Comput Optim Appl (2007) 38: 61–80
DOI 10.1007/s10589-007-9037-9

Regularized estimation for preference disaggregation in multiple criteria decision making

Michael Doumpos · Constantin Zopounidis

Published online: 17 May 2007

Abstract Disaggregation methods have been extensively used in multiple criteria decision making to infer preferential information from reference examples, using linear programming techniques. This paper proposes simple extensions of existing formulations, based on the concept of regularization which has been introduced within the context of the statistical learning theory. The properties of the resulting new formulations are analyzed for both ranking and classification problems and experimental results are presented demonstrating the improved performance of the proposed formulations over the ones traditionally used in preference disaggregation analysis.

Keywords Multiple criteria decision making · Preference disaggregation · Regularization · Linear programming

1 Introduction

The implementation of several multiple criteria decision making (MCDM) methods requires the decision maker (DM) to explicitly define a considerable amount of specific preferential information, such as the relative importance of the criteria, preference, indifference thresholds, etc. Obtaining such information from the DM is not an easy task.

Preference disaggregation analysis (PDA) has been extensively used over the past two decades to resolve this difficulty [6]. Instead of asking the DM to provide details on his/her preferential system, PDA employs a regression-like process to infer the

M. Doumpos (✉) · C. Zopounidis
Technical University of Crete, Dept. of Production Engineering and Management,
Financial Engineering Laboratory, University Campus, 73100 Chania, Greece
e-mail: mdoumpos@dpem.tuc.gr

C. Zopounidis
e-mail: kostas@dpem.tuc.gr

required information through the analysis of the DM's judgments on some reference alternatives. The DM provides a global evaluation of these alternatives, usually expressed either by ranking them from the most preferred to the least preferred ones, or by assigning them to preference classes. Such a global evaluation is implicitly based on the preferential system of the DM. Thus, the identification of the criteria aggregation model that best fits the DM's global judgments on the reference alternatives should be equivalent to the direct specification of detailed preferential information by the DM.

This PDA process is formulated as an optimization problem. Given the general form of a criteria aggregation model $f(\mathcal{A})$ defined by a set of parameters $\mathcal{A}$, the objective is to identify the optimal values for the parameters in $\mathcal{A}$ that minimize the observed deviations between the model's outputs and the DM's global evaluation of the reference alternatives. For instance, in the case where f is a value function, the set of parameters $\mathcal{A}$ includes the trade-off coefficients for the criteria and the corresponding marginal value functions [7]. In an outranking relation model, $\mathcal{A}$ may involve the criteria weights, preference, indifference and veto thresholds, etc. [11].

Several methodologies implementing this framework have been proposed for different kinds of problem formulations and MCDM methods. In the case of value functions, the UTA method has been proposed for ranking problems [5], with several extensions in classification problems [13]. In the case of outranking relations models such techniques have been proposed in [9, 10] for classification problems. An extensive review of existing PDA techniques can be found in [6].

However, since the implementation of a PDA framework is based on an optimization process, it is important to understand its limitations as far as it concerns the inferences that can be made on the basis of the resulting criteria aggregation model. Of course, a model that poorly fits the DM's global judgment on the reference alternatives cannot be accepted as a basis of the decision aiding process. The opposite, however, is not generally true. The quality of the model can only be ascertained by validating its performance on new cases, others than the ones in the reference set. Furthermore, the quality of the model should be assessed in terms of its structural parameters. This involves an analysis of whether the estimated parameters are in accordance with the DM's preferential system. Clearly, the design of PDA procedures that meet these requirements is of major importance in order to have a solid basis for implementing the decision aiding process, which follows model development.

The objective of this paper is to introduce regularized estimation techniques into existing optimization formulations which have been proposed within the context of PDA. The concept of regularization has been introduced in statistical learning theory as a methodology to address the trade-off between model complexity and generalizing performance [4]. The properties of the new formulations developed for ranking and classification problems are analyzed providing some insight on the relationship between the data characteristics and the quality of the resulting models. Experimental results are also given, demonstrating the improved performance of the new formulations over the existing ones. The analysis is given for value function models, but the results have implications for other MCDM aggregation forms too.

The rest of the paper is organized as follows. Section 2 introduces the main concepts of PDA and the methodologies used for developing additive value models for

ranking and classification problems. Section 3, presents the modification of the existing methodologies on the basis of the regularization principle and discusses the properties of the new formulations. Section 4, presents some computational results on the performance of the new formulations, and finally Sect. 5 concludes the paper and discusses some future research directions.

