On The Stability of Feature Selection in the Presence of Feature Correlations

Konstantinos Sechidis[1], Konstantinos Papangelou[1], Sarah Nogueira[2], James Weatherall[2], and Gavin Brown[1]

[1] School of Computer Science, University of Manchester, M13 9PL, UK
{konstantinos.sechidis,konstantinos.papangelou,gavin.brown}@manchester.ac.uk
[2] Criteo, Paris, France
s.nogueira@criteo.com
[3] Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK.
James.Weatherall@astrazeneca.com

Abstract. Feature selection is central to modern data science. The 'stability' of a feature selection algorithm refers to the sensitivity of its choices to small changes in training data. This is, in effect, the *robustness* of the chosen features. This paper considers the estimation of stability when we expect strong pairwise correlations, otherwise known as feature *redundancy*. We demonstrate that existing measures are inappropriate here, as they systematically *underestimate* the true stability, giving an overly pessimistic view of a feature set. We propose a new statistical measure which overcomes this issue, and generalises previous work.

1 Introduction

Feature Selection (FS) is central to modern data science—from exploratory data analysis, to predictive model building. The overall question we address with this paper is "*how can we quantify the reliability of a feature selection algorithm?*". The answer to this has two components — first, how *useful* are the selected features when used in a predictive model; and second, how *sensitive* are the selected features, to small changes in the training data. The latter is known as *stability* [9]. If the selected set varies wildly, with only small data changes, perhaps the algorithm is not picking up on generalisable patterns, and is responding to noise. From this perspective, we can see an alternative (and equivalent) phrasing, in that we ask *"how reliable is the set of chosen features?"* — i.e. how likely are we to get a different recommended feature set, with a tiny change to training data. This is particularity important in domains like bioinformatics, where the chosen features are effectively hypotheses on the underlying biological mechanisms.

There are many measures of stability proposed in the literature, with a recent study [14] providing a good summary of the advantages and disadvantages of each. The particular contribution of this paper is on how to estimate stability in the presence of *correlated features*, also known as feature *redundancy*. We will demonstrate that any stability measure not taking such redundancy into account

necessarily gives a *systematic under-estimate* of the stability, thus giving an overly pessimistic view of a given FS algorithm. This systematic under-estimation of stability can have a variety of consequences, depending on the application domain. In biomedical scenarios, it is common to use data-driven methods to generate candidate biomarker sets, that predict disease progression [16]. If we are comparing two biomarker sets, we might estimate their stability, judge one to be unstable, and discard it. However, if there are background feature correlations, and thus we are overly conservative on the stability, we might miss an opportunity.

We provide a solution to this problem, with a novel stability measure that takes feature redundancy into account. The measure generalises a recent work [14] with a correction factor that counteracts the systematic under-estimation of stability. Since the selection of a FS algorithm can be seen as a multi-objective optimisation problem we show how the choice of a stability measure changes the Pareto-optimal solution[1]. Additionally, we demonstrate the utility of the measure in the context of biomarker selection in medical trials, where strong correlations and necessary robustness of the choices are an unavoidable part of the domain.

2 Background

We assume a dataset $\mathcal{D} = \{\mathbf{x}_n, y_n\}_{n=1}^N$, with a d-dimensional input $\mathbf{x}$. The task of feature selection is to choose a subset of the dimensions, of size $k \ll d$, subject to some constraints; typically we would like to select the smallest subset that contains all the relevant information to predict y.

2.1 Estimating the Stability of Feature Selection

Let us assume we take $\mathcal{D}$ and run some feature selection algorithm, such as L1 regularization where we take non-zero coefficients to be the 'selected' features, or ranking features by their mutual information with the target [3]. When using all N datapoints, we get a subset of features: $s_{\mathcal{D}}$. We would like to know the *reliability* of the chosen feature set under small perturbations of the data. If the algorithm changes preferences drastically, with only small changes in the training data, we might prefer not to trust the set $s_{\mathcal{D}}$, and judge it as an 'unstable' set.

To quantify this, we repeat the same selection procedure M times, but each time leaving out a small random fraction δ of the original data. From this we obtain a sequence $S = \{s_1, s_2, ..., s_M\}$, where each subset came from applying a FS algorithm to a different random perturbation of the training data. At this point it turns out to be more notationally and mathematically convenient to abandon the set-theoretic notation, and use instead a matrix notation. We can treat the sequence S as an $M \times d$ binary matrix, where the d columns represent whether or not (1/0) each feature was chosen on each of the M repeats. For example, selecting from a pool of $d = 6$ features, and $M = 4$ runs:

[1] The software related to this paper will be available at: https://github.com/sechidis

$$\mathcal{Z} = \begin{array}{c} \begin{matrix} Z_1 & Z_2 & Z_3 & Z_4 & Z_5 & Z_6 \end{matrix} \\ \begin{pmatrix} \mathbf{1} & 0 & \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 \\ \mathbf{1} & 0 & 0 & \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 & \mathbf{1} & 0 & 0 \end{pmatrix} \end{array} \begin{array}{ll} ...\mathbf{z}_1, & \text{selections on 1st run} \\ ...\mathbf{z}_2, & \text{selections on 2nd run} \\ & ... \end{array} \tag{1}$$

