Compare and combine different importance ranking methods for feature selection: a gentle review

M. Marinescu[†], L. Martino[⋆], G. Villacrés[†], S. G. Arcidiacono[⋆], Ó. Barquero[†]
[†] Universidad Rey Juan Carlos, Madrid, Spain.
[⋆] Università degli Studi di Catania, Italy.

April 27, 2025

Abstract

Feature selection remains a highly relevant and actively researched topic across signal processing, statistics, and machine learning. It has gained new relevance recently, especially because of renewed interest in the so-called Shapley values. However, beyond the Shapley values, many possibilities exist to measure (explicitly or implicitly) the importance of a variable for a specific task. Given a measure of importance, we can obtain a ranking of the input features (involved, e.g., in a regression or classification problem), as provided by an algorithm and/or expert system. Consequently, it is also necessary to evaluate the obtained rankings, for instance to identify the most effective ranking method or to aggregate all results into an average ranking, akin to an ensemble average of expert opinions. In this work, we provide an exhaustive review of several scoring functions and techniques designed for evaluating the ranking methods with or without an available ground-truth. Moreover, the work contains some novel elements such as the use of other famous indices, for instance, the Gini coefficient and effective sampling size (ESS) measures. It is important to remark that the paper incorporates insights from a variety of sources across diverse scientific disciplines, including computational statistics, quantitative economics, and machine learning. Finally, we test the described schemes in a controlled experiment on feature selection, in order to compare different ranking methods and to assess their performance and robustness.
Keywords. Ranking methods, Shapley values, feature selection, stochastic multicriteria acceptability analysis, Gini coefficient, ensemble of experts.

1 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 renewed interest in the so-called Shapley values is a clear

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

example [8].
Feature importance can be defined in multiple ways across both regression and classification problems [1, 9]. Given an importance measure, we can build a ranking of the involved variables (e.g., from the most important to the least important) by applying an algorithm and/or using the output of an expert system. This selection problem consists of two main theoretical parts: first, ranking the variables; and second, determining the effective number of variables to finally use in a parsimonious model (see, e.g., [7, 10, 11] for the second part). Clearly, by changing the definition or the computation 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. Moreover, it is often desirable to compute an 'averaged rank' that accounts for all available information in a comprehensive manner. 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), or to be able to properly combine the RM results when no ground-truth is available.
This work presents a comprehensive survey aimed at describing and analyzing a variety of scoring functions designed to evaluate the performance of different ranking methods, in both scenarios, *with and without the presence* of ground-truth. Specifically, the primary objective can be summarized as the task of 'ranking the ranking methods' obtained from an ensemble of expert systems. Furthermore, we investigate methods for combining multiple rankings to generate an averaged rank, along with the associated uncertainty measures.
We start with the simplest scoring functions and gradually increase the complexity step by step [12–14]. We also discuss the benefits and drawbacks of the scoring functions, along with their relationships and differences. Each score is described in detail, guiding practitioners to apply them effectively in future work. We introduce also normalized versions of the scores to allow the comparison among different scoring functions. We highlight the possible use of other famous indices in the literature to be applied as a scoring function: an example is the Gini coefficient [15] and the effective sample size formulas [16–18]. A simple running example is used to facilitate the understanding of the engaged readers. It is important to remark that this work incorporates insights from a variety of sources across diverse scientific literature, from computational statistics, quantitative economics, and machine learning. All these diverse methodologies are presented in a joint-unique framework and are described with the same notation within the feature selection context.
Finally, we test all the described scoring schemes in a synthetic regression experiment, considering several alternative ranking methods within a variable selection context [1]. The compared RMs belong all to the family of the so-called *wrapper methods* [19–21]. The primary goal of the experiment is to evaluate the performance of the described scoring functions, both with and without ground-truth. A secondary outcome of the experiment, due to the provided analysis, is that it offers insights into the compared ranking methodologies, indicating which ones perform better or worse in terms of feature selection.

2 Problem statement and main notation of the work

In this work, we describe several scoring functions that allow us to measure the performance of an RM when ground-truth is available and also when is not available. In the first part, i.e., Section 3, we assume the knowledge of a ground-truth ranking. In section 4, we consider strategies when the ground-truth is not available. Next, we describe the main notation used in this work.
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 in 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, whereas x_2 is the worst variable in terms of importance. We recall that the scenario when the ground-truth is not available is also addressed in Section 4.

Generally, we can obtain different rankings for feature selection, each one based (explicitly or implicitly) on different feature importance measures. We desire to score these rankings according to the ground-truth (when it is available as in Section 3) or without ground-truth (as in Section 4). Namely, the goal is to rank the ranking methods or combine them, e.g., discovering the best and the worst RMs, or finding an averaged ranking summarizing all the information. More specifically, a ranking technique yields a ranking of the features in decreasing order of importance that we denote as

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

where $k_i \in \{1, \ldots, R\}$ ($k_i \neq k_j$ 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. For instance, $k_1 = 6$ would mean that the variable x_6 is in the first place of the ranking (i.e., it is the most important), $k_4 = 1$ would mean that the variable x_1 is in the fourth position of the ranking. We desire to "score" this ranking according to the ground-truth.

Functions returning the position. In the rest of the work, we use two functions $r_{\mathcal{R}}$ and $r_{\mathcal{G}}$ which return the position of a feature/variable in a given ranking (where $\mathcal{R}$ and $\mathcal{G}$ denote the set of the feature sub-indices of the ground-truth and ranking, respectively),

$$r_{\mathcal{R}}(\text{feature sub-index}) = \text{position in } \mathcal{R}, \quad \text{and} \quad r_{\mathcal{G}}(\text{feature sub-index}) = \text{position in } \mathcal{G}. \tag{3}$$

Clearly, by definition, the position of the feature k_i in the ranking set $\mathcal{R}$ is $r_{\mathcal{R}}(k_i) = i$, but $r_{\mathcal{G}}(k_i)$ is not generally determined, and depends on the specific scenario. Similarly, the position of the feature g_i, in the ground-truth $\mathcal{G}$, is $r_{\mathcal{G}}(g_i) = i$ (by definition), but $r_{\mathcal{R}}(g_i)$ depends on the specific ranking that we are analyzing.

