Technical Report
hdl.handle.net/10810/91

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

UNIVERSITY OF THE BASQUE COUNTRY
Department of Computer Science and Artificial Intelligence

Statistical model for the reproducibility in ranking based feature selection

Ari Urkullu Villanueva
Aritz Pérez Martínez
Borja Calvo Molinos

October, 2017

San Sebastian, Spain
www.ehu.eus/ccia-kzaa

Statistical model for the reproducibility in ranking based feature selection

Ari Urkullu Villanueva
Department of Computer Science
and Artificial Intelligence,
University of the Basque Country
UPV/EHU

Aritz Pérez Martínez
Department of Data Sciences,
Basque Center
for Applied Mathematics

Borja Calvo molinos
Department of Computer Science
and Artificial Intelligence,
University of the Basque Country
UPV/EHU

Abstract

Recently, concerns about the reproducibility of scientific studies have been growing among the scientific community, mainly due to the existing large quantity of irreproducible results. This has reach such an extent that a perception of a reproducibility crisis has spread through the scientific community (Baker, 2016). Among others, researchers point out “insufficient replication in the lab, poor oversight or low statistical power” as the reasons behind this crisis. Indeed, the A.S.A. warned almost two years ago that the problem derived from an inappropriate use of some statistical tools (Wasserstein & Lazar, 2016). Motivated to work on this reproducibility problem, in this paper we present a framework that allows to model the reproducibility in ranking based feature subset selection problems. In that context, among n features that could be relevant for a given objective, an attempt is made to choose the best subset of a prefixed size $i \in \{1, \dots, n\}$ through a method capable of ranking the features. In this situation, we will analyze the reproducibility of a given method which is defined as the consistency of the selection in different repetitions of the same experiment.

1 Introduction

In order to explain the context for which the model proposed in this paper has been developed, we will use a running example: The biomarker selection in genomic studies. A biomarker consists of “any substance, structure, or process that can be measured in the body or its products and influence or predict the incidence of outcome or disease”[1]. So, in the search for undiscovered biomarkers, many biomedical researches measure a large amount of candidate biomarkers (n) in individuals belonging to different groups. The search for candidate biomarkers typically uses some method to quantify the differences between groups (e.g., a statistical test) and selects a subset based on that quantification (e.g., setting a threshold, fixing the size of the subset, etc.).

In our work, we focus on problems with only two groups, control and disease, in which the selection of candidate biomarkers is made through a ranking of all the candidates so as to identify the i top ranked candidates. In this context, the expected reproducibility of the results of the method can be assessed as the similarity of different subsets obtained in several repetitions of the same experiment.

In summary, our main target is to analyze the methods by how reproducible the results they generate are. It is convenient to remark that, although we use the biomarker selection problem as an example, the same process is followed in other contexts where features are ranked and selected, e.g., ranking based feature subset selection.

This paper is organized as follows. In Section 2 we will explain a procedure to estimate empirically the reproducibility of the results of the method through two repetitions of the same experiment. Section 3 exposes the modeling of reproducibility curves. Section 4 poses how the model can be fitted to empirical data. Then, in Section 5 the experimentation in which the model has been fitted to different experimental data is explained and its results are described. Finally in Section 6 the main conclusions that have been drawn from this research and the future work possibilities will be discussed.

[1] Definition by the World Health Organization: http://www.inchem.org/documents/ehc/ehc/ehc222.htm

2 Empirical analysis of the reproducibility

In this section we present a measure which allows to assess the reproducibility of the results of a given method $\mathcal{M}$. In addition, we pose a procedure to estimate its expected value in scenarios where it can not be analytically derived. Finally, in order to illustrate its use, we estimate the expected value of the mentioned measure for the results of two classical statistical tests applied to real (genomic) datasets.

2.1 The reproducibility curves

Let us assume that we have a set of candidate biomarkers $\boldsymbol{X} = (X_1, \ldots, X_n)$ and a medical condition C which takes binary values $c \in \{+, -\}$ distributed according to some unknown probability distribution p over $(\boldsymbol{X}, C)$. Let us have N^+ i.i.d. samples drawn from $p(\boldsymbol{X}|C = +)$ and N^- i.i.d. samples drawn from $p(\boldsymbol{X}|C = -)$ in a dataset D to which a given method $\mathcal{M}$ is applied so as to obtain a ranking of the candidate biomarkers. We denote that ranking as a permutation $\sigma = (\sigma_1, \ldots, \sigma_n)$, where $\sigma_i = j$ denotes that X_j is in the i-th position of the ranking. Let us denote by $\sigma_{\leq i}$ the first i elements of σ, $\sigma_{\leq i} = (\sigma_1, \ldots, \sigma_i)$.

Let σ and σ' be two rankings obtained from two different datasets D and D' of samples obtained as explained above (i.e., N^+ drawn from $p(\boldsymbol{X}|C = +)$ and N^- drawn from $p(\boldsymbol{X}|C = -)$). We can define the random variable[2] L_i, which consists of the amount of coincidences between two top-i rankings derived from two random datasets. That is, given σ and σ', an observation of the random variable L_i can be computed as:

$$l_i = |\sigma_{\leq i} \cap \sigma'_{\leq i}|. \tag{1}$$

In addition, we define the random variable R_i, referring to it as top-i reproducibility, as the random variable L_i divided by i. Consequently, an observation of R_i can be computed as:

$$r_i = \frac{l_i}{i}. \tag{2}$$

We denote as ρ_i the expected top-i reproducibility given $\mathcal{M}$ and given the procedure of sampling datasets presented above:

$$\rho_i = \int_D \int_{D'} r_i p(D) p(D') dD dD', \tag{3}$$

where, abusing the notation, $p(D)$ represents the probability of D, which, given that D is made of N^+ i.i.d. samples from $p(X|C = +)$ and N^- i.i.d. samples from $p(X|C = -)$, is the product of the probabilities of these samples. In words, we define the expected top-i reproducibility as the expected proportion of candidate biomarkers that are present in both $\sigma_{\leq i}$ and $\sigma'_{\leq i}$ derived from any pair of datasets D and D' sampled as aforementioned. We define the expected reproducibility curve (or simply reproducibility curve) of $\mathcal{M}$ for pairs of datasets where each has N^+ samples of $p(X|C = +)$ and N^- samples of $p(X|C = -)$ as the sequence of points[3] $(0, 0)$, $(1, \rho_1)$, $(2, \rho_2)$, $\ldots$, (n, ρ_n).

