Knowledge and Information Systems manuscript No.
(will be inserted by the editor)

Statistical model for reproducibility in ranking based feature selection

Ari Urkullu · Aritz Pérez · Borja Calvo

Received: date / Accepted: date

Abstract The stability of feature subset selection algorithms has become crucial in real-world problems due to the need for consistent experimental results across different replicates. Specifically, in this paper, we analyze the reproducibility of ranking based feature subset selection algorithms. When applied to data, this family of algorithms builds an ordering of variables in terms of a measure of relevance. In order to quantify the reproducibility of ranking based feature subset selection algorithms, we propose a model that takes into account all the different sized subsets of top-ranked features. The model is fitted to data through the minimization of an error function related to the expected values of Kuncheva's consistency index for those subsets. Once it is fitted, the model provides practical information about the feature subset selection algorithm analyzed, such as a measure of its expected reproducibility or its estimated area under the receiver operating characteristic curve regarding the identification of relevant features. We test our model empirically using both synthetic and a wide range of real data. The results show that our proposal can be used to analyze feature subset selection algorithms based on rankings in terms of their reproducibility and their performance.

Keywords Feature selection · Stability · Reproducibility · High dimensionality

Ari Urkullu
Paseo Manuel de Lardizabal, 1, 20018, Donostia, Gipuzkoa, Spain
Tel.: +34-943-018070
Fax: +34-943-015090
E-mail: ari.urkullu@ehu.eus
Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU)
ORCID: 0000-0002-8597-3260

Aritz Pérez
Alameda Mazarredo, 14, 48009, Bilbao, Bizkaia, Spain
Department of Data Science, Basque Center for Applied Mathematics (BCAM)
ORCID: 0000-0002-8128-1099

Borja Calvo
Paseo Manuel de Lardizabal, 1, 20018, Donostia, Gipuzkoa, Spain
Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU)
ORCID: 0000-0001-9969-9664

1 Introduction

Due to the large quantity of irreproducible results, concern has arisen to such an extent that a perception of a reproducibility crisis has spread through the scientific community [4]. Among other factors, researchers point out insufficient replication in the original laboratory, poor oversight, and low statistical power or poor analysis as the reasons for this crisis. Moreover, researchers identify better understanding of statistics, better mentoring and more robust designs as some of the possible solutions to boost reproducibility. Indeed, the American Statistical Association (ASA) warned recently about the problems derived from the inappropriate use of some statistical tools [41].

In this work, we tackle the feature selection problem, a problem in which the previously mentioned concerns regarding reproducibility are also present. Specifically, in this paper, we focus on problems in which the selection of features is made through a ranking (derived through a feature subset selection algorithm based on rankings) of all the features so as to identify the i top-ranked features. In brief, the aim of this paper is the proposal and testing of a statistical approach that enables the analysis of the reproducibility of ranking based feature subset selection (RFSS) algorithms. In order to illustrate and test our proposal, in addition to the use of synthetic data, we use several real datasets. Some of these belong to the biomarker selection problem because it is a problem that belongs to the area in which the reproducibility crisis is most apparent [3, 13].

In summary, our proposal for RFSS algorithms consists of a framework that allows us to quantify the reproducibility of the outcomes of such algorithms. In addition, it also enables us to gather information of the performance of a given feature subset selection method. Specifically, feature subset selection methods can be seen as classifiers that assess the features they select as interesting, and the rest of the features as non-interesting. In the case of RFSS methods, the resulting rankings can be seen as orderings of the features according to how likely they are to be relevant from the point of view of the given RFSS methods. Consequently, at least in dichotomous problems, the performance of a given RFSS algorithm can be assessed through the AUC in terms of the identification of relevant features.

Briefly, our proposal can be summarized as follows. Our approach takes as input two rankings of features provided by a feature ranking algorithm when applied to different datasets sampled from the same population. With such input, our proposal starts with the assessment of the similarity between the two rankings through the computation of a curve. Succinctly, such a curve measures the consistency index of Kuncheva [23] for each possible pair of top-ranked feature subsets (of the same size) derived from the two rankings. Next, the proposed reproducibility model, which considers that each feature has one of two different possible degrees of relevance, is fitted to the curve. Finally, the parameters of the fitted model enable us to obtain an AUC in terms of the detection of relevant features for the given method. We hypothesize that the estimated AUC derived from the parameters of the model is correlated with the estimation of the true AUC (which should be derived from the data), a hypothesis we have checked during the experimentation with synthetic data. Such information is valuable for scientific research in terms of aiming for efficient research regarding time, effort and money.

This paper is organized as follows. First, Section 2 focuses mainly on describing the process under study in this work and on presenting research on the reproducibility in the feature selection problem. In Section 3, we will explain a procedure to empirically estimate the reproducibility of the results of a RFSS method through two repetitions of the same experiment. Section 4 exposes the model for reproducibility curves. Section 5 poses how the model can be fitted to empirical data and how further information of interest can be derived from the parameters of the fitted model. Then, in Section 6, the experimentation conducted both with synthetic data and real data is explained and the results are described. Finally, in Section 7, the main conclusions that have been drawn from this research and future work possibilities will be discussed.

2 Background and related work

This section is divided into two subsections. First, Subsection 2.1 is dedicated to the description of the process under study in this work. Secondly, Subsection 2.2 focuses on the description of research related to the problem under analysis in this paper.

2.1 Background

Currently, there is no firm consensus about what reproducibility means exactly [4,13]. In this work, we stick to the definition of the reproducibility of results provided by Goodman et al. [13]. They basically stated that reproducibility of results consists of obtaining the same results as a given prior study when conducting an independent study collecting new data and following, as closely as possible, the procedures of the prior study. Specifically, we focus on reproducibility of results when the different datasets collected in different studies are sampled from the same population. In such conditions, although reproducing experimental results may seem a trivial task initially, in a recent *Nature*'s survey more than half of the researchers declared to have failed to reproduce their own experiments [4]. Moreover, the stochasticity that the sampling procedure generally implies makes it difficult to state whether the original results have been reproduced or not. Furthermore, there is no consensus either on what a successful reproduction of results consists of [13]. For the sake of brevity, from now on we also refer to such reproducibility of results under those conditions simply as reproducibility.

The aforementioned concerns regarding reproducibility are also present in the feature selection problem. Specifically, in order to assess the reproducibility of a given feature selection algorithm, the stability (sensitivity to variations in the data) of its outcomes is measured [8,19,20,29,30,32,34]. Given a certain problem in which there is a specific objective (e.g., classification, clustering, knowledge discovery, ...), the feature selection problem is normally conceived as a subproblem or step in which, among all the features available, the aim is generally to select the most interesting features regarding the objective, while discarding the rest. That search of the most interesting features may be carried out for many reasons [8,14,34], such as the improvement of the interpretability or the generalization capability of a given classification model. The many algorithms that have been proposed to tackle the feature selection problem can usually be divided into three major types according to which sort of outcome they produce: weight-scores for the features, rankings of features or subsets of features [8,19,20,29,30,32,34]. Another frequent way to classify feature selection algorithms is according to the relationship they keep with the learning methods, normally distinguishing three categories: filter, wrapper and embedded methods [9,14,34]. Moreover, another possibility is to group the feature selection algorithms according to whether the algorithms ignore dependencies between features (univariate) or not (multivariate) [19,20, 34].

When the feature selection problem is posed in an environment in which the sample size is small and high dimensionality occurs, the concerns regarding reproducibility are greater.

The main reason is that those two circumstances favor the variability of the outcomes the feature selection algorithms produce, an occurrence which may likely have a great impact on the reproducibility of the results achieved regarding the specific objective pursued (e.g., classification, clustering, knowledge discovery, ...). In practice, when dealing with such a problem in which instances are divided into two or more groups of interest, and in which both high dimensionality and small sample size occur, a common procedure consists of applying a univariate filter feature selection method capable of generating rankings of features. That method, which normally is a statistical test or a heuristic measure, is mainly used to quantify how each feature behaves differently throughout the different groups of instances provided, thus enabling the construction of a ranking. Once the ranking is ready, a subset of features, which is usually composed of top-ranked features, is selected (e.g., setting a threshold or fixing the size of the subset). Indeed, this feature subset selection step is generally seen as a filtering process, in which the great majority of the features are filtered out, and in which the usefulness of the remaining ones is yet to be checked.

Let us recall that within the feature selection problem, the reproducibility crisis is most apparent in the biomarker selection problem [3,13]. On one hand, the relevant features (true biomarkers) are normally expected to be far fewer than the features (candidate biomarkers). On the other hand, both high dimensionality and small sample size are present to such an extent that the feature selection becomes not only a convenient step, but also an indispensable one [20,35]. Those two facts combined increase the difficulty of the task consisting of the detection of relevant features. In fact, in that context (biomarker selection), it is not unusual for relevant features identified in a study to later turn out to be invalid [3,17]. As aforementioned for this kind of problems (in which the amount of relevant features is far greater than the amount of irrelevant features and the amount of features is far greater than the amount of instances), the subset of selected features is analyzed and checked in further, more costly in terms of time, effort and money, studies so as to validate them [34].

2.2 Related work

The aforementioned concern regarding reproducibility in the feature selection problem has promoted research in the matter. Such research has been carried out mainly through the assessment of the reproducibility of the outcomes of the feature selection algorithms. Specifically, such assessment is generally conducted through the measurement of the stability of the outcomes of the feature selection algorithms, a stability that can be defined as the sensitivity of the feature selection outcomes to (small) variations in the datasets [8,19,20,29,30, 32,34].

The related work is divided into five subsections. Subsections 2.2.1, 2.2.2, 2.2.3 and 2.2.4 briefly describe the most relevant research lines regarding the work we pose in this paper. Subsection 2.2.5 describes Kuncheva's consistency index due to its central role in this work. For a more general view on the feature selection problems and on the stability of feature selection algorithms, we refer the reader to [6,7,24] and [3,17,22], respectively.

2.2.1 Stability measures

One research line within the area of reproducibility in feature selection consists of the study and development of stability measures for the outcomes of the feature selection algorithms. Kalousis et al. [19,20] analyzed the behavior of different stability measures using several feature selection algorithms in different biomedical problems. In the context of stability measures that deal with the outcomes of feature subset selection algorithms, Nogueira & Brown [29,30] studied the desirable properties that such stability measures should have. They also analyzed popular stability measures used in feature subset selection, such as the Jaccard-Tanimoto index [11], the adjusted stability measure of Lustgarten [26], Kuncheva's consistency index [23] and many others. That analysis was conducted by identifying which properties (e.g., correction for chance) would be desirable in a feature selection measure, and by later checking which properties were satisfied by each measure among those that were analyzed. Nogueira et al. [32] studied the properties of Spearman's rho as a stability measure for rankings of features, and provided insights on its properties. Chelvan & Perumal [8] conducted an experimental comparison of stability measures for feature selection algorithms, testing them on different datasets. Alelyani et al. [2] analyzed the sensibility of several feature selection stability measures to the variability of the incoming datasets. They also proposed a technique so as to assess the stability of feature selection algorithms while taking into account the variability of the incoming datasets.

2.2.2 Ensemble methods

Another attractive research line within the same field is the study and development of ensemble methods so as to increase the stability and in order to observe the stability under controlled circumstances. Guyon & Elisseeff [14] described a procedure to achieve stable feature selection based on the union of subsets of features selected in several bootstraps of a given dataset under analysis. Dunne et al. [12] proposed an ensemble solution in order to increase the stability of the feature subsets selected by wrapper-based approaches, which they tested in biomedical problems and object recognition problems. Saeys et al. [34] designed ensemble versions of different feature selection algorithms seeking to improve the stability of their single versions while maintaining a similar performance. Therefore, they conducted experiments using biomedical data and compared single and ensemble versions of different feature selection algorithms not only in terms of stability, but also in terms of performance. Abeel et al. [1] conducted an extensive analysis of ensemble feature selection within a framework they designed to analyze the stability of feature selection algorithms. Haury et al. [16] compared the single version and different ensemble versions of feature selection algorithms in terms of their influence on both performance and stability in breast cancer prognosis problems. They assessed the influence on performance in terms of the performances achieved by several supervised classification algorithms trained on the features selected by the feature selection algorithms. In contrast, to assess the influence on stability they measured the consistency between subsets of selected features derived from different samples.

