Scoring functions to evaluate the rankings methods for variable selection

M. Marinescu†, G. Villacrés†, L. Martino⋆, Ó. Barquero†
† Universidad Rey Juan Carlos, Madrid, Spain.
⋆ Università di Catania, Italy.

Abstract—**Feature selection is a relevant and hot topic in signal processing and machine learning, and it has gained even more relevance in recent years. There exist many possibilities to measure the importance of a variable for a specific task. Given a measure of importance, we can obtain a ranking of the input variables involved in a regression or classification problem. As a consequence, we also need the ability to identify the best ranking method by analyzing the results for a given task or specific data. In this work, we describe and discuss several scoring functions designed for evaluating the ranking methods. We test the scoring functions in a controlled experiment with synthetic data.**

Index Terms—**Feature selection; feature importance; Shapley value; ranking methods; scoring functions.**

I. Introduction

Variable selection, also known as feature selection [1]–[3],[1] is one of the most relevant topics in signal processing, statistics, and machine learning. This topic has received renewed interest in the last few years. More specifically, the way of defining a *feature importance measure* has become a hot research topic nowadays [4]–[7]. The interest in the so-called Shapley values is a clear example [8].

There are many ways of defining a feature's importance in both regression and classification problems [1], [9]. Given an importance measure, we can build a ranking of the involved variables, from the most important to the least important. Theoretically speaking, the complete variable selection problem is formed by two parts: firstly, ranking the variables and secondly deciding the effective number of variables (see, e.g., [7], [10], [11] for the second part). Clearly, by changing the definition of the feature importance measure, we can obtain a different ranking. From a research point of view, it is essential to find the optimal ranking method (RM) for at least a specific task and/or data type. For this goal, we need the ability to compare RMs where a ground-truth is available (i.e., in experiments with simulated data, for instance).

This work is devoted to describing and discussing different possible scoring functions to "judge" the performance of different RMs. Namely, the final goal is *to rank the ranking methods*. We start with the simplest scoring functions and increase the complexity [12]–[14]. We also describe the benefits and drawbacks of each scoring function. We also introduce normalized versions of the scores to allow for the comparison among different scoring functions. A simple running example is used to facilitate the understanding of the interested reader. We finally test all of them in a synthetic regression scenario considering several alternative RMs [1].

II. Framework

Suppose that we have a set of R variables $\mathbf{x} = [x_1, ..., x_R]^\top$ (input vector) that describes the behavior of a related variable y (output). We assume that we have a dataset of N data pairs, $\{\mathbf{x}_n, y_n\}_{n=1}^N$, and we can define a ground-truth ranking of the input features (i.e., the components of $\mathbf{x}$) in decreasing order of importance,

$$\textbf{Ground-truth: } \mathcal{G} = \{g_1, g_2, ..., g_R\}, \tag{1}$$

where $g_j \in \{1, ..., R\}$ with $g_i \neq g_j$ for $i \neq j$, is the sub-index associated to the variable x_{g_j} and j is the correct position of the ranking of the variable x_{g_j}. As an example, with $R = 10$, if $g_1 = 5$ and $g_{10} = 2$ it means that x_5 is the most important variable and x_2 is the worst variable in terms of importance.

Generally, we have several RMs that we can apply for feature selection. Each one is implicitly or explicitly based on a feature importance measure and provides a ranking of the variables. We want to score these results according to the groud-truth (when it is available, as in experiments with artificial data). Namely, the goal is *to rank the ranking methods*, e.g., discovering the best and the worst RMs for feature selection, at least in one specific application. More specifically, a ranking technique yields a ranking of the features in decreasing order of importance that we denote as

$$\textbf{Ranking: } \mathcal{R} = \{k_1, k_2, \dots, k_R\}, \tag{2}$$

where $k_i \in \{1, \dots, R\}$ ($k_i \neq k_j$ for $i \neq j$), indicates the sub-index associated to the variable x_{k_i} and i is the position of x_{k_i} in the resulting ranking. We desire to "score" this ranking according to the ground-truth.

