Journal of Machine Learning Research 18 (2018) 1-54 Submitted 9/17; Revised 2/18; Published 4/18

On the Stability of Feature Selection Algorithms

Sarah Nogueira SARAH.NOGUEIRA@MANCHESTER.AC.UK
Konstantinos Sechidis KONSTANTINOS.SECHIDIS@MANCHESTER.AC.UK
Gavin Brown GAVIN.BROWN@MANCHESTER.AC.UK
School of Computer Science
University of Manchester
Manchester M13 9PL, UK

Editor: Isabelle Guyon

Abstract

Feature Selection is central to modern data science, from exploratory data analysis to predictive model-building. The "stability" of a feature selection algorithm refers to the *robustness* of its feature preferences, with respect to data sampling and to its stochastic nature. An algorithm is 'unstable' if a *small* change in data leads to *large* changes in the chosen feature subset. Whilst the idea is simple, *quantifying* this has proven more challenging—we note numerous proposals in the literature, each with different motivation and justification. We present a rigorous statistical treatment for this issue. In particular, with this work we consolidate the literature and provide (1) a deeper understanding of existing work based on a small set of properties, and (2) a clearly justified statistical approach with several novel benefits. This approach serves to identify a stability measure obeying all desirable properties, and (for the first time in the literature) allowing confidence intervals and hypothesis tests on the stability, enabling rigorous experimental comparison of feature selection algorithms.

Keywords: stability, feature selection

1. Introduction

High-dimensional data sets are the norm in data-intensive scientific domains. In application areas from bioinformatics to business analytics, it is common to collect many more measurements ("features" or "variables") than a study is able to easily cope with. This is a natural consequence of exploratory data analysis, but brings challenges of computational overhead, model interpretability and overfitting. Modern statistical regularisation methods can often control the model fit, but if the task is to identify *meaningful* subsets of features, there is a jungle of heuristics and domain-specific feature selection methods from which to pick, surveyed in several previous works, e.g. Guyon and Elisseeff (2003); Brown et al. (2012). Many authors have addressed the question of how sensitive each feature selection method is, with respect to small changes in the training data. If, with a different sample from the same training data, the chosen subset of features changes radically, then it is regarded as being an *unstable* procedure. Conversely, if the feature subset is almost static with respect to data changes, it is a *stable* procedure. Whilst the intuition here is clear, there is to date no single agreed measure to *quantify* stability, and numerous proposals in the literature.

The first published work to consider the *stability of feature selection procedures* was Kalousis et al. (2005), with extended experimental results published later as Kalousis et al. (2007). A slightly earlier technical report (Dunne et al., 2002) examined the idea in the limited scope of wrapper-based feature selection, but Kalousis et al. (2005) were the first to discuss stability in depth, independent of the particular feature selection algorithm. They defined stability as follows:

> *"We define the stability of a feature selection algorithm as the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution P(X,C). Stability quantifies how different training sets affect the feature preferences."* (Kalousis et al., 2005, pg 2)

Kalousis et al. (2005) provided an excellent review of the issues, which we will not repeat here. One important point is how the feature preferences are represented—as a ranking, a weighting or a subset. Since any ranking or weighting can be thresholded to obtain a subset (which is often the case), the scope of this particular article is the stability of feature *subset* selection, with the other representations left for future work. The seminal work of Kalousis was followed by a flurry of publications in application areas where stability turns out to be critical, such as microarray classification (Davis et al., 2006), molecular profiling (Jurman et al., 2008) and linguistics (Wichmann and Kamholz, 2008). But, perhaps more interesting for this paper, there was also a flurry of *methodological* papers, addressing how best to quantify stability.

1.1 The Problem: How to Quantify and Estimate Stability

The *measurement* of stability is important, as it addresses a fundamental question in data science—*how much can we trust an algorithm?* If tiny changes to initial conditions result in significantly different conclusions, perhaps we should not trust the output as reflective of the true underlying mechanism. This is important, not just for pure interest's sake in machine learning, but a true interdisciplinary challenge. In biomedical fields, this is a proxy for *reproducible research* (Lee et al., 2012) indicating that whatever biological features the algorithm has found are likely to be a data artefact, not a real clinical signal worth pursuing with further resources. Jurman et al. (2008) argue that having a *stable* selected gene set is equally important as their predictive power, while Goh and Wong (2016, pg 1) state:

> *"Identifying reproducible yet relevant features is a major challenge in biological research.*[...] *We recommend augmenting statistical feature selection methods with concurrent analysis on stability and reproducibility to improve the quality of the selected features prior to experimental validation."*

This is the intuitive concept and motivation to study stability. However intuitive, precisely *quantifying* it has proven somewhat challenging. In a literature search, conducted in early 2018, we identified at least 15 different measures used to quantify the stability of feature selection algorithms (Dunne et al., 2002; Shi et al., 2006; Davis et al., 2006; Kalousis et al., 2007; Krízek et al., 2007; Kuncheva, 2007; Yu et al., 2008; Zucknick et al., 2008; Zhang et al., 2009; Lustgarten et al., 2009; Somol and Novovičová, 2010; Guzmán-Martínez and Alaiz-Rodríguez, 2011; Wald et al., 2013; Lausser et al., 2013; Goh and Wong, 2016). Most of these

were justified and evaluated, though there has been little cross-comparison. The question arises: which should we trust, in which situation? If we do not understand the properties of these measures, it leads to a questionable interpretation of the stability values obtained, and questionable research in general. As acknowledged by Boulesteix and Slawski (2009), a multiplicity[1] of methods for stability assessment may lead to publication bias—in that researchers may (hopefully unintentionally) be drawn toward the metric that reports their feature selection algorithm as more stable. Furthermore, rarely do authors acknowledge that the stability value obtained is *an estimate of a true stability*, based on the number of feature sets sampled. Any measure is an *estimator of an underlying random variable*—therefore we should be able to discuss statistical concepts such as the population parameter being estimated and the convergence properties of the estimator. In this paper, we provide such an estimator and a theoretical analysis of its properties.

1.2 Our Approach to the Problem

Our approach to this problem is to propose a small set of properties, describing desirable behaviours from a stability measure. We will argue that the properties are generic enough to be desirable in all reasonable feature selection scenarios, and that they are critical for useful comparison and interpretation of stability values. We proceed to prove whether the 5 properties hold for each of 15 measures already proposed in the literature, and find no measure satisfying all. We then propose a novel measure for which all properties provably hold, in Section 4. The proposed measure has several desirable characteristics which distinguish it from previous proposals:

1. It is based on 5 well-defined properties (Section 3) which we will argue are essential requirements for a stability measure in most (if not all) feature selection scenarios.

2. It has a clean statistical interpretation in terms of the sample variance of a set of Bernoulli variables. The clean interpretation allows us to derive a set of tools for practitioners including

   - confidence intervals for the true stability;
   - a hypothesis test to check if the true stability is above a user-defined threshold;
   - a hypothesis test to compare the stability of two algorithms on a given data set.

3. It is a proper generalization of several existing measures (and therefore the statistical tools we develop are also applicable to those measures).

4. It is computable in linear time as opposed to quadratic, as is the case for most measures in the literature.

5. Given the theoretical and computational properties above, it can reliably be used to select hyperparameters for feature selection algorithms, such as LASSO or Stability Selection (Meinshausen and Bühlmann, 2010).

1. The R package `OmicsMarkeR` provides 7 different options for measuring stability, with no guidance for which to use, in which situation (c.f. `www.rdocumentation.org/packages/OmicsMarkeR`).

In the following sections we explain our framework, first embarking on a brief review of existing literature. For a more thorough review and an extended version of this work, refer to Nogueira (2018). We also provide:

- The code in R and Matlab at `github.com/nogueirs/JMLR2018` for the proposed measure and associated statistical tools. The code for all experiments is also available, enabling reproducible research.
- A Python package and a demonstration notebook using the package at `github.com/nogueirs/JMLR2018/tree/master/python/`
- A demonstration web page at `www.cs.man.ac.uk/~gbrown/stability`

2. Background

We assume a data set of n examples $\{\mathbf{x}_i, y_i\}_{i=1}^n$ where each $\mathbf{x}_i$ is a d-dimensional feature vector and y_i is the associated label. The task of feature subset selection is to identify a subset of the dimensions, of size $k < d$, that conveys the maximum information about the label y. The challenge of feature selection can be tackled in various ways, commonly grouped in three families: filters, wrappers, and embedded methods (Guyon and Elisseeff, 2003). Filters assign a score to a feature subset (or a feature) based on statistics of the data, *independently* of any particular classifier—for example mutual information-based methods. Wrappers, on the other hand, evaluate a subset based on an error criterion (and therefore are classifier-specific). Because there are $2^d - 1$ possible feature subsets, filters and wrappers often use a search procedure such as a forward or backward search to only evaluate some of the subsets. Finally, embedded methods sit in between these two, choosing a subset as an integral part of learning a prediction model—for example LASSO and other penalized likelihood methods. The output of a feature selection algorithm is therefore either a weighting (or scoring) on the features, a ranking on the features or a subset of the features. Sorting the weights naturally gives a ranking on the features, and selecting the top-k ranked features gives a subset of the features. This way, the output of any weighting or ranking feature selection technique can be treated as a subset selection one (while the reverse is not true). The input to any given procedure is the data set, which itself is assumed to be a finite sample from a generating distribution. If the sample varies, logically the selected feature subset may vary—this variation is the *stability*.

It is important to note *why* instability may occur, and what is commonly done about it—the sources of, and solutions to instability. Several authors study how stability is influenced by data characteristics, such as noise (Shanab et al., 2011), data dimensionality, sample size (Alelyani, 2013), imbalance of the data set (Dittman et al., 2012) or feature redundancy (Gulgezen et al., 2009; Wald et al., 2013). Additionally, several works have been proposed to *increase* stability. These include variance reduction frameworks (Han and Yu, 2012), sample weighting (Yu et al., 2012), ensemble feature selection (Ditzler et al., 2015; Saeys et al., 2008) and multi-objective optimization (Baldassarre et al., 2017; Gulgezen et al., 2009; Kalousis et al., 2007).

It can be noted that each of these works, by definition, mandates the authors to *measure* stability—to know whether they have increased it, or decreased it. A typical approach to

measure stability is to first take M bootstrap samples of the provided data set, to apply feature selection to each one of them, and then to measure the variability in the M feature sets obtained. Approaches other than taking bootstraps have been considered, such as noise injection (Wald et al., 2012b; Altidor et al., 2011) or random subsampling (Wald et al., 2012a). However, in this article, we aim at measuring the variation with respect to data sampling. For this reason, we adopt the bootstrap approach due to its well understood properties and familiarity to the community.

Let $\mathcal{Z} = \{s_1, ..., s_M\}$ be a collection of feature sets, where each s_i is a subset of the d features and let a stability measure $\hat{\Phi}$ be a function taking as input $\mathcal{Z}$ and returning a stability value[2]. How can we define $\hat{\Phi}$ so that it measures the variability of the feature sets in $\mathcal{Z}$? In the literature, we find two approaches to this problem—the *similarity-based* approach and the *frequency-based* approach that we introduce in the next two sections.

2.1 Similarity-based Measures

First introduced by Dunne et al. (2002), the similarity-based approach consists in defining $\hat{\Phi}$ as the average pairwise *similarity* between the $M(M-1)$ possible pairs of feature sets in $\mathcal{Z}$, that is

$$\underbrace{\hat{\Phi}(\mathcal{Z})}_{\text{Stability measure}} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \underbrace{\phi(s_i, s_j)}_{\textit{Similarity} \text{ measure}} ,$$

where ϕ is a function taking two feature sets as input and returning a similarity value. The more the feature sets in $\mathcal{Z}$ are *similar* to each other on average, the larger the value of $\hat{\Phi}(\mathcal{Z})$ will be. Therefore, the definition of $\hat{\Phi}$ and its properties critically depend on the choice of a similarity measure ϕ. Dunne et al. (2002) proposed to use the relative Hamming distance between two feature sets as a measure of their similarity. Since it was introduced, this approach gained popularity in the literature and we found to date 8 other proposals of similarity measures to be used in this context (c.f. Appendix 2.1 for more details). Kalousis et al. (2007) proposed to use the Jaccard index, Yu et al. (2008) the Dice-Sørenson index, Zucknick et al. (2008) the Ochiai index and Shi et al. (2006) the *POG* index (*Percentage of Overlapping Genes*). Kuncheva (2007) analysed some of the existing stability measures and demonstrated that they were behaving in undesirable ways. It pioneered the property-based approach, proposing a new similarity measure based on a set of 3 properties, proven to be essential to the correct comparison and interpretation of stability values. Because of its well-known properties, the proposed measure was used in many works that followed. Nevertheless, Kuncheva's measure was only defined for feature selection algorithms that guarantee to return a constant number of features. Other authors proposed to extend this measure to more general scenarios where the number of features selected is not pre-determined by the user (Lustgarten et al., 2009; Wald et al., 2012b). Nevertheless, we will see in Section 3 that the proposed extensions somehow lose some of the desirable properties.

2. We include the 'hat' notation $\hat{\Phi}$ to acknowledge that the value is an *estimate* of an underlying quantity, dependent on the sample size M.

2.2 Frequency-based Measures

An alternative representation for a feature set is to regard the feature choices as a binary string of length d, where a 1 at the f^{th} position means feature X_f has been selected in the set and a 0 means it has not been selected. This representation has driven the frequency-based approach, where one can measure stability by (for example) looking at the frequencies of selection of each feature over the M feature sets. The collection of the M feature sets can therefore be modelled as a binary matrix $\mathcal{Z}$ of size $M \times d$, where a row represents a feature set and a column represents the selection of a given feature over the M repeats as follows

$$\mathcal{Z} = \begin{pmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,d} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ z_{M,1} & z_{M,2} & \cdots & z_{M,d} \end{pmatrix}.$$

In the remainder of this paper, we will denote the observed frequency of selection of feature X_f by $\hat{p}_f = \frac{1}{M}\sum_{i=1}^{M} z_{i,f}$, the mean of the f^{th} column of $\mathcal{Z}$. We also model the selection of the f^{th} feature by a Bernoulli variable Z_f with unknown mean p_f. Since these two alternative representations are equivalent, with a slight abuse of notation, we will interchangeably denote by $\mathcal{Z}$ a collection of M feature sets or the matrix of M binary vectors. Intuitively, frequencies closer to 0 or 1 will indicate higher stability as it will mean that a feature is either selected on almost all M feature sets or on almost none of them; but as we will later see, defining a measure in this category can also prove to be challenging. In total, we identified 6 stability measures in this category (c.f. Appendix A.2 for more details), which are all either a function the observed frequencies of selection of each feature (Davis et al., 2006; Goh and Wong, 2016; Guzmán-Martínez and Alaiz-Rodríguez, 2011) or of the observed frequencies of selection of each feature set (Krízek et al., 2007).

3. A New Set of Properties for Stability Measures

Given the variety of stability measures published, it is sensible to ask whether any one is more valid than the others. This seems somewhat of a philosophical question—what does it mean for one stability measure to be more "correct" than another? To answer this, we adopt the perspective that a measure should (1) provably obey certain *properties* that are desirable in the domain of application and (2) provide capabilities that other measures do not. In this section, we aggregate and generalize the requirements of the literature into a set of 5 properties[3], given in Table 1, that we will show to be critical to the interpretation and comparison of stability values. For each property, we also study which one of the existing measures satisfy the property and we summarize our findings in Table 2. Our results show that none of the existing measures possess all 5 properties, which leads us to Section 4.1 where we propose a new stability measure.

3. As opposed to what has been done in some previous works (Kuncheva, 2007), these properties are functions of the *stability* measure rather than the *set similarity* measure—this will allow to compare the properties of both pairwise and non-pairwise stability measures in a single framework.

1. *Fully defined.* The stability estimator $\hat{\Phi}$ should be defined for any collection $\mathcal{Z}$ of feature sets, thus allowing for the total number of features selected to vary.
2. *Strict Monotonicity.* The stability estimator $\hat{\Phi}$ should be a strictly decreasing function of the sample variances s_f^2 of the variables Z_f.
3. *Bounds.* The stability $\hat{\Phi}$ should be upper/lower bounded by constants not dependent on the overall number of features or the number of features selected.
4. *Maximum Stability* $\leftrightarrow$ *Deterministic Selection.* A measure should achieve its maximum if-and-only-if all feature sets in $\mathcal{Z}$ are identical.
5. *Correction for Chance.* Under the Null Model of Feature Selection H_0, the expected value of $\hat{\Phi}$ should be constant.

Table 1: Proposed Properties for a Stability Measure.

3.1 Fully Defined

The first property, *Fully Defined*, is that a stability measure $\hat{\Phi}$ should be able to cope with any collection $\mathcal{Z}$ of feature sets. We observed that some of the stability measures do not obey this property. More specifically, the measures proposed by Kuncheva (2007), Krízek et al. (2007), Guzmán-Martínez and Alaiz-Rodríguez (2011) and Lausser et al. (2013) are only defined for a feature selection algorithm that would always return a constant number of features (c.f. definitions in Appendix A). Stability measures not having this property will not be defined for a wide class of feature selection algorithms, such as $L1$-regularization, and therefore such stability measures cannot be used to compare the stability of feature selection algorithms of different types.

3.2 Strict Monotonicity

By definition, all 9 similarity measures proposed to quantify stability are a strictly increasing function of the size of the intersection $|s_i \cap s_j|$ between the two sets s_i and s_j given as input (c.f. Appendix A.1). Kuncheva (2007) explicitly states this as a required property for a similarity measure. This property is implicitly defining what similarity (and therefore stability) is. Since the stability is defined as the average pairwise similarities, in the more general case, this property would naturally translate to: *For a given collection of feature sets $\mathcal{Z}$, the stability $\hat{\Phi}$ should be an increasing function of the average pairwise intersection size* $\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} |s_i \cap s_j|$. Even though this property can be applicable to any stability measure (as it is not phrased in terms of a similarity measure any more), it is not straightforward to comprehend the meaning of this property, neither to verify if a stability measure possesses this property, especially for frequency-based measures. We can therefore wonder what this does translate to in the frequency-based representation and how can

we verify that non-pairwise measures possess this property? Theorem 1 bridges the two approaches and justifies our second property, *Strict Monotonicity*[4].

Theorem 1 *The average pairwise intersection between the M feature sets is as a linear function of the sample variances of the selection of each feature. More precisely,*

$$\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} |s_i \cap s_j| = \bar{k} - \sum_{f=1}^{d} s_f^2, \tag{1}$$

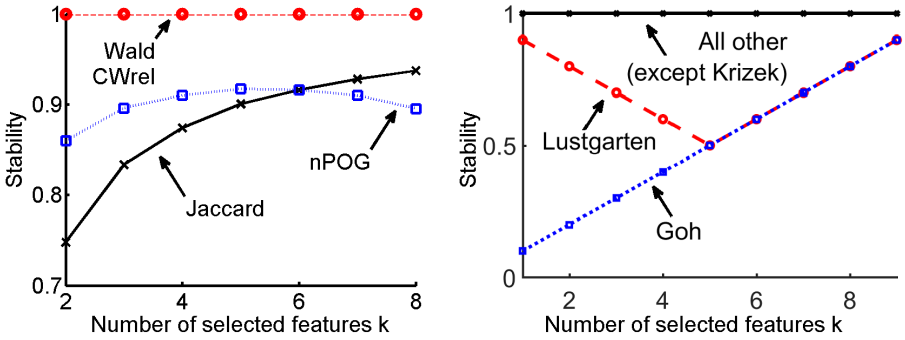
where $s_f^2 = \frac{M}{M-1} \hat{p}_f (1 - \hat{p}_f)$ is the unbiased sample variance of the selection of the f^{th} feature[5] *and where $\bar{k} = \frac{1}{M} \sum_{i=1}^{M} \sum_{f=1}^{d} z_{i,f}$ is the average number of features selected over the M feature sets.*

As we can see from Theorem 1, phrasing Monotonicity in terms of the average pairwise intersections is equivalent to phrasing it in terms of the average variance of the selection of each feature, which led to the definition of this property as stated in Table 1. This property defines *what* a stability measure should measure, rather than stating a necessary condition for a stability measure; therefore other proposals could be made for this purpose. For this reason, we will not discard the measures not having this property. However, it is interesting to note that, as we can see in Table 2, most stability measures have this property, showing some agreement upon the definition of stability across the literature, even if never stated as such. In some way, we can say that most existing measures of the literature implicitly aim at measuring the same quantity. In summary, we showed that: (1) interestingly, most stability measures of the literature possess this property, even though they were not explicitly built to that end, (2) the variance of the selection of each feature is an intuitive and simple way of measuring the variability in the choice features and (3) such a definition will allow us to derive a statistical framework for stability estimates, as we will later see in Section 4.2.