We then choose some measure $\phi(a, b)$ of similarity between the resulting feature sets from two runs, and evaluate the stability from $\mathcal{Z}$, as an average over all possible pairs:

$$\hat{\Phi}(\mathcal{Z}) = \frac{1}{M(M-1)} \sum_i \sum_{j \neq i} \phi(\mathbf{z}_i, \mathbf{z}_j) \tag{2}$$

Let us take for example $\phi(\mathbf{z}_i, \mathbf{z}_j)$ to be a dot-product of the two binary strings. For a single pair, this would correspond to the number of selected features that are common between the two – or the *size of the subset intersection*. Over the M runs, this would correspond to the average subset intersection—so on average, if the feature subsets have large pairwise intersection, the algorithm is returning similar subsets despite the data variations. This of course has the disadvantage that the computation expands quadratically with M, and large M is necessary to get more reliable estimates. Computation constraints aside, if the result indicated sufficiently high stability (high average subset intersection) we might decide we can trust $s_{\mathcal{D}}$ and take it forward to the next stage of the analysis.

A significant body of research, e.g. [5, 9, 10, 17], suggested different similarity measures ϕ that could be used, and studied properties. Kuncheva [11] conducted an umbrella study, demonstrating several undesirable behaviours of existing measures, and proposing an axiomatic framework to understand them. Nogueira et al. [14] extended this, finding further issues and avoiding the pairwise, set-theoretic, definition of ϕ entirely—presenting a measure in closed form, allowing computation in $O(Md)$ instead of $O(M^2 d)$. From the matrix $\mathcal{Z}$, we can estimate various stochastic quantities, such as the average number of features selected across M runs, denoted as $\bar{k}$ and the probability that the feature X_f was selected, denoted as $p_f = \mathbb{E}[Z_f = 1]$. Using these, their recommended stability measure is,

$$\hat{\Phi}(\mathcal{Z}) = 1 - \frac{\sum_f \frac{M}{M-1} \hat{p}_f (1 - \hat{p}_f)}{\bar{k}(1 - \frac{\bar{k}}{d})} \tag{3}$$

The measure also generalises several previous works (e.g. [11]), and was shown to have numerous desirable statistical properties. For details we refer the reader to [14], but the intuition is that the numerator measures the average sample variance, treating the columns of $\mathcal{Z}$ as Bernoulli variables; the denominator is a normalizing term that ensures $\hat{\Phi}(\mathcal{Z}) \in [0, 1]$, as $M \to \infty$.

In the following section we illustrate how stability becomes much more complex to understand and measure, when there are either observed feature correlations, or background domain knowledge on the dependencies between features.

2.2 The Problem: Estimating Stability under Feature Correlations

The example in eq. (1) can serve to illustrate an important point. On each run (each row of $\mathcal{Z}$) the algorithm seems to change its mind about which are the important features—first 1&3, then 2&3, then 1&4, and finally 2&4. Various measures in the literature, e.g. [14] will identify this to be *unstable* as it changes its feature preferences substantially on every run. However, suppose we examine the original data, and discover that features X_1 and X_2 are very *strongly correlated*, as are X_3 and X_4. For the purposes of building a predictive model these are interchangeable, redundant features. What should we now conclude about stability? Since the algorithm always selects one feature from each strongly correlated pair, it always ends up with *effectively the same information* with which to make predictions — thus we should say that it *is in fact perfectly stable*. This sort of scenario is common to (but not limited to) the biomedical domain, where genes and other biomarkers can exhibit extremely strong pairwise correlations. A further complication also arises in this area, in relation to the *semantics* of the features. Certain features may or may not have strong observable *statistical* correlations, but for the purpose of interpretability they hold very similar *semantics* – e.g. if the algorithm alternates between two genes, which are not strongly correlated, but are both part of the renal metabolic pathway, then we can determine that the kidney is playing a stable role in the hypotheses that the algorithm is switching between.

To the best of our knowledge there are only two published stability measures which take correlations/redundancy between features into account, however both have significant limitations. The measure of Yu et al. [19] requires the estimation of a mutual information quantity between features, and the solution of a constrained optimisation problem (bipartite matching), making it quite highly parameterised, expensive, and stochastic in behaviour. The other is nPOGR [20] which can be shown to have several pathological properties [14]. In particular, the measure is not lower-bounded which makes interpretation of the estimated value very challenging – we cannot judge how "stable" a FS algorithm is without a reference point. The nPOGR measure is also very computationally demanding, requiring generation of random pairs of input vectors, and computable in $O(M^2 d)$. To estimate stability in large scale data, computational efficiency is a critical factor.

In the next section, we describe our approach for estimating stability under strong feature correlations, which also allows incorporation of background knowledge, often found in biomedical domains.

3 Measuring Stability in the Presence of Correlations

As discussed in the previous section, a simple stability measure can be derived if we define $\Phi(\cdot, \cdot)$ as the *size of the intersection* between two subsets of feature, and apply eq. (2). The more co-occurring features between repeated runs, the more stable we regard the algorithm to be. It turns out that, to understand stability in the presence of correlated features, we need to revise our concept of subset *intersection*, to one of *effective subset intersection*.