2.2 Estimating the expected reproducibility curve

Unfortunately, in real situations p is unknown and we have a single dataset D of N^+ and N^- samples. In this subsection, we propose a procedure for estimating the expected reproducibility curve in this situation.

Given D, we can split it into two equally sized datasets D^1 and D^2 with $N^+/2$ and $N^-/2$ samples each. Now, the method $\mathcal{M}$ is applied to D^1 and D^2 to rank the candidates, and from these rankings $\boldsymbol{r} = r_1, \ldots, r_n$ is computed. In order to clarify this whole process, a graphical explanation is displayed in Figure 1.

By repeating this procedure t times using different random splits, we obtain t different $\boldsymbol{r}$ vectors. Then, we can make an estimation of the expected reproducibility curve $\boldsymbol{\rho}$ of $\mathcal{M}$ given D, which we denote as $\hat{\boldsymbol{\rho}}$, by averaging them:

$$\hat{\boldsymbol{\rho}} = \frac{1}{t} \sum_{k=1}^{t} \boldsymbol{r}^k, \tag{4}$$

Notice that with this scheme we are actually estimating the reproducibility of a method with datasets of half the size. In this work, we focus on the model and how it can be fitted and, thus, this is not a concern. For practical uses, in the future work section we will discuss alternative estimation procedures.

As an example, we have computed $\hat{\boldsymbol{\rho}}$ for two classical statistical tests, the t-test and the Wilcoxon rank sum test[4] in two different real-life datasets, an ovarian cancer database and a nephropathy database. These

[2] To denote a given random variable, a given observation of it and its expected value, we will use an uppercase letter, its corresponding lowercase letter and its corresponding Greek letter, respectively.

[3] Since the computation of this curve is straightforward given $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$, from here on we refer to this curve simply as $\boldsymbol{\rho}$.

[4] How adequate the selected methods are to rank the

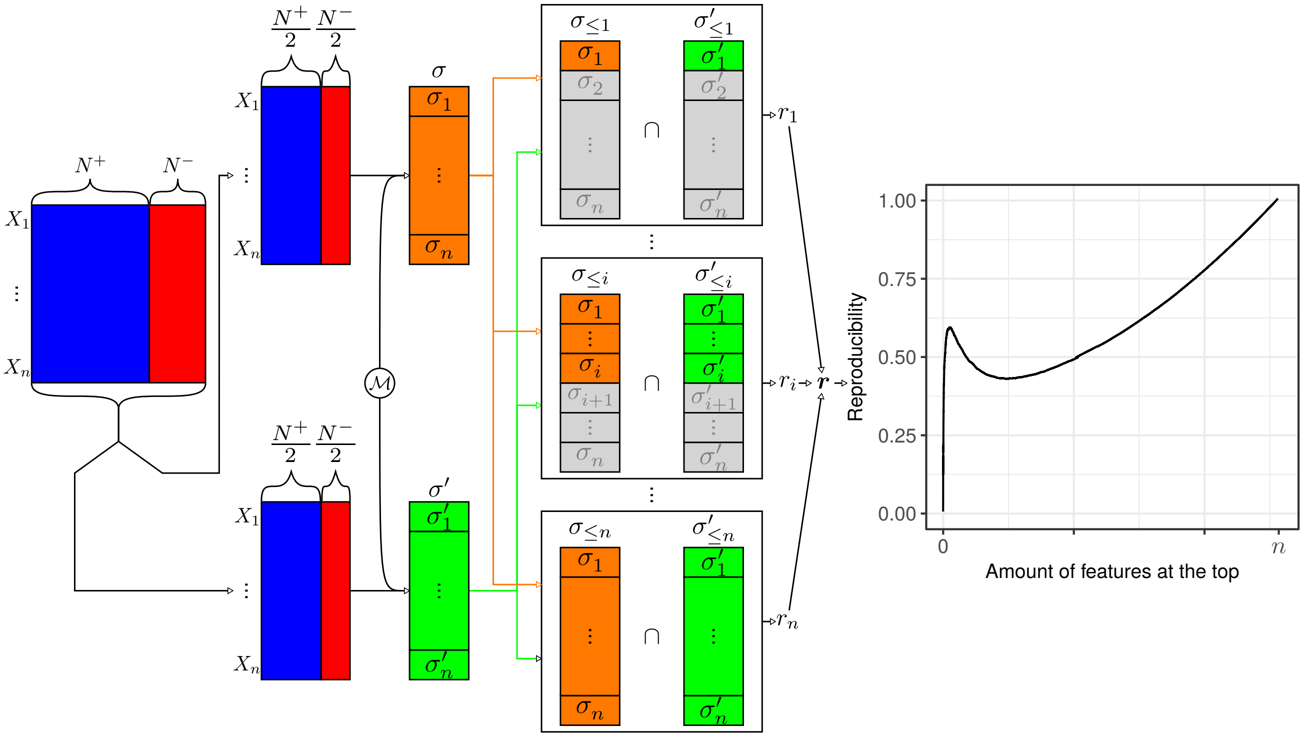


Figure 1: From a dataset to a "reproducibility curve".

two real datasets are DNA methylation datasets with over 27000 candidate biomarkers that are available at the GEO database[5]. Figures 2a and 2b show the different $\hat{\boldsymbol{\rho}}$ corresponding to the t-test and the Wilcoxon test applied to the two real datasets mentioned (with $t = 10$). Additionally, as a reference, Figures 2a and 2b include the $\boldsymbol{\rho}$ corresponding to a method that generates rankings uniformly at random[6].

In Figure 2a, both estimated reproducibility curves corresponding to the t-test and the Wilcoxon rank sum test start by rising very steeply until they flatten out and then each reaches a peak. Then they start decreasing and getting closer to the curve of a uniform random selection, to finally converge asymptotically to it. These results seem to match a scenario in which the methods consistently assess a few candidate biomarkers as more relevant than the rest of the candidate biomarkers. Consequently, they tend to appear in the first positions of the rankings consistently, while the orders of the vast rest of candidate biomarkers are frequently interchanged by the tests. Quite interestingly, we can see that the Wilcoxon estimated reproducibility curve is almost always above the t-test estimated reproducibility curve.

In contrast, in Figure 2b both methods show estimated reproducibility curves similar to a uniform random selection. One possible explanation is that the differences between groups are so small that the tests are barely able to detect them. Subsequently, any change in the sample leads to changes in the rankings the methods produce. An alternative possible cause which would lead to similar consequences could be that the methods are not designed to detect the type of differences underlying the sample (e.g., differences in variance). Another explanation is that the methods do not show preferences towards any feature and, therefore, they produce rankings at random.