3.3 Bounds

This property was stated as necessary in several works. Somol and Novovičová (2010); Zucknick et al. (2008); Guzmán-Martínez and Alaiz-Rodríguez (2011) require a stability measure to be bounded by constants, while Kuncheva (2007) express the same requirement but for a similarity measure. Some of the measures such as Krízek et al. (2007) do not possess this property, since its maximum value depends on the number of features selected k and on the total number of features d (c.f. Appendix A.2). Unbounded measures do not allow for meaningful interpretation of stability values across problems or for different number of features selected, which can be restrictive in many applications.

4. To clarify: strict monotonicity is such that for a function g defined on a set D_g, g is a strictly monotonically (decreasing) function if $\forall x_1, x_2 \in D_g, x_1 < x_2 \Rightarrow g(x_1) > g(x_2)$. A counter-example showing the need for *strict* monotonicity (as opposed to monotonicity only) is to take any constant function: it will be monotonic as it will be an non-decreasing function of $|s_i \cap s_j|$ but cannot be interpreted as the similarity between two feature sets.
5. This expression of the sample variance is derived from a Bernoulli distribution.



(a) Forward implication (b) Backward implication

Figure 1: Illustration of the Maximum property. On the left, demonstration that Wald's measure and CW_{rel} (Somol and Novovičová, 2010) violate the forward implication. On the right, demonstration that Lustgarten's (Lustgarten et al., 2009) and Goh's (Goh and Wong, 2016) measures violate the backward implication[7].

3.4 Maximum Stability ↔ Deterministic Selection

For meaningful *interpretation* of a stability measure and comparison across problems, the range of values of a stability measure should be known and finite. Kuncheva (2007) states that IF two feature sets s_i and s_j are identical THEN their *similarity* is maximal, a desirable behaviour. Similarly, Guzmán-Martínez and Alaiz-Rodríguez (2011) also require that $\hat{\Phi}(\mathcal{Z})$ reaches its maximum whenever all the feature sets in $\mathcal{Z}$ are identical. We make this requirement a bi-implication, which translates in the general case to: $\hat{\Phi}(\mathcal{Z})$ reaches its maximum, *if-and-only-if* all feature sets in $\mathcal{Z}$ are identical, as stated in Table 1. We illustrate the need for the bi-implication below.

We used two scenarios, distinguishing the forward implication from the backward implication. First, we generated a collection of feature sets $\mathcal{Z}$ in which half of the feature sets are $\{X_1, ..., X_k\}$ and the other half are $\{X_1, ..., X_{k-1}\}$. Since there is clearly some variation in the features selected, the selection is not deterministic and we would want stability values not to be equal to their maximum. We plotted stability values against k for $d = 10$ and $M = 100$ in Figure 1a for some stability measures. We can see that Wald's measure (Wald et al., 2013) and CW_{rel} measure (Somol and Novovičová, 2010) still return their maximum value of 1. Therefore, these two measures violate the forward implication of this property. Second, we generated a collection of feature sets $\mathcal{Z}$ in which the same k features are selected on every repeat (for $d = 10$). Since the feature selection is now deterministic, we would want all stability values to be equal to their maximum. We plotted the stability values against the number of features selected k in Figure 1b. Even though the selection is completely deterministic, we can see that Lustgarten's and Goh's measure takes variable stability values depending on the number of selected features k. This shows that the two measures violate the backward implication of this property.

7. We ignored Krizek's measure in the right sub-figure as it is the only measure for which lower values correspond to higher stability and therefore, it should reach its minimum here instead of its maximum.

3.5 Correction for Chance

The fifth property, *Correction for chance*, was a novel concept introduced in the field by Kuncheva (2007) (but already existing in the statistical literature, c.f. Berry et al., 2016). This states that whenever we have *independently drawn subsets* at random, the stability value should be constant in expectation. This property was also later required by Lustgarten et al. (2009); Zhang et al. (2009); Guzmán-Martínez and Alaiz-Rodríguez (2011). When the number of features selected is constant, the notion of purely *random* feature selection is intuitive: it means that on each sample, given the number of selected features k, each one of the $\binom{d}{k}$ feature sets is equally likely to be chosen by the procedure. Now, let us take the case when the procedure does *not guarantee* to return a constant number of features, and thus produces a collection $\mathcal{Z}$ of feature sets, of varying size. We can still define the concept of *randomness* in that case: given the cardinality k_i of the i^{th} set, each one of the $\binom{d}{k_i}$ possible feature sets has an equal probability of being selected. We note that this is the assumption (sometimes implicitly) made by different authors using the concept of correction for chance, e.g. Kuncheva (2007); Lustgarten et al. (2009); Zhang et al. (2009); Guzmán-Martínez and Alaiz-Rodríguez (2011). Since we will use this concept multiple times in the remainder of the paper, we formalize this in Definition 2 and will refer to this assumption as the *Null Model of Feature Selection*.

Definition 2 (The Null Model of Feature Selection H_0) *We define the Null Model of Feature Selection as the situation where all possible permutations of the observed bits in a row of $\mathcal{Z}$ are equally likely. In other words, for all rows i in $\mathcal{Z}$, all subsets of size k_i have an equal probability of being drawn.*

Our final property, *Correction for chance* is that: under the Null Model of Feature Selection H_0, the expected value of $\hat{\Phi}$ should be constant, which for convenience we set to 0 (as done by other authors). The motivation behind such a property is that we would not want a stability measure to reflect the similarity between feature sets that might occur by chance, but only that due to the systematic decision-making of the feature selection algorithm. For illustrative purposes, we reproduced the experiment of Kuncheva (2007) in Figure 2. Let us assume that a feature selection procedure **randomly** selects k features out of $d = 10$ features and that we estimate the stability based on the $M = 100$ feature sets obtained for different values of k. As we can see, even though the feature selection procedure is random and therefore corresponds to a fully unstable situation (i.e. we are under the Null Model of Feature Selection H_0), some stability measures are strongly biased by the feature set size. For instance, we can see that using the Dice similarity measure, the stability systematically increases with the number of features selected, thus being in favour of larger feature sets. On the other hand, Kuncheva's measure that is corrected for chance gives a stability close to 0, no matter what the feature set size is as it is corrected for chance. Non-corrected measures make stability values neither comparable nor interpretable in different settings. To prove whether a measure has this property or not, we derived the value of $\mathbb{E}\left[\hat{\Phi}|H_0\right]$ for each of the existing measures in Appendix C.5 and reported the results in Table 2.

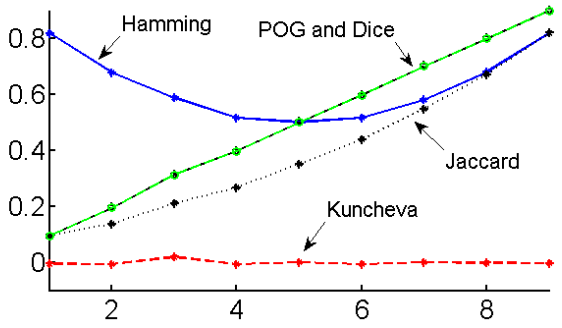


Figure 2: Illustration of the Correction-for-chance property. Stability values using Hamming, *POG*, Jaccard, Dice and Kuncheva similarity measures against the number k of features selected for $M = 100$ repeats.

Name	**Fully defined**	**Monotonicity**	**Bounds**	**Maximum**	**Correction**
Hamming	✓	✓	✓	✓	
Jaccard	✓	✓	✓	✓	
Dice	✓	✓	✓	✓	
Ochiai	✓	✓	✓	✓	
POG	✓	✓	✓	✓	
Kuncheva		✓	✓	✓	✓
Lustgarten	✓	✓	✓		✓
Wald	✓	✓			✓
nPOG	✓	✓		✓	✓
Goh	✓		✓		
Davis	✓		✓		
Krízek				✓	
Guzmán			✓	✓	✓
CW_{rel}	✓	✓	✓		
Lausser		✓	✓	✓	

Table 2: Properties of Stability Measures proposed in the literature 2002–2018. For each of the 15 measures, and for each of the 5 properties, we prove which measure satisfies which property — full proofs available in Appendix C.

3.6 Summary

In the literature, many authors advocated for the need of stability measures with well-known properties. In this section, we aggregated and generalised the properties stated as desirable in the literature into a set of desirable 5 properties[8]—applicable to any stability measure (whether similarity-based or not) and to any any feature selection algorithm (whether it selects a constant number of features or not)—in Table 1. Then, Table 2 summarizes the properties of each one of the stability measures. The first 5 measures are similarity-based measures and as we can see, they all possess all properties except Correction for chance. This weakness, noticed by Kuncheva (2007), gave rise to corrected-by-chance similarity-based measures, which are the 4 following measures of Table 2. As we can see, even though these 4 proposals all possess the Correction-for-chance property, they somehow lost some of the other desirable properties. Finally, the 6 following measures in the table are the frequency-based measures, which are more diverse in terms of the set of properties they satisfy. We note that even though CW_{rel} (Somol and Novovičová, 2010) does not possess the Correction-for-chance property, when the number of features selected is constant, we show in Appendix C.5 that it is asymptotically (as M approaches infinity) corrected for chance. We can therefore conclude that none of the stability measures in the literature possess all desired properties (even when discarding the Monotonicity property). Based on these results, we derive a novel stability measure in the next section, that not only has all desired properties, but also will allow us to develop a statistical framework for the quantification of stability.

4. A Novel Stability Measure

In this section, we propose a measure of stability which provably attains all desirable properties as discussed in the previous section. We recognise the stability measure $\hat{\Phi}$ as an estimator of a random variable, and aim to make explicit the corresponding population parameter Φ. By identifying the sampling distribution, we are able to provide tools for practitioners such as confidence intervals and hypothesis tests. This provides confidence in what the true value may be, and allows us to reliably compare stability across feature selection procedures. In the remainder of the paper, we refer to the stability measure as the *stability estimator* and to the parameter being estimated as the *population stability* or the *true stability*.

4.1 Proposed Stability Estimator

As required by the Monotonicity property, the stability should be a strictly decreasing function of the variances of the selection of each feature—for simplicity, we just negate the mean of the sample variances. As required by Correction for Chance, we rescale it by its expected value under the Null Model of Feature Selection. Finally for convenience of

8. Another property sometimes stated as desirable in the literature concerns the symmetry of the similarity measure ϕ (i.e. $\phi(s_i, s_j) = \phi(s_j, s_i)$) (Alelyani, 2013; P. and Perumal, 2016; Zucknick et al., 2008). We note that for any non-symmetric similarity measure ϕ, taking the arithmetic mean of $\phi(s_i, s_j)$ and $\phi(s_j, s_i)$ gives a symmetric similarity measure holding the same average pairwise value. Thus, the symmetry of a similarity measure ϕ is of little importance when comparing the properties of the corresponding stability measure $\hat{\Phi}$.

interpretation, we ensure that (asymptotically) the value is in the range $[0, 1]$ by taking one minus the resulting expression. Making use of Theorem 3 and by linearity of the expectation, this gives us our stability estimator as given in Definition 4.

Theorem 3 *Under the Null Model of Feature Selection H_0, for all f, the expected value of the sample variance of Z_f is $\mathbb{E}\left[s_f^2|H_0\right] = \frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)$.*

Definition 4 (Novel Measure) *We define the stability estimator as*

$$\hat{\Phi}(\mathcal{Z}) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\mathbb{E}\left[\frac{1}{d}\sum_{f=1}^{d} s_f^2|H_0\right]} = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)}, \tag{2}$$

where $s_f^2 = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$ is the unbiased sample variance of the selection of the f^{th} feature.

We now verify that the proposed measure possesses all 5 properties. First, by construction, we can see that the measure is defined for all collections $\mathcal{Z}$, unless the denominator $\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)$ is equal to zero. This happens whenever $\bar{k} = 0$ or $\bar{k} = d$, which are the two limit cases in which the algorithm does not select any features over the M feature sets in $\mathcal{Z}$ or selects all of the features on every feature set of $\mathcal{Z}$. Since by definition, a feature selection procedure will always select a non-empty proper subset of the set of all available features, the first property *Fully defined* holds for the proposed definition of $\hat{\Phi}$. Second, by construction, $\hat{\Phi}$ is a linear function of the sample variances s_f^2 with a strictly negative slope. Therefore, the Monotonicity property holds for $\hat{\Phi}$. Third, to conform to the Bounds property—by inspection, one can see the numerator and denominator in the measure are positive quantities, therefore the upper bound is 1. For the lower bound, Appendix D shows that the minimum of $\hat{\Phi}$ is equal to $-\frac{1}{M-1}$. This shows that $\hat{\Phi}$ is bounded by -1, but is asymptotically bounded by 0 (as M approaches infinity). Fourth, let us assume for some $\mathcal{Z}$, the measure achieves its maximum, $\hat{\Phi}(\mathcal{Z}) = 1$. This state is equivalent to $\sum_{f=1}^{d} s_f^2 = 0$. Since the variance is a positive quantity, this is in turn equivalent to $s_f^2 = 0$ for all f. This case corresponds to the situation where each column of $\mathcal{Z}$ contains either all 1s or all 0s. Thus $\hat{\Phi}$ is equal to its maximum if-and-only-if all feature sets in $\mathcal{Z}$ are identical. Fifth, under the Null Model of Feature Selection H_0, by linearity of the expectation and using Theorem 3, $\mathbb{E}\left[\hat{\Phi}|H_0\right] = 0$. Therefore, the proposed measure is corrected for chance.

Theorem 5 shows that the proposed measure of stability $\hat{\Phi}$ is a generalization of some other widely used measures to feature sets of varying cardinality, hence being consistent with previous literature. As a result, all statistical tools of Section 4.2 are also valid for Kuncheva's measure and for the other equivalent measures, hence unifying some of the theory on the measurement of stability. The reformulation of Kuncheva's measure using our definition also provides a computational advantage, being $O(Md)$ instead of $O(M^2d)$ (which is the computational complexity of all pairwise measures).

Theorem 5 *When the number of features selected is constant:*

- *The stability estimator $\hat{\Phi}$ is equal to the stability measures proposed by Kuncheva (2007), Wald et al. (2013) and to nPOG (Zhang et al., 2009).*

- *The stability estimator $\hat{\Phi}$ and CW_{rel} (Somol and Novovičová, 2010) are asymptotically equivalent.*

In the next section, we provide the population parameter Φ estimated by $\hat{\Phi}$, that is the *true* stability of the feature selection procedure considered.

4.2 Statistical Tools

In the previous section, we proposed a new stability measure that possesses all desirable properties and is a generalization of some of the existing measures of the literature. We can also ask how it relates to other parts of the literature and which other fields deal with similar problems. This section demonstrates and exploits a relationship between stability and inter-rater agreement (Fleiss, 1971).

4.2.1 Viewing Stability as Inter-rater Agreement

Imagine a medical scenario—we have M doctors (more formally called *raters*) assigning a nominal category $\{1, 2, ..., q\}$ to each member of a set of d patients (called *subjects*). A useful indication of the agreement of the M raters is given by *inter-rater agreement coefficients*. We can view stability in this light—when the number of categories q is equal to 2, and each row of $\mathcal{Z}$ represents a rater, placing the d subjects into category 0 or 1. Interestingly[9], in this special case, we prove with Theorem 6 that a popular measure of inter-rater agreement, Fleiss' Kappa (Fleiss, 1971) reduces to our estimator, Definition 4. As a result, any statistical result previously derived for Fleiss' Kappa also holds for $\hat{\Phi}$. Using this relationship, we can use the work of Gwet (2008) which shows the asymptotic normality of Fleiss' Kappa, hence guaranteeing the validity of confidence intervals and hypothesis tests for large samples.

Theorem 6 *When there are only two categories (0/1), Fleiss' Kappa is equal to $\hat{\Phi}(\mathcal{Z})$.*

4.2.2 The Sampling Distribution of Stability

Let us assume each row of the matrix $\mathcal{Z}$ is an independent sample from the joint distribution $(Z_1, ..., Z_d)$, where Z_f is a Bernoulli variable with unknown population parameter p_f, where we make no assumption of independence between d covariates. In the original paper, Fleiss (1971) derives the variance of Fleiss' Kappa, but only when Φ is equal to 0, which is of little use in our case. Later on, Gwet (2008) provides a variance estimate and the asymptotic distribution of Fleiss' Kappa in the general case. In his work, Gwet (2008) assumes that the raters (samples) and subjects (features) are sampled from a larger population and then derives the variance due to the sampling of raters and the variance due to the sampling of subjects. Using the multivariate Central Limit Theorem and a linear approximation of $\hat{\Phi}$, Gwet (2008) shows that $\hat{\Phi}$ is asymptotically normal. Gwet (2008) also verifies the validity of this result for the construction of confidence intervals with Monte Carlo simulations. In our case, we assume that there is no sampling of the subjects and that the number of categories

9. Another interesting relationship that could be used in future work is that Fleiss' Kappa has also been linked to the Intra-Class Correlation Coefficient (ICC) in the binary case (Fleiss et al., 2004), so any result that applies to the ICC can also be applied to the proposed stability estimator.

$q = 2$. Under these assumptions, the variance due to the sampling of subjects derived by Gwet (2008) becomes zero and the asymptotic distribution of the stability estimator $\hat{\Phi}$ becomes the one given by Theorem 7.

Theorem 7 (Asymptotic Distribution) *As $M \to \infty$, the statistic $\hat{\Phi}$ weakly converges to a normal distribution, that is*

$$\frac{\hat{\Phi} - \Phi}{\sqrt{v(\hat{\Phi})}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

where:

- $\Phi = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} p_f(1-p_f)}{\bar{p}(1-\bar{p})}$ *is the mean of the estimator $\hat{\Phi}$ in which $\bar{p} = \frac{1}{d}\sum_{f=1}^{d} p_f$ is the average mean parameter of the d Bernoulli variables;*
- $v(\hat{\Phi}) = \frac{4}{M^2}\sum_{i=1}^{M}(\hat{\Phi}_{(i)} - \hat{\bar{\Phi}}_{(.)})^2$ *is an estimate of the variance of $\hat{\Phi}$ in which:*
  - $\hat{\Phi}_{(i)} = \frac{1}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)}\left[\frac{1}{d}\sum_{f=1}^{d} z_{i,f}\hat{p}_f - \frac{k_i\bar{k}}{d^2} + \frac{\hat{\Phi}}{2}\left(\frac{2\bar{k}k_i}{d^2} - \frac{k_i}{d} - \frac{\bar{k}}{d} + 1\right)\right]$;
  - *and $\hat{\bar{\Phi}}_{(.)}$ is the average value of $\hat{\Phi}_{(i)}$, that is $\frac{1}{M}\sum_{i=1}^{M}\hat{\Phi}_{(i)}$.*

This asymptotic distribution allows us to identify $\hat{\Phi}$ as being an estimator of an unknown population quantity Φ. In the remainder of the paper, we refer to $\hat{\Phi}$ as being the sample stability and to Φ as being the population (or *true*) stability. The asymptotic convergence shows that $\hat{\Phi}$ is a consistent estimator of the population stability Φ. This means that as M approaches infinity, we are assured that the stability estimator $\hat{\Phi}$ will converge in probability to the population stability Φ.