Running example. Before starting with the description of the possible scoring functions, we introduce an example that we will use throughout the rest of the paper. Considering $R = 5$ features, and a ranking with subindices:

$$\textit{Example – Ground-truth: } \mathcal{G}_E = \{3, 1, 2, 5, 4\}. \tag{3}$$

This work has been partially supported by the "PIAno di inCEntivi per la RIcerca di Ateneo 2024/2026" (UPB 28722052159) and PIACERI Starting Grant BA-GRAPH (UPB 28722052144) of the University of Catania, by Comunidad de Madrid within the 2023-2026 agreement with Universidad Rey Juan Carlos for the granting of direct subsidies for the promotion, encouragement of research and technology transfer, line of Action A Emerging Doctors, under Project OrdeNGN (Ref. F1177), and by project POLI-GRAPH - Grant PID2022-136887NB-I00 funded by MCIN/AEI/10.13039/501100011033.

[1] In this work, we use the terms "variable" and "feature" as synonymous.

Namely, the variable x_3 is the most important, whereas x_4 is the least important. Moreover, in this example, we assume that a ranking scheme provides as a result the following ranking,

$$\textit{Example – Ranking: } \mathcal{R}_E = \{3, 1, 5, 4, 2\}. \tag{4}$$

We can observe that the first two variables are correctly ranked, whereas the last three variables have been wrongly positioned by the employed ranking technique.

III. Scoring Functions

In this section, we present different possible scoring functions starting from the simplest ones in terms of complexity.

A. Baseline scores

1) Match counting: We simply count the number of correct elements in the ranking. Let us define a binary variable

$$I_j = \begin{cases} 1 & \text{if} \quad k_j = g_j, \\ 0 & \text{if} \quad k_j \neq g_j. \end{cases} \tag{5}$$

Then, the final score is defined as

$$S = \sum_{j=1}^{R} I_j. \tag{6}$$

In the case of $\mathcal{G}_E$ and $\mathcal{R}_E$, we have $S = 2$. We can normalize this score by dividing by R, i.e., $0 \leq \frac{S}{R} \leq 1$. We define the normalized score as $\bar{S} = \frac{S}{R}$.

2) Permutation distance: This scoring function is defined as the (minimum) number of permutations one should realize starting from $\mathcal{R}$ until obtaining $\mathcal{G}$. Let P be the number of permutations required. Then, the score is defined as

$$S = (R - 1) - P. \tag{7}$$

This measure goes from $R - 1$ (perfect matching) to the worst case scenario, which corresponds to 0. The normalized score is, $\bar{S} = 1 - \frac{S}{R-1}$. In the case of the example $\mathcal{R}_E$ and $\mathcal{G}_E$, we have: $S = 2$ and $\bar{S} = 0.5$.

B. Distance scores

1) Distance summing: The previous scores do not take into account the distance between the right and wrong positions, and any errors are penalized by 1. The idea is to perform the following steps:

- For $j = 1, ..., R$:

1) Given j, find in $\mathcal{R}$ the position i^* such that $k_{i^*} = g_j$.
2) Compute the distance

$$d_j = |i^* - j|. \tag{8}$$

- Finally, compute the average

$$D = \frac{1}{R} \sum_{j=1}^{R} d_j. \tag{9}$$

We want a score such that the higher its value, the better the RM is. We can achieve this by computing

$$S = D^{\max} - D, \tag{10}$$

where

$$\begin{aligned} D^{\max} &= \frac{1}{R} \sum_{j=1}^{R} d_j^{\max}, \\ &= \frac{1}{R} \sum_{j=1}^{R} |R - 2j + 1|, \\ &= \frac{1}{R} \sum_{j=1}^{\lfloor R/2 \rfloor} 2(R - 2j + 1) = \begin{cases} \frac{R}{2} & R \text{ even} \\ \frac{1}{R}\frac{R^2-1}{2} & R \text{ odd}, \end{cases} \end{aligned} \tag{11}$$

where we have used $d_j^{\max} = |R - 2j + 1|$. Note that $D^{\max}$ corresponds to the worst-case scenario when the model features are ordered the other way around in comparison to the ground-truth. Finally, the normalized score is defined as

$$\bar{S} = \frac{S}{D^{\max}} = 1 - \frac{D}{D^{\max}}. \tag{12}$$

In the case of the example with $\mathcal{G}_E$ and $\mathcal{R}_E$ we have: $S = 1.6$ and $\bar{S} = 1 - \frac{4}{4+2+0+2+4} \approx 0.666$.