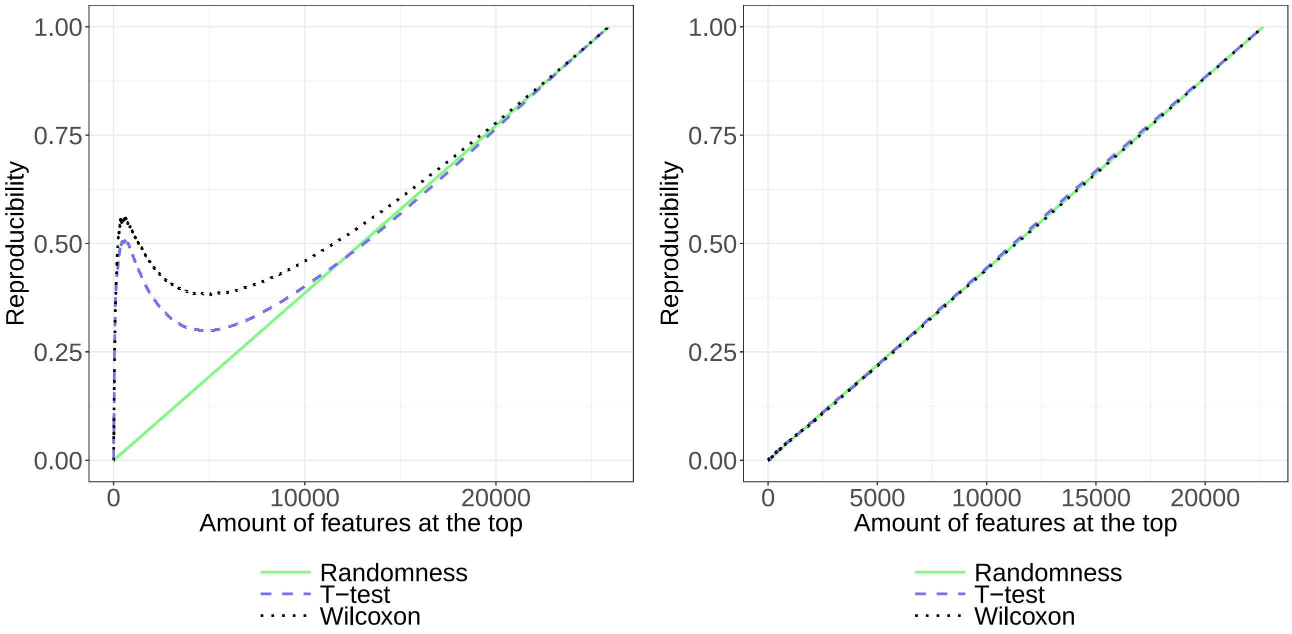
3 Modeling the reproducibility curves

Although the estimated reproducibility curves can be used for analyzing the reproducibility of a method that generates rankings of candidate biomarkers, it would be interesting to statistically model the reproducibility curves, in order to gain insights into the reproducibility of a given method by fitting the parameters of the model to the observed behavior. In this section, we propose a simple yet powerful model that can be used

candidates is irrelevant for our purpose of showing how our statistical approach to the reproducibility problem works. These methods have been selected because they are classical approaches to the problem.

[5] http://www.ncbi.nlm.nih.gov/geo; ovarian cancer database has accession number GSE19711; nephropathy database has accession number GSE20067.

[6] It can be easily calculated given that, for a uniform random selection, $l_i \sim \text{Hypergeometric}(l_i; n, i, i)$ and, thus, $\rho_i = i/n$.



(a) Ovarian cancer database

(b) Nephropathy database

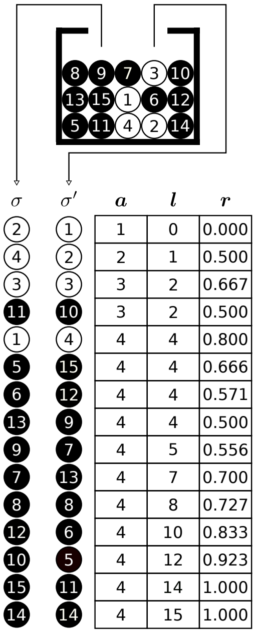
Figure 2: Example of the estimated reproducibility curves in the ovarian cancer database (a) and in the nephropathy database (b), together with the expected reproducibility curve when the candidates are selected randomly.

in order to analyze reproducibility curves.

The proposed model is based on an urn with n balls representing the n candidate biomarkers. The model assumes that there are two types of candidate biomarkers, those that present differences which are detectable by the method under study and those that do not. The candidates that are detectable are represented as white balls while the non-detectable candidates are represented as black balls. A complete extraction of the balls in the urn represents a ranking of the candidates and it will be denoted by a permutation, σ.

As a way of simplifying the model, we will assume that the amount of white balls in any top-i ranking $\sigma_{\leq i}$ is the same and we will denote it as a_i, for $i \in \{1, \ldots, n\}$. In concordance, the sequence of the amounts of white balls extracted is denoted as $\boldsymbol{a} = (a_1, \ldots, a_n)$. Taking by convention $a_0 = 0$, it is convenient to mention that, due to the nature of the process, a_i must be equal to or greater by one than a_{i-1} for $i \in \{1, \ldots, n\}$. A diagram is shown in Figure 3 so as to clarify this whole explanation.

Under the proposed model, it is assumed that the probability of extracting any specific white ball given that the extracted ball color is white is the same for each of the remaining white balls in the urn; an analogous assumption is done regarding the extraction of black balls. When $\boldsymbol{a}$ is known, that assumption makes



σ	σ'	$\boldsymbol{a}$	$\boldsymbol{l}$	$\boldsymbol{r}$
2	1	1	0	0.000
4	2	2	1	0.500
3	3	3	2	0.667
11	10	3	2	0.500
1	4	4	4	0.800
5	15	4	4	0.666
6	12	4	4	0.571
13	9	4	4	0.500
9	7	4	5	0.556
7	13	4	7	0.700
8	8	4	8	0.727
12	6	4	10	0.833
10	5	4	12	0.923
15	11	4	14	1.000
14	14	4	15	1.000

Figure 3: From the urn with two types of balls to $\boldsymbol{r}$.