4.2.3 Confidence Intervals

The asymptotic convergence to a normal distribution allows us to derive approximate confidence intervals for the population stability Φ, given below in Corollary 8. Though the provided confidence intervals are only approximate, we will see in Section 5.2 that even for relatively small values of M, the given intervals still have a good coverage probability.

Corollary 8 (Confidence Intervals) *A $(1-\alpha)\%$-approximate confidence interval for Φ is*

$$\left[\hat{\Phi} - z^*_{\left(1-\frac{\alpha}{2}\right)}\sqrt{v(\hat{\Phi})}, \hat{\Phi} + z^*_{\left(1-\frac{\alpha}{2}\right)}\sqrt{v(\hat{\Phi})}\right],$$

where $z^*_{\left(1-\frac{\alpha}{2}\right)}$ is the inverse cumulative of a standard normal distribution at $1-\frac{\alpha}{2}$.

4.2.4 Hypothesis Testing

In a first scenario, let us assume a practitioner applies a feature selection procedure to M bootstrap samples, generating a matrix $\mathcal{Z}$ of size $M \times d$, and computes the stability estimate $\hat{\Phi}(\mathcal{Z})$. How can we know whether the true stability Φ is significantly greater than

a fixed value Φ_0? This can be defined formally in terms of a null hypothesis significance test.

Is the Population Stability Φ Greater than a Given Value Φ_0? In this case the hypothesis tested is

$$\begin{cases} H_0 : \Phi = \Phi_0 \\ H_1 : \Phi > \Phi_0 \end{cases}$$

Under H_0, $\Phi = \Phi_0$ and therefore the statistic $V_M = \frac{\hat{\Phi} - \Phi_0}{\sqrt{v(\hat{\Phi})}}$ is asymptotically standard normal (c.f. Theorem 7). Therefore we can apply a one-tail test as follows:

1. Compute the statistic V_M.
2. Reject H_0 if $V_M \geq z^*_{(1-\alpha)}$, where the critical value $z^*_{(1-\alpha)}$ is the $(1-\alpha)$th percentile of the standard normal distribution.

In addition, it is very common to compare stability values between algorithms. For example, Saeys et al. (2008) conclude that *"RELIEF is one of the less stable algorithms"* and *"Random Forests clearly outperform other feature selection methods regarding robustness"*. So given two stability estimates $\hat{\Phi}(\mathcal{Z}_1)$ and $\hat{\Phi}(\mathcal{Z}_2)$, can we conclude that the true stability of the first is significantly different than the second?

Do Two Feature Selection Algorithms Have Identical Stabilities? Let $\mathcal{Z}_1$ and $\mathcal{Z}_2$ be the output of two feature selection procedures. In this case, we wish to test the following hypothesis

$$\begin{cases} H_0 : \Phi_1 = \Phi_2 \\ H_1 : \Phi_1 \neq \Phi_2 \end{cases}$$

Using the asymptotic distribution of $\hat{\Phi}(\mathcal{Z}_1)$ and of $\hat{\Phi}(\mathcal{Z}_2)$ given by Theorem 7, we can derive Theorem 9. Using the given test statistic T_M, we reject H_0 if $|T_M| \geq \theta$, where θ is the $\left(1 - \frac{\alpha}{2}\right)^{th}$ percentile of the standard normal distribution.

Theorem 9 *The test statistic for comparing stabilities is*

$$T_M = \frac{\hat{\Phi}(\mathcal{Z}_2) - \hat{\Phi}(\mathcal{Z}_1)}{\sqrt{v(\hat{\Phi}(\mathcal{Z}_1)) + v(\hat{\Phi}(\mathcal{Z}_2))}}.$$

Under H_0, the statistic T_M asymptotically (as M approaches infinity) follows a standard normal distribution.

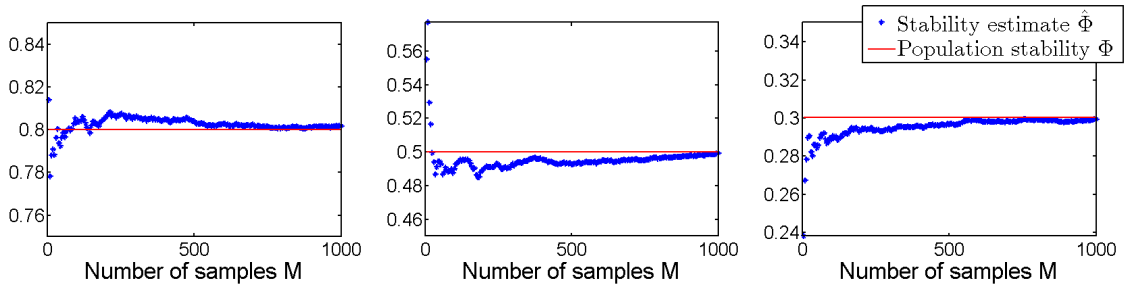
5. Empirical Validation of the Statistical Tools

Fleiss et al. (2004) propose a benchmark scale for interpretation of the value of Fleiss' Kappa. We will use the same scale for $\hat{\Phi}(\mathcal{Z})$, provided in Table 3. Stability values above 0.75 represent an excellent agreement of the feature sets beyond chance, while values below 0.4 represent a poor agreement between sampled feature sets.

In the remainder of this section, we verify the tools of the previous section, using toy data for a population stability Φ in each one of the categories. To be able to generate

Φ	**Strength of Agreement**
< 0.40	Poor
0.40 to 0.75	Intermediate to good
> 0.75	Excellent

Table 3: Benchmark scale for stability.



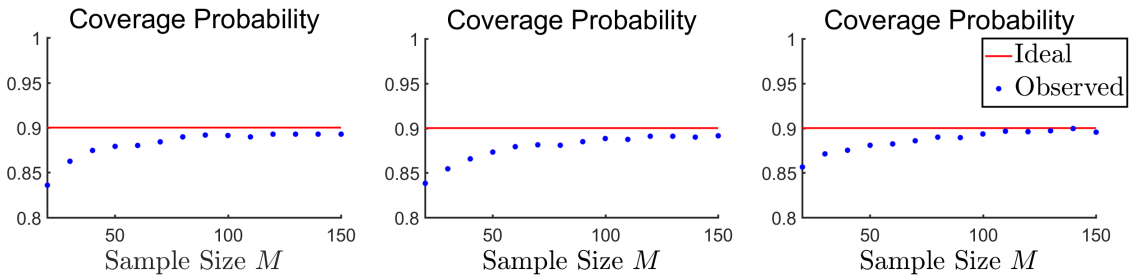
(a) Case 1: $\Phi = 0.8$ (b) Case 2: $\Phi = 0.5$ (c) Case 3: $\Phi = 0.3$

Figure 3: Consistency of the stability estimate $\hat{\Phi}(\mathcal{Z})$ for the 3 toy cases. As M increases, the value of $\hat{\Phi}(\mathcal{Z})$ gets closer to the population parameter Φ.

Bernoulli variables with a specified population stability Φ_0, we first need to chose d Bernoulli parameters $p_1, ..., p_d$ such that $\Phi = \Phi_0$. We picked 3 test cases for the values of Φ equal 0.8, 0.5 and 0.3 and for $d = 100$.

5.1 Validation of Consistency of the Estimator

In this section, we empirically show the consistency of the stability estimator $\hat{\Phi}$ in the 3 test cases described. In each case, we take M samples from the $d = 100$ Bernoulli variables with mean parameters $(p_1, ..., p_d)$. This gives us a binary matrix $\mathcal{Z}$ of size $M \times d$. We then plot the value of the stability estimate $\hat{\Phi}(\mathcal{Z})$ as we increase the number of samples M in Figure 3. As we can see, as the number of samples M increases, the value of $\hat{\Phi}(\mathcal{Z})$ approaches the true stability Φ. We chose similar scales for the 3 test cases. We observe that for relatively small values of M, (generally for $M \leq 10$), the absolute value of the difference $|\hat{\Phi}(\mathcal{Z}) - \Phi|$ is within 5% and for $M \leq 100$, within 1%. Of course, these cannot be used as a general rule of thumb for other chosen parameters p_f and d but it gives an idea of the rate of convergence of the stability estimates. In real applications, the population stability is unknown and we need tools to be able to determine which interval of values the population stability takes. This is the topic of the next section where we study the confidence intervals.



(a) Case 1: $\Phi = 0.8$ (b) Case 2: $\Phi = 0.5$ (c) Case 3: $\Phi = 0.3$

Figure 4: Coverage probabilities for the 3 test cases. We set the nominal level to be 0.90 (90%-confidence interval). The x-axis represents the sample size M and the y-axis the estimated coverage probability with 10000 repeats.

5.2 Validation of Confidence Intervals

The coverage probability of a confidence interval for Φ is the proportion of time the interval built from the data will actually contain the population stability Φ. To verify the results of Theorem 8 providing the confidence intervals, we adopted the following procedure:

1. Compute the population stability Φ using the true Bernoulli parameters $(p_1, ..., p_d)$.
2. Repeat 10,000 times:
   - Take M samples from the d variables $Z_1, ..., Z_d$ with mean $p_1, ..., p_d$.
   - Compute the $(1-\alpha)$-approximate confidence interval using Corollary 8.
3. The estimated coverage probability is the fraction of times (from 10,000) that the true stability Φ was within the confidence intervals.

If we had exact confidence intervals, we should have an exact coverage probability of $(1-\alpha)$. However, the confidence intervals derived are only approximate. Figure 4 gives the estimated coverage probabilities in the 3 test cases for $\alpha = 10\%$, i.e. a 90%-confidence interval.

As we can see, the estimated probability quickly approaches $(1-\alpha) = 0.9$ as expected. In Table 4, we further show the observed coverage probabilities in the 3 test cases for $M = 100$ and for different values of α. The same behaviour can be seen in this table: the values get very close to the expected coverage probability of $(1-\alpha)$. We observed the same behaviour for a larger number of features ($d = 10000$) and a same number of feature sets M.

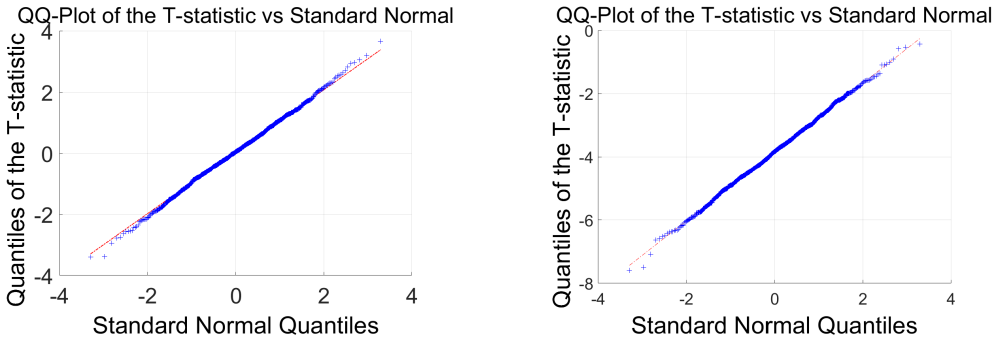
5.3 Validation of the Second Hypothesis Test

In this section, we verify the asymptotic distribution of the test statistic T_M as given by Theorem 9. To verify this result we proceed as follows:

1. Pick two set of parameters $p_1, ..., p_d$ with the desired values of Φ_1 and Φ_2.
2. Repeat 1000 times:

Case 1	$\Phi = 0.8$	98.5%	94.3%	89.0%
Case 2	$\Phi = 0.5$	98.6%	93.8%	89.0%
Case 3	$\Phi = 0.3$	98.6%	94.0%	89.3%
Ideal		99%	95%	90%

Table 4: Coverage probabilities for the 3 test cases with $M = 100$, $d = 100$, estimated via $10,000$ repeats for different nominal confidence intervals.



(a) $\Phi_1 = \Phi_2 = 0.8$ (b) $\Phi_1 = 0.8$ and $\Phi_2 = 0.75$

Figure 5: QQ-plots illustrating the convergence of the statistic T_M to a standard normal distribution when $\Phi_1 = \Phi_2 = 0.8$ [LEFT] and the convergence to a non-standard Gaussian when $\Phi_1 \neq \Phi_2$ [RIGHT]. We took $d = 100$, $M = 1000$ and 1000 repeats. We note that the range of values on the y-axis of the right plot is not the same as the one of the left plot.

- Take M samples from the d variables, for each of $\mathcal{Z}_1$ and $\mathcal{Z}_2$
- Compute the corresponding statistic T_M.

We then order the 1000 estimates of T_M and plot the quantiles against that of a standard normal distribution in a Quantile-Quantile plot (QQ-plot). Figure 5 provides the result for two test cases. In the left sub-figure, the population stabilities are taken to be identical (i.e. $\Phi_1 = \Phi_2$). In that situation, the QQ-plot shows that the quantiles of the test statistic T_M are identical to the ones of a standard normal distribution, which is the result we expected. On the right sub-figure, we chose different population stabilities (i.e. $\Phi_1 \neq \Phi_2$). In this case, the QQ-plot shows that the quantiles are still the ones of a normal distribution but with a mean different from 0. Indeed, if we have a closer look at the right sub-figure, we can see that the range of values taken on the y-axis are all negative. The observed median of the statistic T_M in that case is of -3.8 and the statistic takes values in the interval $[-7.6, -0.4]$. Therefore the statistic T_M is not standard normal in that situation.

6. Experiments

The experiments of this section illustrate how the tools presented in this paper can be used by practitioners to select hyperparameters with higher stability or to compare the stability of different feature selection algorithms. This section contains two sets of experiments[10]:

- Section 6.1 focuses on the LASSO and the Elastic Net, which are both regularized regression models that select features as part of the training process. The Elastic Net is known to yield more stable coefficients than the LASSO in the presence of redundant features (Zhou, 2013). This section shows that the proposed stability measure captures this and that choosing a good trade-off between stability and accuracy can reduce the number of irrelevant features in the model with negligible loss in accuracy.

- Section 6.2 focuses on a popular technique, *Stability Selection* (Meinshausen and Bühlmann, 2010) which we apply to LASSO. The proposed framework defines the stable set as the set of features being selected with high frequency across a set of regularizing parameters. We propose to look at the stability of the *stable set* for different hyperparameters and compare it to the stability of LASSO.

6.1 The Stability of $L1/L2$ Regularized Logistic Regression

In this section, we observe how the degree of redundancy in data can affect the stability of LASSO and Elastic Net, and how we can optimize hyperparameters so that both log-likelihood and stability are taken into account. We show that stability can help recover the *true* set of relevant features. To be able to control the set of relevant features and the redundancy between them, we use a synthetic data set as described in the next section.

6.1.1 Description of the Data Set

We use a synthetic data set (Kamkar et al., 2015)—a binary classification problem, with 2000 instances and $d = 100$ features, where only the first 50 features are relevant to the target class. Instances of the positive class are identically and independently drawn from a normal distribution with mean $\mu_+ = (\underbrace{1, ..., 1}_{50}, \underbrace{0, ..., 0}_{50})$ and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma^*_{50\times 50} & \mathbf{0}_{50\times 50} \\ \mathbf{0}_{50\times 50} & \mathbf{I}_{50\times 50} \end{bmatrix},$$

where $\Sigma^*_{50\times 50}$ is the matrix with ones on the diagonal and ρ (a parameter in $[0, 1]$) controlling the degree of redundancy everywhere else. The mean for the negative class is taken equal to $\mu_- = (\underbrace{-1, ..., -1}_{50}, \underbrace{0, ..., 0}_{50})$. The larger the value of ρ, the more the 50 relevant features will be correlated to each other.

10. You can reproduce all these experiments (in Matlab or R) with the code given at `github.com/nogueirs/JMLR2018`

6.1.2 Stability of LASSO

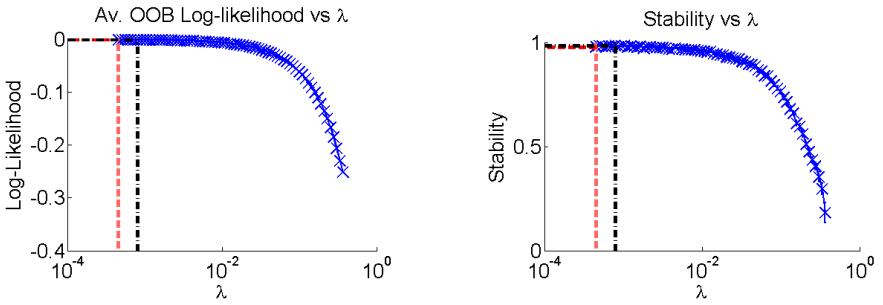
We use a $L1$-regularized logistic regression where λ is the regularizing parameter, influencing the amount of features selected—as λ increases, more and more coefficients are equal to zero and therefore less and less features are selected. We take 2000 samples and divide them into 1000 for model selection (i.e. to select the regularizing parameter λ) and 1000 for selection of the final set of features. The model selection set can be used simply to optimize error, or to optimize error/stability simultaneously—the experiments will demonstrate that the latter provides a lower false positive rate in the final selection of features. We study 4 degrees of redundancy: $\rho = 0$ (no redundancy, the features are independent from each other), $\rho = 0.3$ (low redundancy), $\rho = 0.5$ (medium) and $\rho = 0.8$ (high). We apply $L1$-logistic regression to $M = 100$ bootstrap samples of the data set. We then compute the average out-of-bag (OOB) log-likelihood[11] and the stability of the feature selection.

Figure 6 shows the average log-likelihood (left column) and the stability (right column) versus the regularization parameter λ for the 4 degrees of redundancy chosen. For each degree of redundancy, the pink dashed-line represents the parameter λ maximizing the likelihood and the black one represents the parameter maximizing the stability. In the case of no redundancy, we can see that these two parameters produce similar values of likelihood and stability. But, as we change the degree of redundancy, this is not the case. The result can most strongly be seen in Figure 6b, where $\lambda = 0.0051$ optimizes likelihood, but if we increase to $\lambda = 0.0187$, we sacrifice a negligible amount of likelihood for a quite significant increase in stability to $\hat{\Phi} = 0.63$.

