Computers & Operations Research 71 (2016) 54–70

Contents lists available at ScienceDirect

Computers & Operations Research

journal homepage: www.elsevier.com/locate/caor

Scoring procedures for multiple criteria decision aiding with robust and stochastic ordinal regression

Miłosz Kadziński [a,*], Marcin Michalski [a,b]

[a] Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland
[b] Poznan Supercomputing and Networking Center, Poznań, Poland

ARTICLE INFO

Available online 19 January 2016

Keywords:
Multiple criteria
Scoring procedures
Efficacy measures
Ranking methods
Ordinal regression
Additive value function

ABSTRACT

We propose several scoring procedures for transforming the results of robustness analysis to a univocal recommendation. We use a preference model in form of an additive value function, and assume the Decision Maker (DM) to provide pairwise comparisons of reference alternatives. We adapt single- and multi-stage ranking methods to select the best alternative or construct a complete ranking by exploiting four types of outcomes: (1) necessary preference relation, (2) pairwise outranking indices, (3) extreme ranks, and (4) rank acceptability indices. In each case, a choice or ranking recommendation is obtained without singling out a specific value function. We compare the proposed scoring procedures in terms of their ability to suggest the same recommendation as the one obtained with the Decision Maker's assumed "true" value function. To quantify the results of an extensive simulation study, we use the following comparative measures (including some newly proposed ones): (i) hit ratio, (ii) normalized hit ratio, (iii) Kendall's τ, (iv) rank difference measure, and (v) rank agreement measure. Their analysis indicates that to identify the best "true" alternative, we should refer to the acceptability indices for the top rank(s), whereas to reproduce the complete "true" ranking it is most beneficial to focus on the expected ranks that alternatives may attain or on the balance between how much each alternative outranks and is outranked by all other alternatives.

1. Introduction

In Multiple Criteria Decision Aiding (MCDA), a preference elicitation process consists in an interaction between a Decision Maker (DM) and an analyst, leading the DM to express information on her/his preferences [9]. Such information is represented with a set of compatible parameter values related to the formulation of the assumed preference model. The application of thus inferred model on the set of alternatives should lead to a result which is consistent with the DM's value system.

In the recent years, a growing interest in MCDA has been assigned to preference disaggregation analysis, which is a general methodological framework for the development of a preference model using decision examples provided by the DM [19]. Methods based on the indirect preference information are considered useful for several reasons. Firstly, they require less cognitive effort from the DM, not forcing her/him to express the preferences directly in terms of some technical or complex model parameters. Secondly, reproducing the provided decision examples, these approaches exhibit the link between an inferred preference model and the suggested recommendation.

In this paper, we focus on multiple criteria choice and ranking problems. We use an additive value function as a preference model, and assume the DM to provide some holistic pairwise comparisons of reference alternatives. Thus, we consider the most standard preference disaggregation setting. This setting exhibits a more general problem of potential existence of many instances of the preference model which reproduce all pieces of preference information given by the DM. Although all compatible preference model instances provide the same desired outputs for the reference set, their recommendation can differ significantly when applied to other alternatives.

1.1. Dealing with robustness concern in preference disaggregation methods

In the context of preference disaggregation, four major value-based approaches have been proposed in the literature to deal with the indetermination of the DM's preference model [12].

In Robust Ordinal Regression (ROR), the whole set of compatible value functions is analyzed, thus, addressing the robustness

* Corresponding author. Tel.: +48 61 665 3022.
E-mail addresses: milosz.kadzinski@cs.put.poznan.pl (M. Kadziński), marcinmi@man.poznan.pl (M. Michalski).

http://dx.doi.org/10.1016/j.cor.2016.01.007
0305-0548/

concern (e.g., [15,16,21,25]; for a review, see [9]). To examine how different can be recommendation obtained for these functions, two weak preference relations are considered. Whether for an ordered pair of alternatives there is necessary or possible preference depends on the truth of preference relation for all or at least one compatible value function, respectively. Thus, the necessary preference relation can be considered as robust with respect to the preference information, because it guarantees that a pair of alternatives is compared in the same way whichever compatible model instance is used. Note that ROR is a preference disaggregation extension of some earlier approaches which were developed for dealing with uncertain criteria weights (see, e.g., [10,17,30,32]). While ROR deals with the robustness of pairwise preference relations, Extreme Ranking Analysis (ERA) raises a rank-related perspective [22]. In particular, it delivers the best and the worst ranks attained by each alternative in the set of compatible value functions. Both ROR and ERA determine the necessary, possible, and extreme results using Linear Programming (LP) techniques.

Stochastic Ordinal Regression (SOR) applies Monte-Carlo simulation to derive a representative sample of the set of compatible value functions [26]. Then, it computes a proportion of value functions that confirm preference of one alternative over another, or assign an alternative to some rank from the best to the worst one. These results are materialized with, respectively, pairwise outranking indices and rank acceptability indices [26]. Such outcomes are useful, because in practical decision problems the necessary preference relation leaves many alternatives incomparable, whereas the difference between the extreme ranks may be large. The limitation of SOR is due to an accuracy of simulation-based procedures. Even if the stochastic indices can be estimated to within acceptable error bounds [40], they may still fail to reflect some unlikely, though still possible result. Note that SOR can be seen as a preference disaggregation counterpart of Stochastic Multicriteria Acceptability Analysis (SMAA) (see, e.g., [2,14,28,29,39,41]). Moreover, by exhibiting pairwise outranking indices it refers to the concept of a fuzzy preference relation proposed by Siskos [36].

ROR and SOR have the merit of handling, respectively, all or multiple instances of preference model compatible with the provided preference information. When dealing with a set of compatible value functions, these approaches act with prudence, providing the DM with the possible consequences of her/his partial preference information. This is particularly useful in a constructive learning perspective [9], where the aim is to construct the DM's preferences from scratch and provoke her/him to incrementally enrich the set of exemplary holistic judgments. Nevertheless, the necessary, possible, and extreme results as well as stochastic indices are not clear-cut. In particular, they do not directly generate a univocal recommendation for the set of alternatives. Further, in some decision situations, e.g., in case of big data problems, it is unrealistic to present results of robustness analysis for all considered alternatives [9].

In this perspective, the third stream of research has been focused on selecting a single value function that would approximate the “true” parameter vector of the DM. Such representative instance of the preference model is selected with some mathematical programming techniques. On the one hand, some approaches exploit the polyhedron of compatible value functions to find the one that is “central” [6,13], “mean” [37], or “the most discriminant” [5,11]. On the other hand, yet other procedures construct a single value function making use of the outcomes of ROR or SOR [23,26]. Precisely, these approaches highlight the necessary, possible, extreme, or stochastic consequences of applying the set of compatible value functions. In this way, they select a single function whose representativeness should be understood in the sense of the robustness preoccupation.

The last approach for constructing a recommendation in the context of disaggregation procedures consists in a direct exploitation of the outcomes of robustness analysis, while avoiding to single out a specific preference model instance. In particular, Vetschera [44] proposed different techniques for deriving a complete order of alternatives by exploiting the outcomes of SOR. Let us recall two exemplary procedures in this spirit. On the one hand, we may maximize an average probability of the decided preference relations based on pairwise outranking indices, while ensuring some desired properties of the preference relation. On the other hand, we may account for the smallest probability of assigned ranks while referring to the rank acceptability indices.

1.2. Aim of the paper

Our paper contributes to the stream of research which is focused on transforming the robust results to a univocal recommendation without singling out a specific parameter vector. The main motivation underlying our proposal consists in delivering a recommendation that is easily understandable to non-experts in MCDA, while still accounting for the robustness concern. The aim of this paper is three-fold.

Firstly, we propose several scoring procedures which exploit the results of Robust and Stochastic Ordinal Regression to indicate the best alternative or to order the alternatives from the best to the worst. These approaches implement single- or multi-stage ranking methods in the context of robust and stochastic outcomes. The former assign a score to each alternative, whereas the latter iteratively (downward) apply a particular scoring function on the set of alternatives. The proposed procedures can be divided into four groups based on the type of ranking results they exploit: (1) necessary preference relation, (2) pairwise outranking indices, (3) extreme ranks, and (4) rank acceptability indices.

Although we focus on the outcomes derived from ordinal regression, the proposed scoring procedures are applicable with a set of value functions compatible with other types of incomplete or imprecise preference information. For example, while referring to a set of linear value functions delimited with some linear constraints on the weights, [17,47] determine the strict necessary preference relation, whereas the SMAA methods [28,31] deliver acceptability indices.

The second aim of the paper is to propose some measures for comparing the choice and ranking recommendation delivered by different procedures. Apart from the well-known hit ratio [4] and Kendall's τ [48], we consider the following three measures: (1) normalized hit ratio deriving from the definition of a Jaccard's coefficient [18], (2) rank difference measure investigating the difference between ranks attained by the same alternatives in different approaches, and (3) rank agreement measure which verifies if the alternatives attain exactly the same rank for a pair of compared methods.

Finally, all these measures are used in an experimental comparison of the proposed scoring procedures. Precisely, we investigate the ability to these procedures to suggest the same recommendation as the one obtained with the DM's assumed “true” value function. Even though our framework for the comparative analysis is inspired by [1], many other papers confirm the usefulness of simulation studies for comparing different MCDA approaches or decision rules (see, e.g., [4,33,34,46]). Our results indicate which approaches provide more accurate results in terms of selecting the best “true” alternative or reproducing the “true” complete ranking.

The organization of the paper is the following. In the next section, we remind the existing multiple criteria ranking methods

based on ROR and SOR, and we define the models that are used in our approach. In Section 3, we present several ranking methods which exploit the results of ROR and SOR to order the alternatives from the best to the worst. In Section 4, we discuss five measures of efficacy for comparing the choice or ranking recommendation obtained with two different methods. In Section 5, we discuss the performance of these procedures by referring to the results of an extensive experimental study. The last section concludes the paper.

2. Reminder on robust and stochastic ordinal regression

We use the following notation [9]:

- $A=\{a_1,\ldots,a_i,\ldots,a_n\}$ – a finite set of n alternatives;
- $A^R=\{a^*,b^*,\ldots\}$ – a finite set of reference alternatives on which the DM accepts to express preferences; we assume that $A^R\subseteq A$;
- $G=\{g_1,\ldots,g_j,\ldots,g_m\}$ – a finite set of m evaluation criteria, $g_j: A\to\mathbb{R}$;
- $X_j=\{g_j(a_i),\ a_i\in A\}$ – the set of deterministic evaluations on g_j; we assume, without loss of generality, that the greater $g_j(a_i)$, the better;
- $x_j^1,\ldots,x_j^{n_j(A)}$ – the ordered values of X_j, $x_j^k<x_j^{k+1}, k=1,\ldots,n_j(A)-1$, where $n_j(A)=|X_j|$ and $n_j(A)\le n$.

The DM provides a partial pre-order on the set of reference alternatives A^R, denoted by $\succsim$. In particular, the DM can state that a^* is at least as good as b^* ($a^*\succsim b^*$), a^* is indifferent to b^* ($a^*\sim b^*$), or a^* is strictly preferred to b^* ($a^*\succ b^*$). As a preference model, we use the additive value function:

$$U(a)=\sum_{j=1}^{m}u_j(a) \tag{1}$$

where the marginal value functions $u_j(a)$, $j=1,\ldots,m$, are monotone, non-decreasing and normalized so that the comprehensive value (1) is bounded within the interval [0, 1].

The pairwise comparisons provided by the DM form the input data for the ordinal regression that finds the whole set of value functions able to reconstruct these judgments. Such value functions are compatible with the provided pairwise comparisons. Precisely, a set of such compatible additive value functions $\mathcal{U}^{A^R}_{ROR}$ is defined with the following set of constraints:

$$\left.\begin{array}{l} U(a^*)\ge U(b^*)+\varepsilon, \text{ if } a^*\succ b^* \text{ for } a^*,b^*\in A^R,\\ U(a^*)=U(b^*), \text{ if } a^*\sim b^* \text{ for } a^*,b^*\in A^R,\\ U(a^*)\ge U(b^*), \text{ if } a^*\succsim b^* \text{ for } a^*,b^*\in A^R,\\ u_j(x_j^1)=0,\ \sum_{j=1}^{m}u_j(x_j^{n_j(A)})=1,\\ \text{for all } j=1,\ldots,m \text{ and } k=2,\ldots,n_j(A):\\ \text{for general marginal value functions:}\\ u_j(x_j^k)-u_j(x_j^{(k-1)})\ge 0,\\ \text{for linear marginal value functions:}\\ u_j(x_j^k)=u_j(x_j^{n_j(A)})(x_j^k-x_j^1)/(x_j^{n_j(A)}-x_j^1). \end{array}\right\} E^{A^R}_{ROR} \tag{2}$$

If $E^{A^R}_{ROR}$ is feasible and $\varepsilon^*=\max\varepsilon$, s.t. $E^{A^R}_{ROR}$ is greater than 0, the set of compatible value functions is non-empty. Otherwise, the provided preference information is inconsistent with the assumed preference model.

Note that the interpretation of indifference relation in value-based approaches is very strict, i.e., two alternatives are considered indifferent if and only if their comprehensive values are exactly the same. Reproducing thus defined indifference relation does not pose any problem for the ordinal regression methods which are based on mathematical programming techniques. Moreover, some researchers use this type of preference information on purpose to considerably reduce the set of compatible value functions in ROR and increase the conclusiveness of suggested recommendation [8]. While we will follow this traditional definition, it is worth noting that in SOR, which is based on Monte-Carlo simulation, the exact equality of comprehensive values is unlikely to occur. Thus, one may consider adding some zone of tolerance by introducing the notion of approximate indifference and suitably adapting the definition of considered stochastic indices [45].

When using general non-linear monotonic marginal value functions, all criteria values form their characteristic points. As a result, these functions do not involve any arbitrary or restrictive parametrization, at the same time offering greater flexibility. For easily interpretable linear marginal value functions, there are only two characteristic points corresponding to the extreme performances on a criterion. Nonetheless, only the marginal values for the best performances are considered as variables, while the values assigned to the worst performances are set to zero. Finally, the marginal values assigned to the non-extreme performances are derived from linear interpolation.