2.2.3 Comparison of the stability of different feature selection algorithms

Another interesting research line within the same area consists of the observation and comparison of the stabilities achieved by different feature selection algorithms in a set of problems, thus allowing for the retrieval of information such as an assessment of which method best suits which problem. Kalousis et al. [19,20] also analyzed the behavior of different feature selection algorithms using several stability measures in different biomedical problems. Haury et al. [16] compared several filter and wrapper methods in terms of their stability (and performance), following the same procedure they used to compare the single version and different ensemble versions of feature selection algorithms as previously mentioned. Shanab et al. [36] compared the stabilities of many RFSS algorithms in several cancer datasets while using different sampling techniques. Dernoncourt et al. [9] studied in biomedical problems how and to which extent the stability of the results of feature selection algorithms was affected by different parameters of the datasets.

2.2.4 Reproducibility models

In addition to the aforementioned research lines, it is worth mentioning the work of Li et al. [25] because it is that which is most closely related to our work. In their work, they develop a model which is fitted to a reproducibility measure in order to extract, from that fitting, further information of interest. Specifically, Li et al. [25] assessed the reproducibility across two replicates through curves they defined and which they referred to as correspondence curves. They also explained how to fit a copula mixture model to those curves and derived from the fitted model a reproducibility score that they called Irreproducible Discovery Rate (IDR), which is supposed to be analogous to the False Discovery Rate (FDR).

In fact, it is appropriate to point out that our work shares some similarities with the work of Li et al. [25] in terms of the purpose. To start with, both works share the following two assumptions that support them. First, both papers "assume that each putative signal has been assigned a score that relates to the strength of the evidence for the signal to be real on the corresponding replicate by some data analysis method" [25]. Secondly, both papers made the "assumption that genuine signals are reproducible and noise is irreproducible" [25]. In summary, they assume that the method, in general, fulfills its duty. Moreover, both works use curves, which although different, aim to assess the reproducibility of the outcomes of the feature selection algorithms. In addition, in later steps both works propose models that are fitted to the corresponding curves. However, the models have little to do with each other. On one hand, as aforementioned, they proposed a copula mixture model to infer the reproducibility of the features while considering two types of features, in which the scores of one of the types are conceived as more reproducible than the scores of the other type. On the other hand, our model is based on the conception of each ranking of features as a complete sequence of extractions of balls from an urn with two types of balls, one type representing relevant features and the other one representing irrelevant features. In addition, the correspondence curves to which the model of Li et al. [25] is able to fit to are not corrected for chance (unlike the curves our model fits to). This may lead to biases of the estimation of the reproducibility due to the cardinality of the selected feature subsets. Furthermore, although the two models assess the reproducibility, they assess it through different measures of different magnitudes, a circumstance that allows them to complement each other in such an aspect, and under which a comparison between them can hardly be made. Moreover, Li et al. [25] concluded with the computation of the irreproducible discovery rate, while our work concludes with the gathering of information regarding the true Receiver Operating Characteristic (ROC) curve through its Area Under the Curve (AUC), in terms of the capability of a given feature selection algorithm for selecting relevant features. Specifically, such information is provided through an AUC derivable from our model.

Finally, it is convenient to recall that except for the work of Li et al. [25] presented in this subsection, the rest of the papers described within Subsection 2.2 do not propose any model for the feature selection process from which to derive additional information regarding reproducibility. Moreover, they do not focus on ranking based feature selection as much as on subset based feature selection. Besides, as mentioned, the model proposed by Li et al. [25] and our model are quite different and complimentary, our model having the advantage of being intuitive and easy to interpret. Consequently, all those facts present a research gap in which our research takes place.

2.2.5 *Kuncheva's consistency index*

Kuncheva's consistency index is a feature selection stability measure that enables the assessment of the consistency between pairs of subsets of features. Specifically, given a set X whose cardinality is $|X| = n$ and two subsets $A \in X$ and $B \in X$ whose cardinalities are $|A| = |B| = i$, where $0 < i < n$, Kuncheva's consistency index [23] between those two subsets is defined as follows:

$$r_i(A,B) = \frac{|A \cap B| - \frac{i^2}{n}}{i - \frac{i^2}{n}}. \tag{1}$$

Briefly, Kuncheva's consistency index measures the proportion of features present both in A and B, $|A \cap B|/i$, and it introduces a correction for chance (the $-i^2/n$ terms at both sides of the quotient). This correction for chance has an advantage with respect to other alternatives since it ensures that the expected value of a random feature selection algorithm is constant (0), independently of the sizes of A, B and X.

In fact, Kuncheva's consistency index is a very popular metric due to its advantageous properties that are critical for the interpretation and comparison of stability values [29–31]:

- Strict monotonicity: This property states that the stability measure is an increasing function of the average pairwise intersection size.
- Bounded quantity: This is a bounded measure in the interval [-1,1], and the bounds do not depend on the number of features. Its maximum value (1) is obtained when the sets are equal, $A = B$.
- Correction for chance: As explained previously, this is a measure corrected for chance.

Kuncheva's consistency index is defined for feature subsets of the same cardinality ($|A| = |B|$). In this work, we propose the use of reproducibility curves, the extension of Kuncheva's consistency index to rankings of features.

3 Empirical analysis of the reproducibility

Generally, when a RFSS algorithm ranks a set of n features, the algorithm tends to rank each feature according to its relevance. Namely, the more relevant a feature is, the closer to the top of the ranking it will tend to be placed by the algorithm. That is why our analysis of the reproducibility is based on the assessment of the similarity of two top feature subsets of size i, which are derived from the rankings obtained for two different replicates, for $i \in \{1,\ldots,n\}$, when a given RFSS algorithm is applied. For each possible subset size i, we have chosen to use Kuncheva's consistency index to assess the similarity between pairs of top feature subsets of the same size i. We selected this stability measure for two reasons. First, its previously mentioned advantages enable a straightforward interpretation and comparison of the stability values. Secondly, in our work we aim to assess the stability of subsets of features that have the same cardinality. Note that the resulting vector of n consistency indexes of Kuncheva, each being associated to a different size i of the subsets of top-ranked features, together with the sequence $1,\ldots,n$ can be represented graphically as a curve [23], a curve which we refer to as a "reproducibility curve". It is worth noting that one of the rankings represents the ranking derived from an original study, while the other represents the ranking derived from a reproduction attempt. Consequently, a measurement of the similarity of the two rankings, through the stability of the feature selection, indicates how similar the reproduction attempt and the original study are in terms of the results of the feature selection.

3.1 Formalizing the reproducibility curves

Let us assume that we have a set of features $\boldsymbol{X} = (X_1, \dots, X_n)$ and a class C which takes binary values $c \in \{+, -\}$, where $(\boldsymbol{X}, C)$ is distributed according to some unknown probability distribution p. Let us have a dataset $\boldsymbol{D}$ of N i.i.d. samples according to p. Formally, a RFSS method can be seen as a function f that, given a dataset $\boldsymbol{D}$, maps a set of n elements (features) into an element (permutation) of the symmetric group S_n that represents an ordering of the features:

$$f(\boldsymbol{D}) = \boldsymbol{\sigma}, \tag{2}$$

where $\sigma_i = j$ denotes that feature X_j is ranked as the i-th most relevant feature according to the permutation $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)$. For the sake of simplicity, we also refer to an ordering $\boldsymbol{\sigma}$ as ranking $\boldsymbol{\sigma}$, given that for any ordering $\boldsymbol{\sigma}$ the calculation of its associated ranking is straightforward. From here on we will denote by $\boldsymbol{\sigma}_{\leq i}$ the set formed by the first i elements of $\boldsymbol{\sigma}$, $\boldsymbol{\sigma}_{\leq i} = \{\sigma_1, \dots, \sigma_i\}$.

Next, we define in Equation 3 the function l_i that given two orderings $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ obtains the intersection of the associated two top-i sets, $\boldsymbol{\sigma}_{\leq i}$ and $\boldsymbol{\sigma}'_{\leq i}$.

$$l_i(\boldsymbol{\sigma}, \boldsymbol{\sigma}') = |\boldsymbol{\sigma}_{\leq i} \cap \boldsymbol{\sigma}'_{\leq i}|. \tag{3}$$

We denote this function simply as l_i when it is clear from the context. Let L_i be the random variable associated to l_i for a RFSS method f, where $\boldsymbol{\sigma} = f(\boldsymbol{D})$ and $\boldsymbol{\sigma}' = f(\boldsymbol{D}')$, and $\boldsymbol{D}$ and $\boldsymbol{D}'$ are i.i.d according to p.

We denote as r_i the function that measures Kuncheva's consistency index for the top-i sets of two orderings $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$:

$$r_i(\boldsymbol{\sigma}, \boldsymbol{\sigma}') = \frac{l_i - \frac{i^2}{n}}{i - \frac{i^2}{n}}. \tag{4}$$

Again, we denote this function simply as r_i when it is clear from the context. Let R_i be the random variable associated to r_i for a RFSS method f, where $\boldsymbol{\sigma} = f(\boldsymbol{D})$ and $\boldsymbol{\sigma}' = f(\boldsymbol{D}')$, and $\boldsymbol{D}$ and $\boldsymbol{D}'$ are i.i.d according to p. We call reproducibility curve to a particular realization $\boldsymbol{r} = (r_1, \dots, r_n)$ of $(R_1, \dots, R_n)$ for a given pair of orderings. Specifically, a reproducibility curve $\boldsymbol{r} = (r_1, \dots, r_n)$ is graphically represented as the sequence of points $(0,0), (1, r_1), (2, r_2), \dots, (n, r_n)$.

We denote as ρ_i the expected value of R_i, i.e., $\rho_i = \mathbb{E}_p[R_i]$. Besides, we call expected reproducibility curve (ERC), or simply reproducibility curve when it is clear from the context, to the sequence $\boldsymbol{\rho} = (\rho_1, ..., \rho_n)$.

3.2 Estimating the expected reproducibility curve from data

Unfortunately, in real situations the probability distribution p underlying the data is unknown, and thus we have to estimate the ERC $\boldsymbol{\rho}$, an estimation to which we refer as $\hat{\boldsymbol{\rho}}$, using the available data. To that end, we propose two different sampling strategies for estimating the ERC. Both procedures generate pairs of datasets, and for each pair of datasets an estimated reproducibility curve is calculated. Therefore, by generating a set of pairs of datasets a set of estimated reproducibility curves is calculated, and the ERC is estimated by averaging them.

The proposed methods for generating several pairs of datasets ($\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$) from the available data ($\boldsymbol{D}$) are: i) random splitting (in disjoint halves) of the available data, and ii) bootstrap sampling (i.e., generate datasets of the size of the available data by using uniform random sampling with replacement). In order to reduce the variance of the estimated ERCs, we recommend the use of stratified splitting and sampling (i.e., maintain the class proportion of the original dataset in the generated pairs of datasets). Figure 1 illustrates how to estimate the reproducibility curve using the (stratified) random splitting strategy. Then, by repeating this process and averaging the obtained reproducibility curves, the expectation is estimated. The complete procedure is presented in Algorithm 1: for each iteration of the algorithm (line 2) a reproducibility curve is computed. First, a pair of datasets is obtained by using a random sampling strategy, e.g., random splitting or bootstrapping (line 3). Then, by applying the RFSS algorithm f on the pair of datasets, two rankings of features are obtained (line 4). The two rankings are used to compute the reproducibility curve of the k-th iteration by means of Equations 3 and 4 (line 5). The estimated ERC is computed as the average of the t obtained reproducibility curves (line 7).

The proposed sampling strategies have been selected because, on average, by using the random splitting and bootstrap procedures, we will obtain lower and upper bounds to the ERC. Intuitively, on the one hand, the random splitting generates pairs of datasets of smaller size ($N/2$ instances) that do not share any instance which, on average, leads to pessimistic estimates of the ERC. On the other hand, bootstrap sampling obtains pairs of overlapping datasets of the same size (N) which, on average, leads to optimistic estimates (see [33] for further details on pessimistic and optimistic estimates).

