Journal of Machine Learning Research 22 (2021) 1-57 Submitted 4/20; Revised 4/21; Published 5/21

An Importance Weighted Feature Selection Stability Measure

Victor Hamer VICTOR.HAMER@UCLOUVAIN.BE
Pierre Dupont PIERRE.DUPONT@UCLOUVAIN.BE
UCLouvain - ICTEAM/INGI/Machine Learning Group,
Place Sainte-Barbe 2,
B-1348 Louvain-la-Neuve, Belgium.

Editor: Isabelle Guyon

Abstract

Current feature selection methods, especially applied to high dimensional data, tend to suffer from instability since marginal modifications in the data may result in largely distinct selected feature sets. Such instability strongly limits a sound interpretation of the selected variables by domain experts. Defining an adequate stability measure is also a research question. In this work, we propose to incorporate into the stability measure the importances of the selected features in predictive models. Such feature importances are directly proportional to feature weights in a linear model. We also consider the generalization to a non-linear setting.

We illustrate, theoretically and experimentally, that current stability measures are subject to undesirable behaviors, for example, when they are jointly optimized with predictive accuracy. Results on micro-array and mass-spectrometric data show that our novel stability measure corrects for overly optimistic stability estimates in such a bi-objective context, which leads to improved decision-making. It is also shown to be less prone to the under- or over-estimation of the stability value in feature spaces with groups of highly correlated variables.

Keywords: feature selection, selection stability, bi-objective optimization, bioinformatics, feature importance

1. Introduction

Feature selection, *i.e.* the selection of a small subset of informative and relevant features to be included in a predictive model, has become compulsory for a wide variety of applications due to the appearance of very high dimensional data sets, notably in the biomedical domain (Saeys et al., 2007). Filtering noisy and irrelevant features can avoid overfitting the data and potentially improve predictive performance. Feature selection also allows for the learning of fast and compact models, which are easier to interpret. Such models can then be analyzed by domain experts and are easier to validate. Getting more interpretable models is also a key concern nowadays, and even considered by many as a requirement, when deployed in the medical domain.

Feature selection has been already studied in depth (Tang et al., 2014; Saeys et al., 2007; Kalousis et al., 2007b). Yet, current methods are still somewhat unsatisfactory mainly because of the typical instability they exhibit. Instability here refers to the fact that the

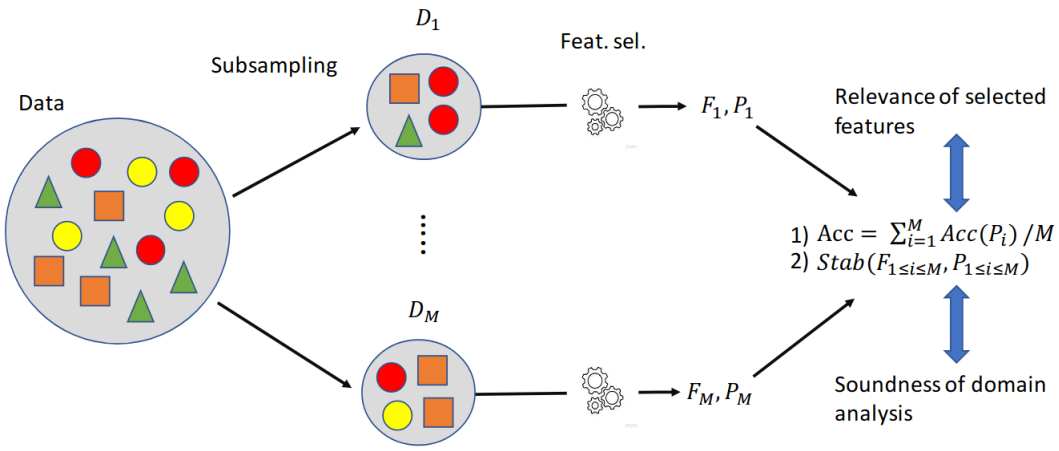


Figure 1: Illustration of the stability problem. The outcome is a measure of the trade-off between predictive accuracy and selection stability. The accuracy relates to the relevance of the selected features while stability is linked to the soundness of the domain analysis.

selected features may drastically change even after marginal modifications of the data, or, more generally, after some fine-tuning of the data production or data analysis pipeline. Figure 1 illustrates such a phenomenon. The initial data set is perturbed[1] to form M different data sets $\mathcal{D}_{1\leq i\leq M}$. A feature subset $\mathcal{F}_{1\leq i\leq M}$ is selected from each of these modified versions of the initial data set. A predictive model $\mathcal{P}_{1\leq i\leq M}$ is then built on each of the feature subsets and evaluated on some test examples. The pipeline depicted in Figure 1 has two non-necessarily competing objectives: 1) a measure of the performance of the predictive models built on the selected features and 2) a measure of the stability of the selected features which is related to the soundness of the domain analysis. Possible additional quality criteria are minimal model size or sparsity. Instability arises when little agreement over the selected features occurs, *i.e.* when the second objective is not met. This prevents a correct and sound interpretation of the selected features and strongly impacts their further validation by domain experts as it reduces their trust towards the proposed features. These experts would often prefer a more stable feature selection algorithm over an unstable and slightly more accurate one (Kalousis et al., 2007a; Saeys et al., 2008b). This is especially true in the biomedical field where reproducibility has proven to be a key challenge (Haibe-Kains et al., 2013). Unlike optimizing the accuracy of predictive models, optimizing selection stability may look trivial since an algorithm always returning an arbitrary but fixed set of features would be stable by design. Yet, such an algorithm is not expected to select informative and predictive features and would thus fail to meet the first objective above. This illustrates that optimizing stability is only well-posed jointly with predictive accuracy which, as in Figure 1, can be measured by the average accuracy of the M learned models.

1. Here it is done by subsampling which is often used to measure such instability, but it could be any small perturbation.

A common approach to estimate the second objective is to measure the stability of the feature subsets, on which are built the predictive models, without considering these models in the stability value. This strategy has been applied in (Sechidis et al., 2019; Hamer and Dupont, 2020). These papers also acknowledge the existence of a Pareto front in the (accuracy, stability) objective space. Considering such a subset selection stability in the model selection can also reduce the number of irrelevant selected features (Nogueira et al., 2017a). Our recent work (Hamer and Dupont, 2020) goes further by jointly optimizing selection stability and predictive performance and by deriving Pareto-optimal compromises using extensions of the well-known recursive feature elimination (RFE) algorithm (Guyon et al., 2002).

In this paper, we demonstrate the limitations that occur when one uses subset stability measures. Instead, we aim at quantifying the stability of a partial feature weighting, where each feature weight represents the importance of the corresponding selected feature in the associated predictive model. In the simplest case corresponding to a (generalized) linear model, the importance of a feature is directly proportional to its associated weight in such a model.[2] Such an objective, in addition to providing more refined feature preferences for domain analysis, is shown to be more adequate in certain situations. This is the case when jointly optimizing selection stability with predictive performance or when the feature space is composed of highly correlated feature groups. Our contributions include

- A visualization tool allowing the intuitive assessment of stability (Section 3).
- A new weight-based stability measure, closely matching the visual interpretation, which provably satisfies several properties (fully defined, bounds, correction for chance and maximum stability $\Leftrightarrow$ deterministic importance), some of which are not fulfilled by current weight-based stability measures (Section 5).
- A generic method to evaluate the importance of a feature in a predictive model (Section 5.1).
- The theoretical justification and experimental validation that current stability measures do not behave adequately in the presence of highly correlated feature groups, while our measure improves this behavior (Section 5.3).
- The introduction (previous work) and extension (current work) of an approach to optimize jointly predictive accuracy and selection stability (Sections 6.1 and 6.2).
- The theoretical justification (Section 6.3) and experimental validation (Section 7.1) that our proposed measure improves decision-making when stability is optimized jointly with predictive accuracy.

2. Related Work

Feature selection techniques are generally split into three categories: filters, wrappers, and embedded methods. *Filters* evaluate the relevance of features independently of the final model, most commonly a classifier, and remove low ranked features. Simple filters (*e.g.*

2. We study the non-linear case in Section 5.1.

t-test or ANOVA) are univariate, which is computationally efficient and tends to produce a relatively stable selection but they plainly ignore the possible dependencies between various features. Information-theoretic methods, such as mRMR (Ding and Peng, 2005) and many others, are based on mutual information between features or with the response, but a robust estimation of these quantities in high dimensional spaces remains difficult. *Wrappers* look for the feature subset that will yield the best predictive performance on a validation set. They are classifier dependent and very often multivariate. However, they can be very computationally intensive and an optimal feature subset can rarely be found. *Embedded methods* select features by determining which features are more important in the decisions of a predictive model. Prominent examples include SVM-RFE (Guyon et al., 2002) and logistic regression with a LASSO (Tibshirani, 1996) or ELASTIC NET penalty (Zou and Hastie, 2005). These methods tend to be more computationally demanding than filters but they integrate into a single procedure the feature selection and the estimation of a predictive model. Yet, they also tend to produce much less stable models. Recently, deep neural networks have started to be used as feature selectors as well (Li et al., 2016; Roy et al., 2015).

Some works specifically study the reasons behind selection instability. Results show that it is mostly caused by the small sample/feature ratio (Alelyani, 2013), noise in the data (Shanab et al., 2012), or imbalanced target variable (Awada et al., 2012) and feature redundancy (Somol and Novovicova, 2010). While all of these reasons clearly play a role, the small sample/feature ratio and feature redundancy are likely the most important ones in a biomedical domain with typically several thousands, if not millions, of sometimes highly correlated features for only a few dozens or hundreds of samples. This is likely why stable feature selection is intrinsically hard in this domain and why existing techniques are still unsatisfactory.

Looking for a stable feature selection also requires a proper way to quantify stability itself. In general, the stability of a feature selection algorithm relates to the robustness of its feature preferences with respect to small modifications of the data. Feature selection algorithms can produce either (a) a subset of selected features, (b) a partial or complete ranking of the features or (c) a weight for each feature which typically assesses the importance of this feature in a predictive model. Each type of feature selection requires dedicated stability measures.

Many subset-based selection stability measures have already been proposed: the Kuncheva index (Kuncheva, 2007), the Jaccard index (Kalousis et al., 2005), the POG (Shi et al., 2006) and nPOG (Zhang et al., 2009b) indices among others. Under such a profusion of different measures, it becomes difficult to justify the choice of a particular index and even more to compare results of works based on different metrics. Furthermore, the large number of available measures can lead to publication bias (researchers may select the index that makes their algorithm look the most stable) (Boulesteix and Slawski, 2009). In the hope of fixing this issue, a more recent work (Nogueira et al., 2017a) lists and analyzes 15 different stability measures. They are compared based on the satisfaction of 5 different properties that a stability measure should comply with. They also propose a novel and unifying index. This index, which we extensively use in this work, is described in more detail in Section 2.1. A related and popular weight-based stability measure, which makes use of the sample Pearson's correlation coefficient (Kalousis et al., 2007a; Nogueira and Brown, 2016), is reviewed in Section 2.2.

In this work, we focus on the study of the feature importances and their variations rather than the feature positions in a ranking. Still, we briefly consider the ranking measure proposed in (Jurman et al., 2008) which is defined from the Canberra distance. This measure is reviewed in Section 2.3. Section 4 illustrates the differences between the three types of stabilities and shows that ranking measures can deviate from our purpose of assessing the stability of the selected features.

Several authors have proposed different approaches to increase stability. For instance, instance-weighting for variance reduction (Han and Yu, 2012) and ensemble methods for feature selection have been proposed (Saeys et al., 2008a; Abeel et al., 2010) and generally increase selection stability. Stability selection (Meinshausen and Bühlmann, 2010) is a particular ensemble method which ultimately selects features with a selection frequency higher than a threshold π_{thr} for at least one regularisation parameter $\lambda \in \Lambda$. While these methods have been shown to increase selection stability, the gain they offer is still limited as they are not designed to search explicitly through a bi-dimensional (accuracy, stability) objective space.

In order to explicitly tune the accuracy/stability trade-off, a hybrid version of the well-known recursive feature elimination algorithm has been recently proposed in (Hamer and Dupont, 2020). In essence, the selection is stabilized by forcing the selection of some features based on univariate criteria which are generally more stable than multivariate selection methods. This approach is more extensively reviewed in Section 6.1. We will then use and extend this work as an illustration of how current stability measures are susceptible to undesirable behaviors when one tries to increase selection stability.

2.1 A Unifying Selection Stability Index

The stability index introduced in (Nogueira et al., 2017a) measures the stability across M selected subsets of features. It can be computed according to Equation (1)

$$\phi = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\frac{\bar{k}}{d} * (1 - \frac{\bar{k}}{d})} \tag{1}$$

with $\bar{k}$ the mean number of features selected from the original d features and $s_f^2 = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$ the estimator of the variance of the selection of the f_{th} feature over the M selected subsets, where $\hat{p}_f$ is the fraction of times feature f has been selected among them. These subsets are typically obtained by resampling M times the learning data. This is the only existing measure satisfying the 5 properties described in (Nogueira et al., 2017a):

- *Fully defined*: the measure is defined for every possible combinations of M feature subsets.
- *Strict monotonicity*: the measure strictly decreases with the selection variance.
- *Bounds*: the measure is bounded by constants.
- *Maximum stability* $\Leftrightarrow$ *deterministic selection*
- *Correction for chance*: under the null model of feature selection, the expected value of the measure is constant.

The null model of feature selection, noted H_0, is defined in (Nogueira et al., 2017a) as *the situation where every possible feature subset has an equal probability to be selected.*

The *correction for chance* property states that, under H_0, ϕ is constant in expectation (here set to 0). The intuition behind this property is that a stability measure should not be influenced by the similarities between selected feature subsets that occur by chance. In addition to the *correction for chance* property, ϕ is formally bounded by -1 and 1 and is asymptotically lower bounded by 0 as $M \to \infty$. This measure is equivalent to the Kuncheva index (KI) (Kuncheva, 2007) when the number of selected features k is constant across the M selected subsets, but it can be computed in $O(Md)$ whereas KI requires $O(M^2 d)$. The Kuncheva index can be computed using Equation (2)

$$KI = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \frac{|\mathcal{F}_i \cap \mathcal{F}_j| - \frac{k^2}{d}}{k - \frac{k^2}{d}} \tag{2}$$

where $|\mathcal{F}_i \cap \mathcal{F}_j|$ is the number of features subset i, $\mathcal{F}_i$, and subset j, $\mathcal{F}_j$, have in common.

Subset stability measures are either similarity-based or frequency-based. Similarity-based measures, such as the Kuncheva index, defines the stability as the average pairwise similarity between pairs of selected feature subsets. Frequency-based measures, such as ϕ, rather use the selection frequencies of each feature in the stability definition. (Nogueira et al., 2017a) bridge the gap between these two families of measures by proving the following result:

$$\frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j| = \bar{k} - \sum_{f=1}^{d} s_f^2. \tag{3}$$

A stability index increasing with the size of the intersection between pairwise feature subsets can then be re-formulated as another index, decreasing with the feature selection variance. In this work, we extend the Kuncheva index to handle a varying number of features and feature importances and we show that it can be re-formulated as a weighted frequency-based measure.

2.2 Pearson's Correlation

A popular weight-based stability measure computes the average correlation between feature weights of different selection runs:

$$\phi_{\text{pears}} = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \rho_{i,j} \tag{4}$$

where

$$\rho_{i,j} = \frac{\sum_{f=1}^{d} (w_{f,i} - \mu_i)(w_{f,j} - \mu_j)}{\sqrt{\sum_{f=1}^{d} (w_{f,i} - \mu_i)^2} * \sqrt{\sum_{f=1}^{d} (w_{f,j} - \mu_j)^2}},$$

with $w_{f,i}$ the weight, or score, associated to feature f in selection run i and μ_i the average feature weight in this run. Nogueira and Brown (2016) prove that, if these weights are either 0 or 1 (indicating the selection of the feature) and if the number of non-zero weights is constant across selection runs, then ϕ_{pears} is equivalent to the Kuncheva index. By extension, it is also equivalent to ϕ in this particular setting.

2.3 The Canberra Distance Between Partial Rankings

The Canberra distance, with location parameter k, evaluates the stability of partial feature rankings (of size k) and is defined as follows (Jurman et al., 2008):

$$\phi^k_{\text{can}} = 1 - \frac{1}{\chi}\frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=i+1}^{M}\sum_{f=1}^{d}\frac{|\min(r_i(f),k+1)-\min(r_j(f),k+1)|}{\min(r_i(f),k+1)+\min(r_j(f),k+1)} \tag{5}$$

with $r_i(f)$ the rank of feature f in ranking i and $\chi = \frac{(k+1)(2d-k)}{d} \times \log(4) - \frac{2kd+3d-k-k^2}{d}$ the term which approximately corrects the measure for chance. This measure naturally penalizes more variability that occurs at the top of the ranking.

3. Feature Stability Maps

In this section, we propose a visualization tool that allows the intuitive estimation of the stability of the learning process outcomes which are here a set of selected features and a predictive model built on them (see Figure 1). Each row of these feature stability maps represents a given decision model, the whole map representing the M models learned over the M resamplings ($M = 30$ here). Each feature is assigned a color and the horizontal extensions of the rectangles measure the importance of the features in the corresponding decision models, here linear ones estimated on each of such resamplings. For clarity, the features are sorted from left to right in decreasing order of selection frequency p_f across runs.

As a domain analysis tool, these so-called feature stability maps allow the identification of the most reliably identified features (the most frequently selected features which are put at the left of such maps) and the most important features in the predictive models (the largest rectangles) that are subsequently built on the selected features. Features combining both properties are likely to be particularly appealing for domain experts. In the example of Figure 2, the green feature seems particularly interesting, as it is selected in every run and also matters the most in the predictive models. In contrast, the mauve and orange features are selected in most of the selection runs but are much less important in the predictive models. They look thus less appealing for subsequent analysis as they actually matter less in the involved process.

4. Motivation for weight-based Measures

The primary goal of increasing stability is the improvement of the domain experts confidence towards the learning (and selection) algorithms and more specifically, their outcomes. In this paper, we propose a measure (ϕ_{iw}, formally defined in Section 5) which weights the contributions of the selected features in the selection stability by their relative importance in the associated predictive models. We motivate here on several examples that such a weighted stability is beneficial for the primary goal stated above.

A first example of feature stability map is displayed in Figure 3. The learning algorithm selects the same 20 variables in each run. They are combined in a multivariate predictive model where each feature plays an approximately equal role. This map would be perfectly interpretable by domain experts.

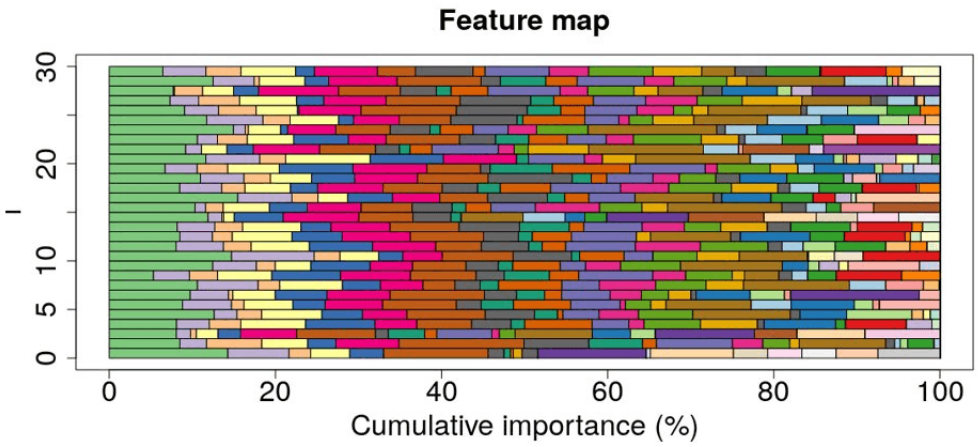


Figure 2: Example of feature stability map. Each row represents a given predictive model (*e.g.* a classifier or a regression model). Features are assigned unique colors and the horizontal extensions of the rectangles measure their importance in the models. The features are sorted from left to right in decreasing order of selection frequency p_f.

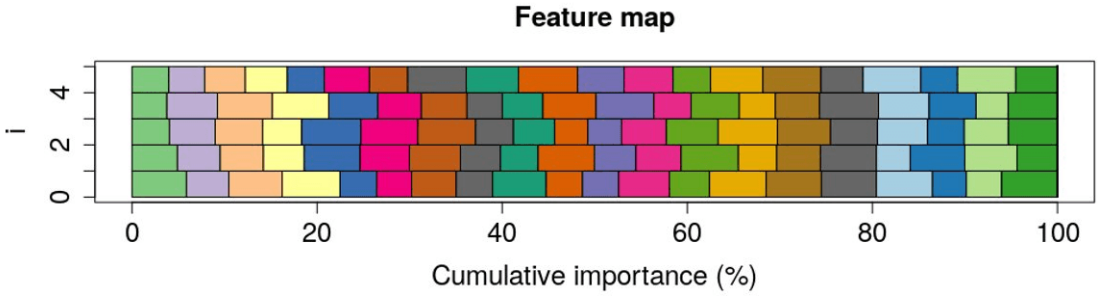


Figure 3: Feature stability map indicating strong interpretability. The measures ϕ, ϕ_{pears} and ϕ_{iw} are high but the ranking measure ϕ_{can} is not ($\phi = 1$, $\phi_{\text{pears}} = 0.97$, $\phi_{\text{iw}} = 0.9$ and $\phi_{\text{can}} = 0.47$).

In this case, both subset-based and weight-based stability measures are very high, even equal to 1 for subset-based measures such as ϕ (assuming $d > 20$ here). For ranking stability measures, this is not the case, as the ranking between the 20 selected features is (by design in this toy example) random. To compute stability values, we assume that the number of input features d tends to ∞.