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 work, to help the understanding of the interested reader. Considering $R = 5$ features, and a ranking with sub-indices:

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

Namely, the variable x_3 is the most important whereas x_4 is the least important (for some specific analyzed task). Moreover, in this running example, we assume that a ranking method provides as a result the following ranking,

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

We can observe that the first two variables are correctly ranked according the ground-truth $\mathcal{G}_E$, whereas the last three variables have been not well-ranked. The most significant error involves variable x_2, which is misplaced by two positions from its correct ranking. Whereas, the other two variables, x_4 and x_5, are only one position away from their correct location.

3 Scoring functions when the ground-truth is available

In this section, we present different possible scoring functions starting with the simplest ones in terms of complexity. Clearly, all the scores described below can be converted trivially in *partial* scores focusing only on the first positions of the ranking (i.e., analyzing a subset of the ranking), if required. The progression of ideas behind the scores in this section is the following:

(a) just counting the errors with equal importance;

(b) then the idea is to take into account the distance between the true position and the wrong position;

(c) and finally also try to take into account that errors in the first positions are more critical than errors in the last positions.

3.1 Baseline and distance scores

Match counting. Perhaps the simplest idea is simply to count the number of correct elements in the ranking. Let us define a binary variable

$$I_j = \begin{cases} 1 & \text{if} \quad k_j = g_j, \\ 0 & \text{if} \quad k_j \neq g_j. \end{cases} \tag{6}$$

Then, the final score is defined as

$$S = \sum_{j=1}^{R} I_j. \tag{7}$$

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 $\bar{S} = \frac{S}{R}$. In the running example, we have $\bar{S} = \frac{2}{5} = 0.4$.

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

$$S = (R - 1) - P. \tag{8}$$

This measure goes from $R - 1$ (perfect matching, i.e., $P = 0$) to the worst-case scenario which corresponds to 0 (i.e., $P = R - 1$). 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 $\bar{S} = 0.5$, with $P = 2$.

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 $d_j = |i^* - j|$.
3. Finally compute the average $D = \frac{1}{R} \sum_{j=1}^{R} d_j$.

Since we desire a score such that the higher its value, then the better the RM is, we have to compute a $D_{\texttt{max}}$. We can achieve this by computing

$$S = D_{\texttt{max}} - D, \tag{9}$$

where

$$\begin{aligned} D_{\texttt{max}} = \frac{1}{R} \sum_{j=1}^{R} d_j^{(\max)} &= \frac{1}{R} \sum_{j=1}^{R} |R - 2j + 1|, \\ &= \frac{1}{R} \sum_{j=1}^{\lfloor R/2 \rfloor} 2(R - 2j + 1) = \begin{cases} \frac{R}{2} & R \text{ even} \\ \frac{1}{R} \frac{R^2-1}{2} & R \text{ odd,} \end{cases} \end{aligned} \tag{10}$$

where we have used $d_j^{(\max)} = |R - 2j + 1|$. Note that $D_{\texttt{max}}$ 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

$$\bar{S} = \frac{S}{D_{\texttt{max}}} = 1 - \frac{D}{D_{\texttt{max}}}. \tag{11}$$

In the case of the example with $\mathcal{G}_E$ and $\mathcal{R}_E$ we have: $S = 1.6$ and $\bar{S} = 1 - \frac{4}{12} = 2/3$.

Generalized weighted distance. The previous score does not consider 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, $\bar{\varphi}_1, \ldots, \bar{\varphi}_R$, such that $\sum \bar{\varphi}_i = 1$. The resulting score is then

$$S = \left(\sum_{j=1}^{R} \bar{\varphi}_j \left(d_j^{(\max)} \right)^{\alpha} \right)^{1/\alpha} - \left(\sum_{j=1}^{R} \bar{\varphi}_j d_j^{\alpha} \right)^{1/\alpha}, \tag{12}$$

where $\alpha > 0$. The normalized score is defined in the same way as Eq. (11). To penalize more the errors in the first positions of the ranking, we can assign weights satisfying $\bar{\varphi}_1 > \bar{\varphi}_2 > \ldots > \bar{\varphi}_R$. However, a clear drawback of this approach is that the choice of these weights is subjective.

3.2 Correlation-based scores

A measure of 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. Observe that, in this way, we cover the points **(a)** and **(b)** described at the beginning of this section.
We recall the two defined functions $r_{\mathcal{R}}$ and $r_{\mathcal{G}}$, which return the position of a feature/variable in a given ranking ($\mathcal{R}$ or $\mathcal{G}$). Clearly, by definition, the position of the feature k_i, in the ranking set $\mathcal{R}$, is $r_{\mathcal{R}}(k_i) = i$. However, $r_{\mathcal{G}}(k_i)$ is not generally determined and depends on the position of k_i in the ground-truth. For simplicity, Table 1 shows the positions of the features in the running example.

Table 1: 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 for ordinal variables [12,13]. In the following, we describe the Spearman and Kendall correlations.

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

$$S = \frac{\text{cov}[r_{\mathcal{R}}, r_{\mathcal{G}}]}{\sigma_{r_{\mathcal{R}}} \sigma_{r_{\mathcal{G}}}}, \tag{13}$$

where $\text{cov}[\cdot,\cdot]$ denotes the covariance, in this case of the data $r_{\mathcal{R}}, r_{\mathcal{G}}$, and $\sigma_{r_{\mathcal{R}}}$, $\sigma_{r_{\mathcal{G}}}$ represent their respective standard deviations. Note that $-1 \leq S \leq 1$. In the running examples, a clear positive association between the positions, $(r_{\mathcal{R}}(k_i), r_{\mathcal{G}}(k_i))$, $i = 1, ..., R$, can be observed, achieving $S = 0.7$. 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. To have a score always normalized between 0 and 1, we can set

$$\bar{S} = \frac{1}{2}(S + 1), \tag{14}$$

so that $\bar{S} = 1$ when $S = 1$, and $\bar{S} = 0$ when $S = -1$. Thus, for the running example, we have $\bar{S} = 0.85$.

Kendall's τ correlation. This method computes the so-called Kendall correlation, which measures the correlation by computing the number of concordant pairs [12]. 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 $\binom{R}{2}$:

$$S = \frac{\text{n}^{\text{o}} \text{ concordant pairs - n}^{\text{o}} \text{ discordant pairs}}{\binom{R}{2}}, \quad \text{and} \quad \bar{S} = \frac{1}{2}(S + 1). \tag{15}$$

In Table 2, we show the calculation of the Kendall correlation for the running example, yielding a result of $S = 0.6$, and hence $\bar{S} = 0.8$. 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 $[-1, 1]$. For this reason, we again define $\bar{S} = \frac{1}{2}(S + 1)$.

