QUANTIFYING THE STABILITY OF FEATURE SELECTION

A THESIS SUBMITTED TO THE UNIVERSITY OF MANCHESTER
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
IN THE FACULTY OF SCIENCE & ENGINEERING

2018

By
Sarah Nogueira
School of Computer Science

Contents

Word Count: 25300

List of Tables

List of Figures

Notation

d	total number of features
X_f	feature f
n	number of samples in the data set
M	the number of bootstrap samples and the number of feature sets
$\mathcal{Z}$	the binary feature selection matrix of size $M \times d$ or the collection of M feature sets
$z_{i,f}$	the binary coefficient of $\mathcal{Z}$ at the i^{th} row and f^{th} column
Z_f	Bernoulli variable corresponding to the selection of the f^{th} feature
k_i	number of features selected on the i^{th} feature set
$\bar{k}$	the average number of features selected over the M feature sets in $\mathcal{Z}$
p_f	the population mean of Z_f
$\hat{p}_f$	observed frequency of selection of feature X_f, also the sample mean of variable Z_f
s_f^2	the sample variance of Z_f
$\hat{\Phi}(\mathcal{Z})$	a stability estimate
Φ	the population stability

Abstract

QUANTIFYING THE STABILITY OF FEATURE SELECTION
Sarah Nogueira
A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy, 2018

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 and axiomatic 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 of an approach, enabling rigorous comparison of feature selection algorithms.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

I would like to express my deep gratitude to my supervisor Prof. Gavin Brown, who has taught me how to do research, to teach, to write and to communicate my ideas. I would like to thank him for his great patience and for the dedication he has for his students. I would like to thank him for believing in my work, for giving me this amazing opportunity and making me the confident scientist I am today.

I would also like to thank the Centre for Doctoral Training (CDT) in Computer Science, which is funded by an Engineering and Physical Sciences Research Council (EPSRC) grant [EP/I028099/1].

I would like to thank all the people who supported me in my research. First, Dr. Kostas Sechidis who has been helping me from the first day I joined the MLO group and who followed all the steps of my research. He has become a real mentor to me and more importantly a close friend. Second, I would like to thank Dr. Nikos Nikolaou and Dr. Henry Reeve for all the great discussions we had, for their insights on my work, for making the office such a nice place–and for always making the best jokes :-) I would also like to thank Dr. Adam Pocock who took the time to give me such a detailed feedback on my work. I would like to thank our new member Kostas Papangelou for his help and I wish him the best of luck for the remainder of his PhD. Finally I would like to thank all the other people from the MLO group with whom it was a pleasure to share the office space, with a special thought for Jon with whom I had an amazing time in Italy.

I would like to thank the APT group for all the Friday after-works. In particular I would like to thank Ioanna, Thanos and Serhat, with whom I always had great fun. I would like to send my best regards to the visitors we had from Coruña, Diego, Veronica, Gabi and Laura with whom I spent memorable times in in Edinburgh and in Manchester.

I would like to thank all the friends I had during my years in Manchester. More specifically, I would like to thank Alkis for having been such a good friend and for

having helped me settling in UK. I would like to thank Eric and Danya for the beautiful times we spent together during my first year. I would like to thank Idoia who was always there for me in the bad times and in the good ones. I would like to thank Anatoli and Yaman for all the memories that I will keep with me. I would like to thank Mark, who was my confidant and a true friend to me.

I would also like to thank the ones who supported me from far away. My cousins Olegário and Nelson, my dear friends Julia and Jean, and all my other friends and relatives from Portugal and France.

Last but not least, I would like to thank from the bottom of my heart, my mother Bertelina, my sister Elisa, my uncle and my aunt Sebastião and Sandra, my cousins Victor and Christophe for all the love and strength they gave me in life and for always encouraging me to pursue my dreams.

Chapter 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. Stolovitzky [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 small change to 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 later published 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. [2007] 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 $P(X,C)$[1]. 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, the scope of this particular thesis 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 thesis, 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."*

[1] Where $P(X,C)$ is the joint probability distribution of class and training instances.

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., 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]. Each of these was justified and evaluated, though there has been little cross-comparison. The issue arises as to which one 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 of methods for stability assessment may lead to publication bias—in that researchers[2] 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 thesis, 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 then proceed to prove whether the 5 properties hold for each of 15 measures already proposed in the literature, and find no single measure satisfying all properties. We then derive a novel measure and estimator for which all properties provably hold (Chapter 5), with the following properties:

1. It is based on 5 simple properties (Chapter 4) which we will argue are essential

[2]Without appropriate guidance, the multiplicity of measures is rather overwhelming to the practitioner—the R package `OmicsMarkeR` provides 7 different options for measuring stability `www.rdocumentation.org/packages/OmicsMarkeR/versions/1.4.2/topics/pairwise.stability`

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 stability for two different algorithms on a data set.

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

In the following chapters we explain our framework, first embarking on a thorough critique of existing literature, including theoretical characterisation of many existing stability measures. 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/python/`
- A demonstration web page at `http://www.cs.man.ac.uk/~gbrown/stability`

1.3 Structure of this Thesis

The structure of this thesis is as follows. Chapter 2 provides background material on the problem of feature selection and on the quantification of stability. Chapter 3 provides an analysis of the existing measures of stability and provides insights on the context in which they have been introduced. Chapter 4 motivates the need for a property-based approach, refines the existing properties of the literature to more general cases and compares all the existing measures in terms of properties. Chapter 5 proposes a novel stability measure, showing that the measure possesses all desired properties, is a generalization of some other of the existing measures and provides statistical tools such as confidence intervals and hypothesis testing machinery. Chapter 6 first validates the different statistical tools from the previous chapter by a set of illustrative experiments. Then, it provides a set of experiments on artificial and real data sets, showing how the proposed framework can be used to tune hyperparameters, to confidently compare the stability of different algorithms and showing how the use of stability can benefit to a classic feature selection scenario. Complete proofs of the theorems that are not already given in the main body of this thesis are provided in Appendix A.

1.4 Contributions

The contributions of this thesis are:

1. A literature review of the existing stability measures.
2. A review of the properties required by the literature for a stability measure.
3. A generalized set of properties, applicable to any stability measure, that aggregates the different requirements made in previous literature.
4. A study of the properties of each one of the existing stability measures.
5. A novel stability measure that satisfies all properties and is a true generalization of some other existing measures.
6. A clean statistical interpretation of the proposed measure.
7. A link between the proposed measure and the statistical literature.
8. The population parameter being estimated and a statistical framework allowing the use of confidence intervals and hypothesis tests on the *true* value of stability.

9. A set of experiments illustrating the use of the statistical tools in practice and showing that:

    - in some scenarios, the stability of a feature selection procedure can be improved without any loss in accuracy of the resulting model;
    - enforcing stability can help avoiding the selection of irrelevant features.

The work presented in this thesis has resulted in two publications, with one journal submission under review:

Nogueira and Brown [2015]: Sarah Nogueira and Gavin Brown. Measuring the stability of feature selection with applications to ensemble methods. In *Multiple Classifier Systems - 12th International Workshop, MCS*, 2015

Nogueira and Brown [2016]: Sarah Nogueira and Gavin Brown. Measuring the stability of feature selection. In *ECML/PKDD*, pages 442–457, 2016

Nogueira et al. [2018]: Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection. *Under review*, 2018

We published one other work related on the topic of stability of feature selection for feature rankings but it is not part of this thesis:

Nogueira et al. [2017]: Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the use of spearman's rho to measure the stability of feature rankings. In *Pattern Recognition and Image Analysis: 8th Iberian Conference (IbPRIA)*. Springer International Publishing, 2017

Chapter 2

Literature Review

In this chapter, we first provide a literature review on feature extraction techniques (and more specifically on feature selection) and then we discuss stability issues in feature selection.

2.1 Feature Extraction

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 corresponding label. Building a predictive model using the whole set of features can lead to high-variance models that have over-fitted to the data, to a poor interpretability of these models, and can be very expensive in terms of computational costs. *Feature extraction* consists in using a combination of the initial set of d features to build a new set of features in a lower dimensional space. The new set of features obtained intends to be informative enough to build a predictive model or to be used by a domain-expert for interpretation. Two popular examples of feature extraction techniques are *Principal Component Analysis* and *Linear Discriminant Analysis* that use a linear projection of the initial set of features. In this thesis, we exclusively focus on a type of feature extraction called *feature selection*. The task of feature selection is to identify a subset of the features, of size $k < d$, that conveys the maximum information about the label. In the remainder of this section, we first describe feature selection techniques in general and then we further detail two categories of feature selection techniques, as they will be extensively used later in this thesis.

2.1.1 General

The challenge of feature selection can be tackled in various ways, commonly grouped in three feature selection families: filters, wrappers, and embedded methods [Guyon and Elisseeff, 2003].

Let us assume we want to evaluate the *predictive power* of all possible feature subsets according to some metric to select he best one. To do this, filters assign a score each feature set (possibly made of only one feature) based on statistics of the data *independently* of any particular model—for example mutual-information-based methods. They tend to be computationally more efficient than wrappers and embedded methods as they do not require to build predictive models. Wrappers, on the other hand, evaluate each feature subset based on the *error* of the predictive model using the given feature subset, and therefore, are model-specific. If we have d features, there are $2^d - 1$ possible feature sets to evaluate and assigning a score or building a predictive model for each possible subset is most often intractable. For this reason, filters and wrappers are often used along with a *search strategy* allowing to evaluate some of the feature subsets only. Popular examples of such strategies are greedy forward selection (or greedy backward elimination) in which the best feature is added (or the worst feature is deleted) at each round. Other common techniques include genetic algorithms or simulated annealing. Some filter techniques also propose to evaluate features individually and to select the top-k features with the highest scores. 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 as we will later see in Section 2.1.2.

A feature selection algorithm can have three different types of outputs:

1. A weighting (sometimes called scoring) on the features giving the relevancy of each feature to the target variable (e.g. the mutual information of each feature with the class label or the coefficients of a linear regression).

2. A ranking on the features giving the relative importance of each feature to the target variable (e.g. with recursive feature elimination). A weighting on the feature can be converted into a ranking, but the contrary is not possible.

3. A subset of the features indicating which features are relevant to the target variable (e.g. with any wrapper method). A feature set can be obtained by putting a threshold on a ranking (e.g. by selecting the top-k features), but the contrary is not possible.

In the following two sections, we review two categories of feature selection algorithms as they will be later used in Section 6.2 of this thesis. First, we focus on regularized methods which are a set of methods widely used to reduce over-fitting of predictive models and that sometimes select feature as part of the learning process (and therefore belong to the category of embedded feature selection procedures). This category of feature selection algorithms is particularly interesting as they are well-studied in the literature and several works show why some of them are expected to produce more stable *regression coefficients* in the presence of feature redundancy [Barla et al., 2008, Zhou, 2013]. It will be therefore interesting to observe the stability of their feature selection in different data scenarios. As such methods may select a different number of features on different samples of data, this is also an interesting scenario in the quantification of stability, as not all stability measures are defined in that scenario (c.f. Section 4.2). Second, we focus on some *information-theoretic-based* feature selection methods. These methods belong to the category of filter feature selection techniques and use mutual information between variables in different ways so that feature redundancy in the selected features is avoided. This scenario is also of interest as redundancy is known to be a source of instability and some of the experiments of Section 6.2 will have a closer look at these techniques.

2.1.2 Regularized Methods

In this section, we review some existing regularized methods methods in the context of a binary classification problem. Binary classification problems consist in predicting the class of a given example, when there is two classes. Let us pick one of the classes and call it "1" and the other "0". For example, given the characteristics of a patient, a classification problem could be to decide whether the patient is ill (class 1) or healthy (class 0).

Let $\mathbf{X}$ be the random vector for the d input variables and Y the binary response random variable. We want to model the conditional probability $\Pr(Y = 1|\mathbf{X})$ that the outcome Y is 1, given the input variables. A regression model for $\Pr(Y = 1|\mathbf{X})$ is given by

$$\Pr(Y = 1|\mathbf{X}) = \frac{1}{1 + \exp(-\beta_0 - \mathbf{X}\beta)},$$

where β_0 and $\beta = (\beta_1, ..., \beta_d)^T$ are the parameters we want to learn using our data. For simplicity, let us denote the conditional probability $\Pr(Y = 1|\mathbf{X})$ by $p(\mathbf{X})$ for some

function p. The estimation of the $d+1$ unknown parameters is done by maximum-likelihood estimation. Then for a test data point $\mathbf{x}_{test}$, we can compute the probability $p(\mathbf{x}_{test})$ that it belongs to class 1 and make a prediction $\hat{y}$ for that test example based on that probability (for example, take $\hat{y}=1$ if $p(\mathbf{x}_{test}) \geq 0.5$). The higher the log-likelihood, the better the model fits the data. Assuming that the n observations $\{\mathbf{x}_i, y_i\}_{i=1}^n$ are independent, the conditional likelihood is

$$\prod_{i=1}^{n} \Pr(Y = y_i | \mathbf{X} = \mathbf{x}_i).$$

Since $y_1, ..., y_n$ is a sequence of Bernoulli trials,

$$\Pr(Y = y_i | \mathbf{X} = \mathbf{x}_i) = \begin{cases} p(\mathbf{x}_i) & \text{if } y_i = 1 \\ 1 - p(\mathbf{x}_i) & \text{if } y_i = 0 \end{cases}.$$

Therefore, the likelihood can be re-written as

$$\prod_{i=1}^{n} p(\mathbf{x}_i)^{y_i} (1 - p(\mathbf{x}_i))^{(1-y_i)}.$$

Taking the logarithm of the above equation, we get the log-likelihood $\mathcal{LL}(\beta_0, \beta)$ of the parameters β_0 and β given the data as follows [1]

$$\mathcal{LL}(\beta_0, \beta) = \sum_{i=1}^{n} y_i \log p(\mathbf{x}_i) + (1 - y_i) \log(1 - p(\mathbf{x}_i)).$$

Then the optimal set of parameters is the solution to $\underset{\beta_0, \beta}{\operatorname{argmax}} \{\mathcal{LL}(\beta_0, \beta)\}$. Because this results in a system of $d+1$ non-linear equations, a numerical approach is typically used. This approach results in a model with $d+1$ parameters, which lacks interpretability, might use irrelevant features and might overfit. For this reason, additional constraints on the parameters can be added so that some of the coefficients β_f are equal to 0, hence reducing the dimensionality of the output model. The features having a non-zero corresponding coefficient form the set of selected features. In the next three sections, we will see in particular three regularized methods adding constraints to the objective function: the LASSO, the Ridge Regression and the Elastic Net [Hastie et al., 2001].

[1] This is commonly called the cross-entropy loss.

Logistic LASSO

One popular technique selecting features is the Least Absolute Shrinkage and Selection Operator (LASSO) [Tibshirani, 1994] . It solves the logistic regression problem adding an additional constraint as follows

$$\hat{\beta}^{lasso} = \underset{\beta_0,\beta}{\operatorname{argmax}} \{\mathcal{LL}(\beta_0,\beta)\} \text{ subject to } ||\beta||_1 \leq t, \tag{2.1}$$

where $||\beta||_1 = \sum_{f=1}^{d} |\beta_f|$ is the standard $L1$-norm and where t is a prespecified free parameter that determines the amount of regularisation. Taking the Lagrangian form of the above, we get

$$\hat{\beta}^{lasso} = \underset{\beta_0,\beta}{\operatorname{argmax}} \left\{ \frac{2}{n}\mathcal{LL}(\beta_0,\beta) - \lambda||\beta||_1 \right\}, \tag{2.2}$$

where $\lambda||\beta||_1$ is called the regularizing term and where $\lambda \geq 0$ is the regularizing parameter[2]. The $L1$-regularizing term will ensure that some of the coefficients β_f will be equal to 0. The hyperparameter λ controls the weight of the regularization. As we increase the value of λ, the number of coefficients equal to 0 will increase and therefore, less features will be included in the model. Therefore, LASSO implicitly performs feature selection: if $\beta_f \neq 0$, the f^{th} feature is selected in the model and if $\beta_f = 0$, it is not selected. In the next section, we introduce the Ridge Regression. Even though Ridge Regression is a regularized method that does not implicitly select features, its principles will be later used in in Section 2.1.2.

Ridge Regression

In this section, we describe the Ridge Regression [Hoerl and Kennard, 1988]. Similarly to LASSO, Ridge Regression uses a regularizing term that shrinks the coefficients β by imposing a penalty on their size as follows

$$\hat{\beta}^{ridge} = \underset{\beta_0,\beta}{\operatorname{argmax}} \left\{ \frac{2}{n}\mathcal{LL}(\beta_0,\beta) \right\} \text{ subject to } \lambda||\beta||_2 \leq t. \tag{2.3}$$

[2]There is a one-to-one correspondence between t in Equation 2.1 and λ in Equation 2.2 [Hastie et al., 2001].

where $||\beta||_2 = \sum_{f=1}^{d} \beta_f^2$ is the standard $L2$-norm. This solution is equivalent to

$$\hat{\beta}^{ridge} = \underset{\beta_0,\beta}{\operatorname{argmax}} \left\{ \frac{2}{n} \mathcal{LL}(\beta_0, \beta) - \lambda ||\beta||_2 \right\}.$$

The size constraint on the coefficients reduces the variance of the resulting model and is often used to avoid the overfitting of the regression to the data. Furthermore, Ridge Regression is known to have a grouping effect [Zhou, 2013]: if a group of variables are highly correlated among themselves, they will have similar coefficients. Nevertheless, ridge regression does not set any regression coefficient to 0, and therefore, does not select features. In the next section, we present a combination of Ridge Regression and LASSO, possessing characteristics of both methods.

Elastic Net

The Elastic Net is a compromise between the ridge regression and the LASSO, using both a $L1$ and a $L2$ regularizing term in the objective function as follows

$$\hat{\beta}^{enet} = \underset{\beta_0,\beta}{\operatorname{argmax}} \left(\frac{2}{n} \mathcal{LL}(\beta_0, \beta) - \lambda \left[\frac{(1-\alpha)}{2} ||\beta||_2 - \alpha ||\beta||_1 \right] \right),$$

where λ controls the weight of the overall regularization and where α controls the weight given to the $L1$ term. When $\alpha = 1$, this reduces to LASSO and when $\alpha = 0$, it reduces to Ridge Regression. The Elastic Net possesses the advantages of both techniques: it keeps the grouping effect of the ridge regression and selects the features as part of the learning procedure like LASSO [Kamkar et al., 2016, Baldassarre et al., 2017]. Therefore, we expect groups of correlated features to have similar coefficients and therefore to be either all selected or all not selected.

As some numerical techniques are sensitive to initial conditions and since only a finite subset of the data is provided to the user, the estimated parameters $\hat{\beta}$ might present some variability. This means that not the same features might be assigned zero-coefficients on different samples, causing the instability of the feature selection. In Section 6.2.1, we will compare the stability of the feature selection performed by LASSO and Elastic Net on a set of test cases with several degrees of feature redundancy.

In the next section, we look at some of the existing information-theoretic-based filtering criteria .

2.1.3 Information-theoretic-based Criteria

Let us assume we have two categorical random variables X and Y taking values in their respective alphabets $\mathcal{X}$ and $\mathcal{Y}$. The population *mutual information* between X and Y is defined as

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x;y)\ln\frac{p(x;y)}{p(x)p(y)},$$

where $p(x;y) = \Pr(X = x, Y = y)$ is the probability mass function of the joint distribution of X and Y and where $p(X)$ and $p(Y)$ are the probability mass functions of the marginal distributions. Since these probabilities are often unknown, we can use a point estimate $\hat{I}(X;Y)$ of the mutual information using the data is as follows

$$\hat{I}(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} \hat{p}(x;y)\ln\frac{\hat{p}(x;y)}{\hat{p}(x)\hat{p}(y)},$$

where $\hat{p}(x)$, $\hat{p}(y)$ and $\hat{p}(x;y)$ are the maximum-likelihood estimates of the probabilities [Paninski, 2003]. In the following sections, we discuss three information-based feature selection techniques using this point estimate.