2 Disaggregation methods in MCDM

The most widely used form of criteria aggregation model in MCDM is the additive value function. Assuming that K criteria $x_1, x_2, \ldots, x_K$ are used in a multicriteria evaluation context, the global value (overall performance) of an alternative $\mathbf{x}_i$, is defined as:

$$V(\mathbf{x}_i) = \sum_{k=1}^{K} p_k v_k(x_{ik}), \tag{1}$$

where x_{ik} denotes the description of alternative $\mathbf{x}_i$ on criterion x_k, $p_k \geq 0$ is the trade-off coefficient for criterion x_k (the normalization $p_1 + p_2 + \cdots + p_K = 1$ is often used) and v_k is the corresponding marginal value function. The marginal value functions define a monotone mapping of each criterion to a value scale (usually in $[0, 1]$) such that:

$$\left.\begin{array}{lll} v_k(x_{ik}) > v_k(x_{jk}) & \Leftrightarrow & \mathbf{x}_i \succ_k \mathbf{x}_j \\ v_k(x_{ik}) = v_k(x_{jk}) & \Leftrightarrow & \mathbf{x}_i \sim_k \mathbf{x}_j \\ v_k(x_{*k}) = 0, & & v_k(x_k^*) = 1 \end{array}\right\}, \tag{2}$$

where $\succ_k$ and $\sim_k$ denote the preference and indifference relations for criterion x_k, and x_{*k}, x_k^* are the least and most preferred values of criterion x_k, respectively.

In such an evaluation context an alternative $\mathbf{x}_i$ is preferred to an alternative $\mathbf{x}_j$ if and only if $V(\mathbf{x}_i) > V(\mathbf{x}_j)$ and the alternatives are indifferent if and only if $V(\mathbf{x}_i) = V(\mathbf{x}_k)$.

Setting $u_k = w_k v_k$, the additive function (1) can be equivalently written in a simpler form as follows:

$$V(\mathbf{x}_i) = \sum_{k=1}^{K} u_k(x_{ik}). \tag{3}$$

In this additive model the marginal value functions are normalized such that $u_k(x_{*k}) = 0$ and $u_k(x_k^*) = p_k$. Such an additive model can be linear if the marginal value functions have a linear form, or non-linear if non-linearity is assumed for the marginal value functions.

The development of the additive model (3) involves the specification of the marginal value functions. Direct procedures such as the mid-value point technique [7] can be used for this purpose, but their implementation is often limited due to significant cognitive effort required by the DM. Alternatively, PDA methodologies are applicable. In this case, it is assumed that the DM provides a global evaluation for a sample of M reference alternatives. In the case of ranking problems this requires a ranking of

the reference alternatives from the most to the least preferred ones, whereas in classification problems an assignment of the alternatives in predefined preference classes is defined.

In the ranking case, without loss of generality, it can be assumed that the reference alternatives are ranked from the best $\mathbf{x}_1$ to the worst $\mathbf{x}_M$, into N indifference classes $I_1 \succ I_2 \succ \cdots \succ I_N$. The m_n alternatives in I_n are indifferent to each other and they are preferred over the alternatives in I_{n+1}. The model that best fits the DM's ranking on a set A of reference alternatives can be constructed from the solution of a mathematical programming problem of the following general form [5]:

$$\begin{aligned} \min \quad & \sum_{i=1}^{M} y_i \\ \text{subject to:} \quad & V(\mathbf{x}_i) - V(\mathbf{x}_{i+1}) + y_i - y_{i+1} \geq \delta, \quad \text{if } \mathbf{x}_i \succ \mathbf{x}_{i+1}, \\ & V(\mathbf{x}_i) - V(\mathbf{x}_{i+1}) + y_i - y_{i+1} = 0, \quad \text{if } \mathbf{x}_i \sim \mathbf{x}_{i+1}, \\ & V(\mathbf{x}_*) = 0, \qquad V(\mathbf{x}^*) = 1, \\ & y_i \geq 0, \quad \forall i = 1, 2, \ldots, M. \end{aligned} \tag{4}$$

The first constraint applies to pairs of reference alternatives $(\mathbf{x}_i, \mathbf{x}_{i+1})$ for which $\mathbf{x}_i \succ \mathbf{x}_{i+1}$ ($0 < \delta \ll 1$ is a constant used to impose the strict inequality $V(\mathbf{x}_i) > V(\mathbf{x}_{i+1})$). The second constraint applies to pairs of reference alternatives $(\mathbf{x}_i, \mathbf{x}_{i+1})$ for which $\mathbf{x}_i \sim \mathbf{x}_{i+1}$. In both cases, the error variables y represent the deviations between the model's results and the predefined ranking of the reference alternatives. Normalization constraints are also imposed, as well as constraints to ensure the monotonicity of the marginal value functions.

