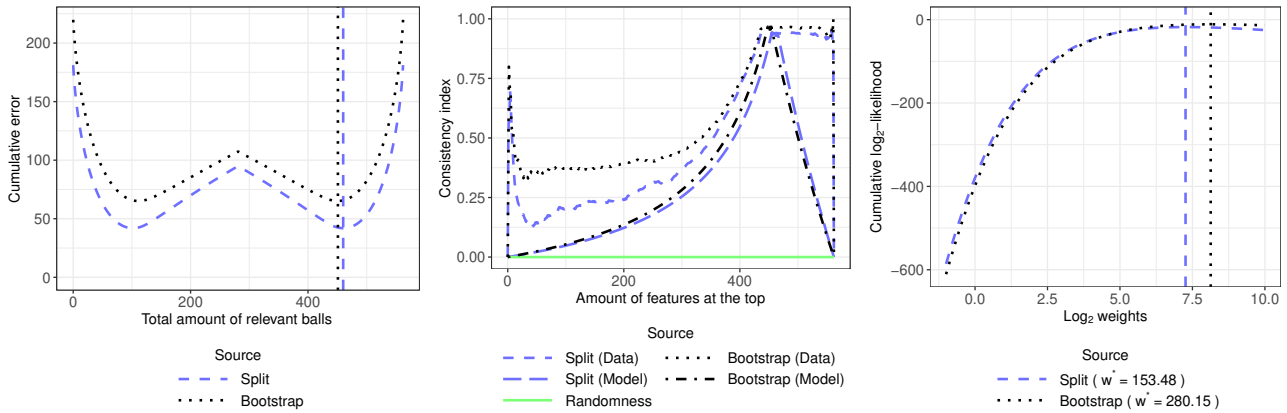
(a) **(b)** **(c)**

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



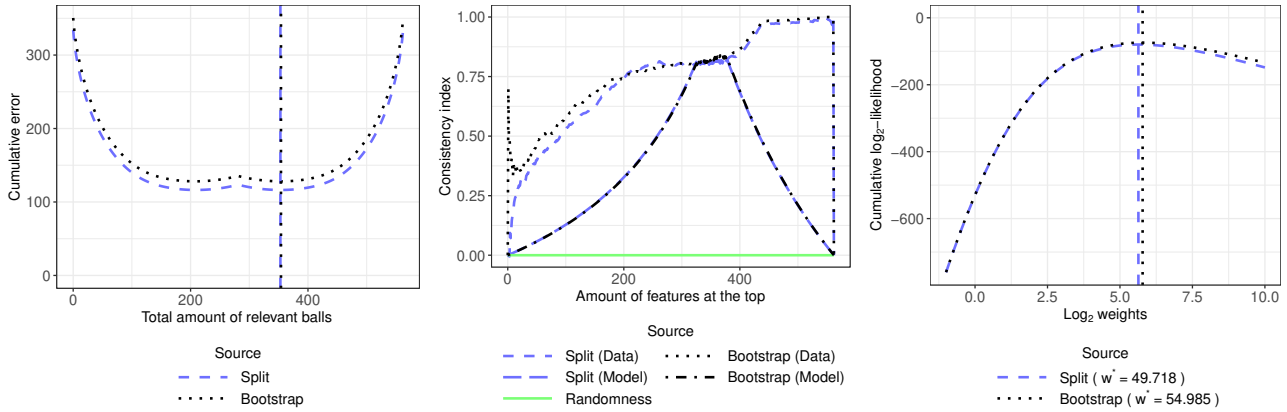
(a) **(b)** **(c)**

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



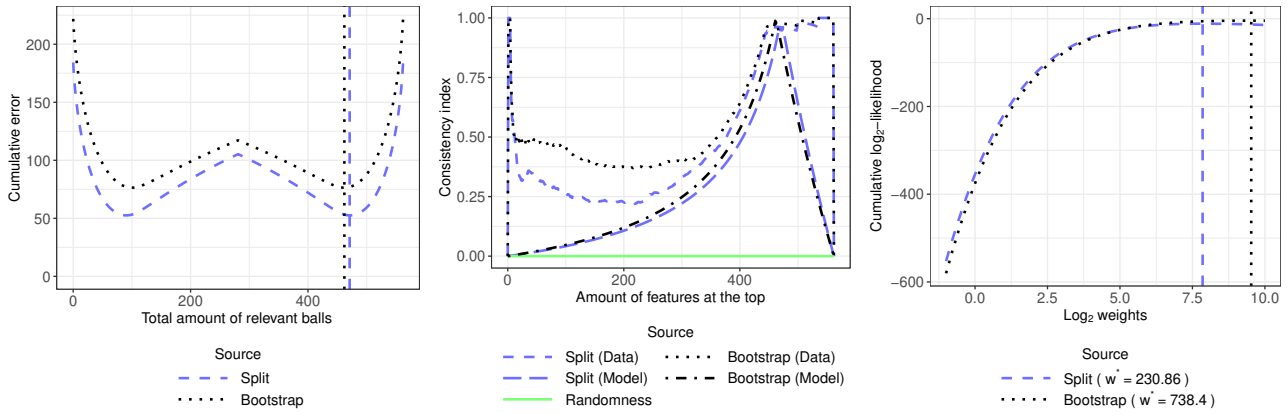
(a) **(b)** **(c)**