2.1. Necessary and possible preference relations

Robust Ordinal Regression applies all compatible value functions $\mathcal{U}^{A^R}_{ROR}$, and defines two binary relations in the set of all alternatives A [16]:

- Necessary weak preference relation, $\succsim^N$, that holds for a pair of alternatives $(a,b)\in A\times A$, in case $U(a)\ge U(b)$ for all compatible value functions;
- Possible weak preference relation, $\succsim^P$, that holds for a pair of alternatives $(a,b)\in A\times A$, in case $U(a)\ge U(b)$ for at least one compatible value function.

The following linear programs (LPs) need to be solved to assess whether the relations hold:

$$\text{Maximize}: \quad \varepsilon$$
$$\text{s.t.} \left.\begin{array}{l} U(b)-U(a)\ge\varepsilon,\\ E^{A^R}_{ROR}, \end{array}\right\} E^N(a,b) \tag{3}$$

and

$$\text{Maximize}: \quad \varepsilon$$
$$\text{s.t.} \left.\begin{array}{l} U(a)-U(b)\ge 0,\\ E^{A^R}_{ROR}. \end{array}\right\} E^P(a,b) \tag{4}$$

Note that we treat ε as a variable, and optimize its value to check whether some hypothesis about comparison of a and b is verified in the set of compatible value functions defined with $E^{A^R}_{ROR}$. Assuming the set of compatible value functions is non-empty (i.e., $E^{A^R}_{ROR}$ is feasible for some $\varepsilon>0$; see problem (2)), $a\succsim^N b$ if $\varepsilon^*_N=\max\ \varepsilon$, s.t. $E^N(a,b)$, is not greater than 0, and $a\succsim^P b$ if $E^P(a,b)$ is feasible and $\varepsilon^*_P=\max\ \varepsilon$, s.t. $E^P(a,b)$, is greater than 0.

2.2. Pairwise outranking indices

Pairwise outranking index $POI(a,b)$ is, for a pair of alternatives $(a,b)\in A\times A$, the share of compatible value functions for which a is not worse than b [31,26]. Consequently, for any $(a,b)\in A\times A$:

$$POI(a,b)\in[0,1] \quad \text{and} \quad POI(a,b)+POI(b,a)\ge 1,$$

and for any $a\in A$, $POI(a,a)=1$. Following the SMAA methods [31,41], we consider a Monte-Carlo estimation of *POIs* [42,43].

2.3. Extreme ranks

To identify the best $P^*(a) = \min_{U \in \mathcal{U}^{A^R}_{ROR}} (rank(U,a))$ and the worst $P_*(a) = \max_{U \in \mathcal{U}^{A^R}_{ROR}} (rank(U,a))$ ranks that a particular alternative $a \in A$ can attain, [22] proposed ERA consisting of the following Mixed-Integer Linear Programming (MILP) models:

$$\text{Minimize}: \quad P^*(a) = 1 + \sum_{b \in A \setminus \{a\}} v_b$$
$$\text{s.t.} \quad \left.\begin{array}{l} U(a) - U(b) + v_b \geq 0, \text{ for all } b \in A \setminus \{a\}, \\ E^{A^R}_{ROR}, \end{array}\right\} E^{A^R}_{max}, \tag{5}$$

and

$$\text{Maximize}: \quad P_*(a) = 1 + \sum_{b \in A \setminus \{a\}} v_b$$
$$\text{s.t.} \quad \left.\begin{array}{l} U(b) - U(a) \geq v_b - 1, \text{ for all } b \in A \setminus \{a\}, \\ E^{A^R}_{ROR}, \end{array}\right\} E^{A^R}_{min}, \tag{6}$$

where v_b is a binary variable associated with the comparison of a to $b \in A \setminus \{a\}$, and ε involved in $E^{A^R}_{ROR}$ is set to an arbitrarily small positive value.

2.4. Rank acceptability indices

The rank acceptability index $RAI(a,k) \in [0,1]$, for alternative $a \in A$ and rank $k = 1, \ldots, n$ is defined as the expected share of compatible value functions that grant alternative a rank k [28,41,26]. Similarly to *POI*s, in what follows we consider a Monte-Carlo estimation of *RAI*s. For each $a \in A$, the following proposition holds:

$$\sum_{k=1}^{n} RAI(a,k) = 1.$$

3. Scoring procedures for robust and stochastic ordinal regression

In this section, we present several scoring procedures which exploit the results of Robust and Stochastic Ordinal Regression to order the alternatives from the best to the worst. We consider two generic score-based ranking methods parametrized by a set of alternatives A, a valued relation $\tilde{R}$ defined over A, and a scoring function sf (for a review, see [38]):

- a single-stage ranking method $\gtrsim^1(A, \tilde{R}, sf)$ assigning a score $sf(a, A, \tilde{R})$ to each alternative $a \in A$; it orders all alternatives from the best to the worst based on the attained scores, i.e., the higher the score, the better (lower) the rank (the potential ties are not decided; thus, alternatives with the same score are ranked ex-aequo);
- a multi-stage ranking method $\gtrsim^i(A, \tilde{R}, sf)$ ranks set A of alternatives by applying a scoring function sf iteratively (downward) on this set; thus, it is used first to identify subsets of alternatives with the same score $sf(a, A, \tilde{R})$; then, the method resolves each tie individually using the same scoring function sf with a scope limited to the subset A' of alternatives ranked ex aequo in all previous stages; the method attempts to break the ties until the number of alternatives in a tied subset is equal to one (i.e., there is no tie anymore) or the cardinality of such tied subset has not changed in the two following two stages (i.e., it is impossible to break the tie); let us emphasize that the scores obtained in different stages are not aggregated but rather used in a lexicographic manner to decide a comprehensive ranking.

To enhance understanding of the difference between single- and multi-stage ranking methods, in Table 2 we provide the scores and comprehensive rankings obtained with a multi-stage method applying a scoring function $sf(a, A', \tilde{R}) = \min_{b \in A' \setminus \{a\}} \tilde{R}(a,b)$. We exploit the relation $\tilde{R}(a,b)$ (see Table 1) defined over set $A = \{a, b, c, d\}$. On the one hand, a single-stage ranking method ends with a ranking $\{a, b\} \succ \{c, d\}$. On the other hand, a multi-stage method is able to additionally break a tie between c and d, however, still failing to discriminate between a and b.

Table 1
The relation $\tilde{R}(a,b)$ employed to illustrate the use of a multi-stage ranking method.

$\tilde{R}(a,b)$	a	b	c	d
a	1.0	0.3	0.5	0.9
b	0.3	1.0	0.7	0.6
c	0.2	0.8	1.0	0.4
d	0.5	0.2	0.6	1.0

Table 2
Scores and comprehensive rankings obtained with a multi-stage ranking method.

Stage	Scores	Comprehensive ranking
I	$sf(a, A, \tilde{R}) = 0.3$, $sf(b, A, \tilde{R}) = 0.3$	$\{a, b\} \succ \{c, d\}$
	$sf(c, A, \tilde{R}) = 0.2$, $sf(d, A, \tilde{R}) = 0.2$	$A^i_I = \{a, b\}, A^{ii}_I = \{c, d\}$
II	$sf(a, A^i_I, \tilde{R}) = 0.3$, $sf(b, A^i_I, \tilde{R}) = 0.3$	$\{a, b\} \succ d \succ c$
	$sf(c, A^{ii}_I, \tilde{R}) = 0.4$, $sf(d, A^{ii}_I, \tilde{R}) = 0.6$	$A^i_{II} = \{a, b\}, A^{ii}_{II} = \{d\}, A^{iii}_{II} = \{c\}$

In general, the final rank of each alternative depends on the exploited relation $\tilde{R}$, the underlying scoring function sf, and the way of deciding the ties. Each of the following four subsections is devoted to exploitation of different results. When it comes to ROR, we exploit the necessary preference relation and extreme ranks. As for SOR, these are rather pairwise outranking and rank acceptability indices. In each subsection, we first define a set of scoring functions sf. Then, we justify which of them will be considered throughout the paper. Finally, we define the underlying ranking methods.

3.1. Scoring functions based on necessary relation

Let us consider a finite set of alternatives A and a function $NEC(a,b)$ indicating the truth or falsity of the necessary relation $\gtrsim^N$ defined over $A \times A$ in the following way:

$$NEC(a,b) = \begin{cases} 1 & \text{if } a \gtrsim^N b, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Then, the score of any alternative $a \in A'$ with respect to the alternatives in $A' \subseteq A$ can be calculated using one of the following scoring functions adapted from [3]:

- "max *in favor based on NEC*", which assigns to $a \in A'$ a score of 1 if it is necessarily preferred to some $b \in A' \setminus \{a\}$, and a score of 0, otherwise:

$$MF(a, A', NEC) = \max_{b \in A' \setminus \{a\}} NEC(a,b); \tag{8}$$

- "min *in favor based on NEC*", which assigns to $a \in A'$ a score of 1 if it is necessarily preferred to all $b \in A' \setminus \{a\}$, and a score of 0,

otherwise:

$$mF(a,A',NEC)=\min_{b\in A'\setminus\{a\}} NEC(a,b); \quad (9)$$

- "*sum in favor based on NEC*", which assigns to $a\in A'$ a score equal to the number of $b\in A'\setminus\{a\}$ to which a is necessarily preferred:

$$SF(a,A',NEC)=\sum_{b\in A'\setminus\{a\}} NEC(a,b); \quad (10)$$

- "max *against based on NEC*", which assigns to $a\in A'$ a score of 0 if there is no $b\in A'\setminus\{a\}$ which is necessarily preferred to it, and a score of -1, otherwise:

$$MA(a,A',NEC)=-\max_{b\in A'\setminus\{a\}} NEC(b,a); \quad (11)$$

- "min *against based on NEC*", which assigns to $a\in A'$ a score of 0 if there is some $b\in A'\setminus\{a\}$ which is not necessarily preferred to it, and a score of -1, otherwise (i.e., if all $b\in A'\setminus\{a\}$ are necessarily preferred to a):

$$mA(a,A',NEC)=-\min_{b\in A'\setminus\{a\}} NEC(b,a); \quad (12)$$

- "*sum against based on NEC*", which assigns to $a\in A'$ a score equal to the negation of the number of $b\in A'\setminus\{a\}$ which are necessarily preferred to a:

$$SA(a,A',NEC)=-\sum_{b\in A'\setminus\{a\}} NEC(b,a); \quad (13)$$

- "max *difference based on NEC*", which assigns to $a\in A'$ a score of 1 if it is strictly necessarily preferred to some $b\in A'\setminus\{a\}$, a score of -1 if all $b\in A'\setminus\{a\}$ are strictly necessarily preferred to it, and a score of 0 if none of the two above scenarios holds for it (i.e., it is incomparable or indifferent in terms of $\succsim^N$ with some $b\in A'\setminus\{a\}$):

$$MD(a,A',NEC)=\max_{b\in A'\setminus\{a\}} [NEC(a,b)-NEC(b,a)]; \quad (14)$$

- "min *difference based on NEC*", which assigns to $a\in A'$ a score of 1 if it is strictly necessarily preferred to all $b\in A'\setminus\{a\}$, a score of 0 if it is incomparable or indifferent in terms of $\succsim^N$ with all $b\in A'\setminus\{a\}$, and a score of -1 if some $b\in A'\setminus\{a\}$ is strictly necessarily preferred to it:

$$mD(a,A',NEC)=\min_{b\in A'\setminus\{a\}} [NEC(a,b)-NEC(b,a)]; \quad (15)$$

- "*sum of differences based on NEC*", which assigns to $a\in A'$ a score equal to the difference between the number of alternatives over which a is necessarily preferred and the number of alternatives which are necessarily preferred to a:

$$SD(a,A',NEC)=\sum_{b\in A'\setminus\{a\}} [NEC(a,b)-NEC(b,a)]. \quad (16)$$

Although all these functions, in general, provide different scores, in what follows we will take into account only the most complex, and, thus, the most discriminating ones, i.e., $MD(a,A',NEC)$, $mD(a,A',NEC)$, and $SD(a,A',NEC)$. In particular, $MD(a,A',NEC)$ ($mD(a,A',NEC)$) can be perceived as a joint consideration of $MF(a,A',NEC)$ and $mA(a,A',NEC)$ ($mF(a,A',NEC)$ and $MA(a,A',NEC)$), whereas $SD(a,A',NEC)=SF(a,A',NEC)+SA(a,A',NEC)$.

Each of these scoring functions will be considered as a parameter of both single- and multi-stage ranking methods. Thus, we will consider the following ranking methods exploiting the necessary preference relation: $\succsim^1(A,NEC,MD)$, $\succsim^i(A,NEC,MD)$, $\succsim^1(A,NEC,mD)$, $\succsim^i(A,NEC,mD)$, $\succsim^1(A,NEC,SD)$, and $\succsim^i(A,NEC,SD)$.

