# Measuring the Stability of Feature Selection Supplementary Material

Sarah Nogueira and Gavin Brown

School of Computer Science, University of Manchester,
Manchester M13 9PL, UK
{sarah.nogueira,gavin.brown}@manchester.ac.uk

This document is the supplementary material of [10]. In section 1, we provide a list of the notations used in the paper. In section 2, we provide the formal definitions of the similarity measures and the stability measures discussed. In section 3, we provide the proof of the table of properties (Table 1). In section 5, we provide the proofs of the 3 theorems in the paper and of Equation 3.

## 1 Notations

We shortly remind the notations of the paper.

- $M$ the number of bootstrap samples taken, also the number of feature sets.
- $d$ the total number of features
- $\mathcal{A} = [s_1, ..., s_M]^T = (x_{i,f})_{\substack{i\in\{1,...,M\}\\ f\in\{1,...,d\}}}$ is a binary matrix where $\mathbf{s}_i$ is the $i^{th}$ feature set in $\mathcal{A}$ and where $x_{i,f} = 1$ if the $f^{th}$ feature has been selected in the $i^{th}$ set, 0 otherwise.
- $\forall i \in \{1, ..., M\}, k_i = |\mathbf{s}_i|$ is the cardinality of set $\mathbf{s}_i$ (i.e. the number of features selected in $\mathbf{s}_i$). When all feature sets in $\mathcal{A}$ are of identical cardinality, we will simply denote it by $k$.
- $r_{i,j} = |\mathbf{s}_i \cap \mathbf{s}_j|$ the number of features that $\mathbf{s}_i$ and $\mathbf{s}_j$ have in common.
- $\mathbb{E}_{\nabla}[r_{i,j}] = \frac{k_i k_j}{d}$ is the expected size of the intersection of a procedure randomly selecting two sets of cardinality $k_i$ and $k_j$.
- $\phi_{name}$ stands for a similarity measure.
- $\hat{\Phi}_{name} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1\\ j\neq i}}^{M} \phi_{Name}(\mathbf{s}_i, \mathbf{s}_j)$ is the resulting stability measure using similarity measure $\phi_{Name}$.
- $\hat{p}_f = \frac{1}{d} \sum_{i=1}^{M} x_{i,f}$ is the observed frequency of occurrence of the $f_{th}$ feature in $\mathcal{A}$.

## 2 Measures

### 2.1 Similarity measures

The *Jaccard index* [4] (a modified version of the *Taminoto distance*) is defined as:

$$\phi_{Jaccard}(\mathbf{s}_i, \mathbf{s}_j) = 1 - Tanimoto(\mathbf{s}_i, \mathbf{s}_j) = \frac{|\mathbf{s}_i \cap \mathbf{s}_j|}{|\mathbf{s}_i \cup \mathbf{s}_j|} = \frac{r_{i,j}}{k_i + k_j - r_{i,j}}.$$

A similarity measure based on the relative Hamming distance [2] is given as:

$$\phi_{Hamming}(\mathbf{s}_i,\mathbf{s}_j) = 1 - \frac{|\mathbf{s}_i \setminus \mathbf{s}_j| + |\mathbf{s}_j \setminus \mathbf{s}_i|}{d} = 1 - \frac{k_i + k_j - 2r_{i,j}}{d},$$

where $|\mathbf{s}_i \setminus \mathbf{s}_j|$ is the number of features selected in $\mathbf{s}_i$ and not selected in $\mathbf{s}_j$.

The *Dice coefficient* [13]:

$$\phi_{Dice}(\mathbf{s}_i,\mathbf{s}_j) = \frac{2|\mathbf{s}_i \cap \mathbf{s}_j|}{|\mathbf{s}_i| + |\mathbf{s}_j|} = \frac{2r_{i,j}}{k_i + k_j}.$$

We can point out that the Dice coefficient is also called the $F_1$-*score* in the binary classification literature, and is used to measure the degree of agreement between the set of true labels $\mathbf{s}_1$ and the set of predicted labels $\mathbf{s}_2$.