Mutual Information Maximization

The Mutual Information Maximization (MIM) algorithm ranks each feature $X_1,...,X_d$ according to their mutual information with the target class Y. Then, the algorithm selects the top-k features with the highest mutual informations. In some data scenarios, there might be redundancy between the features. Let us assume that the first and the second feature are identical and both relevant to the target class. As they will both have identical rankings, if one is selected into the final set, the other one will be selected as well using the MIM algorithm. This phenomenon tends to produce feature sets of highly correlated features with redundant information, which tends not to be optimal. To tackle this, several variants of this algorithm have been proposed in the literature. We present two of them below.

Joint Mutual Information

The Joint Mutual Information (JMI) algorithm uses a forward selection approach where at each step it adds to the current feature set $\mathcal{S}$ the feature X_f having the maximal joint

mutual information with $\mathcal{S}$, that is

$$J_{JMI}(X_f) = \sum_{X_{f'} \in \mathcal{S}} \hat{I}(X_f X_{f'}; Y).$$

The algorithm stops when the set reaches a pre-defined cardinality k. The idea behind this algorithm is to take into account redundancy by focusing "*on increasing complementary information between features.*"[Brown et al., 2012][p.31]. Another algorithm correcting for feature redundancy is presented below.

Minimum Redundancy Maximum Relevance (mRMR)

The mRMR algorithm [Peng et al., 2005] is another algorithm taking into account redundancy between the features. Let us assume we are in a forward selection set-up where $\mathcal{S}$ is the set of currently selected features and where $\bar{\mathcal{S}}$ is its complementary (i.e. that is the set of currently non-selected features). At each step, it will add to $\mathcal{S}$ the feature X_f in $\bar{\mathcal{S}}$ maximizing the criterion

$$J_{mRMR}(X_f) = \hat{I}(X_f, Y) - \frac{1}{|\mathcal{S}|} \sum_{X_{f'} \in \mathcal{S}} \hat{I}(X_f; X_{f'}).$$

The idea behind this criterion is to identify the feature set with minimum redundancy and maximal relevancy.

In these three information-theoretic-based algorithms, the data available is used to get mutual information estimates, which are in return used to perform feature selection. The variance of these estimates will therefore play a role in the stability of the overall feature selection procedure. If on slightly different data, the mutual information estimates vary, a different feature set might be returned, causing the instability of the feature selection procedure. Furthermore, estimating mutual information can prove to be challenging as we need probability estimates $\hat{p}(x,y)$ for all x and y in their respective alphabets and this might be another source of instability.

In the next section, we define and disambiguate the meaning of stability in this thesis and provide a brief overview of the research questions around stability in the literature.

2.2 Stability of Feature Selection

In high dimensional data sets, a feature selection procedure is typically applied to obtain a data set with reduced dimensionality. Then the reduced data set is fed to a predictive model. In predictive models with a continuous output, there is often two measures of quality a user might want to look at: the *bias* and the *variance* of the predictive model. The bias measures how far on average the predicted output is from the true output and the variance measures the sensitivity of the output to perturbations on the data. The variance of a predictive model has been a topic of interest in the Machine Learning community and there exists several approaches to reduce the variance of a model, such as bagging or boosting. In linear regressions, there exists closed forms giving the variance of the estimated regression coefficients under certain assumptions. In addition, the expected generalisation error of some predictive models has been shown to be a linear combination of the bias and the variance of the model; showing the strong relationship between the two quantities. The study of bias and variance in predictive models does not consider the pre-processing step of feature selection, which plays an important role in the overall pipeline.

For a very long time, the only way to assess the quality of a feature selection procedure was to look at its predictive power, i.e. how *useful* the selected features are to predict the target variable. As highlighted in the previous section, there is a plethora of algorithms that have this purpose. In wrapper and embedded methods, the predictive power of a feature set is measured by the accuracy of a model using the selected features. Filter methods propose to use a proxy of the predictive power by using an easy-to-compute statistic for faster computation. More recently, the stability of the feature selection procedure was introduced as another way to assess the quality of a feature selection procedure. Stability can be seen as the analogous concept to the variance of a regression model: an unstable feature selection procedure with *good* predictive power is a sign of bad generalization as the feature selection has overfitted to a particular data set. Even though stability has only been a recent topic of interest in the community, it has been shown to be critical in many application. In early cancer detection, stability of the identified markers is a strong indicator of reproducible research [He and Yu, 2010, Lee et al., 2012] and therefore selecting a stable feature set of markers is said to be equally important as their predictive power [Jurman et al., 2008]. The study of the stability of feature rankers is also important in many fields. In information retrieval, ranking systems on search engines are expected to be robust to spam Dwork et al. [2001]. In bioinformatics, where by nature the training samples are usually small, the

removal of only one example on the training set can cause substantially different rankings making the feature rankings non-interpretable and not reliable for clinical use. For this reason, robust feature rankers have become a major requirement in the field of gene selection, biomarker identification or molecular profiling [Jurman et al., 2008, Boulesteix and Slawski, 2009, Abeel et al., 2010, Wald et al., 2012b, Dittman et al., 2013].

Stability is commonly defined in the literature as the *sensitivity of the feature selection procedure to training set perturbations*[Kalousis et al., 2007]. In the literature we came across several interpretation of what is meant by *perturbations*. As we will later see in Section 2.2.3, the term stability is often used in different context. For example, Altidor et al. [2011] refers to the robustness of a procedure to different levels of noise, while Shen et al. [2012] refers to the consistency of the feature sets obtained when the data is partitioned in non-overlapping chunks. For clarity, it is important to disambiguate the meaning of stability in this thesis. Hereafter, whenever we will refer to stability, we will use the meaning given by Definition 1.

Definition 1 *We define the stability of a given feature selection procedure as the variability of its output with respect to data sampling.*

We will later see in Sections 2.2.3 and 5.2.2, that depending on the sampling approach taken to quantify stability, the quantity measured does not necessarily match this definition. Even though some works use different sampling techniques, the work presented in this thesis up to Section 5.2 (statistical tools) still applies. The proposed stability measure given in Section 5 can be used in any scenario and its properties would be unchanged. Nevertheless, the statistical tools provided in Section 5.2 make the assumption that we are looking at the variability with respect to data sampling.

Another important point is to distinguish between stochastic and deterministic feature selection procedure. A deterministic procedure will always produce the same output if given the same data set as input. Therefore, the stability of feature selection is *only* measuring its sensitivity to sampling of the data in this case. In a stochastic procedure, the output might not be the same every time the procedure is applied to a same data set, due to the stochastic nature of the algorithm (that might be due to sensitivity to random initialization of parameters for example). In that situation, stability refers to both source of variation: the variation due to the intrinsic stochastic nature of the algorithm and the one inferred by the sampling of the data.

The research on stability is mainly articulated around 3 questions:

- What are the sources of instability?
- How can we produce stable feature selection algorithms?
- How to quantify stability?

In this section, we review some of the existing literature for each research question.

2.2.1 Sources of Instability

It is important to note *why* instability may occur, and what is commonly done about it—the sources of, and solutions to instability. Most empirical studies aim at comparing the stability of different feature selection algorithms in different settings or at observing the impact of data characteristics on stability. Several works argue that stability is mostly data-dependent and provide empirical studies looking at the effect of data characteristics on stability. A thorough study in this topic are the works of Alelyani [2013] and Alelyani et al. [2011], which study the impact of the dimensionality d, of the number of selected features k, of the sample size n or of the underlying data distribution. Other authors study how stability is influenced by the imbalance of the data set [Dittman et al., 2012] or by feature redundancy [Gulgezen et al., 2009, Wald et al., 2013]. Since all these aspects have empirically shown to play a role on the stability of feature selection procedures, a variety of frameworks has been proposed to tackle these sources of instability. In the next section, we review some of them.

2.2.2 Making Algorithms More Stable

We can categorize the frameworks for stability in 3 main approaches:

1. ensemble feature selection;
2. data variance reduction;
3. heuristic-based approaches that allow to pick a trade-off between stability and predictive power.

Ensemble Feature Selection

Ensembles of feature selection procedures are with no doubt the most popular way of making existing feature selection more stable. The underlying idea comes from the

field of *ensemble learning* where different classifiers are combined to vote on the class of an example. Ensembles of classifiers have empirically shown to reduce the variance without being at the cost of *performance* in many machine learning applications. Recently, this approach has been used in the context of feature selection, where several feature selectors are combined in order to reduce the variance of the feature selection output, hence making the overall feature selection more robust [Dunne et al., 2002, Saeys et al., 2008, Abeel et al., 2010].

Ensemble feature selection has two main components: a *randomization technique* to promote diversity in the feature selection outputs and an *aggregation technique* to combine the outputs. For randomization, *data perturbation* is classically applied to the original data set, so that M data samples are generated. Then the given feature selection procedure is applied to each one of the sampled data sets, generating M outputs. Data perturbation can be both operated through using different *sampling techniques* or distinct *feature subspaces*. The most common approaches are to take random subsamples of the data (e.g. Davis et al. [2006], Abeel et al. [2010]) or to take bootstrap samples[3] (e.g. Bach [2008]). Alternatively, Shen et al. [2012] propose to partition the data in non-overlapping chunks when the amount of available data is sufficient. Other techniques introduce further randomization in the feature selection process itself (e.g. by only allowing the feature selection to select from a random subset of the features at each step, as done in random forests [Breiman et al., 1984]). Some other works propose to use different feature selection algorithms on the same data set instead of using a sampling approach in order to get various feature selection outputs [Bouaguel et al., 2016].

To combine the M outputs, an aggregation technique is then used. Aggregation techniques vary with the type of output of the feature selection algorithm. For feature selection algorithms returning a subset of the features, the most common aggregation technique is to put a threshold on the number of times a feature has been selected across the M outputs. Another aggregation technique for feature sets is the work of Ditzler et al. [2014] which proposes a statistical testing framework to aggregate the features. It performs a statistical test on the number of times each feature has been selected to determine whether it belongs to the relevant feature set or not. Another technique that could be seen as an ensemble feature selection technique is the work of [Meinshausen and Bühlmann, 2010]. In this work, a framework called *Stability Selection* proposes

[3] A bootstrap sample is a random sample made of n examples taken from the given data set with replacement.

an aggregation technique based on the feature sets obtained with different regularizing parameters in LASSO. We will further discuss this framework in Section 6.2.2.

Data Variance Reduction

As data variance is known to be a source of instability, some works propose to reduce the variance of the data before performing feature selection. Han and Yu [2012] propose a variance reduction framework where samples that contribute more to the quantity of interest (such a quantity could be the mutual information with the target variable) are given more weight than the instances contributing less before performing feature selection. They provide an empirical evaluation of this instance-weighting approach for two feature selection algorithms.

Choosing a Trade-off

The *bias* of the selected feature subset represents how far on average the feature set returned is from the *true* relevant feature set. Since the *true* feature set is unknown, this is typically estimated by some measure-of-fit of a predictive model or by some statistical criterion (e.g. the mutual information). The variance of the feature set is measured by its stability. Many works acknowledge the need for a good trade-off between stability and predictive performance of the selected features [Haury et al., 2011, Dessì et al., 2015]. As opposed as for regression models with a continuous outcome, there is no explicit bias-variance decomposition for feature selection. For this reason, heuristic-based approaches propose to use as an objective function a combination of the performance of a model using the selected features and of the stability of the feature selection procedure. For example, Davis et al. [2006] defines an objective function as being a linear combination of the performance and of the stability as follows

$$\gamma \times Stability + (1-\gamma) \times Perf.$$

where $Perf$ is the performance of the output model built, $Stability$ is the stability of the feature selction procedure and where γ is a hyperparameter adjusted according to the relative importance that we want to give to stability and performance[4]. Similarly, Saeys et al. [2008] proposes to automatically balance stability and classification performance

[4]In this paper, γ is not optimized and is set to 0.5 to give equal importance to the predictive power and the stability of the feature selection procedure.

using as an objective function the harmonic mean of the two as

$$\frac{(\gamma'^2+1)\times Stability\times Perf}{\gamma'^2 Stability + Perf},$$

where γ' also controls the relative importance of stability and classification performance[5]. In Section 6.2, we will optimize both criteria and show that sometimes this can be done without loss of predictive power. Other heuristics to improve stability have been proposed. Gulgezen et al. [2009] proposes to modify the MID (Mutual Information Difference) optimization rule of the mRMR algorithm to improve stability. The modified MID incorporates a user-tuned parameter so that the user can tune the chosen amount of trade-off between relevancy and redundancy of the final feature set.

To be able to study the sources of instability and to carry out empirical works on the stability of feature selection, an agreed-upon measure, with known properties and behaviour is needed. Indeed, the use of different measures might lead to different conclusions and the choice of a measure is often biased by previous literature. Hence, the quantification of stability is critical to any empirical evaluation. In the next section, we review the existing frameworks to quantify stability, which will be the core topic of this thesis.

2.2.3 Quantifying Stability

Any empirical evaluation of stability, by definition, mandates the authors to *measure* stability—to know whether they have increased it, or decreased it.

As discussed earlier, feature selection procedures can have 3 types of outputs: *a weighting* on the features, a *ranking* on the features or a *feature set*. Consequently, there exist stability measures for each type of output.

As aforementioned, the stability of feature rankings is an important topic. Many stability measures have been proposed to the quantification of the stability of feature rankers, such as the average pairwise Spearman's rank correlation coefficient [Kalousis et al., 2007] or the Canberra distance [Jurman et al., 2008]. Nevertheless, we point that (1) a ranking can always be converted in a feature set by applying a threshold on the ranks (while the contrary is not possible) and (2) in many applications, the rankings are used to extract a feature set that will be fed to a predictive model. Another important

[5] In this work, γ' is also set so that stability and performance are equally important ($\gamma' = 1$).

issue is that the variability of the ranks of irrelevant features provides noisy information. Indeed, let us assume that features 1-5 are relevant features while features 6-10 are irrelevant. We would like a ranking algorithm to be stable for the relevant features, but it would not matter if the ranks of features 6-10 were different on different data samples as long as they are ranked after the relevant ones. On that account, many works apply a threshold on the rankings and look at the stability of the produced feature sets rather than measuring the stability of the rankings themselves [Altidor et al., 2011, Kuncheva et al., 2012, Wald et al., 2012c, Dessì and Pes, 2015]. Furthermore, some algorithms (such as forward feature selection) only provide a partial ranking on the features. Defining a measure in that situation can be challenging and simply selecting the top-k ranked features to compute the stability is a much simpler solution. For all these reasons, the remainder of this thesis focuses on the more general case of the quantification of stability when dealing with feature sets only. Nonetheless, a generalization of the proposed framework of this thesis to feature rankings is discussed in Section 7.2.2 as a possible avenue of investigation.

A typical approach to measure stability is twofold. First, we take M samples of the provided data set, to apply feature selection to each one of them, hence getting M feature sets. Second, a function $\hat{\Phi}$ taking as input the collection of the M feature sets is used in order to measure their variability. This is commonly called a *stability measure*. In this section, we discuss these two aspects by first very briefly over-viewing the existing stability measures for feature selection (as a much deeper analysis is given in the next chapter) and then we discuss the sampling techniques used in the literature.

Stability Measures

Let us assume we have a collection $\mathcal{Z} = \{s_1, ..., s_M\}$ of M feature sets, where each s_i is a subset of the features, how could we measure their variability? In the literature, we have distinguished two main approaches to this problem—the *similarity-based* approach and the *frequency-based* approach. A very natural way to represent the output of feature selection is as a *subset* of all the features. This representation has led to the similarity-based approach, introduced by Dunne et al. [2002]. Let ϕ be a function that takes as an input two feature sets s_i and s_j and that returns a value measuring the *similarity* between these two sets (e.g. the Jaccard Index). Then the stability $\hat{\Phi}(\mathcal{Z})$ is defined as the average pairwise similarities between the $M(M-1)$ possible pairs of

feature sets[6] in $\mathcal{Z}$:

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

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. The more the feature sets in $\mathcal{Z}$ are *similar* to each other on average over the M runs, 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. Many such similarity measures have been used to quantify stability. An alternative representation is to regard the feature choices as a binary string (1 for selected and 0 for not selected). This 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. Since any feature subset of a set of d features can equivalently be represented by a binary vector of length d, with a slight abuse of notation, we will interchangeably denote by $\mathcal{Z}$ a collection of M feature sets or the matrix made of M binary vectors (that we will formalize later) as they are two alternative representations of the same object. We depict this equivalence in Figure 2.1. In the next chapter, we review some all existing measures in each category, providing a cross-comparison and a critical analysis. We then look at the properties of these measures stated as necessary by the literature.

$$\begin{array}{lcl}
s_1 = \{X_1, X_2, X_3\} & \longleftrightarrow & \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ & & \cdots & & \\ & & \cdots & & \\ & & \cdots & & \\ & & \cdots & & \\ & & \cdots & & \\ 1 & 1 & 0 & 1 & 1 \end{bmatrix} \\
s_2 = \{X_2, X_3, X_5\} & \longleftrightarrow & \\
\\
\\
\\
\\
\\
s_M = \{X_1, X_2, X_4, X_5\} & \longleftrightarrow &
\end{array}$$

Figure 2.1: Two alternative (but representationally equivalent) forms for considering stability: set notation [LEFT] and binary matrix notation [RIGHT]. Starting from set notation encourages thinking about stability in terms of set intersection/union operations, whereas starting from the binary matrix encourage the statistical view, which we follow in this thesis.

[6] When the similarity measure ϕ is symmetric, this is usually more simply stated as $\frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \phi(s_i, s_j)$.

Sampling Technique

The first step to quantify the stability of feature selection is to apply the feature selection procedure to M samples of the data. So, one choice has to be made towards the sampling technique to get the M samples.

Most works propose to use bootstrap samples to measure stability of feature selection techniques. However, other works propose to use other sampling techniques. For example, Wald et al. [2012c] and Altidor et al. [2011] inject different levels of noise in the data to obtain the M samples. The aim of this approach here is to measure the effect of different levels of noise in the data on the stability. Wald et al. [2012a] proposes an algorithm to get M samples with a fixed amount of overlap and study how the level of overlap affects the stability of different feature selection algorithms. Davis et al. [2006] and Abeel et al. [2010] propose to take random sub-samples of the data (which consists in taking examples without replacement); while Shen et al. [2012] proposes to partition the data in non-overlapping chunks. The motivations behind the use of a particular sampling technique often remains unclear. In this thesis, we adopt the bootstrap approach due to its well understood properties and familiarity to the community. Furthermore, we will show in Section 5.2.2 why the bootstrap approach is relevant in the context of the statistical framework we propose. We emphasize that the proposed stability measure and its properties (c.f. Section 5.1) *do not* depend on the sampling technique used. Nevertheless, as later explained in Section 5.2.2, the statistical tools provided in this thesis are only guaranteed to be valid when using bootstrap sampling.

In the next chapter, we provide a deep analysis of the existing stability measures for feature sets.

Chapter 3

Analysis of Existing Stability Measures

In this chapter, we critically review several published stability measures and provide a description of the different properties stated as desirable for a stability measure in the literature. The purpose of this review is to demonstrate the overwhelming variety in the literature, and motivate the need for a *property-based* approach.