Table 2: 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
S	$\frac{(8-2)}{10}$	0.6

3.3 Partial scores

A way to take into account the importance of the first positions in the rankings without defining arbitrary weights is to define partial scores. For the sake of simplicity, we describe the underlying idea only for the *match counting* score, but it can be extended to the other scoring functions. Then, recalling the simply count the number of correct elements in the ranking,

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

and the partial score at step r is

$$S_r = \sum_{j=1}^{r} I_j. \tag{16}$$

We can normalize the partial score by dividing by R, i.e., $\bar{S}_r = \frac{S_r}{R}$. It is interesting to plot $\bar{S}_r$ versus r and compare the ranking for each r. See Figure 1(a) for an example.

3.4 Scores based on cumulative functions

The cumulative idea given in the previous section, for the partial scores, can be generalized as follows. We define a cumulative function C such that $C(j)$ counts how many features in the set $\mathcal{R}_{1:j} = \{k_1, \ldots, k_j\}$ are contained in the first j features of the ground-truth $\mathcal{G}_{1:j} = \{g_1, \ldots, g_j\}$. Mathematically, consider the indicator function

$$G(i|j) = \begin{cases} 1 & \text{if} \qquad k_i \in \{g_1, \ldots, g_j\}, \quad \text{and} \quad i \leq j, \\ 0 & \text{in any other case,} \end{cases} \tag{17}$$

for $j = 1, ..., R$. It is important to note that $G(i|j)$ can be one even if the variable k_i is not perfectly located (i.e., even if there is not a "perfect match"). This is a very interesting property: for instance, if the first $j = 3$ variables (among 20 possible variables) in a ground-truth are $\{g_1 = 20, g_2 = 3, g_3 = 5\}$ and the obtained ranking is $\{k_1 = 5, k_2 = 20, k_3 = 16\}$, this means that $G(i = 1|j = 3) = 1$, and $G(i = 2|j = 3) = 1$. This interesting property is related to the consideration **(b)** at the beginning of Section 3.

After computing all the $G(i|j)$ values, we compute the cumulative function as $C(j) = \sum_{i=1}^{j} G(i|j)$, that satisfies the following properties:

(a) $C(j) \leq j$,

(b) $C(j) \leq C(j+1)$,

(c) $C(R) = R$.

Hence, C is a monotonically non-decreasing function that satisfies $C(R) = R$, for any possible ranking $\mathcal{R}$. In a perfect case scenario, i.e., a perfect match with the ground-truth, we would have $C(1) = 1, C(2) = 2, \ldots, C(R) = R$. Namely, the perfect match case (ideal scenario) is given by

$$C_{\texttt{ideal}}(j) = j, \quad \text{for all } j = 1, ..., R. \tag{18}$$

To measure the discrepancy from this ideal scenario, as in the previous section, i.e.,

$$D_C = \left(\sum_{j=1}^{R} \bar{\varphi}_j (j - C(j))^{\alpha} \right)^{1/\alpha}, \tag{19}$$

and the final score is

$$S = D_C^{(\max)} - D_C, \tag{20}$$

where $D_C^{(\max)}$ is the D_C measure evaluated in the worst case scenario (i.e., considering the opposite ranking with respect to the ground-truth). The normalized score is then defined as:

$$\bar{S} = 1 - \frac{D_C}{D_C^{(\max)}}, \tag{21}$$

Other discrepancy measures from the uniform distribution. The cumulative sum can be easily converted into a cumulative function by the normalization $\widetilde{C}(j) = \frac{1}{R} C(j)$ for all $j = 1, ..., R$ and add also $\widetilde{C}(0) = 0$. The ideal case corresponds to the cumulative function of a uniform probability mass function (pmf). Clearly, any index that measures the distance between a cumulative discrete function $\widetilde{C}$ and the ideal cumulative function of a uniform pmf, $\widetilde{C}_{\texttt{ideal}}$ (that is a straight line from 0 to R) can be employed. Some relevant examples of indices in the literature that measure the discrepancy between cumulative functions are given below:

- **Gini Index (GI)**. GI has been introduced as a way to measure the income inequality in a population [15]. There are several formulations of the Gini coefficient [22] (for related indices see also [23]). Generally, the GI takes a value between 0 and 1. A convenient formulation for this work is the following:

$$\text{GI} = 1 - \frac{2}{R} \left(\frac{1}{2} + \sum_{j=1}^{R-1} \widetilde{C}(j) \right) = 1 - \frac{\text{ENV}}{R}. \tag{22}$$

  where ENV is the index of "effective number of variables" proposed in other contexts [23]. A normalized score for the ranking $\mathcal{R}$ is then defined as

$$\bar{S} = 1 - \text{GI} = \frac{\text{ENV}}{R}. \tag{23}$$

- **Kolmogorov-Smirnov (KS) statistic**. One can use Kolmogorov-Smirnov statistic (KS) that measures the maximum difference between the two cumulative functions,

$$D_{\texttt{KS}} = \sup \left| \widetilde{C}(j) - j \right|, \quad j = 1, ..., R. \tag{24}$$

The KS statistic also ranges between 0 and 1. Hence, the final score is

$$\bar{S} = 1 - D_{\texttt{KS}}. \tag{25}$$

- **Based on effective sample size (ESS).** Let us define the corresponding probability mass function (pmf) as

$$\bar{\rho}_j = \widetilde{C}(j) - \widetilde{C}(j-1), \quad j = 1, ..., R, \text{ and} \tag{26}$$

$$\widetilde{C}(0) = 0. \tag{27}$$

There are several effective sample size (ESS) expressions that actually are discrepancy measures between the pmf defined by $\bar{\rho}_j$ and a discrete uniform distribution [16–18]. We show two famous examples,

$$\text{ESS} = \frac{1}{\sum_{j=1}^{R} \bar{\rho}_j^2}, \quad \text{and/or} \quad \text{ESS} = \frac{1}{\max \bar{\rho}_j}. \tag{28}$$

These two formulas, as other possible examples, have been widely applied in ecology, economics, and social science (see [18]). Note that both $1 \leq \text{ESS} \leq R$. Thus, the score will be

$$\bar{S} = \frac{1}{R-1} \left(\text{ESS} - 1\right), \tag{29}$$

so that $0 \leq \bar{S} \leq 1$.

Any other distance or divergence between probability distributions can be employed (e.g., the KL divergence [24]).