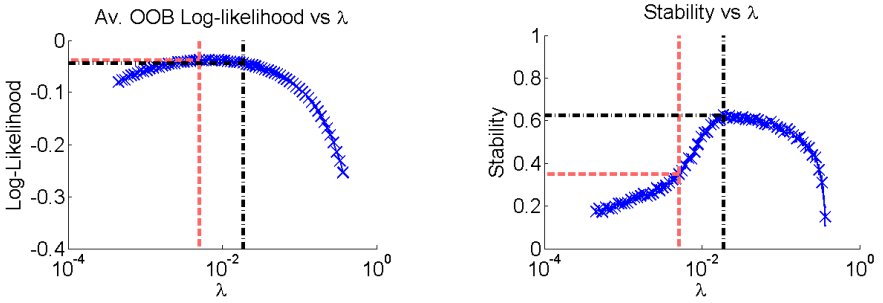
Figure 7 shows an alternative view of these results, plotting stability against likelihood. When there is no redundancy in the data (sub-figure (a)), stability seems to be an increasing function of the likelihood. For higher levels of redundancy, this results in at least *two* values of λ which achieve the *same likelihood*, but clearly one results in higher stability. In practice, to choose a hyperparameter λ, we have to pick a trade-off between likelihood of the model and stability. For this purpose, we can identify the pareto front of likelihood and stability, which is the set of points such that there is no other point with higher likelihood and higher stability. Hence, each point in the pareto front represents a different trade-off between likelihood and stability. Figure 8 summarizes the pareto fronts for the 4 degrees of redundancy. In a classic scenario, we would pick the value of λ that maximizes the likelihood only, which corresponds to the rightmost point in the figure. However, we can see that sacrificing a small amount of likelihood allows us to considerably increase stability. As we increase the degrees of redundancy, we see the best case for stability is lower. Nevertheless, all the points in a given pareto front have a similar likelihood and a similar misclassification rate. All these observations show that *stability can potentially be increased without loss of predictive power.*

We also observe that pursuing stability may help identifying the *relevant* set of features. Figure 9 gives the observed frequencies of selection $\hat{p}_f$ of each feature over the $M = 100$ bootstraps—where features $1-50$ are relevant, and $51-100$ are irrelevant in the case of high redundancy. We picked the value of λ that maximized the likelihood for the left sub-figure and a value of λ in the pareto front of stability and likelihood for the right sub-figure. On the right figure, all 50 irrelevant features have a frequency of selection equal to 0 (i.e. the

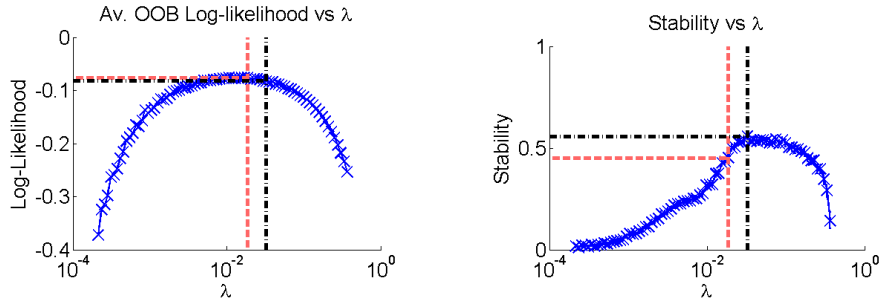
11. In all the presented results, the log-likelihood is rescaled by the number of examples n.



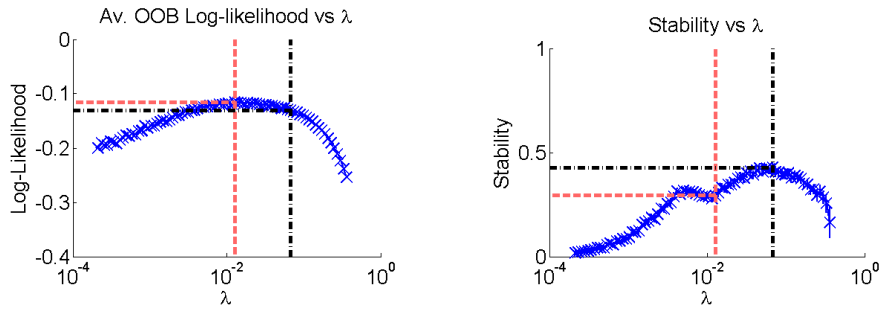
(a) No redundancy



(b) Low redundancy

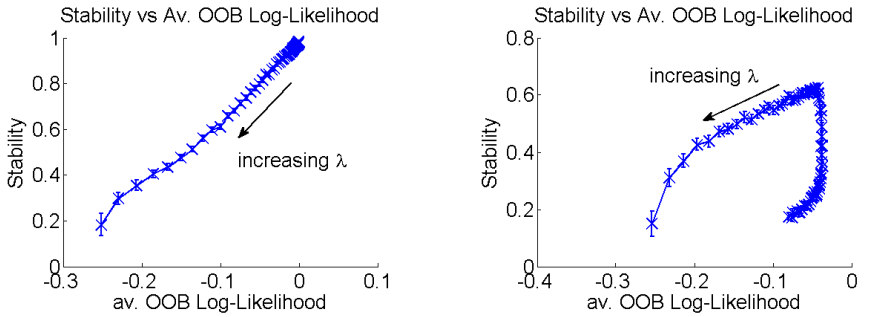


(c) Medium redundancy



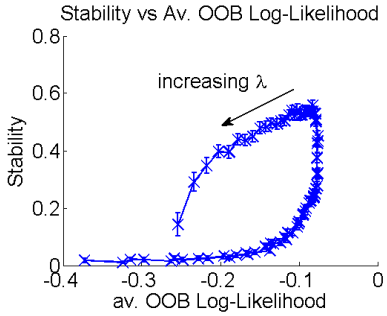
(d) High redundancy

Figure 6: Average OOB log-likelihood [left column] and stability [right column] against the regularizing parameter λ for 4 degrees of redundancy. For each degree of redundancy, the pink dashed-line corresponds to the λ value that maximizes the likelihood and the black one corresponds to the value of λ that maximizes stability. As we can see, by choosing latter parameter λ, we gain in stability with only a small loss in likelihood.

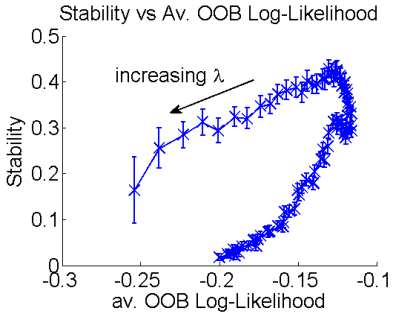


(a) No redundancy

(b) Low redundancy



(c) Medium redundancy



(d) High redundancy

Figure 7: Stability (with 95%-confidence intervals) against average OOB log-likelihood for 4 degrees of redundancy.

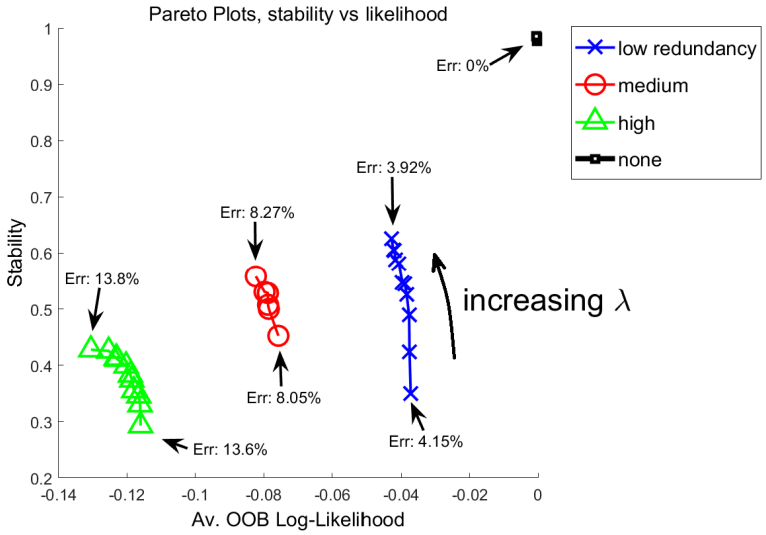
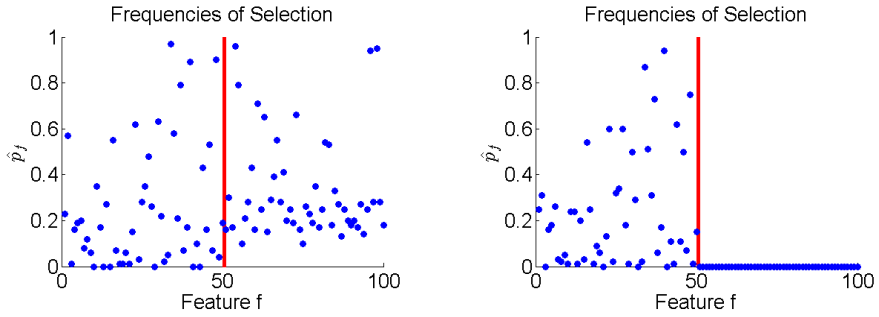


Figure 8: Summary of the pareto fronts for the 4 degrees of redundancy. The average OOB misclassification error is given for the two extreme points of each pareto front.



(a) $\lambda = 1.3 \times 10^{-2}$ (b) $\lambda = 6.8 \times 10^{-2}$

Figure 9: The observed frequencies of selection $\hat{p}_f$ of each feature optimizing only the likelihood [LEFT] and choosing a trade-off between stability and likelihood [RIGHT] for high redundancy ($\rho = 0.8$). The features on the left of the red vertical line correspond to the 50 relevant features and the ones on the right to the 50 irrelevant ones.

false positive rate 0), which means they have not been selected on any of the $M = 100$ samples. This is not the case when only maximizing the likelihood: we cannot discriminate the set of relevant features from the set of irrelevant ones by looking at the frequencies of selection $\hat{p}_f$. Using the 1000 holdout set, we applied $L1$-regularised logistic regression using these two values of λ. Table 5 provides the false positives and false negatives for the 4 degrees of redundancy. The results show a decrease in the false positives when optimizing both stability and likelihood while having a limited effect on the false negatives.

Redundancy	Optimizing likelihood	Optimizing both
none	$FP = 0$, $FN = 0$	IDEM
low	$FP = 14$, $FN = 19$	$FP = 0$, $FN = 20$
medium	$FP = 0$, $FN = 26$	$FP = 0$, $FN = 26$
high	$FP = 12$, $FN = 38$	$FP = 0$, $FN = 39$

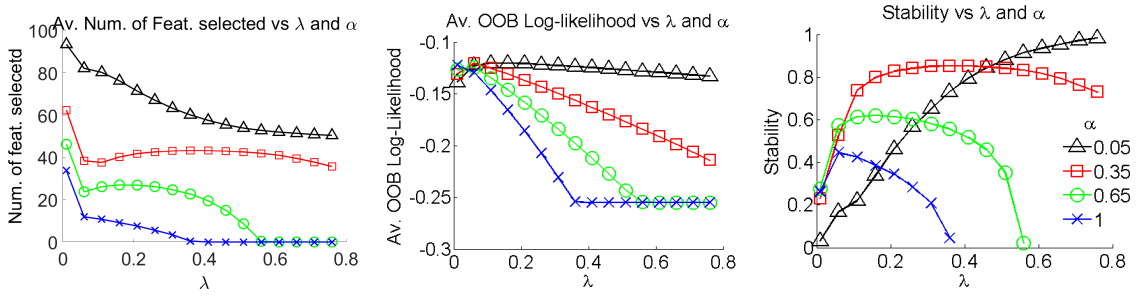
Table 5: False positives and false negatives of the final feature set for different degrees of redundancy ρ when optimizing only the likelihood against when optimizing both likelihood and stability.

6.1.3 Stability of the Elastic Net

$L1$-regularization has the effect of forcing regression coefficients to zero, hence selecting a subset of the available features. $L2$-regularization (*ridge regression*), is known to have a grouping effect: correlated features will have similar coefficients (Zhou, 2013). The Elastic Net is a convex combination of a $L1$ and a $L2$-regularization. It has two parameters, α and

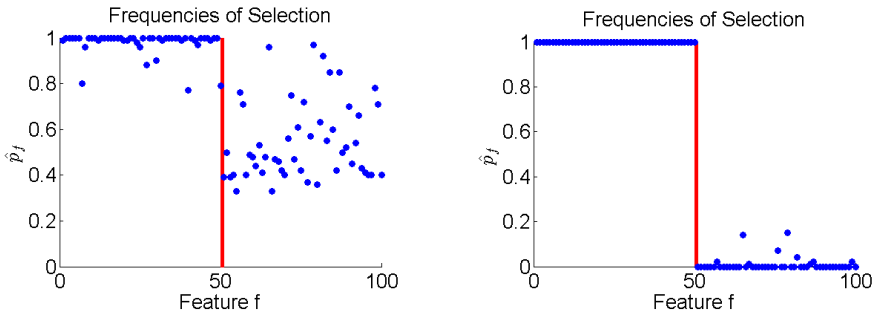
λ, where λ controls the overall weight of regularization and where α controls the balance between the two regularizing terms. When $\alpha = 1$, it becomes $L1$ (LASSO), and when $\alpha = 0$ it is $L2$ (ridge regression). As α varies, the Elastic net blends between the two and offers the advantages of both techniques—it forces some of the coefficients to be zero like LASSO while having the grouping effect of the ridge regression. Correlated features are a source of instability (Gulgezen et al., 2009; Wald et al., 2013), as feature selection procedures will tend to select a different feature from a same group of correlated features on different samples. Therefore, we expect the Elastic net to mitigate this, hence increasing stability.

In this section, we reproduce some of the experiments of the last section for the Elastic Net, optimizing the two regularizing parameters α and λ. We proceed as before, taking $M = 100$ bootstraps, and focus on the most challenging case of high redundancy ($\rho = 0.8$). We confirm in Figure 10a that as λ increases, the overall regularization increases and less features are selected. Figures 10b and 10c respectively give the average OOB log-likelihood and the stability against the values of λ for different values of α. We can see that no matter what is the value of λ chosen, $\alpha = 0.05$ has a higher likelihood than the other values of α in most cases and reaches high stability (greater than 0.90) for values of λ greater than 0.56. Interestingly, for these values of α and λ, we can see in Figure 10a that the number of features selected is around 50, which is the total number of relevant features. Let us have a closer look at $\alpha = 0.05$. If we were only optimizing the likelihood, we would pick $\lambda = 0.05$, which yields a stability of 0.34. The corresponding average misclassification error is 14%. If we wanted to also optimize stability, we could sacrifice a small amount of likelihood by picking $\lambda = 0.76$ which yields a stability of 0.98. The corresponding average OOB misclassification error is also 14%. Figure 11 gives the observed frequencies of selection $\hat{p}_f$ for $\lambda = 0.16$ on the left sub-figure (which is the value of λ that maximizes the likelihood) and for $\lambda = 0.76$ on the right sub-figure (which is a value of λ optimizing both the likelihood and the stability). We can see on the right sub-figure that when also optimizing stability, the whole set of relevant features is selected on each one of the $M = 100$ samples and irrelevant features are only rarely selected. On the left sub-figure, even though the likelihood for the given hyperparameters and the misclassification error are similar, we can see that non-relevant features are a lot more often selected in the model.



(a) Av. num. of feat. selected $\bar{k}$ (b) Likelihood (c) Stability

Figure 10: Plots against λ where each line corresponds to a different value of α in the high redundancy case ($\rho = 0.8$). We pick the $\alpha = 0.05$ as it reaches a higher likelihood for most values of λ and it can also achieve high stability.



(a) $\alpha = 0.05$ and $\lambda = 0.16$ (b) $\alpha = 0.05$ and $\lambda = 0.76$

Figure 11: The observed frequencies of selection $\hat{p}_f$ of each feature optimizing only the likelihood [LEFT] and choosing a trade-off between stability and likelihood [RIGHT] for high redundancy ($\rho = 0.8$). The features on the left of the red vertical line correspond to the 50 relevant features and the ones on the right to the 50 irrelevant ones. We can see that optimizing both stability and likelihood [RIGHT] helps recovering the set of relevant features.

6.1.4 Conclusions

In this section, we proposed a methodology to select hyperparameters using stability along with the error. On the data set used, we showed that it is possible to select a hyperparameter yielding much higher stability values without loss of predictive power. When the stability is optimized along with the likelihood of the model, the false positive rate was lower. Using the Elastic Net, we were able to achieve high stability for a similar loss than using LASSO. Enforcing high stability had the effect of discriminating the set of relevant features, helping to *recover* the true set of relevant features.

6.2 How Stable is Stability Selection?

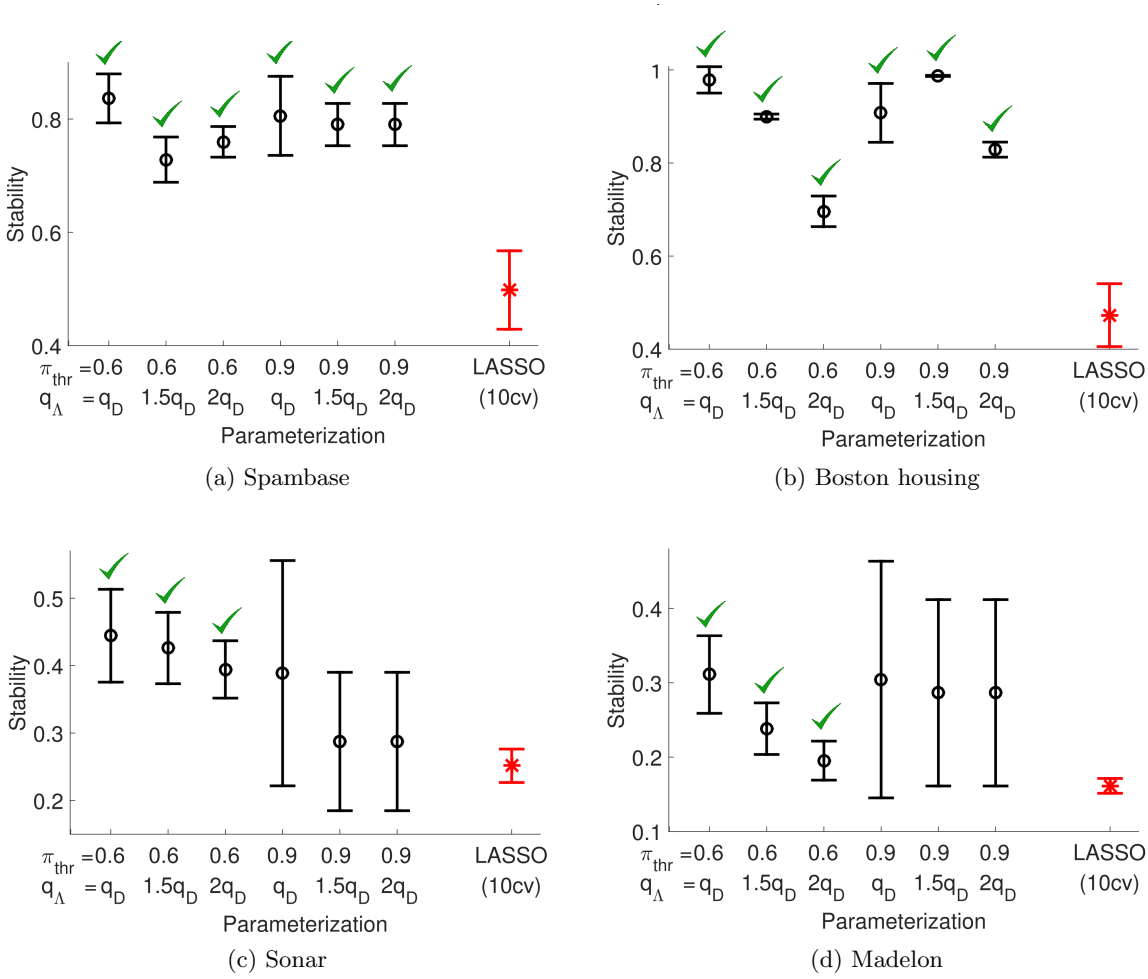
In high dimensional data sets, picking a regularizing parameter λ that recovers the *true* set of relevant features has proven to be challenging. For this reason, Meinshausen and Bühlmann (2010) introduced a technique called "*Stability Selection*", a popular and generic approach that can also be used for solving other problems of structure estimation such as graphical modelling. In this section, we focus on the use of Stability Selection in the context of feature selection with LASSO. It proposes to apply LASSO to M random sub-samples of size $\lfloor \frac{n}{2} \rfloor$ of the original data set (where n is the sample size) for a set of regularizing parameters $\lambda \in \Lambda$, where Λ is a subset of $\mathbb{R}^+$. This method considers the frequencies of selection of each feature $\hat{p}_f$ for each value of $\lambda \in \Lambda$ and defines the **set of stable variables** as *the set of all variables having a frequency of selection* $\hat{p}_f \geq \pi_{thr}$ *for at least one of the regularizing parameters* $\lambda \in \Lambda$ (where π_{th} is a user-defined threshold). Then they propose to use the identified stable set as an approximation of the true relevant set.