3.1 Similarity-based Measures

In this section, we review the existing similarity-based measures used in the context of stability. Table 3.1 summarizes all measures, along with their bounds.

3.1.1 Jaccard Index (Set Intersection/Union)

Kalousis et al. [2005] used the Tanimoto similarity, equivalent to the Jaccard index:

$$\phi(s_i, s_j) = \frac{r_{i,j}}{|s_i \cup s_j|},$$

where $r_{i,j} = |s_i \cap s_j|$ is the size of the intersection between feature sets s_i and s_j and where $|s_i \cup s_j|$ is the size of their union. This measure takes the size of the intersection between the two sets, normalized by the size of their union—representing the proportion of agreement between the two sets. It is interesting to note that several other measures also use the size of the intersection as a core component, but normalize it in different ways. The measures of Shi et al. (*POG*) and Zucknick et al. [2008] (Ochiai's index) reviewed below are examples in this class.

3.1.2 Dice-Sørenson Index (Normalized Set Intersection)

The Dice-Sørenson index which was first used in the context of stability by Yu et al. [2008] is defined as

$$\phi(s_i, s_j) = \frac{2r_{i,j}}{|s_i| + |s_j|}.$$

With a simple re-writing, this can be seen to normalize the intersection by the average size of the two feature sets.

3.1.3 Percentage Overlapping Genes (*POG*)

Shi et al., pg 5 study the replicability of experiments in microarrays – and propose to measure the percentage of overlapping genes. This results in a measure similar to the Dice-Sørenson index, though normalized by the size of just one of the feature sets:

$$\phi(s_i, s_j) = \frac{r_{i,j}}{|s_i|}.$$

As a result, *POG* is non-symmetric. Interestingly, Ochiai's index (below) is a symmetrised version of this measure.

3.1.4 Ochiai's index (Geometric Mean of the Ratio of Shared Features)

Zucknick et al. [2008] proposes the use of Ochiai's index, defined as the geometric mean between $\frac{r_{i,j}}{|s_i|}$ and $\frac{r_{i,j}}{|s_j|}$, that is

$$\phi(s_i, s_j) = \frac{r_{i,j}}{\sqrt{|s_i||s_j|}}.$$

3.1.5 Normalized Hamming Similarity (Symmetric Set Difference)

Most measures regard similarity as a measure of agreement on the *selected* features only. Another way to quantify similarity is to examine agreement on the *unselected* features as well. This is achieved using the Hamming *similarity* between subsets—proposed by Dunne et al. [2002]. The (normalized) Hamming distance between subsets is written in set operations as

$$\phi(s_i, s_j) = 1 - \frac{|s_i \setminus s_j| + |s_j \setminus s_i|}{d}.$$

This measure represents the proportion of agreement of the two sets in the non-selected and the selected features (i.e. the proportion of features that are in neither of the sets or in both).

3.1.6 Set Intersection as a Hypergeometric Random Variable

Kuncheva's Measure

Kuncheva [2007] recognised that natural variations in training data made the selection of features a *random process*, and proposed to model the intersection between feature sets s_i and s_j as a hypergeometric random variable. Let us assume that a feature set of fixed size k is repeatedly drawn, where each of the $\binom{d}{k}$ possible sets has an equal probability of being drawn. Under this assumption, the size of the set intersection $r_{i,j}$ follows a hypergeometric distribution, with expectation $\frac{k^2}{d}$. The expectation can be interpreted as the size of the intersection *due to chance*. Based on this observation, Kuncheva's proposal is to correct the size of the intersection by this term so that the similarity value between any two feature subsets can be interpreted as their *agreement above chance*. This value is then re-scaled by its maximum value such that the measure yields values in the interval $[-1,1]$. This yielded a similarity measure where a value above 0 can be interpreted as a similarity greater than that due to chance, given by

$$\phi(s_i,s_j) = \frac{\text{Observed } r_{i,j} - \text{Expected } r_{i,j}}{\text{Maximum } r_{i,j} - \text{Expected } r_{i,j}} = \frac{r_{i,j} - \frac{k^2}{d}}{k - \frac{k^2}{d}}. \tag{3.1}$$

It is important to distinguish Kuncheva's measure from others discussed so far. The similarity measures presented so far have all been deployed to quantify stability, and the choice of a particular measure is often biased by previous literature in a particular community or the motivation behind the choice is often unclear, often unstated. In contrast, Kuncheva pioneered an *axiomatic* approach, judging candidate stability measures objectively on their *properties*, such as whether they have a known upper bound, or whether they are, as Equation 3.1, corrected for the agreement due to chance. We look deeper at this approach in Section 3.3 and Chapter 4.

One limitation of Kuncheva's analysis is that the procedure is assumed to always choose a subset of fixed size, k. It is instructive to consider the alternative, when we may pick sets of differing size, k_i and k_j. In this case, the hypergeometric distribution simply has an expectation of $\frac{k_i k_j}{d}$. As we will see below, it turns out not to be so straightforward to generalise the similarity measure appropriately, as several proposals

have been made with that objective.

Lustgarten's Measure

Lustgarten et al. [2009] extended Kuncheva's measure by allowing the number of selected features to vary, proposing the following similarity measure:

$$\phi(s_i, s_j) = \frac{\text{Observed } r_{i,j} - \text{Expected } r_{i,j}}{\text{Maximum } r_{i,j} - \text{Minimum } r_{i,j}} = \frac{r_{i,j} - \frac{k_i k_j}{d}}{\min(k_i, k_j) - \max(0, k_i + k_j - d)}.$$

This measure still yields values in $[-1;1]$. Note that Lustgarten et al. [2009] deviated from the original definition of Kuncheva [2007] by dividing the numerator by the range of $r_{i,j}$ instead of dividing it by its maximum value.

Wald's Measure

Wald et al. [2013] also extended Kuncheva's measure by allowing the number of selected features to vary, proposing the similarity measure

$$\phi(s_i, s_j) = \frac{\text{Observed } r_{i,j} - \text{Expected } r_{i,j}}{\text{Maximum } r_{i,j} - \text{Expected } r_{i,j}} = \frac{r_{i,j} - \frac{k_i k_j}{d}}{\min(k_i, k_j) - \frac{k_i k_j}{d}}. \tag{3.2}$$

While this appears to be the natural generalisation of Kuncheva's measure as it has the same formulation, in Section 4.2, we will see that it is not a perfect generalisation as it looses some of the properties of Kuncheva's measure.

3.1.7 Normalized *POG* (*nPOG*)

In the bioinformatics literature, Zhang et al. [2009] define the normalized Percentage of Overlapping Genes (*nPOG*) that measures the agreement between gene lists *above chance* as

$$\phi(s_i, s_j) = \frac{\text{Observed } r_{i,j} - \text{Expected } r_{i,j}}{\text{Size } s_i - \text{Expected } r_{i,j}}. \tag{3.3}$$

It is interesting to note that this has almost the exact same formulation as Kuncheva's and Wald's measures, though derived and argued for in a completely different body of literature. The required expectation being less well known in this community, they propose to estimate its value "*by the average of the scores for 10000 pairs of random lists*" of size k_i and k_j [Zhang et al., 2009, pg 1664]. Nevertheless, it is interesting to see how from independent bodies of literature, the same correction was proposed.

Table 3.1: 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, Chelvan 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 $r_{i,j} = 0$).

First used in	**Name**	**Measure**	$[\min, \max]$
Dunne et al. [2002]	Hamming	$1 - \frac{\lvert s_i \setminus s_j \rvert + \lvert s_j \setminus s_i \rvert}{d}$	$[0,1]$
Kalousis et al. [2005]	Jaccard	$\frac{\lvert s_i \cap s_j \rvert}{\lvert s_i \cup s_j \rvert}$	$[0,1]$
Yu et al. [2008]	Dice-Sørenson	$\frac{2\lvert s_i \cap s_j \rvert}{\lvert s_i \rvert + \lvert s_j \rvert}$	$[0,1]$
Zucknick et al. [2008]	Ochiai	$\frac{\lvert s_i \cap s_j \rvert}{\sqrt{\lvert s_i \rvert \lvert s_j \rvert}}$	$[0,1]$
Shi et al.	POG	$\frac{\lvert s_i \cap s_j \rvert}{\lvert s_i \rvert}$	$[0,1]$
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	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{k_i - \frac{k_i k_j}{d}}$	$[1-d,1]$

Kuncheva's measure was the seminal work in this 'random variable' approach, inspiring the work of Wald et al. [2013] (and others) with intuitive modifications. However, as we will see in Section 4.2, the modifications change the properties of the measures, in unexpected and undesirable ways. Lemma 1 states this formally.

Lemma 1 (Equivalences Between Measures) *If we select a fixed number of features, k, the measures derived by Wald et al. [2013] and by Zhang et al. [2009] are equal to*

Kuncheva's stability measure, that is

$$\hat{\Phi}_{Kuncheva} = \hat{\Phi}_{Wald} = \hat{\Phi}_{nPOG}.$$

Proof. *When replacing the cardinalities k_i and k_j by the constant k in Equations 3.2 and 3.3, we can see that Wald and nPOG similarity measures reduce to Kuncheva's measure. Therefore, the corresponding stability measures are identical when the number of features selected is constant and equal to k.*

All the stability measures presented in this section are based on set-theoretic notation. Indeed, simply referring to feature *subsets*, we think naturally of *set* operations. However, in the next section we discuss an alternative approach which is more amenable to a statistical treatment.

3.2 Frequency-based Stability Measures

An alternative representation of a feature set is 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, as illustrated in Figure 2.1. 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 samples 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 this section, we review the existing measures using this representation, which we will call the *frequency-based* measures, as they all look at the frequency of occurrence of a feature, or a feature set. We denote the frequency of selection of the f^{th} feature as $\hat{p}_f = \frac{1}{M}\sum_{i=1}^{M} z_{i,f}$, which is the sample mean of the f^{th} column of $\mathcal{Z}$. Table 3.2 summarizes all measures in this category, along with their bounds.

3.2.1 Average Frequency of Selection

Goh and Wong [2016] propose to use the frequency of selection, averaged over all features,

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

3.2.2 Corrected Frequency of Selection

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

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

where F is the number of features selected in at least one of the M feature sets (in other words, $F = |\cup_{i\in\{1,\ldots,M\}} s_i|$) and where α is a hyperparameter chosen by the user.

3.2.3 Entropy of Feature Sets

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

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

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. The use of this measure implies a relatively large number feature sets in the collection $\mathcal{Z}$, as we need frequency estimates for every feature set of size k, which makes it challenging to estimate with reasonable values of M. Furthermore, we note hat this measure also assumes that the number of features selected by the feature selection algorithm is constant.

3.2.4 Entropy of the Selection of Each Feature

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.

3.2.5 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

$$\hat{\Phi}(\mathcal{Z}) = \frac{d\left(M\bar{k} - D + \sum_{f=1}^{d} M\hat{p}_f(M\hat{p}_f - 1)\right) - (M\bar{k})^2 + D^2}{d\left(H^2 + M(M\bar{k} - H) - D\right) - (M\bar{k})^2 + D^2}, \tag{3.4}$$

where $D = (M\bar{k})\ mod\ d$ and $H = (M\bar{k})\ mod\ M$. As we can see, this definition is a function of the observed frequencies of selection $\hat{p}_f$ of each feature. Surprisingly, we found that CW_{rel} could also be seen as the extension of Kuncheva's stability measure. Indeed, Theorem 1 shows that when the procedure selects a constant number of features, the two measures are asymptotically equivalent.

Theorem 1 (Asymptotic Equivalence of CW_{rel} and Kuncheva) *When the number of features selected is constant and equal to k, the relative weighted consistency is asymptotically equivalent to Kuncheva's stability measure.*

$$\hat{\Phi}_{CW_{rel}} \underset{M \to +\infty}{\sim} \hat{\Phi}_{Kuncheva}.$$

Proof. *First, when the number of selected features is constant, $\bar{k} = k$ which is an integer and therefore $H = 0$, which simplifies the expression of $\hat{\Phi}_{CW_{rel}}(\mathcal{Z})$. Then, by re-writing $\hat{\Phi}_{CW_{rel}}(\mathcal{Z})$ as a function of the average pairwise intersections $r_{i,j}$ between the feature sets in $\mathcal{Z}$, we show that $\lim_{M \to +\infty} \frac{\hat{\Phi}_{CW_{rel}}(\mathcal{Z})}{\hat{\Phi}_{Kuncheva}(\mathcal{Z})} = 1$, which proves that the two stability measures are asymptotically equivalent. The full proof of this theorem is given in Appendix A.1.*

3.2.6 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. Because of the i^2 term, the features with a higher observed frequency of selection $\hat{p}_f$ will contribute more to the stability value than the features with a lower frequency of selection.

Table 3.2: Non-pairwise stability measures proposed in the literature. In Section 5, we propose a novel measure in this class.

First used in	**Measure**	$[\min, \max]$
Goh and Wong [2016]	$\frac{1}{d}\sum_{f=1}^{d} \hat{p}_f$	$[0,1]$
Davis et al. [2006]	$\max\left(0, \frac{1}{F}\sum_{f=1}^{d} \hat{p}_f - \alpha\frac{med(k_1,\ldots,k_M)}{d}\right)$	$[0,1]$
Krízek et al. [2007]	$-\sum_{s_i \in \mathcal{Z}} \hat{p}(s_i)\log_2 \hat{p}(s_i)$	$[0, \log(\min(M, \binom{d}{k}))]$
Guzmán-Martinez (2011)	$1 - \frac{D_{JS}(\mathbf{q}_1,\ldots,\mathbf{q}_M)}{D^*_{JS}(\mathbf{q}_1,\ldots,\mathbf{q}_M)}$	$[0,1]$
Somol and Novovičová [2010] CW_{rel}	$\frac{d\left(M\bar{k} - D + \sum_{f=1}^{d} M\hat{p}_f(M\hat{p}_f - 1)\right) - (M\bar{k})^2 + D^2}{d\left(H^2 + M(M\bar{k} - H) - D\right) - (M\bar{k})^2 + D^2}$	$[0,1]$
Lausser et al. [2013]	$\frac{1}{M^2 k}\sum_{i=1}^{M} i^2 \sum_{f=1}^{d} \mathbb{1}\{\sum_{j=1}^{M} z_{j,f} = i\}$	$\left[\frac{1}{M}, 1\right]$

3.2.7 Summary

In total, we identified 15 distinct stability measures. Some measures are similar to each other, while some appear completely different. The various notations have obscured equivalences between them. In the next section, we step back and look at objectively desirable properties.

3.3 The Properties of 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. Kuncheva [2007] pioneered the property-based approach, proposing a set of 3 properties that a similarity measure[1] should possess as given in the box below. We remind the reader that the *similarity* measure $\phi(\cdot,\cdot)$ between two feature sets is averaged over all pairs to obtain a *stability* measure $\hat{\Phi}$.

Kuncheva's properties for a Similarity Measure ϕ

1. *Monotonicity*. For a fixed subset size, k, and number of features, d, the larger the intersection between the subsets, the higher the value of the consistency index.
2. *Limits*. The index should be bounded by constants which do not depend on k or d. The maximum value should be attained when the two subsets are identical, i.e., for $r_{i,j} = k$.
3. *Correction for chance*. The index should have a constant value for [random] independently drawn subsets of features of the same cardinality, k.

The first property, *Monotonicity*, ensures that the similarity value increases with the size of the intersection of the two sets. This property is *implicitly* defining what stability is: a stable feature selection procedure will produce feature sets with a large intersection. Although Kuncheva [2007] is most well-known for introducing the *Correction for chance* (3^{rd} property), we regard *Monotonicity* as the most fundamental property as it is implicitly defining the concept of stability.

The second property, *Limits*, concerns the upper/lower bounds of the similarity measure. However, we point out that it contains *two clauses*. First, it states that the similarity values should be bounded by constants not depending on k or d. Second, when two subsets are identical, the maximum value should be attained. However as

[1] We have mentioned the variety of terminology in the literature. Kuncheva refers to ϕ as a "consistency index", whereas in most other works it is called a "similarity measure".

we will show, *it is possible for the first clause to hold while the second does not, and vice versa*. Thus, part of *Limits* could hold, while the other does not. This will be important in the following sections.

The third property, *Correction for chance*, was a novel concept introduced in the field by Kuncheva (but already existing in the statistical literature [Berry et al., 2016]). This states that whenever we have *independently drawn subsets* at random, the stability value should be constant in expectation.

Since this seminal work, the feature selection community has largely converged around this approach and many other authors have identified these 3 properties as desirable if not essential. Among them, we can cite the following works:

- Somol and Novovičová [2010], Zucknick et al. [2008], Guzmán-Martínez and Alaiz-Rodríguez [2011] require a stability measure to be bounded by constants;
- Guzmán-Martínez and Alaiz-Rodríguez [2011] require that $\hat{\Phi}(\mathcal{Z})$ reaches its maximum whenever all the feature sets in $\mathcal{Z}$ are identical;
- Zhang et al. [2009], Lustgarten et al. [2009], Guzmán-Martínez and Alaiz-Rodríguez [2011] require a measure to be corrected by chance.

As is evident, the literature after Kuncheva followed the property-based approach and motivated the need for the same properties. This shows that unknown properties of stability measures are an issue in the community and strengthens the need for these properties.

In this section, we motivated the need for a property-based approach and showed that the construction of several measures has been based on the properties derived by Kuncheva [2007]. We would like to compare the existing measures in terms of properties and verify which stability measures verify which properties. In this approach, we faced the issue that the properties proposed by Kuncheva [2007] are given for similarity measures ϕ (e.g. Jaccard Index) and hence cannot be applied to non-similarity-based measures (e.g. CW_{rel} [Somol and Novovičová, 2010]).

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

First, properties for ϕ do not necessarily imply properties for $\hat{\Phi}$. For instance, a similarity measure ϕ that is not bounded by constants independent of k and d can result

in a stability measure $\hat{\Phi}$ that is bounded. Second, transposing these properties for the stability $\hat{\Phi}$ is not necessarily straightforward. For instance, we can wonder how the property *Monotonicity* defined for a pair of feature sets would translate when looking at the whole collection $\mathcal{Z}$ of feature sets. This stresses the need of an equivalent phrasing for the stability measure $\hat{\Phi}$, which will be the topic of the next chapter.

We identified two other desirable properties that recur across the literature. The first property is the need for a measure that is also defined in scenarios in which the number of selected features is not constant (and this is the reason why other works have proposed *extensions* of Kuncheva's measure as we have seen in Section 3.1 [Zhang et al., 2009, Lustgarten et al., 2009, Yu et al., 2008]). The second one concerns the symmetry of the similarity measure (i.e. $\phi(s_i, s_j) = \phi(s_j, s_i)$) [Alelyani, 2013, Chelvan 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 $\hat{\Phi}$. So the symmetry of ϕ is of little importance when comparing the properties of the stability measure $\hat{\Phi}$.

In the next chapter we shift our perspective from the *similarity* measure to the *stability* measure and refine the axioms in the literature so they apply to a stability measure $\hat{\Phi}$ for any size of feature sets. We then study the properties of all existing measures and illustrate why these properties are necessary to the comparison and the interpretability of stability values.

Chapter 4

A New Set of Properties for Stability Measures

We propose 5 general properties which we will argue are desirable in most (if not all) application domains. In this chapter, we refine and aggregate the list of desired properties of the literature for a stability measure $\hat{\Phi}$ in the general case where the number of features selected might be variable. This will allow us to show which non-pairwise stability measures possess which of the properties.