3.2. Scoring functions based on pairwise outranking Indices

Let us consider a finite set of alternatives and pairwise outranking indices *POI*s defined over $A\times A$. Then, the score of any alternative $a\in A'$ with respect to the alternatives in $A'\subseteq A$ can be calculated using one of the following functions:

- "max *in favor based on POIs*", which assigns to $a\in A'$ the maximal *POI* derived from its comparison with some $b\in A'\setminus\{a\}$; thus, it refers to a pairwise comparison which is the most favorable for a in terms of *POI* with a being a predecessor in an ordered pair:

$$MF(a,A',POI)=\max_{b\in A'\setminus\{a\}} POI(a,b); \quad (17)$$

- "min *in favor based on POIs*", which assigns to $a\in A'$ the minimal *POI* derived from its comparison with some $b\in A'\setminus\{a\}$; thus, it takes into account a pairwise comparison which is the least advantageous for a in terms of *POI* with a being a predecessor in an ordered pair:

$$mF(a,A',POI)=\min_{b\in A'\setminus\{a\}} POI(a,b); \quad (18)$$

- "*sum in favor based on POIs*", which assigns to $a\in A'$ the sum of *POI*s derived from its comparison with all $b\in A'\setminus\{a\}$; thus, it expresses how much a is outranking all other alternatives in A':

$$SF(a,A',POI)=\sum_{b\in A'\setminus\{a\}} POI(a,b); \quad (19)$$

- "max *against based on POIs*", which assigns to $a\in A'$ a score equal to a negation of the maximal *POI* derived from a comparison of some $b\in A'\setminus\{a\}$ with a; thus, it takes into account a pairwise comparison which is the least favorable for a with a being a consequent in an ordered pair:

$$MA(a,A',POI)=-\max_{b\in A'\setminus\{a\}} POI(b,a); \quad (20)$$

- "min *against based on POIs*", which assigns to $a\in A'$ a score equal to a negation of the minimal *POI* derived from a comparison of some $b\in A'\setminus\{a\}$ with a; thus, it takes into account a pairwise comparison which is the most advantageous for a with a being a consequent in an ordered pair:

$$mA(a,A',POI)=-\min_{b\in A'\setminus\{a\}} POI(b,a); \quad (21)$$

- "*sum against based on POIs*", which assigns to $a\in A'$ a score equal to a negation of the sum of *POI*s derived from a comparison of all $b\in A'\setminus\{a\}$ with a; thus, it expresses how much a is outranked by all other alternatives in A':

$$SA(a,A',POI)=-\sum_{b\in A'\setminus\{a\}} POI(b,a); \quad (22)$$

- "max *difference based on POIs*", which assigns to $a\in A'$ the maximal difference between $POI(a,b)$ and $POI(b,a)$ derived from

a comparison of a with some $b \in A' \setminus \{a\}$:

$$MD(a, A', POI) = \max_{b \in A' \setminus \{a\}} [POI(a,b) - POI(b,a)]; \quad (23)$$

- "min *difference based on POIs*", which assigns to $a \in A'$ the minimal difference between $POI(a,b)$ and $POI(b,a)$ derived from a comparison of a with some $b \in A' \setminus \{a\}$:

$$mD(a, A', POI) = \min_{b \in A' \setminus \{a\}} [POI(a,b) - POI(b,a)]; \quad (24)$$

- "*sum of differences based on POIs*", which assigns to $a \in A'$ the sum of differences between $POI(a,b)$ and $POI(b,a)$ derived from a comparison of a with all $b \in A' \setminus \{a\}$; thus, it reflects the balance between how much a outranks and is outranked by all other alternatives in A':

$$SD(a, A', POI) = \sum_{b \in A' \setminus \{a\}} [POI(a,b) - POI(b,a)]. \quad (25)$$

In a sampling procedure, such as rejection sampling [26] or Hit-And-Run [42,43], it is extremely unlikely to hit a function for which $U(a) = U(b)$ for any $a, b \in A$. Thus, when considering the estimations of pairwise outranking indices for a pair $a, b \in A$, the following proposition holds in practice: $POI(a,b) = 1 - POI(b,a)$. Consequently, we can indicate triples of scoring procedures based on *POIs* providing equivalent outcomes. These are distinguished by an aggregation operator. For example, when considering sum in favor, sum against, and sum of differences based on *POIs*, the orders of alternatives derived from the analysis of $SF(a, A', POI)$, $SA(a, A', POI)$, and $SD(a, A', POI)$ are the same. In fact, when assuming $POI(a,b) = 1 - POI(b,a)$, *SA* and *SD* can be presented as follows:

$$SA(a, A', POI) = SF(a, A', POI) - |A'| \quad \text{and} \quad SD(a, A', POI) = 2 \cdot SF(a, A', POI) - |A'|.$$

The same holds for the triples of scoring functions using either max or min as an aggregation operator. In what follows, we will consider only a single scoring function from each triple, i.e., $MD(a, A', POI)$, $mD(a, A', POI)$, and $SD(a, A', POI)$.

Again, these three scoring functions will be considered as a parameter of both single- and multi-stage ranking methods. Thus, we will consider the following ranking methods exploiting the pairwise outranking indices: $\succsim^1(A, POI, MD)$, $\succsim^i(A, POI, MD)$, $\succsim^1(A, POI, mD)$, $\succsim^i(A, POI, mD)$, $\succsim^1(A, POI, SD)$, and $\succsim^i(A, POI, SD)$.

3.3. Scoring functions based on extreme ranks

Let us consider a finite set of alternatives and extreme ranks $P^*(a)$ and $P_*(a)$ defined for each $a \in A$. Then, the score of alternative $a \in A' \subseteq A$ can be calculated using one of the following functions based on the extreme ranks:

- "*the best rank*", which assigns to $a \in A'$ a score equal to a negation of the best rank attained by a, i.e., the better (the less) $P^*(a)$, the higher it is ranked:

$$B(a, A', [P^*, P_*]) = -P^*(a); \quad (26)$$

- "*the worst rank*", which assigns to $a \in A'$ a score equal to a negation of the worst rank attained by a, i.e. the better (the less) $P_*(a)$, the higher it is ranked:

$$W(a, A', [P^*, P_*]) = -P_*(a); \quad (27)$$

- "*the best-worst rank*", which assigns to $a \in A'$ a score equal to a negation of the best rank $P^*(a)$ attained by a so that the possible ties are decided with respect to the worst ranks $P_*(a)$; this procedure can be formalized as a scoring function in the following way:

$$BW(a, A', [P^*, P_*]) = -|A'| \cdot P^*(a) - P_*(a); \quad (28)$$

- "*the worst-best rank*", which assigns to $a \in A'$ a score equal to a negation of the worst rank $P_*(a)$ attained by a so that the possible ties are decided with respect to the best ranks $P^*(a)$; this procedure can be formalized as a scoring function in the following way:

$$WB(a, A', [P^*, P_*]) = -|A'| \cdot P_*(a) - P^*(a). \quad (29)$$

The latter two scoring functions account for the ties. Although they can be perceived as multi-stage ranking methods, we have formulated them as single-stage procedures. In fact, for alternatives with the same score obtained in the first stage, they consider additional information rather than apply the same scoring function limited to the alternative judged as indifferent. For consistency of notation, we denote the underlying ranking procedures as single-stage ones: $\succsim^1(A, [P^*, P_*], B)$, $\succsim^1(A, [P^*, P_*], W)$, $\succsim^1(A, [P^*, P_*], BW)$, and $\succsim^1(A, [P^*, P_*], WB)$. Note, however, that one cannot define their multi-stage variants.

3.4. Scoring functions based on rank acceptability indices

Let us consider a finite set of alternatives and rank acceptability indices *RAIs* defined for each $a \in A$ and $k = 1, \ldots, n$. Then, the score of alternative $a \in A' \subseteq A$ can be calculated using one of the following functions based on *RAIs*:

- "*expected rank ER*", which assigns to $a \in A'$ a score equal to a negation of an estimate of its expected rank:

$$E(a, A', RAI) = -\sum_{k=1}^{n} k \cdot RAI(a,k); \quad (30)$$

- "*the best RAI*", which assigns to $a \in A'$ a score equal to a negation of the highest rank k for which $RAI(a,k)$ is greater than zero so that the possible ties are decided with respect the value of $RAI(a,k)$; thus, we first take into account the best possible rank of a estimated with *RAIs*, and then account for the probability of attaining this rank (the higher it is, the better); this procedure can be formalized as a scoring function as follows:

$$B(a, A', RAI) = -\min_{k=1,\ldots,n} \{RAI(a,k) > 0\} + RAI(a, \min_{k=1,\ldots,n} \{RAI(a,k) > 0\}); \quad (31)$$

- "*the worst RAI*", which assigns to $a \in A'$ a score equal to a negation of the lowest rank k for which $RAI(a,k)$ is greater than zero so that the possible ties are decided with respect the value of $RAI(a,k)$; thus, we first take into account the worst possible rank of a estimated with *RAIs*, and then account for the probability of attaining this rank (the higher it is, the worse); this procedure can be formalized as a scoring function as follows:

$$W(a, A', RAI) = -\max_{k=1,\ldots,n} \{RAI(a,k) > 0\} - RAI(a, \max_{k=1,\ldots,n} \{RAI(a,k) > 0\}). \quad (32)$$

Analogously as for the scoring procedures based on the extreme ranks, we denote the underlying ranking methods as single-stage ones: $\succsim^1(A, RAI, E)$, $\succsim^1(A, RAI, B)$, and $\succsim^1(A, RAI, W)$.

A brief summary of the 19 ranking methods that will be used in an experimental study is presented in Appendix A.

4. Measures of efficacy

In this section, we discuss five measures for comparing the recommendation obtained with two different methods. They will be subsequently used for comparing the outcomes suggested by the DM's "true" value function and a ranking method parameterized with a particular scoring function. The proposed measures are related to the performance of methods in terms of the choice or ranking accuracy. The former refers to indicating the best alternative, whereas the latter takes into account rank-order correlation or similarity.

In Section 4.1, we define some ranking and pairwise preference functions. They are subsequently employed in Section 4.2 in the definition of efficacy measures.

4.1. Ranking and pairwise preference functions

When using the scoring procedure SP, the rank of each alternative $a \in A$ derives from a weak preference relation $\succsim^{SP}$ established on A by SP. To ensure comparability between different orders and enhance interpretability of tied ranks, for each alternative $a \in A$ it is reasonable to consider its best $r^*(SP,a)$ and worst $r_*(SP,a)$ ranks that it would otherwise occupy rather than a single arbitrarily selected rank in $[r^*(SP,a), r_*(SP,a)]$ (e.g., only the best rank or the average one).

The best rank of alternative a is defined with the ranking function referring to $\succsim^{SP}$:

$$r^*(SP,a) = |A| - \sum_{b \in A\setminus\{a\}} h^{\geq}(SP,a,b), \quad \text{where} \tag{33}$$

$$h^{\geq}(SP,a,b) = \begin{cases} 1 & \text{if } a \succsim^{SP} b, \\ 0 & \text{otherwise,} \end{cases} \tag{34}$$

where $a \succsim^{SP} b$ holds if a is assigned a score at least as good as a score of b in all stages in which they have been directly compared against each other, thus, attaining the rank which is not worse than the rank of b.

Analogously, the worst rank of a is defined with the following function:

$$r_*(SP,a) = 1 + \sum_{b \in A\setminus\{a\}} h^{\geq}(SP,b,a). \tag{35}$$

If there are no ex aequo ranks, then for all $a \in A$, $r^*(SP,a) = r_*(SP,a)$. Otherwise, for some $a \in A$, $r^*(SP,a) < r_*(SP,a)$. For example, when taking into account a ranking: $a \succ^{SP} b \sim^{SP} c$, the ranks are as follows: $r^*(SP,a) = r_*(SP,a) = 1$, $r^*(SP,b) = r^*(SP,c) = 2$, and $r_*(SP,b) = r_*(SP,c) = 3$. Let $SP(r)$ be a set of alternatives that attain r-th rank when using SP, i.e.:

$$SP(r) = \{a \in A : r^*(SP,a) \leq r \leq r_*(SP,a)\}. \tag{36}$$

Further, the pairwise preference relation for a pair of alternatives $(a,b) \in A \times A$ is defined with the following preference function:

$$p(SP,a,b) = \begin{cases} 1 & \text{if } a \succ^{SP} b, \\ 0.5 & \text{if } a \sim^{SP} b, \\ 0 & \text{if } a = b \vee b \succ^{SP} a, \end{cases} \tag{37}$$

where $a \succ^{SP} b$ holds if a is assigned a score better than a score of b in the last stage in which they have been compared against each other (this implies that a is ranked better than b), whereas $a \sim^{SP} b$ is verified if a and b attain exactly the same scores in all stages, which implies a tie between a and b.

4.2. Measuring choice and ranking accuracy

In this section, we define the measures for quantifying the agreement between recommendation delivered by different procedures. First, we focus on two measures applicable in the context of choice problems, where the task consists in selecting a single best alternative. Second, we refer to the ranking perspective, and discuss three measures accounting for rank-order correlation or similarity.

Hit Ratio (*HR*): When it comes to the choice perspective, the similarity between recommendations delivered by two scoring procedures has been traditionally materialized with a hit ratio [4,1]. When comparing SP' and SP'', a hit occurs if both these procedures rank the same alternative at the very top. Such hit function can be defined as follows:

$$HR(SP',SP'') = \begin{cases} 1 & \text{if } SP'(r=1) \cap SP''(r=1) \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases} \tag{38}$$

While hit function is easily interpretable, its major disadvantage consists in indicating a full agreement even if both procedures rank ex aequo at the very top different sets of alternatives with a non-empty intersection. In an extreme case, $HR(SP',SP'') = 1$, if one of the procedures indicates all alternatives as equally desirable.

In our view, the above mentioned drawback questions using $HR(SP',SP'')$ as a reliable measure for comparing choice recommendations obtained with different methods. Nevertheless, to make its shortcomings evident we will account for it in the experimental study.

A hit ratio can be defined as an average value of the $HR(SP',SP'')$ function obtained throughout the simulation runs. The same remark applies to the remaining efficacy measures presented below. Thus, we will only discuss how to compute these measures for a single simulation run.

Normalized Hit Ratio (*NHR*): While still considering an agreement with respect to the best indicated alternative, in the normalized hit ratio we prevent the major drawback of HR by introducing a revised hit function. It admits partial agreement which is now computed as the size of the intersection of sets of alternatives indicated as the most desirable by SP' and SP'' divided by the size of the union of these sets, i.e.:

$$NHR(SP',SP'') = \frac{|SP'(r=1) \cap SP''(r=1)|}{|SP'(r=1) \cup SP''(r=1)|}. \tag{39}$$

If SP' and SP'' indicate the best alternative unanimously, $HR(SP',SP'') = NHR(SP',SP'')$. Further, if both procedures rank the same alternatives at the top (i.e., $SP'(1) = SP''(1)$), then $NHR(SP',SP'') = 1$. However, if they admit the same alternative to be the best one, but differ with respect to other top ranked alternatives, then $NHR(SP',SP'') < 1$ while $HR(SP',SP'') = 1$. For example, if a and b are ranked ex aequo first by SP', whereas SP'' ranks ex aequo at the top a and c, then $NHR(SP',SP'') = 1/3$, thus, indicating a partial agreement. On the other extreme, if $SP'(r=1) \cap SP''(r=1) = \emptyset$, then $NHR(SP',SP'') = 0$. Note that a definition of the $NHR(SP',SP'')$ function is inspired by a Jaccard coefficient [18], which is commonly used in data mining and information retrieval to measure the overlap of two sets.

Let us now focus on the three measures for quantifying the agreement between complete rankings. Each of these coefficients takes a different perspective for analyzing the rank similarity in the same spirit as, e.g., the necessary and possible preference relations and extreme ranks provide different viewpoints on the robustness of ranking recommendation observed for all compatible value functions. For all proposed measures, if the agreement between the two rankings is perfect, they have value one.