The proposed technique is effectively an ensemble feature selection technique where the final set is made of all features having a high frequency of selection $\hat{p}_f$ for at least

one the chosen regularizing parameter. The main contribution of Stability Selection is that it provides a control over false discovery error rates (i.e. the number of irrelevant features identified as relevant), and as a result, a principled way to choose the amount of regularization for variable selection. In relation to our work, we can point out the following interesting facts about this work: (1) it uses the concept of frequency of selection $\hat{p}_f$ to detect relevant features; (2) it uses the underlying idea that the *stable set* does not only help recovering the true feature set but also will be more robust to the choice of regularizing parameters; (3) It provides an upper bound and an exact control of the number of false positives (i.e. the number of irrelevant features falsely selected). Therefore, this work implicitly uses the idea that selecting stable features in the final set will help recover the *true* set of relevant features, which is intimately linked to the results of the previous section where we have shown that enforcing stability could potentially reduce the number of false positives. Intuitively, the final set of variables picked by Stability Selection should be more stable in the sense of our definition $\hat{\Phi}(\mathcal{Z})$, as they select the variables showing a consensus across multiple repeats of the data with perturbations and for different regularizing parameters.

In this section, we use our proposed measure to *quantify* just how stable their *stable set* can be. To that end, we look at how much the final set picked by Stability Selection varies in the context of LASSO and we will show on 4 data sets that it will indeed yield more stable results (in the sense of $\hat{\Phi}(\mathcal{Z})$) than its non-ensemble version (LASSO). Nevertheless, we remind the reader that these experiments are purely illustrative of the concepts discussed in the paper and do not claim to be an exhaustive empirical study. Stability selection possesses 3 hyperparameters[12]: (1) the cut-off value π_{thr}, (2) the average number of features selected q_Λ over the all values of $\lambda \in \Lambda$ and (3) the set of regularizing parameters Λ (where the two last hyperparameters are dependent). We used the values suggested by the original authors: that is $\pi_{thr} \in (0.6 - 0.9)$ and q_Λ around $\sqrt{0.8d}$. Figure 12 compares the two approaches for variable selection in 4 data sets, three binary classification problems (Spambase/Sonar/Madelon) and one regression (Boston housing). To derive the 95%-interval estimate of stability, we ran the algorithms on $M = 100$ bootstraps, and used the tools presented in Section 4.2.3. The first observation is that, no matter the parametrisation, the average stability of stability selection is always higher than the stability of LASSO. Furthermore, we performed hypothesis tests at a level of significance of 5% to check for which hyperparameters Stability Selection achieved higher stability than LASSO. A green tick indicates where the null hypothesis (equal stability) was rejected. On the Sonar and Madelon data sets, the stability of Stability Selection is consistently higher than LASSO, but can present high variability for some hyperparameters (as shown by the large confidence intervals). In those cases, it failed to reject the null hypothesis. These experiments highlight the importance of statistical significance when quantifying stability and the need for such statistical tools.

12. We note that the first two hyperparameters listed hereafter effectively control the upper bound on the amount of false positives.



(a) Spambase

(b) Boston housing

(c) Sonar

(d) Madelon

Figure 12: Comparing the stability of LASSO and different parametrisations of Stability Selection in four classification/regression data sets. For LASSO (red star), we optimised the regularisation parameter using 10-fold cross validation and the *one-standard-error* rule—picking up the most parsimonious model within one standard error of the minimum (Hastie et al., 2009). For stability selection (black circle), we explored different parameters: for the cut-off threshold $\pi_{\text{thr}} \in \{0.60, 0.90\}$, while for the average number of selected variable $q_\Lambda \in \{q_D, 1.5q_D, 2q_D\}$, where $q_D = \sqrt{0.8d}$ is the default value. We performed hypothesis tests to check whether the stability of Stability Selection is significantly different from LASSO. The green tick means that the null hypothesis (i.e. the population stabilities are equal) has been rejected at level of significance 5%.

7. Conclusions and Future Work

In this section, we present the conclusions and then consider possible areas of future work.

7.1 Conclusions

We have provided a rigorous statistical treatment for the concept of stability in feature selection. Following a property-based approach suggested in previous works, we identified a set of 5 properties—and argued for them as desirable in most (if not all) scenarios. Then, we compared all existing stability measures in terms of properties. It emerged that no existing measure satisfied all desirable properties—to counter this, we constructed a novel measure, Definition 4. This turns out to be a generalization of several existing measures, and possesses all 5 properties. In addition, it provides new capabilities, not previously possible in the literature. We provided confidence intervals and hypothesis testing of *true* stability, and finally, we showed how these tools could be used to choose hyperparameters. An important conclusion is that optimizing stability can be potentially achieved without significant loss of accuracy, and can help identifying the *true* underlying set of features.

7.2 Future Work

In some data scenarios, the user might not want to measure the instability due to the redundancy of the features. In that scenario, the user is more interested in knowing whether features belonging to a same group of correlated features have been consistently selected, rather than looking at the selection of each feature independently. For this purpose, stability measures taking into account feature redundancy have been proposed in the literature. The relative POG (also called $POGR$) and the relative $nPOG$ (also called $nPOGR$) are both extensions of the POG measure and of the $nPOG$ measure respectively (Zhang et al., 2009) as they both reduce to their original version in the case of no redundancy. This case study is not in the scope of this paper. Nevertheless, we note that since they reduce to POG and $nPOG$, these two measures will also not possess the 5 properties. Future work might consider extending the measure we propose to take redundancy into account.

Another avenue of investigation could be the extension of the present work to other types of feature selection outputs—such as feature rankings or feature weights. A popular measure to quantify the stability of feature rankings is the pairwise Spearman's Rho. In the case of untied ranks, Nogueira et al. (2017) show that this measure can be re-written as

$$1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{V_{rank}},$$

where s_f^2 is the sample variance of the rank of feature X_f and where V_{rank} is the expected variance under the assumption that each ranking was generated at random (i.e. each ranking is equally likely). The above equation has a similar form to the measure proposed in this paper and provides a promising direction for unifying stability of ranking and subset selection.

Acknowledgments

This research was conducted with support from the Centre for Doctoral Training (CDT) in Computer Science, funded by Engineering and Physical Sciences Research Council (EPSRC) grant [EP/I028099/1]. GB was supported by the EPSRC LAMBDA project [EP/N035127/1]. We would like to thank Dr. Adam Pocock for his valuable feedback on our work.

Appendix A. Existing Stability Measures

In the remainder of the appendices, we will add a subscript to the stability $\hat{\Phi}$ giving the author or the name of the measure for disambiguation. Whenever the subscript is omitted, we refer to our proposed stability measure as given by Definition 4.

In this section, we first review the existing similarity-based stability measures and then we focus on the frequency-based ones.

A.1 Similarity-based Measures

We remind the reader that given a similarity measure ϕ between two feature sets s_i and s_j, the resulting stability $\hat{\Phi}(\mathcal{Z})$ is taken as the average pairwise similarities between the feature sets in $\mathcal{Z}$, that is

$$\hat{\Phi}(\mathcal{Z}) = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \phi(s_i, s_j).$$

For simplicity, we introduce notations that will be used in the remainder of the appendices. Let $r_{i,j}$ be the short notation for $|s_i \cap s_j|$, the size of the intersection between feature sets s_i and s_j. Let k_i denote the size of feature set s_i (when the size of the feature set is assumed constant, we will simply denote it k). Table 6 provides all 9 similarity measures used in the literature in the context of stability along with their minimum and maximum value. These definitions will be later used in the proof of properties in Appendix C.

A.2 Frequency-based Measures

In this section, we give the definitions of the frequency-based measures. As these require to introduce a lot of new notations, we do not summarize them in a table like we did for similarity-based measures.

A.2.1 Goh's Measure

Goh and Wong (2016) propose to use the frequency of selection, averaged over all features, that is $\hat{\Phi}_{Goh}(\mathcal{Z}) = \frac{1}{d} \sum_{f=1}^{d} \hat{p}_f$. This measure take values in $[0, 1]$.

First used in	**Name**	**Measure**	[min, max]
Dunne et al. (2002)	Hamming	$1 - \frac{\lvert s_i \setminus s_j \rvert + \lvert s_i \setminus s_j \rvert}{d}$	$[0, 1]$
Kalousis et al. (2005)	Jaccard	$\frac{r_{i,j}}{\lvert s_i \cup s_j \rvert}$	$[0, 1]$
Yu et al. (2008)	Dice-Sørenson	$\frac{2r_{i,j}}{k_i + k_j}$	$[0, 1]$
Zucknick et al. (2008)	Ochiai	$\frac{r_{i,j}}{\sqrt{k_i k_j}}$	$[0, 1]$
Shi et al. (2006)	POG	$\frac{r_{i,j}}{k_i}$	$[0, 1]$
Kuncheva (2007)	Consistency	$\frac{r_{i,j} - \frac{k^2}{d}}{k - \frac{k^2}{d}}$	$[-1, 1]$
Lustgarten et al. (2009)	Lustgarten	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{min(k_i, k_j) - max(0, k_i + k_j - d)}$	$[-1, 1]$
Wald et al. (2013)	Wald	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{min(k_i, k_j) - \frac{k_i k_j}{d}}$	$[1 - d, 1]$
Zhang et al. (2009)	nPOG	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{k_i - \frac{k_i k_j}{d}}$	$[1 - d, 1]$

Table 6: Similarity measures proposed in the literature 2002–2018, using the pairwise formulation. In some cases the measure is extremely simple, (e.g. percentage overlap of features) and authors are chosen simply as the first known usage of the measures in the context of stability. We note that as opposed as what can be found in some literature (Alelyani, 2013; P. and Perumal, 2016), the minimum for Wald's and $nPOG$ similarity measures is equal to $1 - d$ and not 0 (and is reached for $k_i = 1$, $k_j = d - 1$ and $r_{i,j} = 0$).

A.2.2 Davis' Measure

Davis et al. (2006, pg2) penalize the frequency to *"account for the artificial increase in stability that occurs with increasingly long gene signatures"*, as follows

$$\hat{\Phi}_{Davis}(\mathcal{Z}) = \max\left(0,\ \frac{1}{F}\sum_{f=1}^{d}\hat{p}_f - \alpha\frac{median(k_1, ..., k_M)}{d}\right), \tag{3}$$

where F is the number of features selected at least in one of the M feature sets (in other words, $F = |\cup_{i\in\{1,...,M\}} s_i|$) and where α is a hyperparameter chosen by the user. This measure also takes values in the interval $[0, 1]$.

A.2.3 Krizek's Measure

Krízek et al. (2007) treat each possible of the $\binom{d}{k}$ feature sets of k features as a random variable and estimate its Shannon entropy as

$$\hat{\Phi}_{Krizek}(\mathcal{Z}) = -\sum_{s_i\in\mathcal{Z}} \hat{p}(s_i)\log_2 \hat{p}(s_i),$$

where $\hat{p}(s_i)$ is the frequency of occurrence of subset s_i in $\mathcal{Z}$ over all the $\binom{d}{k}$ possible combinations of k features taken amongst d features. It takes values in $\left[0, \log(\min(M, \binom{d}{k}))\right]$.

A.2.4 Guzmán's Measure

A measure is proposed by Guzmán-Martínez and Alaiz-Rodríguez (2011), using frequencies to compute Jensen-Shannon divergences. Originally created for feature rankings, they extend it to feature sets of k features (top-k lists of genes in the literature), using the JS-divergence, as follows

$$\hat{\Phi}(\mathcal{Z}) = 1 - \frac{D_{JS}(\mathbf{q}_1, ..., \mathbf{q}_M)}{D^*_{JS}(\mathbf{q}_1, ..., \mathbf{q}_M)}.$$

Here, each $\mathbf{q}_i$ is a distribution over features, formed by taking the bitstring on the i^{th} row of the matrix $\mathcal{Z}$, and dividing through by k, the number of bits set. D^*_{JS} is a normalizing term—according to the authors *"the divergence value for a [feature set] that is completely random"*. This ensures the value is in $[0, 1]$, but most interestingly, this is yet another work correcting the measure for chance, as introduced by Kuncheva.

A.2.5 Relative Weighted Consistency CW_{rel}

With a property-based analysis, Somol and Novovičová (2010) constructed a new stability measure called the relative weighted consistency (CW_{rel}) as

$$\hat{\Phi}(\mathcal{Z}) = \frac{d\left(M\bar{k} - D + \sum_{f=1}^{d} M\hat{p}_f(M\hat{p}_f - 1)\right) - (M\bar{k})^2 + D^2}{d\left(H^2 + M(M\bar{k} - H) - D\right) - (M\bar{k})^2 + D^2}, \tag{4}$$

where $D = (M\bar{k})\ mod\ d$ and $H = (M\bar{k})\ mod\ M$.

A.2.6 Lausser's Measure

Finally, Lausser et al. (2013) proposed a measure for feature sets of fixed size k as follows:

$$\hat{\Phi}(\mathcal{Z}) = \frac{1}{M^2 k}\sum_{i=1}^{M} i^2 a^{(i)},$$

where $a^{(i)} = \sum_{f=1}^{d} \mathbb{1}\{\sum_{j=1}^{M} z_{j,f} = i\}$ is the number of features selected exactly i times.

Appendix B. Proof of Theorems

In this section, we provide the proofs or proof sketches for the theorems and corollaries in the paper. Before we proceed to the proofs, we give the following equation, that will be repeatedly used in the proofs

$$\sum_{f=1}^{d} \hat{p}_f = \frac{1}{M} \sum_{i=1}^{M} \sum_{f=1}^{d} z_{i,f} = \frac{1}{M} \sum_{i=1}^{M} k_i = \bar{k}. \tag{5}$$

B.1 Proof of Theorem 1

In this appendix we prove the following theorem from Section 3.

Theorem 1 *The average pairwise intersection between the M feature sets is as a linear function of the sample variances of the selection of each feature, as follows*

$$\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} r_{i,j} = \bar{k} - \sum_{f=1}^{d} s_f^2, \tag{6}$$

where $s_f^2 = \frac{M}{M-1} \hat{p}_f (1 - \hat{p}_f)$ is the sample variance of the selection of the f^{th} feature.

Proof. To prove Theorem 1, we start by calculating the average pairwise size of the intersection and show that we get the results presented.

$$\begin{aligned}
\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} r_{i,j} &= \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} r_{i,j} - \frac{1}{M(M-1)} \sum_{i=1}^{M} r_{i,i} \\
&= \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} r_{i,j} - \frac{1}{M(M-1)} \sum_{i=1}^{M} k_i \\
&= \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} r_{i,j} - \frac{1}{M-1} \bar{k}
\end{aligned}$$

Since the i^{th} feature set $\mathcal{Z}_{(i,:)}$ and the j^{th} feature set $\mathcal{Z}_{(j,:)}$ are binary vectors, the size of their intersection $r_{i,j}$ is the number of 1s occurring at the same position in both vectors. In other words, $r_{i,j}$ is the dot product of the two feature sets, that is $\mathcal{Z}_{(i,:)} \cdot \mathcal{Z}_{(j,:)} = \sum_{f=1}^{d} z_{i,f} z_{j,f}$. By substituting in the previous equation, we get

$$\begin{aligned}
\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} r_{i,j} &= \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j=1}^{M} \sum_{f=1}^{d} z_{i,f} z_{j,f} - \frac{1}{M-1} \bar{k} \\
&= \frac{1}{M(M-1)} \sum_{f=1}^{d} \left(\sum_{i=1}^{M} z_{i,f} \right) \left(\sum_{j=1}^{M} z_{j,f} \right) - \frac{1}{M-1} \bar{k} \\
&= \frac{1}{M(M-1)} \sum_{f=1}^{d} \left(\sum_{i=1}^{M} z_{i,f} \right)^2 - \frac{1}{M-1} \bar{k}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{M(M-1)} \sum_{f=1}^{d} (M\hat{p}_f)^2 - \frac{1}{M-1}\bar{k} \\
&= \frac{M}{M-1} \sum_{f=1}^{d} \left(\hat{p}_f^2 - \hat{p}_f + \hat{p}_f\right) - \frac{1}{M-1}\bar{k} \\
&= \frac{M}{M-1} \sum_{f=1}^{d} -\hat{p}_f(1-\hat{p}_f) + \frac{M}{M-1} \sum_{f=1}^{d} \hat{p}_f - \frac{1}{M-1}\bar{k} \\
&= -\frac{M}{M-1} \sum_{f=1}^{d} \hat{p}_f(1-\hat{p}_f) + \frac{M}{M-1}\bar{k} - \frac{1}{M-1}\bar{k} \\
&= \bar{k} - \sum_{f=1}^{d} s_f^2.
\end{aligned}
$$

■

B.2 Proof of Theorem 3

In this appendix we prove the following theorem from Section 4.1.

Theorem 3 *Under the Null Model of Feature Selection H_0, for all f, the expected value of the sample variance of Z_f is* $\mathbb{E}\left[s_f^2 | H_0\right] = \frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)$.

Proof. Since s_f^2 is the unbiased sample variance of the Bernoulli variable Z_f, we have $\mathbb{E}\left[s_f^2 | H_0\right] = Var[Z_f | H_0]$, which is the value of $p_f(1-p_f)$ under the Null Model of Feature Selection H_0.

To calculate p_f under H_0, we will start by carrying out the calculations on a simple example to clarify what is the sample space we are looking at under H_0. Let us assume we observe the matrix

$$\mathcal{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}.$$

By definition, under H_0, for each row i, all permutations of the bit-string on row i are equally likely. For the first row, we have 3 possible permutations that are 100, 010, 001 and for the second row, we also have 3 possible permutations that are 110, 101, 011. Therefore, under H_0, all the $N = 3 \times 3 = 9$ following matrices are equally likely to be observed:

$$
\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}, \quad
\begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \end{pmatrix}, \quad
\begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix},
$$

$$
\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad
\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}, \quad
\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix},
$$

$$
\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad
\begin{pmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}, \quad
\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}.
$$

Since all these matrices are equally likely to be observed under H_0, to calculate p_f, we can simply count the proportion of times feature Z_f equals 1 over the $MN = 2 \times 9 = 18$ rows. Here we would get $p_f = \frac{9}{18} = \frac{1}{2}$ for any f.