Algorithm 1 The pseudo-code of the algorithm used for estimating the expected reproducibility curve $\hat{\boldsymbol{\rho}}$. Depending on the random sampling procedure used to generate the pair of datasets, e.g., random splitting or bootstrapping, different estimates can be implemented (line 3).

```
1: procedure ESTIMATING $\boldsymbol{\rho}$
   Input: Dataset $\boldsymbol{D}$, RFSS algorithm $f$, number of repetitions $t$.
   Output: Estimated expected reproducibility curve $\hat{\boldsymbol{\rho}}$.
2:     for $k = 1$ to $t$ do
3:         Generate $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$ from $\boldsymbol{D}$ by a random sampling strategy.
4:         Apply $f$ to $\boldsymbol{D}^{(1)}$ and $\boldsymbol{D}^{(2)}$ to get the rankings of features $\boldsymbol{\sigma}^{(1)}$ and $\boldsymbol{\sigma}^{(2)}$.
5:         Using Equations 3 and 4 with $\boldsymbol{\sigma}^{(1)}$ and $\boldsymbol{\sigma}^{(2)}$ compute the reproducibility curve $\boldsymbol{r}^k$.
6:     end for
7:     return the average of the reproducibility curves $\boldsymbol{r}^k$ for $k = 1,...,t$.
8: end procedure
```

Finally, seeking to further illustrate the two explained sampling procedures and their outcomes, the following subsection exposes its application to a real dataset as an example.

3.3 An illustrative example

In this example, we have computed $\hat{\boldsymbol{\rho}}$ for two classical statistical tests, the t-test and the Wilcoxon rank sum test in a real-life biomedical dataset. Please note that how adequate

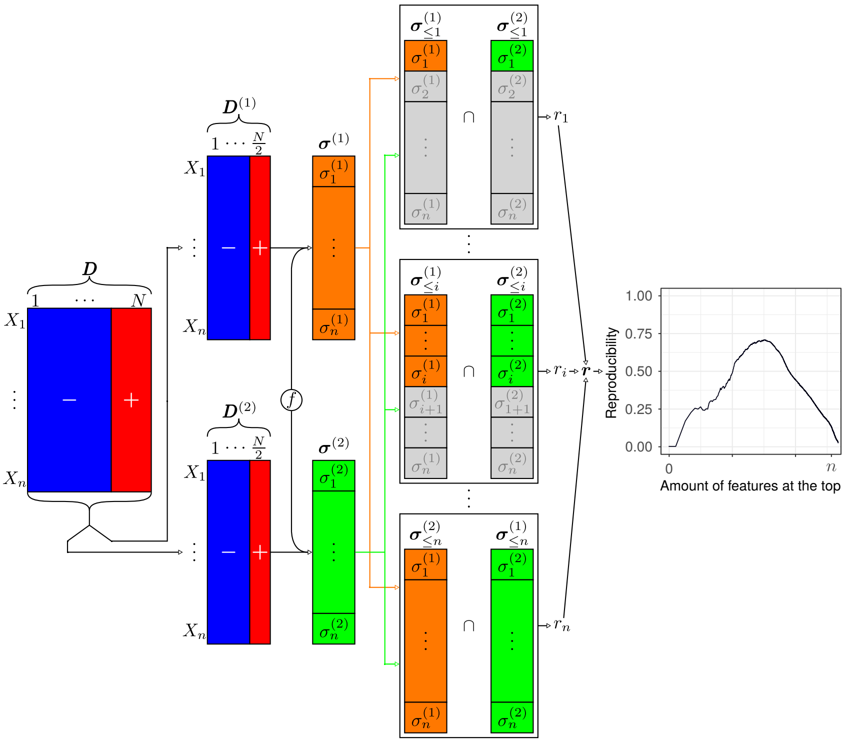
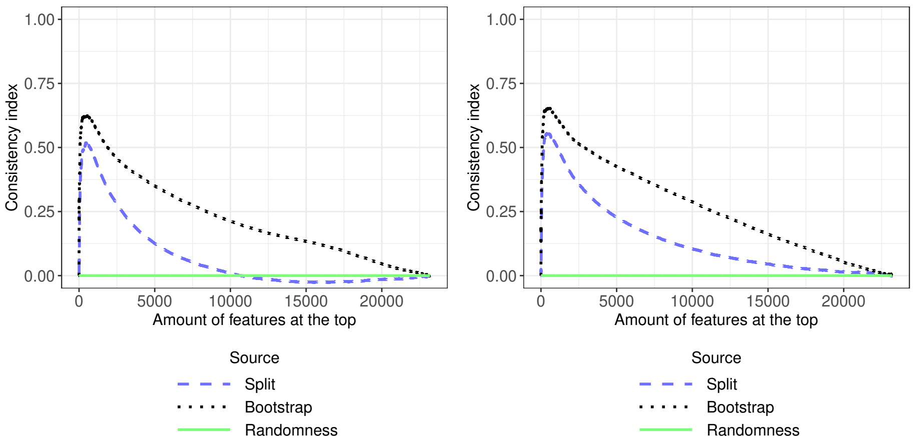

$\boldsymbol{D}$
1 ⋯ N
X_1
X_n
−
+
$\boldsymbol{D}^{(1)}$
$\boldsymbol{D}^{(2)}$
$1 \cdots \frac{N}{2}$
$\boldsymbol{\sigma}^{(1)}$
$\boldsymbol{\sigma}^{(2)}$
f
$\boldsymbol{\sigma}^{(1)}_{\leq 1}$
$\boldsymbol{\sigma}^{(2)}_{\leq 1}$
$\boldsymbol{\sigma}^{(1)}_{\leq i}$
$\boldsymbol{\sigma}^{(2)}_{\leq i}$
$\boldsymbol{\sigma}^{(2)}_{\leq n}$
$\boldsymbol{\sigma}^{(1)}_{\leq n}$
$\cap$
r_1
r_i
$\boldsymbol{r}$
r_n
Reproducibility
1.00
0.75
0.50
0.25
0.00
0
n
Amount of features at the top

Fig. 1 This figure illustrates the computation of a reproducibility curve using the available data by using the (stratified) random splitting strategy, where there are N^+ and N^- samples from the positive and negative classes, respectively

the selected methods are to rank the candidates is irrelevant for our purpose of showing how our statistical approach to the reproducibility problem works. The reason why these methods have been selected is because they are classical approaches to the feature subset selection problem. The dataset consists of ovarian cancer cases and controls [38] which have their Deoxyribonucleic Acid (DNA) methylation values measured over 27000 candidate biomarkers. This dataset is available at the Gene Expression Omnibus (GEO) database (http://www.ncbi.nlm.nih.gov/geo, where the ovarian cancer dataset has accession number GSE19711). In order to enhance the quality of the data provided by that database, we carried out a preprocessing of the data (see supplementary material) based on that done by Wang et al [40].

Figure 2 shows the different $\hat{\boldsymbol{\rho}}$ corresponding to the t-test and the Wilcoxon test applied to the real dataset mentioned (with $t = 10$). Specifically, in Subfigures 2a and 2b, both estimated ERCs start by rising steeply until they flatten out and then each reaches a peak. Then they start decreasing and getting closer to the straight line of a uniform random selection, and finally they meet the same consistency index value (0) when the top-n reproducibility is computed. 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 rest of candidate biomarkers are frequently interchanged by the tests.



(a) **(b)**

Fig. 2 Reproducibility plots for the ovarian cancer database when the t-test is applied (2a) and when the Wilcoxon test is applied (2b)

In addition to what has been aforementioned regarding how interesting each position i of a reproducibility curve is, it should be noted that the top-i at which the mentioned peak is reached is interesting. It is interesting because, under the assumption that there are two types of features (biomarkers and non-biomarkers), such a peak serves as a heuristic that provides hints regarding the size of each of those two subsets of features.

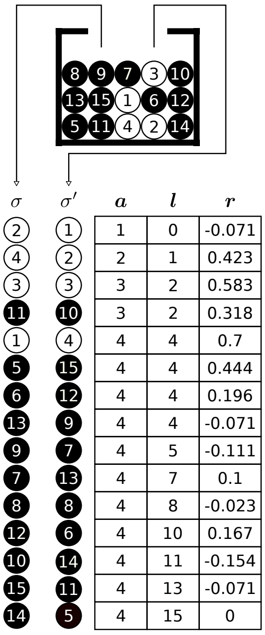
4 Modeling the reproducibility curves

In this section, we present a simple and intuitive statistical model for the ERCs. This model will allow us to analyze the reproducibility of a RFSS algorithm when it deals with real data (see Section 5 for further details).

The proposed model is based on an urn with n balls representing the n features. A complete sequential extraction of the balls in the urn represents a ranking of the features and it is denoted by a permutation, $\boldsymbol{\sigma}$.

In the feature selection problem, the goal is to select the most relevant features following a given criterion. In this sense, the proposed model assumes that features are divided into relevant features (the most relevant ones) and irrelevant features (the less relevant ones). This simplification reduces the problem of feature subset selection to the problem of detecting the relevant features while discarding the irrelevant ones. That assumption leads to an interpretable model from a feature subset selection problem point of view.

As a way of simplifying the model, we will assume that, in any extraction, the amount of relevant balls in any top-i random ranking $\boldsymbol{\sigma}_{\leq i}$ 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 relevant balls extracted is denoted as $\boldsymbol{a} = (a_1,\ldots,a_n)$. Taking by convention that $a_0 = 0$, due to the nature of the process, a_i must be equal to or greater by one than a_{i-1} for $i \in \{1,\ldots,n\}$. Figure 3 shows a scheme of the process with two extractions, representing the relevant balls as white balls and representing the irrelevant balls as black balls. In order to clarify the content of this figure,



σ	σ'	$\boldsymbol{a}$	$\boldsymbol{l}$	$\boldsymbol{r}$
2	1	1	0	-0.071
4	2	2	1	0.423
3	3	3	2	0.583
11	10	3	2	0.318
1	4	4	4	0.7
5	15	4	4	0.444
6	12	4	4	0.196
13	9	4	4	-0.071
9	7	4	5	-0.111
7	13	4	7	0.1
8	8	4	8	-0.023
12	6	4	10	0.167
10	14	4	11	-0.154
15	11	4	13	-0.071
14	5	4	15	0

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

remember that $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ are two rankings obtained from two different datasets as explained in Section 3, that $\boldsymbol{a}$ is the sequence of amounts of relevant balls extracted, that $\boldsymbol{l} = l_1, \ldots, l_n$ is the coincidences vector (see Equation 3), and that $\boldsymbol{r} = r_1, \ldots, r_n$ is the reproducibility vector (see Equation 4).

Under the proposed model, for the sake of simplicity, it is assumed that the probability of extracting any specific relevant ball is the same for each of the remaining relevant balls in the urn; an analogous assumption is made regarding the extraction of irrelevant balls. Namely, for each type of feature, relevant and irrelevant, it is assumed that the goodness of the features of the same type in terms of the score being used (derived by using f) is the same on average, i.e., that they have the same relevance according to f. When $\boldsymbol{a}$ is known, that assumption makes it easy to derive theoretically the ERC $\boldsymbol{\rho} = \{\rho_1, \ldots, \rho_n\}$ from $\boldsymbol{a}$. Specifically, bearing in mind Equation 4, ρ_i can be expressed for $i \in \{1, \ldots, n\}$ as:

$$\rho_i = \frac{\lambda_i - \frac{i^2}{n}}{i - \frac{i^2}{n}}, \tag{5}$$

where λ_i is the expected amount of coincidences between any two top-i ranks. In order to derive λ_i from $\boldsymbol{a}$, first we decompose it as the sum of the expected amount of coincident relevant balls, which we denote as λ_i^a, and the expected amount of coincident irrelevant balls, which we denote as λ_i^b. Namely:

$$\rho_i = \frac{\lambda_i^a + \lambda_i^b - \frac{i^2}{n}}{i - \frac{i^2}{n}}. \tag{6}$$

In fact, the random variable L_i^a, whose expectation is λ_i^a, 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 total amount of relevant balls), the amount of successes in the population (the relevant balls extracted in the first sequence until the first i extractions are made) and the amount of draws (the relevant balls extracted in the second sequence until the first i extractions are made), respectively. Consequently, the expected value is:

$$\lambda_i^a = \frac{a_i^2}{a_n}. \tag{7}$$

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:

$$\lambda_i^b = \frac{(i-a_i)^2}{n-a_n}. \tag{8}$$

Finally, replacing in Equation 6 the terms λ_i^a and λ_i^b with their expressions of Equations 7 and 8 respectively, the expected top-i reproducibility ρ_i under the model represented by $\boldsymbol{a}$ can be calculated as:

$$\rho_i = \frac{\frac{a_i^2}{a_n} + \frac{(i-a_i)^2}{n-a_n} - \frac{i^2}{n}}{i - \frac{i^2}{n}}. \tag{9}$$

Note that the expected top-i reproducibility under the proposed model for $i \in \{1,\ldots,n\}$ is symmetric regarding the relative amount of relevant and irrelevant balls (i.e., switching the labels has no effect in the model). However, in many practical scenarios, such as the biomarker selection, the relevant features (relevant balls, a_n) are far less than the irrelevant ones (irrelevant balls, $n-a_n$).