In solving this optimization problem the marginal value functions are assumed to be piecewise linear. This involves the definition of $b_k + 1$ subintervals $[\beta_0^k, \beta_1^k], [\beta_1^k, \beta_2^k], \ldots, [\beta_{b_k}^k, \beta_{b_k+1}^k]$ in the scale of criterion x_k (with $\beta_0^k = x_{*k}$ and $\beta_{b_k+1}^k = x_k^*$). The marginal value at a point β_t^k of criterion x_k is then:

$$u_k(\beta_t^k) = \sum_{s=1}^{t} \left[u_k(\beta_s^k) - u_k(\beta_{s-1}^k) \right] = \sum_{s=1}^{t} d_{sk} \tag{5}$$

and the marginal value $u_k(x_{ik})$ for an alternative $\mathbf{x}_i$ with $x_{ik} \in [\beta_{t-1}^k, \beta_t^k]$ is defined by a linear interpolation between $u_k(\beta_{t-1}^k)$ and $u_k(\beta_t^k)$:

$$u_k(x_{ik}) = \mathbf{w}_{ik}\mathbf{d}_k, \tag{6}$$

where $\mathbf{d}_k = (d_{1k}, d_{2k}, \ldots, d_{b_k k})^\top$ and $\mathbf{w}_{ik}$ is a row vector whose entry w_{ik}^t ($t = 1, 2, \ldots, b_k$) is defined as:

$$w_{ik}^t = \begin{cases} 0 & \text{if } x_{ik} < \beta_{t-1}^k \,, \\ \frac{x_{ik} - \beta_{t-1}^k}{\beta_t^k - \beta_{t-1}^k} & \text{if } x_{ik} \in [\beta_{t-1}^k, \beta_t^k] \,, \\ 1 & \text{if } x_{ik} > \beta_t^k \,. \end{cases} \tag{7}$$

Therefore, denoting by $\mathbf{w}_i$ the row vector $[\mathbf{w}_{i1}\ \mathbf{w}_{i2}\ \ldots\ \mathbf{w}_{iK}]$ and by $\mathbf{d}$ the column vector $[\mathbf{d}_1\ \mathbf{d}_2\ \ldots\ \mathbf{d}_K]^\top$, the global value of alternative $\mathbf{x}_i$ according to (3) is written as:

$$V(\mathbf{x}_i) = \mathbf{w}_i \mathbf{d}. \tag{8}$$

With this expression, (4) can now be re-written as the following linear program:

$$\begin{aligned} \min \quad & \mathbf{e}_1^\top \mathbf{y} \\ \text{subject to:} \quad & (\mathbf{w}_i - \mathbf{w}_{i+1})\mathbf{d} + y_i - y_{i+1} \geq \delta, \quad \text{if } \mathbf{x}_i \succ \mathbf{x}_{i+1}, \\ & (\mathbf{w}_i - \mathbf{w}_{i+1})\mathbf{d} + y_i - y_{i+1} = 0, \quad \text{if } \mathbf{x}_i \sim \mathbf{x}_{i+1}, \\ & \mathbf{e}^\top \mathbf{d} = 1, \\ & \mathbf{d}, \mathbf{y} \geq \mathbf{0}, \end{aligned} \tag{9}$$

where $\mathbf{e}$ and $\mathbf{e}_1$ are column vectors of ones. The first two constraints are equivalent to the first two constraints of problem (4). The constraint $\mathbf{e}^\top \mathbf{d} = 1$ normalizes the developed additive model between 0 and 1, whereas the non-negativity constraint $\mathbf{d} \geq \mathbf{0}$ ensures the monotonicity of the marginal value functions.