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



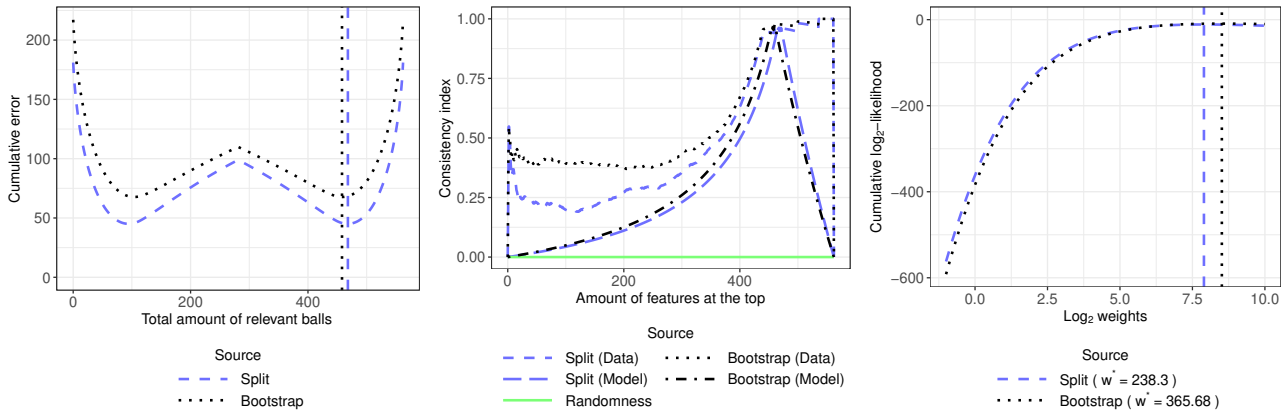
(a) **(b)** **(c)**

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



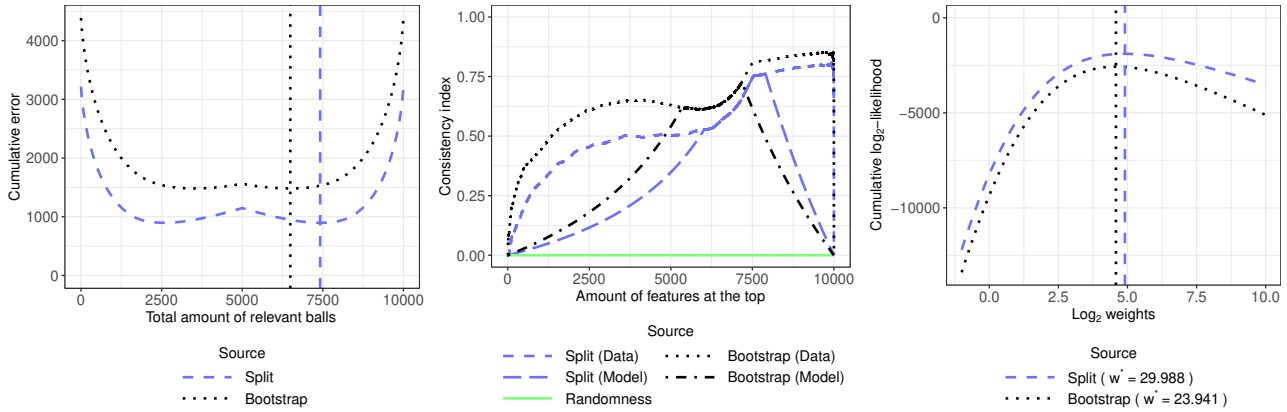
(a) **(b)** **(c)**

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