Figure 4 illustrates a particularly interesting scenario which has been drawn from our experiments. As features are sorted from left to right in decreasing selection frequency, left features are selected the most often and consequently are the most responsible for the apparent (subset-)selection stability which is good in this example ($\phi = 0.72$). However, it appears that the 15 features that are selected the most (among the 20 selected features

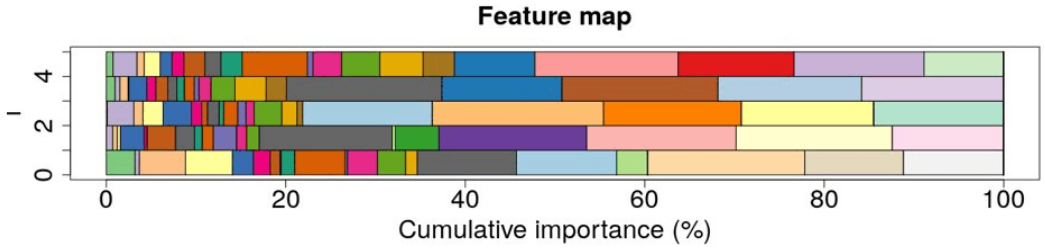


Figure 4: Example of a feature stability map for which ϕ would be drastically overestimated. The measures ϕ_{pears} and ϕ_{iw} (and ϕ_{can} to a lesser extent) correct this phenomenon ($\phi = 0.72$, $\phi_{\text{pears}} = 0.13$, $\phi_{\text{can}} = 0.39$ and $\phi_{\text{iw}} = 0.21$).

here) have a cumulative importance that is only around 25%. A domain expert would expect the most frequently selected features to be particularly useful to the task of interest (= high importance). In Figure 4, it appears to be the opposite, as some of them only play a marginal role in the predictive models.

A second kind of undesirable instability is depicted in Figure 5. In this example, even though the same subset is selected in each run, feature importance is highly varying across selection runs. Like the classical feature selection instability, a strong instability of the importance of the selected features is likely to deteriorate the interpretability of the selected features and the trust of domain experts towards their actual relevance. Subset-based stability measures are unable to grasp these nuances. Situations similar to the ones depicted in Figures 4 and 5 can naturally occur when selection stability is optimized jointly with predictive accuracy, as is studied further in Section 7.1. One can also see that ϕ_{can} is higher in Figure 5 than in Figure 3 which is orthogonal to our purpose of measuring the stability of the selected features importance. We focus our study on subset and weight-based measures for this reason and because dedicated ranking stability measures are not designed to compare rankings of different sizes (which occurs when the number of selected features varies from run to run). For situations where domain experts would rather be interested in a feature ranking, we refer the reader to (Urkullu et al., 2020; Jurman et al., 2008; Nogueira et al., 2017b; Kumar and Vassilvitskii, 2010).

As far as Figures 3, 4 and 5 are concerned, ϕ_{pears} is able to correctly identify instability. However, we note later that ϕ_{pears} lacks some important properties (see Table 2). Furthermore, it behaves inadequately in the setting depicted in Figure 6, which can naturally arise when the feature space is composed of highly correlated feature groups, as detailed in Section 5.3. In Figure 6, perfectly stable features have a cumulative importance of 50% while the other half of importance belongs to 5 features different in each run. Arguably, stability should be close to 0.5 in this setting which is the precise value of our proposed measure ϕ_{iw}. Subset-based measures such as ϕ naturally overestimate stability while the weight-based ϕ_{pears} underestimates it. This phenomenon and the problems that arise from it are studied further in Section 5.3.

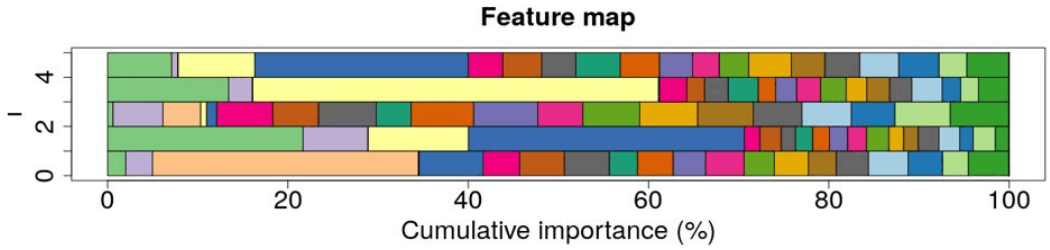


Figure 5: Feature stability map where feature importance is highly unstable. The subset stability ϕ does not account for this instability and is maximal ($\phi = 1$), while ϕ_{can} is the highest among the maps presented here ($\phi_{\text{can}} = 0.55$). Once again, ϕ_{pears} and ϕ_{iw} are able to assess this instability ($\phi_{\text{pears}} = 0.42$, $\phi_{\text{iw}} = 0.53$).

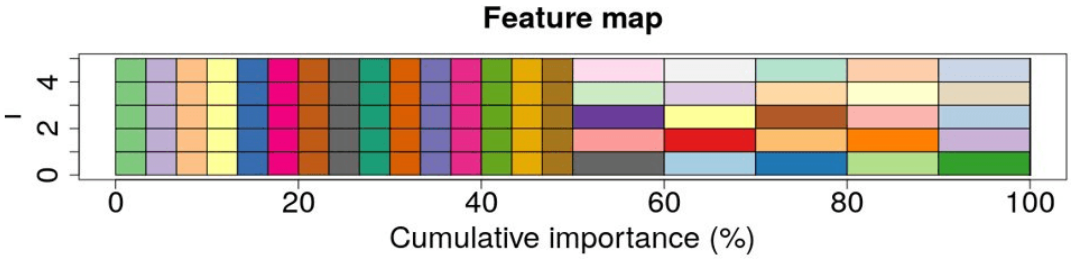


Figure 6: Feature stability map with half the importance space being perfectly stable while the other half is perfectly unstable. The measure ϕ overestimates stability as the number of stable features is high, while ϕ_{pears} underestimates it ($\phi = 0.75$, $\phi_{\text{pears}} = 0.25$, $\phi_{\text{can}} = 0.51$ and $\phi_{\text{iw}} = 0.5$).

5. An Importance Weighted Stability

In this section, we extend the Kuncheva index to handle a varying number of selected features and to incorporate feature importances. We pose

$$\phi_{\text{iw}} = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \frac{|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} - |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}}}{\bar{k} - C} \tag{6}$$

with

$$|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} = \sum_{f \in \mathcal{F}_i \cap \mathcal{F}_j} \min\left(I_{f,i}, I_{f,j}\right), \; |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}} = \frac{1}{d} \sum_{f \in \mathcal{F}_i, f' \in \mathcal{F}_j} \min(I_{f,i}, I_{f',j})$$

and

$$C = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}},$$

$I_{f,i}$ being the importance of the selected feature f in predictive model number i. In essence, the similarity between two selection runs is defined as the sum of the common importance

that selected features have between both decision models. The overall stability value is then the average of the pairwise similarities (normalized and corrected for chance). This can be visually estimated with the help of the feature stability maps, as it corresponds to the overlap of the same colors across rows. This new stability ϕ_{iw} corrects for the undesirable instability of Figures 4 and 5 as these overlaps are low in both cases. As previously stated, it is exactly equal to 0.5 in Figure 6 because the overlaps extend to exactly half of the feature stability map.[3] Such a stability requires an importance evaluation function: $\mathcal{I} : \{\mathcal{F}, \mathcal{P}\} \rightarrow \mathbb{R}$. This function is formally defined in Section 5.1. We normalize the feature importances in each selection run such that $\sum_f I_{f,i} = \bar{k}, \forall 1 \leq i \leq M$. As such a normalization would be undefined for a selection run i with $k_i = 0$, we pose $|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} = |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}} = 0$ if $k_i = 0 \oplus k_j = 0$ and $\bar{k}$ (the average number of selected features) if $k_i = k_j = 0$, with $\oplus$ the XOR operator.

The corrective term C in the definition of ϕ_{iw} (Equation 6) can be computed in $O(M^2\bar{k} + M\overline{k\log(k)})$ time by Algorithm 2 in Appendix C, with $\overline{k\log(k)} = \frac{1}{M}\sum_{i=1}^{M} k_i \log(k_i)$, assuming that the feature selection algorithm produces unsorted feature importances.[4] The overall time complexity to compute ϕ_{iw} from feature importance is also $O(M^2\bar{k} + M\overline{k\log(k)})$, as computing the pairwise intersections $|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}$ requires only $O(M^2\bar{k})$.

5.1 Evaluating Feature Importance

The evaluation of the importance of features in a predictive model is the root of embedded feature selection algorithms. For linear models, one can use the simple function

$$\mathcal{I}_{\text{lin}}(f, \mathbf{w}) = ||\mathbf{w}||_0 \times \frac{|w_f|}{||\mathbf{w}||_1} \tag{7}$$

where $\mathbf{w}$ represents the weight vector of the model. A linear RFE builds linear models and iteratively drops the features whose importance are the lowest according to Equation (7). For non-linear SVM models, one can still attribute an importance to each feature by computing how much this feature contributes to the margin of the SVM (Guyon et al., 2002). This can be done using Equation (8)

$$\mathcal{I}_{\text{svm}}(f, \boldsymbol{\alpha}) \propto |W^2(\boldsymbol{\alpha}) - W^2_{(-f)}(\boldsymbol{\alpha})|, \quad W^2_{-f}(\boldsymbol{\alpha}) = \sum_{k,l} \alpha_k \alpha_l y_k y_l (\mathbf{x}_k^{-f} \cdot \mathbf{x}_l^{-f}) \tag{8}$$

where $\mathbf{x}_k^{-f}$ denotes the training point k without the feature f, y_k is the label of training point k (± 1) and the α_k's are the solutions to the SVM dual problem. Similarly to the linear RFE algorithm, the non-linear SVM-RFE iteratively drops features with the lowest importance according to Equation (8). This process produces the same ranking as Equation (7) for linear SVMs (Guyon et al., 2002). For random forest classifiers, feature importance can be measured by randomly permuting the features in the out-of-bag samples and computing the predictive accuracy decrease these permutations cause (Breiman, 2001; Paul et al., 2013). For black-box classifiers (*e.g.* DNNs), and to provide a unified framework, we define the importance of a feature f in a predictive model p, or the sensitivity of model p to the

3. We suppose that d tends to ∞ in Figure 6 which here implies $C \rightarrow 0$.
4. Otherwise, the time complexity becomes $O(M^2\bar{k})$.

feature f as the inverse of the smallest noise applied to f necessary to flip the decision of model p, averaged over the n learning examples. Formally,

$$\mathcal{I}_n(f,p) \triangleq \frac{1}{n}\sum_{i=1}^{n}\frac{k_p \times \mathcal{I}(f,p,\mathbf{x_i})}{\sum_{f'\in\mathcal{F}_p}^{d}\mathcal{I}(f',p,\mathbf{x_i})}, \quad \mathcal{I}(f,p,\mathbf{x_i}) = \frac{\sigma_f}{\delta_{\mathbf{x_i},p,f}} \tag{9}$$

where k_p is the number of features used by model p, $\delta_{\mathbf{x_i},p,f}$ is the smallest additive change (in absolute value) required to feature f such that the decision of the predictive model p on example $\mathbf{x_i}$ changes, and σ_f the standard deviation of feature f. Intuitively, if one can change feature f by large amounts without perturbing the decisions of the model (here thought as a classifier), then f is not important in the decisions. On the contrary, if a small change to f causes a lot of decision switches, then the model is highly sensitive to it. Theorem 1 states that computing feature importance using Equation (7) or (9) for linear models is equivalent when the selected features are normalized to unit variance.

Theorem 1 *For a linear decision model p with weights $\mathbf{w}$, evaluated from n learning examples with k_p features normalized to unit variance,*

$$\mathcal{I}_{lin}(f,\mathbf{w}) = ||\mathbf{w}||_0 \times \frac{|w_f|}{||\mathbf{w}||_1} = \mathcal{I}_n(f,p).$$

Proof As features are normalized to unit variance, $\sigma_f = 1,\ \forall f$. The decision function of a linear model p with weights $\mathbf{w}$ can be written $D(\mathbf{w},\mathbf{x}) = \text{sign}(\sum_{f=1}^{k_p} w_f \times x_f + w_0)$. The smallest change $\delta_{\mathbf{x_i},p,f}$ to a feature f required to flip the decision of data point $\mathbf{x_i}$ is the change required such as to make $D(\mathbf{w},\mathbf{x_i}) = 0$. Then,

$$\delta_{\mathbf{x_i},p,f} = \frac{D(\mathbf{w},\mathbf{x_i})}{|w_f|} \Rightarrow \mathcal{I}(f,p,\mathbf{x_i}) = \frac{|w_f|}{D(\mathbf{w},\mathbf{x_i})} \Rightarrow \mathcal{I}_n(f,p) = \frac{1}{n}\sum_{i=1}^{n}\frac{k_p \times \frac{|w_f|}{D(\mathbf{w},\mathbf{x_i})}}{\sum_{f'\in\mathcal{F}_p}^{d}\frac{|w_{f'}|}{D(\mathbf{w},\mathbf{x_i})}}$$

$$\Rightarrow \mathcal{I}_n(f,p) = \frac{1}{n}\sum_{i=1}^{n} k_p \times \frac{|w_f|}{||\mathbf{w}||_1} = k_p \times \frac{|w_f|}{||\mathbf{w}||_1} \triangleq ||w_0|| \times \frac{|w_f|}{||\mathbf{w}||_1} \triangleq \mathcal{I}_{lin}(f,\mathbf{w}).$$

■

With feature importance defined by Equation (9), computing the importances for all M selection runs can be done in $O(M\bar{k})$ for linear models (according to Theorem 1) and in $O(M\bar{k}n)$ in the non-linear case. As stability estimation (with ϕ_{iw}) requires $O(M^2\bar{k} + M\overline{k\log(k)})$, the time complexity of the joint process of evaluating feature importance and computing stability is $O(M^2\bar{k} + M\overline{k\log(k)} + M\bar{k}n)$ in general and $O(M^2\bar{k} + M\overline{k\log(k)} + M\bar{k}) = O(M^2\bar{k} + M\overline{k\log(k)})$ when dealing with linear predictive models.

5.2 Properties

In this section, we show that our proposed stability measure ϕ_{iw} satisfy the following properties, adapted from (Nogueira et al., 2017a). Two new desirable properties for a stability measure are then defined in Sections 5.3 and 6.3, and proved for ϕ_{iw} in appendix.

- **Property 1** *Fully defined: the measure is defined for every possible importance combinations.*
- **Property 2** *Maximum stability* $\Leftrightarrow$ *deterministic importance*
- **Property 3** *Bounds: the measure is bounded by constants not dependent on the overall number of features d or on the average number of features selected $\bar{k}$.*
- **Property 4** *Correction for chance: the measure is constant in expectation (here set to 0) when features are selected randomly.*

Our proposed measure, ϕ_{iw}, is fully defined: it is defined everywhere except when no feature is ever selected, or every feature is always selected with an equal importance. In both cases, one can hardly speak of feature selection. The measure is maximal whenever the same feature subset is always selected and the importances of the selected features are constant across runs. It is lower bounded by $\frac{-1}{M-1}$ and upper bounded by 1. As the number of runs M is greater than or equal to 2, ϕ_{iw} is always bounded by -1 and 1 which is necessary for relevant comparisons, and is asymptotically lower bounded by 0 as M tends to ∞. The measure is also corrected for chance as its expected value is constant (here set to 0) whenever features are selected at random. These properties are proved in Appendix A.

Theorem 2, proved in Appendix B, shows more clearly the similitude between ϕ_{iw} and the Kuncheva index whenever the importance of all selected features is evenly distributed between them in any given run.

Theorem 2 *Whenever the importance of all selected features is evenly distributed between them in any given run,*

$$\phi_{iw} = \frac{\mu_M[\frac{\bar{k}|\mathcal{F}_i \cap \mathcal{F}_j|}{\max(k_i,k_j)}] - \frac{\bar{k}}{d}\mu_M[\frac{k_i k_j}{\max(k_i,k_j)}]}{\bar{k} - \frac{\bar{k}}{d}\mu_M[\frac{k_i k_j}{\max(k_i,k_j)}]}$$

with $\mu_M(g(i,j)) = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} g^*(i,j)$, $g^*(i,j) = 1$ *if* $k_i = k_j = 0$, $g(i,j)$ *otherwise.*

The correction for chance term of the Kuncheva index, $\frac{\bar{k}^2}{d}$, is extended here to $\frac{\bar{k}}{d}\mu_M[\frac{k_i k_j}{\max(k_i,k_j)}]$ to handle a varying number of selected features. Also, the subset intersection between selection runs i and j, $|\mathcal{F}_i \cap \mathcal{F}_j|$, is weighted by the term $\frac{\bar{k}}{\max(k_i,k_j)}$, such that selection runs with a high or low number of selected features influence the overall stability value by the same amount. We show in Section 5.3 that this property is convenient when dealing with groups of correlated variables.

Theorem 3, proved in Appendix B as well, further shows that whenever the number of selected features is constant across runs, our proposed measure degenerates into the Kuncheva index and thus, into ϕ.

Theorem 3 *Whenever the importance of all selected features is evenly distributed between them in any given run and the number of selected features is constant across the M runs,*

$$\phi_{iw} = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \frac{|\mathcal{F}_i \cap \mathcal{F}_j| - \frac{\bar{k}^2}{d}}{\bar{k} - \frac{\bar{k}^2}{d}},$$

which is the usual expression of the Kuncheva index.

Theorem 2 and Theorem 3 indicate that, when the predictive models approximately use their features equally, our measure ϕ_{iw} behaves in a similar manner to existing measures (as is validated in Section 7.2). However, we show in Sections 5.3 and 6.3 that current measures are vulnerable to undesirable behaviors in certain situations, while our measure is more robust in this regard.

We further show in Appendix B that the measure ϕ_{iw} can be re-stated in a frequency-based form, as

$$\phi_{\text{iw}} = 1 - \frac{\frac{M}{M-1}(\bar{k} - \sum_{f=1}^{d} I_f^* p_f^2)}{\bar{k} - C} = 1 - \frac{\sum_{f=1}^{d} I_f^* s_f^2 + \frac{M}{M-1}(\bar{k} - \sum_f I_f^* p_f)}{\bar{k} - C} \tag{10}$$

with a properly normalized global feature importance $I_f^* = \sum_{i,j=1}^{M} \frac{\min(I_{f,i}, I_{f,j})}{|z_{f,\cdot} \neq 0|^2}$, where $|z_{f,\cdot} \neq 0|$ is the number of runs where feature f is selected. This new formulation makes even more explicit the corrections for the instabilities of Figure 4 and 5: the selection variance s_f^2 of features with a high importance I_f^* accounts more in the overall stability such that the overestimation of stability, when frequently selected features are not used for prediction, is corrected (Figure 4). Furthermore, having features with highly varying importances in different selection runs is penalized by the term $(\bar{k} - \sum_f I_f^* p_f)$ which is equal to zero only when $I_{f,i} = I_f^*, \forall i$.

5.3 Stability in the Presence of Highly Correlated Feature Groups

In this section, we analyze the behavior of stability measures when the feature space is composed of groups of highly correlated features. We first review a recently proposed measure, which explicitly aims at dealing with feature correlations, and then study its behavior, along with ϕ, ϕ_{pears}, and ϕ_{iw} in the presence of correlated feature groups.

Sechidis et al. (2019) generalize the index ϕ such as to accurately measure selection stability in the presence of high correlation between variables. The idea behind the measure is that an algorithm that tends to select different features should not be considered unstable if these features are highly correlated to each other, as the *effective* extracted information is the same. To this goal, they define the *effective* similarity between two selection runs as the generalized inner product

$$|\mathcal{F}_i \cap \mathcal{F}_j|_{\mathcal{C}} = \mathbf{z}_i \mathcal{C} \mathbf{z}_j$$

where $z_{i,f}$ is the Bernouilli variable which is equal to 1 when feature f is selected in selection run i and where the elements $c_{f,f'} \geq 0$ of the matrix $\mathcal{C}$ represent the correlation between feature f and f'. The more correlated the selected variables in run i and j are to each other, the bigger the similarity between these runs. They then proved the following result:

$$\frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\mathcal{C}} = \bar{k}_{\mathcal{C}} - tr(\mathcal{C}S) \tag{11}$$

where $\bar{k}_{\mathcal{C}} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{z}_i \mathcal{C} \mathbf{z}_i$ and with S the variance-covariance matrix of $\mathcal{Z}$, the matrix with the elements $z_{i,f}$. Equation (11) is analogous to Equation (3) when $\mathcal{C}$ is the identity matrix. The following frequency-based measure can then be derived

$$\phi_C = 1 - \frac{tr(\mathcal{C}S)}{tr(\mathcal{C} \sum^0)} \tag{12}$$

with $\sum^0$ the matrix that normalizes the measure. For further details, we refer the reader to (Sechidis et al., 2019). Some other stability measures designed to correctly handle high correlated variables are proposed in (Yu et al., 2008; Zhang et al., 2009a).

While our measure ϕ_{iw} does not take directly feature correlations into account, we illustrate here, using experiments on simulated data, that evaluating ϕ_{iw} can be beneficial when the feature space is composed of correlated groups. We use an artificially generated data set with $N = 5$ groups of variables. Each group contains c features that are highly correlated to each other (average correlation of $\rho^g \gg 0$). In addition to these feature groups, the data set contains $l = 1000$ variables. Feature values are sampled from two multivariate normal distributions using the `mvrnorm` R package. Positive examples ($n_+ = 100$) are sampled from a first distribution, centered on $\boldsymbol{\mu}_+$, a vector with $\mu_{+,f} = \mu_+^g$ if feature f belongs to one of the $N = 5$ correlated groups, $\mu_+^{\neg g}$ otherwise. Negative examples ($n_- = 100$) are sampled from a second distribution, centered on $\boldsymbol{\mu}_- = -\boldsymbol{\mu}_+$. Both distributions have unit variance. We consider three scenarios with different values of μ_+^g, $\mu_+^{\neg g}$ and ρ^g, specified in Table 1. In all scenarios, features inside a correlated group are very relevant to the binary prediction task, while features outside such groups are less but still marginally relevant. For feature selection, we use the group LASSO (scenarios 1 and 2), configured such that it selects all features inside a group or none of them, and the standard LASSO (scenario 3). We set the regularization parameter λ of the LASSO and group LASSO such as to select approximately 40 features when the size c of the correlated groups is equal to 1. The $N = 5$ feature groups are thus expected to be selected in most of the $M = 30$ selection runs while the selection of the additional features should be unstable. The experiment is repeated 10 times using different generative seeds for the data sets and the mean stability values are reported in Figure 8 as a function of c, the size of the correlated groups.

We first study Figure 7 which represents the cumulative importance of the features that are selected by the group LASSO in scenario 1 when $c = 1$ (top) and $c = 10$ (bottom). Clearly, the group LASSO gives more importance to the features of the 5 groups when $c = 1$ as they are more relevant (by design). When $c = 10$ however, the importance of the features inside the correlated groups is reduced, such that the cumulative importance of each group is approximately the same as in the $c = 1$ case. In the same spirit as in (Sechidis et al., 2019) where the authors argue that the alternate selection of highly correlated features should not influence stability, we argue that the size of the correlated groups c should not drastically influence stability either (whenever the group importance is independent of c), as the *effective* extracted information is unchanged.