In the case of classification problems, the reference alternatives are classified in N ordered classes $c_1 \succ c_2 \succ \cdots \succ c_N$. An alternative is classified in class c_n if and only if $V(\mathbf{x}_i) \in (h_n, h_{n-1})$, where $h_1 > h_2 > \cdots > h_N > 0$ are value thresholds that distinguish the classes such that $h_{n-1} - h_n \geq s$, with $s \geq \delta$ a user-defined positive constant. Obviously there is no need to define h_0 and h_N because $V(\mathbf{x}_i) \in [0, 1]$. In this case the linear programming problem (9) is transformed as follows [2]:

$$\begin{aligned} \min \quad & \mathbf{e}_1^\top (\mathbf{y}^+ + \mathbf{y}^-) \\ \text{subject to:} \quad & \mathbf{w}_i \mathbf{d} - h_n + y_i^+ \geq \delta, && \forall \mathbf{x}_i \in \{c_1, \ldots, c_{N-1}\}, \\ & \mathbf{w}_i \mathbf{d} - h_{n-1} - y_i^- \leq -\delta, && \forall \mathbf{x}_i \in \{c_2, \ldots, c_N\}, \\ & h_{n-1} - h_n \geq s, && \forall n = 2, \ldots, N-1, \\ & \mathbf{e}^\top \mathbf{d} = 1, \\ & \mathbf{d}, \mathbf{h}, \mathbf{y}^+, \mathbf{y}^- \geq \mathbf{0}. \end{aligned} \tag{10}$$

The first constraint defines the violations of the lower bound of each class (this applies only to the reference alternatives that belong to the classes $c_1, \ldots, c_{N-1}$), whereas the second constraint defines the violations of the upper bound of each class (this applies only to the reference alternatives that belong to the classes $c_2, \ldots, c_N$).

In both linear programs (9) and (10) post-optimality techniques are used to explore the existence of multiple optimal or near optimal solutions as described in [2, 5].

3 Regularization for preference disaggregation

3.1 General framework

Regularization is a well-known principle in statistical learning with many applications mainly in neural networks and support vector machines [8, 12]. Whereas the

traditional model development approach is based solely on the fit of the model on some training data, regularization also considers the simplicity of the model. Models that use too many parameters are highly complex and overfitting is possible. On the other hand, models that use a small number of parameters exhibit a more smooth behavior, thus (often) providing better generalizing performance.

On the basis of this framework, in a regularization context, the development of a model f is based on the minimization of the following general regularized loss function:

$$\mathcal{L} = J(f) + \lambda \sum_{i=1}^{M} L\big(\hat{f}(\mathbf{x}_i), f(\mathbf{x}_i)\big) \tag{11}$$

where J is a penalty function used to stabilize (smooth/simplify) the model f, L is a loss function measuring the deviation between the actual and estimated result for alternative $\mathbf{x}_i$ (denoted by $f(\mathbf{x}_i)$ and $\hat{f}(\mathbf{x}_i)$, respectively) and $\lambda > 0$ is a constant representing the trade-off between model fit and complexity. For instance, for a linear regression model $f(\mathbf{x}_i) = \mathbf{x}_i\mathbf{d}$, the regularized function (11) can be expressed as follows:

$$\mathcal{L} = \|\mathbf{d}\|^2 + \lambda \sum_{i=1}^{M} \big(\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)\big)^2. \tag{12}$$

3.2 Ranking

Adopting this approach with the L_1 norm $|\cdot|$, instead of the L_2 norm $\|\cdot\|$, the linear program (9) in the ordinal regression case is transformed as follows:

$$\begin{aligned} \min \quad & \mathbf{e}^\top\mathbf{d} + \lambda\mathbf{e}_1^\top\mathbf{y} \\ \text{subject to:} \quad & (\mathbf{w}_i - \mathbf{w}_{i+1})\mathbf{d} + y_i - y_{i+1} \geq \delta, \quad \text{if } \mathbf{x}_i \succ \mathbf{x}_{i+1}, \\ & (\mathbf{w}_i - \mathbf{w}_{i+1})\mathbf{d} + y_i - y_{i+1} = 0, \quad \text{if } \mathbf{x}_i \sim \mathbf{x}_{i+1}, \\ & \mathbf{d}, \mathbf{y} \geq \mathbf{0}. \end{aligned} \tag{13}$$

Compared to the linear program (9) in the above formulation the normalization $\mathbf{e}^\top\mathbf{d} = 1$ is no longer used as a constraint. Instead, the term $\mathbf{e}^\top\mathbf{d}$ is introduced in the objective function which is expressed in a similar form to the regularized loss function (12).

The removal of the normalization constraint makes the parameter δ a simple scaling constant that does not affect the resulting model. Assuming that $(\mathbf{d}^*, \mathbf{y}^*)$ is the optimal solution of (13) for some δ, then for any other value δ' the optimal solution is simply scaled by a factor of δ'/δ.