4.1 Refining the Properties

The first property we argue for is that any stability measure $\hat{\Phi}$ should be able to cope with any collection $\mathcal{Z}$ of feature sets—that is, it has a defined value even when the number of selected features varies with data perturbations. This way, we could compare the stability of feature selection algorithms of different types. Stability measures not having this property will not be applicable for a wide class of feature selection algorithms, such as LASSO (c.f. Section 2.1.2). The remaining 4 properties are all generalized from the ones of Kuncheva [2007]—which, as we have seen in the previous chapter, encompass all the properties described as desirable by many authors. We call this property *Fully defined*.

The first property proposed by Kuncheva [2007] states that when the number of features selected is constant, then the measure ϕ should be an increasing function of the size of the intersection of the two sets. 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 $\mathcal{Z}$ with feature set sizes $k_1,...,k_M$, the stability measure $\hat{\Phi}$ should be an increasing function of the average pairwise intersections* $\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M} r_{i,j}$. However, a question arises in the more general case: what this does translate to in the binary representation and how can we verify that non-similarity-based measures possess this property? Theorem 2 bridges the two approaches and gives us our second property which generalizes Kuncheva's Monotonicity property[1].

Theorem 2 *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*

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

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[2] *and where $\bar{k} = \frac{1}{M}\sum_{i=1}^{M} z_{i,f}$ is the average number of feature selected over the M feature sets.*

Proof. *In this proof, we use two key equations. First, we prove that $\sum_{f=1}^{d}\hat{p}_f = \bar{k}$ (c.f. Equation A.1). Then, we note that the intersection $r_{i,j}$ between feature sets s_i and s_j is the dot product of the two binary vectors representing the feature sets, that is $r_{i,j} = \sum_{f=1}^{d} z_{i,f}z_{j,f}$. Starting from the right hand side of the equation, we use these two results and by factorizing the expression appropriately, we end up in the right-hand side of the equation. The full proof of this theorem is given in Appendix A.2.*

As we can see from Theorem 2, 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. We also note that 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. Nevertheless, we will show 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

[1] By monotonicity, we mean strict monotonicity, that is for a function g defined on a set D_g, g is a strictly 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 increasing function of the intersection $r_{i,j}$ but cannot be interpreted as the similarity between two feature sets.

[2] This expression of the sample variance is derived from a Bernoulli distribution.

the choice features and (3) that definition allows to derive a statistical framework for stability estimates.

Kuncheva's Limits property contains two clauses. We noted earlier that it is possible for a measure to satisfy one clause but not the other. In fact, we will later see in Table 4.1 of Section 4.2 that Davis et al. [2006] measure satisfies the first clause and not the second one while the *nPOG* measure satisfies the second and not the first one. Since this is the case, we propose to split this property into two distinct requirements. Firstly, the property states that the similarity between two feature sets should be bounded by constants. This naturally implies that the stability measure $\hat{\Phi}$ should be bounded by constants, which will be our third property called *Bounds*. For meaningful *interpretation* of a stability measure and comparison across problems, the range of values of a stability measure should be known and finite. Secondly, Kuncheva's second property states that IF two feature sets s_i and s_j are identical THEN their similarity $\phi(s_i, s_j)$ is maximal, which is something we would want. But if the stability is maximal, this does not guarantee that the two sets are identical. This is something desirable, since we would like to be able to differentiate the situation where feature sets are all identical from the situation where they are not. Therefore, we made Kuncheva's property stronger by making it a double implication. When looking at the stability measure, this translates to: All feature sets in $\mathcal{Z}$ are identical *if-and-only-if* (IFF) $\hat{\Phi}(\mathcal{Z})$ reaches its maximum. This gives us our third property called *Maximum*.

Before we move on to the property of *Correction for chance*, we introduce some terminology in relation to the concept of *randomness*. When the number of features selected is constant, the notion of *random* selection is intuitive: this 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 in which the procedure does *not guarantee* to return a constant number of features, and thus produces a collection $\mathcal{Z}$ of feature sets of variable size. We can still define the concept of *randomness*: 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 the different authors using the concept of correction for chance, e.g. Kuncheva [2007], Zhang et al. [2009], Lustgarten et al. [2009], Guzmán-Martínez and Alaiz-Rodríguez [2011]. Since we will use this concept multiple times in the remainder of the thesis, we formalize this in Definition 2 and will refer to this as the *Null Model of Feature Selection*[3].

[3]The given definition holds for any $i \in \{1, ..., M\}$.

Definition 2 (The Null Model of Feature Selection H_0) *If we draw a subset of k_i features from a set of size d, we define the* Null Model of Feature Selection *as the situation where all possible $\binom{d}{k_i}$ feature sets of size k_i have an equal probability of being drawn.*

We denote the Null Model of Feature Selection as H_0. Thus, in this notation, Kuncheva's similarity can be re-written as

$$\phi(s_i, s_j) = \frac{r_{i,j} - \mathbb{E}\left[r_{i,j}|H_0\right]}{\max(r_{i,j} - \mathbb{E}\left[r_{i,j}|H_0\right])},$$

where $\mathbb{E}\left[r_{i,j}|H_0\right]$ is an abbreviated notation standing for the expected value of $r_{i,j}$ given that *both* s_i and s_j are feature sets drawn under H_0.

The last proposed property, *Correction for chance*, will be that under the Null Model of Feature Selection H_0, the expected value of $\hat{\Phi}$ should be constant, which we set for convenience to 0. 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 the one due to the systematic decision-making of the feature selection algorithm. To illustrate this, imagine that we have 3 features in total X_1, X_2 and X_3 and that we have a random procedure selecting 2 features. Since the procedure is random, it will return one of the following three subsets on each repeat: $\{X_1, X_2\}$, $\{X_1, X_3\}$ or $\{X_2, X_3\}$ with equal probabilities, as these are all possible subsets made of 2 features. As we can see, all six pairs of subsets share at least one feature, even though the algorithm is randomly picking up the features. If we were to measure the similarity of any pair of these subsets by a non-corrected measure, since they overlap, we would not get a minimum value. As shown by Kuncheva [2007], this makes the stability values biased by the feature set size.

In this section we have generalized the set of desirable properties (summarized in the box below) so that it can be applied to any stability measure (whether it is similarity-based or not) and to any any feature selection algorithm (whether it selects a constant number of features or not). This more general approach will allow us to compare all existing measures in terms of properties and will allow us to observe which measure possesses which one of the properties, which is the topic of the next section.

Our proposed properties for a stability measure $\hat{\Phi}$

1. *Fully defined.* The stability estimator $\hat{\Phi}$ should be defined for any collection $\mathcal{Z}$ of feature sets.
2. *Monotonicity.* The stability estimator $\hat{\Phi}$ should be a strictly decreasing function of the sample variance s_f^2 of the selection of each feature.
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.* The stability $\hat{\Phi}(\mathcal{Z})$ 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.

4.2 Which of these properties hold for each measure?

In this section, we study the properties of each one of the existing stability measures of the literature. We briefly comment and illustrate each property and summarize these findings in Table 4.1. All proofs (or proof sketches) for this section are provided in Appendix B. Our results show that although these properties have been previously stated as desirable in the literature, none of the existing measures possesses all 5 of them.

4.2.1 Fully Defined

As we can see from their definitions (c.f. Tables 3.1 and 3.2), 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 and therefore do not have this property.

4.2.2 Strict Monotonicity

As explained in the previous section, this property could be regarded as implicitly defining what is stability and therefore other proposals could be made. For this reason, we will not discard the measures not having this property. In spite of that, it is interesting to note that, as we can see in Table 4.1, 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, which is why we will later propose in Chapter 5 a measure that also possesses this property.

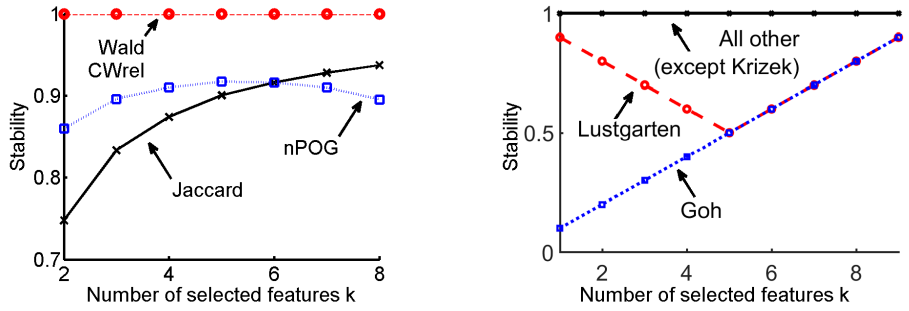
4.2.3 Bounds

Some of the measures such as Krízek's measure [Krízek et al., 2007] do not possess this property, since we can see from Table 3.2 that its maximum value depends on the number of features selected k and on the total number of features d. Unbounded measures will not allow a meaningful interpretation of stability values across problems or for different number of features selected, which can be restrictive in many applications.

4.2.4 Maximum Stability $\leftrightarrow$ Deterministic Selection

To illustrate this property, we used two scenarios, distinguishing the forward implication from the backward implication. First, we generated a collection of feature sets $\mathcal{Z}$ in which all 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 expect stability values not to be equal to their maximum. We plotted stability values against k for $d = 10$ and $M = 100$ in Figure 4.1a for some stability measures. We can see that Wald and CW_{rel} 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 expect all stability values to be equal to their maximum. We plotted the stability values against the number of features selected k in Figure 4.1b. Even though the selection is completely deterministic, Lustgarten takes variable stability values depending on the number of selected features k. This shows that Lustgarten's measure

violates the backward implication of this property.



(a) Forward implication (b) Backward implication

Figure 4.1: 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 [Lustgarten et al., 2009] violates the backward implication[4].

4.2.5 Correction for Chance

For illustrative purposes, we reproduced the experiment of Kuncheva [2007] in Figure 4.2. Let us assume that a feature selection procedure **randomly** selects k features out of $d = 10$ features and that we measure the stability based on $M = 100$ feature sets 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 k. For instance, we can see that using the Dice similarity measure, the stability systematically increases with the number of features selected, hence being in favour of larger feature sets. For $k = 9$ features selected, the stability value using the Dice index is about 0.9 which can be interpreted as a significantly large value as that measure takes values in the interval $[0, 1]$. On the other hand, Kuncheva's measure that is corrected by chance gives a stability close to 0, no matter what is the feature set size as it is corrected by 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 B.5. For a stability measure to be corrected by chance, $\mathbb{E}\left[\hat{\Phi}|H_0\right]$ should be constant and not depend on k or d.

[3]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.

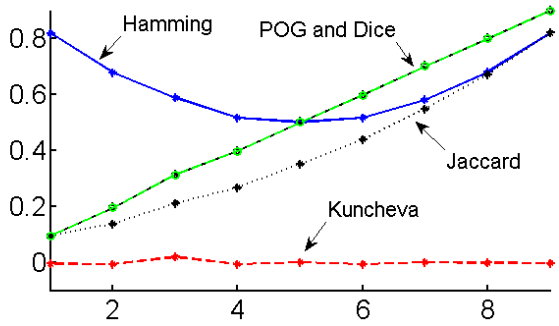


Figure 4.2: Illustration of the Correction for chance property. Stability values using Hamming, POG [Shi et al.], Jaccard, Dice and Kuncheva similarity measures against the number k of features selected for $M = 100$ repeats.

Finally, we summarized the properties of each one of the stability measures in Table 4.1. The first 5 measures are similarity-based measures and as we can see, they all possess all properties except Correction for chance. This flaw, as noticed by Kuncheva [2007] gave rise to corrected-by-chance similarity-based measures, which are the 4 following measures of Table 4.1. 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. Then the 6 last measures of 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 B.5 that it is asymptotically (as M approaches infinity) corrected by 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 propose a novel stability measure in the next chapter, that not only has all desired properties, but also will allow us to develop a statistical framework for the quantification of stability.

Table 4.1: 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 B.

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

Chapter 5

A Novel Stability Measure

In this chapter, we propose a measure of stability which provably retains all desirable properties as discussed in the previous chapter. 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 different feature selection procedures. For this reason, in the remainder of the thesis, we refer to the stability measure as the *stability estimator* and to the parameter being estimated as the *population stability* or the *true stability*.

5.1 Proposed Stability Estimator

As required by the property of Monotonicity, 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 interpretation, we ensure that (asymptotically) the value is in the range $[0,1]$ by taking one minus the resulting expression. Then, by making use of Theorem 3, this gives us our stability estimator as given in Definition 3.

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

Proof. *We start by showing that under H_0, given the cardinality k_i of the i^{th} feature set, every feature X_f is equally likely to be chosen with probability $\frac{k_i}{d}$. Then, using the*

law of total probabilities, we get that under H_0, $p_f = \frac{k_i}{d}$, which finally gives us that $\mathbb{E}\left[s_f^2|H_0\right] = p_f(1-p_f) = \frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)$, since s_f^2 is the unbiased sample variance of the Bernoulli variable Z_f. The full proof is given in Appendix A.3.

Definition 3 (Novel Measure) *We define the stability estimator as*

$$\hat{\Phi}(\mathcal{Z}) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)}, \tag{5.1}$$

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.

In the next section, we verify that the proposed measure possesses all 5 properties.

Properties of the Proposed Measure

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 when $\bar{k} = d$, which are the two limit cases in which the algorithm does not select any features over the M feature sets or selects all of the features in every feature set. 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 B.6 shows that the minimum of $\hat{\Phi}$ is equal to $-\frac{1}{M-1}$. This means that $\hat{\Phi}$ is bounded by -1, but is asymptotically bounded by 0 (as M approaches infinity).

Fourth, let us assume that for some $\mathcal{Z}$, the measure achieves its maximum, that is $\hat{\Phi}(\mathcal{Z}) = 1$. 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 the result of Theorem 3, we have that $\mathbb{E}\left[\hat{\Phi}|H_0\right] = 0$. Therefore, the proposed measure is corrected by chance.

We have noted earlier that there are some equivalences between existing measures in the literature—e.g. Wald's measure and *nPOG* are both generalizations of Kuncheva's measure to feature sets of variable cardinality. In Theorem 4, we show that the proposed measure of stability $\hat{\Phi}$ is also a generalization of Kuncheva's measure to feature sets of variable cardinality. This means that all the results of the following sections such as the asymptotic distribution, the confidence intervals, the hypothesis tests 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 provides a computational advantage, being $O(Md)$ instead of $O(M^2d)$ (which is the computational complexity of all pairwise measures).

Theorem 4 *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. *Using Theorem 2, we show that Kuncheva's stability measure can be re-written identically to the proposed stability estimator $\hat{\Phi}$. Then, using the results of Lemma 1 and of Theorem 1, we get the equivalences with all the other measures. The full proof of this theorem is given in Appendix A.4.*

5.2 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 will demonstrate and exploit a relationship between stability and inter-rater agreement [Fleiss et al., 1971].

5.2.1 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 Z represents a rater, placing the d subjects into category 0 or 1. Interestingly[1], in this special case, we prove with Theorem 5 that a popular measure of interrater agreement, Fleiss' Kappa [Fleiss et al., 1971] reduces to our estimator, Definition 3. This means that any statistical result previously derived for Fleiss' Kappa also holds for $\hat{\Phi}$.

Theorem 5 *When there are only two categories (0/1), Fleiss' Kappa [Fleiss et al., 1971] is equal to* $\hat{\Phi}(Z)$.

Proof. *To prove this, we start from the definition of Fleiss' Kappa as given in the original paper [Fleiss et al., 1971] and show that when the number of categories is equal to 2, it reduces to the proposed definition of stability* $\hat{\Phi}$. *The full proof of this theorem is given in Appendix A.5.*

Using this relationship, we can use the work of Gwet [2008] which derived tools such as the asymptotic distribution of Fleiss' Kappa, confidence intervals for the population parameter and hypothesis tests. Nevertheless, as earlier mentioned in Section 2.2.3, different sampling techniques than bootstrap have been used in the literature. The definition of our measure, as well as its properties do not depend on the sampling technique used, but we do not guarantee the validity of the statistical tools in all settings. In the next section, we comment on the bootstrap approach, showing that it is a valid approach to be able to use the work of Gwet [2008] providing statistical tools for Fleiss' Kappa.

5.2.2 The Bootstrap Approach

We could wonder which is the most relevant sampling technique to use and why. In this section, we provide some motivations in favour of the bootstrap approach.

In machine learning applications, the user is only given a finite subset of the whole data $\mathcal{D}$: this is our data set $\mathcal{D}_n$ made of n examples. By quantifying the stability of a feature selection procedure, we aim at quantifying the variability of the feature selection procedure due to the fact that we only have access to a sample $\mathcal{D}_n$ of the

[1] 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.

whole data $\mathcal{D}$. To formalize this, let g be a function that takes as in input the data set $\mathcal{D}_n$ and returns a feature set. If we apply the feature selection g to the available data set $\mathcal{D}_n$, we only get an estimate of $g(\mathcal{D})$. This estimate $g(\mathcal{D}_n)$ has an unknown *bias* and *variance*. When we quantify stability, in the sense of Definition 1, we aim at estimating the later. So, how can we estimate the variance of $g(\mathcal{D})$?

Since the underlying distribution generating $\mathcal{D}$ is unknown, a popular non-parametric technique to estimate the variance of $g(\mathcal{D}_n)$ is the *bootstrap estimation*, which consists in the following steps [Efron and Tibshirani, 1993]:

1. Take M bootstrap samples of the given data set $\mathcal{D}_n$.

2. Apply feature selection to each sample. This produces a matrix $\mathcal{Z}$ of size $M \times d$.

3. Compute the *bootstrap covariance matrix* equal to

$$\begin{pmatrix} s_1^2 & & & \\ & s_2^2 & & \\ & & \ddots & \\ & & & s_d^2 \end{pmatrix},$$

   where s_f^2 is the bootstrap variance estimate of the selection of the f^{th} feature and where the omitted elements are the bootstrap covariances.

Therefore, when using the bootstrap approach as a sampling technique, the sample variance $s_f^2 = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$ is in fact the bootstrap estimate of the variance of the selection of the f^{th} feature. In that light, when using bootstrap sampling, the proposed stability estimator $\hat{\Phi}$, which is a function of the bootstrap variance estimates s_f^2, is indeed measuring the variation *with respect to data sampling*, as given by Definition 1. Another motivation for the use of the bootstrap approach is given below.

Let us assume that the data set $\mathcal{D}_n$ is a random sample taken from the whole data set $\mathcal{D}$. The principle behind bootstrap sampling is to reproduce this sampling procedure: each bootstrap sample is in turn an independent random sample of $\mathcal{D}_n$, and therefore the rows of $\mathcal{Z}$ are a random sample from a larger population. Indeed, a random sample is by definition a sample where each example is drawn from the original sample *with replacement* [Efron and Tibshirani, 1993]. The work of Fleiss et al. [1971] that derives the asymptotic distribution of Fleiss' Kappa assumes that the raters (the rows of the matrix $\mathcal{Z}$ in our case), are a random sample from a larger population of raters. This

condition is therefore verified when using bootstrap sampling, which makes the use of all the statistical tools presented later in this section valid in that case.

We further note that some feature selection procedures are not deterministic. This means that when given an identical input, the output of the feature selection may vary due to the random initialization of parameters. Examples of such stochastic techniques are feature selection using the variable importance [Guyon and Elisseeff, 2006] of features with random forests or any other feature selection technique depending on initial initialisation of parameters. In that case, the feature selection procedure g is a function of both the input data and of the chosen parameters. Therefore $\hat{\Phi}$ will reflect on both sources of instability: the variance with respect to data sampling, but also, the variability due to the intrinsic stochastic nature of the feature selection procedure.

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.