Scenario	μ_+^g	$\mu_+^{\neg g}$	ρ^g	method
1	0.35	0.05	0.8	group LASSO
2	0.5	0.05	0.8	group LASSO
3	0.5	0.05	0.95	LASSO

Table 1: Experimental settings for the three studied scenarios. The relevance of features inside one of the $N = 5$ correlated groups is related to μ_+^g while the relevance of features outside any group ($\sim \mu_+^{\neg g}$) is constant across scenarios. The average intra-group correlation is ρ^g and inter-group correlation is negligible. In scenarios 2 and 3, features inside correlated groups are very relevant and are selected in (nearly) all runs.

Figure 8 compares ϕ, ϕ_{iw}, ϕ_{pears} (with the feature weights $w_{f,i}$ as the importances $I_{f,i}$) and ϕ_C (with the correlation matrix $\mathcal{C}$ such that $c_{f,f'} = 1$ iff features f and f' belong to the same group, which is consistent with the authors proposal of thresholding the true correlation) when the sizes of the correlated groups c vary.

When the group LASSO is used (Figures 8a and 8b), the standard stability ϕ increases when the number of correlated variables inside each group grows, which is undesirable. Increasing the number of correlated variables increases ϕ because their small selection variance is counted more than once. With c_g, the size of the correlated group g, p_g its selection frequency and $s_g^2 = p_g(1 - p_g)$, its selection variance (we assume here $M \to \infty$ to simplify calculations),

$$\phi = 1 - \frac{\sum_g \sum_{f \in g} s_g^2}{\sum_g p_g c_g} \triangleq 1 - \frac{\sum_g c_g \times s_g^2}{\sum_g p_g c_g}. \tag{13}$$

In Equation (13), a variable outside any correlated group is considered as being in a group of size 1. In the above scenarios, $c_g = c$ for the 5 groups and $c_g = 1$ for all the other variables. The contribution of a group to the variance term of ϕ is proportional to its size, as the importance reduction of the features inside the groups is not taken into account. This behavior is more pronounced in Figure 8b (scenario 2) than in Figure 8a (scenario 1), as the variables inside correlated groups are more relevant, causing the selection variance of the correlated groups to be even smaller.

The Pearson's correlation measure ϕ_{pears} exhibits another behavior: it starts higher than the other measures and gradually decreases. When the total number of features d tends to ∞, the correlation $\rho_{i,j}$ between feature importances of two runs i and j satisfies

$$\rho_{i,j} = \frac{\sum_f I_{f,i} I_{f,j}}{\sqrt{\sum_{f=1}^d I_{f,i}^2} * \sqrt{\sum_{f=1}^d I_{f,j}^2}}. \tag{14}$$

The contribution of a feature in the numerator of $\rho_{i,j}$ increases quadratically with its importance. When $c = 1$, it is dominated by the stability of the $N = 5$ feature groups, as their importance is the highest. Then, as c increases, the importance of each feature inside the groups is cut by c (as illustrated by Figure 7), meaning that the sum of squared of the importance inside each group $\sum_{f \in g} I_{f,i}^2$ decreases by a factor c as well. This implies an

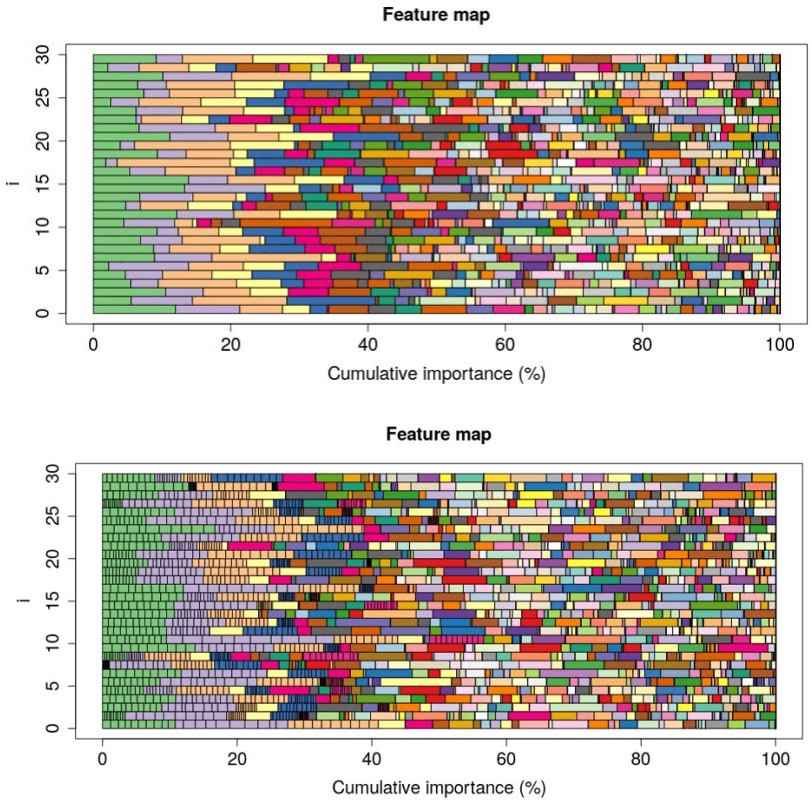
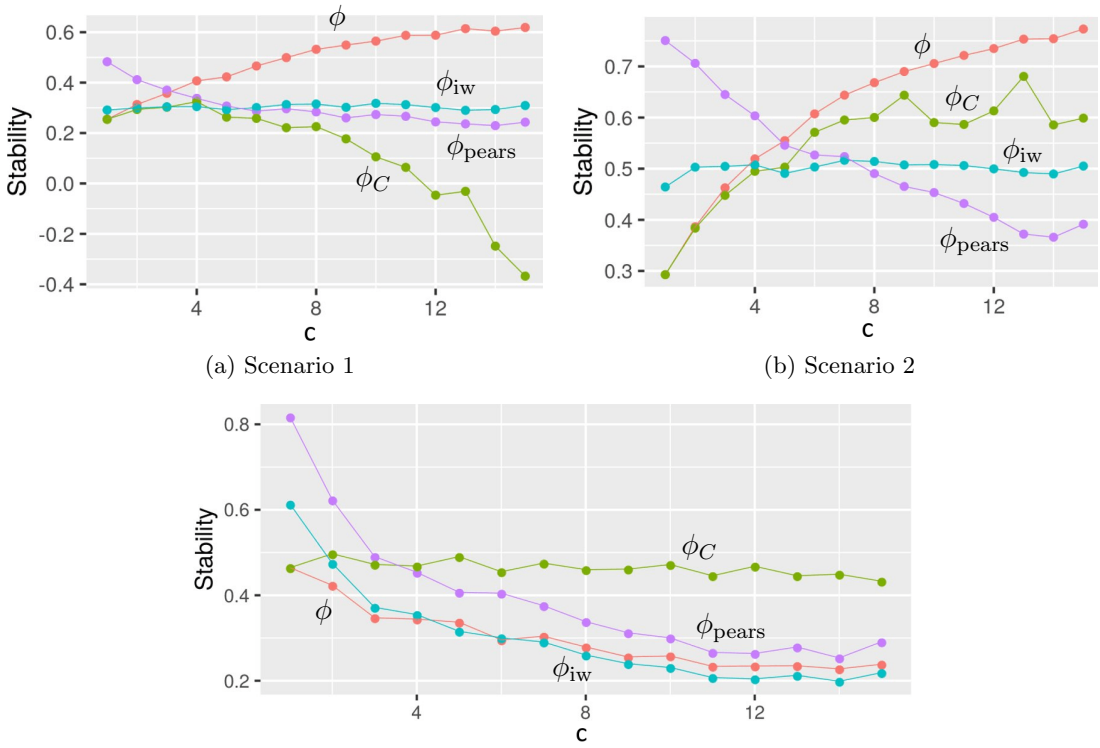


Figure 7: Feature stability maps of the group LASSO (scenario 1) when the size of the correlated groups c is equal to 1 (top) and 10 (bottom). The group LASSO regularization is chosen to select ≈ 40 features on average when $c = 1$, but the specific number of selected features here varies across runs. As the cumulative importance of each group is approximately constant in both feature stability maps, their stability should be similar.

unintuitive result: the contribution of a correlated group to ϕ_{pears} is inversely proportional to its size. As a consequence, when c grows, stability is more and more dominated by the out of group features, which have a larger selection variance. As was the case for ϕ, this behavior is accentuated in Figure 8b (scenario 2), where the importance of correlated groups is larger (as is their relevance).

These results can be compared to the behavior of ϕ_{iw}, where features contribute to the stability proportionally to their importance. Hence, as the sum of the importance of each group remains constant with respect to its size, so is their contribution to stability. As a consequence, ϕ_{iw} remains approximately constant with c. Property 5 formalizes this result.

Property 5 *Group size independence: whenever perfectly correlated feature groups are selected as a whole, stability depends only on the groups' cumulative importance, not on the group sizes.*



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

Figure 8: Experimental comparison of the stability measures ϕ, ϕ_C, ϕ_{pears} and ϕ_{iw} in the presence of highly correlated feature groups, in function of c, the size of such groups. The group LASSO is used for feature selection in (a)(scenario 1) and (b)(scenario 2), the LASSO in (c)(scenario 3). Given the design of these experiments, the stability value should not depend on c.

As detailed in Appendix A, Property 5 strictly holds for ϕ_{iw} if the global importance of each group is distributed among its features deterministically across runs. Unlike the other measures, ϕ_{iw} correctly assesses the relative importance of each group in the global stability value but may underestimate the within-group stability if the above assumption is violated.[5]

The measure ϕ_C stands out from the others by displaying different behaviors in Figure 8a (scenario 1) and 8b (scenario 2). When the correlated groups are almost always selected (Figure 8b), ϕ_C behaves like ϕ, *i.e.* it increases with c. However, when correlated groups have non-negligeable selection variance, (Figure 8a), ϕ_C first starts to improve as before, but as c continues to increase, it turns out that ϕ_C tends to $-\infty$. Nonetheless, ϕ_C is the only measure studied here where different features inside a correlated group are considered equivalent. When the LASSO is used instead of the group LASSO, ϕ_C remains approximately constant, as the LASSO generally selects a single, or a few arbitrary features inside each group. This is illustrated in Figure 8c (scenario 3). The other measures ϕ, ϕ_{iw} and ϕ_{pears}

5. As shown in Figure 7, it is respected with the group LASSO. The global importance of each group is approximately evenly distributed among their features.

all decrease when c increases as the selection of the few selected features inside each group becomes more and more unstable.

6. Stability Optimization

In this section, we study a second scenario that illustrates the limits of subset-based stability measures: the joint optimization of stability and predictive accuracy. Firstly, we review and extend a recently proposed approach for joint optimization. Then, we demonstrate that optimizing subset-based measures, such as ϕ, can sometimes lead to situations with poor interpretability. Next, we argue why considering ϕ_{iw} is more adequate in this context.

6.1 Hybrid-RFE

Hamer and Dupont (2020) optimize the selection stability ϕ jointly with the predictive accuracy in a bi-objective framework. Pareto-optimal trajectories are derived, from which domain experts can choose a particular compromise based on their personal preferences. The trajectories are obtained by pre-selecting some features based on a stable univariate criterion, before running the multivariate recursive feature elimination (RFE) algorithm which then selects the most appropriate additional features.

This methodology is summarized in Algorithm 1. Firstly, a set of *stable features*, S_N, is found as the top-N features based on a univariate criterion (lines 3,4). Univariate filters tend to be more stable than multivariate methods as they do not take feature interdependencies into account. These N features are then forced to be selected at each iteration of the RFE, which selects, in a multivariate fashion, the most appropriate additional features. It does so by iteratively minimizing the logistic loss (line 7), ranking every feature (but the ones of the stable set) based on the absolute value of their weight $\mathbf{w}$ in the learned decision function (line 8), and dropping the one feature with minimal weight (line 9), until the desired number of features k is reached.[6] Finally, it learns the final decision function by minimizing the logistic loss on the k selected features (line 10), possibly with a different regularization constant λ_f. The difference between this approach and the classic RFE is that the features in S_N are never dropped and are thus always present in the final model. To take advantage of this knowledge, one can apply differential shrinkage on these features to increase their importance in the multivariate selection (line 7, with $\odot$ the element-wise product). The intensity of this differential shrinkage is dictated by the meta-parameter $\epsilon \leq 1$ used in line 5.

If the set of *stable features*, S_N, is robust, then increasing N, the number of features selected beforehand, is expected to increase the overall selection stability at the cost of a possible decrease in predictive accuracy. If $N = 0$, this hybrid-RFE is equivalent to the classical RFE, for which no feature is pre-selected, except that the logistic loss is considered here instead of the default hinge loss. This logistic loss choice is motivated by the stability gains it offers, as studied in Appendix E. When $N = k$, the approach becomes equivalent to a purely univariate filter.

6. For computational reasons, it is common to drop a fraction of the remaining features instead of a single one at each iteration. We opt here to drop 20% of the remaining features at each pruning step.

Algorithm 1 Hybrid RFE.

procedure SelectFeatures($N, \lambda, \epsilon, \lambda_f$)
  $\mathcal{F} \leftarrow$ the set of all features
  $r_f \leftarrow$ univariate criterion rank of each feature (descending order)
  $S_N \leftarrow \{f : r_f \leq N\}$
  $\beta_f \leftarrow \epsilon$ if $f \in S_N, 1$ otherwise
  while $|\mathcal{F}| > k$ **do**
    $\mathbf{w}^* \leftarrow \text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + \exp^{-y_i(\mathbf{w}\mathbf{x_i})}) + \lambda||\boldsymbol{\beta} \odot \mathbf{w}||_2$
    $\mathbf{r}^* \leftarrow$ rank features $\{f \in \mathcal{F} \setminus S_N\}$ on $|w_f^*|$ in descending order
    $\mathcal{F} \leftarrow \mathcal{F} \setminus \{f : r_f^* = |\mathcal{F}| - N\}$
  $\mathbf{w}^* \leftarrow \text{argmin}_{\mathbf{w}} \sum_{i=1}^{n} \log(1 + \exp^{-y_i(\mathbf{w}\mathbf{x_i})}) + \lambda_f||\mathbf{w}||_2$
  return $(\mathcal{F}, \mathbf{w}^*)$

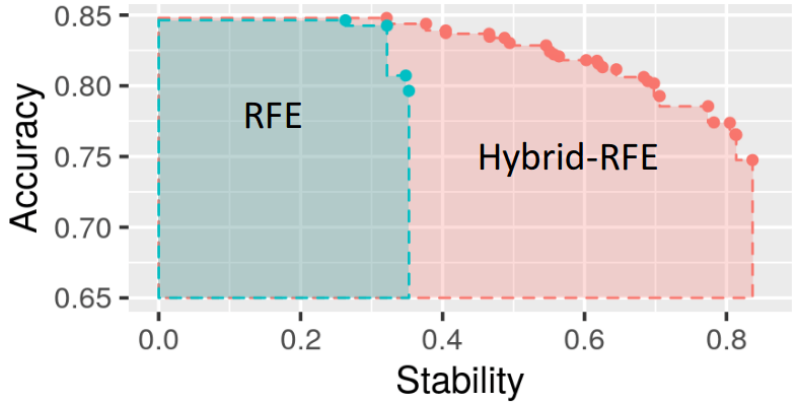


Figure 9: Typical Pareto-optimal curves of the RFE (blue) and the hybrid-RFE (red). Far better (accuracy, stability) trade-offs are reachable with the hybrid-RFE.

We use a linear combination of the supervised Welch's t-test ratio (Welch, 1947) and the unsupervised sample variance as the univariate criterion. Figure 9 depicts a typical result of the hybrid-RFE approach on a micro-array data set with $k = 20$. For more details, we refer the reader to our previous work (Hamer and Dupont, 2020).

The plot represents the areas dominated by the Pareto-optimal curves that can be drawn by model selection on the regularization parameters λ and λ_f of the standard logistic RFE (blue) and by the hybrid-RFE (red). The hybrid-RFE is able to increase the selection stability by considerable amounts and dominates simple model selection. The original paper further shows that the hybrid-RFE is mostly sensitive to the stability of the stable set S_N, and less to its predictive accuracy.

6.2 An Extension to the Hybrid-RFE

We propose an extension to the hybrid-RFE which aims at increasing stability by modulating the importance of the selected features. Indeed, since ϕ_{iw} depends on the M final decision

models, one can increase it independently of the selection process itself. In this section, we estimate which selected features are likely to be frequently selected across the M resampling runs and apply differential shrinkage on them to increase their importance in the decision models. From the left-hand side of Equation (10), repeated here for convenience,

$$\phi_{\text{iw}} = 1 - \frac{\frac{M}{M-1}(\bar{k} - \sum_{f=1}^{d} I_f^* p_f^2)}{\bar{k} - C},$$

it follows that increasing I_f^* of frequently selected features (features with a high p_f) increases stability. In order to increase $I_f^* = \frac{1}{M^2}\sum_{i,j}^{M} \min(I_{f,i}, I_{f,j})$, one must increase the importance $I_{f,i}$ of feature f in multiple selection runs jointly.

As RFE drops features iteratively, it is possible to measure, for each feature, how close they are to the elimination during the selection process. We define an overall frequency score, which estimates the true selection frequency of the selected features

$$\text{sc}_f = \frac{|\mathcal{F}_0|}{|\mathcal{F}_\#|} \prod_{s=1}^{\#\text{ pruning steps}} \frac{|\mathcal{F}_s| - r_f(s) + 1}{|\mathcal{F}_{s-1}| - r_f(s) + 1} \tag{15}$$

with $\mathcal{F}_s$ the set of selected features after s RFE pruning steps, and $r_f(s)$ the rank of feature f at step s (RFE ranks features at each iteration and drops a fraction of the least relevant ones). This score has the convenient properties of being bounded by 0 and 1, and to be independent of the number of pruning steps. Indeed, assuming that a feature f has a constant ranking $r_f(s) = r_f \leq k$, then

$$\text{sc}_f = \frac{|\mathcal{F}_0|}{|\mathcal{F}_\#|}\Big(\frac{|\mathcal{F}_\#| - r_f + 1}{|\mathcal{F}_0| - r_f + 1}\Big)$$

which does not depend on the number of steps.

Figure 10 depicts the correlation between the selection frequency p_f of feature f, across $M = 100$ selection runs, and the average frequency score sc_f given by Equation (15), on two typical micro-array data sets (`singh` (left) and `chiaretti` (right) which are introduced in Section 7.1). The average frequency score sc_f is computed only over the runs for which f is selected (the score is not defined for the other runs). Features which are selected more often tend to have higher frequency scores (the (Pearson) correlations between both variables are 0.68 and 0.54 for `singh` and `chiaretti`, respectively). We then apply the following regularization in order to reduce the importance of the selected features with a low frequency score sc_f:

$$\mathcal{R} = \lambda_f * (1 + \alpha * (1 - \text{sc}_f))||\mathbf{w}||_2 \tag{16}$$

with α a meta-parameter determining the amplitude of the differential shrinkage. Features with a low frequency score are more regularized which is expected to decrease their weight and thus their importance in the linear predictive models.

6.3 Theoretical Analysis

The hybrid-RFE, introduced in Section 6.1, combines the selection of N features which are first chosen based on a univariate criterion, and the selection of $k - N$ features which are

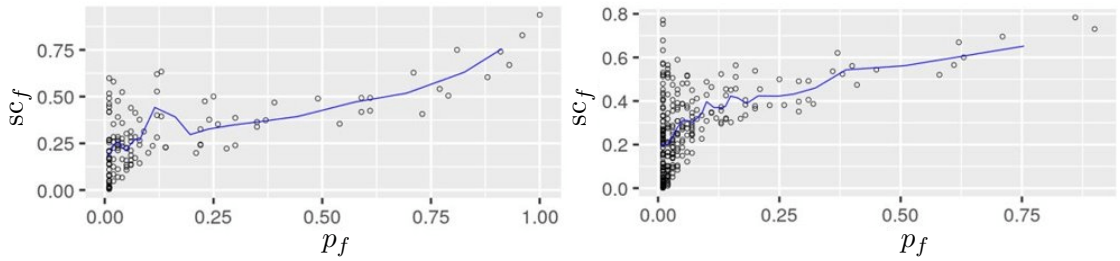


Figure 10: Correlation between the selection frequency p_f of feature f with its average frequency score sc_f on `singh` (left) and `chiaretti` (right).

then found, in a multivariate fashion, by Algorithm 1. In this section, we generalize this idea and provide a theoretical analysis of the evolution of the stability when Q selection methods are combined ($Q = 2$ for the hybrid-RFE). From this analysis, it follows that, unlike the other measures, the stability ϕ_{iw} is particularly suitable to be optimized along with predictive accuracy.

Consider the following scenario. For each selection run m, we learn Q independent linear models on non-overlapping selected feature sets $\mathcal{F}_q(m)$: $\sum_{f\in\mathcal{F}_q(m)} w_f(m)\times x_f \geq t_q(m)$, with the normalization $\sum_{f\in\mathcal{F}_q(m)} w_f(m) = \bar{k}_q$. Each of the Q models is found by a given selection method. The overall prediction model is defined as a fixed linear combination of the Q linear models: $\sum_q \delta I_q \sum_{f\in\mathcal{F}_q(m)} w_f(m)\times x_f \geq t(m)$. The parameter δI_q modulates the importance of the selected features from particular selection methods. In this scenario, Property 6 states that it should be possible to decompose stability in multiple terms, respectively capturing the stability of each selection method, noted here ϕ^q_{iw}. These terms are weighted by $\frac{\bar{k}_q}{\bar{k}}\delta I_q$, such that a selection method q has more weight in the overall stability when it is responsible for a large fraction of the selected features and when these features are important in the combined predictive model. This property is proved for ϕ_{iw} in Appendix A under the assumption that d tends to ∞.

Property 6 *Importance weighted decomposition: when combining the non-overlapping selected feature sets of Q different methods, which produce Q models, in a single predictive model, if features selected by method q have their importance multiplied by δI_q in the combined predictive model of each selection run, stability can be expressed as a weighted sum of the Q prior stabilities:*

$$\phi_{\mathrm{iw}} = \sum_{q=1}^{Q} \frac{\bar{k}_q}{\bar{k}} \delta I_q \phi^q_{\mathrm{iw}}$$

with ϕ^q_{iw} the stability of method q alone, and δI_q the factor such that $I_{f,i} = \delta I_q \times I^q_{f,i}$.