Kendall's τ: Firstly, we analyze the similarity of the constructed rankings from the point of view of pairwise preference relations. This rank-order correlation is materialized with Kendall's τ [48] defined as follows:

$$\tau(SP', SP'', n) = 1 - 4\frac{d_k(SP', SP'')}{n \cdot (n-1)}, \tag{40}$$

where $d_k(SP', SP'')$ is a Kendall's distance between the rankings obtained with SP' and SP'':

$$d_k(SP', SP'') = 0.5 \sum_{(a,b) \in A \times A} |p(SP', a, b) - p(SP'', a, b)|. \tag{41}$$

Note that $d_k(SP', SP'')$ can be interpreted as the number of pairwise preference violations. Overall, Kendall's τ is in the range $[-1, 1]$, where -1 indicates that one ranking negates all pairwise preference relations of the other one, and $+1$ is regarded as a perfect correspondence between the two orders. In this way, the interpretation of τ in terms of the probabilities of observing the agreeable and non-agreeable pairwise preference relations is straightforward.

Since calculations for Kendall's τ are based on the concordant and discordant pairs, they fail to account directly for the ranks attained by alternatives. However, as noted in [24], these positions are of major interest to the DMs when analyzing the final ranking. Therefore, it seems reasonable to additionally account for the rank-oriented measures for quantifying an agreement between different orders.

Rank Difference Measure (RDM): The second measure related to the ranking perspective takes into account the ranks attained by each alternative with different approaches. The Rank Difference Measure is defined as follows:

$$RDM(SP', SP'', n) = 1 - \frac{\sum_{a \in A} |\overline{r}(SP', a) - \overline{r}(SP'', a)|}{\max_{diff}^{rank}(n)}, \tag{42}$$

where

$$|\overline{r}(SP', a) - \overline{r}(SP'', a)| = \frac{\sum_{r' = r^*(SP', a)}^{r_*(SP', a)} \sum_{r'' = r^*(SP'', a)}^{r_*(SP'', a)} |r' - r''|}{[r_*(SP', a) - r^*(SP', a) + 1] \cdot [r_*(SP'', a) - r^*(SP'', a) + 1]} \tag{43}$$

and

$$\max_{diff}^{rank}(n) = \begin{cases} \lfloor n/2 \rfloor \cdot n & \text{if } n \text{ is even,} \\ \lceil n/2 \rceil \cdot (n-1) & \text{if } n \text{ is odd.} \end{cases} \tag{44}$$

Let us first explain the idea underlying definition of $|\overline{r}(SP', a) - \overline{r}(SP'', a)|$ referring to some examples. If a is ranked 5-th and 7-th by, respectively, SP' and SP'', then this measure is equal to $|5-7|/(1 \cdot 1) = 2$. However, if a is ranked ex aequo 5–6 by SP' and 7–9 by SP'', then the rank difference for a is equal to:

$$[|5-7| + |5-8| + |5-9| + |6-7| + |6-8| + |6-9|]/(2 \cdot 3) = 2.5.$$

Further, $\max_{diff}^{rank}(n)$ indicates the maximal sum of rank differences between two complete orders of n alternatives.

When it comes to interpretability of $RDM(SP', SP'')$, it takes values in the range [0, 1], with 1 meaning that each alternative is ranked the same by SP' and SP'', and 0 indicating that the sum of rank differences observed for all alternatives is maximal. While this coefficient is simple to compute and explain, some critics can be raised as to the comparability of rank differences. Indeed, we assume that the difference between any pair of consecutive ranks is the same, while the actual interpretation of ranks is ordinal.

Rank Agreement Measure (RAM): The third measure for comparing two rankings is measuring the overlap of sets of alternatives attaining the same position in both rankings. Precisely, for each of n ranks, we verify if it is attained by the same alternative in the recommendation delivered by SP' and SP'', or from another perspective, if each alternative attains the same rank with SP' and SP''. Thus, the Rank Agreement Measure is defined as follows:

$$RAM(SP', SP'', n) = \frac{1}{n} \sum_{r=1}^{n} RA(SP', SP'', r), \tag{45}$$

where $RA(SP', SP'', r)$ is a rank agreement degree for r-th rank:

$$RA(SP', SP'', r) = \frac{|SP'(r) \cap SP''(r)|}{|SP'(r) \cup SP''(r)|}. \tag{46}$$

The main advantage of $RA(SP', SP'', r)$ consists in generalizing the idea underlying $NHR(SP', SP'')$ function to any rank $r = 1, \ldots, n$. Moreover, RAM can be truncated to consider only k top ranks, for $k < n$. Finally, let us note that the two rankings may be very similar in terms of agreeable pairwise preference relations or low differences between ranks attained by the same alternatives, but the agreement quantified with $RAM(SP', SP'', n)$ may be still low if the alternatives do not attain exactly the same ranks for SP' and SP''. Thus, RAM is more demanding than Kendall's τ or RDM with respect to indicating high agreement values.

In Appendix B, the use of the proposed ranking procedures and efficacy measures is illustrated with an example of the car choice problem, called Thierry's choice [7].

5. Comparative analysis

In this section, we demonstrate the performance of 19 scoring procedures for Robust and Stochastic Ordinal Regression in terms of their ability to suggest the same recommendation as the one obtained with the DM's assumed "true" value function.

5.1. Simulation design

The simulation study has been conducted in the following way:

Step 1: Generate a simulated decision problem with a pre-defined number of criteria and alternatives. The performances are generated randomly from an independent uniform distribution on the (0,1) range.

Step 2: Simulate the DM's "true" value function. Sample the function from a uniform distribution so that it satisfies the monotonicity and normalization constraints. Then, determine the "true" ranking for the simulated decision problem, and draw a pre-defined number of pairwise comparisons consistent with this ranking [42]. Overall, we considered 144 different problem settings (see Table 3). These are distinguished by the assumed type of marginal value functions (either linear (L) or general (G)) as well as the numbers of criteria (m; ranging from 3 to 8), alternatives (n; ranging from 4 to 15), and pairwise comparisons (r; ranging from 2 to 18).

Step 3: Determine the necessary and possible preference relations, extreme ranks, pairwise outranking and rank acceptability indices. The stochastic indices are derived from the analysis of 1000 compatible value functions obtained with the algorithm discussed in [26].

Step 4: Rank the alternatives using 19 scoring procedures introduced in Section 3.

Table 3
Different problem settings considered in the experimental analysis.

Type of marginal value functions	Number of criteria	Number of alternatives	Number of pairwise comparisons
(L,G)	(3,4,6,8)	4	(2,4)
(L,G)	(3,4,6,8)	6	(2,4,6,8)
(L,G)	(3,4,6,8)	(8,10)	(4,6,8,10)
(L,G)	(3,4,6,8)	15	(8,10,14,18)

Step 5: Compare the recommendations delivered by each of the considered scoring procedures and the DM's "true" value function in terms of five efficacy measures discussed in Section 4.2.

In the following subsections, we refer to the average values of efficacy measures obtained in the experimental study. These are discussed separately for linear and general value functions. Each of these average values is derived from 72,000 runs ($72,000 = 72 \cdot 100 \cdot 10$; 72 problem settings determined by the considered combination of m, n, and r; 100 performance matrices for each unique triple (m, n, r); 10 DM's "true" value functions for each generated performance matrix).

The detailed results for all considered problem settings distinguished by the assumed type of marginal value functions, the numbers of criteria, alternatives, and pairwise comparisons, are provided in the e-Appendix. In Appendix C, we provide some exemplary detailed outcomes, and observe the general impact of the individual problem settings on the efficacy measures.

Table 4
Simulation results of the average hit ratio and normalized hit ratio when using either linear or general marginal value functions (the rank attained by each method in terms of a given efficacy measure is provided in the brackets).

Scoring procedure	Hit ratio		Normalized hit ratio	
	Linear	General	Linear	General
$\succsim^1(A, NEC, MD)$	0.9894 (1)	0.9743 (1)	0.1861 (18)	0.1971 (18)
$\succsim^1(A, NEC, mD)$	0.9335 (3)	0.9290 (3)	0.5627 (16)	0.4411 (16)
$\succsim^1(A, NEC, SD)$	0.7713 (15)	0.6991 (16)	0.6922 (12)	0.6180 (12)
$\succsim^i(A, NEC, MD)$	0.8115 (6)	0.7367 (6)	0.6559 (14)	0.5659 (14)
$\succsim^i(A, NEC, mD)$	0.9335 (3)	0.9290 (3)	0.5627 (16)	0.4411 (16)
$\succsim^i(A, NEC, SD)$	0.7713 (15)	0.6991 (16)	0.6922 (12)	0.6180 (12)
$\succsim^1(A, POI, MD)$	0.9894 (1)	0.9743 (1)	0.1861 (18)	0.1971 (18)
$\succsim^1(A, POI, mD)$	0.7747 (8)	0.7143 (8)	0.7746 (2)	0.7142 (2)
$\succsim^1(A, POI, SD)$	0.7733 (10)	0.7116 (10)	0.7732 (6)	0.7115 (5)
$\succsim^i(A, POI, MD)$	0.7366 (18)	0.6569 (18)	0.7366 (7)	0.6569 (7)
$\succsim^i(A, POI, mD)$	0.7747 (8)	0.7143 (8)	0.7746 (2)	0.7142 (2)
$\succsim^i(A, POI, SD)$	0.7733 (11)	0.7115 (12)	0.7733 (4)	0.7115 (6)
$\succsim^1(A, RAI, E)$	0.7733 (11)	0.7116 (10)	0.7732 (5)	0.7115 (4)
$\succsim^1(A, RAI, B)$	0.7751 (7)	0.7146 (7)	0.7751 (1)	0.7146 (1)
$\succsim^1(A, RAI, W)$	0.7248 (19)	0.6558 (19)	0.7247 (8)	0.6558 (8)
$\succsim^1(A, [P^*, P_*], B)$	0.9332 (5)	0.9289 (5)	0.5700 (15)	0.4467 (15)
$\succsim^1(A, [P^*, P_*], W)$	0.7714 (14)	0.7062 (14)	0.6946 (11)	0.6262 (11)
$\succsim^1(A, [P^*, P_*], BW)$	0.7725 (13)	0.7064 (13)	0.6970 (9)	0.6266 (9)
$\succsim^1(A, [P^*, P_*], WB)$	0.7713 (17)	0.7062 (14)	0.6961 (10)	0.6264 (10)

5.2. Simulation results

5.2.1. Measures of efficacy for multiple criteria choice

In Table 4, we provide the average results for Hit Ratio and Normalized Hit Ratio for both linear and general value functions. Although the absolute values do differ, the rankings of scoring procedures for both types of value functions are consistent to a great extent.

When it comes to HR, the best performing methods are $\succsim^1(A, NEC, MD)$ and $\succsim^1(A, POI, MD)$. This confirms the validity of a heuristic indicating that the "true" best alternative is among these which prove their evident superiority over some other alternatives, e.g., by being preferred to them for all compatible value functions. In fact, such a heuristic is valid for about 98% considered problem settings. On the other hand, the worst performing procedures in terms of HR refer to the worst ranks attained by the alternatives (see, e.g., $\succsim^i(A, RAI, W)$ and $\succsim^1(A, [P^*, P_*], WB)$). This observation contradicts an intuitive rule that the better an alternative is in the worst case (i.e., for the least advantageous compatible value function), the more it is valued for the DM.

Interestingly, the experiments indicate that when trying to maximize an average HR, one should rather reduce the amount of information that is taken into account to select the best alternative. Note that the single-stage procedures (e.g., $\succsim^1(A, POI, MD)$) or procedures referring to just a single measure (e.g., $\succsim^1(A, [P^*, P_*], B)$) perform much better than, respectively, the multi-stage procedures (e.g., $\succsim^i(A, POI, MD)$) or procedures accounting for several measures in the lexicographic order (e.g., $\succsim^1(A, [P^*, P_*], BW)$). This is due to an internal definition of HR. Precisely, the procedures which rank several alternatives ex-aequo at the very top artificially increase their chance to hit the "true" best alternative. In an extreme case, for a procedure that ranks all alternatives at the very top, the value of HR would be always equal to one. This clearly shows why the use of HR should be neglected.

The above described characteristic is penalized when computing NHR. As a result, the ranking of scoring procedures implied

Table 5
The statistical significance p for the one-sided Wilcoxon test comparing the average Hit Ratio and Normalized Hit Ratio for five top ranked procedures (0.00^+ is used for denoting $p < 0.001$).

Hit ratio

Linear value functions					General value functions				
	H_2^L	H_3^L	H_4^L	H_5^L		H_2^G	H_3^G	H_4^G	H_5^G
$H_1^L = \succsim^1(A, NEC, MD)$	1.000	0.00^+	0.00^+	0.00^+	$H_1^G = \succsim^1(A, NEC, MD)$	1.000	0.00^+	0.00^+	0.00^+
$H_2^L = \succsim^1(A, POI, MD)$	–	0.00^+	0.00^+	0.00^+	$H_2^G = \succsim^1(A, POI, MD)$	–	0.00^+	0.00^+	0.00^+
$H_3^L = \succsim^1(A, NEC, mD)$	–	–	1.000	0.00^+	$H_3^G = \succsim^1(A, NEC, mD)$	–	–	1.000	0.042
$H_4^L = \succsim^i(A, NEC, mD)$	–	–	–	0.00^+	$H_4^G = \succsim^i(A, NEC, mD)$	–	–	–	0.042
$H_5^L = \succsim^1(A, [P^*, P_*], B)$	–	–	–	–	$H_5^G = \succsim^1(A, [P^*, P_*], B)$	–	–	–	–

Normalized hit ratio

Linear value functions					General value functions				
	N_2^L	N_3^L	N_4^L	N_5^L		N_2^G	N_3^G	N_4^G	N_5^G
$N_1^L = \succsim^1(A, RAI, B)$	0.072	0.072	0.001	0.001	$N_1^G = \succsim^1(A, RAI, B)$	0.219	0.219	0.00^+	0.00^+
$N_2^L = \succsim^1(A, POI, mD)$	–	1.000	0.002	0.001	$N_2^G = \succsim^1(A, POI, mD)$	–	1.000	0.00^+	0.00^+
$N_3^L = \succsim^i(A, POI, mD)$	–	–	0.002	0.001	$N_3^G = \succsim^i(A, POI, mD)$	–	–	0.00^+	0.00^+
$N_4^L = \succsim^i(A, POI, SD)$	–	–	–	0.335	$N_4^G = \succsim^1(A, RAI, E)$	–	–	–	0.074
$N_5^L = \succsim^1(A, RAI, E)$	–	–	–	–	$N_5^G = \succsim^1(A, POI, SD)$	–	–	–	–

by an average NHR differs significantly from the one obtained for HR. On the one hand, the top ranked procedures in terms of HR are now ranked at the very bottom with clearly worst performance measures. Both $\succsim^1(A, NEC, MD)$ and $\succsim^1(A, POI, MD)$ tend to select a large subset of all alternatives as the most desirable, not being sufficiently discriminative for real-world decision aiding.