3.5 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, one possible way to address this situation, while still being able to apply the scoring methods described in the previous sections, is to rearrange the ground-truth (accounting for the possible ties) so that it is as close as possible to the ranking that must be evaluated. Namely, we permute within the position of the ties to find the ground-truth sequence that is the closest to the ranking that we need to evaluate. After finding the closest ground-truth sequence, all the scoring functions can be directly applied, as described previously.

4 Scoring without ground-truth

4.1 Compare and combine in the absence of ground-truth

If a ground-truth is not available, the evaluation of several ranking methods becomes more complex, especially if we desire to avoid arbitrary decisions that can adulterate the final considerations. To analyze and/ aggregate (i.e., combine) different rankings when a ground-truth is not available, the main strategy is the so-called *stochastic multicriteria acceptability analysis* (SMAA) [25, 26]. SMAA is a decision-making scheme used when we need to evaluate different options (or alternatives) and there is a lack of precise information about preferences. It has been applied in different contexts: finance and investment, medical decision-making, and other industrial applications (such as product design and project evaluation, where expert opinions vary), to name a few.
The underlying idea is to combine the results of the different rankings without fixing or choosing any arbitrary element that can affect the final analysis (such as weights, scores/rewards, etc.). The SMAA procedure is very simple and powerful at the same time: it is a powerful analytical tool that enables a comprehensive analysis. For instance, the method allows us to calculate an averaged ranking (by employing possibly different *aggregation functions*), along with an associated measure of uncertainty. The resulting analysis also enables checking the robustness of the ranking positions of each variable. All the technical details are given below.

4.2 SMAA procedure

Let us consider having obtained M different rankings of features,

$$\mathcal{R}_m = \{k_1^{(m)}, ..., k_R^{(m)}\}, \quad m = 1, ..., M, \tag{30}$$

from different statistical or machine learning algorithms. Recall that:

- we assume that we have R features, $x_1, ..., x_R$, so that $k_i^{(m)} \in \{1, ..., R\}$;
- by definition, the position of the feature k_i in the ranking set $\mathcal{R}_m$ is $r_{\mathcal{R}}(k_i^{(m)}) = i$, and $k_i^{(m)} \neq k_k^{(m)}$ for $i \neq k$.

For each one of the R variables, we can build vectors containing the positions of the variable in each of the M rankings, i.e.,

$$\begin{aligned}
\mathbf{p}_1 &= [p_{1,1}, p_{1,2}, ..., p_{1,M}]^\top, \\
\mathbf{p}_2 &= [p_{2,1}, p_{2,2}, ..., p_{2,M}]^\top, \\
&\vdots \\
\mathbf{p}_j &= [p_{j,1}, p_{j,2}, ..., p_{j,M}]^\top, \\
&\vdots \\
\mathbf{p}_R &= [p_{R,1}, p_{R,2}, ..., p_{R,M}]^\top,
\end{aligned} \tag{31}$$

where $p_{j,m}$ denotes the position of the j-th variable, i.e., x_j, in the m-th ranking. The vector $\mathbf{p}_j$ contains all the positions of the k-th variable, i.e., x_j, in the different rankings. Then, the standard SMAA procedure is given below:

- For $n = 1, ..., N$:
  - Draw a $1 \times M$ vector $\bar{\mathbf{w}}^{(n)} = [\bar{w}_1^{(n)}, ..., \bar{w}_M^{(n)}]$ uniformly from the simplex of dimension M, i.e., where $\sum_{m=1}^{M} \bar{w}_m^{(n)} = 1$, i.e., they are normalized. It can be done as described in [27, Chapter 6].
  - For $j = 1, ..., R$:
    - Compute one of the two types of the so-called "aggregations" for the k-th variable:

$$\text{weighted mean:} \quad a_j^{(n)} = \bar{\mathbf{w}}^{(n)} \mathbf{p}_j = \sum_{m=1}^{M} \bar{w}_m^{(n)} p_{j,m}, \tag{32}$$

$$\text{or weighted median:} \quad a_j^{(n)} = \text{median}\left(\{p_{j,m}, \bar{w}_m^{(n)}\}\right). \tag{33}$$

  - The aggregated position values $a_j^{(n)}$ are sorted in increasing order, i.e.,

$$a_{i_1}^{(n)} \leq a_{i_2}^{(n)} ... \leq a_{i_R}^{(n)}, \tag{34}$$

    where each $i_r \in \{1, ..., R\}$, and we define the n-th mean position of the j-th variable as

$$\bar{p}_j^{(n)} = \{\text{the value of } r \text{ such that } i_r = j\}, \qquad j = 1, ..., R. \tag{35}$$

Thus, for each feature, we can study the empirical probability mass function defined by the sample positions $\bar{p}_j^{(n)}$, with $n = 1, ..., N$. Hence, we can compute the empirical probabilities

$$b_j(i) = \frac{\#\{\bar{p}_j^{(n)} = i\}}{N}, \qquad i, j \in \{1, ..., R\}. \tag{36}$$

that are called "rank acceptability indices" in the literature. The range of $b_j(i)$ is clearly $[0, 1]$, meaning that the greater its value, the greater is the probability that the j-th feature achieves the position i. For instance, a value $b_i(2) = 0.64$ means that the variable x_i achieves the second position 64% of the time. Table 6 in the numerical experiment shows examples of $b_j(i)$ by using the mean aggregation function given above.

4.3 Average rank positions, uncertainty information and scores

As suggested in [28], these empirical probabilities $b_j(i)$ can be used to assign expected positions to each variable x_j as follows

$$E(j) = \sum_{i=1}^{m} i \; b_j(i). \tag{37}$$

The values above are *expected positions*. To compute a unique average ranking, we rank the variables x_i according to their expected positions $E(x_i)$, using them as scores. Table (8) shows the two average rankings obtained by using the weighted average and the weighted median aggregations. Measures of uncertainty, such as the variance, can also be computed as

$$\mathrm{var}(j) = \sum_{i=1}^{m} \big(i - E(j)\big)^2 \, b_j(i). \tag{38}$$

Moreover, the median or other quantiles can be obtained based on the empirical probabilities $b_j(i)$. Confidence intervals on $E(j)$ can be computed by bootstrap, as well.

Scoring the RMs without groundtruth. Finally, note that if we treat the average ranking as a ground-truth, we can apply all the scoring functions described in the previous sections to evaluate the different ranking methods.