As far as the hybrid-RFE is concerned, whenever the N features of the stable set are less relevant, combining both selections reduces their importance in the predictive models and increases the importance of the $k - N$ multivariate features ($\delta I_{q=\mathrm{univariate}} < 1$ and $\delta I_{q=\mathrm{RFE}} > 1$). This gives more weight to $\phi^{q=\mathrm{RFE}}_{\mathrm{iw}}$ in Property 6, thus limiting the stability increase provided by the forced univariate selection of the N features.

Hamer and Dupont (2020) optimize jointly the measure ϕ with the predictive accuracy and show that the quality of the reachable compromises is highly dependent on the stability of the stable set, rather than on its predictive performance. In general, better compromises are obtained when the N pre-selected features are stable, even if not relevant to the prediction task. This is caused by the fact that, unlike ϕ_{iw}, the selection of stable, yet marginally important features increases ϕ. Indeed, a similar result to Property 6 can be derived for ϕ when d tends to ∞:

$$\phi = \sum_{q=1}^{Q} \frac{\bar{k}_q}{\bar{k}} \phi^q, \tag{17}$$

which does not depend on δI_q. As a consequence, a good (accuracy, ϕ) compromise is not necessarily meaningful, as ϕ could have been artificially increased by the selection of stable features which marginally take part in the predictive models. Subset measures, in general, can not be optimized soundly with predictive accuracy, as is shown experimentally in Section 7.1.

The measure ϕ_{pears} can not be decomposed in multiple terms under the scenario described by Property 6. Still, if we assume that the q methods distribute importance evenly among their respective selected features (*i.e.* $I^q_{f,i} = 1,\ \forall f, i, q$), the following decomposition holds when d tends to ∞:

$$\phi_{\text{pears}} = \frac{\sum_{q=1}^{Q} \delta I_q^2 \bar{k}_q \phi^q_{\text{pears}}}{\sum_{q=1}^{Q} \delta I_q^2 \bar{k}_q}. \tag{18}$$

In this specific case, the relative contributions of the prior stabilities ϕ^q_{pears} to the combined stability ϕ_{pears} are proportional to δI_q^2. Due to this δI_q^2 factor, the stability ϕ_{pears}, like ϕ_{iw}, can not be increased by the selection of stable, yet marginally important features. However, as shown in Section 5.3 and on the illustrative example below, this quadratic dependency can have some negative consequences.

Consider the following example with $Q = 2$. Both selection methods distribute importance evenly among their respective selected features. The first method always selects the same 15 features and is thus perfectly stable: $\phi^1 = \phi^1_{\text{pears}} = \phi^1_{\text{iw}} = 1$. The second method always select 5 features but these 5 features never overlap across the M selection runs. Method 2 is perfectly unstable: $\phi^2 = \phi^2_{\text{pears}} = \phi^2_{\text{iw}} = 0$, when d tends to ∞. Assume that we combine these two methods and obtain the feature stability map depicted in Figure 6. This map is obtained with $\delta I_1 = \frac{2}{3}$ and $\delta I_2 = 2$. In this scenario, according to Equation (17), $\phi = \frac{15}{20} \times \phi^1 + \frac{5}{20} \times \phi^2 = 0.75$. According to Equation (18), $\phi_{\text{pears}} = \frac{1}{4} \times \phi^1_{\text{pears}} + \frac{3}{4} \times \phi^2_{\text{pears}} = 0.25$. Finally, according to Property 6, $\phi_{\text{iw}} = \frac{1}{2} \times \phi^1_{\text{iw}} + \frac{1}{2} \times \phi^2_{\text{iw}} = 0.5$, which is the preferable stability value for such a feature stability map. Table 2 summarizes the different desirable properties of the considered stabilities.

7. Experiments

In this experimental section, we study the behavior of the stability measures ϕ, ϕ_{pears} and ϕ_{iw} in the context of joint optimization with predictive accuracy (Section 7.1). We evaluate the stability of classical feature selection approaches according to these measures in Section 7.2 before briefly comparing their sampling distributions in Section 7.3. In this

Property	ϕ		ϕ_{pears}	
Fully defined	✓		✓	
Upper bound	✓	≤ 1	✓	≤ 1
Lower bound	✓	$\geq \frac{-1}{M-1}$	✓	$\geq \frac{-1}{M-1}$
Maximum	✓	$\Leftrightarrow$ Det. sel.	✗	$\Leftrightarrow$ Linear dep.
Corrected for H_0	✓		✓	
Group size independence	✗	The contribution of a group to ϕ is prop. to its size	✗	The contribution of a group to ϕ_{pears} decreases with its size
Interchangeable correlated features	✗		✗	
Importance weighted decomposition	✗	**Non**-weighted decomposition	✗	No general decomposition

Property	ϕ_{iw}		ϕ_C	
Fully defined	✓		✓	
Upper bound	✓	≤ 1	✗	
Lower bound	✓	$\geq \frac{-1}{M-1}$	✗	
Maximum	✓	$\Leftrightarrow$ Det. imp.	✗	
Corrected for H_0	✓		✓	
Group size independence	✓	The contribution of a group to ϕ is independent of its size	✗	Complex
Interchangeable correlated features	✗		✓	
Importance weighted decomposition	✓		✗	No general decomposition

Table 2: Summary of the properties verified by the stability measures under study. The measure ϕ cannot grasp nuances brought by feature importance and do not take variable correlation into account. The measure ϕ_C extends ϕ to deal with feature correlations at the expense of the bounds and maximum property. The correlation measure ϕ_{pears} takes feature importance into account in a way that undesirable behaviors can occur (notably in the presence of large correlated groups of variables). Furthermore, it does not satisfy the maximum property as only a perfect linear dependency is sufficient to make the measure maximal. Our proposed measure satisfy the *group size independence* and *importance weighted decomposition* properties which makes it more robust in different scenarios. Still, ϕ_{iw} lacks the ability to consider highly correlated features as interchangeable (the alternate selection of highly correlated features creates instability).

experimental part, we focus on the impact of feature importance on stability rather than on the effect of feature correlation. We do not evaluate ϕ_C here for this reason, and because it is not bounded which makes any comparison difficult.

7.1 Case Study: Decision Making for Cancer Diagnosis

Let us first study the compromises between accuracy and stability which are achievable with the hybrid-RFE on five micro-array and one mass spectrometric data sets. We show here that decision-making is heavily influenced by the choice of the stability measure. In particular, we show that the Pareto-optimal front is different for each measure and that optimizing ϕ or ϕ_{pears} can lead to unsatisfactory feature stability maps. Furthermore, optimizing both of these measures usually gives a false sense of stability which also hinders appropriate decision-making. The studied data sets are summarized in Table 3. They all have a small n (number of samples) to d (number of features) ratio, which generally causes feature selection methods to be particularly unstable. The learning task consists in predicting whether or not a patient is suffering from the corresponding disease. As is often done when dealing with high dimensional data sets, the feature space is first pre-filtered by removing the features with lowest variance (except for `alon` and `gravier`, for which such a pre-filtering has already been performed). The amount of pre-filtering is found such as to maximize the predictive performance of the classical RFE ($N = 0$) and is kept constant for all experiments. To measure the accuracy and stability obtained with a given set of meta-parameters, we use the classical bootstrap protocol which draws with replacement M samples of the same size as the original data set. Each model is evaluated on the out-of-bag examples and the mean classification accuracy is reported. The selection stability is evaluated using Equation (1)(ϕ), (4)(ϕ_{pears}) and (6)(ϕ_{iw}), over the $M = 100$ resamplings.

We perform experiments using the hybrid-RFE with the additional meta-parameter α, introduced in Section 6.2. The N pre-selected features are ranked according to the considered univariate criterion and are put, in that order, on top of the RFE ranking at each iteration, such that their frequency score sc_f given by Equation (15) is the highest. Increasing α is thus expected to increase the importance of these N pre-selected features in the predictive models, as they are less regularized (Equation 16). To limit our analysis to a subset of all Pareto-optimal points, we assume here that a domain expert aims at maximizing the objective function $o(A, S) = \gamma A + (1 - \gamma)\phi_{\_}$, with A the accuracy, $\phi_{\_}$ a particular stability, and $0 \leq \gamma \leq 1$ a parameter representing the domain expert's affinity towards accuracy versus stability. Intuitively, a given γ value implies a willingness to sacrifice a point in accuracy if stability can be increased by more than $\frac{\gamma}{1-\gamma}$. Such an objective function restricts our analysis to the convex hull of the Pareto-optimal curve.

We study in Figure 11 the achievable compromises in the (accuracy, stability) objective space using the hybrid-RFE on the `singh` data set. The plain points represent all Pareto-optimal trade-offs when ϕ (red), ϕ_{pears} (green) and ϕ_{iw} (blue) are used to assess stability. They are obtained with the hybrid-RFE with different sets of meta-parameters (N ,λ ,ϵ and λ_f, see Algorithm 1). These plain points are the same across the three subfigures. The three stabilities start roughly equal at the left of their respective Pareto front, and our measure ϕ_{iw} can be increased much less than ϕ and ϕ_{pearson}. The three measures are provably equivalent whenever the importance of the selected features are always 1. A strong difference among

name	data	year	n	d	disease	d after fil.
`alon`	micro-array	1999	62	2000	colon cancer	2000
`borovecki`	micro-array	2005	31	22283	Huntington's	1000
`singh`	micro-array	2002	102	12600	prostate cancer	1250
`gravier`	micro-array	2010	168	2905	breast cancer	2905
`chiaretti`	micro-array	2004	111	12625	leukemia	5000
`arcene`	mass-spectra	2003	198	10000	ovarian/prostate cancer	5000

Table 3: Information on used data sets, from the UCI machine learning repository (`arcene`) and from the `datamicroarray` R package for the others.

the values taken by ϕ, ϕ_{pears} and ϕ_{iw} suggests that the selection moves away from this base scenario. We illustrate this phenomenon in the rest of this section. Most importantly, we show that optimizing ϕ or ϕ_{pears} respectively leads to situations where $\phi \gg \phi_{\text{iw}}$ and $\phi_{\text{pears}} \gg \phi_{\text{iw}}$, with poorly interpretable feature stability maps.

To this goal, consider the circled points in Figure 11. They are the compromises (A, ϕ) (red), (A, ϕ_{pears}) (green) and (A, ϕ_{iw}) (blue) maximizing $\gamma A + (1-\gamma)\phi_{\text{-}}$ for some $0 \leq \gamma \leq 1$ with $\phi_{\text{-}} = \phi$ (Figure 11a), $\phi_{\text{-}} = \phi_{\text{pears}}$ (Figure 11b) and $\phi_{\text{-}} = \phi_{\text{iw}}$ (Figure 11c). For instance, the blue circled points in Figure 11a are the compromises in the (A, ϕ_{iw}) objective space that correspond to the convex hull of the (A, ϕ) Pareto-optimal curve. These blue circled points depart from the (A, ϕ_{iw}) Pareto-optimal curve which consists of the plain blue points. This implies that the Pareto-optimal curves of ϕ and ϕ_{iw} are obtained with different meta-parameters choices, otherwise the two blue curves would coincide in Figure 11a. Figure 11a also illustrates that increasing ϕ is not guaranteed to increase ϕ_{iw} or ϕ_{pears}, as the blue and green circled curves sometimes move back towards lower stability values. The feature stability maps annotated to Figure 11a clearly show that optimizing ϕ tends to reduce the importance of frequently selected features in the predictive models. In other words, ϕ is best increased here by the selection of stable features, yet marginally used for prediction. Figure 11b shows that maximizing ϕ_{pears} is not guaranteed to increase ϕ or ϕ_{iw}. Optimizing ϕ_{pears} tends to give large, yet highly varying importances to frequently selected features, which also hinders sound domain analysis. Finally, optimizing ϕ_{iw} provides a much nicer feature stability map where features have reasonably constant importance across runs. This stability map has been obtained by forcing the selection of the 5 features with highest sample variance and by applying a small differential shrinkage based on the frequency score sc_f ($\alpha = 0.1$).

The choice of the stability measure also influences the predictive performance of the chosen compromise. On `singh`, for $0.4 \leq \gamma \leq 0.6$, the chosen compromises correspond to the feature stability maps to the right of each subfigure. Using ϕ as stability measure gives the illusion of increasing stability by large amounts, thus more accuracy is sacrificed (here $A \approx 0.889$ for ϕ (Figure 11a), $A \approx 0.930$ for ϕ_{pears} (Figure 11b) and $A \approx 0.934$ for ϕ_{iw} (Figure 11c)). Optimizing ϕ, and ϕ_{pears} (to a lesser extent here), leads to unsatisfactory stability maps with lower predictive performance.

Analogous results to the ones presented in Figure 11 for the other data sets are presented in Appendix D. On most data sets, the Pareto-optimal compromises depend on the choice

of stability. Even when this is not the case, the choice of measure strongly affects the willingness of sacrificing accuracy which ultimately leads to different chosen Pareto-optimal compromises.

7.2 Stability of Standard Feature Selection Methods

The hybrid-RFE algorithm (Algorithm 1 in Section 6.1) is designed to navigate through the (accuracy, stability) objective space. In Section 7.1, we have shown that using ϕ_{iw} as the stability measure improves decision-making in such a context. To broaden our analysis, we study in this section the stability of common feature selection methods: logistic regression with the LASSO or ELASTIC NET penalty, random forests, the RELIEF algorithm, and the standard SVM or logistic RFE, which are not designed to explore such a bi-objective space. Results show that our proposed measure ϕ_{iw} behaves similarly to ϕ and ϕ_{pears} (but still provides some additional insights). This indicates that ϕ_{iw} keeps the correct behavior of well-known measures in standard cases (while being more robust to extreme situations, as demonstrated in Sections 5.3 and 7.1).

We use additional data sets which are briefly described in Table 4. Biomedical data sets (from Table 3) are now pruned to 5000 features. We aim at selecting $\min(20, \sqrt{d})$ features, while we set M to 30. We study each selection method independently (aggregated results are provided in Appendix E).

name	year	n	d	name	year	n	d
`ionosphere`	1989	350	34	`gastro`	2016	76	698
`sonar`	NA	207	60	`lsvt`	2014	126	310
`breast`	1995	568	30				

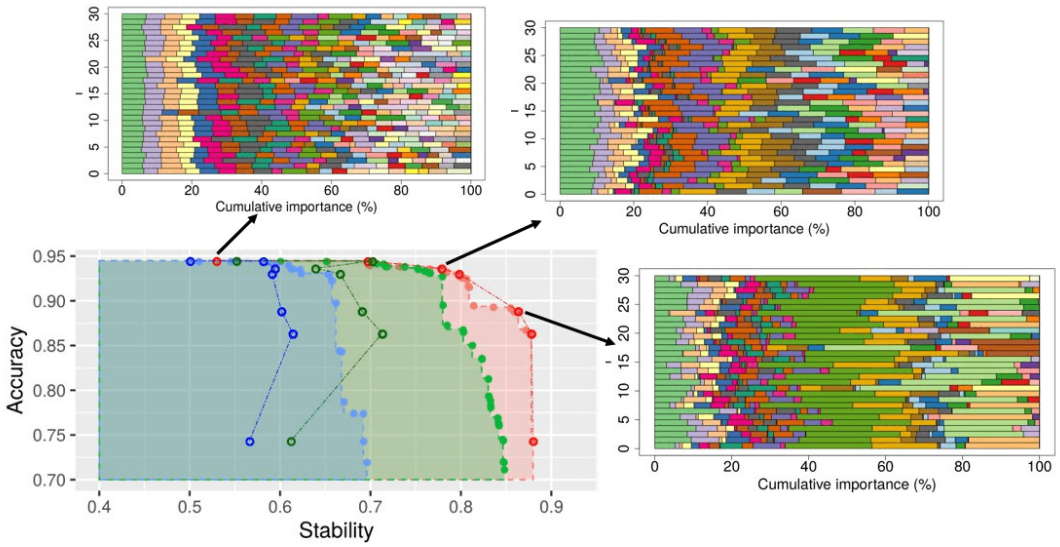
Table 4: Information on the data sets used in this section (in addition to those introduced in Table 3), from the UCI machine learning repository.

7.2.1 The Lasso

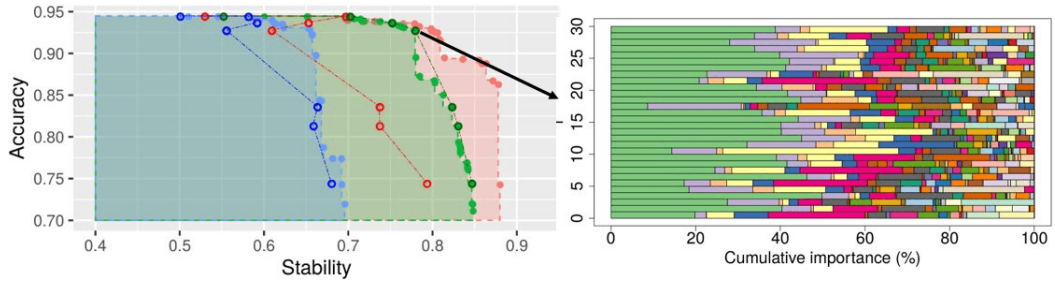
The LASSO, used in the context of logistic regression, finds the linear model $\mathbf{w}$ minimizing

$$\sum_{i=1}^{n} \log(1 + \exp^{-y_i(\mathbf{w}\mathbf{x_i})}) + \lambda ||\mathbf{w}||_1.$$

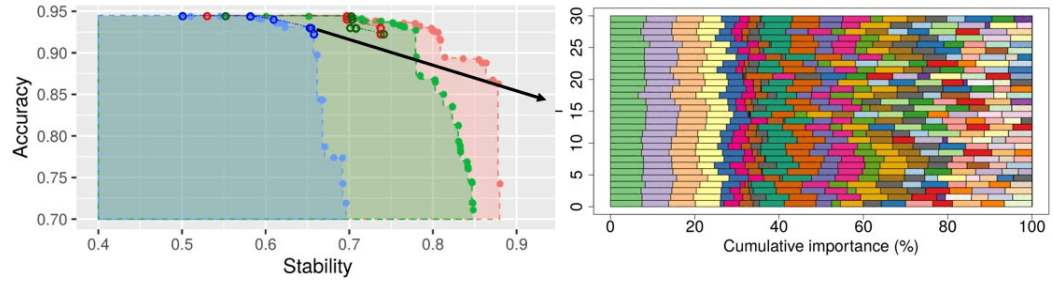
The larger the regularization parameter λ, the fewer features are selected (features with a non-zero weight). The LASSO is known as being an unstable feature selection approach. This is confirmed in our experiments (see Appendix E for comparative results). Nonetheless, Figure 12 illustrates that the LASSO tends to give more importance (width of the rectangles) to frequently selected features (features at the left of the maps). This positively affects ϕ_{iw}. Yet, another kind of instability is often observed with the LASSO: the importance given to each selected feature varies significantly from one selection run to another, which is penalized by ϕ_{iw}. This behavior tends to make ϕ_{iw} lower than ϕ_{pears} as the latter is less sensitive to such an instability. Table 5 summarizes the compromises achievable by the LASSO for the three considered stabilities.



(a) Optimizing ϕ



(b) Optimizing ϕ_{pears}



(c) Optimizing ϕ_{iw}

Figure 11: Pareto-optimal curves (plain points) for ϕ (red), ϕ_{pears} (green) and ϕ_{iw} (blue), obtained with the hybrid-RFE on the `singh` data set. The circled points are the compromises (A, ϕ) (red), (A, ϕ_{pears}) (green) and (A, ϕ_{iw}) (blue) maximizing $\gamma A+(1-\gamma)\phi_{\_}$ for some $0 \leq \alpha \leq 1$ with $\phi_{\_} = \phi$ (a), $\phi_{\_} = \phi_{\text{pears}}$ (b) and $\phi_{\_} = \phi_{\text{blue}}$ (c). The map to the right of each subfigure is chosen with $0.40 \leq \gamma \leq 0.60$ (a), $0.35 \leq \gamma \leq 0.70$ (b) and $0.35 \leq \gamma \leq 0.80$ (c). The optimal trade-off found along Pareto-curves in the (accuracy, stability) space clearly depends on the stability measure used and results into strongly different feature stability maps.

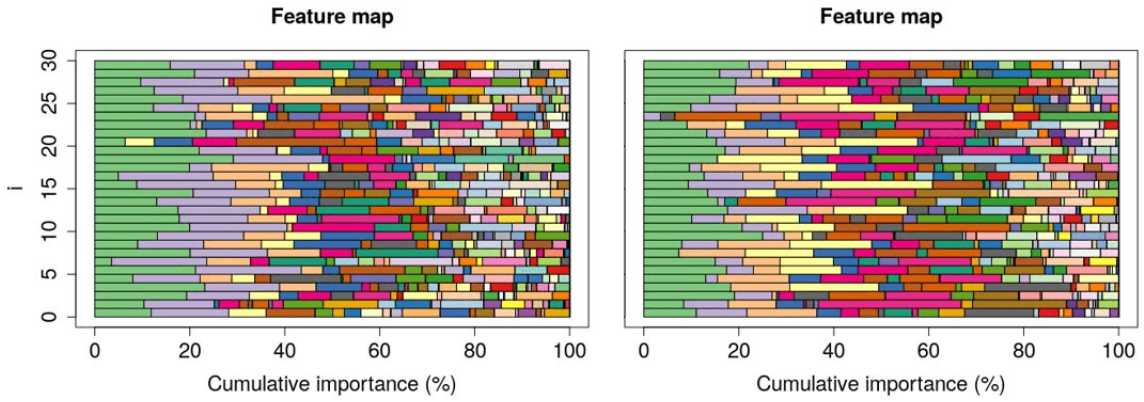


Figure 12: Typical feature stability map of the LASSO (on `singh` (left) and `lsvt`(right)). The LASSO regularization parameter is chosen to select $\min(20, \sqrt{d})$ features on average, but the specific number of selected features here varies across runs. The LASSO gives larger importances (width of the rectangles) to frequently selected features (features at the left of the maps). Yet, feature importance is highly varying.

data	A	ϕ	ϕ_{pears}	ϕ_{iw}	data	A	ϕ	ϕ_{pears}	ϕ_{iw}
`ion`	0.85	0.74	0.91	0.76	`son`	0.73	0.44	0.52	0.37
`bre`	0.95	0.71	0.91	0.78	`gas`	0.82	0.27	0.28	0.21
`lsv`	0.84	0.4	0.51	0.38	`alo`	0.8	0.2	0.23	0.17
`sin`	0.91	0.31	0.57	0.37	`chi`	0.81	0.3	0.36	0.27
`gra`	0.73	0.18	0.22	0.15	`arc`	0.72	0.14	0.21	0.14
`bor`	0.94	0.12	0.16	0.10					

Table 5: Stability of the LASSO on all considered data sets.