5 Fitting the model to empirical data

This section is divided into three subsections. In the first one, given an estimated ERC from a given RFSS algorithm, a procedure based on dynamic programming to find the sequence $\boldsymbol{a}$ that best fits a given estimated ERC $\hat{\boldsymbol{\rho}}$ is described. In the second one, a procedure is presented to estimate quantitatively how often the relevant balls tend to be ranked closer to the top than the irrelevant balls. Finally, using the model, we show how to estimate the quality of the orderings produced by the RFSS algorithm.

5.1 Fitting $\boldsymbol{a}$ to empirical data

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 in order to derive information of interest. For instance, the estimation of the amount of relevant balls a_n can be interpreted as the amount of features that present differences detectable by the given method when dealing with the given dataset.

Before explaining the fitting process, it is convenient to recall the set of constraints that any given sequence $\boldsymbol{a}$ must satisfy so as to be feasible according to our model. 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 can assess the difference between a given estimated reproducibility curve $\hat{\boldsymbol{\rho}}$ and the ERC given a particular $\boldsymbol{a}$ (see Equation 9):

$$E(\hat{\boldsymbol{\rho}}, \boldsymbol{a}) = \sum_{i=1}^{n} e_i(\hat{\rho}_i, a_i, a_n), \tag{10}$$

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 9). Consequently, we have:

$$e_i(\hat{\rho}_i, a_i, a_n) = \left(\hat{\rho}_i - \frac{\frac{a_i^2}{a_n} + \frac{(i-a_i)^2}{n-a_n} - \frac{i^2}{n}}{i - \frac{i^2}{n}}\right)^2. \tag{11}$$

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, which we denote as $\boldsymbol{a}^*$:

$$\boldsymbol{a}^* = \arg\min_{\boldsymbol{a} \in A} E(\hat{\boldsymbol{\rho}}, \boldsymbol{a}). \tag{12}$$

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. Secondly, each subproblem is solved by using dynamic programming through the following recursive function:

$$E_{a_i}^i(\hat{\boldsymbol{\rho}}) = e_i(\hat{\rho}_i, a_i, a_n) + \min(E_{a_i}^{i-1}(\hat{\boldsymbol{\rho}}), E_{a_i-1}^{i-1}(\hat{\boldsymbol{\rho}})), \tag{13}$$

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. Hence, the sequence $\boldsymbol{a}^*$ that minimizes the cumulative error can be found by searching for the $\boldsymbol{a}$ sequence whose associated cumulative error is the minimum among the $n+1$ computed ones. Note that while each subproblem is being resolved, 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 13.

In order to complement the explanation given for the fitting process that enables the calculation of $\boldsymbol{a}^*$ for a given $\hat{\boldsymbol{\rho}}$, i.e., the problem posed in Equation 12, we present the pseudo-code of that process in Algorithm 2.

Next, we provide a brief description of Algorithm 2 (a detailed explanation can be found in the supplementary material):

- The variables: n stores the amount of balls, $\boldsymbol{S}$ stores in its columns the best solutions for the different subproblems, $\boldsymbol{e}$ stores the errors of the best solutions for the different subproblems, $\boldsymbol{E}_m$ stores the cumulative errors described in Equation 13 for the m-th subproblem, $\boldsymbol{P}_m$ store the paths that enable the retrieval of the best solution $\boldsymbol{a}$ for the m-th subproblem, and $\boldsymbol{a}^*$ stores the best solution for the whole problem.
- The loops: Algorithm 2 solves the problem using three nested loops: The outer one iterates through different subproblems, the middle one through different positions of the sequences of the amounts of relevant balls, and the inner one through different amounts of relevant balls. The outer loop only needs to cover half of the subproblems due to the aforementioned symmetry regarding the relative amount of relevant and irrelevant balls.
- The steps:
  - n, $\boldsymbol{S}$ and $\boldsymbol{e}$ are initialized.
  - Inside the outer loop, in each iteration:

Algorithm 2 The pseudo-code of the algorithm used for computing the sequence of the amounts of relevant balls $\boldsymbol{a}^*$ of minimum error.

```
procedure COMPUTING $\boldsymbol{a}^*$
Input: Estimated expected reproducibility curve $\hat{\boldsymbol{\rho}}$.
Output: Sequence of the amounts of relevant balls $\boldsymbol{a}^*$ that minimizes the cumulative error function.
    $n = \text{length}(\hat{\boldsymbol{\rho}})$
    $\boldsymbol{S} = \text{Zeros}(n, \lfloor n/2 \rfloor + 1)$
    $\boldsymbol{e} = \text{Zeros}(\lfloor n/2 \rfloor + 1)$
    for $m = 0$ to $\lfloor n/2 \rfloor$ do
        $\boldsymbol{E}_m = \text{Infinites}(n+1, m+1)$
        $\boldsymbol{E}_m[0,0] = 0$
        $\boldsymbol{P}_m = \text{Zeros}(n+1, m+1)$
        for $i = 1$ to $n$ do
            for $j = 0$ to $min(i,m)$ do
                if $j = 0$ then
                    $\boldsymbol{E}_m[i,j] = \boldsymbol{E}_m[i-1,j] + e_i(\hat{\rho}_i, j, m)$
                    $\boldsymbol{P}_m[i,j] = j$
                end if
                if $j = i$ then
                    $\boldsymbol{E}_m[i,j] = \boldsymbol{E}_m[i-1,j-1] + e_i(\hat{\rho}_i, j, m)$
                    $\boldsymbol{P}_m[i,j] = j-1$
                end if
                if $j \neq 0$ and $j \neq i$ then
                    if $\boldsymbol{E}_m[i-1,j] < \boldsymbol{E}_m[i-1,j-1]$ then
                        $\boldsymbol{E}_m[i,j] = \boldsymbol{E}_m[i-1,j] + e_i(\hat{\rho}_i, j, m)$
                        $\boldsymbol{P}_m[i,j] = j$
                    else
                        $\boldsymbol{E}_m[i,j] = \boldsymbol{E}_m[i-1,j-1] + e_i(\hat{\rho}_i, j, m)$
                        $\boldsymbol{P}_m[i,j] = j-1$
                    end if
                end if
            end for
        end for
        $\boldsymbol{e}[m] = \boldsymbol{E}_m[m,n]$
        $\boldsymbol{S}[.,m] = \text{get\_subproblem\_best\_solution}(\boldsymbol{P}_m)$
    end for
    $\boldsymbol{a}^* = \text{get\_problem\_best\_solution}(\boldsymbol{S}, \boldsymbol{e})$
    return $\boldsymbol{a}^*$
end procedure
```

    - $\boldsymbol{E}_m$ and $\boldsymbol{P}_m$ are initialized.
    - The middle and inner loops are executed to solve the m-th subproblem, filling $\boldsymbol{E}_m$ and $\boldsymbol{P}_m$ accordingly (considering Equation 13).
    - The best error achieved in the m-th subproblem is stored in $\boldsymbol{e}$.
    - From the filled matrix of paths $\boldsymbol{P}_m$ the sequence $\boldsymbol{a}$ that is the best solution for subproblem m is derived and stored in the m-th column of $\boldsymbol{S}$.
- Given $\boldsymbol{e}$, the best solution for the whole problem, $\boldsymbol{a}^*$, can be found within the solutions stored in $\boldsymbol{S}$.

Regarding the computational complexity, in order to find $\boldsymbol{a}^*$, $\lceil n/2 \rceil$ dynamic programming problems are solved, one for each possible value of a_n (considering the aforementioned symmetry). 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 $\boldsymbol{a}^*$ has a computational complexity of $\mathcal{O}(n^3)$.

5.2 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 $\boldsymbol{a}^*$ using the process underlying the non-central hypergeometric distribution of Wallenius [39]. Therefore, the $\boldsymbol{a}^*$ computed in the previous subsection becomes an input for the process carried out in this subsection, which uses $\boldsymbol{a}^*$ to derive further data of interest.

In this process we have an urn with relevant and irrelevant balls, but each type has an associated weight that biases the extraction. The balls are extracted sequentially and, at each step, the probability of extracting a relevant ball will be the total weight of the remaining relevant balls divided by the total weight of all 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 each irrelevant ball is 1 and the weight of each relevant ball (or simply referred to as the weight) is w.

Therefore, in this second stage we see a given $\boldsymbol{a}$ as a summary of the outcome of a complete sequence of draws that follows the process described above. Consequently, the likelihood of $\boldsymbol{a}$ given w can be seen as the product of the probabilities of obtaining a relevant or an irrelevant ball at each step of the sequence of extractions given w. That is, the likelihood of $\boldsymbol{a}$ given w can be expressed as:

$$\begin{aligned}\mathscr{L}(\boldsymbol{a}|w) = &\prod_{i=1}^{n} \frac{(a_i - a_{i-1}) \cdot w \cdot (a_n - a_{i-1})}{w \cdot (a_n - a_{i-1}) + n - (i - 1 - a_{i-1})} + \\ &\frac{(1 - (a_i - a_{i-1})) \cdot (n - (i - 1 - a_{i-1}))}{w \cdot (a_n - a_{i-1}) + n - (i - 1 - a_{i-1})},\end{aligned} \tag{14}$$

where $a_i - a_{i-1}$ determines, whether in extraction i, a relevant 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 search in the interval $(0, \infty)$ of proper values of w for the w that maximizes the likelihood of $\boldsymbol{a}$, which we denote as w^*. This piece of information is very important due to its interpretation: For a fixed a_n, the more reproducible the outcomes of the method, the higher the value of its w. The weight also summarizes in a single scalar value the degree of mixing of the relevant and irrelevant balls in the sequence of extractions, with w becoming further away from 1 as the mixing decreases.

Unfortunately, there is no analytical solution to Equation 14, but an approximate value of w^* can be calculated through a search based on numerical analysis, such as, for instance, Brent's method [5]. We choose to use Brent's method because it is an efficient method that behaves well when dealing with the problem under analysis. Besides, an approximation of w^* can be very quickly obtained compared with the search for $\boldsymbol{a}^*$ of the previous stage of the fitting process.

5.3 Deriving information from $\boldsymbol{a}^*$

Now we explore a simple yet potentially useful idea which aims to obtain information about the performance of the method under analysis through the outcomes of the fitted model.

To start with, it is convenient to remember that feature subset selection algorithms can be seen as classifiers that assess as interesting the features they select and the rest of them as non-interesting. Specifically, the rankings produced by RFSS methods can be seen as orderings of the features according to the feature selection methods. At least in dichotomous problems, that concept enables us to assess the performance of feature selection algorithms in terms of the AUC regarding the capability of a given feature selection algorithm for selecting relevant features, as long as we know which are relevant and which are not. For instance, let us have a given ranking σ of a set of dichotomous features in terms of their relevance (relevant/irrelevant) and let us know which are relevant and which are not. Note that ranking σ can be cut into two parts at every interstice of it. For a given cut, the topmost features can be predicted to be relevant while the bottommost features can be predicted to be irrelevant. Then, contrasting such predictions with the true conditions of the features of ranking σ enables the calculation of a true positive rate and of a false positive rate for each cut. Therefore, we can use this procedure to derive a ROC curve and to compute its AUC (data AUC).