3.1 Subset Intersection and *Effective* Subset Intersection

We take again the example from eq. (1). We have $\mathbf{z}_1 = [1, 0, 1, 0, 0, 0]$, and $\mathbf{z}_2 = [0, 1, 1, 0, 0, 0]$. The subset intersection, given by the inner product is $\mathbf{z}_1 \, \mathbf{z}_2^T = 1$, due to the selection of the third feature. But, as mentioned, perhaps we learn that in the original data, X_1 and X_2 are strongly correlated, effectively *interchangeable* for the purposes of building a predictive model. When comparing the two subsets, X_1 and X_2 should be treated similarly, thus increasing the size of the intersection to 2. Hence, we do not have a simple subset intersection, but instead an *effective* subset intersection, based not on the *indices* of the features (i.e. X_1 vs X_2) but instead on the *utility* or *semantics* of the features.

We observed that the intersection between two subsets s_i and s_j, i.e. the two rows $\mathbf{z}_i$ and $\mathbf{z}_j$ of the binary matrix $\mathcal{Z}$, can be written as an inner product: $r_{i,j} = |s_i \cap s_j| = \mathbf{z}_i \, \mathbb{I}_d \, \mathbf{z}_j^T$ where $\mathbb{I}_d$ is the $d \times d$ identity matrix. We can extend this with a *generalised* inner product, where the inner product matrix will capture the feature relationships.

Definition 1 (Effective subset intersection). *The "effective" subset intersection with correlated features is given by the generalised inner product:*

$$r_{i,j}^{\mathbb{C}} = |s_i \cap s_j|_{\mathbb{C}} = \mathbf{z}_i \, \mathbb{C} \, \mathbf{z}_j^T$$

The inner product matrix $\mathbb{C}$ *has diagonal elements set to* 1*, while the off-diagonals capture the relationships between pairs of features, i.e.*

$$\mathbb{C} = \begin{bmatrix} 1 & c_{1,2} & \dots & c_{1,d} \\ c_{2,1} & 1 & \dots & c_{2,d} \\ \vdots & \vdots & \vdots & \vdots \\ c_{d,1} & c_{d,2} & \dots & 1 \end{bmatrix} \tag{4}$$

with $c_{f,f'} = c_{f',f} > 0 \; \forall \; f \neq f'$.

The entries of the matrix $\mathbb{C}$ *could* be absolute correlation coefficients $c_{f,f'} = |\rho_{X_f, X_{f'}}|$ thus capturing redundancy as explained by the data. But in general we emphasise that entries of $\mathbb{C}$ are not necessarily statistical correlations between features. For example, $\mathbb{C}$ could be a binary matrix, where $c_{f,f'} = \delta(|\rho_{X_f, X_{f'}}| > \theta)$, or constructed based on domain knowledge, thus capturing redundancy as explained by domain experts (e.g. two biomarkers appearing in the same metabolic pathway). The following theorem shows why we are guaranteed to underestimate the stability, if feature redundancy is not taken into account.

Theorem 1. *The effective intersection is greater than or equal to intersection,*

$$|s_i \cap s_j|_{\mathbb{C}} \quad \geq \quad |s_i \cap s_j|$$

The proof of this can be seen by relating the "traditional" intersection $|s_i \cap s_j|$ and the "effective" intersection as follows:

Lemma 1. *The effective intersection can be written,*

$$|s_i \cap s_j|_{\mathbb{C}} \;=\; |s_i \cap s_j| \;+\; \sum_{f=1}^{d} \sum_{\substack{f'=1 \\ f' \neq f}}^{d} c_{f,f'} z_{i,f} z_{j,f'}$$

If all entries in $\mathbb{C}$ are non-negative, we have $r^{\mathbb{C}}_{i,j} \geq r_{i,j}$ — without this correction, we will systematically under-estimate the true stability.

The set-theoretic interpretation of stability is to be contrasted with the binary matrix representation $\mathcal{Z} \in \{0,1\}^{M \times d}$. Nogueira et al. [14] proved the following result, bridging these two conceptual approaches to stability. The average subset intersection among M feature sets can be written,

$$\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} |s_i \cap s_j| \quad = \quad \overline{k} - \sum_{f=1}^{d} \widehat{\text{var}}(Z_f)$$

where $\overline{k}$ is the average number of features selected over M rows, and $\widehat{\text{var}}(Z_f) = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$, i.e. the unbiased estimator of the variance of the Bernoulli random feature Z_f. Then a stability measure defined as an increasing function of the intersection can be equivalently phrased as a decreasing function of the variance of the columns of the selection matrix, thus bridging the set-theoretic view with a probabilistic view. This property is also known as monotonicity [11, 14] and is a defining element of a stability measure. In the presence of redundancy we instead would like our measure to be an increasing function of the effective intersection. The following theorem bridges our set-theoretic view with the statistical properties of the selection matrix in the presence of feature redundancy captured in the matrix $\mathbb{C}$.