The POG measure (*Percentage of Overlapping Genes*) [9]:

$$\phi_{POG}(\mathbf{s}_i,\mathbf{s}_j) = \frac{|\mathbf{s}_i \cap \mathbf{s}_j|}{|\mathbf{s}_i|} = \frac{r_{i,j}}{k_i}.$$

Kuncheva's similarity measure [7] (also called the *consistency index*) is defined only for feature sets $\mathbf{s}_i$ and $\mathbf{s}_j$ such that $|\mathbf{s}_i| = |\mathbf{s}_j|$ as follows:

$$\phi_{Kuncheva}(\mathbf{s}_i,\mathbf{s}_j) = \frac{r_{i,j} - \mathbb{E}_{\nabla}[r_{i,j}]}{max(r_{i,j}) - \mathbb{E}_{\nabla}[r_{i,j}]}.$$

Lustgarten's measure [8]:

$$\phi_{Lust}(\mathbf{s}_i,\mathbf{s}_j) = \frac{r_{i,j} - \mathbb{E}_{\nabla}[r_{i,j}]}{min(k_i,k_j) - \max(0, k_i + k_j - d)}.$$

The *nPOG* measure (a normalized version of the POG measure) [9]:

$$\phi_{nPOG}(\mathbf{s}_i,\mathbf{s}_j) = \frac{r_{i,j} - \mathbb{E}_{\nabla}[r_{i,j}]}{k_i - \mathbb{E}_{\nabla}[r_{i,j}]}$$

Wald's measure [12]:

$$\phi_{Wald}(\mathbf{s}_i,\mathbf{s}_j) = \frac{r_{i,j} - \mathbb{E}_{\nabla}[r_{i,j}]}{min(k_i,k_j) - \mathbb{E}_{\nabla}[r_{i,j}]}.$$

### 2.2 Stability measures

For every similarity measure $\phi_{Name}$ presented in the section above, the corresponding stability measure is taken as the average pairwise similarities between the sets in $\mathcal{A}$ as follows:

$$\hat{\Phi}_{name} = \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \phi_{Name}(\mathbf{s}_i,\mathbf{s}_j).$$

Krízek [6] defines a stability measure for feature sets of identical cardinality as follows:

$$\gamma(\mathcal{A}) = -\sum_{j=1}^{C(d,k)} \overline{G_j}\ \log \overline{G_j},$$

where $\overline{G_j}$ is the frequency of occurrence of subset $j$ in $\mathcal{A}$ over all the $C(d, k)$ possible combinations of $k$ features taken amongst $d$ features. Its minimum is 0 and its upper bound has been shown to depend on $d$ and $k$. It is only defined when $\mathcal{A}$ is made of feature sets of identical cardinality $k$. The use of such a measure implies a relatively big number feature subsets in the sequence $\mathcal{A}$, as we need frequency estimates for every possible choice of $k$ features taken out of $d$ features, which limits its use in practice. We also note that this measure is the only one for which low values correspond to high stability.

Somol's measure [11] is an improved version of two other measures (the Consistency Measure $C$ and the Weighted Consistency Measure $CW$ [11]) that we omitted for simplicity. This measure is constructed so that it is not *subset-size biased* and is defined for any feature set size:

$$CW_{rel}(\mathcal{A}) = \frac{d\left[\sum_{i=1}^{M} k_i - D + \sum_{f=1}^{d} M\hat{p}_f(M\hat{p}_f - 1)\right] - \left(\sum_{i=1}^{M} k_i\right)^2 + D^2}{d\left[H^2 + M\left(\sum_{i=1}^{M} k_i - H\right) - D\right] - \left(\sum_{i=1}^{M} k_i\right)^2 + D^2},$$

where $D = \left(\sum_{i=1}^{M} k_i\right) mod\ d$ and $H = \left(\sum_{i=1}^{M} k_i\right) mod\ M$ and $\hat{p}_f$ is the frequency of occurrence of the $f^{th}$ feature in the $M$ feature sets.

## 3 Proof of Properties

### 3.1 Fully defined

This property directly follows from the definitions of the stability measures.