In order to seek information about the data AUC of a given method, we use the sequence $\boldsymbol{a}^*$ estimated in our model to derive an AUC from the model (model AUC). The process to derive a model AUC from $\boldsymbol{a}^*$ is similar to the aforementioned process to derive a data AUC from σ. Specifically, this time, the "true condition" (according to the model) of the top-i feature is considered to be relevant if $a_i^* - a_{i-1}^* = 1$ and irrelevant otherwise, for $i \in \{1, \ldots, n\}$ and assuming that by convention $a_0^* = 0$.

Unfortunately, in real life situations, we do not know which features are relevant and which are not and, hence, there is no way to obtain the data AUC. However, through the analysis of the reproducibility based on our proposal, we can obtain the model AUC. Evidently, we cannot trivially assume that it is an estimation of the true AUC (data AUC), but through experimentation with synthetic data, we have observed that they are correlated.

It should be noted that the reproducibility curves (to which the model is fitted) are derived from the outcomes of the feature ranking algorithm (after applying Kuncheva's consistency index). Consequently, the reliability of a given model AUC is as good as the capability of the feature ranking algorithm to fulfill its duty (to rank the relevant features before the irrelevant ones). However, let us recall that the two previously mentioned assumptions that this work shares with the work of Li et al [25] limit the impact of such an issue: First, it is assumed that "each putative signal has been assigned a score that relates to the strength of the evidence for the signal to be real on the corresponding replicate by some data analysis method". Secondly, it is assumed that "genuine signals are reproducible and noise is irreproducible".

6 Experimentation

This section is divided into three subsections. In the first two subsections, in order to illustrate the model and its use, the proposed model is fitted to the estimated ERC $\hat{\boldsymbol{\rho}}$, using both synthetic and real world datasets. Finally, in the last subsection a discussion of the results obtained from the experimentation is carried out.

6.1 Experimentation with synthetic data

Fitting the model to synthetic data enables the appropriateness of the model to be checked in controlled scenarios. Besides, the relationship between the AUC of $\boldsymbol{a}^*$ (model AUC) and the AUCs of $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ (data AUCs) can also be assessed through the experimentation with synthetic data.

The synthetic data used in the experimentation belong to a supervised classification problem with a binary class variable, $C \in \{+,-\}$. Each problem consists of 1000 features. Specifically, we have two types of features: the irrelevant ones whose distribution is independent from the class variable, and the relevant ones whose distribution is not independent from the class variable. In order to approximate real world scenarios, the number of relevant and irrelevant features is unbalanced: 50 relevant features and 950 irrelevant features. For the sake of simplicity, we have assumed that features are conditionally independent given the class and that the features conditioned to each class value are distributed according to a normal density function. In order to estimate the relevance curve from data, we have generated pairs of datasets, $\boldsymbol{D}$ and $\boldsymbol{D}'$ i.i.d. according to p. Specifically, 100 samples are generated for each of the 2 groups and for each of the 1000 features per dataset.

In particular, we have designed two different scenarios for synthetic data. In one of the scenarios, the relevant features show differences in location among groups, while in the other one the relevant features show differences both in location and spread. Finally, each scenario is composed of 21 different configurations. In each of those, the relevant features show increasing degrees of differences among groups. Consequently, different configurations within the same scenario pose problems with different difficulties. In particular, the 21 different difficulties aim to cover a wide range of situations. Namely, they range from situations in which the relevant features can be distinguished with ease from irrelevant ones to situations in which the relevant features are virtually indistinguishable from irrelevant ones.

In each configuration, the 950 non-relevant features are drawn from the normal distribution tagged in Figure 4 as “Difficulty 21” for both groups. The remaining 50 relevant features are drawn from the normal distribution tagged as “Difficulty 21” for one group, and from the normal distribution whose tag matches the difficulty of the given configuration for the other group (see further details in the supplementary material, including the specific parameter values of all these distributions). For each configuration of synthetic data, the estimated ERC $\hat{\boldsymbol{\rho}}$ has been calculated as explained in Section 3.1 (with $t = 32$).

Regarding the AUC, for each configuration, on one hand, the model AUC of $\boldsymbol{a}^*$ is computed, while, on the other hand, the average of the data AUCs of $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ is calculated. It is convenient to recall that $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ are permutations in which the indexes of the features appear in the order in which they are ranked. Besides, given those permutations and given that it is known which features are relevant and which are not, a data AUC can be derived for each permutation. Consequently, the average of the data AUCs derived from $\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$ is an average of 64 data AUC values in each configuration, since $t = 32$ and since in each run two data AUCs can be obtained from the two rankings generated within that run. After all these calculations, for each configuration a pair of values is obtained, the AUC derived from the model ($\boldsymbol{a}^*$) and the AUC derived from the data ($\boldsymbol{\sigma}$ and $\boldsymbol{\sigma}'$), thus obtaining 21 pairs of AUC values for each scenario and each method. Finally, in each combination of scenario and method, the Kendall rank correlation coefficient between the 21 AUC values derived from the model and the 21 AUC values derived from the data is calculated as an assessment of the relationship between them. We chose to use the Kendall rank correlation coefficient because the kind of correlation both measures might keep is unknown and because it is appropriate to deal with correlated quantities that might present tied cases, through the $\tau - b$ statistic [21].

As regards the ranking methods used during the experimentation with synthetic data, two different ranking methods are used. The first one is based on the t-test while the second one is based on the Wilcoxon rank sum test.

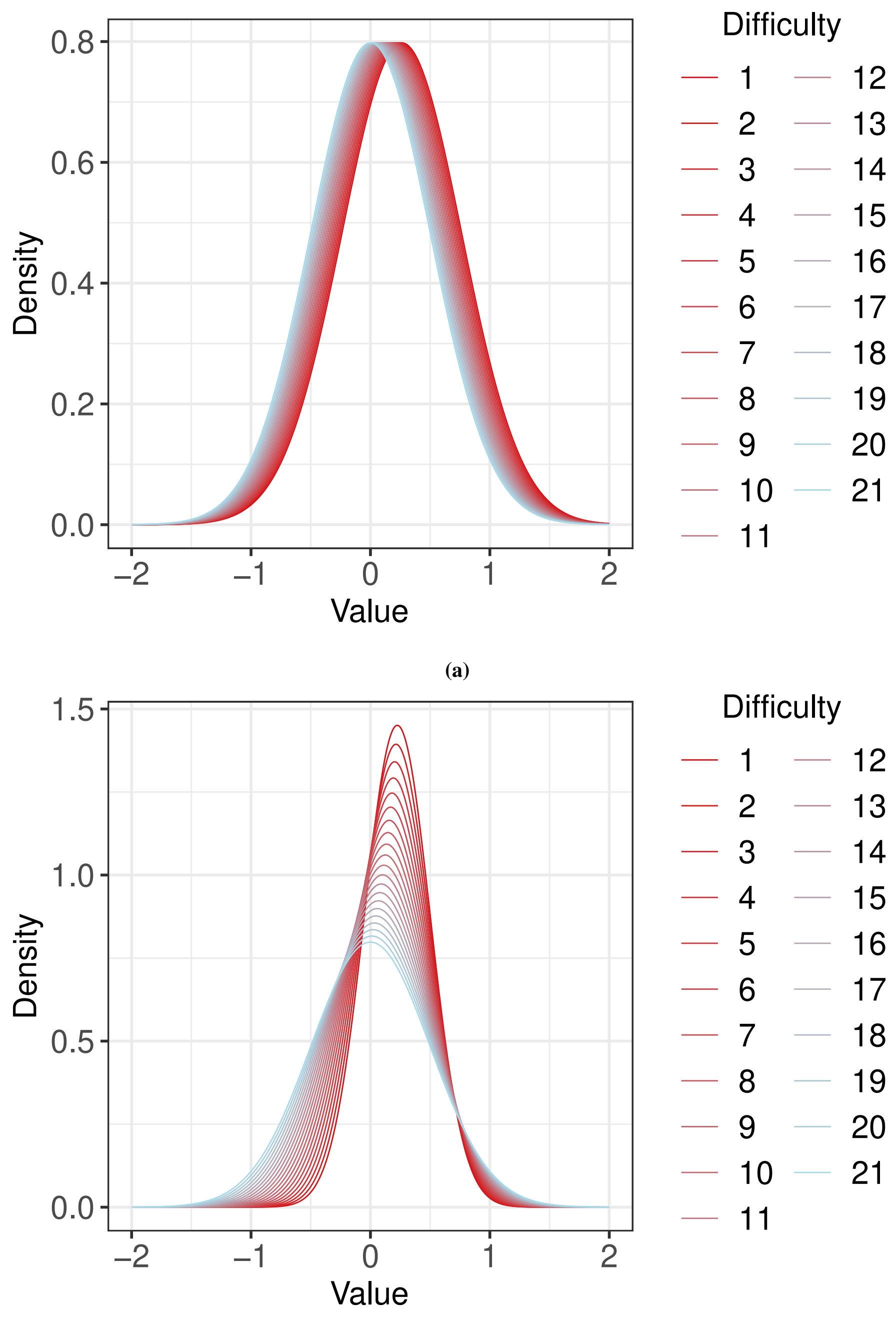
In order to display the results of the experimentation with synthetic data, different types of plots have been generated:

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

Problem	Difficulty	w^*	
		T-test	Wilcoxon test
Location	1	11.518	5.849
Location	8	5.345	3.016
Location	15	1.036	0.980
Location & spread	1	122.460	12.492
Location & spread	8	4.110	4.742
Location & spread	15	1.024	1.211

- Error plots (Subfigures 5a, 5b, 5c, 6a, 6b and 6c): They display in their abscissa axis the total amount of relevant balls (a_n) that correspond to different dynamic programming problems. The ordinate axis shows the cumulative errors of the optimum solution for each of these problems. The vertical dashed and dotted 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.
- Reproducibility plots (Subfigures 5d, 5e, 5f, 6d, 6e and 6f): They have already been presented in Section 3.2 in Figure 2. This time, both the reproducibility curves derived from the data and the reproducibility curves derived from the model (from $\boldsymbol{a}^*$) are shown. Regarding the reproducibility curves derived from the model, it is important to clarify that, the more mixed the two types of features (relevant/irrelevant) in $\boldsymbol{a}^*$ are, the flatter and more similar to a random reproducibility curve the reproducibility curve associated to $\boldsymbol{a}^*$ will be. Besides, the more separated the two types of features in $\boldsymbol{a}^*$ are, the more peaky the reproducibility curve associated to $\boldsymbol{a}^*$ will be. Specifically, it will likely have a single peak that will tend to reach 1 at the top-i equal to the amount of relevant balls in $\boldsymbol{a}^*$.
- Weight plots (Subfigures 5g, 5h, 5i, 6g, 6h and 6i): 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 the log-likelihoods of $\boldsymbol{a}^*$ are maximum, namely, w^*. In addition, the values of these w^* are displayed in the legends. Let us clarify that the weight w^* reflects and assesses quantitatively the degree of mixing of the two different types of features (relevant/irrelevant) in $\boldsymbol{a}^*$. Namely, the more separated the two types of features are in $\boldsymbol{a}^*$, the more distant from 1 the value of w^* will tend to be. Conversely, the more mixed the two types of features are in $\boldsymbol{a}^*$, the closer to 1 the value of w^* will tend to be.
- Correlation plots (Subfigures 7a 7c, 7b and 7d): They display for a combination of method and synthetic scenario the 21 pairs of AUC values (derived from the model and from the data).

The results of the experimentation with synthetic data are shown in Figures 5, 6 and 7, and in Tables 1 and 2. Notice that, due to space limitations, only a representative subset of the results of the experimentation with synthetic data has been displayed (difficulties 1, 8 and 15), while the remaining plots are available in the supplementary material along with further details of the experimentation conducted.