2) Generalized weighted distance: The previous score does not take into account the importance of each feature, and we could also change the type of distance. A generalized distance is considered, and to penalize more the errors in first positions, we may assign some weights, $\bar{w}_1, \ldots, \bar{w}_R$, such that $\sum \bar{w}_i = 1$. The resulting score is then

$$S = \left(\sum_{j=1}^{R} \bar{w}_j (d_j^{\max})^\alpha \right)^{1/\alpha} - \left(\sum_{j=1}^{R} \bar{w}_j d_j^\alpha \right)^{1/\alpha}, \tag{13}$$

where $\alpha > 0$. The normalized score is defined in the same way as Eq. (12). To penalize more the errors in the first positions of the ranking, we can assign weights satisfying $\bar{w}_1 > \bar{w}_2 > \ldots > \bar{w}_R$. For instance:

- Rational decay: consider $b = \frac{R(R+1)}{2}$, the sum of $1 + 2 + \ldots + R$. Then we can assign the weights $\bar{w}_1 = \frac{R}{b}, \bar{w}_2 = \frac{R-1}{b}, \ldots, \bar{w}_R = \frac{1}{b}$. That is, each weight is proportional to the position of the feature in the ground truth. This gives primal importance to the first position in the ranking.
- Exponential decay: assign weights $\bar{w}_j \propto \exp\{-j\lambda\}$, $j = 1, \ldots, R$, for a chosen decay rate λ. This exponential decay results in failures in the last position, typically residual.

A drawback is that the choice of these weights is subjective.

C. Correlation based methods

A way to measure the association between ordinal datasets is to study the correlation between their positions. In our problem, this means studying the association between the positions of each feature in the ground-truth set and in the ranking set.

We denote the position of the feature i in the ranking set as $r_{\mathcal{R}}(k_i) = i$ (is i by definition) and the position of the feature i in the ground-truth set as $r_{\mathcal{G}}(k_i)$. For the sake of illustration, Table I shows the positions of the features in the running example.

TABLE I
POSITION OF EACH FEATURE IN THE RUNNING EXAMPLE

Feature	Ground-truth position $r_{\mathcal{G}}(k_i)$	Method - position $r_{\mathcal{R}}(k_i) = i$
$k_1 = 3$	1	1
$k_2 = 1$	2	2
$k_3 = 5$	4	3
$k_4 = 4$	5	4
$k_5 = 2$	3	5

Given the positions of the features, we can use some correlation measures proposed in the literature as scoring functions [12], [13]. In the following, we describe the Spearman and Kendall correlations.

1) Spearman's correlation: it is the classical Pearson correlation coefficient applied to the positions. Consider the points $(r_{\mathcal{R}}(k_i), r_{\mathcal{G}}(k_i))$, $i = 1, ..., R$ (note that $r_{\mathcal{R}}(k_i) = i$). Then, the Spearman's correlation is the Pearson correlation coefficient over these positions, i.e.,

$$\bar{S} = \frac{\text{cov}[r_{\mathcal{R}}, r_{\mathcal{G}}]}{\sigma_{r_{\mathcal{R}}} \sigma_{r_{\mathcal{G}}}}, \tag{14}$$

where cov denotes the covariance of the data $r_{\mathcal{R}}, r_{\mathcal{G}}$ and $\sigma_{r_{\mathcal{R}}}$ $\sigma_{r_{\mathcal{G}}}$, their standard deviations. Note that $-1 \leq \bar{S} \leq 1$. It can be easily shown that in our running example:

$$\bar{S} = 0.7.$$

For the sake of illustration, we have plotted the pairs $(r_{\mathcal{R}}(k_i), r_{\mathcal{G}}(k_i))$, $i = 1, ..., R$ in Figure 1. A clear positive association between the positions can be observed, resembling the obtained result $\bar{S} = 0.7$. Generally, a value of 1 means a totally correct model, a value of -1 is an incorrect model with variables positioned the other way around, and a value of 0 means a nominal random association. In case a RM has very low performance, being worse than a random ranking assignment, we can assign any negative values of Spearman correlation to 0, in order to avoid negative scores.