### 3.2 Bounds

The bounds of most similarity measures (and therefore stability measures) can be easily found in the literature. Jaccard, Hamming, Dice, $POG$ and $CW_{rel}$ are all in the interval $[0, 1]$ [1], [11]. Kuncheva and Pearson measures are in the interval $[-1, 1]$ ([3], [7]). Lustgarten's similarity measure can also easily be shown to be in $[-1, 1]$.

Contrarily to some misconceptions in the literature, $nPOG$ and Wald's similarity measures have a minimum of $1 - n$ and a maximum of 1 therefore not verifying the property of bounds. Krízek's stability measure is shown to take values in the interval $[0, \log(\min(M, C(d, k)))]$ [5].

## 4 Maximum

### Deterministic Selection → Maximum Stability

Let us assume that all the feature sets in $\mathcal{A}$ are identical of cardinality $k$, therefore $r_{i,j} = k$ and by definition, we have that $\phi_{Jaccard} = \phi_{Hamming} = \phi_{Dice} = \phi_{POG} = \phi_{Kuncheva} = \phi_{Wald} = 1$.

In that case, $\phi_{Lust} = \frac{k-\frac{k^2}{d}}{k-max(0,2k-d)}$, meaning that its maximum depends on the values of $k$ and $n$ as shown by Figure 1 of the paper.

### Maximum Stability → Deterministic Selection

Showing that Lustgarten and Wald's stability measure do not have this property can easily be done with a counter-example as done in the paper (c.f. Figure 2).

Let us show that the property is true for Krízek $\gamma$. The maximum stability for that measure corresponds to a value of 0.

$$\begin{aligned}
&\gamma(\mathcal{A}) = 0\\
&\Rightarrow -\sum_{j=1}^{C(d,k)} \overline{G_j}\ \log \overline{G_j} = 0\\
&\Rightarrow \forall j \in \{1,...,C(d,k)\}, \overline{G_j}\ \log \overline{G_j} = 0 \ \text{ since all elements of the sum are negative,}\\
&\Rightarrow \forall j \in \{1,...,C(d,k)\}, \overline{G_j} = 0 \ \text{ or } \ \overline{G_j} = 1\\
&\Rightarrow \text{All feature sets in } \mathcal{A} \text{ are identical.}
\end{aligned}$$

All other measures have a maximum of stability of 1. We therefore assume that $\hat{\Phi}(\mathcal{A}) = max(\hat{\Phi}) = 1$ and we want to show that this implies that all feature sets in $\mathcal{A}$ are identical.

$$\begin{aligned}
&\hat{\Phi}(\mathcal{A}) = 1\\
&\Rightarrow \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1\\ j\neq i}}^{M} \phi(\mathbf{s}_i, \mathbf{s}_j) = 1\\
&\Rightarrow \sum_{i=1}^{M} \sum_{\substack{j=1\\ j\neq i}}^{M} \phi(\mathbf{s}_i, \mathbf{s}_j) = M(M-1)\\
&\Rightarrow \forall i \in \{0,1\}^d, \forall j \in \{0,1\}^d, j \neq i, \phi(\mathbf{s}_i, \mathbf{s}_j) = 1.
\end{aligned}$$

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

### 4.1 Correction for Chance

Using the linearity of the expected value and the definitions given in section 2, we get that: $\mathbb{E}_{\nabla}[\phi_{Kuncheva}] = \mathbb{E}_{\nabla}[\phi_{Lust}] = \mathbb{E}_{\nabla}[\phi_{nPOG}] = \mathbb{E}_{\nabla}[\phi_{Wald}] = 0$ and therefore the stability measures using these similarity measures also have an expected value of 0 and have the property of correction for chance. $CW_{rel}$ is by construction *subset-size unbiased* and is shown empirically to hold the property of correction for chance [11].

When the FS procedure is randomly selecting feature sets of cardinality $k$, the expected value of the frequency of occurrence of a feature set is equal to $\frac{1}{C(d,k)}$. Therefore Krizek's stability measure is not corrected by chance as its expected value when the FS is random will depend on $k$ and $d$.