5.2.3 The Sampling Distribution of Stability

Let us assume each row of the matrix 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 the d covariates. In the original paper, Fleiss et al. [1971] derive 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 $q = 2$. 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 6.

Theorem 6 (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 value of the* d *parameters of each one 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 5 showing the equality between our proposed measure and Fleiss' Kappa and the work of Gwet [2008] that derives the asymptotic distribution of Fleiss' Kappa.*

This asymptotic distribution allows us to identify $\hat{\Phi}$ as being an estimator of an unknown population quantity Φ. In the remainder of the thesis, 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.

5.2.4 Confidence Intervals

The asymptotic convergence to a normal distribution allows us to derive approximate confidence intervals for the population stability Φ as given by Corollary 1. Though the provided confidence intervals are only approximate, we will see in Section 6.1.2 that for relatively small values of M, the given intervals still have good coverage probability.

Corollary 1 (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* $1-\frac{\alpha}{2}$*.*

Proof. *This corollary directly follows from Theorem 6, c.f. Theorem on Normal-based confidence intervals from Wasserman [2010].*

5.2.5 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 hypotheses tested are

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

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 6). Therefore we can apply a one-tail test using the following two steps:

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 $\hat{\Phi}(Z_1)$ and of $\hat{\Phi}(Z_2)$ given by Theorem 6, we can derive Theorem 7.

Theorem 7 *The test statistic for comparing stabilities is*

$$T_M = \frac{\hat{\Phi}(Z_2) - \hat{\Phi}(Z_1)}{\sqrt{v(\hat{\Phi}(Z_1)) + v(\hat{\Phi}(Z_2))}}.$$

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

Proof. *We know from Theorem 6 that $\hat{\Phi}(Z_1)$ is asymptotically normal with unknown mean Φ_1 and variance σ_1^2 and that $\hat{\Phi}(Z_2)$ is asymptotically normal with unknown mean Φ_2 and variance σ_2^2. Therefore, the difference $\hat{\Phi}(Z_2) - \hat{\Phi}(Z_1)$ is normal with unknown mean $\Phi_2 - \Phi_1$ and with variance $\sigma_1^2 + \sigma_2^2$. Under the null hypothesis H_0, $\Phi_2 - \Phi_1 = 0$ and we estimate this variance by $v(\hat{\Phi}(Z_1)) + v(\hat{\Phi}(Z_2))$ using the result of Theorem 6. This gives us the asymptotic distribution of the statistic T_M.*

Using the 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.

In the next chapter, we first provide some experiments empirically showing the validity of the statistical tools proposed in this chapter. Then, we conduct a set of experiments showing their use in practice in different feature selection scenarios.

Chapter 6

Experiments

In this chapter, we first provide an empirical validation of the statistical tools proposed in Section 5.2. Then, we pursue with 3 sets of experiments, illustrating the use of the proposed tools in practice and showing the advantages of the optimization of stability along with the predictive power of the selected features.

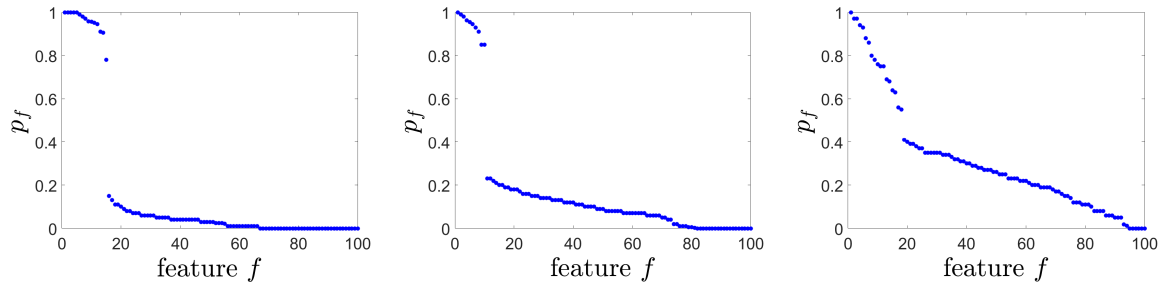
6.1 Empirical Validation of the Statistical Tools

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 6.1. 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.

Table 6.1: Benchmark scale for stability.

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

In the remainder of this section, we verify the tools of the previous chapter, using toy data for a population stability Φ in each one of the categories. To be able to generate 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. We picked the total number of features as $d = 100$, and we summarize the chosen population parameters $p_1, ..., p_d$ in Figure 6.1.



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

Figure 6.1: Three toy cases. We give the population parameters of the chosen Bernoulli variables $p_1, .., p_d$. The x-axis represents the feature index f taking values from 1 to $d = 100$. The y-axis represents the population parameter value p_f. The parameters have been chosen so that $\Phi = 0.8$ *[LEFT]*, $\Phi = 0.5$ *[MIDDLE]* and $\Phi = 0.3$ *[RIGHT]*. For visualization purposes, we sorted the parameters p_f in decreasing order.

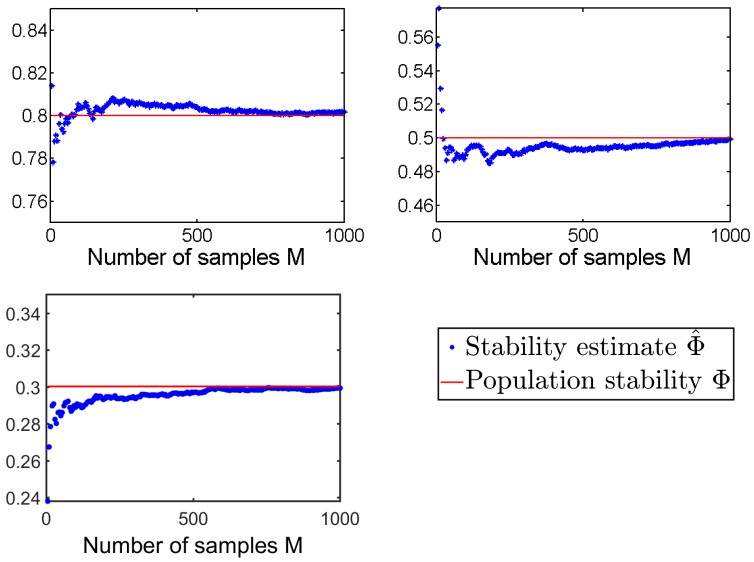


Figure 6.2: Consistency of the stability estimate $\hat{\Phi}(\mathcal{Z})$ for the 3 test cases presented in Figure 6.1. We can see that as M increases, the value of $\hat{\Phi}(\mathcal{Z})$ gets closer to the population parameter Φ.

6.1.1 Validation of Consistency of the Estimator

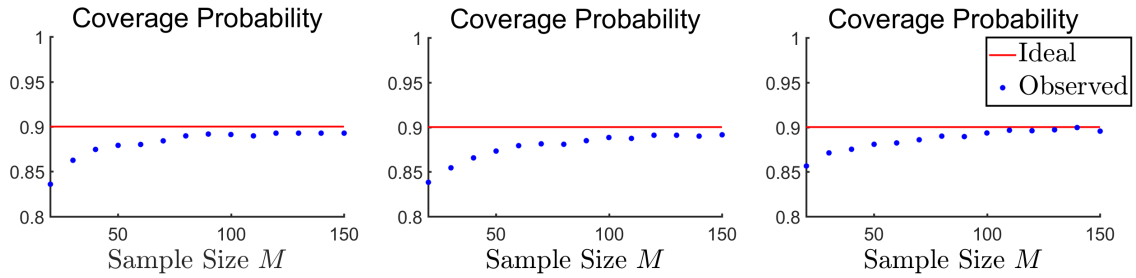
In this section, we empirically show the consistency of the stability estimator $\hat{\Phi}$ in the 3 test cases described. In each case, we take M samples from the $d = 100$ Bernoulli variables with mean parameters $(p_1, ..., p_d)$. This gives us a binary matrix $\mathcal{Z}$ of size $M \times d$. We then plot the value of the stability estimate $\hat{\Phi}(\mathcal{Z})$ as we increase the number of samples M in Figure 6.2. As we can see, as the number of samples M gets larger,

the value of $\hat{\Phi}(\mathcal{Z})$ gets closer to 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 lower than 5% and for $M \leq 100$, it is lower than 1%. Of course, these cannot be used as 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 the 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.

6.1.2 Validation of Confidence Intervals

The coverage probability of a confidence interval for Φ is the proportion of the time that the interval built from the data will actually contain the population stability Φ. To verify the results of Corollary 1 providing the confidence intervals, we adopted the following procedure and plotted the results in Figure 6.3:

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.
   - Compute the $(1-\alpha)$-approximate confidence interval.
3. The estimated coverage probability is the fraction of times (from 10,000) that the true stability Φ was within the confidence intervals.



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

Figure 6.3: 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 represents the estimated coverage probability with 10000 repeats.

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 6.3 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 6.2, we further gave 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 $(1-\alpha)$. The same properties can be seen again in Table 6.3, for $d = 10000$ features, and $M = 100$ samples.

Table 6.2: Coverage probabilities for the 3 test cases with $M = 100$, $d = 100$, estimated via $10,000$ repeats for different nominal confidence intervals.

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 6.3: Coverage probabilities for the 3 test cases with $M = 100$, $d = 10000$, estimated via $10,000$ repeats for different nominal confidence intervals.

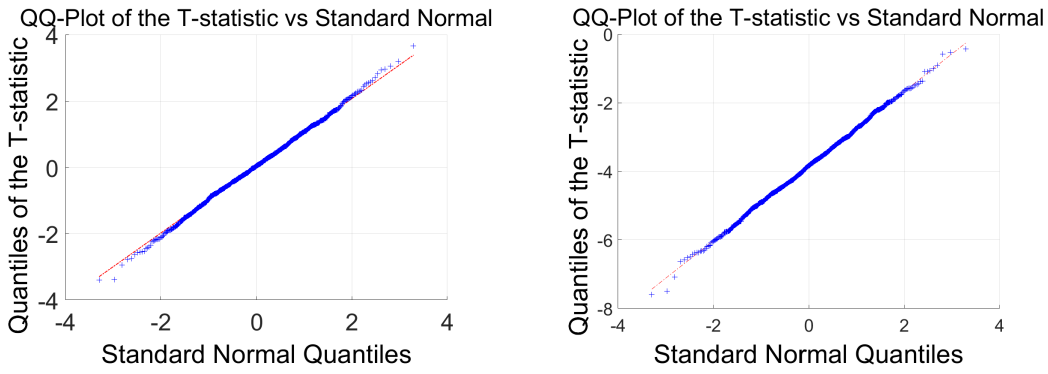
Case 1	$\Phi = 0.8$	98.8%	94.3%	88.8%
Case 2	$\Phi = 0.5$	98.6%	94.5%	88.8%
Case 3	$\Phi = 0.3$	99.0%	94.9%	90.2%
Ideal		99%	95%	90%

6.1.3 Validation of the Second Hypothesis Test

In this section, we verify the asymptotic distribution of the test statistic T_M as given by Theorem 7. To verify this result we proceed as follows:

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



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

Figure 6.4: 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.

6.2 Stability in Practice

The experiments of this section are illustrative of how the tools presented in this thesis 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[1]:

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

- Section 6.2.1 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 will show that the proposed stability measure captures this, and will show how 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 6.2.2 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.

We point out that all feature selection procedures used in this chapter are deterministic. Indeed, in the experiments of Section 6.2.1, we use the Matlab function `lassoglm` for the logistic LASSO and for the elastic net, which is an implementation of the cyclic coordinate descent (CCD) algorithm proposed by Friedman et al. [2010]. The experiments of Section 6.2.2 use the `glmnet` R package, which is also an implementation of the CCD algorithm. Finally, in the experiments of Section 6.2.3, we use mutual information criteria as presented in Section 2.1.3, which are all deterministically calculated from the data.

6.2.1 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 will also see that stability can help recover the *true* set of relevant features. To be able to control the true set of relevant features and the degree of redundancy between them, we use a synthetic data set as described in the next section.

Description of the data set

We use a synthetic data set [Kamkar et al., 2015]—a binary classification problem, with $n = 2000$ 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.

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 use $M = 100$ bootstrap samples of the data set and apply $L1$-logistic regression to each one of the samples. We then compute the average out-of-bag (OOB) log-likelihood[2] and the stability of the feature selection.

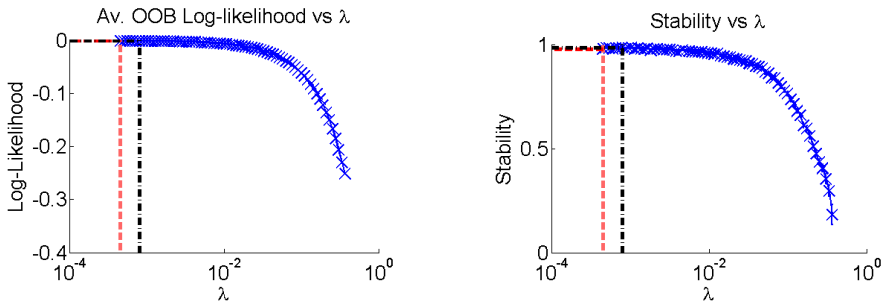
Figure 6.5 shows the average log-likelihood (left column) and the stability (right column) versus the regularization parameter λ for the 4 degrees of redundancy chosen. For each degree of redundancy, the pink dashed-line represents the parameter λ maximizing the likelihood and the black one represents 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 6.5b, 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 $\hat{\Phi} = 0.63$.

An alternative view of these results is shown in Figure 6.6, plotting stability against

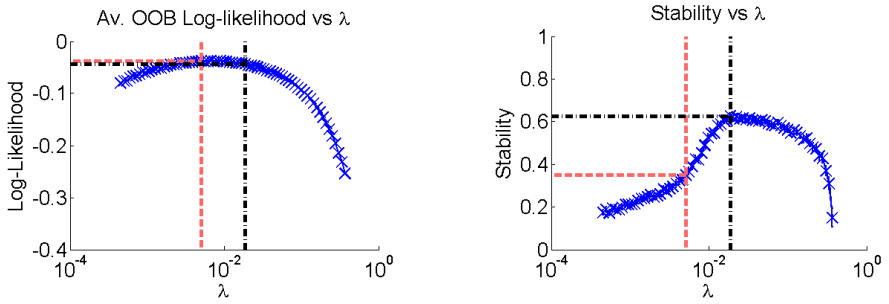
[2] In all the presented results, the log-likelihood is rescaled by the number of examples n.

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, we can see that this is not the case. 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 represents a different trade-off between likelihood and stability. We plotted the pareto fronts for each degree of redundancy in Figure 6.7. In a classic scenario, we would pick the value of λ that maximizes the likelihood only, which corresponds to the rightmost points in the figure. However, we can see from the 3 sub-figures, that sacrificing a small amount of likelihood allows us to considerably increase stability. Figure 6.8, we summarized the pareto fronts for the 4 degrees of redundancy. When there is no redundancy, the pareto front consists of three overlaid points (as they correspond to very similar values of stability and likelihood) which all correspond to an average OOB misclassification error equal to 0%. As we increase the degrees of redundancy in the data, 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.*

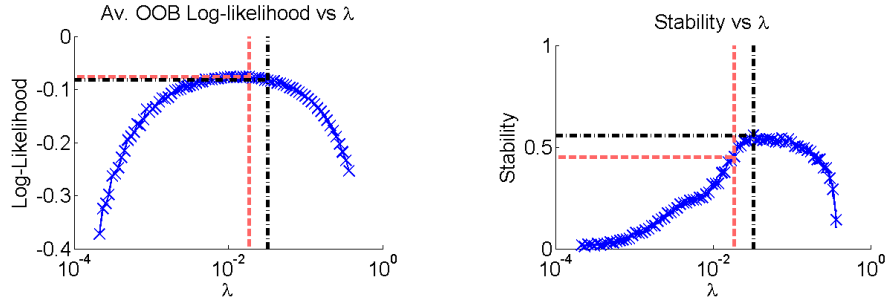
We also observe that pursuing stability may help us to identify the *true* set of features, when there is one. Figure 6.9 gives the observed frequencies of selection $\hat{p}_f$ of each feature over the $M = 100$ bootstraps—where features $1-50$ are relevant, and $51-100$ are noise/irrelevant features. This is in the case of high redundancy for two situations—first, when optimizing the likelihood only (left sub-figure), and second when optimizing both the likelihood and the stability (right sub-figure). We can see on the right figure that all 50 irrelevant features have a frequency of selection equal to 0 (i.e. the false positive rate 0), which means they have not been selected on any of the $M = 100$ repeats. This is not the case when only optimizing 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 the λ parameter that maximizes likelihood and the one that maximizes stability. Table 6.4 provides the false positives and false negatives for the 4 degrees of redundancy. The results show a decrease in the false positives when optimizing stability while having a limited effect on the false negatives.



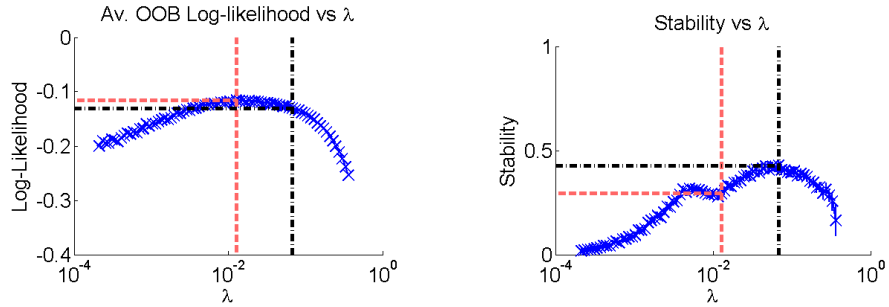
(a) No redundancy



(b) Low redundancy

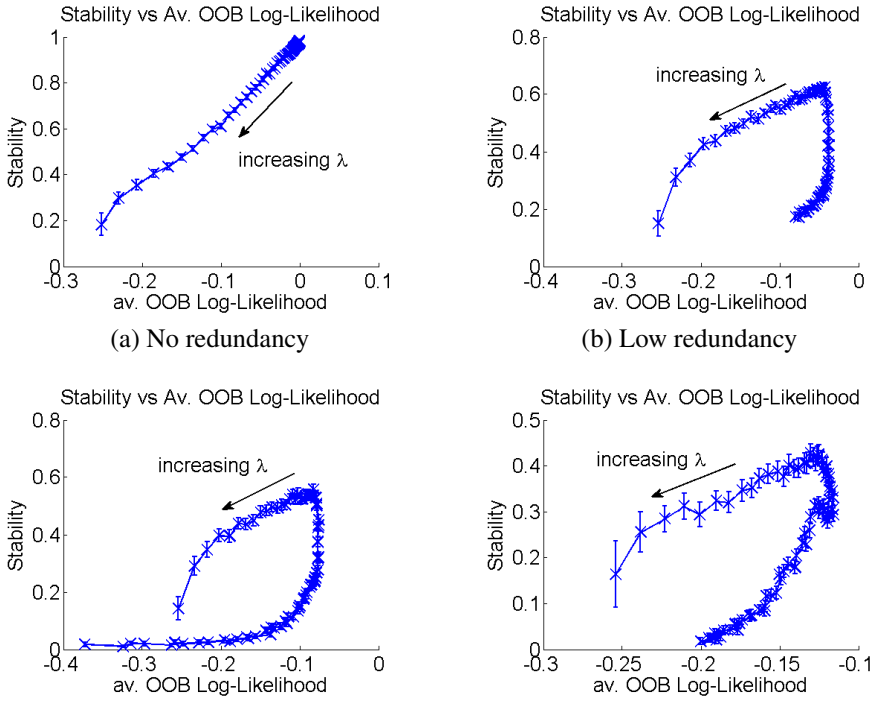


(c) Medium redundancy



(d) High redundancy

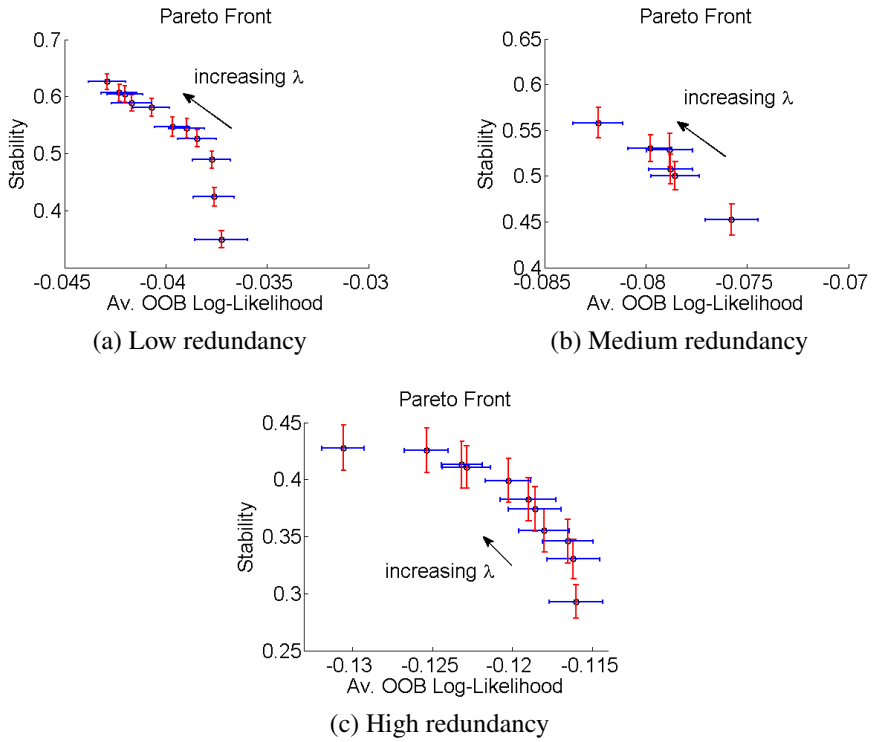
Figure 6.5: Average OOB log-likelihood [left column] and stability [right column] against the regularizing parameter λ for 4 degrees of redundancy. For each degree of redundancy, the pink dashed-line corresponds to the λ value that maximizes the likelihood and the black one corresponds to the value of 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 6.6: Stability (with 95%-confidence intervals) against average OOB log-likelihood for 4 degrees of redundancy.