5 Numerical Experiments

To evaluate the different score methods described previously, we utilize the RMs based on wrapper methods [1, 8, 19]. We first consider the use of a ground-truth in Section 5.1, and then we assume that the ground-truth is not available in Section 5.2.

5.1 Numerical experiment with ground-truth

5.1.1 Data generation

We create a synthetic dataset to rank variables under controlled conditions, using the RMs described in Section 5. 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 3.

Table 3: 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}	$z = 0.5x_8 + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1), \quad \frac{z\text{-mean}(z)}{\text{std}(z)}$
x_{12}	$z = 0.5x_{10} + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1), \quad \frac{z\text{-mean}(z)}{\text{std}(z)}$
x_{14}	$z = x_5 + \epsilon, \quad \epsilon \sim \mathcal{N}(0,1), \quad \frac{z\text{-mean}(z)}{\text{std}(z)}$
x_{17}	$z = 0.2x_2 + u, \quad u \sim \mathcal{U}([0,1]), \quad \frac{z\text{-mean}(z)}{\text{std}(z)}$

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

True model: The corresponding observations were generated as follows

$$\begin{aligned} y_n =& 0.6x_2 + 0.6x_3 - 0.2x_4 + 0.1x_5 - 0.3x_7 + 0.1x_8 \\ &+ 0.8x_9 - 0.3x_{11} + 0.3x_{12} + 0.3x_{14} + 0.5x_{15} + 0.9x_{16} \\ &+ 0.2x_{17} - 0.3x_{18} - 0.5x_{19} + 0.6x_{20}. \end{aligned} \tag{39}$$

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. (39) 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 3. Some variables, like x_2 and x_3, as well as x_7 and x_{11}, share identical coefficients but follow different distributions. This design introduces collinearity and redundant information, creating a robust dataset for evaluating model performance.

5.1.2 Ground-truth

In this work, we define the *importance of feature* as the module of the coefficient in the true linear model in Eq. (39). Hence, following this definition, the variables (showing only their sub-indices) are ranked in a decreasing order of importance (i.e., obtained by sorting in decreasing order the absolute values of the coefficients in the true model), i.e.,

	Ground-truth																			
Pos	1^{th}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}								
	16	9	(2, 3, 20)			(15, 19)		(7, 11, 12, 14, 18)												
Pos	13^{th}	14^{th}	15^{th}	16^{th}	17^{th}	18^{th}	19^{th}	20^{th}												
	(4, 17)		(5, 8)		(1, 6, 10, 13)															

where indices within the parentheses $(\cdot, \cdots, \cdot)$ 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.

5.1.3 Application of the ranking methods (RMs)

We apply the RMs based on wrapper methods [1, 8, 19]. 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). A more detailed description of these ranking methods is given in [1, Sec. III A].
Each RM applied in this work uses an internal model. We define the same parametric model that is used to generate the data as the internal model, to assess the different RMs for the dataset described in Section 5.1.1. The relationship between inputs and outputs is studied using the linear parametric model,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}. \tag{40}$$

We apply a regularized least squares (LS) estimator $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^\top\mathbf{y}$, where $\lambda = 0.5$ and $\mathbf{I}$ is a diagonal unit matrix ($\lambda \neq 0$ only to avoid numerical issues). Hence, the predicted output according to the model is $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. To evaluate the model's performance, we use the Euclidean-norm to compute the error. Note that the internal model is linear, as is the true model. Thus, we remove the issue of model mismatch, and we can focus on the comparison of the RMs. The obtained rankings are shown in Table 4.

Table 4: Rankings of the variables/features. We show the indices of the corresponding variable in a decreasing order of importance. For instance, x_{16} is the most relevant for all the RMs. Colored cells show the correctly detected positions (taking into account ties).

RMs	**Ranking**																			
Pos	1^{th}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	13^{th}	14^{th}	15^{th}	16^{th}	17^{th}	18^{th}	19^{th}	20^{th}
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

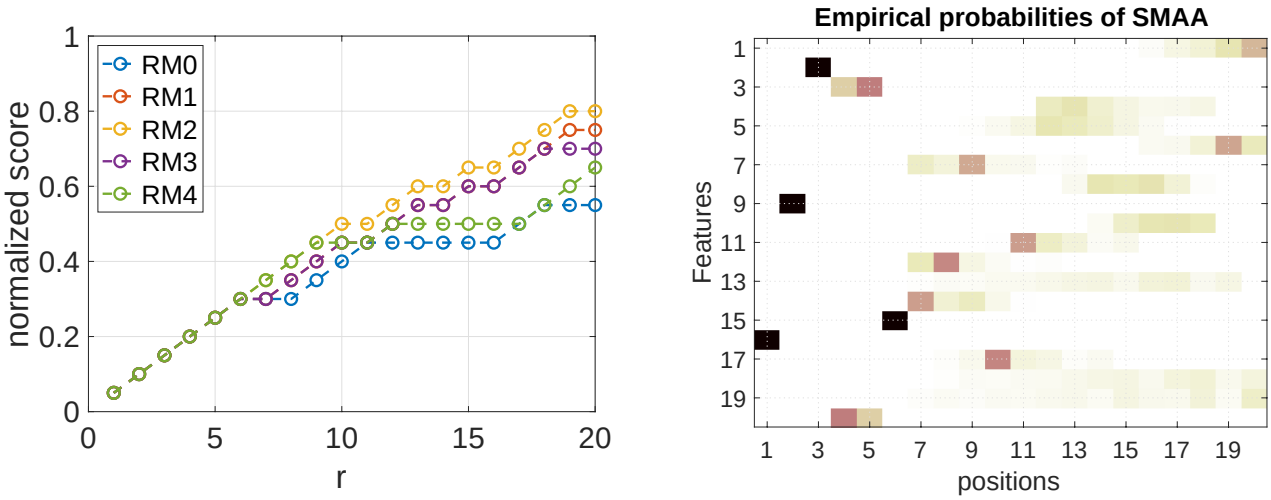
5.1.4 Scores for each ranking method using the ground-truth

This section shows the score obtained by using the scoring functions in Section 3, allowing the comparison of the different RMs. With this aim, we first define the ground-truth of our true model in Eq. (39). We focus on normalized scores to allow the comparison among RMs and also among scoring functions. The following scoring functions are applied:

- S1: Match counting,
- S2: Permutation distance,
- S3: Distance summing,
- S4: Generalized weighted distance, with $\alpha = 2$ and rational decay weights (defined as $\bar{\varphi}_1 = \frac{R}{b}, \bar{\varphi}_2 = \frac{R-1}{b}, \ldots, \bar{\varphi}_R = \frac{1}{b}$, where $b = \frac{R(R+1)}{2}$),
- S5: Spearman Correlation,
- S6: Kendall Correlation,
- S7: Cumulative measure, with $\alpha = 1$, and uniform weights,
- S8: Gini-based score in Eq. (23),
- S9: based on KS statistic in Eq. (25).
- S10: based on Eq. (29) and using ESS $= \frac{1}{\sum_{j=1}^{R} \bar{\rho}_j^2}$.