Using the fact that $\mathbb{E}_{\nabla}[r_{i,j}] = \frac{k_i k_j}{d}$ and the linearity of the expected value, we can see that the other similarity measures will have an expected value depending on $k_i$, $k_j$ and $d$ and therefore do not have the property of correction for chance.

## 5 Proofs of theorems

**Theorem 1.** *For all $(i,j) \in \{1,...,M\}^2$, the sample Pearson's coefficient can be re-written:*

$$\phi_{Pearson}(\mathbf{s}_i, \mathbf{s}_j) = \frac{r_{i,j} - \mathbb{E}_{\nabla}[r_{i,j}]}{d\ \upsilon_i \upsilon_j} = \frac{r_{i,j} - \frac{k_i k_j}{d}}{d\ \upsilon_i \upsilon_j},$$

*where $\forall i \in \{1,...,M\}, \upsilon_i = \sqrt{\frac{k_i}{d}(1-\frac{k_i}{d})}$. Therefore it possesses the property of correction for chance.*

**Proof.**

We remind that the sample Pearson's correlation coefficient between two feature sets $\mathbf{s}_i$ and $\mathbf{s}_j$ is by definition:

$$\phi_{Pearson}(\mathbf{s}_i, \mathbf{s}_j) = \frac{\frac{1}{d}\sum_{f=1}^{d}(x_{i,f} - \bar{x}_{i,.})(x_{j,f} - \bar{x}_{j,.})}{\sqrt{\frac{1}{d}\sum_{f=1}^{d}(x_{i,f} - \bar{x}_{i,.})^2}\sqrt{\frac{1}{d}\sum_{f=1}^{d}(x_{j,f} - \bar{x}_{j,.})^2}}, \quad (1)$$

where $\forall i \in \{1,...,M\}, \bar{x}_{i,.} = \frac{1}{d}\sum_{f=1}^{d} x_{i,f} = \frac{k_i}{d}$ since there are $k_i$ features selected in $\mathbf{s}_i$.
Let us calculate the denominator term:

$$\begin{aligned}\frac{1}{d}\sum_{f=1}^{d}(x_{i,f} - \bar{x}_{i,.})^2 &= \frac{1}{d}\sum_{f=1}^{d}(x_{i,f}^2 - 2\bar{x}_{i,.}x_{i,f} + \bar{x}_{i,.}^2) \\ &= \left(\frac{1}{d}\sum_{f=1}^{d} x_{i,f}^2\right) - \frac{2}{d}\bar{x}_{i,.}\sum_{f=1}^{d} x_{i,f} + \frac{1}{d}d\bar{x}_{i,.}^2.\end{aligned}$$

As $x_{i,f}$ is binary (equal to 0 or 1), we have that $(x_{i,f})^2 = x_{i,f}$. Therefore:

$$\begin{aligned}
\frac{1}{d}\sum_{f=1}^{d}(x_{i,f} - \bar{x}_{i,f})^2 &= \left(\frac{1}{d}\sum_{f=1}^{d} x_{i,f}\right) - \frac{2}{d}\bar{x}_{i,.}k_i + \bar{x}_{i,.}^2 \\
&= \bar{x}_{i,.} - 2\bar{x}_{i,.}^2 + \bar{x}_{i,.}^2 \\
&= \bar{x}_{i,.}(1 - \bar{x}_{i,.}) \\
&= \frac{k_i}{d}\left(1 - \frac{k_i}{d}\right) = \upsilon_i^2.
\end{aligned}$$