2) Kendall correlation: this method computes the so-called Kendall correlation, which measures the correlation by computing the number of concordant pairs. Two pairs of observations $(r_{\mathcal{R}}(k_i), r_{\mathcal{R}}(k_j))$ and $(r_{\mathcal{G}}(k_i), r_{\mathcal{G}}(k_j))$ are *concordant* if either $r_{\mathcal{R}}(k_i) > r_{\mathcal{G}}(k_i)$ and $r_{\mathcal{R}}(k_j) > r_{\mathcal{G}}(k_j)$ both holds simultaneously, or $r_{\mathcal{R}}(k_i) < r_{\mathcal{G}}(k_i)$ and $r_{\mathcal{R}}(k_j) < r_{\mathcal{G}}(k_j)$ holds jointly. Specifically, this correlation computes the difference between the number of concordant pairs and the ones that are not, normalized by the number of total pairs $\binom{R}{2}$:

$$\bar{S} = \frac{\text{n}^{\text{o}} \text{ concordant pairs - n}^{\text{o}} \text{ discordant pairs}}{\binom{R}{2}}.$$

In Table II, we show the calculation of the Kendall correlation for the running example, giving the result of $\bar{S} = 0.6$.

Since the maximum number of concordant pairs is $\binom{R}{2}$ and the same holds for the number of discordant pairs, the Kendall correlation is also bounded in the interval $[-1, 1]$. Both Kendall and Spearman correlations are special cases of a more general concept called *generalized rank correlations* [12], which is a family of statistics that measures ordinal association between variables.

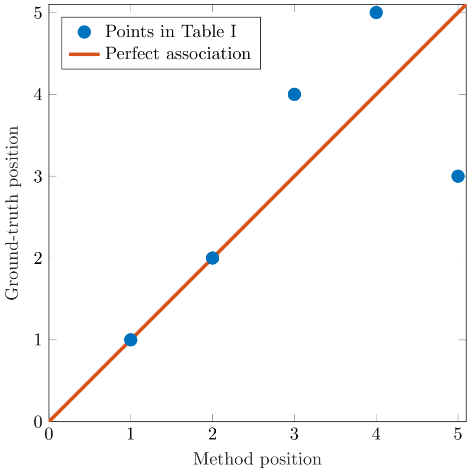

Fig. 1. Association between ground-truth and running example. x-axis represents the position of features in the method, whereas y-axis is the position of features in the ground-truth. The line represents a perfect association, whereas the set of points $(r_{\mathcal{R}}(k_i), r_{\mathcal{G}}(k_i))$, $i = 1, ..., R$, are the pairs in Table I, i.e., $(1, 1)$, $(2, 2)$, $(3, 4)$, $(4, 5)$ and $(5, 3)$.

TABLE II
COMPUTATION OF KENDALL CORRELATION FOR THE RUNNING EXAMPLE.

Index	**Pairs** $(r_{\mathcal{R}}(k_i), r_{\mathcal{R}}(k_j)); (r_{\mathcal{G}}(k_i), r_{\mathcal{G}}(k_j))$	**Concordant**
1	(1, 2), (1, 2)	Yes
2	(1, 3), (1, 4)	Yes
3	(1, 4), (1, 5)	Yes
4	(1, 5), (1, 3)	Yes
5	(2, 3), (2, 4)	Yes
6	(2, 4), (2, 5)	Yes
7	(2, 5), (2, 3)	Yes
8	(3, 4), (4, 5)	Yes
9	(3, 5), (4, 3)	No
10	(4, 5), (5, 3)	No
$\bar{S}$	$\frac{(8-2)}{10}$	**0.6**