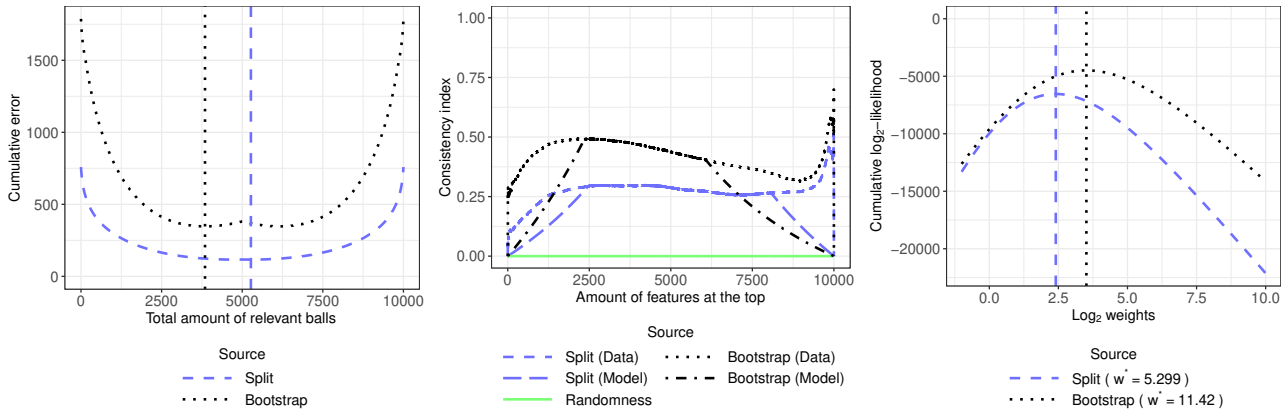
(a) **(b)** **(c)**

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



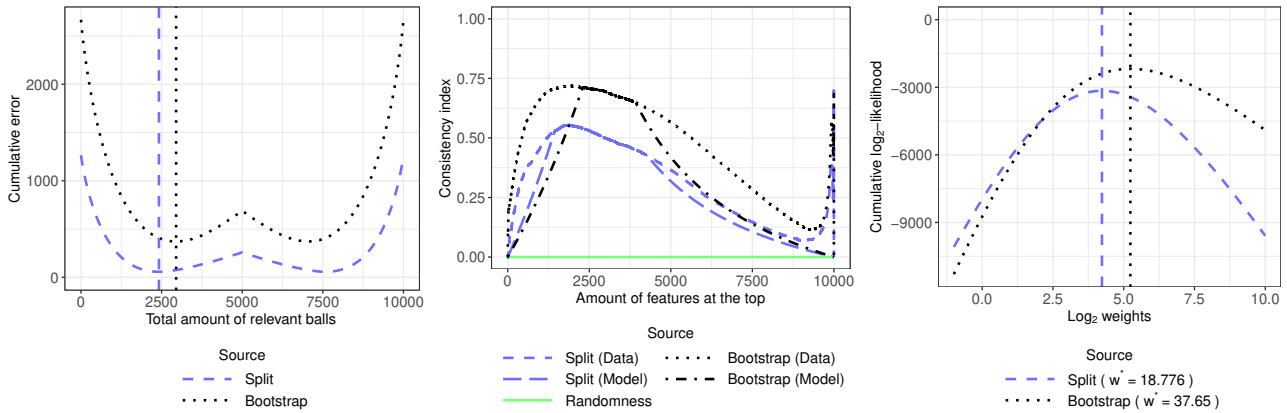
(a) **(b)** **(c)**

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



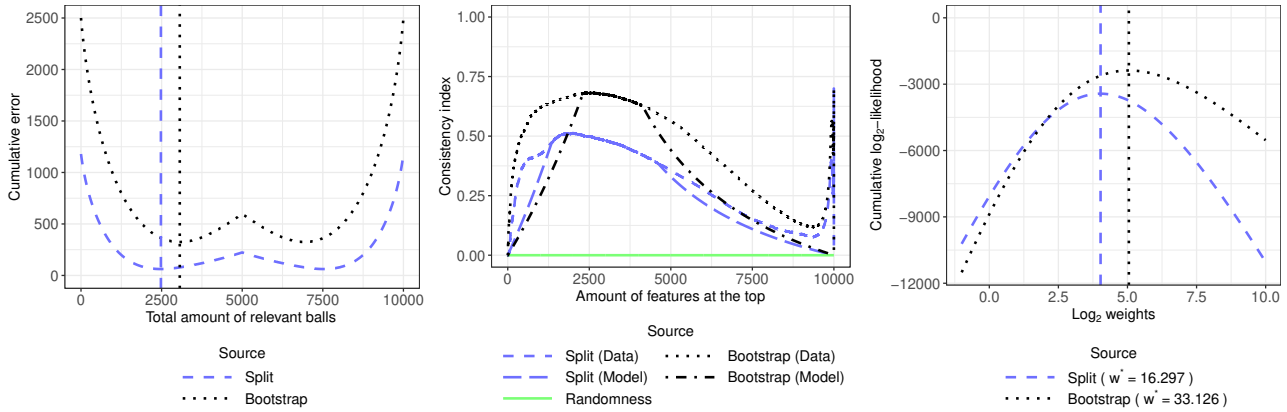
(a) **(b)** **(c)**