(a)

(b)

Fig. 4 Distributions used in the scenario of differences in location (4a) and in the scenario of differences in both location and spread (4b)

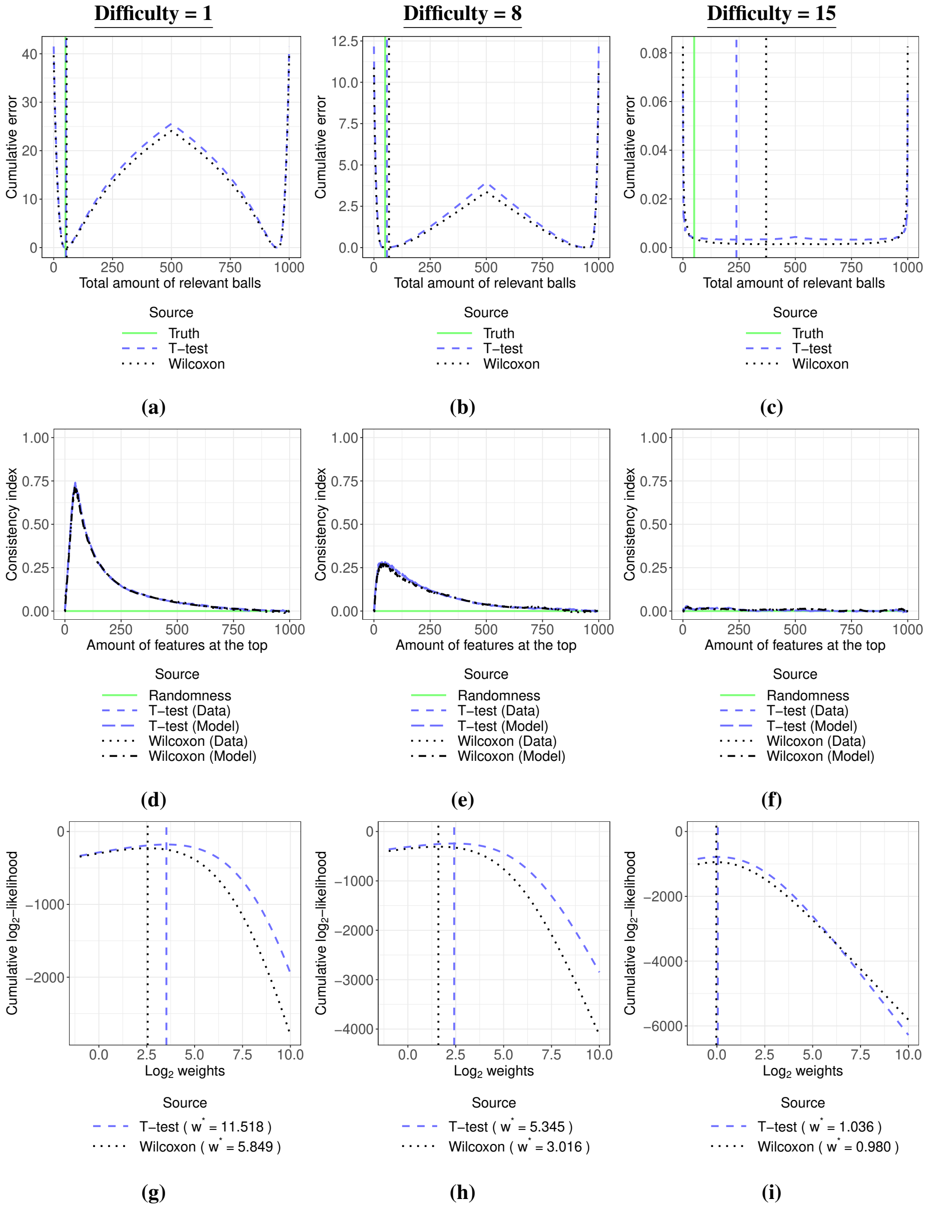
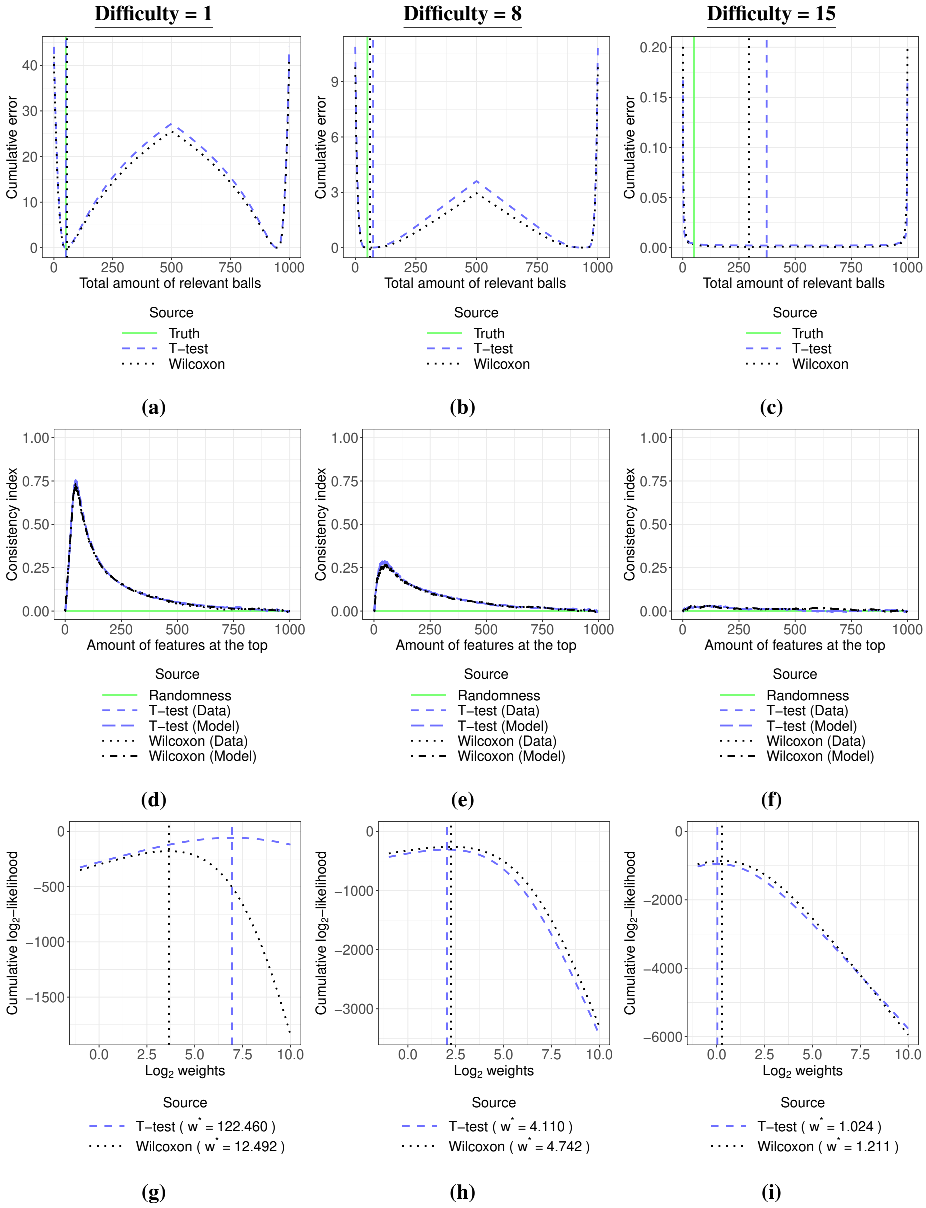


Fig. 5 Error plots (5a, 5b, 5c), reproducibility plots (5d, 5e, 5f) and weight plots (5g, 5h, 5i) for different difficulty configurations, 1 (5a, 5d, 5g), 8 (5b, 5e, 5h) and 15 (5c, 5f, 5i) respectively, in the differences in location scenario

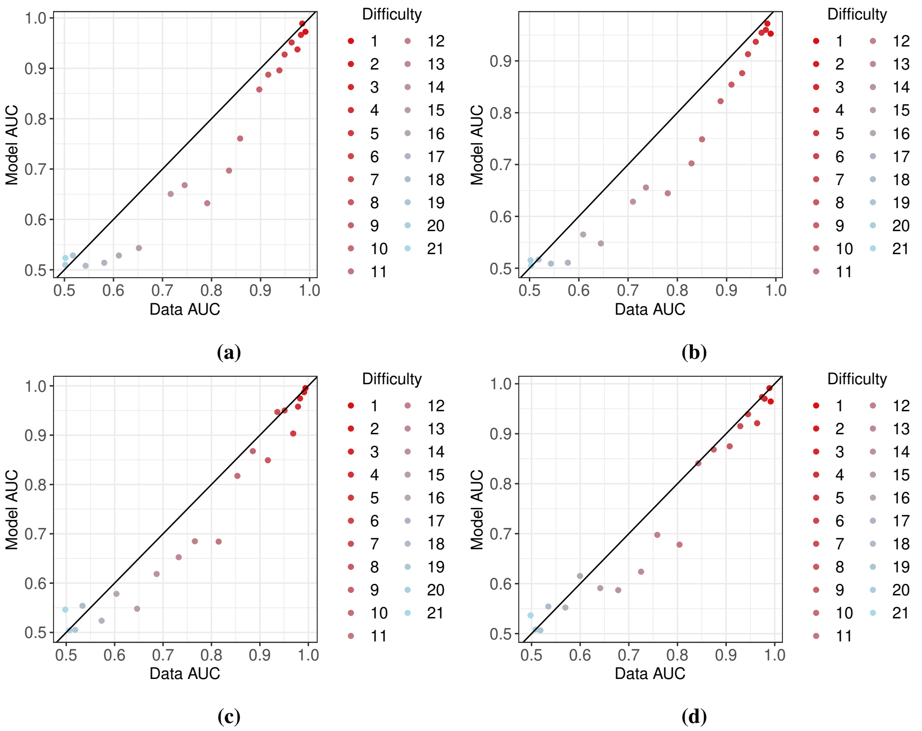


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 6 Error plots (6a, 6b, 6c), reproducibility plots (6d, 6e, 6f) and weight plots (6g, 6h, 6i) for different difficulty configurations, 1 (6a, 6d, 6g), 8 (6b, 6e, 6h) and 15 (6c, 6f, 6i) respectively, in the differences both in location and spread scenario



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

Fig. 7 Correlation plots for the (7a and 7c) t-test and the (7b and 7d) Wilcoxon rank sum test in the scenario of differences in location (7a and 7b) and in the scenario of both differences in both location and spread (7c and 7d)

Table 2 Kendall correlations for the different combinations of methods and problems when dealing with synthetic data

Problem	$\tau - b$	
	T-test	Wilcoxon test
Location	0.905	0.905
Location & spread	0.905	0.876

6.2 Experimentation with real data

In the experimentation with real data, four different ranking methods are used. Apart from the t-test and the Wilcoxon rank sum test, a ranking method based on the mutual information and a ranking method based on the coefficients of a linear SVM are also used. Specifically, the last two methods have been frequently used in the literature to tackle the feature selection within supervised classification problems.

Regarding the datasets, we have used 5 different datasets, four obtained from the UCI Repository [10] and another obtained from the GEO Repository. In Table 3 we display the amounts of features and instances that each dataset has. For the interested reader, a description of each dataset and an explanation of the preprocessing applied to each dataset is provided in the supplementary material.

Table 3 Amounts of features and instances that each real dataset has

Database	Amount of features	Amount of instances		
		Group 1	Group 2	Total
Breast cancer [27,37]	30	357	212	569
Mice protein expression [18]	77	570	510	1080
SECOM [28]	591	1463	104	1567
Arcene [15]	10000	502	398	900
Ovarian cancer [38]	27578	274	266	540

For each of the real datasets, the two procedures exposed (see Subsection 3.2) to estimate the ERC and the four ranking methods previously mentioned are applied. Given the amount of results collected, for the sake of brevity in this paper we will only show a representative subset of the obtained results when dealing with real data in Figures 8 to 9 and in Tables 4 and 5, leaving the rest of the results in the supplementary material for the interested reader. Please bear in mind that, unfortunately, since the labels of the real data are unknown, we cannot compute the data AUC with which to compare the model AUC.