Theorem 2. *The effective average pairwise intersection among the M subsets can be written:*

$$\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} |s_i \cap s_j|_{\mathbb{C}} \quad = \quad \overline{k}_{\mathbb{C}} - \text{tr}(\mathbb{C}\mathbf{S})$$

where $\overline{k}_{\mathbb{C}} = \sum_{f=1}^{d} \sum_{f'=1}^{d} c_{f,f'} \overline{z}_{f,f'}$ the effective average number of features selected over M runs. The unbiased estimator of the covariance between Z_f and $Z_{f'}$ is $\widehat{\text{cov}}(Z_f, Z_{f'}) = \frac{M}{M-1}(\hat{p}_{f,f'} - \hat{p}_f \hat{p}_{f'})$, $\forall\ f, f' \in \{1...d\}$, while $\mathbf{S}$ is an unbiased estimator of the variance-covariance matrix of $\mathcal{Z}$.

Proof: Provided in Supplementary material Section A.

We are now in position to introduce our new measure, which based on the above theorem should be a decreasing function of $\text{tr}(\mathbb{C}\mathbf{S})$. There a final element that needs to be taken into account—we need to normalise our estimation to bound it so that it can be interpretable and comparable between different FS approaches, developed in the next section.

3.2 A Stability Measure for Correlated Features

Based on the previous sections, we can propose the following stability measure.

Definition 2 (Effective Stability). *Given a matrix of feature relationships $\mathbb{C}$, the effective stability is*

$$\hat{\Phi}_{\mathbb{C}}(\mathcal{Z}) \quad = \quad 1 - \frac{\mathrm{tr}(\mathbb{C}\mathbf{S})}{\mathrm{tr}(\mathbb{C}\mathbf{\Sigma^0})},$$

where $\mathbf{S}$ is an unbiased estimator of the variance-covariance matrix of $\mathcal{Z}$, i.e. $\mathbf{S}_{f,f'} = \widehat{\mathrm{Cov}}(Z_f, Z_{f'}) = \frac{M}{M-1}(\hat{p}_{f,f'} - \hat{p}_f\hat{p}_{f'})$, $\forall\ f, f' \in \{1...d\}$, while $\mathbf{\Sigma^0}$ is the matrix which normalises the measure.

To derive a normaliser, we need to estimate the variance/covariance under the *Null Model* of feature selection [14, Definition 3]. The Null Model expresses the situation where there is *no preference* toward any particular subset, and all subsets of size k have the same probability of occurrence, thus accounting for the event of a completely random selection procedure. For a detailed treatment of this subject we refer the reader to the definition of this, by Nogueira et al. [14].

Theorem 3. *Under the Null Model, the covariance matrix of $\mathcal{Z}$ is given by:*

$$\mathbf{\Sigma^0} = \begin{bmatrix} \mathrm{var}\left(Z_1 \middle| H_0\right) & \dots & \mathrm{cov}\left(Z_1, Z_d \middle| H_0\right) \\ \vdots & \ddots & \vdots \\ \mathrm{cov}\left(Z_d, Z_1 \middle| H_0\right) & \dots & \mathrm{var}\left(Z_d \middle| H_0\right) \end{bmatrix},$$

where the main diagonal elements are given by: $\mathrm{var}\left(Z_f \middle| H_0\right) = \frac{\overline{k}}{d}\left(1 - \frac{\overline{k}}{d}\right)$ *and the off-diagonal elements, $f \neq f'$ are:* $\mathrm{cov}\left(Z_f, Z_{f'} \middle| H_0\right) = \frac{\overline{k^2} - \overline{k}}{d^2 - d} - \frac{\overline{k}^2}{d^2}$

Proof: Provided in Supplementary material Section B.

It can immediately be seen that the proposed measure is a generalisation of Nogueira et al. [14], as it reduces to eq. (3) when $\mathbb{C}$ is the identity, in which case $\mathrm{tr}(\mathbb{C}S) = \sum_i \mathrm{var}(z_i)$. At this point we can observe that when $\mathbb{C} = \mathbb{I}_d$ we implicitly assume the columns of the selection matrix to be independent variables hence considering only their variance. In contrast, our measure accounts additionally for all pairwise covariances weighted by the coefficients of the matrix $\mathbb{C}$. As we already discussed these coefficients can be seen as our confidence on the correlation between the columns of the selection matrix as explained by the data (using for example Spearman's correlation coefficient) or by domain experts.

Finally, we can summarise the protocol for estimating the stability of a FS procedure in a simple algorithm shown in Algorithm 1. We also compare the computational time of our measure against nPOGR, as the dimensionality of the feature set increases—shown in fig. 1—we observe that our measure is as expected, orders of magnitude faster to compute.

In the next section, we demonstrate several cases where incorporating prior knowledge and using our proposed stability measure, we may arrive to completely different conclusions on the reliability of one FS algorithm versus another, hence potentially altering strategic decisions in a data science pipeline.

Algorithm 1: Recommended protocol for estimating FS stability.