7.2.2 Elastic-Net penalty

The ELASTIC NET penalty is a direct generalization to the LASSO penalty which minimizes the linear combination of the L1 and L2 loss

$$\sum_{i=1}^{n} \log(1 + \exp^{-y_i(\mathbf{w}\mathbf{x_i})}) + \lambda_1(\lambda_2||\mathbf{w}||_1 + (1-\lambda_2)||\mathbf{w}||_2). \tag{19}$$

It is purely equivalent to the LASSO when $\lambda_2 = 1$. Figure 13 studies the dependency of the accuracy and stability on the parameter λ_2. Each line depicts the evolution of a stability measure with (from left to right) a decreasing λ_2 parameter. All three stability measures increase when departing from the pure LASSO selection ($\lambda_2 = 1$). On some data sets (notably `singh`, `alon`, `lsvt`), increasing the L2 regularization also first increases the accuracy. Then, as λ_2 continues to decrease, the accuracy starts to drop. On other data sets, such as `gravier`, the accuracy drops directly when departing from the LASSO selection.

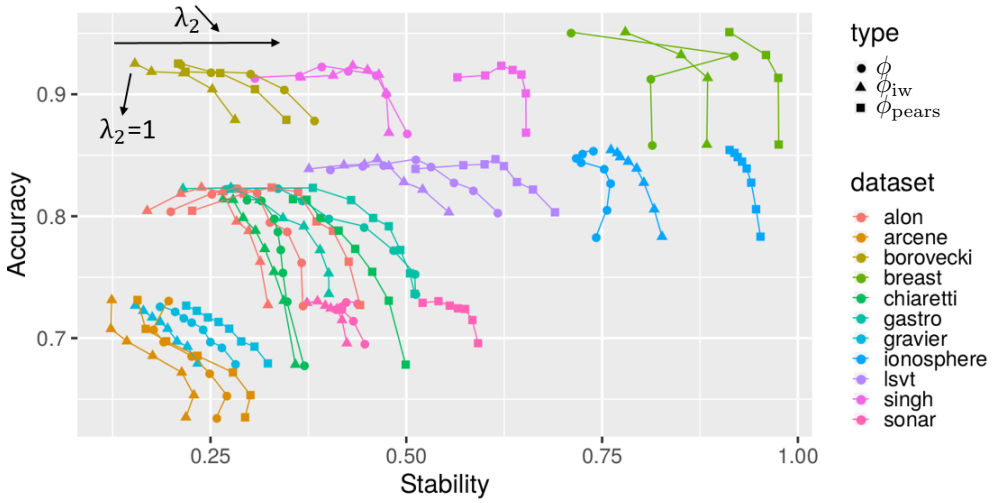


Figure 13: Evolution of the (accuracy, stability) compromises obtained with logistic regression with an ELASTIC NET penalty. The reported stability measures are ϕ, ϕ_{pears} and ϕ_{iw}. Each line starts with $\lambda_2 = 1$ (LASSO) which decreases by 0.1 at each point. Stability is increased when λ_2, the weight of the L1 loss in Equation (19), decreases. This stability increase sometimes comes at the cost of predictive accuracy.

7.2.3 Logistic RFE

The logistic RFE is illustrated by Algorithm 1, without any feature pre-selection (N=0). Equation (19) with $\lambda_2 = 0$ is iteratively minimized and the least significant features are dropped at each iteration. After the selection procedure, which uses a given regularization parameter λ, Equation (19) is minimized again with a different regularization λ_f for learning the predictive model. Increasing λ_f tends to increase ϕ_{iw} and ϕ_{pears} as it reduces the instability of the importance of the selected features. As λ_f does not influence the identity of the selected features, ϕ remains unchanged with its variations. This phenomenon is illustrated in Figure 14 which has been obtained on the `lsvt` data set. Increasing the regularization parameter λ used during the selection improves all stability measures. When λ_f is low (left of each line), ϕ_{iw} is much lower than ϕ due to the instability that occurs during the learning of the final model. Increasing λ_f first increases both the accuracy and ϕ_{iw} before the accuracy finally starts to drop, as a too strong regularization prevents the learning of an adequate model. We observed a very similar behavior for the SVM-RFE algorithm, even though the latter is generally more unstable (see Appendix E).

7.2.4 Random Forests

Random forests can be used for feature selection as well. Feature importance is computed by randomly permuting the features in the out-of-bag samples of each of the T trees and by computing the predictive accuracy decrease these permutations imply (Breiman, 2001; Paul

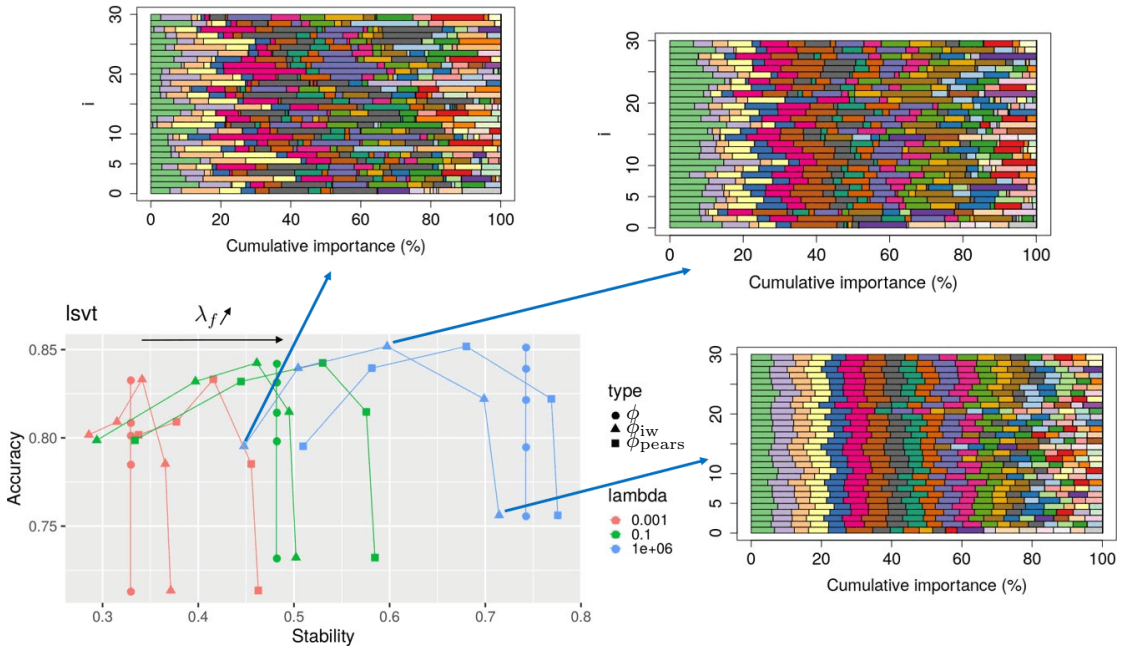


Figure 14: Typical results of the logistic RFE (on `lsvt`). The reported stability measures are ϕ, ϕ_{pears} and ϕ_{iw}. Each line starts (from the left) with a low final regularization parameter λ_f, which is gradually increased. A larger λ_f stabilizes feature importance and improves ϕ_{pears} and ϕ_{iw}. Increasing the regularization λ improves all three stability measures.

et al., 2013). The features whose removal causes the largest accuracy decrease are selected and a new random forest is learned on those features only.

Figure 15 depicts the (accuracy, stability) trade-offs that are achievable on all data sets. Stability is clearly increased when the forest size grows before converging for $T \approx 1000$. Accuracy also tends to be increased and to converge faster than stability which is consistent with the results of (Paul et al., 2012). Figure 16 illustrates typical feature stability maps obtained with random forests, here on `gastro`, with $T = 10$ (left) and $T = 3000$ (right). Clearly, increasing the size of the forests both stabilizes the selection of the feature subsets and the importance of the selected features. Large forests produce much more interpretable feature stability maps.

7.2.5 Relief algorithm

The RELIEF algorithm is a multi-variate filter approach (Kira et al., 1992). Feature scores are computed based on feature value differences between neighbor examples. We consider here a version of the RELIEF where K nearest instances have equal weights.[7] The predictive model is a standard K nearest neighbor classifier with majority voting. As can be seen in Figure 17, stability is largely influenced (and often increased) by the number K of considered

7. We use the `CORElearn` R package for this purpose.

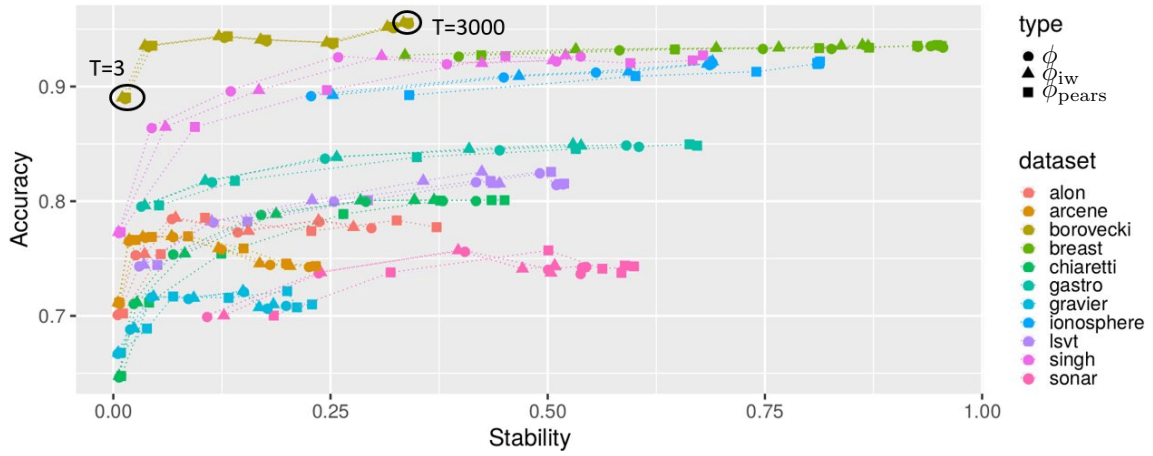


Figure 15: (accuracy, stability) compromises reachable with random forests. The reported stability measures are ϕ, ϕ_{pears} and ϕ_{iw}. Each line starts (from the left) with forests of only three trees. The forest size is gradually increased up to 3000 trees. Larger forests are more stable and have a better predictive accuracy.

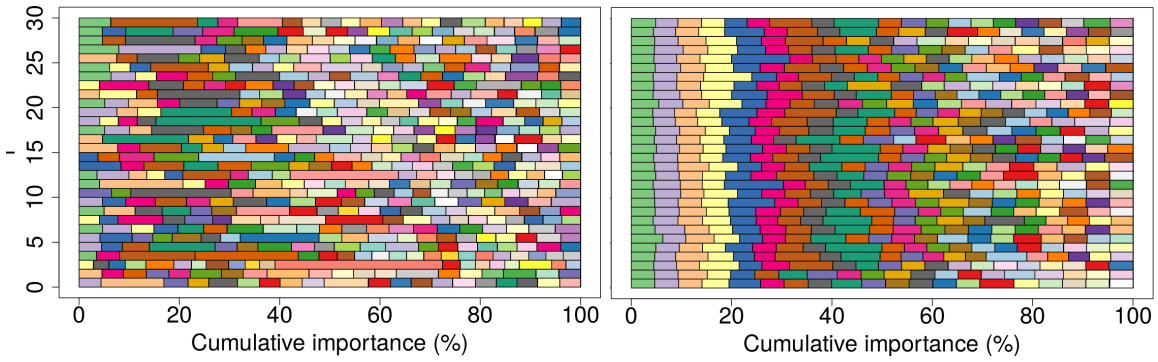


Figure 16: Feature stability maps of random forests on `gastro` with $T = 10$ (left) and $T = 3000$ (right). Increasing the size of the random forests produces more interpretable feature stability maps.

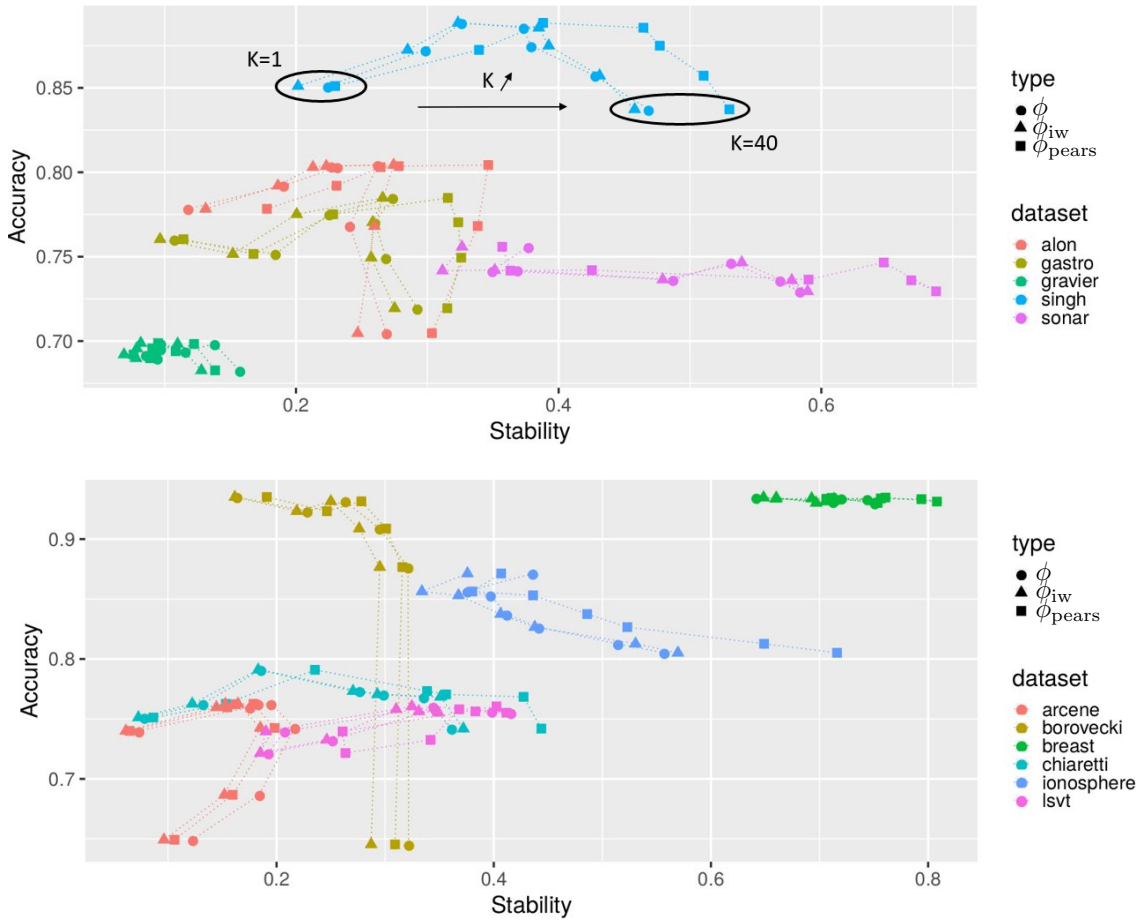


Figure 17: (accuracy, stability) compromises of the RELIEF algorithm. The reported stability measures are ϕ, ϕ_{pears} and ϕ_{iw}. Each line starts (from the left) with only one neighbor considered. This number is gradually increased up to 40 which increases stability in most cases.

neighbors. Typical features maps obtained with the RELIEF (here on `sonar` with $K = 1$ (left) and $K = 25$ (right)) are depicted in Figure 18. In the right map, ϕ_{iw} is higher than ϕ, as frequently selected features have high importances and these importances are quite constant across selection runs.

7.3 Variability of the Stability Measures

Taking into account the importance of selected features in the stability value creates an additional degree of variability that could result in unreliable stability estimation when the number of selection runs M is small. We show here that this added variability is usually small. Furthermore, the variability of our measure ϕ_{iw} is generally lower than the variability of ϕ_{pears}. Figure 19 depicts the variation coefficients $c_v(\triangleq \frac{\sigma}{\mu})$ of ϕ, ϕ_{pears} and ϕ_{iw}, when 200 stability estimates are measured using $M = 5$ with the LASSO. As the LASSO exhibits

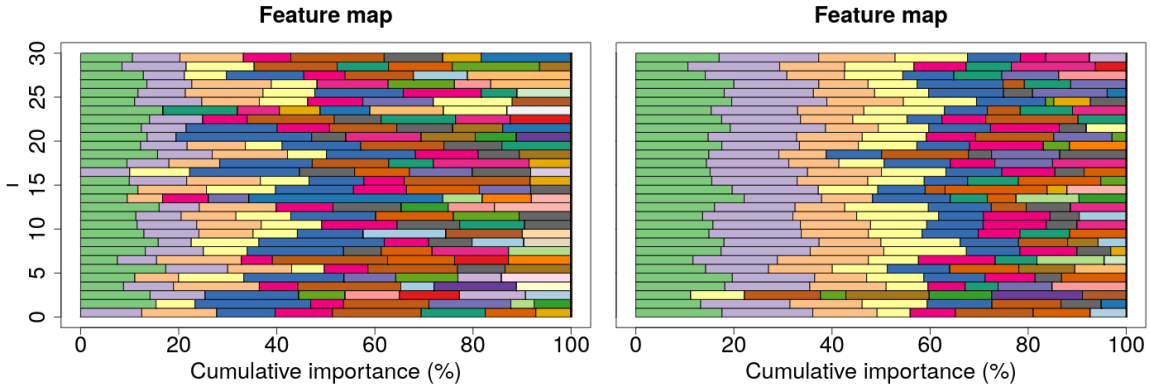


Figure 18: Feature stability maps of the RELIEF on `sonar` with $K = 1$ (left) and $K = 25$ (right).

the largest feature importance variability, it best highlights the differences between stability measures. On 6 of the 11 data sets, $c_{v,\phi} < c_{v,\phi_{\text{iw}}} < c_{v,\phi_{\text{pears}}}$, there is no noticeable difference on 4, and on `ionosphere` only, $c_{v,\phi_{\text{pears}}} < c_{v,\phi_{\text{iw}}} < c_{v,\phi}$. Using importance-based measures indeed creates an additional uncertainty in the stability value, but our proposed measure ϕ_{iw} is, in general, less impacted than ϕ_{pears}. The sampling distributions of all three measures are approximately normally distributed, as illustrated in Figure 20.

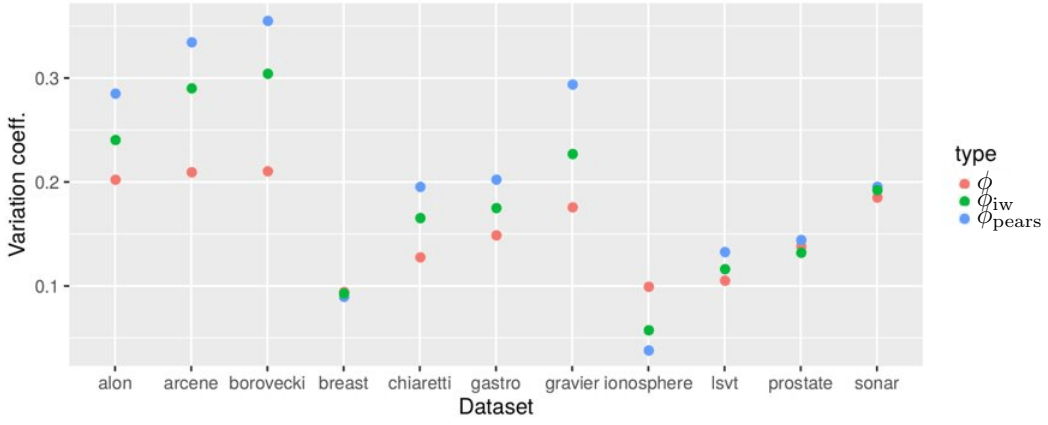


Figure 19: Variation coefficients of the stability measures obtained with the LASSO regression on $M = 1000$ selection runs. Stability estimates are found by aggregating groups of 5 runs (thus producing 200 of such estimates). In most cases, $c_{v,\phi} < c_{v,\phi_{\text{iw}}} < c_{v,\phi_{\text{pears}}}$.

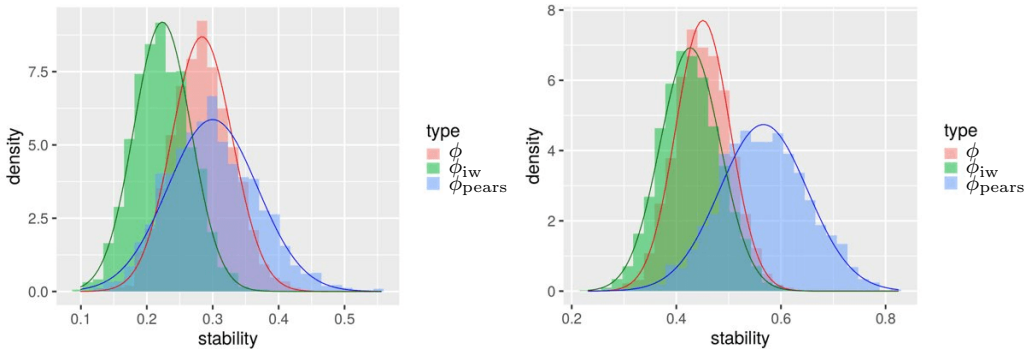


Figure 20: Sampling distributions of the stability measures obtained with the LASSO regression on $M = 1000$ selection runs, for `gastro` (left) and `lsvt` (right). Stability estimates are found by aggregating groups of 5 runs (thus producing 200 of such estimates). These sampling distributions are approximately normally distributed.

8. Conclusions and Perspectives

The typical instability of standard feature selection methods is a key concern nowadays as it reduces the interpretability of the predictive models as well as the trust of domain experts towards the selected feature subsets. Such experts would often prefer a more stable feature selection algorithm over an unstable but possibly slightly more accurate one.

To tackle this issue, predictive accuracy and selection stability are here jointly optimized in a bi-objective framework. Pareto-optimal trajectories are derived, from which domain experts can choose a particular compromise based on their personal preferences. In this context, adequately measuring stability is necessary for domain experts to make informed decisions. While most measures proposed in the literature study the stability of feature subsets, we incorporate into the stability value the selected features importance in predictive models. In particular, we propose a generic way to evaluate the importance of each selected feature in a predictive model and to consider the variability of this importance as an additional source of instability.