Results. It is worth noting that all ranking methods failed to correctly place the features x_{10} and x_{17}. This may be because x_{17} is correlated with x_2, which is one of the relevant features. Moreover, x_{10} is correlated with x_{12}. RM0, RM1, and RM3 struggle with x_{19}, placing it erroneously among the lowest ranks. Whereas, RM2 and RM4 accurately identify the appropriate position of x_{19}. The feature x_{18} is well-ranked only by RM4. Note that x_{18} and x_{19} have associated both a negative coefficient in the model.
The resulting normalized scores are presented in Table 5. Moreover, Figure 1(a) depicts the partial scores of each RM using the match counting score, $\bar{S}_r = \frac{1}{R} \sum_{j=1}^{r} I_j$. Analyzing Figure 1(a), we can observe that up to the 6-th position, all the RMs perform equally well. Until the 9-th position, RM2 and RM4 obtain the maximum possible score. However, from 9-th to 17-th positions, RM4 has just one "match" and the rest are errors. Finally, RM4 is able to detect correctly the positions of the last 3 variables. Figure 1(a) also shows the increase of uncertainty in the last positions of the rankings. The final scores $\bar{S}_R$ with $r = R = 20$ in Figure 1(a) correspond to the normalized scores of S1.
More generally, we observe that RM1, RM2, and RM4 consistently demonstrate the best performance, most often ranking first or second across the majority of scoring functions. While RM4 tends to make more errors in the middle and lower ranks positions that are typically less relevant, its overall performance remains strong. Notably, the scoring function S10 ranks RM4 the lowest, contrary to the trend observed with the other scoring functions. RM3 and RM0 exhibit

the worst performance. Note that the score functions S5 and S6, based on correlations, and S7, based on a cumulative measure and uniform weights, coincide in their classifications for the RMs. Similarly, S1, S2, which both focus solely on identifying correct versus incorrect features, yield almost identical classifications. The slight difference is because S2 produces ties in its scores, whereas S1 does not, though ties are theoretically possible in both. Additionally, there are features (such as x_{19}) with a negative coefficient but a quite large module. In this case, the RMs often struggle to provide a good rank. Indeed, e.g., regarding x_{19}, all RMs except RM2 and RM4 rank this feature incorrectly, while RM2 and RM4 correctly identify its importance and the relative correct position. Surprisingly, RM0 (jointly with RM3) which is related to the Shapley values, seems to be the worst RM [8].



(a) (b)

Figure 1: **(a)** Partial scores of each RM using the match counting score, $\bar{S}_r = \frac{1}{R}\sum_{j=1}^{r} I_j$. Until the 6-th position, all the RMs perform equally well. Until the 9-th position, RM2 and RM4 obtain the maximum possible score. However, from 9-th to 17-th, RM4 has just one "match" and the rest are errors. Finally, RM4 can correctly detect the positions of the last 3 variables. **(b)** Empirical probabilities $b_j(i)$ of SMAA for x_j, given in Table 6, obtained with weighted mean as aggregation function. Darker colored squares represent probabilities close to 1, whereas lighter colored squares represent probabilities close to 0.

5.2 Numerical experiment without ground-truth

In this section, we consider that the ground-truth is not available and apply SMAA to assess the different RMs. First of all, we compute the empirical probabilities $b_j(i)$ that represent the probability that the j-th feature achieves the position i. Table 6 and Figure 1(b) provide the values of $b_j(i)$. In Figure 1(b), darker colored squares represent probabilities close to 1, whereas lighter colored squares represent probabilities close to 0. We can easily observe that:

Table 5: Normalized scores $\bar{S}$ for each RM. The highlighted cells show the highest scores.

	Exact match		**Based on Distances**		**Based on Corr.**		**Based on Cumulative funct.**			
RMs	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
RM0	0.55	0.63	0.80	0.66	0.89	0.84	0.80	0.90	0.90	0.82
RM1	0.75	0.79	0.87	0.69	0.92	0.88	0.87	0.94	0.95	0.90
RM2	0.80	0.84	0.91	0.82	0.96	0.92	0.91	0.95	0.95	0.90
RM3	0.70	0.79	0.79	0.64	0.87	0.81	0.79	0.89	0.90	0.82
RM4	0.65	0.79	0.92	0.88	0.98	0.94	0.92	0.96	0.90	0.76

- The features x_{16}, x_9, x_2, x_{15} obtain the ranking positions $1^{st}, 2^{nd}, 3^{rd}, 6^{th}$ respectively, with an empirical probability of 1;
- the rest variables present a greater dispersion, specially x_{13}, x_{18} and x_{19}.

Figure 2 also depicts the histograms corresponding to the empirical probabilities $b_j(i)$ for $j \in \{7, 12, 13, 19\}$, i.e., for the features x_7, x_{12}, x_{13} and x_{19}. Note the dispersion in x_{13} and x_{19}. Recall that x_{13} is not contained in the model, whereas x_{19} should be located in 6-th or 7-th positions. RM2 and RM4 accurately identify the appropriate position of x_{19}, in contrast to RM0, RM1, and RM3, which erroneously place it among the lowest ranks. Consequently, the SMAA evaluation of x_{19} reflects the bias introduced by these inaccurate rankings.
We can obtain expected positions and corresponding variances for each variable x_j [28], using the empirical probabilities $b_j(i)$, as illustrated below:

$$E(j) = \sum_{i=1}^{m} i\ b_j(i), \qquad \text{var}(j) = \sum_{i=1}^{m} \big(i - E(j)\big)^2\ b_j(i).$$

The values $E(j)$ and the standard deviations $\sqrt{\text{var}(j)}$ are given in Table 7. Then, to compute the average ranking (AvR), we order the variables x_i according to their scores $E(i)$. The AvR is shown Table 8 and, indirectly also in Table 7. More specifically, Table 7 provides the average position of each feature, whereas Table 8 shows the indices of the variables ordered in decreasing order of importance. Note that AvR is quite close to the ground-truth (even without using it). Thus, AvR could be employed as an approximated ground-truth, when it is not available.
Finally, we have computed the Kendall correlations between the average ranking (AvR) and the RMs are given in Table 9, where we can see a high correlation between AvR and RM4. Therefore, the SMAA procedure attributes significant importance to RM4 in the construction of AvR. Observe that RM4 identifies the appropriate positions of x_{13}, x_{18}, and x_{19}, which are the most difficult variables according to the uncertainty provided by SMAA. This also illustrates the ability of SMAA to identify variables where the RMs struggle, and/or features that are inherently challenging due to the structure of the problem.