On the other hand, the best performing procedure in terms of NHR is $\succsim^1(A, RAI, B)$. This indicates that the "true" best alternative is most often ranked first for the greatest number of compatible value functions (i.e., it has the greatest *RAI* for the top rank). The second best procedure is $\succsim^1(A, POI, mD)$ which, in turn, suggests that analysis of the least beneficial pairwise comparison for each alternative may be very useful when identifying the best alternative in set *A*. Indeed, such a maximin approach is implemented in many fields which are closely related to MCDA (e.g., the maximin rule [35] (also called Simpson–Kramer rule) is extensively used in the computational social choice [27]). Note that the best performing scoring procedures manage to indicate the "true" best alternative for over 77% (71%) of considered problem settings with linear (general) value functions.

Statistical significance: Table 5 summarizes the results of the one-sided Wilcoxon test for verifying if a significant difference occurs in the average HR and NHR values for different procedures. To save space, we focus on the five best performers for each unique pair of an efficacy measure and an assumed type of marginal value functions. The results of a comparison for all procedures are provided in the e-Appendix.

When it comes to HR, the hypothesis on the same average values cannot be rejected for comparison of the following pairs: $[\succsim^1(A, NEC, MD), \succsim^1(A, POI, MD)]$ and $[\succsim^1(A, NEC, mD), \succsim^i(A, NEC, mD)]$. Thus, these procedures are ranked ex-aequo first and third, respectively. Moreover, the two best performing procedures significantly outperform all the competitors.

As far as NHR is concerned, there is no statistical difference between three top performers even when assuming the significance level of 0.05. However, these top ranked procedures (i.e., $\succsim^1(A, RAI, B)$, $\succsim^1(A, POI, mD)$, and $\succsim^i(A, POI, mD)$) significantly outperform all other approaches at the significance level of 0.01.

5.2.2. Measures of efficacy for multiple criteria ranking

In Table 6, we provide the average results for Kendall's τ, Rank Difference and Rank Agreement Measures for both linear and general value functions. For the three considered measures and both types of considered value functions, the relative performance of all scoring procedures is consistent to a great extent.

The best performing methods in terms of all considered efficacy measures for multiple criteria ranking are $\succsim^i(A, POI, SD)$, $\succsim^1(A, POI, SD)$, and $\succsim^1(A, RAI, E)$. Let us remind that $\succsim^i(A, POI, SD)$ and $\succsim^1(A, POI, SD)$ exploit the matrix of *POI*s which provides information on the shares of compatible value functions confirming a weak preference relation for each pair of alternatives. As a result, these procedures prefer alternatives which on average outrank all other alternatives more than being outranked by them. Note that it is exceptional that $SD(a, A, POI)$ is the same for at least two alternatives. Thus, the results provided by $\succsim^i(A, POI, SD)$ and $\succsim^1(A, POI, SD)$ are the same in the vast majority of cases (then, the multi-stage procedure is limited to a single-stage one). Equally desirable results for multiple criteria ranking are provided by $\succsim^1(A, RAI, E)$. This intuitive procedure exploits the matrix of *RAI*s in a comprehensive way, and orders the alternatives with respect to the estimates of their expected ranks.

When comparing the orders derived from the best perforrming scoring procedures and the DM's "true" value function, the average values of the efficacy measures for all considered problems settings indicate:

- about 90% consistency (Kendall's $\tau \sim 0.8$) in terms of the observed pairwise preference relations,
- over 80% consistency when accounting for the difference between ranks attained by the alternatives (see RDM),
- around 50% consistency when analyzing if the alternatives are ranked at exactly the same positions (see RAM).

Thus, although Kendall's τ and RDM confirm that the best performing procedures reproduce the majority of preference relations and manage to rank the alternatives rather closely to their positions in the DM's "true" ranking, RAM indicates that these positions are often not exactly the same.

Table 6
Simulation results of the average Kendall's τ, Rank Difference Measure, and Rank Agreement Measure when using either linear or general marginal value functions (the rank attained by each method in terms of a given efficacy measure is provided in the brackets).

Scoring procedure	Kendall's τ		Rank difference measure		Rank agreement measure	
	Linear	General	Linear	General	Linear	General
$\succsim^1(A, NEC, MD)$	0.3491 (18)	0.3665 (19)	0.5512 (18)	0.5585 (19)	0.2294 (18)	0.2185 (18)
$\succsim^1(A, NEC, mD)$	0.3477 (19)	0.3666 (18)	0.5503 (19)	0.5585 (18)	0.2292 (19)	0.2183 (19)
$\succsim^1(A, NEC, SD)$	0.7604 (10)	0.6949 (10)	0.8133 (10)	0.7666 (10)	0.4695 (8)	0.3932 (10)
$\succsim^i(A, NEC, MD)$	0.7067 (14)	0.6222 (15)	0.7744 (14)	0.7150 (15)	0.4152 (14)	0.3384 (14)
$\succsim^i(A, NEC, mD)$	0.7060 (15)	0.6224 (14)	0.7739 (15)	0.7152 (14)	0.4145 (15)	0.3381 (15)
$\succsim^i(A, NEC, SD)$	0.7604 (10)	0.6949 (10)	0.8133 (10)	0.7666 (10)	0.4695 (8)	0.3932 (10)
$\succsim^1(A, POI, MD)$	0.3821 (16)	0.4171 (17)	0.5748 (16)	0.5936 (17)	0.2701 (16)	0.2669 (17)
$\succsim^1(A, POI, mD)$	0.3805 (17)	0.4176 (16)	0.5737 (17)	0.5940 (16)	0.2694 (17)	0.2680 (16)
$\succsim^1(A, POI, SD)$	0.8167 (3)	0.7680 (3)	0.8547 (3)	0.8185 (3)	0.5588 (3)	0.4807 (3)
$\succsim^i(A, POI, MD)$	0.7889 (4)	0.7255 (5)	0.8340 (4)	0.7875 (5)	0.5190 (4)	0.4347 (4)
$\succsim^i(A, POI, mD)$	0.7884 (5)	0.7259 (4)	0.8336 (5)	0.7878 (4)	0.5181 (5)	0.4343 (5)
$\succsim^i(A, POI, SD)$	0.8167 (1)	0.7680 (1)	0.8547 (1)	0.8185 (1)	0.5588 (1)	0.4807 (1)
$\succsim^1(A, RAI, E)$	0.8167 (2)	0.7680 (2)	0.8547 (2)	0.8185 (2)	0.5588 (2)	0.4807 (2)
$\succsim^1(A, RAI, B)$	0.7826 (7)	0.7231 (7)	0.8289 (7)	0.7859 (6)	0.5047 (7)	0.4289 (7)
$\succsim^1(A, RAI, W)$	0.7831 (6)	0.7231 (6)	0.8292 (6)	0.7858 (7)	0.5051 (6)	0.4292 (6)
$\succsim^1(A, [P^*, P_*], B)$	0.7302 (13)	0.6533 (13)	0.7907 (13)	0.7367 (13)	0.4347 (13)	0.3586 (13)
$\succsim^1(A, [P^*, P_*], W)$	0.7308 (12)	0.6537 (12)	0.7911 (12)	0.7370 (12)	0.4349 (12)	0.3599 (12)
$\succsim^1(A, [P^*, P_*], BW)$	0.7618 (9)	0.6962 (9)	0.8135 (9)	0.7670 (8)	0.4678 (10)	0.3936 (9)
$\succsim^1(A, [P^*, P_*], WB)$	0.7619 (8)	0.6962 (8)	0.8137 (8)	0.7670 (9)	0.4678 (11)	0.3938 (8)

Table 7
The statistical significance p for the one-sided Wilcoxon test comparing the average Kendall's τ, Rank Difference Measures, and Rank Agreement Measure for five top ranked procedures (0.00^+ is used for denoting $p < 0.001$).

Kendall's τ

Linear value functions					General value functions				
	τ_2^L	τ_3^L	τ_4^L	τ_5^L		τ_2^G	τ_3^G	τ_4^G	τ_5^G
$\tau_1^L = \succsim^i(A, POI, SD)$	0.038	0.024	0.00^+	0.00^+	$\tau_1^G = \succsim^i(A, POI, SD)$	0.306	0.183	0.00^+	0.00^+
$\tau_2^L = \succsim^1(A, RAI, E)$	–	0.537	0.00^+	0.00^+	$\tau_2^G = \succsim^1(A, RAI, E)$	–	0.186	0.00^+	0.00^+
$\tau_3^L = \succsim^1(A, POI, SD)$	–	–	0.00^+	0.00^+	$\tau_3^G = \succsim^1(A, POI, SD)$	–	–	0.00^+	0.00^+
$\tau_4^L = \succsim^i(A, POI, MD)$	–	–	–	0.281	$\tau_4^G = \succsim^i(A, POI, mD)$	–	–	–	0.271
$\tau_5^L = \succsim^i(A, POI, mD)$	–	–	–	–	$\tau_5^G = \succsim^i(A, POI, MD)$	–	–	–	–

Rank difference measure

Linear value functions					General value functions				
	D_2^L	D_3^L	D_4^L	D_5^L		D_2^G	D_3^G	D_4^G	D_5^G
$D_1^L = \succsim^i(A, POI, SD)$	0.125	0.049	0.00^+	0.00^+	$D_1^G = \succsim^i(A, POI, SD)$	0.170	0.087	0.00^+	0.00^+
$D_2^L = \succsim^1(A, RAI, E)$	–	0.333	0.00^+	0.00^+	$D_2^G = \succsim^1(A, RAI, E)$	–	0.174	0.00^+	0.00^+
$D_3^L = \succsim^1(A, POI, SD)$	–	–	0.00^+	0.00^+	$D_3^G = \succsim^1(A, POI, SD)$	–	–	0.00^+	0.00^+
$D_4^L = \succsim^i(A, POI, MD)$	–	–	–	0.279	$D_4^G = \succsim^i(A, POI, mD)$	–	–	–	0.346
$D_5^L = \succsim^i(A, POI, mD)$	–	–	–	–	$D_5^G = \succsim^i(A, POI, MD)$	–	–	–	–

Rank agreement measure

Linear value functions					General value functions				
	A_2^L	A_3^L	A_4^L	A_5^L		A_2^G	A_3^G	A_4^G	A_5^G
$A_1^L = \succsim^i(A, POI, SD)$	0.371	0.102	0.00^+	0.00^+	$A_1^G = \succsim^i(A, POI, SD)$	0.137	0.017	0.00^+	0.00^+
$A_2^L = \succsim^1(A, RAI, E)$	–	0.129	0.00^+	0.00^+	$A_2^G = \succsim^1(A, RAI, E)$	–	0.109	0.00^+	0.00^+
$A_3^L = \succsim^1(A, POI, SD)$	–	–	0.00^+	0.00^+	$A_3^G = \succsim^1(A, POI, SD)$	–	–	0.00^+	0.00^+
$A_4^L = \succsim^i(A, POI, MD)$	–	–	–	0.151	$A_4^G = \succsim^i(A, POI, MD)$	–	–	–	0.125
$A_5^L = \succsim^i(A, POI, mD)$	–	–	–	–	$A_5^G = \succsim^i(A, POI, mD)$	–	–	–	–

The middle performers in terms of all three measures are scoring procedures exploiting either the "sum of differences based on NEC" or the best and/or the worst ranks. This suggests that the analysis limited only to the exact results (rather than stochastic ones) does not guarantee a satisfactory agreement with the DM's "true" ranking. The worst performing procedures are $\succsim^1(A, NEC, MD)$ and $\succsim^1(A, NEC, mD)$. In general, they discriminate between three groups of alternatives, using only the following scores: -1, 0, or $+1$. Due to this limitation, these procedures are not able to approximate the complete "true" ranking of the DM.

Statistical significance: Table 7 summarizes the results of the one-sided Wilcoxon test for verifying if a significant difference occurs in the average Kendall's τ, RDM, and RAM values for different procedures. Again, we focus on the five best performers only, and provide the complete results in the e-Appendix.

On the one hand, when considering the rank-related efficacy measures produced by the top three performers ($\succsim^i(A, POI, SD)$, $\succsim^i(A, RAI, E)$, and $\succsim^1(A, POI, SD)$), there is no statistical difference between them at the level of 0.01. In some cases this difference is confirmed though at the weaker level of 0.05 or 0.1. On the other hand, the three best procedures significantly outperform all other approaches already at the level of 0.01 in terms of all considered measures of efficacy and types of marginal value functions.

6. Conclusions

The outcomes of an experimental study lead us to indicating some desired properties of ranking methods. Obviously, our remarks are inherently limited by the focus on approximating some assumed "true" DM's choice or ranking.

Firstly, the best performing methods are problem specific. In choice problems, it is most beneficial to account for the acceptability indices for the top rank(s) ($\succsim^1(A, RAI, B)$) or the least advantageous pairwise comparison against some other alternative ($\succsim^1(A, POI, mD)$ and $\succsim^i(A, POI, mD)$). These procedures favor alternatives which prove their superiority over all remaining alternatives for the greatest share of compatible value functions. In ranking problems, in turn, it is useful to analyze a comprehensive performance in view of all pairwise preference relations or all attained ranks. Precisely, to reproduce the complete "true" ranking one should refer to the balance between how much each alternative outranks and is outranked by all other alternatives ($\succsim^1(A, POI, SD)$ and $\succsim^i(A, POI, SD)$) or to the expected ranks that alternatives attain ($\succsim^1(A, RAI, E)$).

All aforementioned procedures but $\succsim^1(A, POI, mD)$ perform reasonably well for both choice and ranking problems. For example, $\succsim^i(A, POI, mD)$ is ranked second for NHR and fourth/fifth in terms of the ranking oriented measures, whereas the positions attained by $\succsim^1(A, RAI, E)$ are just the opposite. These procedures attempt to differentiate all alternatives by assigning them different scores. Instead, $\succsim^1(A, POI, mD)$ leaves indifferent a large subset of relatively worse alternatives, thus, performing well only in a choice perspective. These results confirm that although in the context of choice problems it might be beneficial to use a min operator or refer to some extreme performances, to approximate a complete ranking one should account for a sum operator or an average performance instead.