it easy to compute, for any given $i \in \{1, \ldots, n\}$, the expected amount of coincidences between $\sigma_{\leq i}$ and $\sigma'_{\leq i}$, which we denote as λ_i. In order to compute this, first we decompose λ_i as the sum of the expected amount of coincident white balls, which we denote as λ_i^a, and the expected amount of coincident black balls, which we denote as λ_i^b. In fact, the random variable beneath λ_i^a, that is, the amount of coincident white balls between $\sigma_{\leq i}$ and $\sigma'_{\leq i}$, follows a hypergeometric distribution: $L_i^a \sim \text{Hypergeometric}(a_n, a_i, a_i)$, where the three parameters represent the population size, the amount of successes in the population and the amount of draws, respectively. In consequence, $\lambda_i^a = a_i^2/a_n$. An analogous procedure can be performed with λ_i^b: $L_i^b \sim \text{Hypergeometric}((n - a_n), (i - a_i), (i - a_i))$ and, thus, $\lambda_i^b = (i - a_i)^2/(n - a_n)$. Finally, λ_i can be computed as the sum of λ_i^a and λ_i^b, leading to the following expected top-i reproducibility under the model represented by $\boldsymbol{a}$:

$$\rho_i = \frac{a_i^2}{i \cdot a_n} + \frac{(i - a_i)^2}{i \cdot (n - a_n)}. \tag{5}$$

Note that the expected top-i reproducibility under the proposed model for $i \in \{1, \ldots, n\}$ is symmetric regarding the relative amount of white and black balls. However, in most practical scenarios, such as the biomarker selection, the relevant features (white balls, a_n) are far less than the irrelevant ones (black balls, $n - a_n$). Besides, it is worth mentioning that the sequence $\boldsymbol{a}$ can be used to compute a_i/i for any size $i \in \{1, \ldots, n\}$, which can be interpreted as the true positive rates associated to the selection of the top-i candidates.

4 Fitting the model to empirical data

This section is divided in two subsections, each one dedicated to a different stage of the fitting process. In the first one, a procedure is described to find the sequence $\boldsymbol{a}$ that best fits a given estimated reproducibility curve $\hat{\boldsymbol{\rho}}$ in terms of a cumulative error function E. In the second one, a procedure is presented to estimate quantitatively how often the white balls tend to be drawn before the black balls.

4.1 Finding the sequence $\boldsymbol{a}$ with the best fit

The main motivation for fitting the model to a given estimated reproducibility curve $\hat{\boldsymbol{\rho}}$ is to analyze the parameters of the fitted model. As aforementioned, sequence $\boldsymbol{a}$ can be interpreted in terms of the true positive rate. Additionally, the estimation of the amount of white balls a_n can be interpreted as the amount of candidate biomarkers that present differences detectable by the given method in the light of the given dataset.

Before explaining how the fitting can be undertaken, it is convenient to recall the set of constraints that any given sequence $\boldsymbol{a}$ must satisfy so as to be feasible. A given sequence $\boldsymbol{a}$ belongs to the set A of all the feasible sequences if and only if $a_i - a_{i-1} \in \{0, 1\}$ for $i \in \{1, \ldots, n\}$, assuming by convention that $a_0 = 0$. With those restrictions in mind, from here on we only deal with feasible sequences, unless explicitly stated otherwise.

In order to begin the fitting of the proposed model, we define a cumulative error function E. This cumulative error function E assesses the difference between a given estimated reproducibility curve $\hat{\boldsymbol{\rho}}$ and the expected reproducibility curve given a particular $\boldsymbol{a}$ (see Equation 5):

$$E(\hat{\boldsymbol{\rho}}, \boldsymbol{a}) = \sum_{i=1}^{n} e_i(\hat{\rho}_i, a_i, a_n), \tag{6}$$

where e_i is the quadratic difference between the estimated top-i reproducibility $\hat{\rho}_i$ and the expected top-i reproducibility ρ_i given $\boldsymbol{a}$ (expressed in Equation 5). Consequently we have:

$$e_i(\hat{\rho}_i, a_i, a_n) = \Big(\hat{\rho}_i - \Big(\frac{a_i^2}{i \cdot a_n} + \frac{(i - a_i)^2}{i \cdot (i - a_n)}\Big)\Big)^2. \tag{7}$$

Now, given the estimated reproducibility curve $\hat{\boldsymbol{\rho}}$, the problem consists of finding the feasible sequence $\boldsymbol{a}$ that minimizes the cumulative error function E:

$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a} \in A} E(\hat{\boldsymbol{\rho}}, \boldsymbol{a}). \tag{8}$$

In order to solve this problem, first we divide it into $n + 1$ subproblems, in each of which a_n has a fixed different value. So, any given subproblem specified by the constraint of a_n having a specific fixed value is solvable using dynamic programming through the next recursive function:

$$E_{a_i}^{i}(\hat{\boldsymbol{\rho}}) = e_i(\hat{\rho}_i, a_i, a_n) + \min(E_{a_i}^{i-1}(\hat{\boldsymbol{\rho}}), E_{a_i-1}^{i-1}(\hat{\boldsymbol{\rho}})), \tag{9}$$

departing from $E_{a_n}^{n}(\hat{\boldsymbol{\rho}})$, where $E_{a_i}^{i}(\hat{\boldsymbol{\rho}}) = \infty$ when $i < a_i$ or when $a_i < 0$ and $E_0^0(\hat{\boldsymbol{\rho}}) = 0$. When the $n + 1$ subproblems are solved, $n + 1$ cumulative error values are available. Additionally, for each subproblem, while it is being solved it is possible to gather the sequence $\boldsymbol{a}$ that solves it by noting the choices made in every step of the recursion in Equation 9. As a result, the sequence $\boldsymbol{a}^*$ that minimizes the cumulative error can be found searching for the $\boldsymbol{a}$ sequence whose associated cumulative error is the minimum among the $n+1$ computed ones.

Regarding the computational complexity, in order to find $\boldsymbol{a}^*$, $n+1$ dynamic programming problems are solved, one for each possible value of a_n. In addition, to solve each of these, n recursions are performed. In the worst cases each dynamic programming problem is solved in $\mathcal{O}(n^2)$, and, thus, the whole search for a^* has a computational complexity of $\mathcal{O}(n^3)$.

4.2 Modeling the differences between types of balls

So far we have modeled the empirical data as a sequence of extractions. With the aim of gathering further information about the reproducibility, we will model the sequence a^* using the process underlying the non-central hypergeometric distribution of Wallenius (Wallenius, 1963).