Let N be the total number of possible matrices under H_0. In the more general case, to compute p_f under H_0, we can count the proportion of times $Z_f = 1$ over the MN rows. We can decompose this into a sum over the row numbers as follows:

$$p_f = \Pr[Z_f = 1|H_0] = \frac{\sum_{i=1}^{M} N \times \Pr[Z_f = 1|\text{on row i }, H_0]}{MN}$$

$$p_f = \frac{1}{M} \sum_{i=1}^{M} \Pr[Z_f = 1|\text{on row i }, H_0], \tag{7}$$

where $\Pr[Z_f = 1|\text{on row i }, H_0]$ is the proportion of times Z_f is equal to 1 on all the permutations of the i^{th} row of $\mathcal{Z}$. Since the i^{th} row of $\mathcal{Z}$ has k_i bits set to 1, this is equal to

$$\frac{\#\{\text{bit-strings with } k_i \text{ 1s where } Z_f = 1\}}{\#\{\text{bit-strings with } k_i \text{ 1s}\}}.$$

The denominator is equal to $\binom{d}{k_i}$. For the numerator, we know X_f is set equal to 1, which means we have now $k_i - 1$ bits left to set to 1 from the remaining $d - 1$ bits. Therefore the numerator is equal to $\binom{d-1}{k_i-1}$. Replacing these in the previous equation, we get that on the i^{th} row, all features have an equal probability of being selected equal to $\frac{\binom{d-1}{k_i-1}}{\binom{d}{k_i}} = \frac{k_i}{d}$. Now, replacing this in Equation (7), we get that under H_0,

$$p_f = \frac{1}{M} \sum_{i=1}^{M} \frac{k_i}{d} = \frac{\bar{k}}{d}.$$

We can verify this in our previous example, where we had $k_1 = 2$, $k_2 = 1$, $d = 3$, which gives $p_f = \frac{\bar{k}}{d} = \frac{1}{2}$ as computed previously.
Using this, we can now compute $\mathbb{E}\left[s_f^2|H_0\right] = Var[Z_f|H_0] = p_f(1 - p_f) = \frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)$. ■

B.3 Proof of Theorem 5

In this appendix we prove the following theorem from Section 4.1.
Theorem 5 *When the number of features selected is constant:*

- *The stability estimator $\hat{\Phi}$ is equal to the stability measures derived by Kuncheva (2007), Wald et al. (2013) and to nPOG (Zhang et al., 2009).*
- *The stability estimator $\hat{\Phi}$ and CW_{rel} (Somol and Novovičová, 2010) are asymptotically equivalent.*

Proof. First, we show that when the number of features selected is constant equal to k, then Kuncheva's measure is equal to the proposed stability measure. The stability measure

defined by Kuncheva (2007) is

$$\hat{\Phi}_{Kuncheva}(\mathcal{Z}) = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \frac{r_{i,j} - \frac{k^2}{d}}{k - \frac{k^2}{d}} = \frac{1}{M(M-1)} \left(\sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \frac{r_{i,j}}{k - \frac{k^2}{d}} \right) - \frac{\frac{k^2}{d}}{k - \frac{k^2}{d}}$$
$$= \frac{1}{k - \frac{k^2}{d}} \left(\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} r_{i,j} \right) - \frac{\frac{k^2}{d}}{k - \frac{k^2}{d}}.$$

Using Theorem 1, we can replace the term between parenthesis in the latter equation by $k - \sum_{f=1}^{d} s_f^2$ (since the number of features selected is constant, $\bar{k} = k$). We get that

$$\hat{\Phi}_{Kuncheva}(\mathcal{Z}) = \frac{1}{k - \frac{k^2}{d}} \left(k - \sum_{f=1}^{d} s_f^2 \right) - \frac{\frac{k^2}{d}}{k - \frac{k^2}{d}} = \frac{k - \frac{k^2}{d}}{k - \frac{k^2}{d}} - \frac{\sum_{f=1}^{d} s_f^2}{k - \frac{k^2}{d}} = 1 - \frac{\frac{1}{d} \sum_{f=1}^{d} s_f^2}{\frac{k}{d}\left(1 - \frac{k}{d}\right)},$$

which is our proposed stability measure for $\bar{k} = k$. Then, using the definitions of Wald and $nPOG$ measures given in Table 6, by replacing the cardinalities of the sets by k in the equations, they reduce to Kuncheva's measure, which proves the first part of the theorem.

We now prove the equivalence of our measure with CW_{rel}. Using Equation (4), we can rewrite CW_{rel} as

$$\hat{\Phi}_{CW_{rel}}(\mathcal{Z}) = \frac{-\frac{1}{d}\frac{M-1}{M} \sum_{f=1}^{d} s_f^2 + \frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right) - \frac{D}{M^2 d}\left(1 - \frac{D}{d}\right)}{\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right) - \frac{D}{M^2 d}\left(1 - \frac{D}{d}\right) + \frac{H^2}{dM^2} - \frac{H}{Md}}. \tag{8}$$

Assuming that the number of features selected is constant equal to k, we then have that $\bar{k} = k$ and hence that $H = (Mk)\ mod\ M = 0$. Therefore the above equation becomes

$$\hat{\Phi}_{CW_{rel}}(\mathcal{Z}) = 1 - \frac{\frac{M-1}{M} \sum_{f=1}^{d} s_f^2}{k\left(1 - \frac{k}{d}\right) - \frac{D}{M^2}\left(1 - \frac{D}{d}\right)}.$$

We have that $D = (Mk)\ mod\ d$ which implies that D is a constant number between 0 and $d-1$. Therefore the limit of the term $\frac{D}{M^2}\left(1 - \frac{D}{d}\right)$ as M approaches infinity is 0. Therefore, taking the limit of the above equation, we get

$$\lim_{M \to \infty} \hat{\Phi}_{CW_{rel}}(\mathcal{Z}) = \lim_{M \to \infty} \left[1 - \frac{\frac{M-1}{M} \sum_{f=1}^{d} s_f^2}{k\left(1 - \frac{k}{d}\right)} \right] = \lim_{M \to \infty} \hat{\Phi}(\mathcal{Z}).$$

This shows that $\lim_{M \to +\infty} \frac{\hat{\Phi}_{CW_{rel}}(\mathcal{Z})}{\hat{\Phi}(\mathcal{Z})} = 1$ and therefore we obtain that $\hat{\Phi}_{CW_{rel}}(\mathcal{Z}) \underset{M \to +\infty}{\sim} \hat{\Phi}(\mathcal{Z})$, which is what we wanted to prove. ■

B.4 Proof of Theorem 6

In this appendix we prove the following theorem from Section 4.2.1.

Theorem 6 *When there are only two categories (0/1), Fleiss' Kappa is equal to* $\hat{\Phi}(\mathcal{Z})$.

Proof. To prove this, we start from the definition of Fleiss' Kappa as given in the original paper (Fleiss, 1971) and show that when the number of categories is equal to 2, it reduces to the proposed definition of stability (c.f. Definition 4). Fleiss (1971) defines Kappa as

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{9}$$

where:

- $\bar{P}_e = \sum_{j=1}^{q} p_j^2$ in which
  - $q = 2$ is the number of categories;
  - $p_j = \frac{1}{Md} \sum_{f=1}^{d} n_{fj}$;
  - n_{fj} is the number of samples that assign f to the j^{th} category. Therefore, in our case, we have $n_{f1} = M\hat{p}_f$ and $n_{f0} = M - M\hat{p}_f$.
- $\bar{P} = \frac{1}{dM(M-1)} \sum_{f=1}^{d} \sum_{j=1}^{q} n_{fj}(n_{fj} - 1)$.

Now we can re-write this using our notation. First, we have that

- $p_1 = \frac{1}{Md} \sum_{f=1}^{d} n_{f1} = \frac{1}{Md} \sum_{f=1}^{d} M\hat{p}_f = \frac{1}{d} \sum_{f=1}^{d} \hat{p}_f = \frac{\bar{k}}{d}$;
- $p_0 = \frac{1}{Md} \sum_{f=1}^{d} n_{f0} = \frac{1}{Md} \sum_{f=1}^{d} (M - M\hat{p}_f) = 1 - \frac{1}{d} \sum_{f=1}^{d} \hat{p}_f = 1 - \frac{\bar{k}}{d}$.

Therefore,

$$\bar{P}_e = p_0^2 + p_1^2 = \frac{\bar{k}^2}{d^2} + \left(1 - \frac{\bar{k}}{d}\right)^2 = \frac{\bar{k}^2}{d^2} + 1 - 2\frac{\bar{k}}{d} + \frac{\bar{k}^2}{d^2} = 1 - 2\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right).$$

Let us now calculate $\bar{P}$.

$$\bar{P} = \frac{1}{dM(M-1)} \sum_{f=1}^{d} \sum_{j=1}^{q} n_{fj}(n_{fj} - 1)$$

$$\bar{P} = \frac{1}{dM(M-1)} \sum_{f=1}^{d} \left(n_{f0}(n_{f0} - 1) + n_{f1}(n_{f1} - 1)\right)$$

$$\bar{P} = \frac{1}{dM(M-1)} \sum_{f=1}^{d} \left((M - M\hat{p}_f)(M - M\hat{p}_f - 1) + M\hat{p}_f(M\hat{p}_f - 1)\right)$$

$$\bar{P} = 1 - \frac{2}{d} \sum_{f=1}^{d} \frac{M}{M-1} \hat{p}_f(1 - \hat{p}_f) = 1 - \frac{2}{d} \sum_{f=1}^{d} s_f^2.$$

Now, substituting the two last equations back into Equation (9), we finally get that

$$\kappa = \frac{1 - \frac{2}{d} \sum_{f=1}^{d} s_f^2 - 1 + 2\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)}{1 - 1 + 2\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)} = \frac{-\frac{1}{d} \sum_{f=1}^{d} s_f^2 + \frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)}{\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)}$$

$$\kappa = 1 - \frac{\frac{1}{d} \sum_{f=1}^{d} s_f^2}{\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right)} = \hat{\Phi}(\mathcal{Z}),$$

which is what we wanted to prove. ■

B.5 Proof of Theorem 7

Theorem 7 (Asymptotic Distribution) *As $M \to \infty$, the statistic $\hat{\Phi}$ weakly converges to a normal distribution:*

$$V_M = \frac{\hat{\Phi} - \Phi}{\sqrt{v(\hat{\Phi})}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

where:

- $\Phi = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} p_f(1-p_f)}{\bar{p}(1-\bar{p})}$ *is the mean of the estimator $\hat{\Phi}$ in which $\bar{p} = \frac{1}{d}\sum_{f=1}^{d} p_f$ is the average mean parameter of the d Bernoulli variables;*
- $v(\hat{\Phi}) = \frac{4}{M^2}\sum_{i=1}^{M}(\hat{\Phi}_{(i)} - \hat{\bar{\Phi}}_{(.)})^2$ *is an estimate of the variance of $\hat{\Phi}$ in which:*
  - $\hat{\Phi}_{(i)} = \frac{1}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)}\left[\frac{1}{d}\sum_{f=1}^{d} z_{i,f}\hat{p}_f - \frac{k_i\bar{k}}{d^2} + \frac{\hat{\Phi}}{2}\left(\frac{2\bar{k}k_i}{d^2} - \frac{k_i}{d} - \frac{\bar{k}}{d} + 1\right)\right]$;
  - *and $\hat{\bar{\Phi}}_{(.)}$ is the average value of $\hat{\Phi}_{(i)}$, that is: $\frac{1}{M}\sum_{i=1}^{M}\hat{\Phi}_{(i)}$.*

Proof. We prove this using Theorem 6 showing the equality between our proposed measure and Fleiss' Kappa and the work of Gwet (2008) that derives the asymptotic distribution of Fleiss' Kappa, as explained in the first paragraph of Section 4.2.2. ■

B.6 Proof of Theorem 9

In this appendix we prove the following theorem from Section 4.2.4.

Theorem 9 *The test statistic for comparing stabilities is*

$$T_M = \frac{\hat{\Phi}(\mathcal{Z}_2) - \hat{\Phi}(\mathcal{Z}_1)}{\sqrt{v(\hat{\Phi}(\mathcal{Z}_1)) + v(\hat{\Phi}(\mathcal{Z}_2))}}.$$

Under H_0, the statistic T_M asymptotically (as M approaches infinity) follows a standard normal distribution.

Proof. We know from Theorem 7 that $\hat{\Phi}(\mathcal{Z}_1)$ is asymptotically normal with unknown mean Φ_1 and variance σ_1^2 and that $\hat{\Phi}(\mathcal{Z}_2)$ is asymptotically normal with unknown mean Φ_2 and variance σ_2^2. Therefore, the difference $\hat{\Phi}(\mathcal{Z}_2) - \hat{\Phi}(\mathcal{Z}_1)$ is normal with unknown mean $\Phi_2 - \Phi_1$ and with variance $\sigma_1^2 + \sigma_2^2$. Under H_0, $\Phi_2 - \Phi_1 = 0$ and we estimate this variance by $v(\hat{\Phi}(\mathcal{Z}_1)) + v(\hat{\Phi}(\mathcal{Z}_2))$ using the result of Theorem 7. This gives us the asymptotic distribution of the statistic T_M. ■

Appendix C. Proof of Properties

In this section, for each one of the 5 properties given in Section 3, we determine which measures possess the property.

C.1 First property: Fully defined

This property directly follows from the definitions of the stability measures given Appendix A. Kuncheva's, Krízek's, Guzmán's and Lausser's measures are only defined when the number of features selected is fixed, and therefore do not possess this property.

C.2 Second property: Monotonicity

Since the proofs will all be similar for similarity-based measures, we first provide the proofs for the similarity based measures and then we look at frequency-based ones.

C.2.1 Similarity-based Measures

We start by calculating the derivative for each one of the 9 the similarity measures and provide the results in Table 7. As we can see, for all 9 similarity measures, assuming that the cardinalities of the feature sets are always in $\{1, .., d-1\}$, we have that $\frac{d\phi(s_i,s_j)}{dr_{i,j}} > 0$ (some derivatives are undefined otherwise, which correspond to the limit cases where no features are selected or all the features are selected). Therefore the derivative of the stability measure $\hat{\Phi}(\mathcal{Z})$ will be positive since

$$\frac{d\hat{\Phi}(\mathcal{Z})}{d\left(\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M} r_{i,j}\right)} = \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\frac{\partial\phi(s_i,s_j)}{\partial r_{i,j}}$$

and since a sum of strictly positive quantities is strictly positive. Therefore, all similarity-based stability measures have the Monotonicity property.

	Hamming	Jaccard	Dice	Ochiai	POG
$\frac{d\phi(s_i,s_j)}{dr_{i,j}}$	$\frac{2}{d}$	$\frac{k_i+k_j}{(k_i+k_j-r_{i,j})^2}$	$\frac{2}{k_i+k_j}$	$\frac{1}{\sqrt{k_i k_j}}$	$\frac{1}{k_i}$

	Kuncheva	Lustgarten	Wald	$nPOG$
$\frac{d\phi(s_i,s_j)}{dr_{i,j}}$	$\frac{1}{k-\frac{k^2}{d}}$	$\frac{1}{\min(k_i,k_j)-\max(0,k_i+k_j-d)}$	$\frac{1}{\min(k_i,k_j)-\frac{k_i k_j}{d}}$	$\frac{1}{k_i-\frac{k_i k_j}{d}}$

Table 7: Derivatives for each one of the similarity measures.

C.2.2 Goh's Measure

Using the definition of Goh's measure (c.f. Appendix A.2) and Equation (5), we have that

$$\hat{\Phi}_{Goh}(\mathcal{Z}) = \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f = \frac{\bar{k}}{d}. \tag{10}$$

This is not a function of the variances of selection of each feature s_f^2, therefore the measure does not have the Monotonicity property.

C.2.3 Davis' Measure

When $\alpha = 0$, this measure reduces to Goh's measure, which does not possess the property. Therefore this measure does not possess the Monotonicity property either.

C.2.4 Krízek's Measure

To prove that this measure does not possess the Monotonicity property, we give a counter-example. Let us assume we have a procedure that selects $k = 2$ features out of $d = 4$

features in total. The two binary matrices $\mathcal{Z}_1$ and $\mathcal{Z}_2$ illustrate two different scenarios with $M = 4$ as follows

$$\mathcal{Z}_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \text{and} \quad \mathcal{Z}_2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

Using Krízek's measure, we get a stability of 1 for both $\mathcal{Z}_1$ and $\mathcal{Z}_2$ (using log base 2). Now by computing the sum of variances of selection of the 4 features for the two test cases, we get $\sum_{f=1}^{4} s_f^2 = \frac{4}{3}$ for $\mathcal{Z}_1$ and $\sum_{f=1}^{4} s_f^2 = \frac{2}{3}$ for $\mathcal{Z}_2$. Therefore $\mathcal{Z}_1$ and $\mathcal{Z}_2$ have the same stability value but different sums of variance, which is a counter-example of the property.

C.2.5 Guzmán's Measure

Before proving this, we re-write Guzmán's measure using our notation (this re-writing will also be useful for later proofs). For feature sets of fixed cardinality k, the stability is defined by Guzmán-Martínez and Alaiz-Rodríguez (2011) as

$$\hat{\Phi}_{Guzman}(\mathcal{Z}) = 1 - \frac{D_{JS}(\mathbf{q}_1, ..., \mathbf{q}_M)}{D^*_{JS}(\mathbf{q}_1, ..., \mathbf{q}_M)},$$

where:

- $D^*_{JS}(\mathbf{q}_1, ..., \mathbf{q}_M) = \log \frac{d}{k}$;
- $D_{JS}(\mathbf{q}_1, ..., \mathbf{q}_M) = \frac{1}{M} \sum_{i=1}^{M} \sum_{f=1}^{d} q_{f,i} \log \frac{q_{f,i}}{\bar{q_f}}$;
- $q_{f,i} = \frac{1}{k}$ if the f^{th} feature is selected on the i^{th} run and 0 otherwise;
- $\bar{q_f} = \frac{1}{M} \sum_{i=1}^{M} q_{f,i}$.

Therefore, using our notation, we get that $q_{f,i} = z_{i,f} \frac{1}{k}$ and that $\bar{q_f} = \frac{1}{M} \frac{1}{k} \sum_{i=1}^{M} z_{i,f} = \frac{\hat{p}_f}{k}$. Therefore,

$$\begin{aligned}
\hat{\Phi}_{Guzman}(\mathcal{Z}) =& 1 - \frac{\frac{1}{kM} \sum_{i=1}^{M} \sum_{f=1}^{d} z_{i,f} \log \frac{z_{i,f}}{\hat{p}_f}}{\log \frac{d}{k}} \\
\hat{\Phi}_{Guzman}(\mathcal{Z}) =& 1 - \frac{\frac{1}{kM} \sum_{i=1}^{M} \sum_{f=1}^{d} z_{i,f} \log z_{i,f}}{\log \frac{d}{k}} + \frac{\frac{1}{kM} \sum_{i=1}^{M} \sum_{f=1}^{d} z_{i,f} \log \hat{p}_f}{\log \frac{d}{k}} \\
\hat{\Phi}_{Guzman}(\mathcal{Z}) =& 1 - \frac{\frac{1}{kM} \sum_{i=1}^{M} \sum_{f=1}^{d} z_{i,f} \log z_{i,f}}{\log \frac{d}{k}} + \frac{\frac{1}{k} \sum_{f=1}^{d} \hat{p}_f \log \hat{p}_f}{\log \frac{d}{k}}.
\end{aligned}$$

Since $z_{i,f}$ is binary, $z_{i,f} \log z_{i,f} = 0$. Therefore, the previous equation becomes

$$\hat{\Phi}_{Guzman}(\mathcal{Z}) = 1 + \frac{\frac{1}{k} \sum_{f=1}^{d} \hat{p}_f \log \hat{p}_f}{\log \frac{d}{k}} = 1 - \frac{\frac{1}{d} \sum_{f=1}^{d} \hat{p}_f \log \hat{p}_f}{\frac{k}{d} \log \frac{k}{d}}. \tag{11}$$

Now, to prove that this measure does not have the Monotonicity property, we will give a counter-example. Let $\mathcal{Z}_1$ and $\mathcal{Z}_2$ be the two following binary matrices

$$\mathcal{Z}_1 = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{pmatrix} \qquad \text{and} \qquad \mathcal{Z}_2 = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix}.$$

We show that using Guzmán's measure, $\mathcal{Z}_1$ has a lower stability than $\mathcal{Z}_2$ but also a lower average variance, thus violating the Monotonicity property. Indeed, we have $\hat{\Phi}_{Guzman}(\mathcal{Z}_1) \simeq 0.24$ and $\hat{\Phi}_{Guzman}(\mathcal{Z}_2) \simeq 0.31$ while we have a sum of variances $\simeq 0.15$ for $\mathcal{Z}_1$ and $\simeq 0.17$ for $\mathcal{Z}_2$.