(a) Low redundancy (b) Medium redundancy

(c) High redundancy

Figure 6.7: Pareto front of stability and log-likelihood with 95%-confidence intervals.

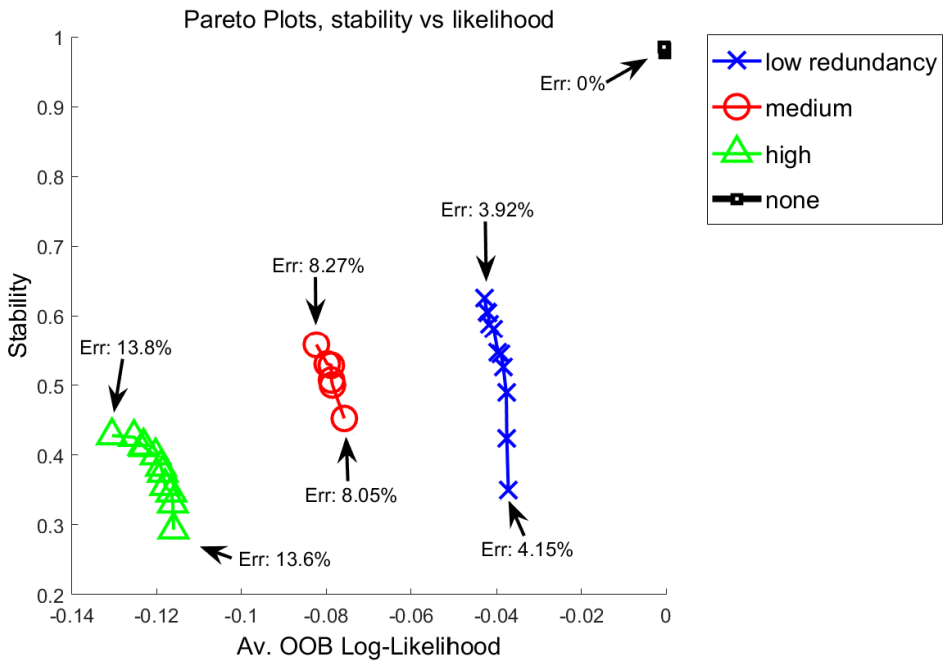
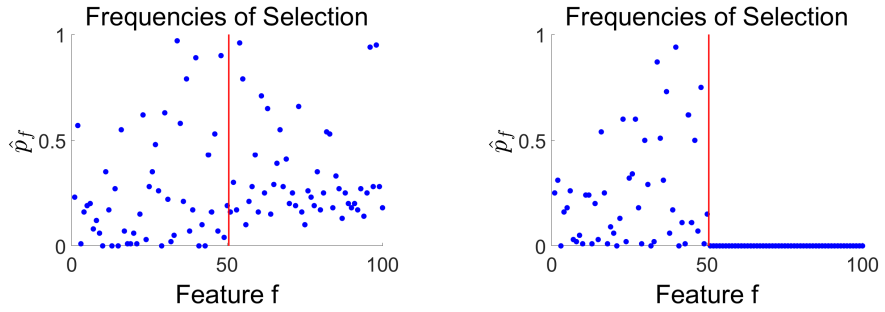


Figure 6.8: Summary of the pareto fronts for the 4 degrees of redundancy. The average OOB misclassification error is given for the two extreme points of each pareto front. For the 3 cases with redundancy, choosing any point on the pareto front will have little effect on the likelihood and the misclassification error, while it can considerably impact on the stability.



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

Figure 6.9: The observed frequencies of selection $\hat{p}_f$ of each feature in the case of high redundancy ($\rho = 0.8$). On the left sub-figure, we took the value of λ that maximized the likelihood. On the right sub-figure, we took a value of λ giving a trade-off between stability and likelihood. 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.

Table 6.4: False positives and false negatives of the final feature set for different degrees of redundancy ρ when optimizing only the likelihood against when optimizing stability.

Redundancy	Optimizing likelihood	Optimizing stability
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	$FP = 12, FN = 38$	$FP = 0, FN = 39$

Stability of the Elastic Net

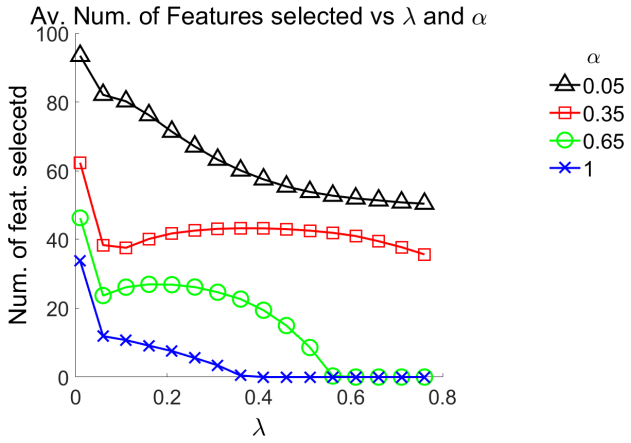
$L1$-regularization has the effect of forcing regression coefficients to zero, hence selecting a subset of the available features. $L2$-regularization (*ridge regression*), is known to have a grouping effect: correlated features will have similar coefficients [Zhou, 2013]. The Elastic Net is a convex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

the group of correlated features on every repeat. Therefore, we expect the Elastic net to mitigate this, hence increasing stability.

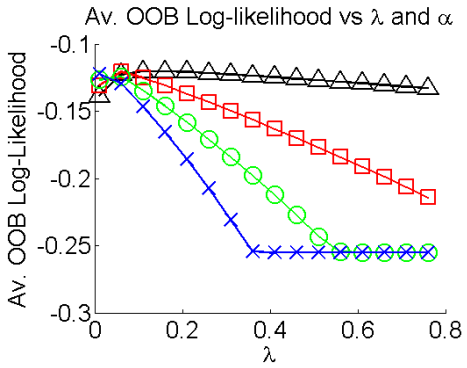
In this section, we reproduce some of the experiments of the last section for the Elastic Net, optimizing the two regularizing parameters α and λ. We proceed as before, taking $M = 100$ bootstraps, and focus on the most challenging case of high redundancy ($\rho = 0.8$). We confirm in Figure 6.10a that as λ increases, the overall regularization increases and less features are selected. Figures 6.10b and 6.10c 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 6.10a that the number of features selected is ~ 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 6.11 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 the value of λ maximizing the stability). We can see on the right sub-figure that when optimizing stability, the set of relevant features is always the set selected on each one of the $M = 100$ repeats 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.

Conclusions

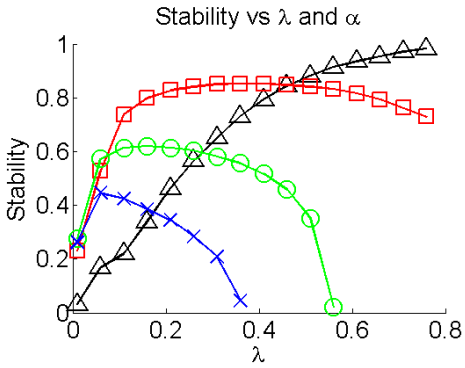
In this section, we proposed a methodology to select hyperparameters using stability along with the error. On the data set used, we showed that it is possible to select a hyperparameter yielding much higher stability values without loss of predictive power. When the stability is optimized along with 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.



(a) Average number of features selected $\bar{k}$



(b) Likelihood

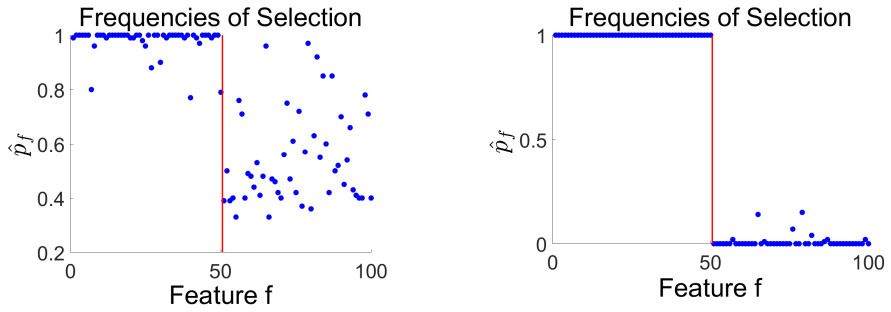


(c) Stability

Figure 6.10: 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.

6.2.2 How Stable is Stability Selection?

In high dimensional data sets, picking a regularizing parameter λ in a LASSO regression 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 take M random sub-samples of size $\lfloor \frac{n}{2} \rfloor$ of the original dataset (where n is the sample size) and to apply LASSO to each one of the sub-samples 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**



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

Figure 6.11: The observed frequencies of selection $\hat{p}_f$ of each feature in the case of high redundancy ($\rho = 0.8$). On the left sub-figure, we took the value of λ that maximized the likelihood. On the right sub-figure, we took a value of λ giving a trade-off between stability and likelihood. 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.

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.

We note that the proposed technique is effectively an ensemble feature selection technique where the final set is chosen to be the set of features having a high probability 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:

- It uses the concept of frequency of selection $\hat{p}_f$ to detect relevant features.
- 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 choices of regularizing parameters (i.e. it shows that the influence of the cut-off parameter π_{thr} and of the set of chosen regularizing parameters Λ is very small).
- It provides an upper bound on the number of false positives (i.e. the number of irrelevant features falsely selected) and shows that Stability Selection allows an exact control of the false positives.

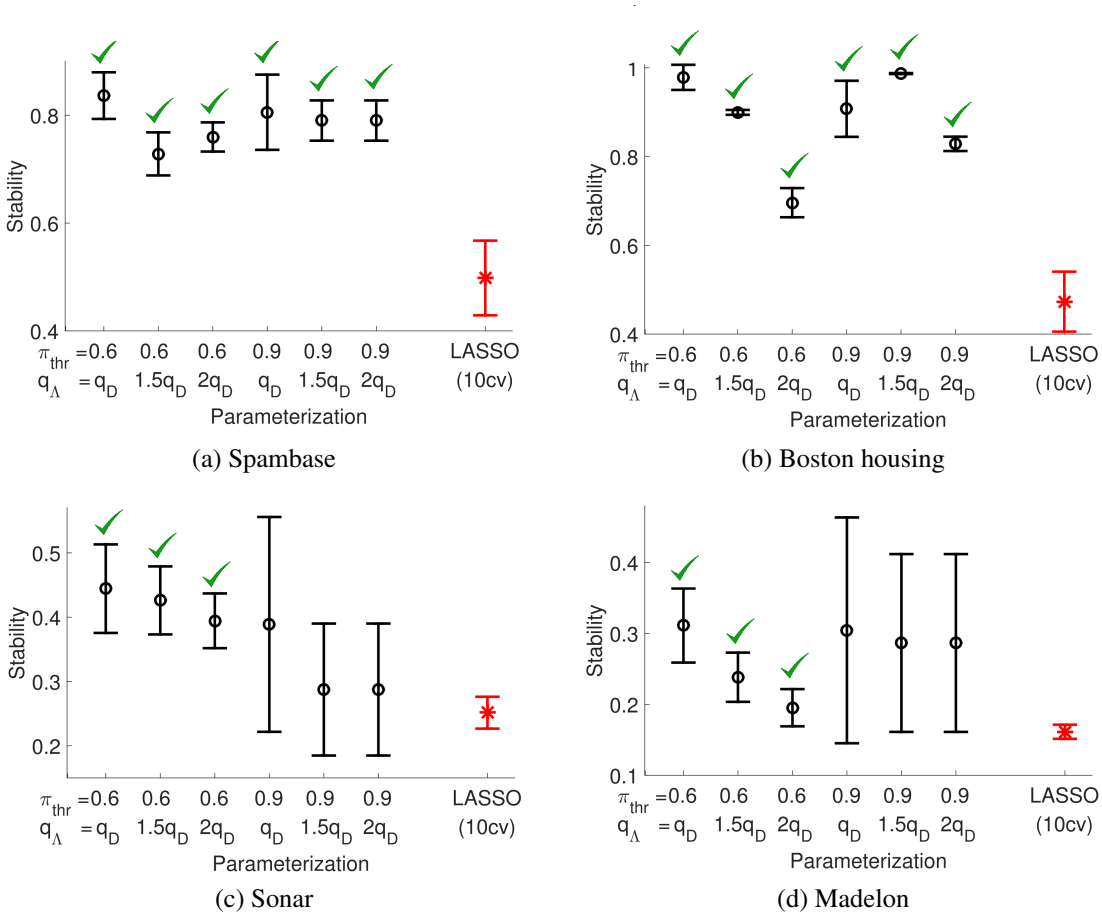
We note that (1) this work implicitly uses the idea that selecting stable features in the

final set will help recover the *true* set of relevant features and that (2) this 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 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 thesis and do not claim to be an exhaustive empirical study. Stability selection possesses 3 hyperparameters[3]: (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 6.12 compares the two approaches for variable selection in four diverse 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 results presented in Section 5.2.4. 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, we can see that 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, we failed to reject the null hypothesis. These experiments highlight the importance of statistical significance when quantifying stability and the need for such statistical tools.

[3]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 6.12: 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%.

6.2.3 Information Theoretic Feature Selection

Brown et al. [2012] studied the properties of a large number of information theoretic criteria and reported the accuracy and stability (using Kuncheva's measure). They concluded "*that the JMI criterion (Yang and Moody, 1999; Meyer et al., 2008) provides the best trade-off in terms of accuracy, stability, and flexibility with small data samples.*" [Brown et al., 2012][p.27]. In this section, we will reproduce some of their experiments and illustrate how these results can be reproduced with significance. Since our measure reduces to Kuncheva's measure when the number of features selected k is constant (c.f. Theorem 5), we can reproduce some of their experiments and compare the stability of the information theoretic criteria with significance by using the confidence intervals and the hypothesis tests of Section 5.2.

We compared 3 information theoretic feature selection procedures: the MIM, the JMI and the mRMR algorithms. For the purpose of these experiments, we used 8 data sets from the UCI repository where each feature has been discretized using 10 bins of equal width[4] and we used the FEAST Toolbox [Pocock and Brown, 2014].

Table 6.5: Description of the 8 UCI data sets used. All the data sets are binary problems. The final column indicates the proportion of positive examples (the majority class being always taken as the positive class). The closer this value is to 50%, the more the data set is balanced.

Data	Examples n	Features d	Ratio (%)
breast	569	30	63
congress	435	16	61
heart	270	13	56
ionosphere	351	34	64
krvskp	3196	36	52
parkinsons	195	22	75
sonar	208	60	53
spect	267	22	79

We used $M = 50$ bootstrap samples and applied the MIM, the JMI and the mRMR algorithm to each sample to select $k = 10$ features. We then computed the stability with a 95% interval and computed the average OOB accuracy using a using a nearest-neighbour classifier with 3 neighbours. We reported the results in Table 6.6. Then for each data set, we performed pairwise hypothesis tests to check whether the stabilities of each pair of algorithms are equal, as given by Figure 6.13 where a star means

[4]Taken from `http://www.cs.man.ac.uk/~gbrown/fstoolbox/`

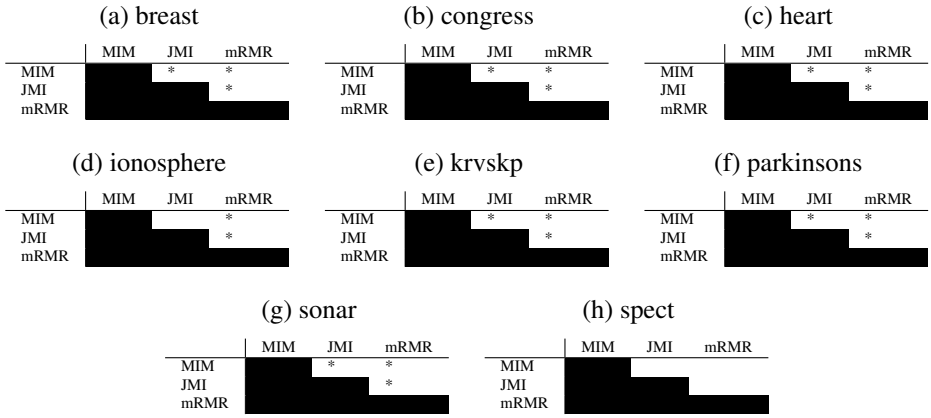
the null hypothesis has been rejected. Using these pairwise tests, we can give a more significant interpretation to Table 6.6. For example, on the spect data set, no significant difference has been found between the 3 algorithms in terms of stability while we can confidently say that there is a winner on the other 7 data sets. One interesting observation, is that the algorithms showing the highest stability were not always the ones with the smallest trade-off, highlighting the trade-off between stability and accuracy. For example, on the soar data set, the MIM was the algorithm with the highest stability while the JMI algorithm has a better accuracy. On the breast data set, the MIM algorithm reaches a perfect stability of 1 while the mRMR algorithm seems to have a better accuracy. As seen in Section 2.1.3, the JMI and mRMR algorithms use a greedy forward selection approach to build a feature set of non-redundant features. As these two algorithms depend on the search procedure and take into account feature redundancy, we expected them to be more sensitive to variations in the data, selecting different correlated features on every repeat. This phenomenon was not observed in our experiments. As we can see in Table 6.6, the MIM has a significant higher stability on 3 of the 8 data sets. Further experiments tuning the number of selected features for each algorithms on these data sets showed similar results with no clear winner in terms of stability.