Table 6: Empirical probabilities $b_j(i)$ in percentages of the variables x_j, when we use the weighted mean as aggregation function. As an example, we can observe that $b_{16}(1) = 1$ and $b_5(10) = 0.05$.

	Probabilities (in percentages) associated to each position																			
	$b(1)$	$b(2)$	$b(3)$	$b(4)$	$b(5)$	$b(6)$	$b(7)$	$b(8)$	$b(9)$	$b(10)$	$b(11)$	$b(12)$	$b(13)$	$b(14)$	$b(15)$	$b(16)$	$b(17)$	$b(18)$	$b(19)$	$b(20)$
x_1	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	2%	10%	15%	27%	46%
x_2	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
x_3	0%	0%	0%	36%	64%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
x_4	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	22%	27%	17%	11%	7%	7%	10%	0%	0%
x_5	0%	0%	0%	0%	0%	0%	0%	0%	1%	5%	10%	26%	23%	18%	11%	5%	1%	0%	0%	0%
x_6	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	5%	6%	15%	52%	23%
x_7	0%	0%	0%	0%	0%	0%	20%	13%	50%	7%	6%	1%	3%	0%	0%	0%	0%	0%	0%	0%
x_8	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	6%	26%	24%	27%	15%	2%	0%	0%
x_9	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
x_{10}	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	4%	18%	26%	27%	24%	0%	0%
x_{11}	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	54%	21%	14%	4%	8%	0%	0%	0%	0%	0%
x_{12}	0%	0%	0%	0%	0%	0%	23%	60%	10%	4%	1%	1%	0%	0%	0%	0%	0%	0%	0%	0%
x_{13}	0%	0%	0%	0%	0%	0%	1%	4%	3%	5%	6%	8%	10%	11%	7%	14%	15%	6%	10%	0%
x_{14}	0%	0%	0%	0%	0%	0%	54%	16%	22%	8%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%
x_{15}	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
x_{16}	100%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
x_{17}	0%	0%	0%	0%	0%	0%	0%	2%	7%	61%	13%	10%	2%	5%	0%	0%	0%	0%	0%	0%
x_{18}	0%	0%	0%	0%	0%	0%	0%	1%	4%	3%	5%	5%	8%	10%	11%	7%	13%	15%	7%	12%
x_{19}	0%	0%	0%	0%	0%	0%	2%	4%	3%	5%	5%	7%	8%	5%	11%	7%	7%	14%	5%	19%
x_{20}	0%	0%	0%	64%	36%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

Table 7: Expected positions and corresponding standard deviations.

Feature (j)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
$E(j)$	19.03	3.00	4.64	14.15	12.90	18.81	8.80	15.24	2.00	16.47	11.88	8.02	14.40	7.86	6.00	1.00	10.50	15.59	15.35	4.36
std(j)	1.11	0.00	0.48	1.94	1.56	1.01	1.30	1.21	0.00	1.17	1.23	0.87	3.14	1.06	0.00	0.00	1.21	3.18	3.75	0.48
Av. Pos.	20	3	5	13	12	19	9	15	2	18	11	8	14	7	6	1	10	17	16	4

Table 8: Average ranking of the variables by SMAA.

Position	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}	6^{th}	7^{th}	8^{th}	9^{th}	10^{th}	11^{th}	12^{th}	13^{th}	14^{th}	15^{th}	16^{th}	17^{th}	18^{th}	19^{th}	20^{th}
Av. Rank	16	9	2	20	3	15	14	12	7	17	11	5	4	13	8	19	18	10	6	1

Table 9: Kendall correlations between the average ranking (AvR) and the RMs.

RM0	RM1	RM2	RM3	RM4
0.0526	0.2269	0.1319	0.1058	0.4127

6 Conclusions

This work presents a comprehensive survey on methodologies for evaluating, comparing, and integrating multiple ranking methods (RMs) obtained by distinct expert systems. To this purpose, we take into account both settings: one in which a ground-truth is available and one in which it is not.

In our analysis, we have explored a spectrum of methodologies from elementary techniques, such as match counting, to more refined and complex measures. We test different RMs (all of them belonging to the family of the wrapper techniques) and the different scoring functions to evaluate the performance of each RM, as well. Experimental results on synthetic data demonstrate that

RM2 and RM4 (both sequential backward procedures) and RM1 (a sequential forward procedure) consistently attain superior performance across various evaluation metrics, highlighting their robustness. Finally, it is interesting to remark that RM0 (denoted as LOCO in the literature), which is very related to the Shapley values [8], seems to be one of the worst RMs according to almost all the score functions.
Regarding the scoring functions, it seems that the score measures based on correlation or cumulative functions can detect relevant behavior in the RMs without relying on subjective choices (such as defining some weights). Furthermore, we have also considered the case when the ground-truth is not available. Specifically, the SMAA procedure allows the computation of averaged rankings (combining the different RMs) along with corresponding measures of uncertainty. This analysis further enables the assessment of the robustness of each RM. From the experiments, we observe that SMAA was able to detect both the strength of RM4 and, for example, the difficulties all RMs had with the 19-th feature, x_{19}.
Last but not least, it is important to remark that a relevant contribution of this survey is its synthesis of insights from a wide range of sources across various scientific disciplines, including computational statistics, quantitative economics, and machine learning, to name a few. These diverse methodologies have been integrated into a unified framework and have been consistently described using a common notation within the context of feature selection. This will undoubtedly pave the way for further research directions and future studies.

Acknowledgments

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.

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.