C.2.6 Relative Weighted Consistency CW_{rel}

Since H, D and $\bar{k}$ only depend on the feature set cardinalities $k_1, ..., k_M$, on M and on d, we can see from Equation (8) that $\hat{\Phi}_{CW_{rel}}(\mathcal{Z})$ is a linear and strictly decreasing function of s_f^2. Therefore, CW_{rel} possesses the Monotonicity property.

C.2.7 Lausser's measure

Similarly to what has been done for Guzmán's measure, we will re-write this measure as a function of the frequencies of selection $\hat{p}_f$. This will help us understand the measure and also will be useful for other proofs involving this measure. We remind the reader that Lausser's measure is defined as

$$\hat{\Phi}_{Lausser}(\mathcal{Z}) = \frac{1}{M^2 k} \sum_{f=1}^{d} \sum_{i=1}^{M} i^2 \mathbb{1}\{\sum_{j=1}^{M} z_{j,f} = i\} = \frac{1}{M^2 k} \sum_{f=1}^{d} \left[\sum_{i=1}^{M} i^2 \mathbb{1}\{M\hat{p}_f = i\} \right].$$

Let us look at the term in between brackets that depends on the row index i. We note that the indicator term $\mathbb{1}\{M\hat{p}_f = i\}$ is equal to 1 only when i is equal to $M\hat{p}_f$ and is equal to 0 for any other value of i in $\{1, ..., M\}$. Therefore we can make the sum over i disappear since it is always equal to $(M\hat{p}_f)^2$. Hence, Lausser's Measure can be re-written as

$$\hat{\Phi}_{Lausser}(\mathcal{Z}) = \frac{1}{M^2 k} \sum_{f=1}^{d} (M\hat{p}_f)^2 = \frac{1}{k} \sum_{f=1}^{d} \hat{p}_f^2. \tag{12}$$

This simple expression helps us understand what is this measure actually measuring. Let us now show that this is a strictly decreasing function of the sum of variances $\sum_{f=1}^{d} \hat{p}_f(1-\hat{p}_f)$. We can re-write the sum of variances as follows

$$\frac{M-1}{M} \sum_{f=1}^{d} s_f^2 = \sum_{f=1}^{d} \hat{p}_f(1-\hat{p}_f) = \sum_{f=1}^{d} \hat{p}_f - \sum_{f=1}^{d} \hat{p}_f^2 = k - \sum_{f=1}^{d} \hat{p}_f^2 = k(1 - \hat{\Phi}_{Lausser}(\mathcal{Z})).$$

Therefore, Lausser's measure has the Monotonicity property.

C.3 Third property: Bounds

In this section, we verify which measures have the Bounds property as given by Table 1.

C.3.1 Similarity-based measures

If a similarity measure ϕ is bounded, i.e. if $\exists(a,b) \in \mathbb{R}^2, a \leq \phi \leq b$, then it follows that the corresponding stability measure will also be bounded. Indeed:

$$a \leq \phi \leq b \quad \Rightarrow \quad a \leq \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \phi(s_i, s_j) \leq b \quad \Rightarrow \quad a \leq \hat{\Phi}(\mathcal{Z}) \leq b.$$

As given in Table 6, we can see that all similarity measures except Wald's measure (Wald et al., 2013) and $nPOG$ measure (Zhang et al., 2009) are bounded. Therefore, we know that their corresponding stability measures will also be bounded.

The contrary is not necessarily true. If a similarity measure is not bounded, this does not imply that the corresponding stability measure is not bounded. Nevertheless, we prove that the stability measures using Wald's and $nPOG$ similarity measures are not bounded using a counter-example. Let us assume we have the following scenario

$$\mathcal{Z} = \begin{bmatrix} 1 & 1 & 1 & 1 & ... & 1 & 0 \\ & & \vdots & & & & \\ 1 & 1 & 1 & 1 & ... & 1 & 0 \\ 0 & 0 & 0 & 0 & ... & 0 & 1 \\ & & \vdots & & & & \\ 0 & 0 & 0 & 0 & ... & 0 & 1 \end{bmatrix},$$

where the first $\frac{M}{2}$ feature sets are all identical and select the first $d-1$ features and the $\frac{M}{2}$ following ones are also identical but only select the first feature. In this situation, using Wald's similarity, the first block will give $\frac{M}{2}\left(\frac{M}{2}-1\right)$ similarities of 1 (as all feature sets in the first block are identical), the second block of feature sets will also give $\frac{M}{2}\left(\frac{M}{2}-1\right)$ similarities of 1 as all feature sets in the second block are also identical. Then the $\frac{M^2}{2}$ remaining pairs of feature sets (coming from the inter block pairs) have an intersection $r_{i,j} = 0$ and therefore a similarity equal to

$$\frac{0 - \frac{d-1}{d}}{1 - \frac{d-1}{d}} = \frac{1-d}{d-d+1} = 1-d.$$

So overall, the stability using Wald's similarity measure is equal to

$$\hat{\Phi}_{Wald}(\mathcal{Z}) = \frac{1}{M(M-1)}\left[2\frac{M}{2}\left(\frac{M}{2}-1\right) + \frac{M^2}{2}(1-d)\right] = \frac{\frac{M}{2}-1}{M-1} + \frac{M}{2(M-1)}(1-d).$$

We can see that that the value of the stability decreases with d. Therefore, we can conclude that Wald's stability measure is not bounded by constants. Using the same scenario, we can similarly show that the $nPOG$ measure is not bounded.

C.3.2 Frequency-based measures

The range of values of all the frequency-based measures are given in the literature and recapitulated in Appendix A.2. Krízek's measure has a maximum depending on M, d and k and therefore is not bounded. All five other frequency measures (CW_{rel}, Davis' and Goh's measures) take values in $[0,1]$ and therefore are bounded.

C.4 Fourth Property: Maximum

In this section, we show which one of the stability measures possess the Maximum property, as given in Table 1.

C.4.1 Similarity-based Measures

For the backward implication (Deterministic Selection $\rightarrow$ Maximum Stability), let us assume that all the feature sets in $\mathcal{Z}$ are identical with cardinality k, therefore $|s_i \cap s_j| = r_{i,j} = k$. By definition, for all similarity measures given in Table 1 except Lustgarten's measure, for all $i, j \in \{1, ..., M\}$, $\phi(s_i, s_j) = 1$ which means that the average pairwise similarity is also 1. Therefore all similarity-based measures have this property except Lustgarten's measure (as it is shown with a counter-example in Figure 1b).

For the forward implication (Maximum Stability $\rightarrow$ Deterministic Selection), showing that Wald's stability measure does not have this property can easily be done with a counter-example as done in the paper (c.f. Figure 1a). All other similarity-based stability measures have a maximum equal to 1. Let us assume that $\hat{\Phi}(\mathcal{Z}) = max(\hat{\Phi}) = 1$. We want to show that this implies that all feature sets in $\mathcal{Z}$ are identical.

$$\begin{aligned}
\hat{\Phi}(\mathcal{Z}) = 1 \Rightarrow \quad & \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \phi(s_i, s_j) = 1 \\
\Rightarrow \quad & \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \phi(s_i, s_j) = M(M-1) \\
\Rightarrow \quad & \forall i \in \{0,1\}^d, \forall j \in \{0,1\}^d, j \neq i, \phi(s_i, s_j) = 1.
\end{aligned}$$

Then using the constraint that $r_{i,j}$ is a natural number less or equal than $\min(k_i, k_j)$ (since that is the maximal possible size of intersection between two sets of size k_i and k_j), it can be shown for Jaccard, Dice, POG, $nPOG$ and Kuncheva, that this implies that $k_i = k_j = r_{i,j}$ which means that $s_i = s_j$.

C.4.2 Goh's Measure

Using Equation (10), we have that $\hat{\Phi}_{Goh}(\mathcal{Z}) = \frac{\bar{k}}{d}$. Therefore, when all feature sets in $\mathcal{Z}$ are identical, $\hat{\Phi}_{Goh}(\mathcal{Z})$ only reaches its maximal value of 1 if all features are selected (i.e., $\hat{p}_f = 1$ for all $f \in \{1, ..., d\}$). Therefore, this measure does not have the Maximum property.

C.4.3 Davis' Measure

Taking $\alpha = 0$, this measure is equal to Goh's measure (seen in the previous section). Therefore this stability measure does not have the Maximum property either.

C.4.4 Krízek's Measure

Let us show that the property is true for Krízek's stability measure. We note this measure is the only one for which lower values correspond to a higher stability and the maximum

stability is reached for a stability of 0.

$$\begin{aligned}
\hat{\Phi}_{Krizek}(\mathcal{Z}) = 0 \Leftrightarrow & -\sum_{s_i \in \mathcal{Z}} \hat{p}(s_i) \log_2 \hat{p}(s_i) = 0 \\
\Leftrightarrow & \ \forall j \in \{1, ..., \binom{d}{k}\}, \hat{p}(s_i) \log_2 \hat{p}(s_i) = 0 \\
\Leftrightarrow & \ \forall j \in \{1, ..., \binom{d}{k}\}, \hat{p}(s_i) = 0 \text{ or } \hat{p}(s_i) = 1 \\
\Leftrightarrow & \ \text{All feature sets in } \mathcal{Z} \text{ are identical.}
\end{aligned}$$

Therefore, Krízek's measure has the Maximum property.

C.4.5 Relative Weighted Consistency CW_{rel}

Using Equation (8), we have

$$\begin{aligned}
\hat{\Phi}_{CW_{rel}}(\mathcal{Z}) = 1 \quad \Leftrightarrow \quad & \frac{-\frac{1}{d}\frac{M-1}{M}\sum_{f=1}^{d} s_f^2 + \frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right) - \frac{D}{M^2 d}\left(1 - \frac{D}{d}\right)}{\frac{\bar{k}}{d}\left(1 - \frac{\bar{k}}{d}\right) - \frac{D}{M^2 d}\left(1 - \frac{D}{d}\right) + \frac{H^2}{M^2 d} - \frac{H}{Md}} = 1 \\
\Leftrightarrow \quad & \sum_{f=1}^{d} s_f^2 = \frac{H}{M-1} - \frac{H^2}{M(M-1)}.
\end{aligned}$$

When all feature sets in $\mathcal{Z}$ are identical, we have $\bar{k} = k$ and therefore $H = (M\bar{k}) \mod M = 0$. Therefore the right-hand side of the above equation is 0 and the left-hand side is also 0. This proves that CW_{rel} possesses the backward implication of the Maximum property.

The forward implication is not true. We give the following counter-example

$$\mathcal{Z} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix},$$

It can easily be shown that $\hat{\Phi}_{CW_{rel}}(\mathcal{Z}) = 1$, even though all rows in $\mathcal{Z}$ are not identical. Therefore CW_{rel} does not have the Maximum property.

C.4.6 Lausser's Measure

To prove that Lausser's measure has this property, we will use the expression given in Equation (12), that is

$$\hat{\Phi}_{Lausser}(\mathcal{Z}) = \frac{1}{k}\sum_{f=1}^{d} \hat{p}_f^2.$$

Let us first assume that all feature sets in $\mathcal{Z}$ are identical. This implies that we will have exactly k features for which the value of $\hat{p}_f$ will be 1 and $d-k$ features for which it will be 0. Hence in that case, $\hat{\Phi}_{Lausser}(\mathcal{Z}) = 1$. Now let us assume that the stability is equal to 1. This means that we have $\sum_{f=1}^{d} \hat{p}_f^2 = k$. The only solution to that is when we have exactly k features with a frequency of selection equal to 1 and the other features have a frequency of selection equal to 0. Therefore Lausser's measure possesses the Maximum property.

C.5 Fifth Property: Correction for Chance

In order to prove the property of Correction for chance, we calculate the expected value of $\hat{\Phi}(\mathcal{Z})$ under the Null Model of Feature Selection H_0 for each one of the existing stability measures. If $\mathbb{E}\left[\hat{\Phi}(\mathcal{Z})|H_0\right]$ is not constant (i.e. if it depends on parameters of the problem like k or d), then it does not have this property.

C.5.1 Similarity-based Measures

It has been shown in the literature that under H_0, the intersection follows a hypergeometric distribution with known expected value equal to $\mathbb{E}\left[r_{i,j}|H_0\right] = \frac{k_i k_j}{d}$ (Lustgarten et al., 2009). Using the linearity of the expectation, we will have that $\mathbb{E}\left[\hat{\Phi}(\mathcal{Z})|H_0\right]$ of the Normalized Hamming distance, the Jaccard index, the Dice-Sørensen index, Ochiai index and the *POG* measures will depend on k_i, k_j and d. Therefore, all these stability measures will not have the property of Correction for chance. All other similarity measures will verify $\mathbb{E}\left[\hat{\Phi}(\mathcal{Z})|H_0\right] = 0$ and will have the property of Correction for chance. We detail all calculations below.

For the Hamming similarity measure, we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\Phi}_{Hamming}(\mathcal{Z})|H_0\right] &= 1 - \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} \frac{k_i + k_j - \mathbb{E}\left[r_{i,j}|H_0\right]}{d} \\
&= 1 - \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} \frac{k_i + k_j - \frac{k_i k_j}{d}}{d} \\
&= 1 - \frac{1}{M(M-1)d} \left[\sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} k_i + \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} k_j - \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} \frac{k_i k_j}{d} \right] \\
&= 1 - \frac{1}{M(M-1)d} \left[M(M-1)\bar{k} + M(M-1)\bar{k} - \frac{1}{d} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} k_i k_j \right] \\
&= 1 - 2\frac{\bar{k}}{d} + \frac{1}{M(M-1)d^2} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j\neq i}}^{M} k_i k_j \\
&= 1 - 2\frac{\bar{k}}{d} + \frac{1}{M(M-1)d^2} \left(M^2\bar{k}^2 - \sum_{i=1}^{M} k_i^2 \right) \\
&= 1 - 2\frac{\bar{k}}{d} + \frac{M}{M-1}\frac{\bar{k}^2}{d^2} - \frac{1}{M(M-1)} \sum_{i=1}^{M} \frac{k_i^2}{d^2}.
\end{aligned}$$

In particular, when the number of features selected is constant equal to k, the above equation becomes $1 - 2\frac{k}{d}\left(1 - \frac{k}{d}\right)$, which depends on k and d. Therefore, this measure measure

does not have the Correction-for-chance property.

For the Jaccard similarity measure, we have

$$\begin{aligned}\mathbb{E}\left[\hat{\Phi}_{Jaccard}(\mathcal{Z})|H_0\right] &= \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\mathbb{E}\left[\frac{r_{i,j}}{k_i+k_j-r_{i,j}}|H_0\right] \\ &= \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\sum_{n=1}^{d}\frac{n}{k_i+k_j-n}\mathbb{P}(r_{i,j}=n|H_0).\end{aligned}$$

Since we know that under H_0, the intersection $r_{i,j}$ follows a central hypergeometric distribution, we have that

$$\mathbb{P}(r_{i,j}=n|H_0)=\frac{\binom{k_i}{n}\binom{d-k_i}{k_j-n}}{\binom{d}{k_j}}.$$

Therefore, the expected value of the average pairwise Jaccard index under the Null Model of Feature Selection H_0 is

$$\mathbb{E}\left[\hat{\Phi}_{Jaccard}(\mathcal{Z})|H_0\right]=\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\sum_{n=1}^{d}\frac{n}{k_i+k_j-n}\frac{\binom{k_i}{n}\binom{d-k_i}{k_j-n}}{\binom{d}{k_j}}.$$

In particular, when the number of features selected is constant equal to k, the above equation becomes $\sum_{n=1}^{d}\frac{n}{2k-n}\frac{\binom{k}{n}\binom{d-k}{k-n}}{\binom{d}{k}}$, which depends on k and d. Therefore this measure does not have the Correction-for-chance property.

For the Dice coefficient, we have

$$\mathbb{E}\left[\hat{\Phi}_{Dice}(\mathcal{Z})|H_0\right] = \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\frac{2\mathbb{E}\left[r_{i,j}|H_0\right]}{k_i+k_j}=\frac{2}{M(M-1)d}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\frac{k_ik_j}{k_i+k_j}.$$

In particular, when the number of features selected is constant equal to k, the above equation becomes $\frac{k}{d}$, which depends on k and d. Therefore this measure does not have the Correction-for-chance property.

For Ochiai's index, we have

$$\begin{aligned}\mathbb{E}\left[\hat{\Phi}_{Ochiai}(\mathcal{Z})|H_0\right] &= \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\frac{\mathbb{E}\left[r_{i,j}|H_0\right]}{\sqrt{k_ik_j}}=\frac{1}{M(M-1)d}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\frac{k_ik_j}{\sqrt{k_ik_j}} \\ &= \frac{1}{M(M-1)d}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\sqrt{k_ik_j}=-\frac{1}{M-1}\frac{\bar{k}}{d}+\frac{1}{M(M-1)}\left(\sum_{i=1}^{M}\sqrt{\frac{k_i}{d}}\right)^2.\end{aligned}$$

In particular, when the number of features selected is constant equal to k, the above equation becomes $\frac{k}{d}$, which depends on k and d. Therefore this measure does not have the Correction-for-chance property.

For *POG* measure, we use its symmetrical version equal to $\frac{r_{i,j}}{2k_i} + \frac{r_{i,j}}{2k_j}$ to carry out calculations. This results in the same stability value.

$$\begin{aligned}\mathbb{E}\left[\hat{\Phi}_{POG}(\mathcal{Z})|H_0\right] &= \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\left(\frac{\mathbb{E}\left[r_{i,j}|H_0\right]}{2k_i} + \frac{\mathbb{E}\left[r_{i,j}|H_0\right]}{2k_j}\right)\\ &= \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\left(\frac{k_i k_j}{2dk_i} + \frac{k_i k_j}{2dk_j}\right) = \frac{\bar{k}}{d},\end{aligned}$$

which depends on the number of feature selected and d. Therefore this measure does not have the Correction-for-chance property.

C.5.2 Frequency-based Measures

We showed in the proof of Theorem 3 (c.f. Appendix B.2), that under the Null Model of Feature Selection H_0, that p_f was equal to $\frac{\bar{k}}{d}$. Therefore, $\mathbb{E}\left[\hat{p}_f|H_0\right] = p_f = \frac{\bar{k}}{d}$. This will be used repeatedly in the proofs below.

For Goh's measure, since we have $\hat{\Phi}_{Goh}(\mathcal{Z}) = \frac{\bar{k}}{d}$ (c.f. Equation 10), this gives $\mathbb{E}\left[\hat{\Phi}_{Goh}(\mathcal{Z})|H_0\right] = \hat{\Phi}_{Goh}(\mathcal{Z}) = \frac{\bar{k}}{d}$, which is not constant. Therefore this measure does not have the property.