Similarly, $\upsilon_j^2 = \frac{1}{d}\sum_{f=1}^{d}(x_{j,f} - \bar{x}_{j,.})^2$. Replacing in Equation 1, we get that:

$$\begin{aligned}
\phi_{Pearson}(\mathbf{s}_i, \mathbf{s}_j) &= \frac{1}{d\upsilon_i\upsilon_j}\sum_{f=1}^{d}(x_{i,f} - \bar{x}_{i,.})(x_{j,f} - \bar{x}_{j,.}) \\
&= \frac{1}{d\upsilon_i\upsilon_j}\sum_{f=1}^{d}(x_{i,f}x_{j,f} - \bar{x}_{j,.}x_{i,f} - \bar{x}_{i,.}x_{j,f} + \bar{x}_{i,.}\bar{x}_{j,.}) \\
&= \left(\frac{1}{d\upsilon_i\upsilon_j}\sum_{f=1}^{d} x_{i,f}x_{j,f}\right) - \frac{\bar{x}_{i,.}\bar{x}_{j,.}}{\upsilon_i\upsilon_j} - \frac{\bar{x}_{i,.}\bar{x}_{j,.}}{\upsilon_i\upsilon_j} + \frac{\bar{x}_{i,.}\bar{x}_{j,.}}{\upsilon_i\upsilon_j} \\
&= \left(\frac{1}{d\upsilon_i\upsilon_j}\sum_{f=1}^{d} x_{i,f}x_{j,f}\right) - \frac{\bar{x}_{i,.}\bar{x}_{j,.}}{\upsilon_i\upsilon_j}.
\end{aligned}$$

As, $x_{i,f}x_{j,f}$ will only be equal to 1 when both $x_{i,f}$ and $x_{j,f}$ are equal to 1, we have that $\sum_{f=1}^{d} x_{i,f}x_{j,f} = |\mathbf{s}_i \cap \mathbf{s}_j| = r_{i,j}$. Therefore:

$$\phi_{Pearson}(\mathbf{s}_i, \mathbf{s}_j) = \frac{r_{i,j}}{d\upsilon_i\upsilon_j} - \frac{\bar{x}_{i,.}\bar{x}_{j,.}}{\upsilon_i\upsilon_j} = \frac{r_{i,j} - d\bar{x}_{i,.}\bar{x}_{j,.}}{d\upsilon_i\upsilon_j} = \frac{r_{i,j} - \frac{k_ik_j}{d}}{d\upsilon_i\upsilon_j} = \frac{r_{i,j} - \mathbb{E}_{\nabla}[r_{i,j}]}{d\upsilon_i\upsilon_j}.$$

**Theorem 2.** *When $k$ is constant, the stability using Pearson's correlation is equal to some other measures, that is:*

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

**Proof.**
Straightforward using the definition of the measures given in section 2.1 and Theorem 1 for $k_i = k_j = k$. Indeed the similarity measures of Kuncheva, Wald, nPOG and Pearson will all be equal to $\frac{r_{i,j} - \frac{k^2}{d}}{k - \frac{k^2}{d}}$. Therefore the stability measures using these similarity measures are all equal.

**Theorem 3.** *The stability estimate $\hat{\Phi}_{Pearson}$ is asymptotically in the interval $[0, 1]$ as $M$ approaches infinity.*

**Proof.**
The upper bound is trivial: $\phi_{Pearson} \leq 1 \Rightarrow \hat{\Phi}_{Pearson} \leq 1$.
We prove the lower bound by showing that :

$$\hat{\Phi}_{Pearson} = \frac{1}{M(M-1)}\frac{1}{d^2}\underbrace{\sum_{f<f'}\left[\sum_{i=1}^{M}\frac{x_{i,f}-x_{i,f'}}{\upsilon_i}\right]^2}_{\geq 0} - \underbrace{\frac{1}{M-1}}_{\underset{M\to+\infty}{\to}0}$$

which gives us that $\underset{M\to+\infty}{lim}[\hat{\Phi}_{Pearson}] \geq 0$. Indeed, we have:

$$\frac{1}{M(M-1)}\frac{1}{d^2}\sum_{f<f'}\left[\sum_{i=1}^{M}\frac{x_{i,f}-x_{i,f'}}{\upsilon_i}\right]^2$$

$$=\frac{1}{M(M-1)}\frac{1}{d^2}\sum_{f<f'}\left[\sum_{i=1}^{M}\frac{x_{i,f}}{\upsilon_i}-\sum_{i=1}^{M}\frac{x_{i,f'}}{\upsilon_i}\right]^2$$