Our proposed measure ϕ_{iw} is shown to improve decision-making when stability is jointly optimized with predictive accuracy. Indeed, we show that using current measures in such a bi-objective context can lead to unsatisfactory results and often gives a false sense of stability which hinders appropriate decision-making. We demonstrate this phenomenon both theoretically and experimentally on micro-array and mass-spectrometric data. The proposed measure is also shown to correct for under- or over-estimation of stability in feature spaces with groups of highly correlated variables. While current measures are very sensitive to the size of such groups, we formally prove and validate on simulated data that our measure is not.

We also propose a simple visualization tool, referred to as feature stability maps, which allows the intuitive estimation of stability across several runs of feature selection. Inspection of such feature stability maps allows the identification of the most reliably identified features

along with the most important features in the predictive models. Features combining both properties are likely to be particularly appealing to domain experts.

Even though our proposed measure is shown to improve the behavior of current measures when the feature space is composed of highly correlated feature groups, it still lacks the ability to consider highly correlated features as interchangeable, *i.e.* the alternate selection of highly correlated features still creates instability. Our future work includes the design of an improved stability measure, which would both directly be function of the selected features correlation and incorporate feature importance.

Appendix A. Proof of Properties

In this appendix, we prove the 6 properties satisfied by our proposed stability measure

$$\phi_{\text{iw}} = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \frac{|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} - |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}}}{\bar{k} - C} \tag{20}$$

with

$$|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} = \sum_{f \in \mathcal{F}_i \cap \mathcal{F}_j} \min\left(I_{f,i}, I_{f,j}\right), \; |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}} = \frac{1}{d} \sum_{f \in \mathcal{F}_i, f' \in \mathcal{F}_j} \min(I_{f,i}, I_{f',j})$$

and

$$C = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}}.$$

Property 1 *Fully defined: the measure is defined for every possible importance combinations.*

Proof
For ϕ_{iw} to be undefined, the corrective term C must be equal to $\bar{k}$, such that the denominator of Equation (20) is equal to 0. As feature importances are normalized such that $\sum_f I_{f,i} = \bar{k}, \forall 1 \leq i \leq M$, $|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}}$ is upper bounded by $\bar{k}$, $\forall i, j$. Indeed, such a normalization implies

$$\sum_{f' \in \mathcal{F}_j} \min(I_{f,i}, I_{f',j}) \leq \bar{k}, \forall f \in \mathcal{F}_i.$$

As there are at most d features selected in run i,

$$|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}} = \frac{1}{d} \sum_{f \in \mathcal{F}_i} \sum_{f' \in \mathcal{F}_j} \min(I_{f,i}, I_{f',j}) \leq \frac{1}{d} d\bar{k} \leq \bar{k}.$$

For the inequality not to be strict, all d features must be selected in both selection runs and the importances must all be equal such that $\min(I_{f,i}, I_{f',j}) = I_{f,i} = I_{f,i} \; \forall f, f'$ or no feature must be selected in both runs (as we posed $|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}} = \bar{k}$ in this situations). It then follows that ϕ_{iw} is well defined unless all features are always selected with a constant ($= 1$) importance in each run, or no feature is ever selected.

■

Property 2 *Maximum stability $\Leftrightarrow$ deterministic importance*

Proof For ϕ_{iw} to be maximal ($= 1$, see Property 3), $|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}$ must be equal to $\bar{k} \; \forall i, j$. This condition is equivalent to $I_{f,i} = I_{f,j} \; \forall f, i, j$. In other words, ϕ_{iw} is maximal when the importance of each feature is constant across selection runs. We refer to such a situation as deterministic. ■

Property 3 *Bounds: the measure is bounded by constants not dependent on the overall number of features d or on the average number of features selected $\bar{k}$.*

Proof Our measure ϕ_{iw}, defined in Equation (6), can be trivially restated as

$$\phi_{\text{iw}} = \frac{\left(\frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}\right) - C}{\bar{k} - C}.$$

As $|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} \leq \bar{k}\ \forall i, j$, this implies $\frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} \leq \bar{k}$. The measure ϕ_{iw} is thus upper bounded by 1. The proof of the lower bound is more complex. Referring to

$$A = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}},$$

we first use the chain of equivalence:

$$\phi_{\text{iw}} \geq \frac{-1}{M-1} \Leftrightarrow \frac{A-C}{\bar{k}-C} \geq \frac{-1}{M-1} \Leftrightarrow \frac{M}{M-1}C - A \leq \frac{\bar{k}}{M-1} \Leftrightarrow M^2C - M(M-1)A \leq M\bar{k}.$$

We maximize $M^2C - M(M-1)A$ using linear programming and show that the optimal solution cannot exceed $M\bar{k}$.

$$\begin{aligned}
\text{maximize} \quad & \frac{M}{M-1}\frac{1}{d} \sum_{i} \sum_{j \neq i} \sum_{f} \sum_{f'} z_{ijff'} - \sum_{i} \sum_{j \neq i} \sum_{f} z_{ijff} \\
\text{subject to} \quad & z_{ijff'} - w_{if} \leq 0 \\
& z_{ijff'} - w_{jf'} \leq 0 \\
& z_{ijff} - w_{if} \geq 0 \\
& z_{ijff} - w_{jf} \geq 0 \\
& \textstyle\sum_f w_{i,f} = \bar{k}\ \forall i \\
& z_{ijff'} \geq 0,\ w_{if} \geq 0
\end{aligned}$$

The first four constraints impose that $z_{ijff'} = \min(w_{i,f}, w_{j,f'})$. As the sign of $z_{ijff'}$ in the objective is positive, at least one of the first two constraints will always be tight. The same is true for the 3rd and 4th constraints, as the sign of z_{ijff} is negative in the objective ($\frac{M}{M-1}\frac{1}{d} - 1 \leq 0$ with $d \geq 2, M \geq 2$). This leads to the following dual formulation.

$$\begin{aligned}
\text{minimize} \quad & \sum_{i} \bar{k} W_i \\
\text{subject to} \quad & y_{ijff'1} + y_{ijff'2} \geq \frac{M}{M-1}\frac{1}{d} \text{ for } f \neq f' \\
& y_{ijff1} + y_{ijff2} \geq \tfrac{M}{M-1}\tfrac{1}{d} - 1 \\
& -\textstyle\sum_{j \neq i} \sum_{f'} y_{ijff'1} - \sum_{i \neq j} \sum_{f'} y_{ijff'2} + W_i \geq 0\ \forall i, f \\
& y_{ijff'.} \geq 0, y_{ijff.} \leq 0
\end{aligned}$$

The assignment $y_{ijff'.} = \frac{M}{M-1}\frac{1}{2d}$, $y_{ijff.} = \frac{M}{M-1}\frac{1}{2d} - \frac{1}{2}$, $W_i = M$ is a solution to the problem. The first two constraints are trivially satisfied. One can check that the third one is also satisfied:

$$-\sum_{j\neq i}\sum_{f'} y_{ijff'1} - \sum_{i\neq j}\sum_{f'} y_{ijff'2} + W_i = (-(M-1)d\frac{M}{M-1}\frac{1}{2d}) \times 2 + M = 0.$$

This solution has an objective value of $M\bar{k}$. As a consequence, the optimal value of the primal formulation is upper bounded by $M\bar{k}$. ■

Property 4 *Correction for chance: the measure is constant in expectation (here set to* 0*) when features are selected randomly.*

Proof The average contribution to $|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}$ that a feature selected in both runs i and j imply is

$$\frac{\sum_{f\in\mathcal{F}_i, f'\in\mathcal{F}_j} \min(I_{f,i}, I_{f,j})}{k_i k_j}.$$

If features were to be selected at random (*i.e.* under the null model of feature selection H_0, defined in Section 2.1 as *the situation where every possible feature subset has an equal probability to be selected.*), there would be, on average, $\frac{k_i k_j}{d}$ of such jointly selected features. Then,

$$|\mathcal{F}_i \cap \mathcal{F}_j|^{H_0}_{\text{iw}} = \frac{1}{d}\sum_{f\in\mathcal{F}_i, f'\in\mathcal{F}_j} \min(I_{f,i}, I_{f,j}) = |\mathcal{F}_i \cap \mathcal{F}_j|^{\text{rand}}_{\text{iw}}.$$

Also, in the special cases where k_i, k_j or both are equal to 0, $|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} = |\mathcal{F}_i \cap \mathcal{F}_j|^{\text{rand}}_{\text{iw}}$. Under H_0, the numerator of Equation (6) would then be equal to 0, making $\phi_{\text{iw}}|H_0 = 0$. ■

Property 5 concerns the behavior of stability when the feature space is composed of highly correlated feature groups, which is studied in details in Section 5.3. It formalizes why our proposed measure ϕ_{iw} can remain approximately independent from the correlated groups sizes in the two first scenarios studied in that section (Figures 8a and 8b), where the group LASSO is used for feature selection. We prove that ϕ_{iw} satisfies Property 5 under the following assumptions: 1) the importance of each group is deterministically distributed among their respective selected features and 2) d tends to ∞.

Property 5 *Group size independence: whenever perfectly correlated feature groups are selected as a whole, stability depends only on the groups' cumulative importance, not on the group sizes.*

Proof According to the assumptions (1) and (2) above, it follows that the importance of feature f, which belong to group g, in selection run i, satisfies

$$I_{f,i} = I^1_{g,i}\frac{\bar{k}}{\bar{k}^1 c_g}\delta_f \tag{21}$$

with $I^1_{g,i}$ the importance of the group g in run i whenever the group size c_g is reduced to 1, $\bar{k}^1$ the average number of selected features in this reference scenario, and δ_f the constant fraction of importance feature f has in group g ($0 \leq \delta_f, \sum_{f \in g} \delta_f = 1$). One can easily verify that Equation (21) satisfy the normalization constraint

$$\sum_{f=1}^{d} I^1_{f,i} = \bar{k^1},\ \forall i \Rightarrow \sum_{f=1}^{d} I_{f,i} = \bar{k},\ \forall i.$$

Let I^*_f be, as in Equation (10), equal to $\sum_{i,j=1}^{M} \frac{\min(I_{f,i}, I_{f,j})}{|z_{f,.} \neq 0|^2}$, where $|z_{f,.} \neq 0|$ is the number of runs where feature f is selected. Similarly, let $I^{*,1}_g = \sum_{i,j=1}^{M} \frac{\min(I^1_{g,i}, I^1_{g,j})}{|z_{f,.} \neq 0|^2}$. It follows that

$$I^*_f = \frac{1}{|z_{f,.} \neq 0|^2} \sum_{i,j} \min(I^1_{g,i} \frac{\bar{k}}{\bar{k}^1 c_g} \delta_f, I^1_{g,j} \frac{\bar{k}}{\bar{k}^1 c_g} \delta_f) = \frac{\frac{\bar{k}}{\bar{k}^1 c_g} \delta_f}{|z_{f,.} \neq 0|^2} \sum_{i,j} \min(I^1_{g,i}, I^1_{g,j}) = \frac{\bar{k}}{\bar{k}^1 c_g} \delta_f I^{*,1}_g.$$

Then, using Equation (10),

$$\sum_g \sum_{f \in g} I^*_f p^2_g = \sum_g p^2_g \frac{\bar{k}}{\bar{k}^1 c_g} \sum_{f \in g} \delta_f I^{1,*}_g = \sum_g p^2_g \frac{\bar{k}}{\bar{k}^1} I^{*,1}_g \Rightarrow$$

$$\phi_{\text{iw}} = 1 - \frac{\bar{k} - \sum_g \sum_{f \in g} I^*_f p^2_g}{\bar{k} - C} = 1 - \frac{\bar{k} - \frac{\bar{k}}{\bar{k}^1} \sum_g I^{*,1}_g p^2_g}{\bar{k}^{c_g} - C} = 1 - \frac{\bar{k}^1 - \sum_g I^{*,1}_g p^2_g}{\bar{k}^1 - C} = \phi^1_{\text{iw}},$$

with ϕ^1_{iw} the stability of the reference scenario where the group sizes c_g are all equal to 1. ■

In Section 6.3, we study the behavior of ϕ, ϕ_{pears} and ϕ_{iw} when different feature selection methods are combined. Combining different selection methods allows the exploration of part of the (accuracy, stability) objective space whenever they produce different (accuracy, stability) trade-offs. This is illustrated in Section 7.1 with the hybrid-RFE (introduced in Section 6.1) which combines the classical logistic RFE and a more stable, yet less accurate, univariate criterion. Property 6 states that the contribution of each of the combined selection methods to the stability should be proportional to the importance of their selected features in the global predictive model. We prove Property 6 for ϕ_{iw} under the assumption that d tends to ∞. One of the consequence of this result is that ϕ_{iw}, unlike ϕ, cannot be increased by the selection of stable, yet marginally important features (see Section 7.1 for practical evaluations).

Property 6 *Importance weighted decomposition: when combining the non-overlapping selected feature sets of Q different methods, which produce Q models, in a single predictive model, if features selected by method q have their importance multiplied by δI_q in the combined predictive model of each selection run, stability can be expressed as a weighted sum of the Q prior stabilities:*

$$\phi_{\text{iw}} = \sum_{q=1}^{Q} \frac{\bar{k}_q}{\bar{k}} \delta I_q \phi_{\text{iw}}^q$$

with ϕ_{iw}^q the stability of method q alone, and δI_q the factor such that $I_{f,i} = \delta I_q \times I_{f,i}^q$.

Proof Assuming that d tends to ∞, we can neglect the correction for chance terms, such that

$$\phi_{\text{iw}} = \frac{2}{M(M-1)} \frac{\sum_{i=1}^{M} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}}{\bar{k}}.$$

As there is no feature intersections between the sets selected by the Q methods, one can write

$$|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} = \sum_q \sum_{f_q} \min(I_{f,i}^q \delta I_q, I_{f,j}^q \delta I_q) = \sum_q \delta I_q \sum_{f_q} \min(I_{f,i}^q, I_{f,j}^q) = \sum_q \delta I_q |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^q.$$

Then,

$$\phi_{\text{iw}} = \frac{2}{M(M-1)} \sum_q \delta I_q \frac{\sum_{i=1}^{M} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^q}{\bar{k}} \Rightarrow$$

$$\phi_{\text{iw}} = \sum_q \frac{\bar{k_q}}{\bar{k}} \delta I_q \frac{2}{M(M-1)} \frac{\sum_{i=1}^{M} \sum_{j=i+1}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^q}{\bar{k_q}} = \sum_q \frac{\bar{k}_q}{\bar{k}} \delta I_q \phi_{\text{iw}}^q$$

To keep the normalization $\sum_f I_{f,i} = \bar{k}, \forall i$ correct, the δI_q values must verify

$$\sum_q \frac{\bar{k_q}}{\bar{k}} \delta I_q = 1$$

such that

$$\sum_f I_{f,i} = \sum_q \delta I_q \sum_{f_q} I_{f,i}^q = \sum_q \delta I_q k_q = \bar{k}.$$

■

Appendix B. Proof of Theorems

In this Appendix, we prove Theorem 2, Theorem 3 and Equation (10).

Theorem 2 *Whenever the importance of all selected features is evenly distributed between them in any given run,*

$$\phi_{iw} = \frac{\mu_M[\frac{\bar{k}|\mathcal{F}_i \cap \mathcal{F}_j|}{\max(k_i,k_j)}] - \frac{\bar{k}}{d}\mu_M[\frac{k_i k_j}{\max(k_i,k_j)}]}{\bar{k} - \frac{\bar{k}}{d}\mu_M[\frac{k_i k_j}{\max(k_i,k_j)}]}$$

with $\mu_M(g(i,j)) = \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=i+1}^{M} g^*(i,j)$, $g^*(i,j) = 1$ *if* $k_i = k_j = 0$, $g(i,j)$ *otherwise.*

Proof As feature importance is constant inside each selection run, according to the normalization $\sum_f I_{f,i} = \bar{k}, \forall i$, the importance of a feature in selection run i is equal to $\frac{\bar{k}}{k_i}$. Starting from $|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} = \sum_f \min(I_{f,i}, I_{f,j})$, we get

$$|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} = \sum_{f \in \mathcal{F}_i \cap \mathcal{F}_j} \min(\frac{\bar{k}}{k_i}, \frac{\bar{k}}{k_j}) = \frac{\bar{k}|\mathcal{F}_i \cap \mathcal{F}_j|}{\max(k_i,k_j)}.$$

Also,

$$|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}} = \frac{1}{d}\sum_{f \in \mathcal{F}_i, f' \in \mathcal{F}_j} \min(I_{f,i}, I_{f,j}) = \frac{1}{d}k_i k_j \min(\frac{\bar{k}}{k_i}, \frac{\bar{k}}{k_i}) = \frac{\bar{k}}{d}\frac{k_i k_j}{\max(k_i,k_j)}.$$

■

Theorem 3 *Whenever the importance of all selected features is evenly distributed between them in any given run and the number of selected features is constant across the M runs,*

$$\phi_{iw} = \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=i+1}^{M}\frac{|\mathcal{F}_i \cap \mathcal{F}_j| - \frac{\bar{k}^2}{d}}{\bar{k} - \frac{\bar{k}^2}{d}},$$

which is the usual expression of the Kuncheva index.

Proof Using Theorem 2 and posing $k_i = \bar{k}\ \forall i$, we get $\frac{\bar{k}}{\max(k_i,k_j)} = 1$ and $\frac{k_i k_j}{d} = \frac{\bar{k}^2}{d}$. Then,

$$\phi_{\text{iw}} = \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=i+1}^{M}\frac{|\mathcal{F}_i \cap \mathcal{F}_j| - \frac{\bar{k}^2}{d}}{\bar{k} - \frac{\bar{k}^2}{d}}$$

which is the usual expression of the Kuncheva index. As the Kuncheva index is equivalent to ϕ when the number of selected features is constant, so is ϕ_{iw}. ■

We further prove the frequency-based restatement of ϕ_{iw} (Equation 10):

$$\phi_{\text{iw}} = 1 - \frac{\frac{M}{M-1}(\bar{k} - \sum_{f=1}^{d} I_f^* p_f^2)}{\bar{k} - C} = 1 - \frac{\sum_{f=1}^{d} I_f^* s_f^2 + \frac{M}{M-1}(\bar{k} - \sum_f I_f^* p_f)}{\bar{k} - C}$$

with p_f the selection frequency of feature f in the M runs, $s_f^2 \triangleq \frac{M}{M-1} p_f(1-p_f)$ its selection variance and $I_f^* = \sum_{i,j=1}^{M} \frac{\min(I_{f,i}, I_{f,j})}{|z_{f,\cdot} \neq 0|^2}$ a properly normalized global feature importance, where $|z_{f,\cdot} \neq 0|$ is the number of runs where feature f is selected. In the following proof, $z_{f,i}$ is the Bernouilli variable indicating the selection or non-selection of feature f in selection run i.

Proof We first develop the average pairwise similarity between selection runs:

$$\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \sum_{f=1}^{d} \min(I_{f,i}, I_{f,j}) z_{f,i} z_{f,j}$$

$$= \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{f=1}^{d} \min(I_{f,i}, I_{f,j}) z_{f,i} z_{f,j} - \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{f=1}^{d} I_{i,f} z_{i,f} z_{i,f} =$$

$$\frac{1}{M(M-1)} \sum_{f=1}^{d} M^2 I_f^* p_f^2 - \frac{\bar{k}}{M-1} = \frac{M}{M-1} \sum_{f=1}^{d} I_f^* p_f^2 - \frac{\bar{k}}{M-1} =$$

$$\frac{M}{M-1} \sum_{f=1}^{d} I_f^* p_f - \sum_{f=1}^{d} I_f^* s_f^2 - \frac{\bar{k}}{M-1}.$$

Injecting into ϕ_{iw},

$$\phi_{\text{iw}} = \frac{2}{M(M-1)} \sum_{i=1, j>i}^{M} \frac{|\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}} - |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}^{\text{rand}}}{\bar{k} - C} = \frac{\frac{1}{M(M-1)}(\sum_{i=1, j\neq i}^{M} |\mathcal{F}_i \cap \mathcal{F}_j|_{\text{iw}}) - C}{\bar{k} - C}$$

$$= \frac{(\frac{M}{M-1} \sum_{f=1}^{d} I_f^* p_f - \sum_{f=1}^{d} I_f^* s_f^2 - \frac{\bar{k}}{M-1}) - C}{\bar{k} - C}$$

$$= \frac{(\frac{M}{M-1} \sum_{f=1}^{d} I_f^* p_f - \sum_{f=1}^{d} I_f^* s_f^2 - \frac{M\bar{k}}{M-1}) + \bar{k} - C}{\bar{k} - C}$$

$$= 1 - \frac{\sum_{f=1}^{d} I_f^* s_f^2 + \frac{M}{M-1}(\bar{k} - \sum_f I_f^* p_f)}{\bar{k} - C}.$$

■

Appendix C. Time Complexity of the Correction for Chance

In this Appendix, we propose an algorithm which computes the corrective term C in the definition of ϕ_{iw} in $O(M^2\bar{k}+M\overline{k\log(k)})$, with $\overline{k\log(k)} = \frac{1}{M}\sum_{i=1}^{M} k_i \log(k_i)$. As a reminder, C is defined as

$$C = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \frac{1}{d} \sum_{f\in\mathcal{F}_i, f'\in\mathcal{F}_j} \min(I_{f,i}, I_{f',j}).$$

Algorithm 2 Computing C.

```
1:  procedure C(F, I, M)                                   ▷ O(M²k̄ + M·overline(k log(k)))
2:      for i ∈ 1:M do                                     ▷ O(M·overline(k log(k)))
3:          I_i ← sort(I_i)                                ▷ O(k_i log(k_i))
4:          Isum_i[l] ← Σ_{p≤l} I_i[p]   ∀1 ≤ l ≤ k_i      ▷ O(k_i)
5:      C ← 0
6:      for i, j ∈ 1 : M do                                ▷ O(M²k̄)
7:          C ← 0, |F_i ∩ F_j|^rand_iw ← 0, c_i = 0
8:          for c_j ∈ 1:k_j do                             ▷ O(k_i + k_j)
9:              while c_i < k_i  &  I_i[c_i + 1] < I_j[c_j] do
10:                 c_i ← c_i + 1
11:             if c_i == 0 then
12:                 |F_i ∩ F_j|^rand_iw += I_j[1] × k_i
13:             else
14:                 if c_i == k_i then
15:                     |F_i ∩ F_j|^rand_iw += Isum_i[k_i]
16:                 else
17:                     |F_i ∩ F_j|^rand_iw += Isum_i[c_i] + (k_i − c_i) × I_j[c_j]
18:         C ← C + |F_i ∩ F_j|^rand_iw / M²
19:     return C/d
```