In this process we have an urn with white and black balls, but each type has a weight that biases the extraction. The balls are extracted sequentially and, at each step, the probability of extracting a white ball will be the total weight of the remaining white balls divided by the total weight of the remaining balls. As any common factor between both weights does not affect the probabilities, we will assume without loss of generality that the weight of a black ball is 1 and the weight of a white ball (or simply referred to as weight) is w.

Therefore, in this second stage we see $\boldsymbol{a}$ as the summary of the outcome of a complete sequence of draws that follows the process described above. In consequence, the likelihood of $\boldsymbol{a}$ given w can be seen as the product of the probabilities of obtaining a white or a black ball at each step of the sequence of extractions given w, the color of the ball depending on what $\boldsymbol{a}$ states for each extraction. This is, assuming the convention $a_0 = 0$, the likelihood of $\boldsymbol{a}$ given w can be expressed as:

$$\mathcal{L}(\boldsymbol{a}|w) = \prod_{i=1}^{n} \frac{(a_i - a_{i-1}) \cdot w \cdot (a_n - a_{i-1})}{w \cdot (a_n - a_{i-1}) + n - (i-1-a_{i-1})} + \frac{(1-(a_i - a_{i-1})) \cdot (n-(i-1-a_{i-1}))}{w \cdot (a_n - a_{i-1}) + n - (i-1-a_{i-1})}, \tag{10}$$

where $a_i - a_{i-1}$ determines whether in extraction i a white ball is extracted or not.

Given an $\boldsymbol{a}$, all the parameters except w are fixed and, as we can compute the likelihood of $\boldsymbol{a}$ given a certain w, we can look for the w that maximizes the likelihood of $\boldsymbol{a}$. This piece of information is very important due to its interpretation: The more reproducible the method, the higher the value of w. The weight also summarizes the degree of mixing between the white and black balls in the sequence of extractions, w becoming further away from 1 as the mixing decreases.

In order to carry out the search, an approximate value of w can be achieved through a search based on numerical analysis, like for instance the method of Brent. This approximation of w can be very quickly achieved compared with the search for $\boldsymbol{a}^*$ of the previous stage of the fitting process.

5 Experimentation

In this section, in order to illustrate the model and its use, we test the proposed model by fitting it to the estimated reproducibility curve $\hat{\boldsymbol{\rho}}$ derived from both synthetic and real data. Fitting the model to synthetic data enables the appropriateness of the model to be checked in controlled scenarios. Fitting the model to real data, enables the model behaviors to be tested in real situations within the context in which the model can be applied. In addition, it also can be used to gather information of the situations.

We have designed four different configurations for synthetic data, derived from the combinations of the two methods and two different scenarios. The two methods are the t-test and the Wilcoxon rank sum test. The two scenarios are defined by the kind of differences that the truly relevant candidate biomarkers show among groups, which may be differences in location or both in location and spread.

In each synthetic data configuration there are just 2 groups and 100 samples per group. For each, 1000 candidate biomarkers are simulated and only 50 of those are truly relevant biomarkers. The 950 non-relevant biomarkers are drawn from a normal distribution with $\mu = 0$ and $\sigma = 0.5$ for both groups, while the 50 truly relevant biomarkers are drawn from different distributions in each group. In one of those groups, the same normal distribution with $\mu = 0$ and $\sigma = 0.5$ is used, but in the other group a different distribution is used. Specifically, a normal distribution with $\mu = 0.35$ and $\sigma = 0.5$ is used if it must be only differences in location, while a normal distribution with $\mu = 0.25$ and $\sigma = 0.25$ is used if it must be differences both in location and spread. For each configuration of synthetic data, the estimated reproducibility curve $\hat{\boldsymbol{\rho}}$ has been estimated as explained in Section 2.2 (with $t = 10$).

Regarding the fitting of the model to real data, we have used the four estimated reproducibilities derived from the use of the statistical tests that appear in Figures 2a and 2b (Section 2.2).

In order to display the results of the models fitting for both the synthetic and the real data, we made four different types of plots per experiment:

- Reproducibility plots (Subfigures 4a, 5a, 6a and 7a): They have already been explained and displayed in Figures 2a and 2b.

- Error plots (Subfigures 4b, 5b, 6b and 7b): They display in their abscissa axis the total amount of white balls (a_n) that correspond to different dynamic programming problems. The ordinate axis shows the cumulative errors of the optimum solution for each problem. The vertical non-solid lines mark the a_n values with the minimum cumulative errors. In addition, in the case of synthetic data, since the true a_n value is known, a vertical solid line marks where that value is located.

- Proportion plots (Subfigures 4c, 5c, 6c and 7c): For each $i \in \{1, \ldots, n\}$ of the abscissa axis, the proportion of white balls in the top i is displayed in the ordinate axis. The lines show the sequences of proportions of the white balls derived from the $\boldsymbol{a}^*$ sequences issued by the model. In addition, in the case of synthetic data, since at which extraction each truly relevant biomarker is extracted is known, analogous curves can be computed departing from the data. Specifically, the 20 different sequences of extractions (2 per partition multiplied by the $t = 10$ repetitions of the partitioning) can be used to estimate the expected sequence of the proportions of extracted white balls. First, from each of those, a sequence of proportions of extracted white balls is computed, and then, the different computed sequences are averaged.

- Log-likelihood plots (Subfigures 4d, 5d, 6d and 7d): They display in their abscissa axis the different possible values of w while showing in the ordinate axis the log-likelihood of the sequence $\boldsymbol{a}^*$ given w. The vertical lines are used to show the locations of the w for which their log-likelihoods of $\boldsymbol{a}^*$ are maximum. In addition, the values of these w are displayed.

The results for synthetic data are shown in Figures 4 and 5, while Figures 6 and 7 show the results for real data.

5.1 Discussion

Now we briefly discuss the results derived from the experimentation. To start with, in the light of the experimentation with synthetic data, it can be seen that, as expected, the proposed model for reproducibility curves fits well to empirical data coming from two types of candidates (with and without differences between groups). In addition, it seems convenient to note that the model fits well regardless of the assumption that the amount of white balls in the top-i of any sequence of extractions is the same for any $i \in \{1, \ldots, n\}$.