Input : A dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, where $\mathbf{x}$ is d-dimensional.
A procedure $f(\mathcal{D})$ returning a subset of features $s_{\mathcal{D}}$, of size $k < d$.
A matrix $\mathbb{C}$, specifying known feature redundancies.

Output : Stability estimate $\hat{\Phi}$, for feature set $s_{\mathcal{D}}$.

Define $\mathcal{Z}$, an empty matrix of size $M \times d$.
for $j := 1$ *to* M **do**
  Generate $\mathcal{D}_j$, a random sample from $\mathcal{D}$ (e.g. leave out 5% rows, or bootstrap)
  Set $s_j \leftarrow f(\mathcal{D}_j)$
  Set the jth row of $\mathcal{Z}$ as the binary string corresponding to selections s_j.

Return stability estimate $\hat{\Phi}(\mathcal{Z})$ using Definition 2.

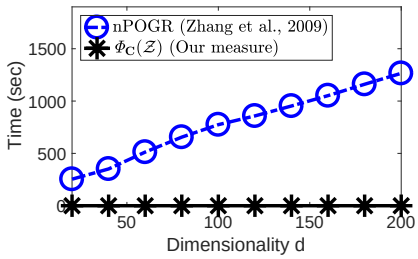


Fig. 1: Computational cost of nPOGR versus our measure as the number of features grow. We generated randomly selection matrices $\mathcal{Z}$ of dimension M×d, with $M = 50$ and various values of d. The proposed measure remains largely unaffected by the dimensionality (taking milliseconds).

4 Experiments

Our experimental study is split in two sections. Firstly we will show how our measure can be used for choosing between different feature selection criteria in real-world datasets. We will apply the protocol described in the previous section to estimate the stability which along with the predictive performance of the resulting feature set can give the full picture on the performance of a FS procedure. Secondly, we will show how we can use stability in clinical trials data to identify robust groups of biomarkers.

4.1 Pareto-Optimality using Effective Stability

In many applications, given a dataset we might wish to apply several feature selection algorithms, which we evaluate and compare. The problem of deciding which FS algorithm we should trust can be seen as a multi-objective optimisation combining two criteria: (1) the features result in high accuracy, and (2) we want

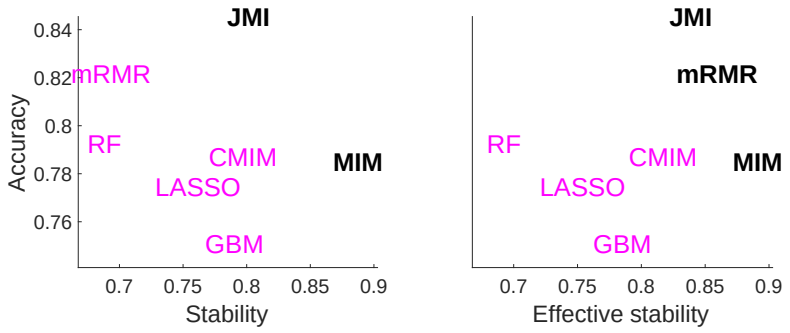
algorithms that generate *stable subsets*, i.e. stable hypotheses on the underlying mechanisms. In this context, we define the Pareto-optimal set as the set of points for which no other point has both higher accuracy and higher stability, thus the members of the Pareto-optimal set are said to be non-dominated [7]. In this section we will explore whether using the proposed stability measure, $\hat{\Phi}_{\mathbb{C}}(\mathcal{Z})$, can result in different optimal solutions in comparison with the original measure, $\hat{\Phi}(\mathcal{Z})$, that ignores feature redundancy.

We used ten UCI datasets and created $M = 50$ versions of each one of them by removing 5% of the examples at random. We applied several feature selection algorithms and evaluated the predictive power of the selected feature sets using a simple nearest neighbour classifier (3-nn). By using this classifier we make few assumptions about the data and avoid additional variance from hyperparameter tuning. For each dataset, we estimated the accuracy on the hold-out data (5%). To ensure a fair comparison of the feature selection methods, all algorithms are tuned to return the top-k features for a given dataset. We chose k to be the 25% of the number of features d of each dataset. Here we provide a short description of the feature selection methods we used and implementation details.

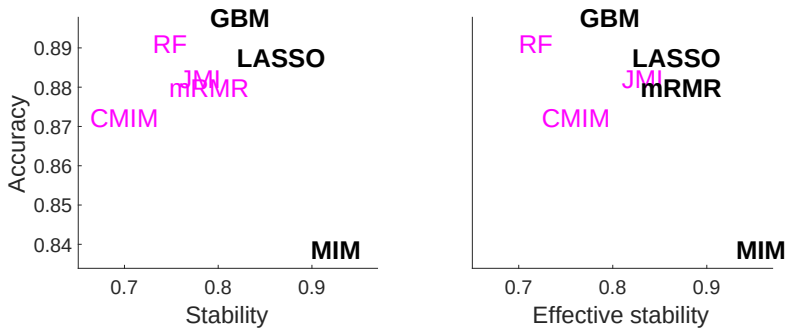
- **Penalized linear model (LASSO):** with the regularisation parameter λ tuned such that we get k non-zero coefficients—these are the selected features.
- **Tree-based methods (RF/GBM):** We used Random Forest (RF) [2] and Gradient Boosted Machines (GBM) with decision stumps [8] to choose the top-k features with highest importance scores. For both algorithms we used 100 trees.
- **Information theoretic methods (MIM/mRMR/JMI/CMIM):** We used various information theoretic feature selection methods, each one of them making different assumptions (for a complete description of the assumptions made by each method we refer the reader to [3]). For example MIM quantifies only the relevancy, mRMR the relevancy and redundancy [15], while the JMI [18] and CMIM [6] the relevancy, the redundancy and the complementarity. To estimate mutual and conditional mutual information terms, continuous features were discretized into 5 bins using an equal-width strategy.