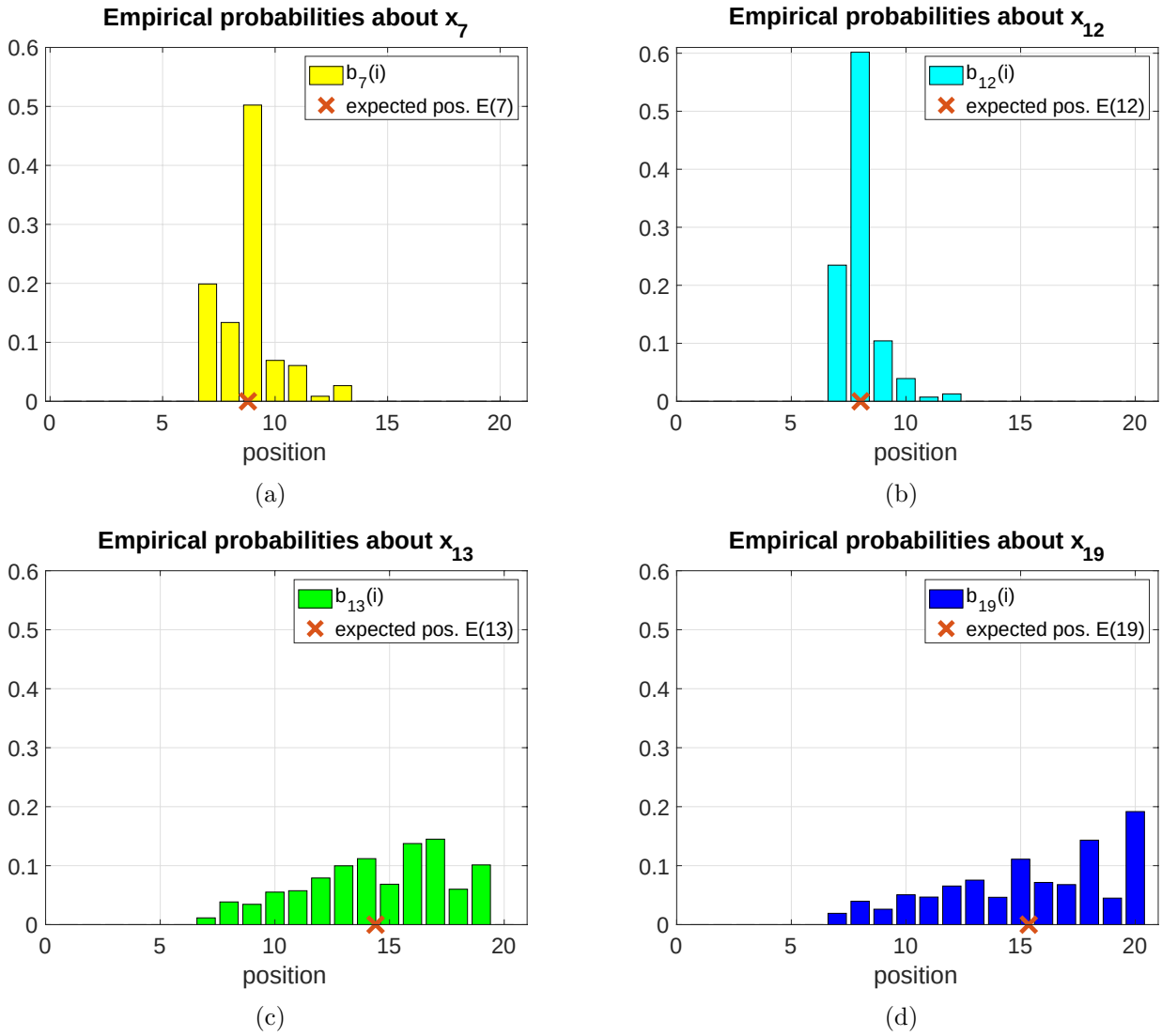


Figure 2: Histograms represent the empirical probabilities $b_j(i)$ obtained by SMAA, for the feature **(a)** x_7, **(b)** x_{12}, **(c)** x_{13}, and **(d)** x_{19}. Recall that x_{13} is not contained in the model, whereas x_{19} should be located in the 6-th or 7-th positions. RM0, RM1, and RM3 incorrectly rank x_{19} near the bottom, whereas RM2 and RM4 correctly identify its position. The SMAA result for x_{19} is influenced by the incorrect rankings produced by RM0, RM1, and RM3.

[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. Vicent 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] C. Gini, "On the Measure of Concentration with Special Reference to Income and Statistics," *Colorade College Publication*, vol. 208, pp. 73–79, 1936.

[16] L. Martino, V. Elvira, and M. F. Louzada, "Effective Sample Size for importance sampling based on the discrepancy measures," *Signal Processing*, vol. 131, pp. 386–401, 2017.

[17] V. Elvira, L. Martino, and C. P. Robert, "Rethinking the effective sample size," *International Statistical Review*, vol. 90, no. 3, pp. 525–550, 2022.

[18] L. Martino and V. Elvira, "Effective sample size approximations as entropy measures," *viXra:2111.0145*, pp. 1–31, 2025.

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

[20] E. Saghapour, S. Kermani, and M. Sehhati, "A novel feature ranking method for prediction of cancer stages using proteomics data," *PLOS ONE*, no. 9, pp. 1–17, 2017.

[21] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-free predictive inference for regression," *Journal of the American Statistical Association*, vol. 113, no. 523, pp. 1094–1111, 2018.

[22] J. L. Gastwirth, "The estimation of the lorenz curve and gini index," *The review of economics and statistics*, pp. 306–316, 1972.

[23] L. Martino, E. Morgado, and R. S. Millán-Castillo, "An index of effective number of variables for uncertainty and reliability analysis in model selection problems," *Signal Processing*, vol. 227, p. 109735, 2025.

[24] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[25] R. Lahdelma, J. Hokkanen, and P. Salminen, "Smaa-Stochastic multiobjective acceptability analysis," *European Journal of Operational Research*, vol. 106, no. 1, pp. 137–143, 1998.

[26] R. Lahdelma and P. Salminen, "SMAA-2: Stochastic multicriteria acceptability analysis for group decision making," *Operations Research*, vol. 49, no. 3, pp. 444–454, 2001.

[27] L. Martino, D. Luengo, and J. Míguez, "Independent random sampling methods," *Springer*, 2018.

[28] M. Kadziński and M. Michalski, "Scoring procedures for multiple criteria decision aiding with robust and stochastic ordinal regression," *Computers & Operations Research*, vol. 71, pp. 54–70, 2016.