The computational complexities are reported in the comments of the main steps of this algorithm. We first sort the selected features by increasing importance for each selection run and store the cumulative importance of the first l features in $Isum_i[l]$. Then, for each pair of runs, we can compute $\sum_{f\in\mathcal{F}_i, f'\in\mathcal{F}_j} \min(I_{f,i}, I_{f',j})$ by traversing each sorted array of features once. The invariant of the `for` loop on line 8 is the following:

$$|\mathcal{F}_i \cap \mathcal{F}_j|^{\text{rand}}_{\text{iw}} = \sum_{f\in\mathcal{F}_i, f'\in\mathcal{F}_j[1:(c_j-1)]} \min(I_{f,i}, I_{f',j}).$$

After the `while` loop on line 9 is executed, c_i is equal to the biggest importance in run i lower than $I_j[c_j]$, the importance of the feature of run j considered at this iteration of the `for` loop. The sum of minimums $\sum_{f\in\mathcal{F}_i, f'=\mathcal{F}_j[c_j]} \min(I_{f,i}, I_{f',j})$ can then be decomposed

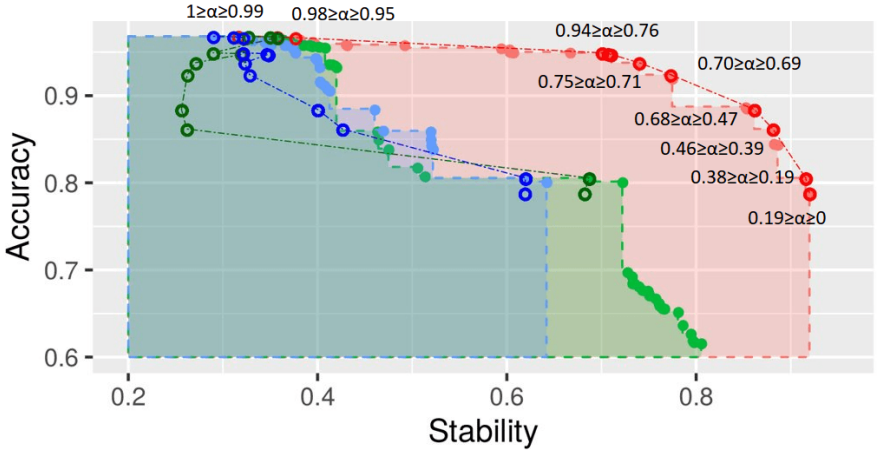
into two terms (the cases where $c_i = 0$ or $c_i = k_i$ are handled explicitly in the algorithm):

$$\sum_{f \in \mathcal{F}_i[1:c_i], f' = \mathcal{F}_j[c_j]} \min(I_{f,i}, I_{f',j}) + \sum_{f \in \mathcal{F}_i[(c_i+1):k_i], f' = \mathcal{F}_j[c_j]} \min(I_{f,i}, I_{f',j}) = Isum_i[c_i] + (k_i - c_i) I_j[c_j].$$

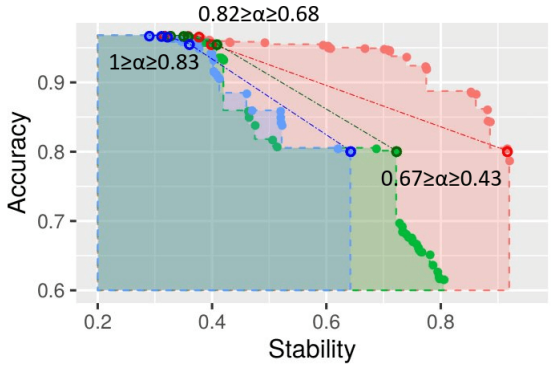
Appendix D. Decision Making for Cancer Diagnosis

This Appendix further studies the behavior of the measures ϕ, ϕ_{pears} and our measure ϕ_{iw} in the context of joint optimization with predictive accuracy. As detailed in Section 7.1, ϕ is best increased by the selection of stable features, yet marginally used for prediction, while optimizing ϕ_{pears} can give highly varying importances to the selected features (example in Figure 11). Optimizing our proposed measure leads to more satisfactory feature stability maps with higher predictive performance.

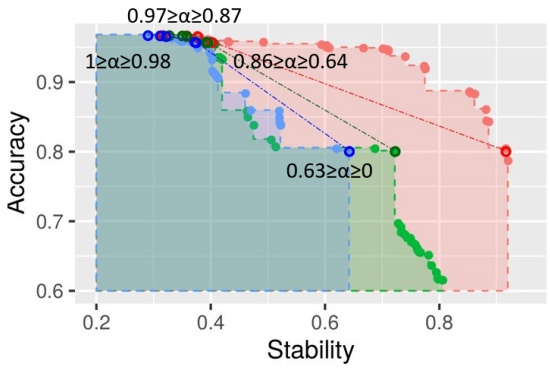
This Appendix presents the results obtained with the hybrid-RFE on the data sets `borovecki` (Figure 21), `chiaretti` (Figure 22), `alon` (Figure 23), `gravier` (Figure 24) and `arcene` (Figure 25). As in Figure 11 in Section 7.1 (on `singh`), the plain points in the figures below represent all Pareto-optimal trade-offs when ϕ (red), ϕ_{pears} (green) and ϕ_{iw} (blue) are used to assess stability. The circled points are the compromises (A, ϕ) (red), (A, ϕ_{pears}) (green) and (A, ϕ_{iw}) (blue) maximizing $\gamma A + (1 - \gamma)\phi_{\_}$ for some $0 \leq \gamma \leq 1$ with $\phi_{\_} = \phi$ (a), $\phi_{\_} = \phi_{\text{pears}}$ (b) and $\phi_{\_} = \phi_{\text{iw}}$ (c). The Pareto-optimal fronts obtained when each measure is optimized are in general different. This illustrates that optimizing ϕ or (to a lesser extent here) ϕ_{pears} is not guaranteed to increase ϕ_{iw}, which can lead to unsatisfactory results (see feature stability maps of Figure 11). Even when Pareto-fronts are approximately identical, the γ values corresponding to each compromise can vary significantly (ex: `chiaretti` (Figure 22) for ϕ_{pears} and ϕ_{iw}) which ultimately leads to different chosen trade-offs.



(a) Optimizing ϕ

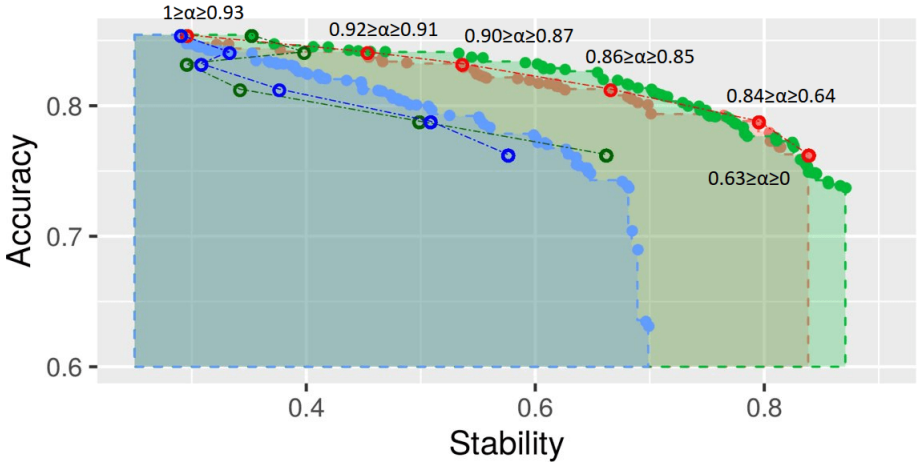


(b) Optimizing ϕ_{pears}

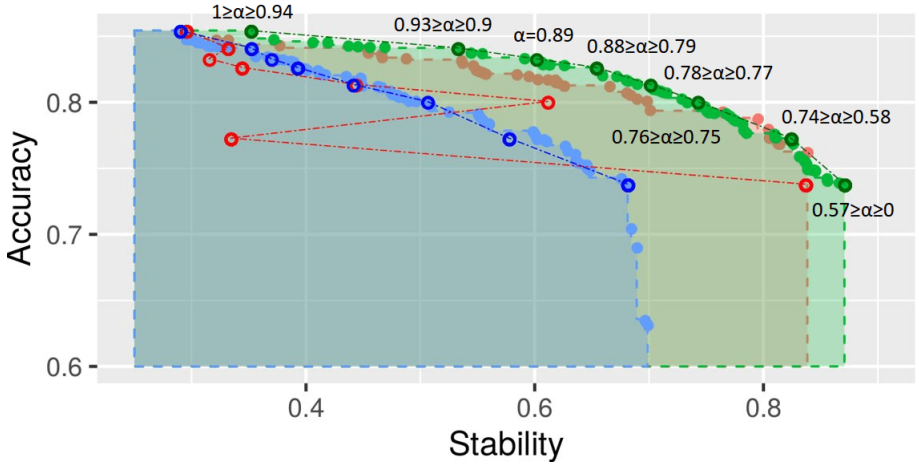


(c) Optimizing ϕ_{iw}

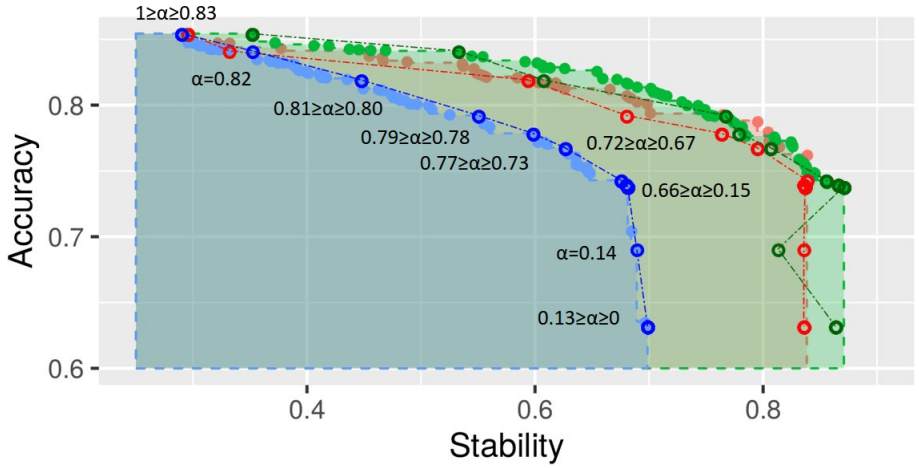
Figure 21: Pareto-optimal curves for ϕ (red), ϕ_{pears} (green) and ϕ_{iw} (blue), obtained with the hybrid-RFE on the `borovecki` data set. From (a), it is clear that increasing ϕ do not necessarily increase ϕ_{pears} and ϕ_{iw}. Furthermore, accuracy is sacrificed for γ values much bigger when ϕ is optimized. Using ϕ to estimate stability gives a false sense of stability and may convince domain experts to sacrifice more predictive performance, even though the more robust measure ϕ_{iw} is barely increased. For this data set, only small differences are observed when optimizing ϕ_{pears} or ϕ_{iw}.



(a) Optimizing ϕ

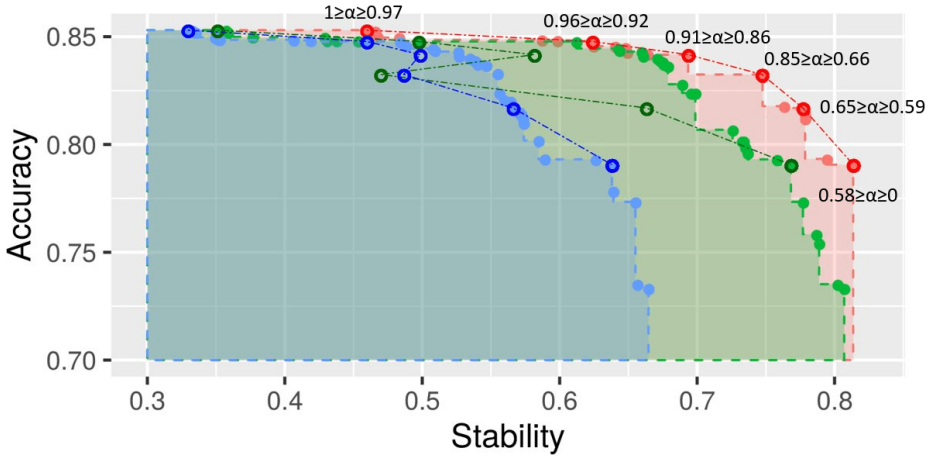


(b) Optimizing ϕ_{pears}

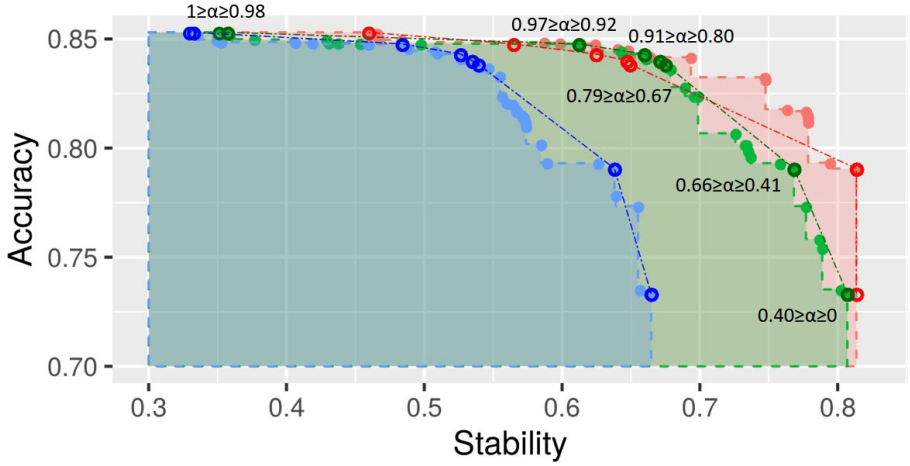


(c) Optimizing ϕ_{iw}

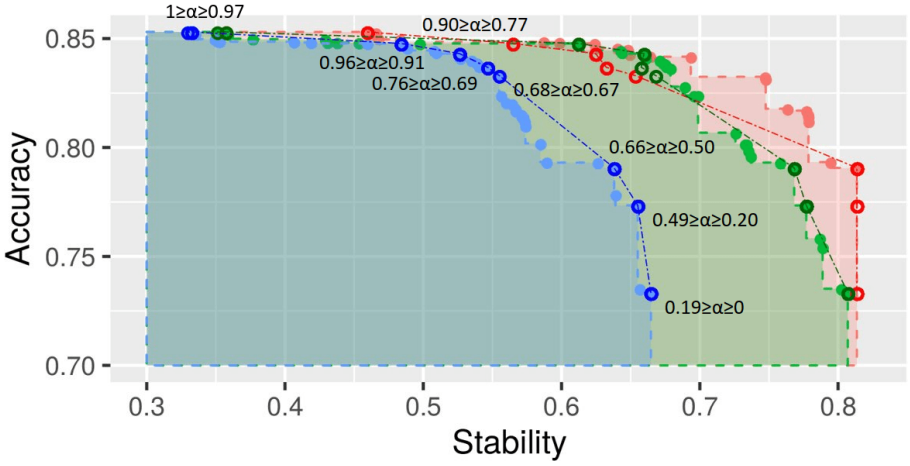
Figure 22: Pareto-optimal curves for ϕ (red), ϕ_{pears} (green) and ϕ_{iw} (blue), obtained with the hybrid-RFE on the `chiaretti` data set.



(a) Optimizing ϕ

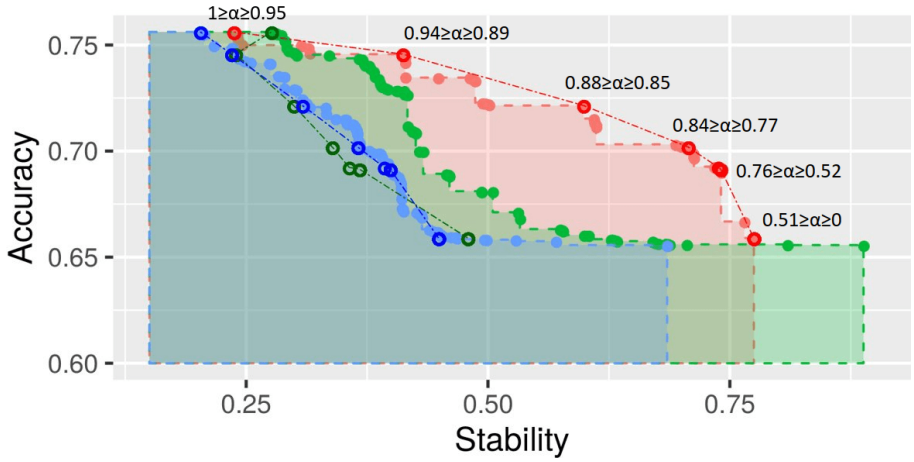


(b) Optimizing ϕ_{pears}

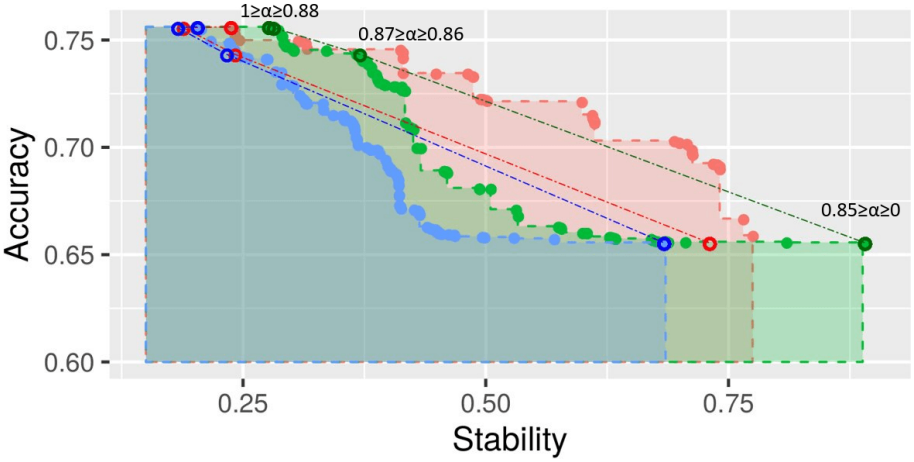


(c) Optimizing ϕ_{iw}

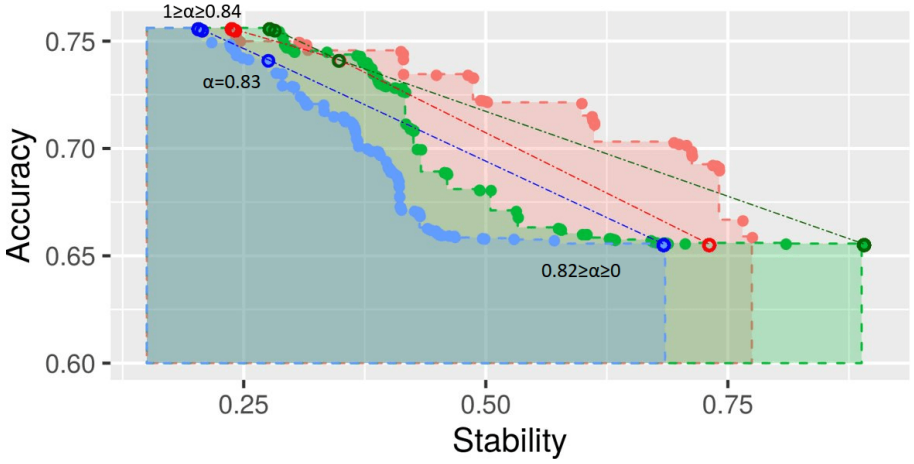
Figure 23: Pareto-optimal curves for ϕ (red), ϕ_{pears} (green) and ϕ_{iw} (blue), obtained with the hybrid-RFE on the `alon` data set.



(a) Optimizing ϕ

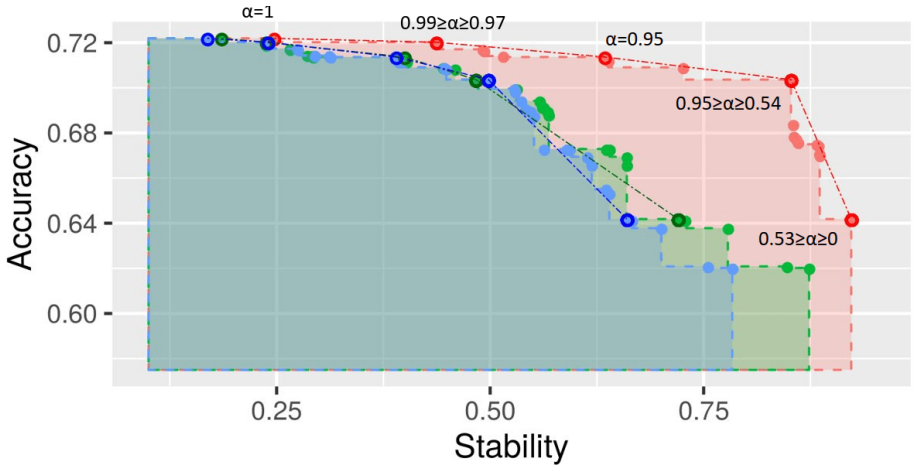


(b) Optimizing ϕ_{pears}

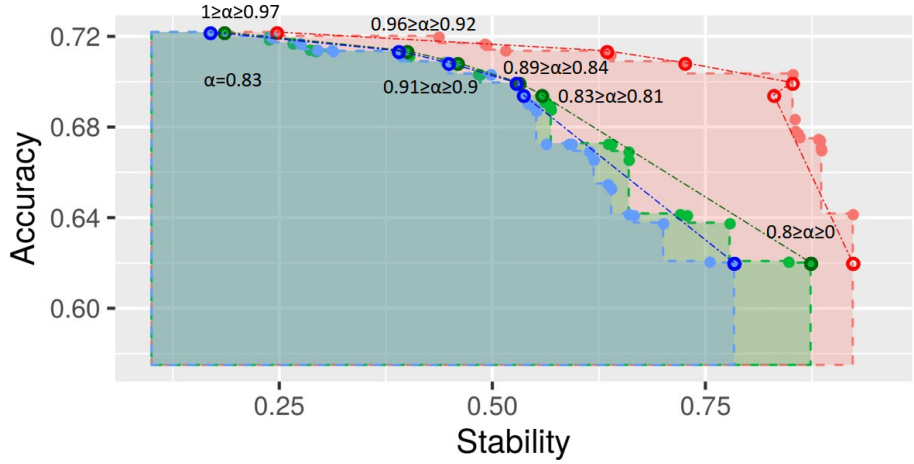


(c) Optimizing ϕ_{iw}

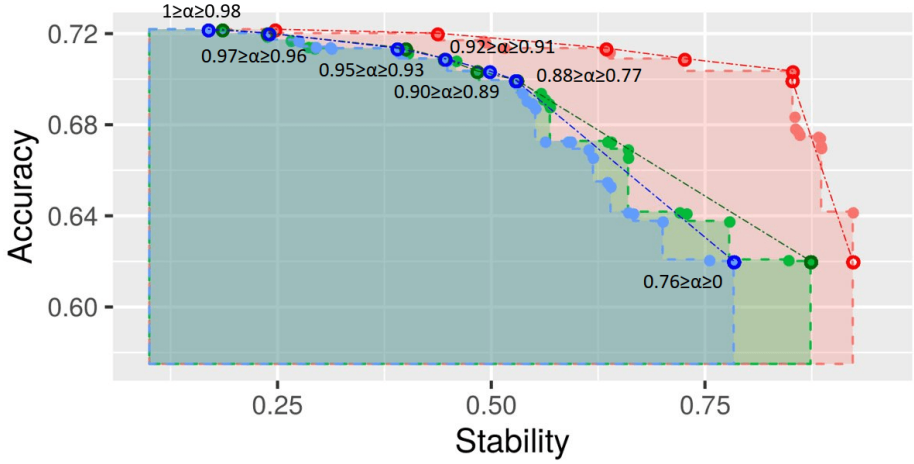
Figure 24: Pareto-optimal curves for ϕ (red), ϕ_{pears} (green) and ϕ_{iw} (blue), obtained with the hybrid-RFE on the `gravier` data set.