The UCI datasets do not contain information about correlated features. In order to take into account possible redundancies we used Spearman's ρ correlation co-efficient to assess non-linear relationships between each pair of features. For estimating the effective stability, we incorporate these redundancies in the $\mathbb{C}$ matrix using the rule: $c_{f,f'} = \delta(|\rho_{X_f,X_{f'}}| > \theta)$. Following Cohen [4], two features X_f and $X_{f'}$ are assumed to be strongly correlated, when the co-efficient is greater than $\theta = 0.5$.

Figure 2 shows the Pareto-optimal set for two selected datasets. The criteria on the top-right dominate the ones on the bottom left and they are the ones that should be selected. We observe that by incorporating prior knowledge (r.h.s. in fig. 2a and fig. 2b) we change our view about the best-performing algorithms in terms of the accuracy/stability trade-off. Notice that mRMR, a criterion that penalizes the selection of redundant features, becomes much more stable using our proposed measure, $\hat{\Phi}_{\mathbb{C}}(\mathcal{Z})$. A summary of the Pareto-optimal solutions for



(a) `sonar`



(b) `ionosphere`

Fig. 2: Accuracy/Stability trade-off between different feature selection algorithms for two UCI datasets. The methods on top right corner are the Pareto-optimal solutions.

all datasets is given in table 1, where we can observe that similar changes occur in most cases.

Furthermore, table 2 shows the non-dominated rank of the different criteria across all datasets. This is computed per data set as the number of other criteria which dominate a given criterion, in the Pareto-optimal sense, and then averaged over the 10 datasets. Similarly to our earlier observations (fig. 2), the average rank of mRMR increases dramatically. Similarly JMI increases its average position, as opposed to MIM that captures only the relevancy.

In the next section, we describe how incorporating prior knowledge about the *semantics* of biomarkers may incur changes on the stability of feature selection in clinical trials.

4.2 Stability of Biomarker Selection in Clinical Trials

The use of highly specific biomarkers is central to personalised medicine, in both clinical and research scenarios. Discovering new biomarkers that carry prognostic

Table 1: Pareto-optimal solutions for 10 UCI datasets. We observe that in most cases incorporating prior knowledge about possible feature redundancies changes the optimal solutions.

Dataset	Pareto-Optimal set (accuracy vs stability)	Pareto-Optimal set (accuracy vs *effective* stability)	Change ?
`breast`	LASSO, MIM	MIM	✓
`ionosphere`	LASSO, GBM, MIM	LASSO, GBM, MIM mRMR	✓
`landsat`	mRMR	JMI	✓
`musk2`	LASSO, MIM	LASSO	✓
`parkinsons`	LASSO, MIM	MIM, mRMR, JMI	✓
`semeion`	GBM, MIM, mRMR, JMI	GBM, mRMR, JMI, CMIM	✓
`sonar`	MIM, JMI	MIM, mRMR, JMI	✓
`spect`	MIM	MIM	
`waveform`	GBM, mRMR	GBM, mRMR	
`wine`	MIM, CMIM	MIM, CMIM	

Table 2: Column 1: Non-dominated Rank of different criteria for the trade-off of accuracy/stability estimated by $\Phi(\mathcal{Z})$. Criteria with a higher rank (closer to 1.0) provide a better tradeoff than those with a lowerrank. Column 2: As column 1 but using our measure $\Phi_{\mathbb{C}}(\mathcal{Z})$ for estimating effective stability.

Accuracy/Stability	Accuracy/Effective stability
MIM (1.6)	mRMR (1.7)
GBM (1.8)	MIM (2)
JMI (2.6)	JMI (2.4)
LASSO (2.7)	GBM (2.4)
mRMR (2.9)	CMIM (2.9)
CMIM (2.9)	LASSO (3.1)
RF (3.1)	RF (3.1)

information is crucial for general patient care and for clinical trial planning, i.e. prognostic markers can be considered as covariates for stratification. A prognostic biomarker is a biological characteristic or a clinical measurement that provides information on the likely outcome of the patient irrespective of the applied treatment [16]. For this task, any supervised feature selection algorithm can be used to identify and rank the biomarkers with respect to the outcome Y. Having *stable* biomarker discovery algorithms, i.e. identifying biomarkers that can be reproduced across studies, is of great importance in clinical trials. In this section we will present a case study on how to evaluate the stability of different algorithms, and how we can incorporate prior knowledge over groups of biomarkers with semantic similarities.