IV. HANDLING POSSIBLE TIES

Ground-truth may present ties among variables. That is, there exists at least one subset of features where any arrangement of them within a set of specific positions is valid/correct. As an example, we may have five features where the third and fourth elements are of equal importance. For instance, we could have as a ground truth $\{g_1 = 5, g_2 = 4, g_{3:4} = [1, 2], g_5 = 3\}$. In this example, we can interpret that we have two possible sequences of ground-truth: $\{g_1 = 5, g_2 = 4, g_3 = 1, g_4 = 2, g_5 = 3\}$ or $\{g_1 = 5, g_2 = 4, g_3 = 2, g_4 = 1, g_5 = 3\}$. Hence, a possible way to deal with this situation and to be able to apply the scoring methods of the previous sections is to rearrange (according to the possible ties) the ground-truth, in such a way as to be closer to the ranking to evaluate. Namely, we permute within the position of the ties to find the ground-truth sequence that is the closest (in a distance) to the ranking that we need to evaluate. After finding the closest ground-truth sequence, all the scoring functions can be directly applied.

V. Experiments

A. Ranking methods for feature selection

To evaluate the different score methods described in Section III, we utilize the RMs based on wrapper methods [1], [8], [15]. Specifically, we use the following RMs: 1) leave-one-covariate-out, called LOCO in the literature (RM0) [8], 2) forward selection adding variables "forward" minimizing an external cost (RM1), 3) backward elimination removing variables "backward" minimizing an external cost (RM2), 4) backward elimination removing the best variable "backward" maximizing an external cost (RM3), and 5) forward selection adding the worst variable, maximizing an external cost (RM4). The last four RMs are described in [1, Sec. III A].

B. Synthetic dataset generation

We create a synthetic dataset to rank variables under controlled conditions, using the RMs described in Section V-A. The RMs are then assessed, knowing the ground-truth, using the different scores. The dataset is structured according to a linear model, with variables selectively included and excluded based on specific criteria. It contains $N = 5000$ observations and $R = 20$ variables, represented as $\mathbf{x} = [x_1, \ldots, x_{20}]$. The details of these variables are provided in Table IV.

TABLE IV
FEATURE GENERATION: SAMPLING FROM A DISTRIBUTION

Variables	Generation / Distribution
x_1, x_2, x_5 x_7, x_{15}, x_{16}, x_{18}, x_{19}	$\mathcal{N}(0,1)$
x_3, x_4, x_8 x_9, x_{10}, x_{13}, x_{20}	$\mathcal{U}\left(\left[-\frac{\sqrt{12}}{2}, \frac{\sqrt{12}}{2}\right]\right)$
x_6	x_2^2
x_{11}	$z = 0.5x_8 + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1)$, $\frac{z - \text{mean}(z)}{\text{std}(z)}$
x_{12}	$z = 0.5x_{10} + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1)$, $\frac{z - \text{mean}(z)}{\text{std}(z)}$
x_{14}	$z = x_5 + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1)$, $\frac{z - \text{mean}(z)}{\text{std}(z)}$
x_{17}	$z = 0.2x_2 + u, \quad u \sim \mathcal{U}([0,1])$, $\frac{z - \text{mean}(z)}{\text{std}(z)}$

True model: The corresponding observations were generated as follows

$$\begin{aligned} y_n = & 0.6x_2 + 0.6x_3 - 0.2x_4 + 0.1x_5 - 0.3x_7 + 0.1x_8 \\ & + 0.8x_9 - 0.3x_{11} + 0.3x_{12} + 0.3x_{14} + 0.5x_{15} + 0.9x_{16} \\ & + 0.2x_{17} - 0.3x_{18} - 0.5x_{19} + 0.6x_{20}. \end{aligned} \tag{15}$$

Note that in this experiment, we have not added noise in the generation of y. It is important to remark that the model in Eq. (15) excludes explicitly the following features: x_1, x_6, x_{10}, and x_{13}. However, x_6 is included as a transformation of x_2, i.e., $x_6 = x_2^2$. Moreover, some variables present linear correlation: x_8 and x_{11}, x_{10} and x_{12}, x_5 and x_{14}, x_2 and x_{17}. Indeed, x_{11}, x_{12}, x_{14}, and x_{17} are obtained with a linear transformation of another variable plus noise as shown in Table IV. Some variables, like x_2 and x_3, as well as x_7 and x_9, share identical coefficients but follow different distributions. This design introduces collinearity and redundant information, creating a robust dataset for evaluating model performance.