Secondly, the experimental results indicate that the ranking methods should exploit the most detailed available information derived from the robustness analysis. When using the same operator, the exploitation of *POI*s or *RAI*s proved to be more advantageous than the exploitation of necessary preference relation or extreme ranks, respectively. For example, the simulation-based procedures such as $\succsim^i(A, POI, SD)$ and $\succsim^1(A, RAI, W)$ are significantly better in terms of NHR, Kendall's τ, RDM, and RAM, than their robust LP-based counterparts $\succsim^i(A, NEC, SD)$ and $\succsim^1(A, [P^*, P_*], W)$, respectively. Since sampling the space of compatible value functions is computationally less demanding than using LP techniques, the aforementioned best performing procedures are potentially usable also in the context of big data problems.

From another perspective, instead of referring to just a single premise for constructing a recommendation, it is beneficial to combine it with other rationales considered in a lexicographic order. For example, the agreement with the true DM's recommendation can be increased when breaking the potential ties obtained when considering only the best ranks ($\succsim^1(A, [P^*, P_*], B)$) with the worst ranks ($\succsim^1(A, [P^*, P_*], BW)$) or the probability of attaining these ranks ($\succsim^1(A, RAI, B)$).

Let us emphasize that exploitation of rank-related results (*RAI*s or extreme ranks) does not imply better performance in terms of rank-related measures (NHR, RDM, or RAM). In the same spirit, there is no link between exploitation of pairwise preference relations (*POI*s and $\succsim^N$) and attaining better results for Kendall's τ.

Finally, it is beneficial to use multi-stage methods only with relations and operators whose specificity offers a greater potential for tie breaking. In the context of ranking problems, an iterative application of a scoring procedure proved to be useful with min and max operators (e.g., $\succsim^i(A, NEC, MD)$ or $\succsim^i(A, POI, mD)$ significantly outperform $\succsim^1(A, NEC, MD)$ or $\succsim^1(A, POI, mD)$, respectively). On the contrary, we have not observed the benefits of using several stages with a sum operator which is more discriminating than min or max operators in terms of scores assigned to the alternatives already in the first stage. The differences between $\succsim^1(A, NEC, SD)$ and $\succsim^i(A, NEC, SD)$ or $\succsim^1(A, POI, SD)$ and $\succsim^i(A, POI, SD)$ are not statistically significant.

When it comes to choice problems, multiple stages improve the agreement in terms of NHR only when used together with the max operator. In this case, they help to systematically reduce the subset of the best alternatives. This does not hold for the min and sum operators. In this case, the subset of the most desirable options is often fully distinguished in the first stage, while in the following stages the method rather discriminates the subset of relatively worse alternatives.

We envisage further experimental analysis where the best performers from the current study will be compared with different procedures for selecting a single (central, mean, representative, or discriminant) value function [23,13] in terms of both consistency with the DM's "true" ranking/choice and robustness of the delivered results [13]. Moreover, we will propose some procedures for constructing a univocal recommendation for multiple criteria sorting problems by exploiting different types of sorting results [20]. Finally, let us remark that in this paper, we focused on the ordinal regression methods with preference information in form of pairwise comparisons and a value-based preference model. Obviously, this study can be extended to other forms of preference information and preference models.

Acknowledgments

The first author acknowledges financial support from the Polish National Science Center (Grant no. DEC-2013/11/D/ST6/03056). This research was supported in part by PL-Grid Infrastructure.

Appendix A. Description of the ranking methods used in the experimental study

In Table A1, we provide a brief summary concerning the ranking methods which are used in the experimental study in Section 5.

Table A1
Concise description of the 19 ranking methods used in the experimental study in Section 5 ("ties not resolved" means that alternatives with the same score are ranked ex-aequo; "ties resolved by using a multi-stage procedure" means that the same scoring procedure is applied iteratively on the subsets of alternatives ranked ex-aequo in all previous stages).

Procedure	Description
$\succsim^1(A, NEC, MD)$	$+1$, if $\exists b \in A\backslash\{a\}$, $a \succ^N b$; 0, if $\exists b \in A\backslash\{a\}$, $a \sim^N b$ or $a ?^N b$; -1, if $\forall b \in A\backslash\{a\}$, $b \succ^N a$; ties not resolved; the highest ranked alternatives are necessarily strictly preferred to some other alternative; the lowest ranked alternatives are necessarily strictly preferred by all other alternatives
$\succsim^1(A, NEC, mD)$	$+1$ if $\forall b \in A\backslash\{a\}$, $a \succ^N b$; 0, if $\forall b \in A\backslash\{a\}$, $a \sim^N b$ or $a ?^N b$; -1, if $\exists b \in A\backslash\{a\}$, $b \succ^N a$; ties not resolved; the highest ranked alternatives are necessarily strictly preferred to all other alternatives; the lowest ranked alternatives are necessarily strictly preferred by some other alternative
$\succsim^1(A, NEC, SD)$	$\sum_{b \in A\backslash\{a\}}[NEC(a,b) - NEC(b,a)]$; ties not resolved; alternatives are ordered according to a difference between numbers of alternatives which are necessarily preferred by them and these which are necessarily preferred to them
$\succsim^i(A, NEC, MD)$	Scores as in $\succsim^1(A, NEC, MD)$; ties resolved by using a multi-stage procedure
$\succsim^i(A, NEC, mD)$	Scores as in $\succsim^1(A, NEC, mD)$; ties resolved by using a multi-stage procedure
$\succsim^i(A, NEC, SD)$	Scores as in $\succsim^1(A, NEC, SD)$; ties resolved by using a multi-stage procedure
$\succsim^1(A, POI, MD)$	$\max_{b \in A\backslash\{a\}}[POI(a,b) - POI(b,a)]$; ties not resolved; alternatives are ordered according to their most advantageous balance of pairwise outranking indices when compared with some other alternative
$\succsim^1(A, POI, mD)$	$\min_{b \in A\backslash\{a\}}[POI(a,b) - POI(b,a)]$; ties not resolved; alternatives are ordered according to their least advantageous balance of pairwise outranking indices when compared with some other alternative
$\succsim^1(A, POI, SD)$	$\sum_{b \in A\backslash\{a\}}[POI(a,b) - POI(b,a)]$, ties not resolved; alternatives are ordered according to a balance between how much they outrank and are outranked by all other alternatives
$\succsim^i(A, POI, MD)$	Scores as in $\succsim^1(A, POI, MD)$; ties resolved by using a multi-stage procedure
$\succsim^i(A, POI, mD)$	Scores as in $\succsim^1(A, POI, mD)$; ties resolved by using a multi-stage procedure
$\succsim^i(A, POI, SD)$	Scores as in $\succsim^1(A, POI, SD)$; ties resolved by using a multi-stage procedure
$\succsim^1(A, RAI, E)$	$-\sum_{k=1}^{n} k \cdot RAI(a,k)$; alternative are ordered according to their expected ranks; the lower, the better
$\succsim^1(A, RAI, B)$	$-\min_{k=1,\ldots,n}\{RAI(a,k) > 0\} = -P_R^*(a)$; ties resolved by $RAI(a, P_R^*(a))$; alternatives are ordered according to their best ranks $P_R^*(a)$ observed in the sample (the lower, the better) with ties broken by the probability of attaining this rank estimated with the underlying rank acceptability index (the higher, the better)
$\succsim^1(A, RAI, W)$	$-\max_{k=1,\ldots,n}\{RAI(a,k) > 0\} = -P_{R,*}(a)$; ties resolved by $RAI(a, P_{R,*}(a))$; alternatives are ordered according to their worst ranks $P_{R,*}$ observed in the sample (the lower, the better) with ties broken by the probability of attaining this rank estimated with the underlying rank acceptability index (the lower, the better)
$\succsim^1(A, [P^*, P_*], B)$	$-P^*(a)$; ties not resolved; alternatives are ordered according to their best possible ranks (the lower, the better)
$\succsim^1(A, [P^*, P_*], W)$	$-P_*(a)$; ties not resolved; alternatives are ordered according to their worst possible ranks (the lower, the better)
$\succsim^1(A, [P^*, P_*], BW)$	$-P^*(a)$; ties resolved by $-P_*(a)$; alternatives are ordered according to their best possible ranks with ties broken by their worst ranks
$\succsim^1(A, [P^*, P_*], WB)$	$-P_*(a)$; ties resolved by $-P^*(a)$; alternatives are ordered according to their worst possible ranks with ties broken by their best ranks

Appendix B. Illustrative example

To illustrate the use of Robust and Stochastic Ordinal Regression along with the scoring procedures proposed in Section 3, let us reconsider an example of the choice of a car, called Thierry's choice [7]. The set of alternatives consists of fourteen cars evaluated with respect to five criteria:

- cost (g_1, in Euro) expressing the estimated expenses incurred by buying and using a car,
- acceleration (g_2, in seconds) indicating the time needed to cover the distance of 1 km from standstill,
- pick up (g_3, in seconds) indicating the time needed to cover the distance of 1 km starting in fifth gear at 40 km/h,
- brakes (g_4) expressing braking quality,
- road-hold (g_5) indicating the quality of road holding.

The first three criteria are minimized, while the last two are maximized. The performances of considered cars are provided in Table B1.

B.1. Preference information

For illustrative purpose, we assume the DM to have provided eight pairwise comparisons of reference alternatives:

$$b \succ h, j \succ g, j \succ h, k \succ j, a \succ m, n \succ e, l \succ b, \quad \text{and} \quad l \succ d. \tag{B.1}$$

We analyze the problem with linear marginal value functions. The pairwise comparisons are derived from the DM's true value function with the following maximal shares of individual criteria in the comprehensive value: $w_1(g_1) = 0.20$, $w_2(g_2) = 0.15$, $w_3(g_3) = 0.10$, $w_4(g_4) = 0.20$, and $w_5(g_5) = 0.35$. The underlying complete order of the alternatives is presented in Table B2.

B.2. Necessary preference relations and pairwise outranking indices

For the provided preference information, the necessary weak preference relation is given in Table B3. There are 44 pairs of alternatives $(a, b) \in A \times A$, $a \neq b$, related by the necessary preference $\succsim^N$ (including these which are directly compared by the DM; note that for some pairs the truth of necessary relation is derived directly from the transitivity of preference relation (e.g., $k \succ j$ and $j \succ h$ imply $k \succsim^N h$)). These relations are robust with respect to the provided preference information, meaning that they hold for all compatible value functions. When analyzing the necessary relation, c, k, and l should be perceived as the best ones, because there is no other alternative necessarily preferred to them. On the contrary, d, f, and m should be considered as the worst cars, because they are not necessarily preferred to any other alternative. Nevertheless, the necessary relation leaves many pairs of alternatives equally desirable, stating that there is at least one compatible value function for which one of them is ranked better than the other, and vice versa.

The matrix of the pairwise outranking indices (see Table B4) provides estimates of the shares of compatible value functions that confirm the possible preference relation. For pairs of alternatives related by the necessary relation, the respective *POI* is 100%. When it comes to pairs related by the necessary incomparability, for some of them one car is preferred to the other for the vast majority of compatible value function. In particular, for a and d, $POI(a, d) = 99.4\%$ and $POI(d, a) = 0.6\%$, and for c and g, $POI(c, g) = 99.7\%$ and $POI(g, c) = 0.3\%$. As for the alternatives indicated as the best ones on the basis of the necessary preference relation, analysis of the pairwise outranking indices supports k in comparison with c and l. However, for some pairs of cars, designating the better one on the basis of *POI*s is not straightforward. For example, for h and i, $POI(h, i) = 44.5\%$ and $POI(i, h) = 55.5\%$. In any case, analyzing *POI*s, we are able to state whether some possible relation is "almost sure", "sure on average", "barely sure", or "almost not possible at all" [26].

B.3. Extreme ranks and rank acceptability indices

Table B5 (columns P^* and P_*) shows the best and the worst ranks of each alternative $a \in A$ for all compatible value functions. There are three potential best alternatives (c, k, and l). Among them k is the least sensitive to the choice of a compatible value function, because its rank may drop to 4, while c and l are ranked sixth in the worst case. Overall, there are two and four other alternatives which are ranked, respectively, second and third in the best case. When it comes to relatively worse alternatives, e, f, and h (d and m) are never ranked in top five (ten).

When analyzing the worst ranks, one may note that nine out of fourteen alternatives may be ranked outside top ten in the worst case. However, only three of them (d, f, and m) are possibly ranked at the very bottom. The average width of the rank interval for the analyzed cars is 6.28. The least variation of the attained positions

Table B1
Performance matrix for the Thierry's choice problem.

ID	Name	g_1	g_2	g_3	g_4	g_5
a	Tipo	18,342	30.7	37.25	2.33	3.0
b	Alfa	15,335	30.2	41.6	2.00	2.5
c	Sunny	16,973	29.0	34.9	2.66	2.5
d	Mazda	15,460	30.4	35.8	1.66	1.5
e	Colt	15,131	29.7	35.6	1.66	1.75
f	Corolla	13,841	30.8	36.5	1.33	2.0
g	Civic	18,971	28.0	35.6	2.33	2.0
h	Astra	18,319	28.9	35.3	1.66	2.0
i	Escort	19,800	29.4	34.7	2.00	1.75
j	R19	16,966	30.0	37.7	2.33	3.25
k	P309-16	17,537	28.3	34.8	2.33	2.75
l	P309	15,980	29.6	35.3	2.33	2.75
m	Galant	17,219	30.2	36.9	1.66	1.25
n	R21t	21,334	28.9	36.7	2.00	2.25

Table B2
Ranking of alternatives obtained with the DM's true value function.

Alternative	a	b	c	d	e	f	g	h	i	j	k	l	m	n
Rank	5	7	3	13	10	12	6	9	11	2	1	4	14	8

Table B3
Matrix of the necessary preference relation.