We focus on the IPASS study [13], which evaluated the efficacy of the drug *gefitinib* (Iressa, AstraZeneca) versus first-line chemotherapy with *carboplatin* (Paraplatin, Bristol-Myers Squibb) plus *paclitaxel* (Taxol, Bristol-Myers Squibb)

Table 3: Top-4 prognostic biomarkers in IPASS for each competing method. The results can be interpreted by domain experts (e.g. clinicians) on their biological plausibility. However, to answer in what extend these sets are reproducible and how they can be affected by small changes in the data (such as patient dropouts) we need to evaluate their stability.

Rank	GBM	CMIM
1	EGFR expression (X_4)	EGFR mutation (X_2)
2	Disease stage (X_{10})	Serum ALP(X_{13})
3	WHO perform. status (X_1)	Blood leukocytes (X_{21})
4	Serum ALT(X_{12})	Serum ALT (X_{12})

in an Asian population of 1217 light- or non-smokers with advanced non-small cell lung cancer. A detailed description of the trial and the biomarkers used in the IPASS study are given in the Appendix A.

In this section we will focus on two commonly used algorithms: Gradient Boosted Machines [8] and conditional mutual information maximisation (CMIM) [6]. GBM sequentially builds a weighted voting ensemble of decision stumps based on single features, while CMIM is an information theoretic measure based on maximising conditional mutual information. These two methods are quite different in nature: for example GBM builds decision trees, while CMIM estimates two-way feature interactions. As a result, they often return different biomarker subsets and choosing which one to take forward in a phased clinical study is an important problem.

Table 3 presents the top-4 prognostic biomarkers derived by each method. We observe that the two methods return significantly different biomarker sets; Which one should we trust? To answer this question we estimate their stability with respect to data variations using $M = 50$ and 5% leave-out. This could simulate the scenario where for some patients we do not know the outcome e.g. they dropped out from the trial. In table 4 we see that when using $\widehat{\Phi}(\mathcal{Z})$, in agreement with data science folklore, GBM is judged a stable method, more so than CMIM.

But, with a closer study of the biomarkers considered in IPASS, there are in fact groups of them which are biologically related: **(Group A)** those that describe the receptor protein EGFR, X_2, X_3, X_4, **(Group B)** those which are measures of liver function, X_{12}, X_{13}, X_{14}, and **(Group C)** those which are counts of blood cells, $X_{20}, X_{21}, X_{22}, X_{23}$. There are also sub-groupings at play here. For instance, given that neutrophils are in fact a type of leukocyte (white blood cell), one may expect X_{21} and X_{22} to exhibit a stronger pairwise correlation than any other pair of cell count biomarkers.

We can take these groupings and redundancies into account by setting to 1, all of the elements in $\mathbb{C}$ matrix that represent pairs of features that belong the the same group. Table 4 compares the effective stability of the two algorithms using our novel measure $\widehat{\Phi}_{\mathbb{C}}(\mathcal{Z})$, which takes into account the groups A, B and C. This time, CMIM is substantially *more stable* than GBM—leading to the

Table 4: Stability and effective stability of GBM and CMIM in IPASS. The instability of CMIM is caused by variations within groups of semantically related biomarkers. When this is taken into account using $\widehat{\Phi}_{\mathbb{C}}(\mathcal{Z})$ the method is deemed more stable than GBM.

	GBM		CMIM
Stability $\widehat{\Phi}(\mathcal{Z})$	**0.87**	$>$	0.68
- within **Group A**	0.96		0.45
- within **Group B**	0.82		0.80
- within **Group C**	0.14		0.43
Effective stability $\widehat{\Phi}_{\mathbb{C}}(\mathcal{Z})$	0.87	$<$	**0.91**

conjecture that the instability in GBM is generated by variations *between groups*, while CMIM is caused by *within-group variations*.

To validate this conjecture, we calculate the stability within each group using $\widehat{\Phi}(\mathcal{Z})$. In table 4 we observe that CMIM has small stability, especially within the groups A and C. The algorithm alternates between selecting biomarkers that are biologically related, hence when we incorporate domain knowledge the effective stability of CMIM increases significantly. Thus, based on our prior knowledge on feature relationships, CMIM is the more desirable prospect to take forward.

5 Conclusions

We presented a study on the estimation of stability of feature selection in the presence of feature redundancy. This is an important topic, as it gives an indication of how reliable a selected subset may be, given correlations in the data or domain knowledge. We showed that existing measures are unsuitable and potentially misleading, also proving that many will systematically under-estimate the stability. As a solution to this, we presented a novel measure which allows us to incorporate information about correlated and/or semantically related features. An empirical study across 10 datasets and 7 distinct feature selection methods confirmed the utility, while a case study on real clinical trial data highlighted how critical decisions might be altered as a result of the new measure.