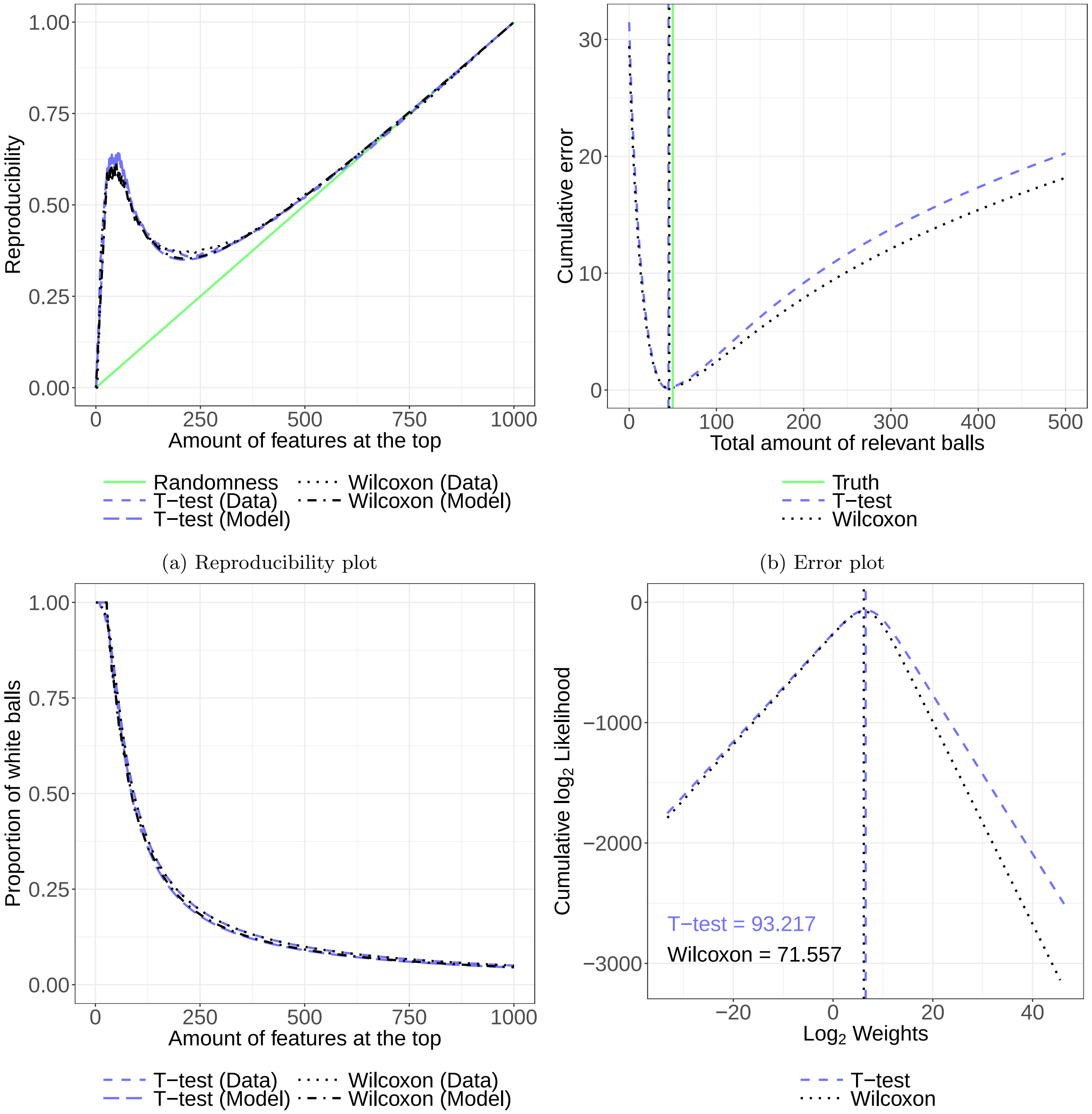
However, the fittings to the real data give results that are not as good as the results achieved with synthetic data. In fact, in real data it is very likely to occur that there are no longer just two types of candidate biomarkers. As a consequence, the model proposed may be too simple to properly represent this situation. Regarding the different weights obtained, a big difference between the two datasets can be seen, the w values issued for the ovarian cancer dataset being far bigger than the w values issued for the nephropathy dataset.

6 Conclusions

Paying attention to the reproducibility of the methods used in scientific studies is essential to ensure sound conclusions. Motivated by this concern, we have presented a statistical approach to analyze the reproducibility of ranking based selection methods. In addition to the model, we have exposed a way in which it can be fitted to experimental data. Then, so as to illustrate its behavior and the fitting process, we have used the problem of biomarker selection as an example. After testing the behavior of the model with both synthetic and real data, we have drawn some conclusions.

Regarding the results of the experimentation with synthetic data, we recall the achievement of good results despite the simplification about the amount of white balls drawn from both sequences of extractions. This implies that the restriction does not have a great impact on the results.

In the case of real data, specifically in the results of the ovarian cancer dataset, the model does not fit the empirical curve correctly. A likely explanation for this is that in real-life problems there are more than just 2 types of candidates and, thus, the model is not flexible enough to represent that situation. In particular, it seems that the model for reproducibility curves tends to issue a total amount of white balls a_n at a point of equilibrium. This idea can be seen in the results for