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



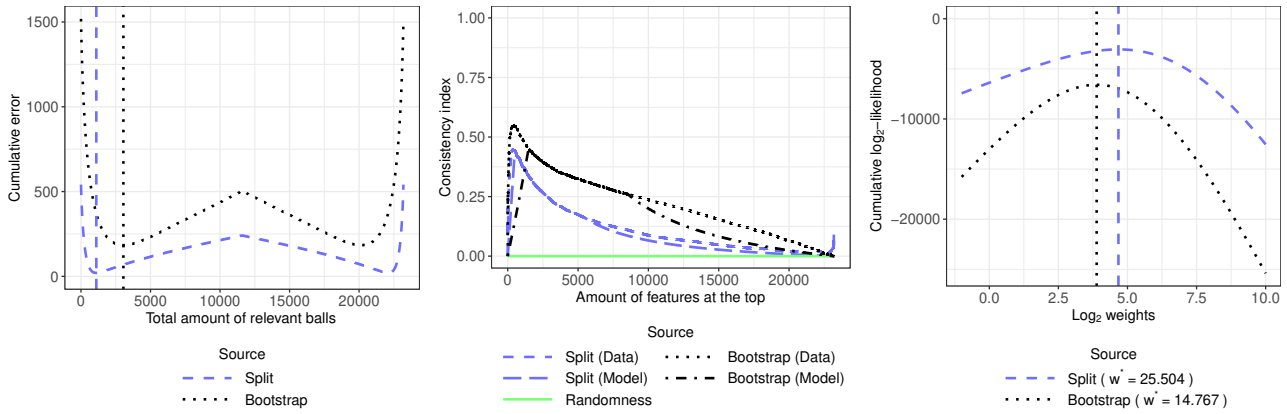
(a) **(b)** **(c)**

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



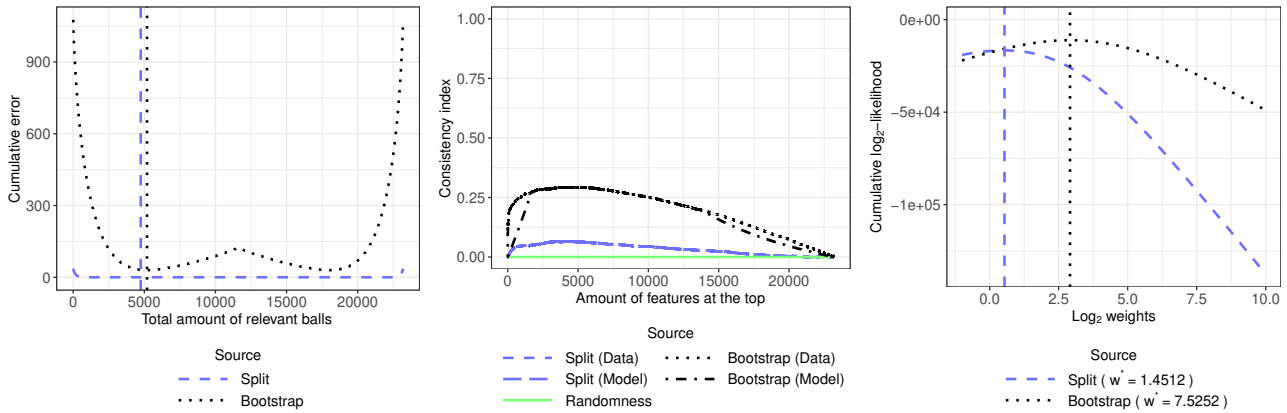
(a) **(b)** **(c)**

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