$$=\frac{1}{M(M-1)}\left[\frac{1}{d}\sum_{f=1}^{d}\left(\sum_{i=1}^{M}\frac{x_{i,f}}{\upsilon_i}\right)^2-\left(\frac{1}{d}\sum_{f=1}^{d}\sum_{i=1}^{M}\frac{x_{i,f}}{\upsilon_i}\right)^2\right]$$

$$=\frac{1}{M(M-1)}\frac{1}{d}\sum_{f=1}^{d}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{x_{i,f}x_{j,f}}{\upsilon_i\upsilon_j}-\frac{1}{M(M-1)}\left(\frac{1}{d}\sum_{i=1}^{M}\frac{1}{\upsilon_i}\left(\sum_{f=1}^{d}x_{i,f}\right)\right)^2$$

$$=\frac{1}{M(M-1)}\frac{1}{d}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{\sum_{f=1}^{d}x_{i,f}x_{j,f}}{\upsilon_i\upsilon_j}-\frac{1}{M(M-1)}\left(\frac{1}{d}\sum_{i=1}^{M}\frac{k_i}{\upsilon_i}\right)^2$$

$$=\frac{1}{M(M-1)}\frac{1}{d}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{r_{i,j}}{\upsilon_i\upsilon_j}-\frac{1}{M(M-1)}\frac{1}{d^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\frac{k_ik_j}{\upsilon_i\upsilon_j}$$

$$=\frac{1}{M(M-1)}\left[\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\frac{r_{i,j}-\frac{k_ik_j}{d}}{d\upsilon_i\upsilon_j}\right]+\frac{1}{M(M-1)}\sum_{i=1}^{M}\frac{r_{i,i}}{d\upsilon_i^2}-\frac{1}{M(M-1)}\frac{1}{d^2}\sum_{i=1}^{M}\frac{k_i^2}{\upsilon_i^2}$$

$$=\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\left(\frac{r_{i,j}-\frac{k_ik_j}{d}}{d\upsilon_i\upsilon_j}\right)+\frac{1}{M(M-1)}\sum_{i=1}^{M}\frac{1}{\upsilon_i^2}\frac{k_i}{d}-\frac{1}{M(M-1)}\sum_{i=1}^{M}\frac{1}{\upsilon_i^2}\frac{k_i^2}{d^2}$$

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

$$=\hat{\Phi}_{Pearson}+\frac{1}{M(M-1)}\sum_{i=1}^{M}\frac{1}{\upsilon_i^2}\left(\frac{k_i}{d}-\frac{k_i^2}{d^2}\right)$$

$$=\hat{\Phi}_{Pearson}+\frac{1}{M-1}$$

**Proof of Equation (3).**

Let $\widehat{Var}(X_f) = \frac{M}{M-1}\hat{p}_f(1-\hat{p}_f)$ be the unbiased sample variance of the variable $X_f$. We want to show that when the cardinality of the feature sets is constant and equal to $k$, we have that:

$$\hat{\Phi}_{Pearson} = 1 - \frac{S}{S_{max}},$$

where $S = \frac{1}{d}\sum_{f=1}^{d}\widehat{Var}(X_f)$ and where $S_{max} = \frac{k}{d}\left(1-\frac{k}{d}\right)$ the maximal value of $S$ given that the FS procedure is selecting $k$ features per feature set. Indeed, we have that:

$$\begin{aligned}
1 - \frac{S}{S_{max}} &= 1 - \frac{\frac{M}{M-1}\frac{1}{d}\sum_{f=1}^{d}\hat{p}_f(1-\hat{p}_f)}{S_{max}} \\
&= 1 - \frac{1}{S_{max}}\frac{M}{M-1}\frac{1}{d}\left[\sum_{f=1}^{d}\hat{p}_f - \sum_{f=1}^{d}(\hat{p}_f)^2\right] \\
&= 1 - \frac{1}{S_{max}}\frac{M}{M-1}\frac{1}{d}\left[k - \sum_{f=1}^{d}\left(\frac{1}{M}\sum_{i=1}^{M}x_{i,f}\right)^2\right] \\
&= 1 - \frac{1}{S_{max}}\frac{M}{M-1}\frac{1}{d}\left[k - \frac{1}{M^2}\sum_{f=1}^{d}\sum_{i=1}^{M}\sum_{j=1}^{M}x_{i,f}x_{j,f}\right] \\
&= 1 - \frac{1}{S_{max}}\frac{M}{M-1}\frac{1}{d}\left[k - \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}\underbrace{\left(\sum_{f=1}^{d}x_{i,f}x_{j,f}\right)}_{r_{i,j}}\right] \\
&= 1 - \frac{1}{S_{max}}\frac{M}{M-1}\frac{1}{d}\left[k - \frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M}r_{i,j}\right] \\
&= 1 - \frac{1}{S_{max}}\frac{M}{M-1}\frac{1}{d}\left[k - \frac{1}{M^2}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}r_{i,j} - \frac{1}{M^2}\sum_{i=1}^{M}r_{i,i}\right] \\
&= 1 - \frac{1}{S_{max}}\frac{M}{M-1}\frac{1}{d}\left[k - \frac{1}{M^2}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}r_{i,j} - \frac{1}{M^2}\sum_{i=1}^{M}k\right] \\
&= 1 - \frac{1}{S_{max}}\frac{M}{M-1}\frac{1}{d}\left[k - \frac{1}{M^2}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}r_{i,j} - \frac{k}{M}\right]
\end{aligned}$$

$$
\begin{aligned}
&= 1 + \frac{1}{S_{max}} \frac{1}{M(M-1)} \frac{1}{d} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} r_{i,j} + \frac{1}{S_{max}} \frac{M}{M-1} \frac{1}{d} \left[ \frac{k}{M} - k \right] \\
&= 1 + \frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \phi_{Pearson} + \frac{1}{S_{max}} \frac{1}{M(M-1)} \frac{1}{d} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} \frac{k^2}{d} - \frac{1}{S_{max}} \frac{k}{d} \\
&= 1 + \hat{\Phi}_{Pearson} + \frac{1}{S_{max}} \frac{k^2}{d^2} - \frac{1}{S_{max}} \frac{k}{d} \\
&= 1 + \hat{\Phi}_{Pearson} - \frac{\frac{k}{d}\left(1 - \frac{k}{d}\right)}{S_{max}} \\
&= \hat{\Phi}_{Pearson}
\end{aligned}
$$

## References

1. Alelyani, S.: On Feature Selection Stability: A Data Perspective. Ph.D. thesis (2013)
2. Dunne, K., Cunningham, P., Azuaje, F.: Solutions to instability problems with sequential wrapper-based approaches to feature selection. Tech. rep., Journal of Machine Learning Research (2002)
3. Edmundson, H.P.: A correlation coefficient for attributes or events. In: Proc. Statistical association methods for mechanized documentation (1966)
4. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. (2007)
5. Křížek, P.: Feature selection: Stability, algorithms, and evaluation. Ph.D. thesis, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic (2008)
6. Krízek, P., Kittler, J., Hlavác, V.: Improving stability of feature selection methods. In: CAIP (2007)
7. Kuncheva, L.I.: A stability index for feature selection. In: Artificial Intelligence and Applications (2007)
8. Lustgarten, J.L., Gopalakrishnan, V., Visweswaran, S.: Measuring stability of feature selection in biomedical datasets. AMIA Annu Symp Proc (2009)
9. MAQC consortium: The MicroArray quality control project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotech. (2006)
10. Nogueira, S., Brown, G.: Measuring the stability of feature selection. In: Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg (2016)
11. Somol, P., Novovičová, J.: Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
12. Wald, R., Khoshgoftaar, T.M., Napolitano, A.: Stability of filter- and wrapper-based feature subset selection. In: International Conference on Tools with Artificial Intelligence. IEEE Computer Society (2013)
13. Yu, L., Ding, C.H.Q., Loscalzo, S.: Stable feature selection via dense feature groups. In: KDD (2008)