In this chapter, we provided a set of experiments validating the statistical tools for stability and illustrating their applications in different feature selection contexts. In the next chapter, we provide the conclusions of this thesis and provide possible directions for future investigation.

Table 6.6: Average OOB error and stability using $M = 50$ bootstrap samples. We provide 95%-confidence intervals for the error and the stability. The values in bold are the highest ones for each data set. The data set having multiple lines in bold are the ones for which the hypothesises test showed no statistically significant difference between values using the pairwise tests of Figure 6.13.

Data		MIM	JMI	mRMR
breast	err	$7.5\% \pm 0.47\%$	$6.8\% \pm 0.54\%$	$6.0\% \pm 0.35\%$
	stab	**1**	$0.87 \pm 2.7\%$	$0.64 \pm 3.0\%$
congress	err	$8.0\% \pm 0.48\%$	$7.1\% \pm 0.47\%$	$7.4\% \pm 0.45\%$
	stab	$\mathbf{0.95 \pm 3.8}\%$	$0.89 \pm 4.3\%$	$0.71 \pm 4.6\%$
heart	err	$21\% \pm 0.91\%$	$22\% \pm 0.93\%$	$22\% \pm 0.97\%$
	stab	0.63 ± 0.05	$\mathbf{0.74 \pm 6.8}\%$	$0.46 \pm 9.9\%$
ionosphere	err	$14\% \pm 1.3\%$	$13\% \pm 0.79\%$	$13\% \pm 0.95\%$
	stab	$0.56 \pm 2.7\%$	$0.57 \pm 4.4\%$	$\mathbf{0.64 \pm 3.2}\%$
krvskp	err	$7.1\% \pm 0.55\%$	$6.7\% \pm 0.48\%$	$7.1\% \pm 0.51\%$
	stab	$0.80 \pm 2.1\%$	$0.84 \pm 2.0\%$	$\mathbf{0.87 \pm 1.3}\%$
parkinsons	err	$16\% \pm 1.1\%$	$12\% \pm 1.1\%$	$12\% \pm 0.10\%$
	stab	$0.72 \pm 2.3\%$	$\mathbf{0.84 \pm 3.9}\%$	$0.47 \pm 3.8\%$
sonar	err	$25\% \pm 1.5\%$	$22\% \pm 1.3\%$	$26\% \pm 1.5\%$
	stab	$\mathbf{0.52 \pm 3.0}\%$	$0.43 \pm 2.7\%$	$0.31 \pm 3.8\%$
spect	err	$24\% \pm 1.2\%$	$23\% \pm 1.1\%$	$22\% \pm 1.3\%$
	stab	$\mathbf{0.34 \pm 3.7}\%$	$\mathbf{0.37 \pm 4.0}\%$	$\mathbf{0.37 \pm 5.4}\%$

Figure 6.13: Pairwise hypothesis tests on stability equality. A star * means the null hypothesis (equality in stability of the two algorithms) is rejected at $\alpha = 0.05$.



Chapter 7

Conclusions and Future Work

In this thesis, we first introduced the topic of stability in feature selection and provided a literature review of feature selection techniques, of research questions around stability and briefly discussed the existing solutions. In this chapter, we summarize the contributions of the thesis by answering the main research questions, as stated in Chapter 1. Then, we provide possible directions for future work.

7.1 What Did We Learn in This Thesis?

We thereafter provide a summary of the contributions of this thesis.

7.1.1 Which Properties a Stability Measure Should Possess?

We answered this question in Chapter 4. There is a wide choice of stability measures in the literature and we could wonder which stability measures we should use and why. In the literature, many measures were proposed to comply with a certain number of properties (e.g. Kuncheva [2007], Lustgarten et al. [2009], Somol and Novovičová [2010], Guzmán-Martínez and Alaiz-Rodríguez [2011]). Nevertheless, stability measures are still being proposed, as none of them seem to comply with all desired properties. Moreover, the properties defined in the literature are sometimes only defined for a specific category of measures, which makes measures of different types difficult to compare in terms of properties. The diversity of the existing measures makes empirical studies on stability difficult to cross-compare and results might be biased by the choice of a specific measure. After a large analysis of the literature, we proposed a set of 5 properties applicable to any stability measure in the general case and encompassing the

needs expressed in the literature. Let $\mathcal{Z}$ be a collection of M feature sets and $\hat{\Phi}$ be the stability function taking $\mathcal{Z}$ as input. We recapitulate the proposed set of properties for $\hat{\Phi}$ hereafter:

1. *Fully defined.* The stability estimator $\hat{\Phi}$ should be defined for any collection $\mathcal{Z}$ of feature sets.

2. *Monotonicity.* The stability estimator $\hat{\Phi}$ should be a strictly decreasing function of the sample variance s_f^2 of the selection of each feature.

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.* The stability $\hat{\Phi}(\mathcal{Z})$ 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.

We further motivated the importance of these properties with various counter-examples and showed why these properties were critical in many scenarios. The set of properties ensures that stability values are interpretable and comparable in different scenarios. For instance, the Correction for chance property ensures that the value of $\hat{\Phi}(\mathcal{Z})$ will not be biased by the number of features selected and that the measure will not take into account the similarity between feature sets due to *chance* only. This new set of properties allowed us to analyse and compare all existing measures in terms of properties.

7.1.2 Is There a Measure of Stability Possessing All Properties?

We reviewed all existing measures in Chapter 3. For each one of the 15 existing measures, we checked which one of the 5 properties they possess and summarized our findings in Table 4.1. The proofs for these results were given in Appendix B. After this analysis, we found out that none of the existing measures possessed all 5 properties. Built on the set of desired properties, we proposed a novel stability measure defined as

$$\hat{\Phi}(\mathcal{Z}) = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d} s_f^2}{V_{set}},$$

where s_f^2 is the sample variance of the selection of the f^{th} feature and where V_{set} is the value of the numerator under the Null Model of Feature Selection (given by Definition 2). This measure can be interpreted as the proportion of agreement between the M feature sets in $\mathcal{Z}$ above the one due to chance alone. We showed that (1) not only this measure possesses all the desired properties but also that, (2) it is a generalization of some of the other existing measures (as given by Theorem 4) that were only defined in specific scenarios. For instance, it is a generalization of the popular Kuncheva's measure to scenarios where the number of features selected may vary when given different data samples. We can also point out that the computational complexity of this measure is lower than most of the existing ones and using this expression for Kuncheva's measure instead of its original pairwise definition is more efficient. This novel measure is hence consistent with previous proposals and possesses the desired properties stated in various bodies of literature.

7.1.3 Can We See Stability as a Random Variable?

The answer to that question is *yes*. In Section 5.1, we showed that the proposed measure is a special case of Fleiss's Kappa [Fleiss et al., 1971] (c.f. Theorem 5). Using this equivalence, we showed that:

- The stability $\hat{\Phi}$ is a consistent estimator of an underlying random variable Φ. As we increase the number of samples M, the estimator asymptotically converges to the population parameter Φ.
- As the number of samples M approaches infinity, the stability $\hat{\Phi}$ weakly converges to a Normal distribution with unknown mean Φ.

To the best of our knowledge, this is the first statistical treatment of a stability measure for feature sets. This allowed us to derive a set of statistical tools, as given in the next section.

7.1.4 Which Statistical Tools for Stability?

We answered this question in Section 5.2. The asymptotic distribution of the stability $\hat{\Phi}$ allowed us to derive a set of statistical tools that can be shown to be useful in many experimental scenarios that we describe in the sections below.

How Confident Are We About the Value of Φ?

As aforementioned, the stability value computed $\hat{\Phi}(\mathcal{Z})$ is only an estimate of the *true* stability Φ. So how confident can we be about the value of the true parameter Φ? We answer this question by providing approximate confidence intervals in Section 5.2.4. For example, if we pick a level of significance equal to 5%, computing the confidence interval will tell us that there is a 95% probability that the true stability parameter Φ belongs to that interval. Since this result is only asymptotic, we further validated this result with a set of test cases in Section 6.1.2. In our experiments, the estimated coverage probability was converging to the expected one for reasonably small sample sizes ($M \geq 30$).

Is the Population Stability Greater than a Given Value?

In many empirical scenarios, it can be useful to test whether the true stability Φ is greater than a given value. Based on previous literature about Fleiss' Kappa, we provided a benchmark scale for interpretation of stability given in Table 6.1. One might therefore want to test if $\Phi > 0.75$ which corresponds to an excellent level of stability. To answer this, we proposed a null hypothesis test in Section 5.2.5.

Have Two Given Algorithms Identical Stabilities?

This is another critical question. In the literature, many works aim at empirically comparing the stabilities of several feature selection algorithms [Kalousis et al., 2007, Kuncheva et al., 2012]. Comparing only estimates in this scenario can be hazardous: due to the variance of the stability estimates, we might observe different stability estimates while the population stabilities are identical. Furthermore, even in scenarios where the stability estimates are quite accurate, it is unlikely that two algorithms having identical stabilities will have identical estimates. For this reason, we proposed a null hypothesis test allowing to decide whether two feature selection algorithms have identical population stabilities in Section 5.2.5. We further validated this result in Section 5.2.5 on a set of test cases and illustrated its use in practice in Section 6.2.2.

7.1.5 Can We Increase Stability Without Loss of Predictive Power?

The purpose of Chapter 6.2 was twofold:

1. It aimed at illustrating how the tools derived in this thesis could be used in practice in empirical scenarios.

2. It showed that in some scenarios, it is possible to pick-up hyperparameters that yield higher stability and at least as good predictive power than when only optimizing the predictive power.

Moreover, to be able to control the set of relevant features, we carried some experiments on artificial data sets in Section 6.2.1. The results showed that when using regularized methods, achieving higher stability could also help recover the true set of features without loss in terms of accuracy.

The material covered in this thesis is applicable to a large variety of problems in feature selection. Nevertheless, as it will be covered in the next section, extensions of this work to other type of outputs or to other case scenarios could be explored in future work.

7.2 Future Work

In this section, we propose two areas of research that could be explored in future work.

7.2.1 Feature redundancy

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] and they both reduce to their original version when the redundancy between the features is null. This case study is not in the scope of this thesis. 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.

7.2.2 Unifying Framework

An important future question is how we could extend the present work to other types of feature selection outputs—such as feature rankings or feature weights. A popular measure in that category is the average pairwise Spearman's Rho (a similarity-based stability measure), used to quantify the stability of feature rankings. In the presence of untied ranks, Nogueira et al. [2017] shows 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 the f^{th} feature and where V_{rank} is the expected variance under the assumption that each ranking was generated at random (i.e. each ranking has equal probability). The above equation has a similar form as the measure in this thesis and provides a promising direction for unifying stability of ranking and subset selection.

Appendix A

Proof of Theorems

In this appendix, we provide the full proofs or proof sketches for the theorems that are not already given in the main body of the thesis. For disambiguation, 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. Whenever the subscript is omitted, we refer to the proposed stability measure as given by Definition 3. Before we proceed to the proofs, in order to make this appendix self-contained, we remind the notations below:

- d is the total number of features;
- $X_1, ..., X_d$ represent the d features;
- M is the number of bootstrap samples (also the number of feature sets);
- Z_f is the Bernoulli variable modelling the selection of the f^{th} feature;
- $\mathcal{Z}$ is the binary feature selection matrix, of size $M \times d$;
- $z_{i,f}$ is the binary value at the i^{th} row and f^{th} column of $\mathcal{Z}$ indicating whether the f^{th} feature has been selected on the i^{th} sample;
- k is the number of selected features when the procedure is returning a constant number of features;
- $k_i = \sum_{f=1}^{d} z_{i,f}$ is the number of selected features on the i^{th} bootstrap samples (i.e. the number of 1s on the i^{th} row of $\mathcal{Z}$);
- $\bar{k} = \frac{1}{M} \sum_{i=1}^{M} k_i$ is the average number of selected features over the M feature sets in $\mathcal{Z}$;

- $\hat{p}_f = \frac{1}{M}\sum_{i=1}^{M} z_{i,f}$ is the observed frequency of selection of the f^{th} feature in $\mathcal{Z}$;
- p_f is the population mean of variable Z_f;
- $s_f^2 = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$ is the sample variance of the selection of the f^{th} feature;
- $r_{i,j}$ is the size of the intersection between the i^{th} and the j^{th} feature set in $\mathcal{Z}$;
- ϕ is a similarity measure, taking two feature sets as input and returning a value in $\mathbb{R}$;
- $\hat{\Phi}(\mathcal{Z})$ is the proposed stability estimator;
- Φ is the population stability being estimated;
- H_0 is the Null Model of Feature Selection, as given by Definition 2.

We also give the following equation, that will be repeatedly used in the proofs

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

A.1 Proof of Theorem 1

In this appendix, we prove the following theorem from Section 3.2.

Theorem 1 *When the number of features selected is constant and equal to k, the relative weighted consistency is asymptotically equivalent to Kuncheva's stability measure.*

$$\hat{\Phi}_{CW_{rel}} \underset{M\to+\infty}{\sim} \hat{\Phi}_{Kuncheva}.$$

Proof. *Using Equation 3.4, it can be shown that CW_{rel} can be re-written as follows*

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

Assuming that the number of features selected is constant equal to k, we then have that $\bar{k} = k$ and hence that $H = (Mk) \bmod M = 0$. 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 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, if we take the limit of the above equation, we get:

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

Using the result of Theorem 2, we have that $\sum_{f=1}^{d}s_f^2 = k - \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}r_{i,j}$. By using this in the previous equation, we get

$$\lim_{M\to\infty}\hat{\Phi}_{CW_{rel}}(\mathcal{Z}) = \lim_{M\to\infty}\left[1-\frac{\frac{M-1}{M}\left(k-\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}r_{i,j}\right)}{k\left(1-\frac{k}{d}\right)}\right].$$

$$\lim_{M\to\infty}\hat{\Phi}_{CW_{rel}}(\mathcal{Z}) = \lim_{M\to\infty}\left[\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\frac{r_{i,j}-\frac{k^2}{d}+\frac{k}{M}}{k\left(1-\frac{k}{d}\right)}\right].$$

Reminding the reader that the stability measure using Kuncheva's similarity is equal to

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

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

A.2 Proof of Theorem 2

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

Theorem 2 *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*

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

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

$$\begin{aligned}
\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M} r_{i,j} &= \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{f=1}^{d} z_{i,f} z_{j,f} - \frac{1}{M-1}\bar{k} \\
&= \frac{1}{M(M-1)}\sum_{f=1}^{d}\sum_{i=1}^{M}\left(z_{i,f}\left(\sum_{j=1}^{M} z_{j,f}\right)\right) - \frac{1}{M-1}\bar{k} \\
&= \frac{1}{M(M-1)}\sum_{f=1}^{d}\left(\sum_{i=1}^{M} z_{i,f}\right)\left(\sum_{j=1}^{M} z_{j,f}\right) - \frac{1}{M-1}\bar{k} \\
&= \frac{1}{M(M-1)}\sum_{f=1}^{d}\left(\sum_{i=1}^{M} z_{i,f}\right)^2 - \frac{1}{M-1}\bar{k} \\
&= \frac{1}{M(M-1)}\sum_{f=1}^{d}\left(M\hat{p}_f\right)^2 - \frac{1}{M-1}\bar{k} \\
&= \frac{M}{M-1}\sum_{f=1}^{d}\hat{p}_f^2 - \frac{1}{M-1}\bar{k} \\
&= \frac{M}{M-1}\sum_{f=1}^{d}\left(\hat{p}_f^2 - \hat{p}_f + \hat{p}_f\right) - \frac{1}{M-1}\bar{k} \\
&= \frac{M}{M-1}\sum_{f=1}^{d} -\hat{p}_f(1-\hat{p}_f) + \frac{M}{M-1}\sum_{f=1}^{d}\hat{p}_f - \frac{1}{M-1}\bar{k} \\
&= -\frac{M}{M-1}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f) + \frac{M}{M-1}\bar{k} - \frac{1}{M-1}\bar{k} \text{ using Eq. A.1}
\end{aligned}$$

$$= \bar{k} - \sum_{f=1}^{d} s_f^2.$$

■

A.3 Proof of Theorem 3

In this appendix, we prove the following theorem from Section 5.1.

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

Proof. *Let $k_1,...,k_M$ be the cardinalities of the M feature sets in $\mathcal{Z}$. Under the Null Model of Feature Selection H_0, for each row i in $\mathcal{Z}$, every feature set of cardinality k_i is equally likely to be chosen. Therefore, each feature X_f is also equally likely to be chosen. In that situation, one way of calculating the probability that feature X_f belongs to the set of selected features is to compute the number of events in which feature X_f is selected divided by the size of the sample space (which is the total number of possible feature sets of size k_i). Therefore, given the cardinality k_i of the i^{th} feature set, the probability of selection p_f of feature X_f is equal to*

$$\frac{\#\{\text{sets of size } k_i \text{ containing } X_f\}}{\#\{\text{sets of size } k_i\}}.$$

The denominator is equal to $\binom{d}{k_i}$. For the numerator, we know X_f is already included in the set, which means we have now $k_i - 1$ features to choose from the remaining $d-1$ features. Therefore the number of features sets containing X_f is $\binom{d-1}{k_i-1}$. Replacing these in the previous equation, given the cardinality k_i of the i^{th} feature set, we get that all features have an equal probability of being selected equal to $\frac{\binom{d-1}{k_i-1}}{\binom{d}{k_i}} = \frac{k_i}{d}$. Finally, using the law of total probability, we have that $p_f = \frac{\bar{k}}{d}$. Therefore, since s_f^2 is the unbiased sample variance of the Bernoulli variable Z_f, we have $\mathbb{E}\left[s_f^2|H_0\right] = Var[Z_f|H_0] = p_f(1-p_f) = \frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)$. ■

A.4 Proof of Theorem 4

In this appendix, we prove the following theorem from Section 5.1.

Theorem 4 *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. *Here, we show that when the feature sets are of constant cardinality equal to k, then Kuncheva's measure is equal to the proposed stability measure. The stability measure defined by Kuncheva [2007] is*

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

Using Theorem 2, 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

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

which is our proposed stability measure when the number of selected features is constant equal to k. Then using the results of Lemma 1 and of Theorem 1, we get the equivalences with all the other measures. ■

A.5 Proof of Theorem 5

In this appendix, we prove the following theorem from Section 5.2.1.

Theorem 5 *When there are only two categories (0/1), Fleiss' Kappa [Fleiss et al.,*

1971] 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 et al., 1971] and show that when the number of categories is equal to 2, it reduces to the proposed definition of stability (c.f. Definition 3). Fleiss et al. [1971] defines Kappa as*

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{A.3}$$

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 $\bar{P}$.

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

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

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

$$\bar{P} = \frac{1}{dM(M-1)} \sum_{f=1}^{d} \left(M^2 - M - 2M^2\hat{p}_f + 2M^2\hat{p}_f^2\right)$$

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

$$\bar{P} = 1 - \frac{2}{d}\sum_{f=1}^{d} s_f^2.$$

Now, substituting the two last equations back into Equation A.3, we finally get that

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

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

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

$$\kappa = \hat{\Phi}(\mathcal{Z}),$$

which is what we wanted to prove. ■

Appendix B

Proof of Properties

In this appendix, for each one of the 5 properties given in Section 4, we determine which measures possess the property.