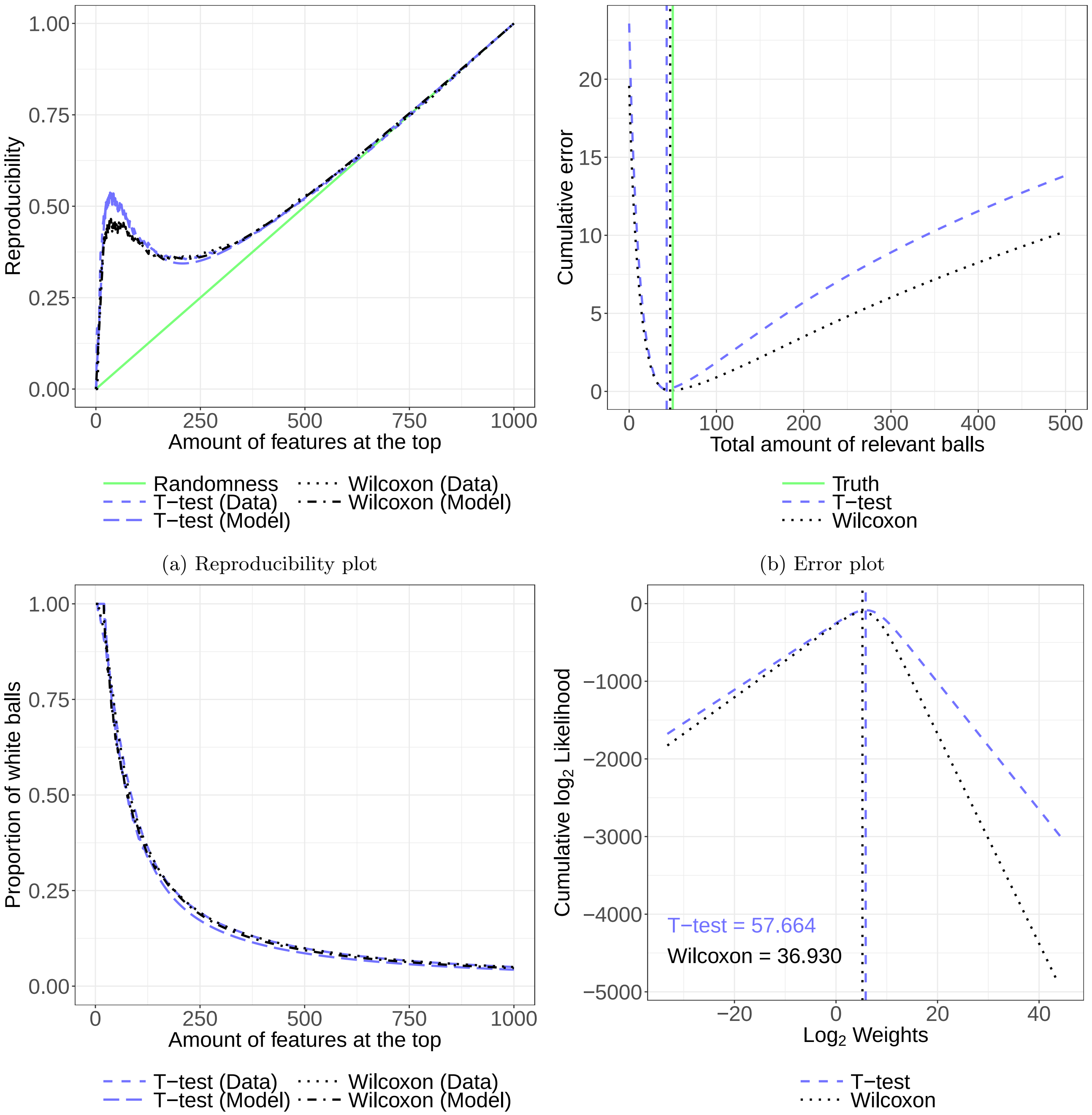
(a) Reproducibility plot

(b) Error plot

(c) Proportion plot

(d) Weight plot

Figure 4: Plots for the synthetic problem in which the data shows only differences in central tendency.



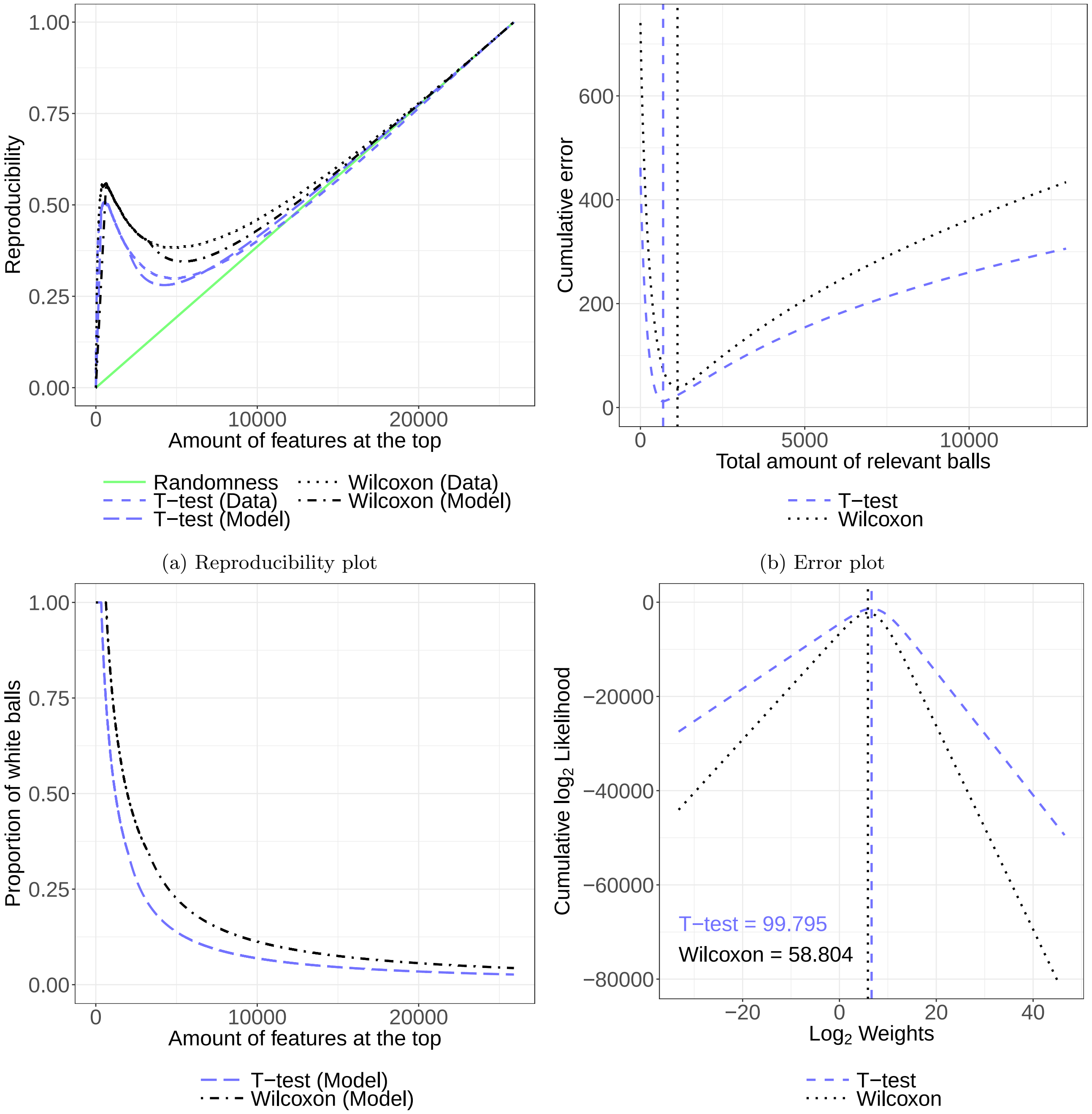
(a) Reproducibility plot

(b) Error plot

(c) Proportion plot

(d) Weight plot

Figure 5: Plots for the synthetic problem in which the data shows differences both in central tendency and variability.