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

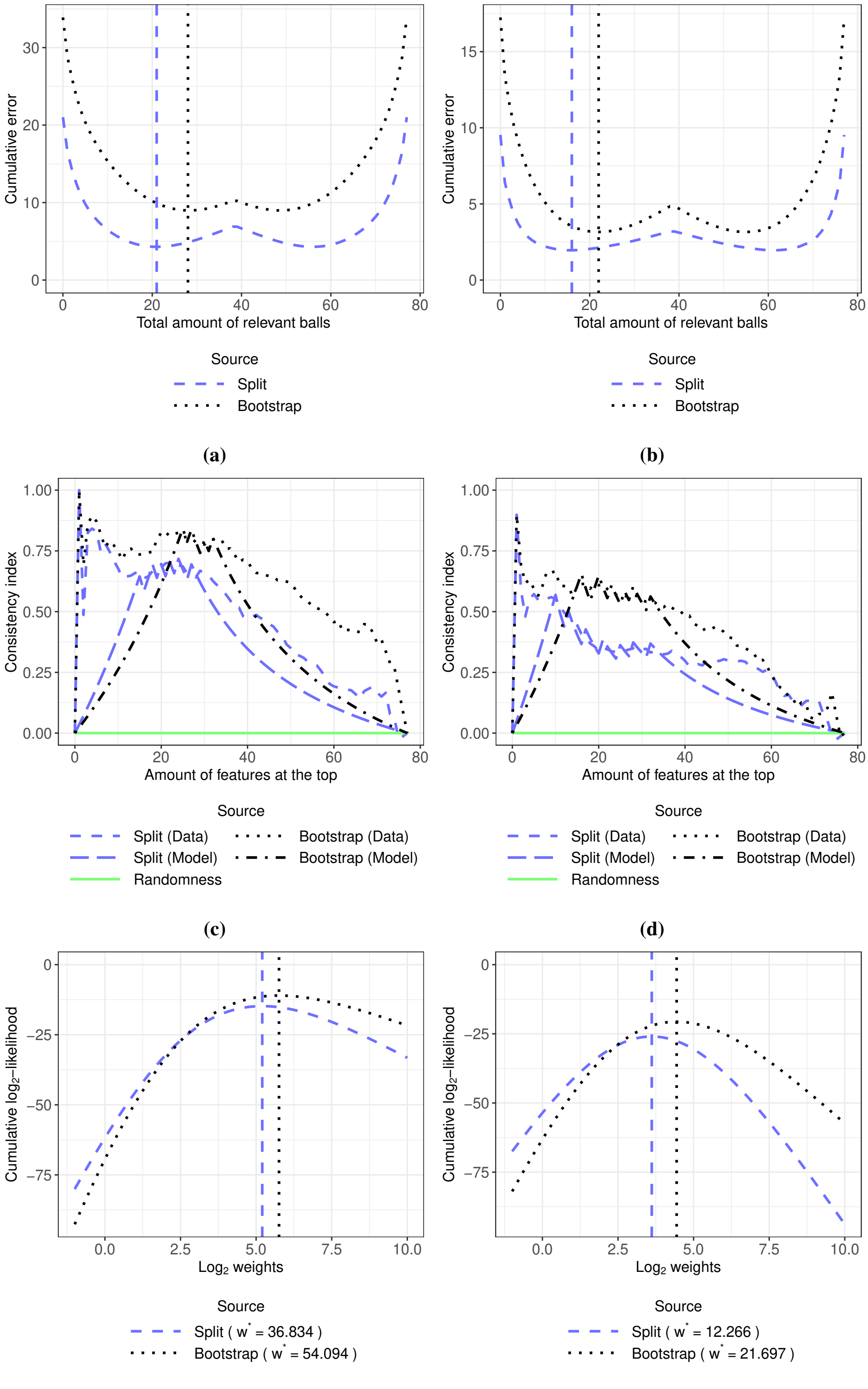
Database	Estimation	Mutual information	SVM	T-test	Wilcoxon test
Breast cancer	Random split	$2.335 \cdot 10^{19}$	3.301	60.500	$2.335 \cdot 10^{19}$
Breast cancer	Bootstrap	$2.335 \cdot 10^{19}$	6.392	58.182	$2.248 \cdot 10^{19}$
Mice	Random split	36.834	12.266	25.282	24.479
Mice	Bootstrap	54.094	21.697	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

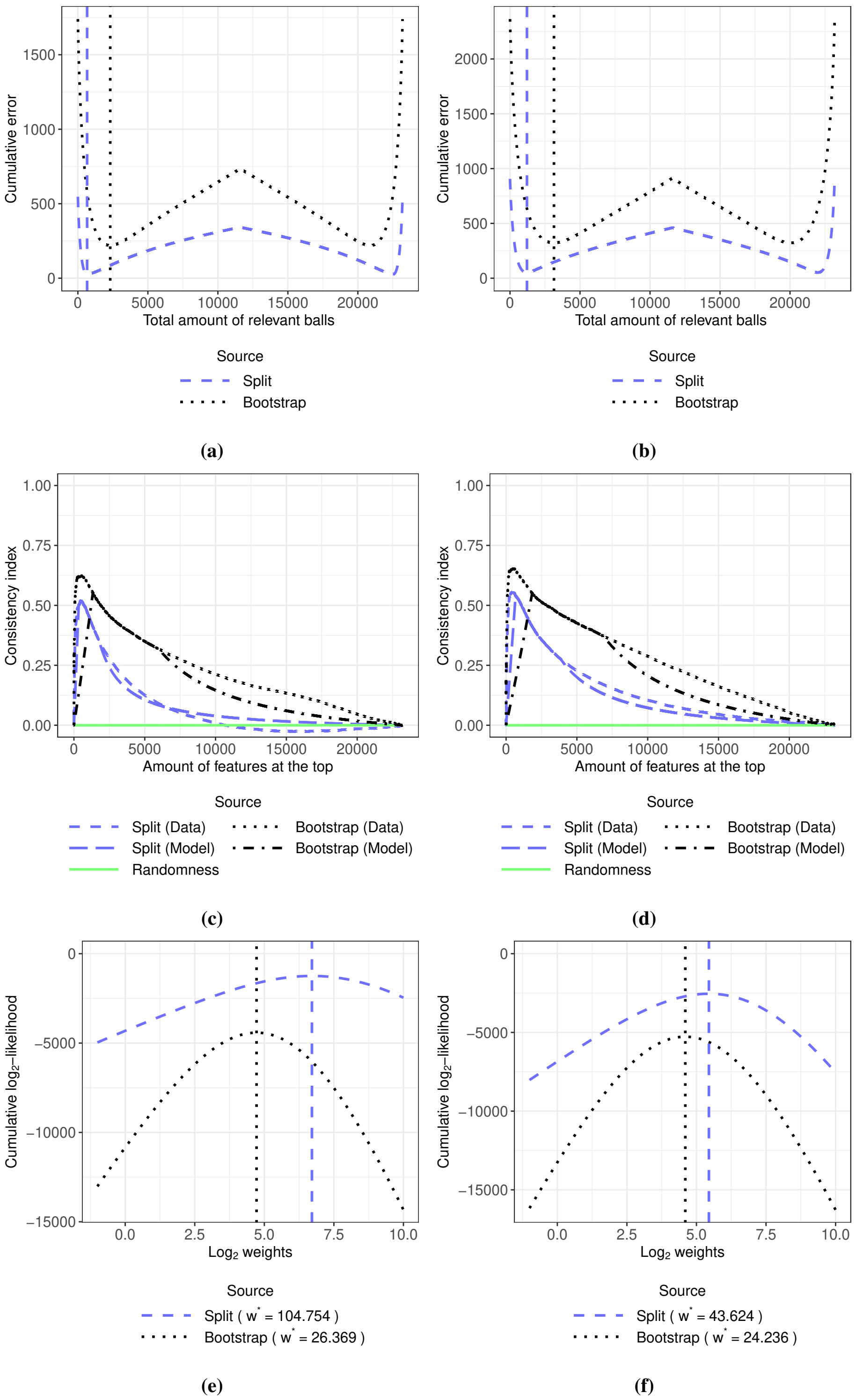
6.3 Discussion

In this subsection, we briefly discuss the results derived from the experiments with synthetic and real data.



(a) **(b)** **(c)** **(d)** **(e)** **(f)**

Fig. 8 Error plots (8a,8b), reproducibility plots (8c, 8d) and weight plots (8e, 8f) for the mice protein expression dataset when the mutual information is used as a ranking method (8a, 8c, 8e) and when the SVM is used as a ranking method (8b, 8d, 8f)



(a) **(b)**

(c) **(d)**

(e) **(f)**

Fig. 9 Error plots (9a,9b), reproducibility plots (9c, 9d) and weight plots (9e, 9f) for the ovarian cancer dataset when the t-test is used as a ranking method (9a, 9c, 9e) and when the Wilcoxon test is used as a ranking method (9b, 9d, 9f)

In the results obtained when dealing with synthetic data, it can be seen (Figures 5 and 6) that the proposed dichotomous model for reproducibility curves fits well to synthetic data coming from two types of features (with and without differences between groups). It is worth noting that the model fits well regardless of the assumption that the amount of relevant balls in the top-i of any sequence of extractions is the same for any $i \in \{1,\ldots,n\}$. Besides, it can be appreciated that, as the scenario becomes more difficult from one configuration to another (columns from left to right in Figures 5 and 6), the difference between the number of truly relevant features and the number of relevant balls identified by the model tends to increase. Such an occurrence makes perfect sense, since the increase in the difficulty is a direct consequence of the two types of features becoming more similar. In fact, as the two types of features become more similar, it becomes harder for the fitted dichotomous model to retrieve an accurate estimation of the amount of each type. Moreover, in the extreme case in which the two types of features are equal, there would not be a way to distinguish them any longer.

Regarding the weights, although it is necessary to point out that the weight values seem to be quite sensitive, in general, their values make sense, approaching 1 as the scenario becomes more difficult from one configuration to another. Specifically, when the difficulty is so high that the feature ranking methods are incapable of ranking relevant features before irrelevant features consistently, the achieved weight values near 1 appropriately represent that lack of tendency to rank relevant features before irrelevant features consistently. In summary, the weights reflect the difficulty to distinguish the relevant features from the irrelevant ones (the higher the w^*, the lower the difficulty). In addition, that occurrence hints that the weights enable the degree of relevance of the relevant features to be quantitatively assessed in an intuitive manner.

Regarding the AUCs, it can be seen qualitatively in Figure 7 and quantitatively in Table 2 that there is a high correlation in terms of the Kendall rank correlation coefficients between the model AUCs and the data AUCs. Namely, the rankings of the algorithms according to the model AUCs and the rankings according to the data AUCs are very similar, and, thus, the estimated model AUC could be used for algorithm comparison purposes. Besides, the model AUC values provide information regarding the degree of relevance of the relevant features. In that sense, the model AUC values complement the weights derived from the model. In addition, a model AUC summarizes a reproducibility curve in a single scalar value, which eases the comparison of RFSS algorithms from the point of view of their reproducibility.

However, the fittings to the real data give results that are not as good as the results achieved with synthetic data. Nonetheless, the fitted model can still provide sensible outcomes (e.g., 9c, 9d) despite its constraints (e.g., only two different types of features, relevant and irrelevant). In fact, it seems that the further a given real dataset is from having just two types of features, the worse the fit of the model (conversely, the closer, the better). In particular, apparently the model for reproducibility curves tends to issue a total amount of relevant balls a_n at a point of equilibrium for several cases (e.g., Subfigures 8c, 8d). Namely, in such cases, 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 relevant 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 actually does.

When dealing with real data, the different weights obtained serve as a quantitative heuristic that eases the assessment of the degree of separation of the different types of features according to $\boldsymbol{a}^*$. Namely, although in sequence $\boldsymbol{a}^*$ the degree of mixing between the different types of features can be observed, the weights summarize that information in a single scalar value. Specifically, it is consistent to obtain the highest weights when the mutual information and Wilcoxon test are used on the breast cancer dataset, given that, according to the reproducibility curves derived from the model, there is a complete separation between the two different types of features identified (see supplementary material). Besides, it is consistent to obtain the lowest weights when the SVM is used in the ovarian cancer dataset while performing random split sampling since the modeled reproducibility curve is the closest one to the reproducibility curve of the randomness.

Regarding the model AUCs derived when dealing with real datasets, all of them are above 0.9, except for half of the cases when the SVM method is applied. Such an occurrence suggests that, in general, each method truly discriminates between the two supposed types of features, although the SVM runs into difficulties in some cases.