(a) Optimizing ϕ



(b) Optimizing ϕ_{pears}



(c) Optimizing ϕ_{iw}

Figure 25: Pareto-optimal curves for ϕ (red), ϕ_{pears} (green) and ϕ_{iw} (blue), obtained with the hybrid-RFE on the `arcene` data set.

Appendix E. Stability of Standard Feature Selection Methods

In this Appendix, the aggregated results of Section 7.2 are depicted. They compare the stability of logistic regression with a LASSO or ELASTIC NET penalty, the logistic RFE, the SVM-RFE, random forests and the RELIEF algorithm. For the sake of readability, we compare ϕ and ϕ_{iw} only. Both stability measures ϕ and ϕ_{iw} have similar values most of the times. This illustrates that ϕ_{iw} tends to behave like ϕ in standard scenarios. Selection methods that are the most present in the overall Pareto front, hence offering a best trade-off between predictive accuracy and stability, are the random forests and the logistic RFE while the LASSO, the RELIEF and the SVM-RFE seems the most unstable methods.

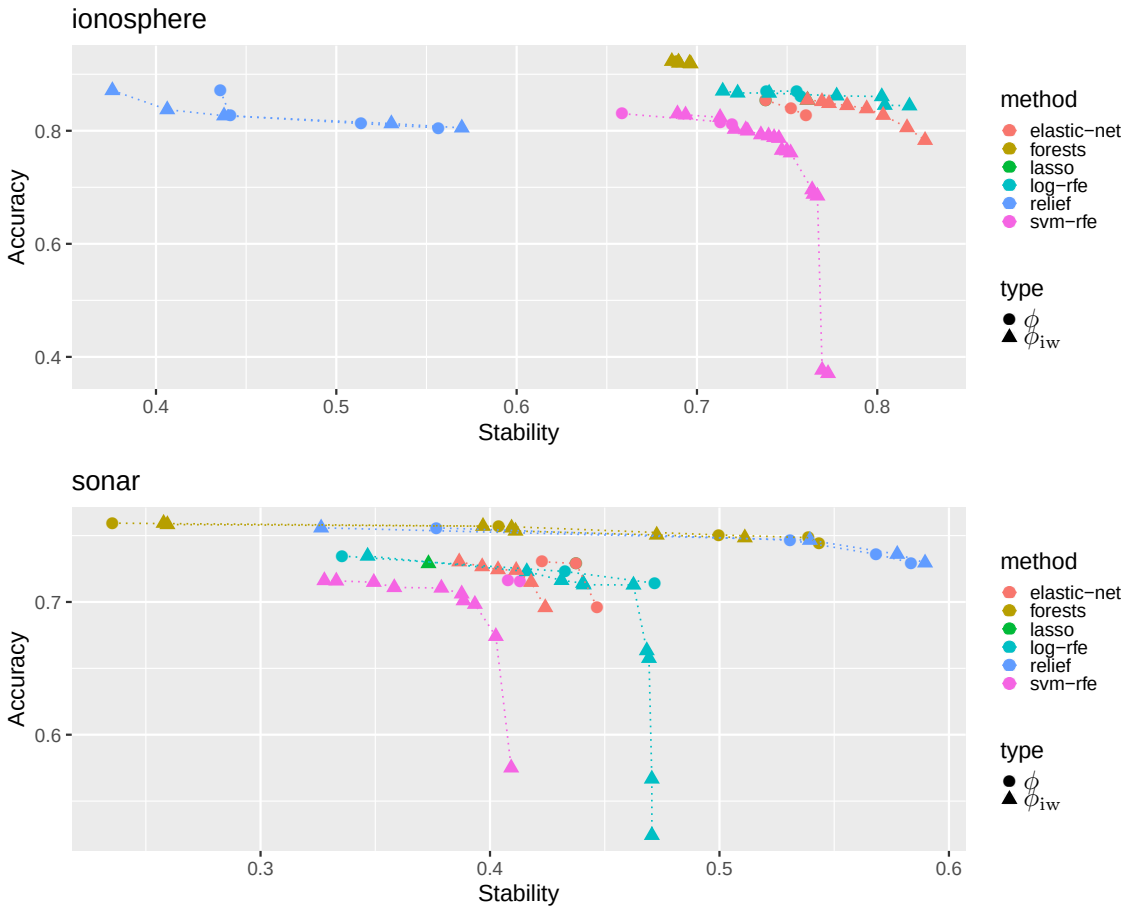


Figure 26: Comparison of ϕ and ϕ_{iw} of the ELASTIC NET, the LASSO, random forests, the logistic RFE, the SVM-RFE and the RELIEF on the `ionosphere` and `sonar` data sets.

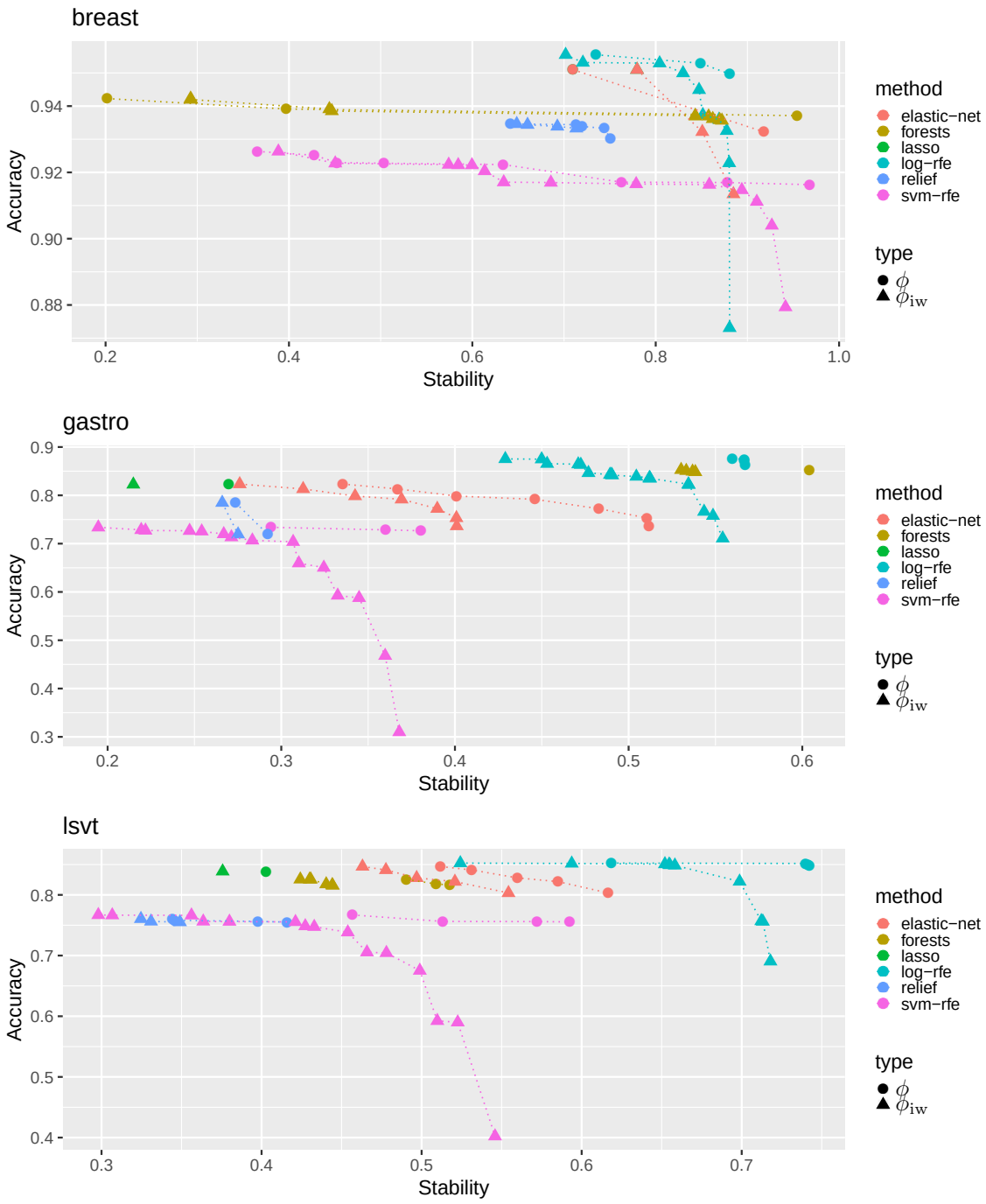


Figure 27: Comparison of ϕ and ϕ_{iw} of the ELASTIC NET, the LASSO, random forests, the logistic RFE, the SVM-RFE and the RELIEF on the `breast`, `gastro` and `lsvt` data sets.

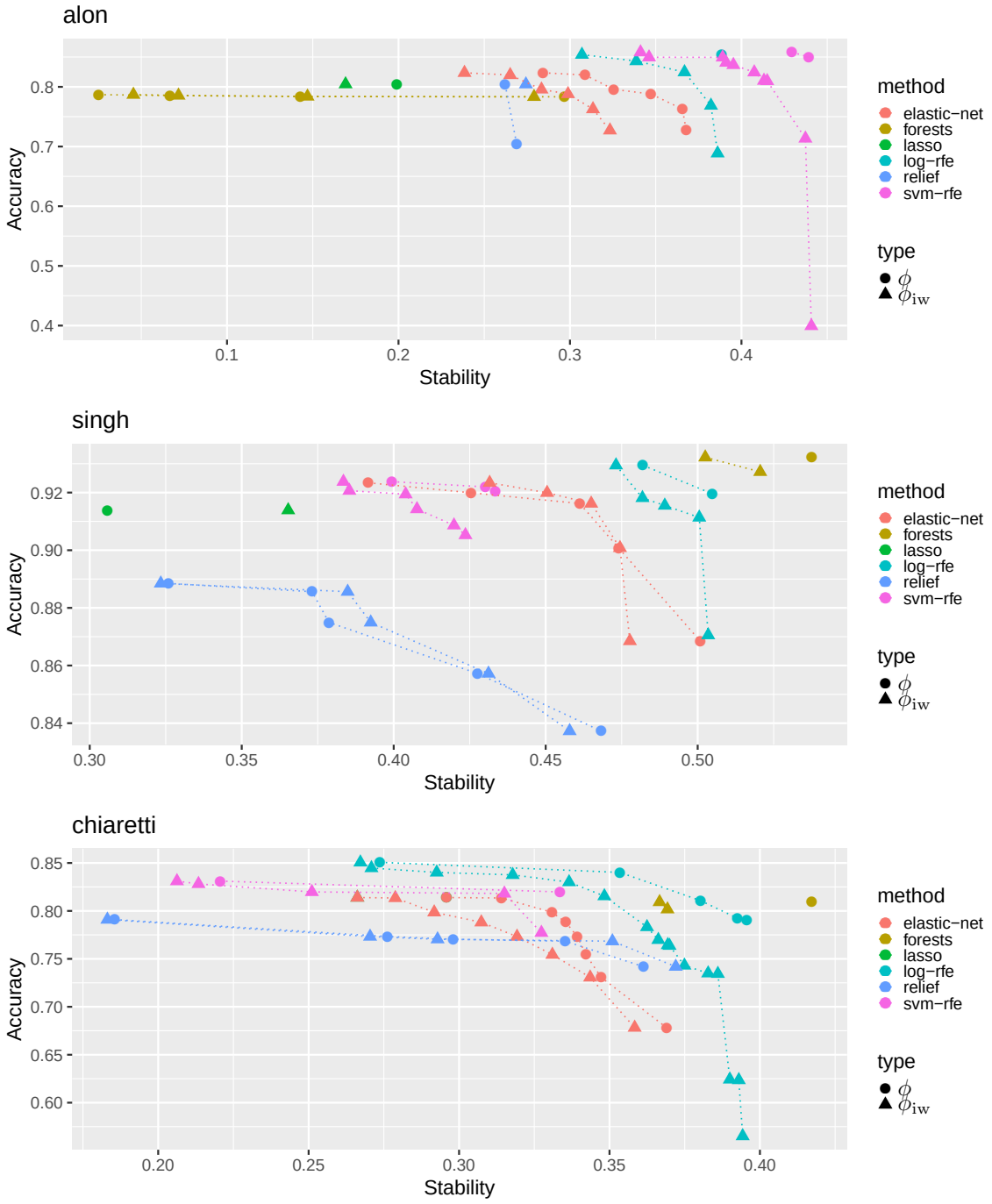


Figure 28: Comparison of ϕ and ϕ_{iw} of the ELASTIC NET, the LASSO, random forests, the logistic RFE, the SVM-RFE and the RELIEF on the `alon`, `singh` and `chiaretti` data sets.

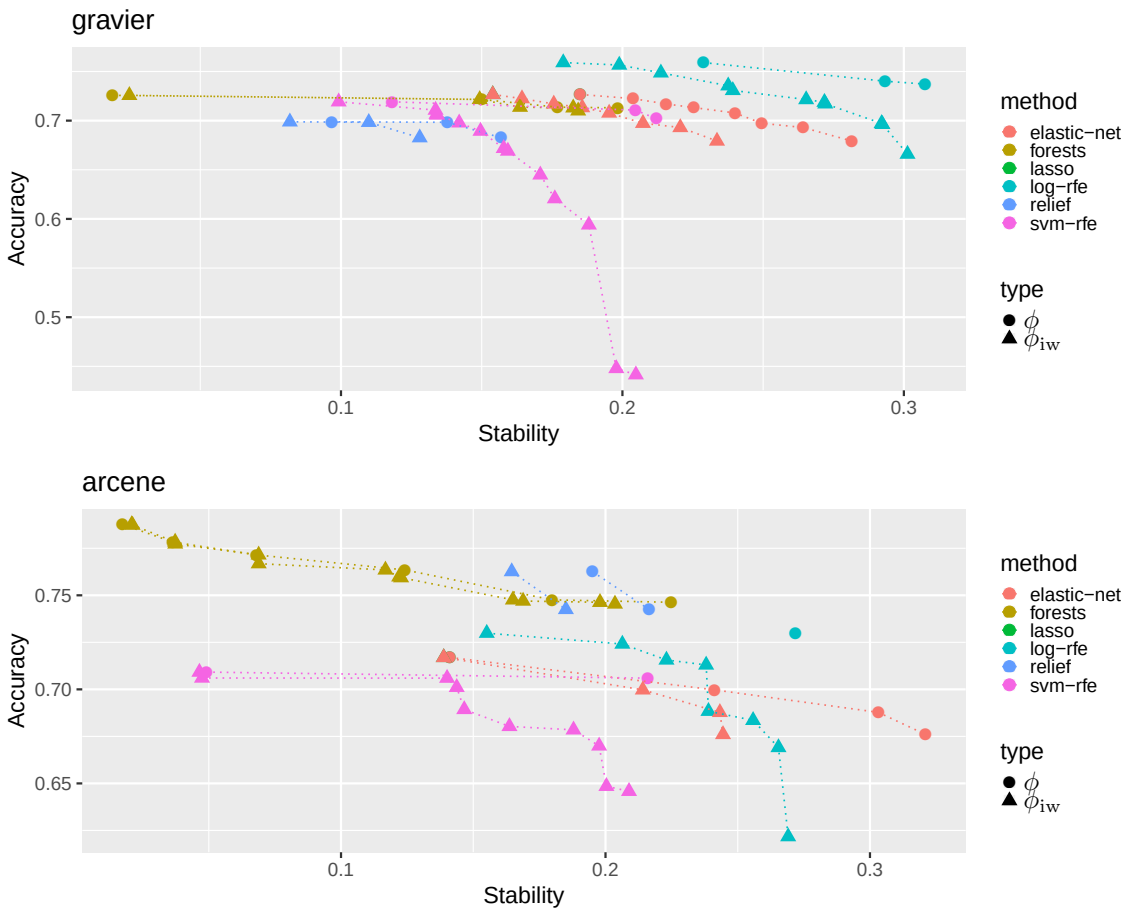


Figure 29: Comparison of ϕ and ϕ_{iw} of the elastic net, the lasso, random forests, the logistic rfe, the svm-rfe and the relief on the `gravier` and `arcene` data sets.

References

Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2010.

Salem Alelyani. *On feature selection stability: A data perspective*. PhD thesis, Arizona State University, 2013.

Wael Awada, Taghi M Khoshgoftaar, David Dittman, Randall Wald, and Amri Napolitano. A review of the stability of feature selection techniques for bioinformatics data. In *International Conference on Information Reuse & Integration (IRI)*, pages 356–363. IEEE, 2012.

Anne-Laure Boulesteix and Martin Slawski. Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10(5):556–568, 2009.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(02):185–205, 2005.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

Benjamin Haibe-Kains, Nehme El-Hachem, Nicolai Juul Birkbak, Andrew C Jin, Andrew H Beck, Hugo JWL Aerts, and John Quackenbush. Inconsistency in large pharmacogenomic studies. *Nature*, 504(7480):389–393, 2013.

Victor Hamer and Pierre Dupont. Joint optimization of predictive performance and selection stability. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2020.

Yue Han and Lei Yu. A variance reduction framework for stable feature selection. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):428–445, 2012.

Giuseppe Jurman, Stefano Merler, Annalisa Barla, Silvano Paoli, Antonio Galea, and Cesare Furlanello. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, 24(2):258–264, 2008.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms. In *International Conference on Data Mining (ICDM)*, pages 8–pp. IEEE, 2005.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12(1):95–116, 2007a.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12(1):95–116, 2007b.

Kenji Kira, Larry A Rendell, et al. The feature selection problem: Traditional methods and a new algorithm. In *Aaai*, volume 2, pages 129–134, 1992.

Ravi Kumar and Sergei Vassilvitskii. Generalized distances between rankings. In *International Conference on World Wide Web (WWW)*, pages 571–580, 2010.

Ludmila I Kuncheva. A stability index for feature selection. In *Artificial Intelligence and Applications*, pages 421–427, 2007.

Yifeng Li, Chih-Yu Chen, and Wyeth W Wasserman. Deep feature selection: theory and application to identify enhancers and promoters. *Journal of Computational Biology*, 23 (5):322–336, 2016.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Sarah Nogueira and Gavin Brown. Measuring the stability of feature selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 442–457. Springer, 2016.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *The Journal of Machine Learning Research*, 18(1):6345–6398, 2017a.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the use of Spearman's rho to measure the stability of feature rankings. In *Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA)*, pages 381–391. Springer, 2017b.

Jérôme Paul, Michel Verleysen, and Pierre Dupont. The stability of feature selection and class prediction from ensemble tree classifiers. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2012.

Jérôme Paul, Michel Verleysen, Pierre Dupont, et al. Identification of statistically significant features from random forests. In *ECML workshop on Solving Complex Machine Learning Problems with Ensemble Methods*, pages 69–80, 2013.

Debaditya Roy, K Sri Rama Murty, and C Krishna Mohan. Feature selection using deep neural networks. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE, 2015.

Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 313–325. Springer, 2008a.

Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 313–325. Springer, 2008b.

Konstantinos Sechidis, Konstantinos Papangelou, Sarah Nogueira, James Weatherall, and Gavin Brown. On the stability of feature selection in the presence of feature correlations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2019.

Ahmad Abu Shanab, Taghi M Khoshgoftaar, Randall Wald, and Amri Napolitano. Impact of noise and data sampling on stability of feature ranking techniques for biological datasets. In *International Conference on Information Reuse and Integration (IRI)*, pages 415–422. IEEE, 2012.

Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise De Longueville, Ernest S Kawasaki, Kathleen Y Lee, et al. The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9):1151, 2006.

Petr Somol and Jana Novovicova. Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1921–1939, 2010.

Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data Classification: Algorithms and Applications*, page 37, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Ari Urkullu, Aritz Pérez, and Borja Calvo. Statistical model for reproducibility in ranking-based feature selection. *Knowledge and Information Systems*, pages 1–32, 2020.

Bernard L Welch. The generalization of student's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

Lei Yu, Chris Ding, and Steven Loscalzo. Stable feature selection via dense feature groups. In *International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD)*, pages 803–811, 2008.

Min Zhang, Lin Zhang, Jinfeng Zou, Chen Yao, Hui Xiao, Qing Liu, Jing Wang, Dong Wang, Chenguang Wang, and Zheng Guo. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, 25(13):1662–1668, 2009a.

Min Zhang, Lin Zhang, Jinfeng Zou, Chen Yao, Hui Xiao, Qing Liu, Jing Wang, Dong Wang, Chenguang Wang, and Zheng Guo. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, 25(13):1662–1668, 2009b.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.