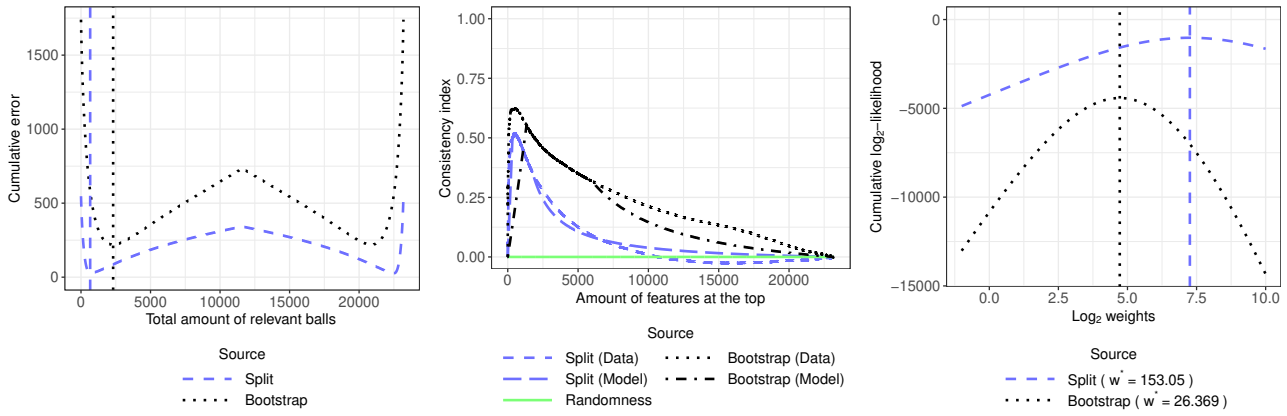
(a) **(b)** **(c)**

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



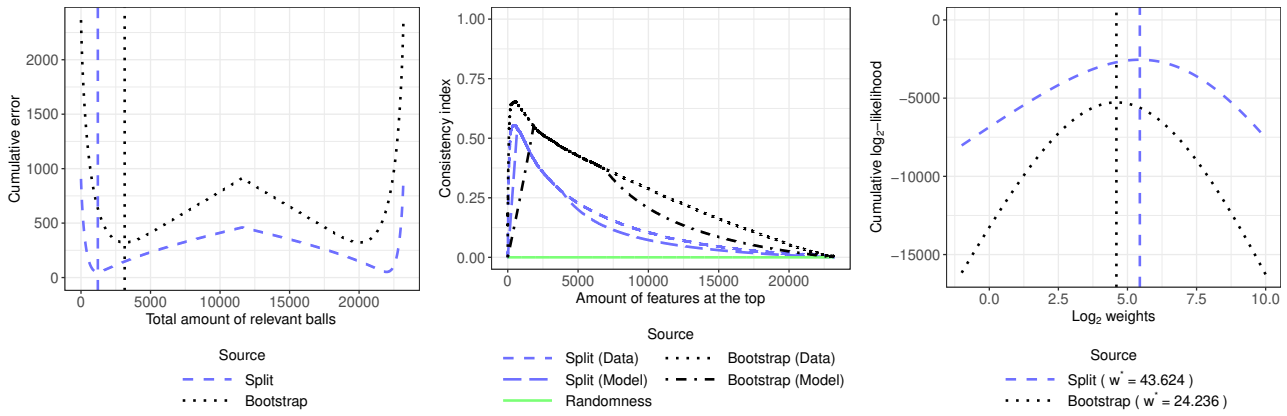
(a) **(b)** **(c)**

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

References

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