B.1 First property: Fully defined

This property directly follows from the definitions of the stability measures given either in Section 3.1. We note that 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.

B.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.

B.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 B.1. 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 B.1: Derivatives for each one of the similarity measures.

B.2.2 Frequency-based Measures

Goh's Measure

Using the definition of Goh's measure (c.f. Table 3.2) and Equation A.1, we have that

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

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

Davis's Measure

This measure adds a penalizing term to the previous measure and depends on a user-defined hyperparameter parameter α. When $\alpha = 0$, we are in the same case as in the previous section. Therefore this measure does not possess the Monotonicity property.

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 Z_1 and Z_2 illustrate two different scenarios with $M = 4$ as follows

$$Z_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad \textit{and} \quad Z_2 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}.$$

Using Krízek's measure, we get a stability of 1 for the two collections of feature sets Z_1 and 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 Z_1 and $\sum_{f=1}^{4} s_f^2 = \frac{2}{3}$ for Z_2. Therefore we can see that Z_1 and Z_2 have the same stability value using Krízek's measure but different sums of variance. Therefore Krízek's measure does not have the Monotonicity property.

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

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

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, we have

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

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

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

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

$$\hat{\Phi}_{Guzman}(\mathcal{Z}) = 1 + \frac{\frac{1}{k}\sum_{f=1}^{d}\hat{p}_f\log\hat{p}_f}{\log\frac{d}{k}} = 1 - \frac{\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f\log\hat{p}_f}{\frac{k}{d}\log\frac{k}{d}}. \tag{B.2}$$

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

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

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 $\mathcal{Z}_2$.

Relative Weighted Consistency CW_{rel} [Somol and Novovičová, 2010]

From Equation A.2, we have

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

Since H, D and $\bar{k}$ only depend on the feature set cardinalities $k_1,\ldots,k_M$, on M and on d, we can see 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.

Lausser's measure

Similarly to what has been done for Guzmán's measure in Section B.2.2, 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 will always be equal to* $(M\hat{p}_f)^2$*. Hence, Lausser's Measure can be re-written as*

$$\hat{\Phi}_{Lausser}(\mathcal{Z}) = \frac{1}{M^2 k} \sum_{f=1}^{d} (M\hat{p}_f)^2 = \frac{1}{k} \sum_{f=1}^{d} \hat{p}_f^2. \tag{B.3}$$

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*

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

Therefore, Lausser's measure has the Monotonicity property.

B.3 Third property: Bounds

In this section, we verify which measures have the Bounds property as given by Section 4.

B.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 3.1, 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.

Wald and *nPOG* measures

In this section, we provide a counter-example to show that Wald's measure is not bounded. Let us assume we have the following scenario

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

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.

B.3.2 Frequency-based measures

The minimum and maximal values of the frequency-based measures are given in the literature and recapitulated in Table 3.2. 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's and Goh's measures) take values in $[0,1]$ *and therefore are bounded.*

B.4 Fourth Property: Maximum

In this section, we show which one of the stability measures possess the Maximum property, as given in Section 4.

B.4.1 Similarity-based Measures

Deterministic Selection → 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 stability measure have this property except Lustgarten's measure (as it is shown with a counter-example in Figure 4.1b).*

Maximum Stability → 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 thesis (c.f. Figure 4.1a). 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* $\mathcal{A}$ *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 $s_i = s_j$.

B.4.2 Frequency-based Measures

Goh's Measure

Using Equation B.1, we have that $\hat{\Phi}_{Goh}(\mathcal{Z}) = \frac{\bar{k}}{d}$. Therefore, when all feature sets in $\mathcal{Z}$ are identical, $\hat{\Phi}(\mathcal{Z})$ 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.

Davis's 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.

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.

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

Therefore, Krízek's measure has the Maximum property.

Relative Weighted Consistency CW_{rel}

Using Equation A.2, we have

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

$$\Leftrightarrow \quad \sum_{f=1}^{d} s_f^2 = \frac{H}{M-1} - \frac{H^2}{M(M-1)}.$$

When all feature sets in $\mathcal{Z}$ *are identical, we have* $\bar{k} = k$ *and therefore* $H = (M\bar{k}) \bmod M = 0$*. 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 $\hat{\Phi}(\mathcal{Z}) = 1$*, even though all rows in* $\mathcal{Z}$ *are not identical. Therefore* CW_{rel} *does not have the Maximum property.*

Lausser's Measure [Lausser et al., 2013]

To prove that Lausser's measure has this property, we will use the expression given in Equation B.3, 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*. Therefore in that case, the stability value is* 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.

B.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.

B.5.1 Similarity-based Measures

We know that $\mathbb{E}\left[r_{i,j}|H_0\right] = \frac{k_i k_j}{d}$ (c.f. Section 4.1). 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's 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 show the calculations below.

Normalized Hamming distance

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

$$= 1 - 2\frac{\bar{k}}{d} + \frac{M}{M-1}\frac{\bar{k}^2}{d^2} - \frac{1}{M(M-1)}\sum_{i=1}^{M}\frac{k_i^2}{d^2}.$$

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 does not have the Correction for chance property.*

Jaccard index

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

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*

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

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.*

Dice coefficient

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

$$= \frac{2}{M(M-1)d} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \frac{k_i k_j}{k_i + k_j}.$$

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.

Ochiai's index

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

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.

POG

Here, we use the symmetrical version of POG 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.

B.5.2 Frequency-based Measures

We first start by proving Theorem 8 that will useful to calculate the value of $\mathbb{E}\left[\hat{\Phi}(\mathcal{Z})|H_0\right]$ *for the different stability measures.*

Theorem 8 (Expected Value of $\hat{p}_f$ ***under*** H_0***)*** *Under the Null Model of Feature Selection H_0, for all $f \in \{1,...,d\}$, the f^{th} column of $\mathcal{Z}$ can be modelled by a Bernoulli random variable with true parameter $p_f = \frac{\bar{k}}{d}$ and the expected value of the sample mean is $\mathbb{E}\left[\hat{p}_f|H_0\right] = p_f = \frac{\bar{k}}{d}$.*

Given the cardinality k_i, under H_0, all features will have an equal probability of being selected equal to $\frac{\binom{d-1}{k_i-1}}{\binom{d}{k_i}} = \frac{k_i}{d}$. Therefore, given the cardinalities of the M feature sets $k_1,...,k_M$, the probability of selection of the f^{th} feature is equal to $p_f = \frac{1}{M}\sum_{i=1}^{M}\frac{k_i}{d} = \frac{\bar{k}}{d}$. Under H_0, the output of the feature selection does not depend on the data and each row of $\mathcal{Z}$ can then be assumed independently drawn for a same distribution. Therefore, $\mathbb{E}\left[\hat{p}_f|H_0\right] = \frac{1}{M}\sum_{i=1}^{M}\mathbb{E}\left[z_{i,f}|H_0\right] = p_f = \frac{\bar{k}}{d}$.

Goh's Measure

Since we have $\hat{\Phi}_{Goh}(\mathcal{Z}) = \frac{\bar{k}}{d}$ (c.f. Equation B.1), this gives $\mathbb{E}\left[\hat{\Phi}_{Goh}(\mathcal{Z})|H_0\right] = \hat{\Phi}(\mathcal{Z}) = \frac{\bar{k}}{d}$, which is not constant. Therefore this measure does not have the property.

Davis's Measure

For this 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.

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 we get that

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

Therefore Krízek's measure is not corrected by chance.

Guzmán-Martinez's Measure

We remind the reader that in Guzmán-Martínez and Alaiz-Rodríguez [2011], 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.*

$$\begin{aligned}\mathbb{E}\left[\hat{\Phi}_{Guzman}(\mathcal{Z})|H_0\right] &= 1 - \frac{\frac{1}{d}}{\frac{k}{d}\log(\frac{k}{d})}\sum_{f=1}^{d}\mathbb{E}\left[\hat{p}_f \log \hat{p}_f|H_0\right] \\ &= 1 - \frac{\frac{1}{d}}{\frac{k}{d}\log(\frac{k}{d})}\sum_{f=1}^{d}\mathbb{E}\left[\frac{\sum_{i=1}^{M} z_{i,f}}{M}\log\frac{\sum_{i=1}^{M} z_{i,f}}{M}|H_0\right]. \qquad \text{(B.4)}\end{aligned}$$

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 B.4,

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

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

$$\mathbb{E}\left[\frac{\sum_{i=1}^{M} z_{i,f}}{M}\log\frac{\sum_{i=1}^{M} z_{i,f}}{M}|H_0\right] = \mathbb{E}\left[\frac{Y_f}{M}\log\frac{Y_f}{M}|H_0\right] = \mathbb{E}\left[g(Y_f)|H_0\right]. \qquad \text{(B.5)}$$

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

$$\begin{aligned}&\mathbb{E}\left[g(Y_f)|H_0\right] \geq g\left(\mathbb{E}\left[Y_f|H_0\right]\right) \\ \Rightarrow\quad &\mathbb{E}\left[g(Y_f)|H_0\right] \geq g\left(M\frac{k}{d}\right) \\ \Rightarrow\quad &\mathbb{E}\left[g(Y_f)|H_0\right] \geq \frac{k}{d}\log\frac{k}{d}\end{aligned}$$

[1] 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.

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

As shown by Equation B.4 and B.5, the left-hand-side term is equal to $\mathbb{E}\left[\hat{\Phi}(\mathcal{Z})|H_0\right]$*, therefore we get*

$$\mathbb{E}\left[\hat{\Phi}(\mathcal{Z})|H_0\right] \leq 0.$$

Since we know that $\hat{\Phi}(\mathcal{Z})$ *is a positive quantity, this gives us that* $\mathbb{E}\left[\hat{\Phi}(\mathcal{Z})|H_0\right] = 0$*.*

Relative Weighted Consistency CW_{rel}

Using Equation A.2, the result of Theorem 3 and by linearity of the expectation, we get

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

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*

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

which is not constant. Nevertheless, we note that when $\lfloor \bar{k} \rfloor = \bar{k}$*, we have* $\mathbb{E}\left[\hat{\Phi}(\mathcal{Z})|H_0\right] \xrightarrow[M\to\infty]{} 0$ *and therefore the relative weighted consistency* CW_{rel} *is asymptotically corrected by chance. This is a result we expect since when the number of selected features is constant, this measure is asymptotically equivalent to Kuncheva's measure (Theorem 1).*

Lausser's Measure

Using the expression of Lausser's measure given in Equation B.3, we have

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

As we have seen in Theorem 8, under the Null Model of Feature Selection, we have that: $p_f = \frac{k}{d}$*. 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 the Correction for chance property.

B.6 Proofs of the Bounds on the Proposed Measure

In this section, we prove the lower bound of the proposed stability measure given in Definition 3. To do so, we first prove the lemma below that will be used later on.

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

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

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

$$
\begin{aligned}
\Rightarrow \quad & \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2 - \left(\frac{\bar{k}}{d}\right)^2 \geq 0 \\
\Rightarrow \quad & \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2 \geq \left(\frac{\bar{k}}{d}\right)^2 \\
\Rightarrow \quad & \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f^2 - \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f \geq \left(\frac{\bar{k}}{d}\right)^2 - \frac{1}{d}\sum_{f=1}^{d}\hat{p}_f \\
\Rightarrow \quad & -\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f) \geq \left(\frac{\bar{k}}{d}\right)^2 - \frac{\bar{k}}{d} \\
\Rightarrow \quad & -\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f) \geq -\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right) \\
\Rightarrow \quad & \frac{\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f)}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)} \leq 1 \\
\Rightarrow \quad & \frac{\frac{1}{d}\frac{M}{M-1}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f)}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)} \leq \frac{M}{M-1} \\
\Rightarrow \quad & 1 - \frac{\frac{1}{d}\frac{M}{M-1}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f)}{\frac{\bar{k}}{d}\left(1-\frac{\bar{k}}{d}\right)} \geq 1 - \frac{M}{M-1} \\
\Rightarrow \quad & \hat{\Phi}(\mathcal{Z}) \geq -\frac{1}{M-1}.
\end{aligned}
$$

As we can see, the measure is lower bounded by -1 *(since* $M \geq 2$*), but is asymptotically bounded by* 0*.*

Bibliography

Thomas Abeel, Thibault Helleputte, Yves Van de Peer, Pierre Dupont, and Yvan Saeys. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics*, 2, 2010.*

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

Salem Alelyani, Huan Liu, and Lei Wang. The effect of the characteristics of the dataset on the selection stability. In IEEE 23rd International Conference on Tools with Artificial Intelligence (ICTAI)*, 2011.*

Wilker Altidor, Taghi M. Khoshgoftaar, and Amri Napolitano. A noise-based stability evaluation of threshold-based feature selection techniques. In IRI'11*, 2011.*

Francis R. Bach. Bolasso: Model consistent lasso estimation through the bootstrap. In Proceedings of the 25th International Conference on Machine Learning*, ICML '08, 2008.*

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, 2017.*

Annalisa Barla, Sofia Mosci, Lorenzo Rosasco, and Alessandro Verri. A method for robust variable selection with significance assessment. In ESANN 2008, 16th European Symposium on Artificial Neural Networks, Bruges, Belgium, April 23-25, 2008, Proceedings*, pages 83–88, 2008.*

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

Waad Bouaguel, Emna Mouelhi, and Ghazi Bel Mufti. New Method for Instance Feature Selection Using Redundant Features for Biological Data*, pages 398–405. Springer International Publishing, 2016.*

Anne-Laure Boulesteix and Martin Slawski. Stability and aggregation of ranked gene lists. Briefings in Bioinformatics*, 2009.*

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. Classification and Regression Trees*. Wadsworth and Brooks, 1984.*

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.*

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

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*, 2006.*

Nicoletta Dessì and Barbara Pes. Stability in Biomarker Discovery: Does Ensemble Feature Selection Really Help? In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 191–200. Springer, 2015.*

Nicoletta Dessì, Barbara Pes, and Marta Angioni. On stability of ensemble gene selection. In Intelligent Data Engineering and Automated Learning - IDEAL 2015 - 16th International Conference Wroclaw, Poland.*, pages 416–423, 2015.*

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).*, 2012.*

David J. Dittman, Taghi M. Khoshgoftaar, Randall Wald, and Amri Napolitano. Classification performance of rank aggregation techniques for ensemble gene selection. In FLAIRS Conference*. AAAI Press, 2013.*

Gregroy Ditzler, Robi Polikar, and Gail Rosen. A bootstrap based neyman-pearson test for identifying variable importance. IEEE Transactions on Neural Networks and Learning Systems*, 2014.*

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.

Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In International Conference on World Wide Web, Proceedings*, 2001.*

Bradley Efron and Robert J. Tibshirani. An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 1993.*

J.L. Fleiss et al. 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.*

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software, Articles*, 33(1):1–22, 2010.*

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 ECML/PKDD*, 2009.*

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

Isabelle Guyon and André Elisseeff. An Introduction to Feature Extraction*. Springer Berlin Heidelberg, 2006.*

Roberto Guzmán-Martínez and Rocío Alaiz-Rodríguez. Feature selection stability assessment based on the jensen-shannon divergence. In European Conference on Machine Learning. *Springer, 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*, 2012.*

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The Elements of Statistical Learning. *Springer, 2001.*

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning: data mining, inference and prediction*. Springer, 2 edition, 2009.*

Anne-Claire C. Haury, Pierre Gestraud, and Jean-Philippe P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PloS one*, 6, 2011.*

Zengyou He and Weichuan Yu. Review article: Stable feature selection for biomarker discovery. Comput. Biol. Chem.*, 2010.*

A Hoerl and R Kennard. Ridge regression, in encyclopedia of statistical sciences, vol. 8, 1988.

Giuseppe Jurman, Stefano Merler, Annalisa Barla, Silvano Paoli, Antonio Galea, and Cesare Furlanello. Algebraic stability indicators for ranked lists in molecular profiling. Bioinformatics*, 2008.*

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms. In IEEE International Conference on Data Mining*, pages 218–255, 2005.*

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst.*, 2007.*

Iman Kamkar, Sunil Kumar Gupta, Dinh Phung, and Svetha Venkatesh. AI 2015: Advances in Artificial Intelligence: 28th Australasian Joint Conference, Canberra, ACT, Australia, Proceedings*, chapter Stable Feature Selection with Support Vector Machines. Springer, 2015.*

Iman Kamkar, Sunil Kumar Gupta, Dinh Phung, and Svetha Venkatesh. Stabilizing l1-norm prediction models by supervised feature grouping. Journal of Biomedical Informatics*, 59(Supplement C):149 – 168, 2016.*

Pavel Krízek, Josef Kittler, and Václav Hlavác. Improving stability of feature selection methods. In CAIP*, 2007.*

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

Ludmila I. Kuncheva, Christopher J. Smith, Yasir Iftikhar Syed, Christopher O. Phillips, and Keir Edward Lewis. Evaluation of feature ranking ensembles for high-dimensional biomedical data: A case study. In ICDM Workshops*. IEEE Computer Society, 2012.*

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*, 2012.*

Jonathan L Lustgarten, Vanathi Gopalakrishnan, and Shyam Visweswaran. Measuring stability of feature selection in biomedical datasets. AMIA Annu Symp Proc*, 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 and Gavin Brown. Measuring the stability of feature selection with applications to ensemble methods. In Multiple Classifier Systems - 12th International Workshop, MCS*, 2015.*

Sarah Nogueira and Gavin Brown. Measuring the stability of feature selection. In ECML/PKDD*, pages 442–457, 2016.*

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the use of spearman's rho to measure the stability of feature rankings. In Pattern Recognition and Image Analysis: 8th Iberian Conference (IbPRIA)*. Springer International Publishing, 2017.*

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection. Under review*, 2018.*

Liam Paninski. Estimation of entropy and mutual information. Neural Comput.*, 15 (6):1191–1253, 2003.*

H. Peng, Fulmi Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2005.*

Adam Pocock and Gavin Brown. Feast, 2014. `http://mloss.org/software/view/386/`*.*

Yvan Saeys, Thomas Abeel, and Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In ECML/PKDD (2)*, 2008.*

Qiang Shen, Ren Diao, and Pan Su. Feature selection ensemble. In Turing-100 - The Alan Turing Centenary, Manchester, UK.*, 2012.*

L. Shi, L. H. Reid, W. D. Jones, et al.

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*, 2010.*

Gustavo Stolovitzky. Gene selection in microarray data: the elephant, the blind men and our algorithms. Current Opinion in Structural Biology*, 13(3):370–376, 2003.*

Robert Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B*, 1994.*

Randall Wald, Taghi M. Khoshgoftaar, and David J. Dittman. A new fixed-overlap partitioning algorithm for determining stability of bioinformatics gene rankers. In 11th International Conference on Machine Learning and Applications, ICMLA*, 2012a.*

Randall Wald, Taghi M. Khoshgoftaar, David J. Dittman, Wael Awada, and Amri Napolitano. An extensive comparison of feature ranking aggregation techniques in bioinformatics. In IRI*. IEEE, 2012b.*

Randall Wald, Taghi M. Khoshgoftaar, and Ahmad Abu Shanab. The effect of measurement approach and noise level on gene selection stability. In 2012 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2012, Philadelphia, PA, USA, October 4-7, 2012*, 2012c.*

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.*

Larry Wasserman. All of statistics : a concise course in statistical inference*. Springer, 2010.*

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 KDD*, 2008.*

Min Zhang, Lin Zhang, Jinfeng Zou, Chen Yao, Hui Xiao, Qing Liu, Jing Wang, Dong Wang, Chenguang Wang, and Zheng Guo. Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. Bioinformatics*, 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.*