$\succsim^N$	a	b	c	d	e	f	g	h	i	j	k	l	m	n
a	**1**	0	0	0	0	0	0	0	0	0	0	0	**1**	0
b	0	**1**	0	**1**	0	0	0	**1**	0	0	0	0	**1**	0
c	0	0	**1**	**1**	**1**	**1**	0	**1**	**1**	0	0	0	**1**	**1**
d	0	0	0	**1**	0	0	0	0	0	0	0	0	0	0
e	0	0	0	**1**	**1**	0	0	0	0	0	0	0	**1**	0
f	0	0	0	0	0	**1**	0	0	0	0	0	0	0	0
g	0	0	0	**1**	0	0	**1**	0	0	0	0	0	**1**	0
h	0	0	0	0	0	0	0	**1**	0	0	0	0	**1**	0
i	0	0	0	**1**	0	0	0	0	**1**	0	0	0	**1**	0
j	0	0	0	**1**	**1**	**1**	**1**	**1**	0	**1**	0	0	**1**	0
k	**1**	0	0	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0	**1**	**1**
l	0	**1**	0	**1**	**1**	**1**	0	**1**	0	0	0	**1**	**1**	0
m	0	0	0	0	0	0	0	0	0	0	0	0	**1**	0
n	0	0	0	**1**	**1**	0	0	0	0	0	0	0	**1**	**1**

Table B4
Matrix of the pairwise outranking indices (in %).

Alt.	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*	*n*
a	100.0	58.1	1.6	99.4	90.9	98.6	39.2	85.5	81.3	0.9	0.0	0.4	100.0	68.1
b	41.9	100.0	0.3	100.0	99.3	99.8	30.6	100.0	78.9	1.5	0.3	0.0	100.0	67.4
c	98.4	99.7	100.0	100.0	100.0	100.0	99.7	100.0	100.0	86.4	39.3	69.7	100.0	100.0
d	0.6	0.0	0.0	100.0	0.0	41.4	0.0	1.8	0.0	0.0	0.0	0.0	95.9	0.0
e	9.1	0.7	0.0	100.0	100.0	89.5	1.1	27.3	30.6	0.0	0.0	0.0	100.0	0.0
f	1.4	0.2	0.0	58.6	10.5	100.0	1.3	8.5	11.0	0.0	0.0	0.0	84.1	2.5
g	60.8	69.4	0.3	100.0	98.9	98.7	100.0	97.0	90.3	0.0	0.0	1.2	100.0	82.9
h	14.5	0.0	0.0	98.2	72.7	91.5	3.0	100.0	44.5	0.0	0.0	0.0	100.0	12.5
i	18.7	21.1	0.0	100.0	69.4	89.0	9.7	55.5	100.0	2.9	0.0	0.4	100.0	28.9
j	99.1	98.5	13.6	100.0	100.0	100.0	100.0	100.0	97.1	100.0	0.0	21.6	100.0	98.7
k	100.0	99.7	60.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	85.6	100.0	100.0
l	99.6	100.0	30.3	100.0	100.0	100.0	98.8	100.0	99.6	78.4	14.4	100.0	100.0	99.5
m	0.0	0.0	0.0	4.1	0.0	15.9	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0
n	31.9	32.6	0.0	100.0	100.0	97.5	17.1	87.5	71.1	1.3	0.0	0.5	100.0	100.0

Table B5
The matrix of the rank acceptability indices (in %) and extreme ranks (P^* and P_*).

Alt.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	P^*	P_*
a	0.0	0.0	0.3	2.1	28.3	25.0	15.0	11.0	7.2	5.7	4.6	0.6	0.2	0.0	3	13
b	0.0	0.2	0.1	1.3	14.2	26.7	27.4	20.1	9.7	0.3	0.0	0.0	0.0	0.0	2	10
c	36.7	32.5	20.1	8.8	1.8	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	6
d	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.4	38.6	58.3	1.7	11	14
e	0.0	0.0	0.0	0.0	0.0	0.1	0.1	3.2	12.1	28.4	51.3	4.8	0.0	0.0	6	12
f	0.0	0.0	0.0	0.0	0.0	0.5	0.3	1.1	2.1	4.6	10.5	39.8	25.4	15.7	6	14
g	0.0	0.0	0.1	1.2	42.0	29.0	15.5	8.5	2.5	0.9	0.2	0.1	0.0	0.0	3	12
h	0.0	0.0	0.0	0.0	0.0	0.1	3.7	13.3	26.9	35.2	14.8	4.7	1.3	0.0	6	13
i	0.0	0.0	0.4	2.5	2.1	6.1	11.7	11.2	19.9	20.4	16.9	8.8	0.0	0.0	3	12
j	0.0	7.0	21.0	65.8	6.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2	6
k	55.2	35.8	8.8	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	4
l	8.1	24.5	48.9	17.0	1.4	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	6
m	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.6	14.8	82.6	12	14
n	0.0	0.0	0.3	1.1	4.2	12.1	26.3	31.6	19.6	4.5	0.3	0.0	0.0	0.0	3	11

is observed for *m*, *k*, and *d*, while the greatest variation can be attributed to the ranks attained by *a*, *g*, and *i*.

The rank acceptability indices are presented in Table B5 (columns 1–14). When it comes to the potentially top alternatives, *k* is ranked first for 55.2% compatible value functions, whereas for *l* the probability of being ranked at the very top is less than 10%. For some cars, it is possible to indicate a single rank that they attain for the majority of compatible value functions. For example, *j* (*m*) is ranked fourth (fourteenth) by 65.8% (82.6%) functions. For yet other cars, analysis of rank acceptability indices enables to narrow down the range of the most probable ranks. For example, for over 98% (91%) of value functions, *b* (*f*) is ranked between 5 and 9 (outside top ten), whereas, in general, its rank interval is [2, 10] ([6, 14]).

B.4. Scoring procedures

In Table B6, we present the results of applying all considered scoring procedures to the Thierry's choice problem. For each procedure, we present the best $r^*(SP, a)$ and the worst $r_*(SP, a)$ ranks attained by each alternative. Additionally, for all procedures except the multi-stage ones, we provide the underlying scores. For the multi-stage procedures, one cannot present a single comprehensive score for each alternative, because in the subsequent stages the alternatives derive scores only from the comparison against a subset of alternatives judged exactly the same in all previous stages.

To emphasize the difference between exemplary single- and multi-stage scoring procedures exploiting the same results, we illustrate the scores attained by the alternatives for $\succsim^i(A, NEC, MD)$ in Fig. B.1. In the first stage, a score of 1 is assigned to 11 alternatives, and a score of 0 is assigned to 3 remaining alternatives. Note that after this stage $\succsim^1(A, NEC, MD)$ returns the final ranking, while $\succsim^i(A, NEC, MD)$ attempts to make the alternatives more comparable. In the second stage, $\succsim^i(A, NEC, MD)$ applies the same scoring function within two subsets of alternatives ranked ex-aequo in the previous stage. On the one hand, 11 alternatives which proved more advantageous are now divided into a group of 6 alternatives with a score of 1 and a group of 5 alternatives with a score of 0. On the other hand, 3 alternatives judged worse in the first stage attain the same score of 0 when compared against each other, and, thus, they are assigned the final ranks (12–14). The iterative application of the scoring function is continued until each alternative is assigned a final rank, i.e. as long as one can resolve the ties.

In Table B7, we provide the efficacy measures for a comparison of the recommendation delivered with 19 compared scoring procedures and the DM's true value function. These results indicate that all procedures managed to identify *k* as the best alternative (thus, $HR=1$). Nevertheless, just 13 of them ranked only *k* at the very top (then, $NHR=1$), whereas the remaining ones indicated either 2 or 10 other alternatives as equally desirable (then, *NHR* is equal to, respectively, 0.33 or 0.091).

Table B6
Outcomes (final ranks and scores (in parentheses)) of the scoring procedures for the Thierry's choice problem. In all computations we interpreted acceptability indices as decimals (e.g., for $\succsim^1(A, RAI, B)$ the score of a is $-3+0.003=-2.997$).

	Necessary relation (NEC)						Extreme ranks ($[P^*, P_*]$)			
	Single-stage ($\succsim^1$)			Multi-stage ($\succsim^i$)			Single-stage ($\succsim^1$)			
Alt.	MD	mD	SD	MD	mD	SD	B	W	BW	WB
a	1–11 (1)	4–14 (−1)	7–9 (0)	7–11	4–8	7–9	6–9 (−3)	10–11 (−13)	9 (−55)	10 (−185)
b	1–11 (1)	4–14 (−1)	5 (2)	4–6	4–8	5	4–5 (−2)	5 (−10)	5 (−38)	5 (−142)
c	1–11 (1)	1–3 (0)	2 (7)	1–3	1–3	2	1–3 (−1)	2–4 (−6)	2–3 (−20)	2–3 (−85)
d	12–14 (0)	4–14 (−1)	13 (−9)	12–14	13–14	13	13 (−11)	12–14 (−14)	13 (−168)	13 (−207)
e	1–11 (1)	4–14 (−1)	10 (−3)	7–11	9–12	10	10–12 (−6)	7–9 (−12)	10 (−96)	9 (−174)
f	12–14 (0)	4–14 (−1)	11–12 (−4)	12–14	9–12	11–12	10–12 (−6)	12–14 (−14)	12 (−98)	12 (−202)
g	1–11 (1)	4–14 (−1)	7–9 (0)	7–11	9–12	7–9	6–9 (−3)	7–9 (−12)	7–8 (−54)	7–8 (−171)
h	1–11 (1)	4–14 (−1)	11–12 (−4)	7–11	9–12	11–12	10–12 (−6)	10–11 (−13)	11 (−97)	11 (−188)
i	1–11 (1)	4–14 (−1)	7–9 (0)	7–11	4–8	7–9	6–9 (−3)	7–9 (−12)	7–8 (−54)	8–8 (−171)
j	1–11 (1)	4–14 (−1)	4 (5)	4–6	4–8	4	4–5 (−2)	2–4 (−6)	4 (−34)	4 (−86)
k	1–11 (1)	1–3 (0)	1 (10)	1–3	1–3	1	1–3 (−1)	1 (−4)	1 (−18)	1 (−57)
l	1–11 (1)	1–3 (0)	3 (6)	1–3	1–3	3	1–3 (−1)	2–4 (−6)	2–3 (−20)	2–3 (−85)
m	12–14 (0)	4–14 (−1)	14 (−11)	12–14	13–14	14	14 (−12)	12–14 (−14)	14 (−182)	14 (−208)
n	1–11 (1)	4–14 (−1)	6 (1)	4–6	4–8	6	6–9 (−3)	6 (−11)	6 (−53)	6 (−157)

	Pairwise outranking indices (POI)						Rank acceptability indices (RAI)		
	Single-stage ($\succsim^1$)			Multi-stage ($\succsim^i$)			Single-stage ($\succsim^1$)		
Alt.	MD	mD	SD	MD	mD	SD	E	B	W
a	1–11 (1.0)	4–14 (−1.0)	6 (1.480)	8	5	6	6 (−6.760)	7 (−2.997)	10 (−13.002)
b	1–11 (1.0)	4–14 (−1.0)	7 (1.400)	5	6	7	7 (−6.800)	5 (−1.998)	5 (−10.003)
c	1–11 (1.0)	2 (−0.214)	2 (10.860)	2	2	2	2 (−2.068)	2 (−0.633)	3 (−6.001)
d	12 (0.918)	4–14 (−1.0)	13 (−10.210)	12	13	13	13 (−12.603)	13 (−10.986)	12 (−14.017)
e	1–11 (1.0)	4–14 (−1.0)	11 (−5.834)	11	11	11	11 (−10.417)	12 (−5.999)	8 (−12.048)
f	13 (0.682)	4–14 (−1.0)	12 (−9.438)	13	12	12	12 (−12.219)	10 (−5.995)	13 (−14.157)
g	1–11 (1.0)	4–14 (−1.0)	5 (2.990)	7	9	5	5 (−6.005)	9 (−2.999)	7 (−12.001)
h	1–11 (1.0)	4–14 (−1.0)	10 (−4.262)	9	10	10	10 (−9.631)	11 (−5.999)	11 (−13.013)
i	1–11 (1.0)	4–14 (−1.0)	9 (−3.088)	10	8	9	9 (−9.044)	6 (−2.996)	9 (−12.088)
j	1–11 (1.0)	4–14 (−1.0)	4 (7.572)	4	4	4	4 (−3.714)	4 (−1.93)	4 (−6.002)
k	1–11 (1.0)	1 (0.214)	1 (11.920)	1	1	1	1 (−1.540)	1 (−0.448)	1 (−4.002)
l	1–11 (1.0)	3 (−0.712)	3 (9.421)	3	3	3	3 (−2.794)	3 (−0.919)	2 (−6.001)
m	14 (−0.682)	4–14 (−1.0)	14 (−12.600)	14	14	14	14 (−13.80)	14 (−11.974)	14 (−14.826)
n	1–11 (1.0)	4–14 (−1.0)	8 (−0.210)	6	7	8	8 (−7.605)	8 (−2.997)	6 (−11.003)

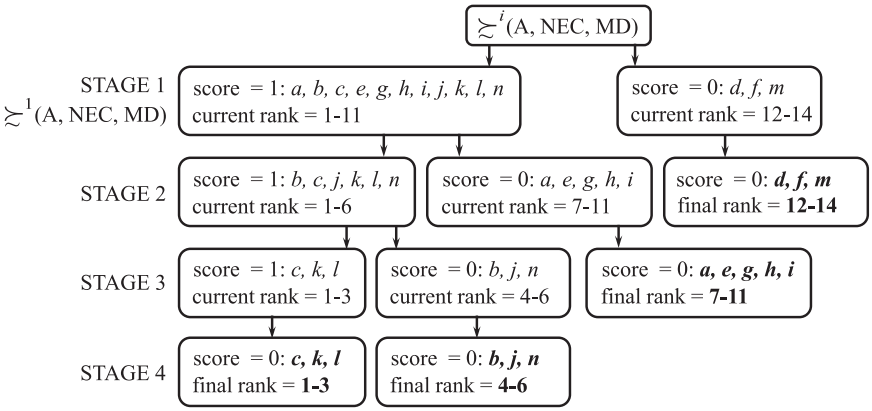

Fig. B.1. Scores and ranks obtained in different stages of $\succsim^i(A, NEC, MD)$ applied to the Thierry's choice problem (stage 1 corresponds to $\succsim^1(A, NEC, MD)$).