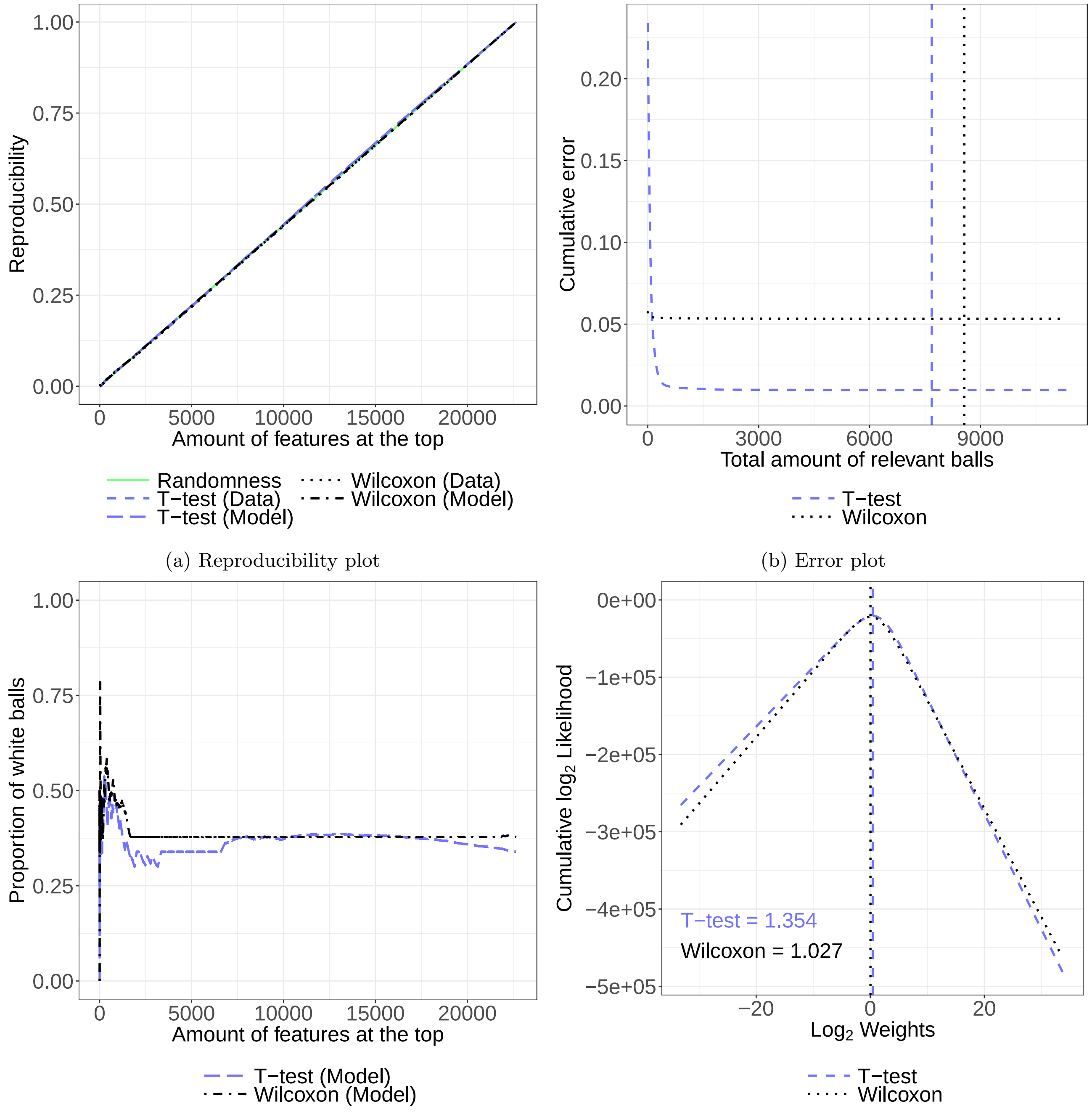
(a) Reproducibility plot

(b) Error plot

(c) Proportion plot

(d) Weight plot

Figure 6: Plots for the ovarian cancer database.



(a) Reproducibility plot

(b) Error plot

(c) Proportion plot

(d) Weight plot

Figure 7: Plots for the nephropathy database.

the ovarian cancer database. Namely, it seems that if a_n was smaller, then the reproducibility curve would rise faster in the first tops. However, after peaking it would fall earlier than it does because it would run out of white balls to extract. Similarly, it seems that if a_n was bigger, then the reproducibility curve would not decrease so fast after the peak, but it would not rise as fast before the peak as it does.

Regarding the different weights obtained, a big difference between the w values issued for the ovarian cancer dataset and the w values issued for the nephropathy dataset can be seen. In fact, in the nephropathy database both weights for the t-test and the Wilcoxon rank sum tests are close to 1, suggesting that both methods rank the candidates almost at random. One possible cause is that both tests are apparently unable to detect differences, this may be because the differences present are small enough or because the nature of the differences is undetectable for the methods. Regarding the w values for the ovarian cancer database, it is noteworthy that although the Wilcoxon rank sum test obtains an a_n greater than the t-test, it gets a lower weight. It is possible that in the tops (or in the majority of them) in which the t-test obtains white balls, the Wilcoxon rank sum test also obtains white balls. Moreover, it is possible that the additional white balls obtained by the Wilcoxon rank sum test are obtained in further tops. Thus, it could be interesting so as to compare the w values better, compare them after computing them for sequences in which the amount of white balls is the same, painting the leftover white balls located in the later tops as black balls.

Our research opens several future lines of work. One possible way to proceed can be to extend the model to more than just two types of balls, which most likely will increase the precision of the fittings in real data at the cost of increasing the computational complexity. As an approach to this, we could first fit the model considering two types of balls and then use its minimum error solution as a departing point for the fitting of a model considering three types of balls, for instance.

Another important line to follow is the estimation of the expected reproducibility curve. At this point we are splitting the dataset into two partitions, leading to an evident (pessimistic) bias in the estimations. In the future, we will explore other schemes, such as bootstrapping, which may ease this problem at the cost of introducing dependencies between the two sets of samples.

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, **533**(7604), 452.

Wallenius, K.T. (1963). *Biased sampling; the noncentral hypergeometric probability distribution* (Doctoral dissertation, Stanford University, California, USA). Retrieved from http://www.dtic.mil/docs/citations/AD0426243

Wasserstein, R.L. & Lazar, N.A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, **70**(2), 129–133.