Model used within the RMs. We define the parametric model that is used to assess the different RMs for the dataset described in Section V-B. The relationship between inputs and outputs is studied using the linear parametric model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}. \tag{16}$$

The regularized least squares (LS) estimator is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}, \tag{17}$$

where $\lambda = 0.5$ and $\mathbf{I}$ is a diagonal unit matrix. Hence, the predicted output according to the model is

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}. \tag{18}$$

the error vector is $\hat{e} = y - \hat{y} = (I - X(X^TX)^{-1}X^T)y$, where $\hat{e} = [\hat{e}_1, \ldots, \hat{e}_N]^T$, and I is an $N \times N$ identity matrix. To evaluate the model's performance, we use the Euclidean-norm to compute the error. Note that the parametric model is linear as the true model. Then, we remove the issue of model mismatc,h and we can focus on the comparison of the RMs. The ranking obtained by each RM is shown in Table V.

C. Scores for each ranking method

This section shows the score obtained by using the scoring functions in Section III, allowing for comparing the different RMs. With this aim, we first define the ground-truth of our true model in (15).

Ground-truth: The variables, considering only their subindices, are ranked in descending order of importance (obtained sorting in decreasing order the absolute values of the coefficients in the true model) as detailed below

$$\begin{aligned} & 16 \to 9 \to (2,3,20) \to (15,19) \to (7,11,12,14,18) \\ & \to (4,17) \to (5,8) \to (1,6,10,13), \end{aligned}$$

where indices within the parentheses $(\cdot, \ldots, \cdot)$ indicate ties in the ranking, meaning the variables inside the parentheses have the same importance in the model. Any permutation of these variables will be considered a correct ranking. We can compare non-normalized scores between RMs, and normalized scores can be used to compare different scores for the same RMs. The following scoring functions are considered:

S1 - Match counting,
S2 - Permutation distance,
S3 - Distance summing,
S4 - Eq. (13), with $\alpha = 2$, $w_i = {}^1/_R$.
S5 - Eq. (13), with $\alpha = \infty$, $w_i = {}^1/_R$.
S6 - Eq. (13), with $\alpha = 2$, rational decay weights,
S7 - Spearman Correlation,
S8 - Kendall Correlation.
The resulting scores are shown in Table III.

TABLE III
NON-NORMALIZED (LEFT) AND NORMALIZED (RIGHT) SCORES OF EACH RM.

RMs	S1	S2	S3	S4	S5	S6	S7	S8
RM0	11 - 0.55	12 - 0.63	7.5 - 0.79	7.66 - 0.66	6 - 0.32	7.65 - 0.66	0.77	0.67
RM1	15 - 0.75	15 - 0.79	8.2 - 0.86	8.26 - 0.72	6 - 0.32	8.00 - 0.69	0.84	0.77
RM2	16 - 0.80	16 - 0.84	8.6 - 0.91	9.19 - 0.80	10 - 0.53	9.45 - 0.82	0.92	0.84
RM3	14 - 0.70	15 - 0.79	7.4 - 0.78	7.34 - 0.64	6 - 0.32	7.43 - 0.64	0.74	0.62
RM4	13 - 0.65	15 - 0.79	8.7 - 0.92	9.90 - 0.86	14 - 0.74	10.13 - 0.88	0.96	0.87

TABLE V
RANKINGS OF THE VARIABLES/FEATURES.

RMs	Ranking
RM0	16 9 2 20 3 15 7 17 12 11 14 4 8 5 10 1 6 13 18 19
RM1	16 9 2 3 20 15 18 14 12 7 17 11 4 8 5 10 1 6 13 19
RM2	16 9 2 3 20 15 19 14 12 7 17 11 4 8 5 10 1 6 13 18
RM3	16 9 2 20 3 15 17 12 7 14 5 11 4 10 8 6 13 1 18 19
RM4	16 9 2 20 3 15 19 18 14 5 17 12 7 11 4 10 8 6 13 1

TABLE VI
RANKING THE RMS. THE SYMBOL * INDICATES SCORE TIES.