When it comes to the ranking perspective, three best procedures ($\succsim^1(A, POI, SD)$, $\succsim^1(A, RAI, E)$, and $\succsim^i(A, POI, SD)$) attained exactly the same results in terms of Kendall's τ, *RDM*, and *RAM*. The comparison of the ranking they delivered with the DM's true order indicates that:

- only 5 $((a,g),(c,j),(j,l),(e,i),(h,i))$ out of 91 pairwise preference relations are inverse; thus, Kendall's $\tau = 1-4\cdot 5/(14\cdot 13)=0.890$;
- the difference of ranks attained by the same alternatives is equal to 10 (for *a*, *c*, *e*, *g*, *h*, and *l* it differs by 1 rank; for *i* and *j* it differs by 2 ranks), whereas the maximal possible difference of ranks for this problem size is 98; thus, $RDM = 1-10/98=0.898$;
- 6 (*b*, *d*, *f*, *k*, *m*, and *n*) out of 14 alternatives are ranked exactly the same; thus, $RAM = 6/14=0.429$.

Table B7
Efficacy measures for comparison of the recommendation delivered by the ranking methods and the DM's true value function (the rank attained by each procedure in terms of a given efficacy measure is provided in the brackets).

Procedure	HR	NHR	Kendall's τ	RDM	RAM
$\succsim^1(A, NEC, MD)$	1.000 (1)	0.091 (18)	0.363 (17)	0.565 (17)	0.143 (16)
$\succsim^1(A, NEC, mD)$	1.000 (1)	0.333 (14)	0.319 (19)	0.531 (19)	0.113 (19)
$\succsim^1(A, NEC, SD)$	1.000 (1)	1.000 (1)	0.758 (5)	0.806 (5)	0.321 (5)
$\succsim^i(A, NEC, MD)$	1.000 (1)	0.333 (14)	0.659 (14)	0.732 (14)	0.162 (15)
$\succsim^i(A, NEC, mD)$	1.000 (1)	0.333 (14)	0.626 (15)	0.721 (15)	0.215 (10)
$\succsim^i(A, NEC, SD)$	1.000 (1)	1.000 (1)	0.758 (5)	0.806 (5)	0.321 (5)
$\succsim^1(A, POI, MD)$	1.000 (1)	0.091 (18)	0.374 (16)	0.571 (16)	0.143 (16)
$\succsim^1(A, POI, mD)$	1.000 (1)	1.000 (1)	0.352 (18)	0.551 (18)	0.136 (18)
$\succsim^1(A, POI, SD)$	1.000 (1)	1.000 (1)	0.890 (1)	0.898 (1)	0.429 (1)
$\succsim^i(A, POI, MD)$	1.000 (1)	1.000 (1)	0.802 (4)	0.837 (4)	0.214 (11)
$\succsim^i(A, POI, mD)$	1.000 (1)	1.000 (1)	0.714 (10)	0.775 (12)	0.214 (11)
$\succsim^i(A, POI, SD)$	1.000 (1)	1.000 (1)	0.890 (1)	0.898 (1)	0.429 (1)
$\succsim^1(A, RAI, E)$	1.000 (1)	1.000 (1)	0.890 (1)	0.898 (1)	0.429 (1)
$\succsim^1(A, RAI, B)$	1.000 (1)	1.000 (1)	0.736 (7)	0.776 (10)	0.286 (8)
$\succsim^1(A, RAI, W)$	1.000 (1)	1.000 (1)	0.714 (10)	0.776 (10)	0.214 (11)
$\succsim^1(A, [P^*, P_*], B)$	1.000 (1)	0.333 (14)	0.725 (9)	0.781 (9)	0.274 (9)
$\succsim^1(A, [P^*, P_*], W)$	1.000 (1)	1.000 (1)	0.692 (13)	0.762 (13)	0.214 (11)
$\succsim^1(A, [P^*, P_*], BW)$	1.000 (1)	1.000 (1)	0.736 (7)	0.806 (5)	0.393 (4)
$\succsim^1(A, [P^*, P_*], WB)$	1.000 (1)	1.000 (1)	0.714 (10)	0.785 (8)	0.321 (7)

Table C1
Exemplary average efficacy measures for $\succsim^1(A, RAI, E)$ illustrating the influence of different problem setting on the consistency with the recommendation derived from the DM's "true" value function (remark that for this procedure $HR=NHR$). The settings are distinguished with the following quadruples: [type of marginal value functions $\in\{L, G\}$, number of criteria $\in\{3, 4, 6, 8\}$, number of alternatives $\in\{4, 6, 8, 10, 15\}$, number of pairwise comparisons $\in\{2, 4, 6, 8, 10, 14, 18\}$].

Increase of the number of criteria					Increase of the number of alternatives				
$[L, m, 10, 4]$	$m=3$	$m=4$	$m=6$	$m=8$	$[L, 4, n, 8]$	$n=6$	$n=8$	$n=10$	$n=15$
NHR	0.738	0.709	0.650	0.646	*HR*	0.883	0.785	0.766	0.704
τ	0.718	0.661	0.606	0.569	τ	0.841	0.767	0.750	0.729
RDM	0.802	0.774	0.750	0.733	*RDM*	0.910	0.849	0.832	0.808
RAM	0.417	0.363	0.326	0.314	*RAM*	0.755	0.550	0.455	0.312

Increase of the number of pairwise comparisons					Increase of the preference model's flexibility				
$[L, 4, 10, r]$	$r=4$	$r=6$	$r=8$	$r=10$		$[L, 4, 10, 8]$	$[G, 4, 10, 8]$	$[L, 8, 6, 6]$	$[G, 8, 6, 6]$
NHR	0.709	0.754	0.766	0.826	*NHR*	0.766	0.699	0.849	0.793
τ	0.661	0.715	0.750	0.787	τ	0.750	0.650	0.729	0.691
RDM	0.774	0.807	0.832	0.854	*RDM*	0.832	0.776	0.839	0.814
RAM	0.363	0.403	0.455	0.499	*RAM*	0.455	0.357	0.604	0.556

Appendix C. Remarks on the influence of different problem settings on the efficacy measures

The analysis of the detailed results for all considered problem settings indicates that on average the consistency with the recommendation derived from the DM's "true" value function:

- increases when more pairwise comparisons are provided by the DM,
- decreases when the number of criteria or alternatives is growing, or when admitting greater flexibility of the assumed preference model (i.e., when moving from linear to general value functions).

Table C1 provides the exemplary efficacy measures derived from the e-Appendix that illustrate these phenomenons for the selected scoring procedure and problem settings.

Appendix D. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.cor.2016.01.007.

References

[1] Ahn B, Park K. Comparing methods for multiattribute decision making with ordinal weights. Comput Oper Res 2008;35(5):1660–70.
[2] Angilella S, Corrente S, Greco S. Stochastic multiobjective acceptability analysis for the Choquet integral preference model and the scale construction problem. Eur J Oper Res 2015;240(1):172–82.
[3] Barrett C, Pattanaik P, Salles M. On choosing rationally when preferences are fuzzy. Fuzzy Sets Syst 1990;34:197–212.
[4] Barron F, Barrett B. Decision quality using ranked attribute weights. Manag Sci 1996;42:1515–23.
[5] Beuthe M, Scannella G. Comparative analysis of UTA multicriteria methods. Eur J Oper Res 2001;130(2):246–62.
[6] Bous G, Fortemps P, Glineur F, Pirlot M. ACUTA: A novel method for eliciting additive value functions on the basis of holistic preference statements. Eur J Oper Res 2010;206(2):435–44.
[7] Bouyssou D, Marchant T, Pirlot M, Perny P, Tsoukias A, Vincke P. Evaluation and decision models: a critical perspective. Kluwer: Kluwer's International Series; 2000.
[8] Branke J, Corrente S, Greco S, Gutjahr W. Using indifference information in robust ordinal regression. In: Gaspar-Cunha A, Henggeler Antunes C, Coello C, editors. Evolutionary multi-criterion optimization. Lecture notes in computer science, vol. 9019. Switzerland: Springer International Publishing; 2015. p. 205–17.
[9] Corrente S, Greco S, Kadziński M, Słowiński R. Robust ordinal regression in preference learning and ranking. Mach Learn 2013;93(2–3):381–422.
[10] Dias L, Climaco J. Additive aggregation with variable interdependent parameters: the VIP analysis software. J Oper Res Soc 2000;51:1070–82.

[11] Doumpos M, Zopounidis C. Regularized estimation for preference disaggregation in multiple criteria decision making. Comput Optim Appl 2007;38(1):61–80.
[12] Doumpos M, Zopounidis C. The robustness concern in preference disaggregation approaches for decision aiding: an overview. In: Rassias TM, Floudas CA, Butenko S, editors. Optimization in science and engineering. New York: Springer; 2014. p. 157–77.
[13] Doumpos M, Zopounidis C, Galariotis E. Inferring robust decision models in multicriteria classification problems: an experimental analysis. Eur J Oper Res 2014;236(2):601–11.
[14] Durbach I, Lahdelma R, Salminen P. The analytic hierarchy process with stochastic judgements. Eur J Oper Res 2014;238(2):552–9.
[15] Figueira J, Greco S, Słowiński R. Building a set of additive value functions representing a reference preorder and intensities of preference: GRIP method. Eur J Oper Res 2009;195(2):460–86.
[16] Greco S, Mousseau V, Słowiński R. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. Eur J Oper Res 2008;191(2):415–35.
[17] Hazen G. Partial information, dominance, and potential optimality in multiattribute utility theory. Oper Res 1986;34(2):296–310.
[18] Jaccard P. The distribution of the flora in the Alpine zone. New Phytol 1912; 11(2):37–50.
[19] Jacquet-Lagrèze E, Siskos Y. Preference disaggregation: 20 years of MCDA experience. Eur J Oper Res 2001;130(2):233–45.
[20] Kadziński M, Ciomek K, Słowiński R. Modeling assignment-based pairwise comparisons within integrated framework for value-driven multiple criteria sorting. Eur J Oper Res 2015;241(3):830–41.
[21] Kadziński M, Corrente S, Greco S, Słowiński R. Preferential reducts and constructs in robust multiple criteria ranking and sorting. OR Spectr 2014;36 (4):1021–53.
[22] Kadziński M, Greco S, Słowiński R. Extreme ranking analysis in robust ordinal regression. Omega 2012;40(4):488–501.
[23] Kadziński M, Greco S, Słowiński R. Selection of a representative value function in robust multiple criteria ranking and choice. Eur J Oper Res 2012;217(3):541–53.
[24] Kadziński M, Greco S, Słowiński R. RUTA: a framework for assessing and selecting additive value functions on the basis of rank related requirements. Omega 2013;41(4):735–51.
[25] Kadziński M, Słowiński R, Greco S. Multiple criteria ranking and choice with all compatible minimal cover sets of decision rules. Knowl-Based Syst 2015;89:569–83.
[26] Kadziński M, Tervonen T. Robust multi-criteria ranking with additive value models and holistic pair-wise preference statements. Eur J Oper Res 2013;228(1):169–80.
[27] Klamler C. On the closeness aspect of three voting rules: Borda – Copeland – Maximin. Group Decis Negot 2005;14(3):233–40.
[28] Lahdelma R, Salminen P. SMAA-2: stochastic multicriteria acceptability analysis for group decision making. Oper Res 2001;49(3):444–54.
[29] Lahdelma R, Salminen P. The shape of the utility or value function in stochastic multicriteria acceptability analysis. OR Spectr 2012;34(4):785–802.
[30] Lee KS, Park KS, Kim SH. Dominance, potential optimality, imprecise information, and hierarchical structure in multi-criteria analysis. Comput Oper Res 2002;29(9):1267–81.
[31] Leskinen P, Viitanen J, Kangas A, Kangas J. Alternatives to incorporate uncertainty and risk attitude in multicriteria evaluation of forest plans. For Sci 2006;52(3):304–14.
[32] Malakooti B. Ranking and screening multiple criteria alternatives with partial information and use of ordinal and cardinal strength of preferences. IEEE Trans Syst Man Cybern – Part A: Syst Hum 2000;30:355–68.
[33] Sarabando P, Dias LC. Multiattribute choice with ordinal information: a comparison of different decision rules. IEEE Trans Syst Man Cybern Part A: Syst Hum 2009;32(3):545–54.
[34] Sarabando P, Dias LC. Simple procedures of choice in multicriteria problems without precise information about the alternatives' values. Comput Oper Res 2010;37(12):2239–47.
[35] Simpson P. On defining areas of voter choice: Professor Tullock on stable voting. Q J Econ 1969;83(3):478–90.
[36] Siskos Y. A way to deal with fuzzy preferences in multicriteria decision problems. Eur J Oper Res 1982;10(3):314–24.
[37] Siskos Y, Yanacopoulos D. UTA STAR – an ordinal regression method for building additive value functions. Investig Oper 1985;5:39–53.
[38] Szelag M, Greco S, Słowiński R. Variable consistency dominance-based rough set approach to preference learning in multicriteria ranking. Inf Sci 2014;277:525–52.
[39] Tervonen T. JSMAA: open source software for SMAA computations. Int J Syst Sci 2014;45(1):69–81.
[40] Tervonen T, Lahdelma R. Implementing stochastic multicriteria acceptability analysis. Eur J Oper Res 2007;178(2):500–13.
[41] Tervonen T, Linkov I, Steevens J, Chappell M, Figueira JR, Merad M. Risk-based classification system of nanomaterials. J Nanopart Res 2009;11(4):757–66.
[42] Tervonen T, van Valkenhoef G, Basturk N, Postmus D. Hit-And-Run enables efficient weight generation for simulation-based multiple criteria decision analysis. Eur J Oper Res 2013;224(3):552–9.
[43] van Valkenhoef G, Tervonen T, Postmus D. Notes on Hit-And-Run enables efficient weight generation for simulation-based multiple criteria decision analysis. Eur J Oper Res 2014;239(3):865–7.
[44] Vetschera R. Deriving rankings from incomplete preference information: a comparison of different approaches. In: 20th conference of the international federation of operational research societies, Barcelona, Spain; 2014.
[45] Vetschera R. Indifference in volume-based methods for decision making under incomplete information. In: 27th European conference on operational research, Glasgow, United Kingdom; 2015.
[46] Vetschera R, Sarabando P, Dias L. Levels of incomplete information in group decision models – a comprehensive simulation study. Comput Oper Res 2014;51:160–71.
[47] White C, Sage A, Dozono S. A model of multiattribute decision making and trade-off weight determination under uncertainty. IEEE Trans Syst Man Cybern 1984;14(2):223–9.
[48] Winkler R, Hay W. Statistics: probability, inference, and decision. New York: Rinehart & Winston; 1985.