For Davis' Measure, we have $\mathbb{E}\left[\hat{\Phi}_{Davis}(\mathcal{Z})|H_0\right] = \hat{\Phi}_{Davis}(\mathcal{Z})$ as well. Therefore this measure does not have the Correction-for-chance property.

For Krízek's Measure, when the feature selection procedure is randomly selecting feature sets of cardinality k, the expected value of the frequency of occurrence of a feature set is equal to $\frac{1}{\binom{d}{k}}$. Therefore

$$\mathbb{E}\left[\hat{\Phi}_{Krizek}(\mathcal{Z})|H_0\right] = -\sum_{j=1}^{\binom{d}{k}}\frac{1}{\binom{d}{k}}\log\frac{1}{\binom{d}{k}} = -\log\frac{1}{\binom{d}{k}} = \log\binom{d}{k}.$$

Therefore Krízek's measure is not corrected for chance.

For Guzmán-Martinez's Measure, the stability measure is said to *"take the value zero for completely random rankings"* (Guzmán-Martínez and Alaiz-Rodríguez, 2011, pg 602), so we expect this measure to possess the Correction-for-chance property. We show this below.

$$\mathbb{E}\left[\hat{\Phi}_{Guzman}(\mathcal{Z})|H_0\right] = 1 - \frac{\frac{1}{d}}{\frac{k}{d}\log(\frac{k}{d})}\sum_{f=1}^{d}\mathbb{E}\left[\hat{p}_f\log\hat{p}_f|H_0\right]$$

$$= 1 - \frac{\frac{1}{d}}{\frac{k}{d}\log(\frac{k}{d})} \sum_{f=1}^{d} \mathbb{E}\left[\frac{\sum_{i=1}^{M} z_{i,f}}{M} \log \frac{\sum_{i=1}^{M} z_{i,f}}{M} | H_0\right]. \tag{13}$$

Under the Null Model of Feature Selection H_0, we have that $z_{i,f}$ follows a Bernoulli distribution with parameter $\frac{k}{d}$. Since we assumed that the samples $z_{1,f}, ..., z_{M,f}$ are independent and identically distributed (i.i.d.), we have that $\sum_{i=1}^{M} z_{i,f}$ follows a Binomial distribution with parameters M and $\frac{k}{d}$. Let $Y_f = \sum_{i=1}^{M} z_{i,f}$. Using this latter equation, we can calculate the expected value term of Equation (13),

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M} z_{i,f}}{M} \log \frac{\sum_{i=1}^{M} z_{i,f}}{M} | H_0\right] = \mathbb{E}\left[\frac{Y_f}{M} \log \frac{Y_f}{M} | H_0\right].$$

Let $g : y \longmapsto \frac{y}{M} \log \frac{y}{M}$, we have

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M} z_{i,f}}{M} \log \frac{\sum_{i=1}^{M} z_{i,f}}{M} | H_0\right] = \mathbb{E}\left[\frac{Y_f}{M} \log \frac{Y_f}{M} | H_0\right] = \mathbb{E}\left[g(Y_f) | H_0\right]. \tag{14}$$

Since g is a convex function[13] of y on the interval $(0, 1]$, we can use Jensen's inequality, which gives

$$\begin{aligned}
&\mathbb{E}\left[g(Y_f)|H_0\right] \geq g\left(\mathbb{E}\left[Y_f|H_0\right]\right) \Rightarrow \mathbb{E}\left[g(Y_f)|H_0\right] \geq g\left(M\frac{k}{d}\right) \Rightarrow \mathbb{E}\left[g(Y_f)|H_0\right] \geq \frac{k}{d}\log\frac{k}{d} \\
\Rightarrow \quad & \frac{1}{d}\sum_{f=1}^{d} \mathbb{E}\left[g(Y_f)|H_0\right] \geq \frac{k}{d}\log\frac{k}{d} \Rightarrow \frac{\frac{1}{d}}{\frac{k}{d}\log(\frac{k}{d})} \sum_{f=1}^{d} \mathbb{E}\left[g(Y_f)|H_0\right] \geq 1 \\
\Rightarrow \quad & \frac{\frac{1}{d}}{\frac{k}{d}\log(\frac{k}{d})} \sum_{f=1}^{d} \mathbb{E}\left[g(Y_f)|H_0\right] \leq -1 \Rightarrow 1 - \frac{\frac{1}{d}}{\frac{k}{d}\log(\frac{k}{d})} \sum_{f=1}^{d} \mathbb{E}\left[g(Y_f)|H_0\right] \leq 0 \\
\Rightarrow \quad & 1 - \mathbb{E}\left[\frac{\frac{1}{d}}{\frac{k}{d}\log(\frac{k}{d})} \sum_{f=1}^{d} g(Y_f)|H_0\right] \leq 0.
\end{aligned}$$

As shown by Equations (13) and (14), the left-hand-side term is equal to $\mathbb{E}\left[\hat{\Phi}_{Guzman}(\mathcal{Z})|H_0\right]$, therefore we get $\mathbb{E}\left[\hat{\Phi}_{Guzman}(\mathcal{Z})|H_0\right] \leq 0$. Since we know that $\hat{\Phi}_{Guzman}(\mathcal{Z})$ is a positive quantity, this gives us that $\mathbb{E}\left[\hat{\Phi}_{Guzman}(\mathcal{Z})|H_0\right] = 0$.

For the Relative Weighted Consistency CW_{rel}, using Equation (8), the result of Theorem 3 and by linearity of the expectation, we get

$$\mathbb{E}\left[\hat{\Phi}_{CW_{rel}}(\mathcal{Z})|H_0\right] = \frac{-\frac{1}{d}\frac{M-1}{M}\sum_{f=1}^{d}\mathbb{E}\left[s_f^2|H_0\right] + \frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right) - \frac{D}{M^2 d}\left(1-\frac{D}{d}\right)}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right) - \frac{D}{M^2 d}\left(1-\frac{D}{d}\right) + \frac{H^2}{M^2 d} - \frac{H}{Md}}$$

13. Indeed, its second derivative $g''(y) = \frac{1}{y \ln a}$ where a is the logarithm base used is non-negative for $y \in (0, 1]$. Therefore g is convex on that interval.

$$= \frac{-\frac{1}{d}\frac{M-1}{M}\sum_{f=1}^{d}\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)+\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)-\frac{D}{M^2d}\left(1-\frac{D}{d}\right)}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)-\frac{D}{M^2d}\left(1-\frac{D}{d}\right)+\frac{H^2}{M^2d}-\frac{H}{Md}}$$
$$= \frac{\frac{1}{M}\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)-\frac{D}{M^2d}\left(1-\frac{D}{d}\right)}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)-\frac{D}{M^2d}\left(1-\frac{D}{d}\right)+\frac{H^2}{M^2d}-\frac{H}{Md}}.$$

Since $H = (M\bar{k}) \mod M$, H is such that $M\bar{k} = \lfloor\bar{k}\rfloor M + H$. Therefore $H = M(\bar{k} - \lfloor\bar{k}\rfloor)$. Replacing in the previous equation, we get

$$\mathbb{E}\left[\hat{\Phi}_{CW_{rel}}(\mathcal{Z})|H_0\right] = \frac{\frac{1}{M}\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)-\frac{D}{M^2d}\left(1-\frac{D}{d}\right)}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)-\frac{D}{M^2d}\left(1-\frac{D}{d}\right)+\frac{(\bar{k}-\lfloor\bar{k}\rfloor)^2}{d}-\frac{\bar{k}-\lfloor\bar{k}\rfloor}{d}},$$

which is not constant. Nevertheless, when $\lfloor\bar{k}\rfloor = \bar{k}$, we have $\mathbb{E}\left[\hat{\Phi}_{CW_{rel}}(\mathcal{Z})|H_0\right] \xrightarrow[M\to\infty]{} 0$ and therefore the relative weighted consistency CW_{rel} is asymptotically corrected for chance. This is a result we expect since when the number of selected features is constant, this measure is asymptotically equivalent to our proposed measure (c.f. Theorem 5).

For Lausser's Measure, using Equation (12), we have

$$\begin{aligned}
\mathbb{E}\left[\hat{\Phi}_{Lausser}(\mathcal{Z})|H_0\right] &= \mathbb{E}\left[\frac{1}{k}\sum_{f=1}^{d}\hat{p}_f^2|H_0\right] = -\frac{1}{k}\sum_{f=1}^{d}\mathbb{E}\left[\hat{p}_f-\hat{p}_f^2-\hat{p}_f|H_0\right] \\
&= -\frac{1}{k}\sum_{f=1}^{d}\mathbb{E}\left[\hat{p}_f(1-\hat{p}_f)|H_0\right]+\frac{1}{k}\sum_{f=1}^{d}\mathbb{E}\left[\hat{p}_f|H_0\right] \\
&= -\frac{1}{k}\frac{M-1}{M}\sum_{f=1}^{d}\mathbb{E}\left[\frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)|H_0\right]+\frac{1}{k}\sum_{f=1}^{d}p_f \\
&= -\frac{1}{k}\frac{M-1}{M}\sum_{f=1}^{d}p_f(1-p_f)+\frac{1}{k}\sum_{f=1}^{d}p_f.
\end{aligned}$$

As shown earlier, $\mathbb{E}\left[\hat{p}_f|H_0\right] = \frac{k}{d}$, therefore

$$\mathbb{E}\left[\hat{\Phi}_{Lausser}(\mathcal{Z})|H_0\right] = -\frac{M-1}{M}\left(1-\frac{k}{d}\right)+1 = \frac{1}{M}+\frac{M-1}{M}\frac{k}{d},$$

which is not constant. Therefore, Lausser's measure does not have this property.

Appendix D. Proof of the Lower Bound of the Proposed Measure

In this section, we prove the lower bound of the proposed stability measure given in Definition 4. To do so, we first prove the lemma below that will be used later on.

Lemma 10 $\frac{1}{d^2}\sum_{f<f'}(\hat{p}_f-\hat{p}_{f'})^2 = \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2 - \left(\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f\right)^2$.

Proof.
Starting from the right-hand term, we get

$$\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2 - \left(\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f\right)^2 = \left(\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2\right) - \frac{1}{d^2}\sum_{f=1}^{d}\sum_{f'=1}^{d}\hat{p}_f\hat{p}_{f'} = \frac{1}{d^2}\sum_{f=1}^{d}\left(d\hat{p}_f^2 - \hat{p}_f\sum_{f'=1}^{d}\hat{p}_{f'}\right)$$
$$= \frac{1}{d^2}\sum_{f=1}^{d}\sum_{f'=1}^{d}\hat{p}_f\left(\hat{p}_f - \hat{p}_{f'}\right).$$

Since the term $(\hat{p}_f - \hat{p}_{f'})$ is equal to zero when $f = f'$, by splitting the sum in two terms, this is equal to

$$\frac{1}{d^2}\sum_{f=1}^{d}\sum_{f'=1}^{f-1}\hat{p}_f\left(\hat{p}_f - \hat{p}_{f'}\right) + \frac{1}{d^2}\sum_{f=1}^{d}\sum_{f'=f+1}^{d}\hat{p}_f\left(\hat{p}_f - \hat{p}_{f'}\right) \quad .$$

The left term $\sum_{f=1}^{d}\sum_{f'=1}^{f-1}\hat{p}_f(\hat{p}_f - \hat{p}_{f'})$ is equal to $\sum_{f=1}^{d}\sum_{f'=f+1}^{d} -\hat{p}_{f'}(\hat{p}_f - \hat{p}_{f'})$. Therefore the previous equation becomes

$$\frac{1}{d^2}\sum_{f=1}^{d}\sum_{f'=f+1}^{d} -\hat{p}_{f'}\left(\hat{p}_f - \hat{p}_{f'}\right) + \frac{1}{d^2}\sum_{f=1}^{d}\sum_{f'=f+1}^{d}\hat{p}_f\left(\hat{p}_f - \hat{p}_{f'}\right)$$
$$= \frac{1}{d^2}\sum_{f=1}^{d}\sum_{f'=f+1}^{d}\left(-\hat{p}_{f'}\left(\hat{p}_f - \hat{p}_{f'}\right) + \hat{p}_f\left(\hat{p}_f - \hat{p}_{f'}\right)\right)$$
$$= \frac{1}{d^2}\sum_{f=1}^{d}\sum_{f'=f+1}^{d}\left(\hat{p}_f - \hat{p}_{f'}\right)^2 = \frac{1}{d^2}\sum_{f<f'}\left(\hat{p}_f - \hat{p}_{f'}\right)^2.$$

■

Since a sum of squares is always positive, using this lemma we have that

$$\frac{1}{d^2}\sum_{f<f'}(\hat{p}_f - \hat{p}_{f'})^2 \geq 0$$
$$\Rightarrow \quad \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2 - \left(\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f\right)^2 \geq 0$$
$$\Rightarrow \quad \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2 - \left(\frac{\bar{k}}{d}\right)^2 \geq 0$$
$$\Rightarrow \quad \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2 \geq \left(\frac{\bar{k}}{d}\right)^2$$
$$\Rightarrow \quad \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2 - \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f \geq \left(\frac{\bar{k}}{d}\right)^2 - \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f$$

$$
\begin{aligned}
\Rightarrow \quad & -\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f) \geq \left(\frac{\bar{k}}{d}\right)^2 - \frac{\bar{k}}{d} \\
\Rightarrow \quad & -\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f) \geq -\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right) \\
\Rightarrow \quad & \frac{\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f)}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)} \leq 1 \\
\Rightarrow \quad & 1 - \frac{\frac{1}{d}\frac{M}{M-1}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f)}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)} \geq 1 - \frac{M}{M-1} \\
\Rightarrow \quad & \hat{\Phi}(\mathcal{Z}) \geq -\frac{1}{M-1}.
\end{aligned}
$$

Hence, $\hat{\Phi}(\mathcal{Z})$ is lower bounded by -1 (as $M \geq 2$), but asymptotically bounded by 0. ■

References

Salem Alelyani. *On Feature Selection Stability: A Data Perspective*. PhD thesis, Arizona State University, 2013.

Wilker Altidor, Taghi M. Khoshgoftaar, and Amri Napolitano. A noise-based stability evaluation of threshold-based feature selection techniques. In *IEEE International Conference on Information Reuse & Integration (IRI'11)*, pages 240–245, 2011.

Luca Baldassarre, Massimiliano Pontil, and Janaina Mouro-Miranda. Sparsity is better with stability: Combining accuracy and stability for model selection in brain decoding. *Frontiers in Neuroscience*, 11:62, 2017.

Kenneth J Berry, Paul W Mielke, Jr, and Janis E Johnston. *Permutation statistical methods: an integrated approach*. Springer, 2016.

Anne-Laure Boulesteix and Martin Slawski. Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10(5):556–68, 2009.

Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13(1):27–66, 2012.

Chad A. Davis, Fabian Gerick, Volker Hintermair, Caroline C. Friedel, Katrin Fundel, Robert Kffner, and Ralf Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–63, 2006.

David J. Dittman, Taghi M. Khoshgoftaar, Randall Wald, and Amri Napolitano. Similarity analysis of feature ranking techniques on imbalanced dna microarray datasets. In *2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1–5, 2012.

Gregory Ditzler, Robi Polikar, and Gail Rosen. A bootstrap based neyman-pearson test for identifying variable importance. *IEEE Transactions on Neural Networks and Learning Systems*, 26(4):880–886, 2015.

Kevin Dunne, Padraig Cunningham, and Francisco Azuaje. Solutions to instability problems with sequential wrapper-based approaches to feature selection. Technical Report TCD-CS-2002-28, Trinity College Dublin, School of Computer Science, 2002.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. *The Measurement of Interrater Agreement*, pages 598–626. John Wiley & Sons, Inc., 2004.

Wilson Wen Bin Goh and Limsoon Wong. Evaluating feature-selection stability in next-generation proteomics. *Journal of Bioinformatics and Computational Biology*, 14(05): 1650029, 2016.

Gokhan Gulgezen, Zehra Cataltepe, and Lei Yu. Stable and accurate feature selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 455–468, 2009.

Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3:1157–1182, 2003.

Roberto Guzmán-Martínez and Rocío Alaiz-Rodríguez. Feature selection stability assessment based on the jensen-shannon divergence. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD)*, pages 597–612. 2011.

Kilem Li Gwet. Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika*, 73(3):407, 2008.

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

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2nd edition, 2009.

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):258–64, 2008.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms. In *IEEE International Conference on Data Mining (ICDM'05)*, 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, 2007.

Iman Kamkar, Sunil Kumar Gupta, Dinh Phung, and Svetha Venkatesh. Stable feature selection with support vector machines. In *Australasian Joint Conference on Artificial Intelligence (AI 2015)*, volume 9457 of *LNCS*, pages 298–308, 2015.

Pavel Krízek, Josef Kittler, and Václav Hlavác. Improving stability of feature selection methods. In *International Conference on Computer Analysis of Images and Patterns (CAIP)*, volume 4673 of *LNCS*, pages 929–936, 2007.

Ludmila I. Kuncheva. A stability index for feature selection. In *Artificial Intelligence and Applications (AIAP'07)*, pages 390–395, 2007.

Ludwig Lausser, Christoph Müssel, Markus Maucher, and Hans A. Kestler. Measuring and visualizing the stability of biomarker selection techniques. *Computational Statistics*, 28 (1):51–65, 2013.

Hae Woo Lee, Carl Lawton, Young Jeong Na, and Seongkyu Yoon. Robustness of chemometrics-based feature selection methods in early cancer detection and biomarker discovery. *Statistical Applications in Genetics and Molecular Biology*, 12(2):207–23, 2012.

Jonathan L Lustgarten, Vanathi Gopalakrishnan, and Shyam Visweswaran. Measuring stability of feature selection in biomedical datasets. *AMIA Annual Symposium Proceedings.*, pages 406–410, 2009.

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

Sarah Nogueira. *Quantifying the Stability of Feature Selection.* PhD thesis, University of Manchester, 2018.

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, 2017.

Mohana Chelvan P. and Karuppasamy Perumal. A survey on feature selection stability measures. *International Journal of Computer and Information Technology*, 5(1):98–103, 2016.

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*, volume 5212 of *LNCS*, pages 313–325, 2008.

Ahmad A. Shanab, Taghi M. Khoshgoftaar, and Randall Wald. Impact of noise and data sampling on stability of feature selection. In *International Conference on Machine Learning and Applications and Workshops (ICMLA)*, pages 172–177, 2011.

Leming Shi, Laura H. Reid, Wendell D. Jones, et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151–61, September 2006.

Petr Somol and Jana Novovičová. 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.

Randall Wald, Taghi M. Khoshgoftaar, and David J. Dittman. A new fixed-overlap partitioning algorithm for determining stability of bioinformatics gene rankers. In *International Conference on Machine Learning and Applications (ICMLA)*, 2012a.

Randall Wald, Taghi M. Khoshgoftaar, and Ahmad Abu Shanab. The effect of measurement approach and noise level on gene selection stability. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2012b.

Randall Wald, Taghi M. Khoshgoftaar, and Amri Napolitano. Stability of filter- and wrapper-based feature subset selection. In *International Conference on Tools with Artificial Intelligence*. IEEE Computer Society, 2013.

Søren Wichmann and David Kamholz. A stability metric for typological features. *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, 61(3):251–262, 2008.

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

Lei Yu, Yue Han, and Michael E. Berens. Stable gene selection from microarray data via sample weighting. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 9(1):262–272, 2012.

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*, 2009.

Ding-Xuan Zhou. On grouping effect of elastic net. *Statistics & Probability Letters*, 83(9): 2108 – 2112, 2013.

Manuela Zucknick, Sylvia Richardson, and Euan A. Stronach. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statistical Applications in Genetics and Molecular Biology*, 7(1), 2008.