A IPASS description

The IPASS study [13] was a Phase III, multi-center, randomised, open-label, parallel-group study comparing gefitinib (Iressa, AstraZeneca) with carboplatin (Paraplatin, Bristol-Myers Squibb) plus paclitaxel (Taxol, Bristol-Myers Squibb) as first-line treatment in clinically selected patients in East Asia who had NSCLC. 1217 patients were balanced randomised (1:1) between the treatment arms, and the primary end point was progression-free survival (PFS); for full details of the trial see [13]. For the purpose of our work we model PFS as a Bernoulli endpoint,

neglecting its time-to-event nature. We analysed the data at 78% maturity, when 950 subjects have had progression events.

The covariates used in the IPASS study are shown in Table 5. The following covariates have missing observations (as shown in parentheses): X_5 (0.4%), X_{12} (0.2%), X_{13} (0.7%), X_{14} (0.7%), X_{16} (2%), X_{17} (0.3%), X_{18} (1%), X_{19} (1%), X_{20} (0.3%), X_{21} (0.3%), X_{22} (0.3%), X_{23} (0.3%). Following Lipkovich et al. [12], for the patients with missing values in biomarker X, we create an additional category, a procedure known as the *missing indicator method* [1].

Table 5: Covariates used in the IPASS clinical trial.

Biomarker	Description	Values
X_1	WHO perform. status	0 or 1, 2
X_2	EGFR mutation status	Negative, Positive, Unknown
X_3	EGFR FISH status	Negative, Positive, Unknown
X_4	EGFR expression status	Negative, Positive, Unknown
X_5	Weight	(0,50] , (50,60] , (60,70] , (70, 80] , (80, $+\infty$)
X_6	Race	Oriental, Other
X_7	Ethnicity	Chinese, Japanese, Other Asian, Other not Asian
X_8	Sex	Female, Male
X_9	Smoking status	Ex-Smoker, Smoker
X_{10}	Disease stage	Locally Advanced, Metastatic
X_{11}	Age	(0, 44] , [45,64] , [65,74] , [75, $+\infty$)
X_{12}	Serum ALT	Low, Medium, High
X_{13}	Serum ALP	Low, Medium, High
X_{14}	Serum AST	Low, Medium, High
X_{15}	Bilirubin	Low, Medium, High
X_{16}	Calcium	Low, Medium, High
X_{17}	Creatinine	Low, Medium, High
X_{18}	Potassium	Low, Medium, High
X_{19}	Sodium	Low, Medium, High
X_{20}	Blood hemoglobin	Low, Medium, High
X_{21}	Blood leukocytes	Low, Medium, High
X_{22}	Blood neutrophils	Low, Medium, High
X_{23}	Blood platelets	Low, Medium, High

References

1. Allison, P.D.: Missing Data. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136 (2001)
2. Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)
3. Brown, G., Pocock, A., Zhao, M.J., Luján, M.: Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. The Journal of Machine Learning Research 13(1), 27–66 (2012)

4. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences (2nd Edition). Routledge Academic (1988)
5. Dunne, K., Cunningham, P., Azuaje, F.: Solutions to instability problems with sequential wrapper-based approaches to feature selection. Tech. Rep. TCD-CS-2002-28, Trinity College Dublin, School of Computer Science (2002)
6. Fleuret, F.: Fast binary feature selection with conditional mutual information. Journal of Machine Learning Research (JMLR) 5, 1531–1555 (2004)
7. Fonseca, C.M., Fleming, P.J.: On the performance assessment and comparison of stochastic multiobjective optimizers. In: International Conference on Parallel Problem Solving from Nature. pp. 584–593. Springer (1996)
8. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
9. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms. In: IEEE International Conference on Data Mining. pp. 218–255 (2005)
10. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. (2007)
11. Kuncheva, L.I.: A Stability index for Feature Sselection. In: Artificial Intelligence and Applications (2007)
12. Lipkovich, I., Dmitrienko, A., D'Agostino Sr., R.B.: Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Statistics in Medicine 36(1), 136–196 (2017)
13. Mok, T.S., et al.: Gefitinib or Carboplatin/Paclitaxel in Pulmonary Adenocarcinoma. New England Journal of Medicine 361(10), 947–957 (2009)
14. Nogueira, S., Sechidis, K., Brown, G.: On the stability of feature selection algorithms. Journal of Machine Learning Research 18(174), 1–54 (2018), `http://jmlr.org/papers/v18/17-514.html`
15. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27(8), 1226–1238 (2005)
16. Sechidis, K., Papangelou, K., Metcalfe, P., Svensson, D., Weatherall, J., Brown, G.: Distinguishing prognostic and predictive biomarkers: An information theoretic approach. Bioinformatics 34(19), 3365–3376 (2018)
17. Shi, L., Reid, L.H., Jones, W.D., et al.: The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nature Biotechnology 24(9), 1151–61 (September 2006)
18. Yang, H.H., Moody, J.: Data visualization and feature selection: New algorithms for nongaussian data. In: NIPS. pp. 687–693 (1999)
19. Yu, L., Ding, C., Loscalzo, S.: Stable feature selection via dense feature groups. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 803–811. ACM (2008)
20. Zhang, M., Zhang, L., Zou, J., Yao, C., Xiao, H., Liu, Q., Wang, J., Wang, D., Wang, C., Guo, Z.: Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. Bioinformatics (2009)