Scoring	Ranking methods				
	1st.	2nd.	3rd.	4th.	5th.
S1	RM2	RM1	RM3	RM4	RM0
S2	RM2	RM1*	RM3*	RM4*	RM0
S3	RM4	RM2	RM1	RM0	RM3
S4	RM4	RM2	RM1	RM0	RM3
S5	RM4	RM2	RM3*	RM1*	RM0*
S6	RM4	RM2	RM1	RM0	RM3
S7	RM4	RM2	RM1	RM0	RM3
S8	RM4	RM2	RM1	RM0	RM3

All input variables are normalized with zero mean and unit variance, ensuring consistent signal power.

Results. We present a summary of the results obtained in Table VI. We observe that RM2 and RM4 show virtually always the best performance, occupying either the first or second position in the ranking among almost all scoring functions. RM4 makes more errors in the central and last positions of the ranking, which are less relevant. RM3 and RM0 exhibit the worst performance. Note that the score functions S7 and S8, based on correlations, coincide in their classifications for the RMs. Similarly, S1, S2, which only focus on correct/incorrect features, give the same classification. Finally, we compare the average of the normalized scores for each RM. We get the following classification: 1) RM4 (0.83), RM2 (0.81), RM1 (0.72), RM3 (0.65), RM0 (0.63). Surprisingly, RM0 (jointly with RM3), which is related to the Shapley values, seems to be the worst RM [8].

VI. CONCLUSIONS

In this work, we evaluated various scoring functions to compare ranking methods (RMs) for feature selection. We explored multiple approaches, from the simplest one, the match counting, to more advanced metrics. We test different RMs and the different scoring functions to evaluate the performance of each RM, as well. Experimental results on synthetic data have shown that RM2 and RM4 (which are two sequential backward procedures) consistently achieved higher scores across different evaluation metrics, indicating their robustness. Moreover, it seems the score measures based on correlation are able to detect relevant behavior in the RMs without subjective choices (such as defining some weights).

REFERENCES

[1] R. San Millán-Castillo, L. Martino, E. Morgado, and F. Llorente, "An exhaustive variable selection study for linear models of soundscape emotions: Rankings and Gibbs analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2460–2474, 2022.

[2] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," *Knowl. Inf. Syst.*, vol. 66, pp. 1575–1637, 2023.

[3] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection - a review and recommendations for the practicing statistician," *Biometrical journal*, vol. 60, no. 3, pp. 431–449, 2018.

[4] D. Wood, T. Papamarkou, M. Benatan, and R. Allmendinger, "Model-agnostic variable importance for predictive uncertainty: an entropy-based approach," *Data Mining and Knowledge Discovery*, vol. 38, no. 6, pp. 4184–4216, 2024.

[5] K. Blesch, D. S. Watson, and M. N. Wright, "Conditional feature importance for mixed data," *AStA Advances in Statistical Analysis*, vol. 108, no. 2, pp. 259–278, 2024.

[6] J. Pries, G. Berkelmans, S. Bhulai, and R. van der Mei, "The berkelmans-pries feature importance method: A generic measure of informativeness of features," *arXiv preprint arXiv:2301.04740*, 2023.

[7] L. Martino, R. S. Millán-Castillo, and E. Morgado, "Spectral information criterion for automatic elbow detection," *Expert Systems with Applications*, vol. 231, p. 120705, 2023.

[8] I. Verdinelli and L. Wasserman, "Feature Importance: A Closer Look at Shapley Values and LOCO," *Statistical Science*, vol. 39, no. 4, pp. 623 – 636, 2024.

[9] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, p. 927312, 2022.

[10] E. Morgado, L. Martino, and R. S. Millán-Castillo, "Universal and automatic elbow detection for learning the effective number of components in model selection problems," *Digital Signal Processing*, vol. 140, p. 104103, 2023.

[11] J. V. Servera, L. Martino, J. Verrelst, J. P. Rivera-Caicedo, and G. Camps-Valls, "Multioutput feature selection for emulation and sensitivity analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–11, 2024.

[12] M. G. Kendall, *Rank correlation methods*, 4th ed. Griffin, 1970.

[13] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904. [Online]. Available: http://www.jstor.org/stable/1412159

[14] J. Goldwasser and G. Hooker, "Statistical significance of feature importance rankings," *arXiv:2401.15800v4*, 2025.

[15] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.