7 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 RFSS algorithms. Specifically, the approach starts with the use of a given RFSS algorithm in different subsamplings and resamplings of a given experimental dataset. Next, the expected Kuncheva's consistency index for each of the subsets of different size of the top-ranked features is computed to build what we refer to as a "reproducibility curve". Then, a novel urn-based model in which an ordering of features is conceived as a full sequence of extractions of balls from the urn is posed. In particular, in the model there are conceived relevant and irrelevant balls, each with different weights related to how likely they are to be drawn. The fitting of the model is composed of two sequential steps. The first step is the identification of the sequence $\boldsymbol{a}^*$ of relevant and irrelevant balls that minimizes an error function regarding the expected Kuncheva's consistency indexes previously calculated. The second step is the search of the weight w^* of the relevant balls that maximizes the likelihood of the identified sequence of balls.

Once the model is fitted to data, it provides practical, intuitive and easy to interpret information regarding the reproducibility and the performance of the given RFSS algorithm. To start with, the computed reproducibility curve estimates the expected reproducibility for each different possible size i of the subset of top-ranked features. The sequence $\boldsymbol{a}^*$ provides an estimate of how many relevant and irrelevant features may be found for each different possible size i of the subset of top-ranked features. The weight w^* assesses the tendency to rank relevant features before irrelevant features and summarizes it in a single scalar value. From the sequence $\boldsymbol{a}^*$, a model ROC curve can be derived, thus enabling us to gather a True Positive Rate (TPR) and a False Positive Rate (FPR) for each different possible size of the subset of top-ranked features. Consequently, the model AUC of that curve can be computed. That model AUC is related to the true data AUC in terms of detection of relevant features, thus providing useful information about it. In summary, the proposed model gathers information regarding both reproducibility and performance of a given RFSS algorithm.

In order to illustrate the behavior of the model for analyzing RFSS algorithms, we have conducted experiments both with synthetic data and real data. The experimentation with synthetic data enables us to test our proposal under controlled circumstances. Briefly, in the experimentation with synthetic data, each feature follows a known distribution and, consequently, we know which features are relevant and which are not. Namely, the experimentation with synthetic data allows us to compare the obtained reproducibility and performance measures with the true values. For the experimentation with real data, we selected several datasets, some of them belonging to the problem of biomarker selection, so that they can serve as an example of a feature selection problem within biomedical research. We chose some of the datasets to belong to the biomarker selection problem because its concerns regarding reproducibility are particularly strong, since some of its usual characteristics (far fewer individuals than features and far fewer true biomarkers than candidate biomarkers) hinder the achievement of reproducible results. In summary, the results of the experimentation with synthetic data and real data show that the proposed model can be used to analyze RFSS algorithms in terms of their reproducibility and their performance.

Regarding the results of the experimentation with synthetic data, apart from the sensitivity of the weights, the model can be well fitted to the reproducibility curves. On one hand, the lesser the difficulty of the problem, the more similar the amount of relevant features identified by the fitted model and the amount of truly relevant features. On the other hand, the true data AUCs of the methods and the model AUCs are highly correlated. This last fact suggests that the model AUCs can retrieve useful information for the selection of RFSS algorithms.

Concerning the results of the experimentation with real data, the proposed model manages to achieve a sensible and logical outcome, reaching a compromise solution. In the feature selection problem, both the identification of as many relevant features as possible and the efficient assignment of the possibly limited resources (time, effort, money, ...) are usually desired. Dealing with that trade-off and deciding the size i of the subset of top-ranked features that is worth researching are two problems that can be eased with our proposal, through the provision of a notion of how many relevant features may be found among the i top-ranked features in terms of the measure of relevance used.

Our research opens several future work lines. One interesting way to proceed consists of the extension of the model to more than just two types of balls, which most likely will increase the goodness 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. This extension can easily be incorporated to the model AUC through the use of the distance τ of Kendall, which, in fact, can be seen as a generalization of the AUC to multipermutations (permutations with repeated elements) that are not just dichotomous.

Another important line to follow is the estimation of the ERC. At this point we have offered an underestimation (random split) and an overestimation (bootstrap) of the ERC. However, it would be interesting to achieve narrower practical bounds through other approaches to the estimation of the ERC. Also, it would be interesting to associate to practical bounds a probability of the ERC to fall within those bounds.

8 Online resources

8.1 Online resource 1 — Online resource 1: Supplementary material

A Portable Document Format (PDF) file which contains supplementary material. It includes details of the fitting process and of the experimentation carried out, and can be found at the following URL: https://github.com/isg-ehu/ari.urkullu

Acknowledgements This work is partially supported by the Basque Government (IT609-13, Elkartek BID3A and Elkartek project 3KIA, KK2020/00049) and the Spanish Ministry of Economy and Competitiveness

MINECO (TIN2016-78365-R and project PID2019-104966GB-I00) and a University-Society Project 15/19 (Basque Government and University of the Basque Country UPV/EHU). Ari Urkullu has been supported by the Basque Government through a predoctoral grant (PRE_2013_1_1313, PRE_2014_2_87, PRE_2015_2_0280 and PRE_2016_2_0314). Aritz Pérez has been partially supported through the BERC 2018-2021 program and the ELKARTEK program (Basque Government), the excellence accreditation Severo Ochoa SEV-2017-0718 (Spanish Ministry of Science, Innovation and Universities) and the project TIN2017-82626-R (Spanish Ministry of Economy and Competitiveness). Borja Calvo has been partially supported by the ELKARTEK program (Basque Government), and the project TIN2016-78365-R (Spanish Ministry of Economy and Competitiveness).

Our work was assisted by the UCI repository by supplying datasets.

References

1. Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y.: Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics **26**(3), 392–398 (2009)
2. Alelyani, S., Zhao, Z., Liu, H.: A dilemma in assessing stability of feature selection algorithms. In: High Performance Computing and Communications (HPCC), 2011 IEEE 13th International Conference on, pp. 701–707. IEEE (2011)
3. Awada, W., Khoshgoftaar, T.M., Dittman, D., Wald, R., Napolitano, A.: A review of the stability of feature selection techniques for bioinformatics data. In: Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on, pp. 356–363. IEEE (2012)
4. Baker, M.: 1,500 scientists lift the lid on reproducibility. Nature News **533**(7604), 452 (2016)
5. Brent, R.P.: Algorithms for minimization without derivatives. Prentice-Hall, Englewood Clifts, New Jersey (1973)
6. Cai, J., Luo, J., Wang, S., Yang, S.: Feature selection in machine learning: A new perspective. Neurocomputing **300**, 70–79 (2018)
7. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering **40**(1), 16–28 (2014)
8. Chelvan, P.M., Perumal, K.: A comparative analysis of feature selection stability measures. In: Trends in Electronics and Informatics (ICEI), 2017 International Conference on, pp. 124–128. IEEE (2017)
9. Dernoncourt, D., Hanczar, B., Zucker, J.D.: Analysis of feature selection stability on high dimension and small sample data. Computational statistics & data analysis **71**, 681–693 (2014)
10. Dua, D., Graff, C.: UCI machine learning repository (2017). URL http://archive.ics.uci.edu/ml
11. Duda, R.O., Hart, P.E.: Pattern classification and scene analysis. A Wiley-Interscience Publication, New York: Wiley, 1973 (1973)
12. Dunne, K., Cunningham, P., Azuaje, F.: Solutions to instability problems with sequential wrapper-based approaches to feature selection. Journal of Machine Learning Research pp. 1–22 (2002)
13. Goodman, S.N., Fanelli, D., Ioannidis, J.P.: What does research reproducibility mean? Science translational medicine **8**(341), 341ps12–341ps12 (2016)
14. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of machine learning research **3**(Mar), 1157–1182 (2003)
15. 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)
16. Haury, A.C., Gestraud, P., Vert, J.P.: The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS one **6**(12), e28210 (2011)
17. He, Z., Yu, W.: Stable feature selection for biomarker discovery. Computational biology and chemistry **34**(4), 215–225 (2010)
18. 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)
19. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms. In: Data Mining, Fifth IEEE International Conference on, pp. 8–pp. IEEE (2005)
20. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and information systems **12**(1), 95–116 (2007)
21. Kendall, M.G.: The treatment of ties in ranking problems. Biometrika **33**(3), 239–251 (1945)
22. Khoshgoftaar, T.M., Fazelpour, A., Wang, H., Wald, R.: A survey of stability analysis of feature subset selection techniques. In: Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on, pp. 424–431. IEEE (2013)
23. Kuncheva, L.I.: A stability index for feature selection. In: Artificial intelligence and applications, pp. 421–427 (2007)

24. Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H.: Feature selection: A data perspective. ACM Computing Surveys (CSUR) **50**(6), 94 (2017)

25. Li, Q., Brown, J.B., Huang, H., Bickel, P.J., et al.: Measuring reproducibility of high-throughput experiments. The annals of applied statistics **5**(3), 1752–1779 (2011)

26. Lustgarten, J.L., Gopalakrishnan, V., Visweswaran, S.: Measuring stability of feature selection in biomedical datasets. In: AMIA annual symposium proceedings, vol. 2009, p. 406. American Medical Informatics Association (2009)

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

28. 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)

29. Nogueira, S., Brown, G.: Measuring the stability of feature selection with applications to ensemble methods. In: International Workshop on Multiple Classifier Systems, pp. 135–146. Springer (2015)

30. Nogueira, S., Brown, G.: Measuring the stability of feature selection. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 442–457. Springer (2016)

31. Nogueira, S., Sechidis, K., Brown, G.: On the stability of feature selection algorithms. The Journal of Machine Learning Research **18**(1), 6345–6398 (2017)

32. Nogueira, S., Sechidis, K., Brown, G.: On the use of spearmans rho to measure the stability of feature rankings. In: Iberian conference on pattern recognition and image analysis, pp. 381–391. Springer (2017)

33. RodríGuez, J.D., Pérez, A., Lozano, J.A.: A general framework for the statistical analysis of the sources of variance for classification error estimators. Pattern recognition **46**(3), 855–864 (2013)

34. Saeys, Y., Abeel, T., Van de Peer, Y.: Robust feature selection using ensemble feature selection techniques. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 313–325. Springer (2008)

35. Saeys, Y., Inza, I., Larrañaga, P.: A review of feature selection techniques in bioinformatics. bioinformatics **23**(19), 2507–2517 (2007)

36. Shanab, A.A., Khoshgoftaar, T.M., Wald, R., Napolitano, A.: Impact of noise and data sampling on stability of feature ranking techniques for biological datasets. In: Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on, pp. 415–422. IEEE (2012)

37. 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)

38. 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)

39. Wallenius, K.: Biased sampling; the noncentral hypergeometric probability distribution. Tech. rep., STANFORD UNIV CA APPLIED MATHEMATICS AND STATISTICS LABS (1963)

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

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

Author Biographies

Ari Urkullu received the MSc degree in computational engineering and intelligent systems from the University of the Basque Country in 2011. From 2011 to 2013, he was a software developer at Fullstep Networks. In 2014 he joined Intelligent Systems Group. From 2014 to 2018, he worked as a predoctoral researcher at the University of the Basque Country. In 2018 he was a temporary lecturer at the Department of Languages and Information Systems at the University of the Basque Country. From 2019 to the present, he has been working at Gestamp, where he currently works as a senior data scientist of the advanced analytics team. Besides, he works to finish his Ph.D. in informatics engineering under the supervision of Borja Calvo and Aritz Pérez. His research interests include bioinformatics, supervised and unsupervised classification, feature selection, model selection and evaluation, and both classification and forecasting of multivariate time series.

Aritz Pérez received his Ph.D. degree in 2010 from the University of Basque Country, Department of Computer Science and Artificial Intelligence. Currently, he is a postdoctoral researcher at the Basque Center for Applied Mathematics. His current scientific interests include supervised, unsupervised and weak classification, probabilistic graphical models, model selection and evaluation, time series, and crowd learning.

Borja Calvo received his master's degree in Biochemistry from the University of the Basque Country in 1999 and, after two years working, he took a bachelor's degree in Computer Science at the same university. In 2004 he earned the bachelor's degree and in 2008 the Ph.D. in Computer Science. After two years as a PostDoc researcher at the Intelligent Systems Group, in 2011 he won his current lecturer position at the Department of Computer Science and Artificial Intelligence of the University of the Basque Country. Currently, he is leading a research project funded by the DGT (spanish traffic agency) aimed at the prediction of car accidents in the basque road network. He is also supervising three Ph.D. students and several master thesis.