Furthermore, with the removal of the normalization constraint $\mathbf{e}^\top\mathbf{d} = 1$, the optimal solution $(\mathbf{d}^*, \mathbf{y}^*)$ normalizes the resulting model in $[0, \mathbf{e}^\top\mathbf{d}^*]$. However, given the above remark on the scaling constant δ, it is easy to see that setting $\delta' = \delta/\mathbf{e}^\top\mathbf{d}^*$ leads to a model normalized in $[0, 1]$. This shows that the resulting rescaled optimal solution of (13) is also optimal to (9) when δ' is used instead of δ.

However, with the removal of the normalization constraint $\mathbf{e}^\top\mathbf{d} = 1$, it is possible that the trivial solution $\mathbf{d} = \mathbf{0}$ will be optimal for (13). The following theorem characterizes the conditions under which this is possible and provides some guidelines for

the proper selection of the parameter λ ($\mathbf{P}$ is the matrix with the differences $\mathbf{w}_i - \mathbf{w}_{i+1}$ for pairs of alternatives $\mathbf{x}_i \succ \mathbf{x}_{i+1}$, and $\mathbf{I}_n$ is the matrix with the differences $\mathbf{w}_i - \mathbf{w}_{i+1}$ between the alternatives in the indifference class I_n).

Theorem 1 *The trivial solution* $\mathbf{d} = \mathbf{0}$ *is optimal to* (13) *if and only if*:

$$\lambda \left[\mathbf{P}^\top \mathbf{m} + \sum_{n=1}^{N-1} \mathbf{I}_n^\top \left(\mathbf{m}'_n + \mathbf{e}_n \left(\sum_{t=1}^{n-1} m_t \right) \right) \right] - \mathbf{I}_N^\top \mathbf{z} \leq \mathbf{e}, \tag{14}$$

where $\mathbf{m} = (m_1, m_1 + m_2, \ldots, m_1 + \cdots + m_{N-1})^\top$, $\mathbf{m}'_n = (1, 2, \ldots, m_n - 1)^\top$ *and* $\mathbf{z} = (z_1, z_2, \ldots, z_{m_N - 1})^\top$ *with* $z_1 \geq -\lambda(M - m_N + 1)$, $z_{m_N - 1} \leq \lambda$ *and* $z_i - z_{i-1} \geq -\lambda$ *for all other* i.

Proof Without loss of generality, assume that the $M-1$ constraints in (13) are written in the same order with the ranking of the alternatives.

At the solution $\mathbf{d} = \mathbf{0}$, $y_i = 0$ for all $\mathbf{x}_i \in I_N$, $y_i = \delta$ for all $\mathbf{x}_i \in I_{N-1}$, $y_i = 2\delta$ for all $\mathbf{x}_i \in I_{N-2}$, etc. Therefore, for the solution $\mathbf{d} = \mathbf{0}$ the optimal objective function value is:

$$\lambda \delta \sum_{n=1}^{N} (N - n) m_n. \tag{15}$$

Let $\mathbf{a}$ be the dual variables associated to the constraints of (13), and denote by $\mathbf{a}_P$ the members of $\mathbf{a}$ associated to the primal constraints involving pairs of alternatives $\mathbf{x}_i \succ \mathbf{x}_{i+1}$ (i.e., dual variables $a_{m_1}, a_{m_1+m_2}, \ldots, a_{M-m_N}$). The remaining dual variables $\mathbf{a}_I$ are associated to the primal constraints involving pairs of alternatives $\mathbf{x}_i \sim \mathbf{x}_{i+1}$. Denoting by $\mathbf{I}$ the matrix of the difference $\mathbf{w}_i - \mathbf{w}_{i+1}$ for pairs of indifferent alternatives $\mathbf{x}_i \sim \mathbf{x}_{i+1}$, the optimality conditions are then written as follows (with $a_0 = a_M = 0$):

$$-\delta \mathbf{e}_P^\top \mathbf{a}_P = \lambda \delta \sum_{n=1}^{N} (N - n) m_n, \tag{16}$$

$$-\mathbf{P}^\top \mathbf{a}_P - \mathbf{I}^\top \mathbf{a}_I \leq \mathbf{e}, \tag{17}$$

$$a_{i-1} - a_i \leq \lambda, \quad i = 1, \ldots, M, \tag{18}$$

$$\mathbf{a}_P \geq 0, \quad \mathbf{a}_I \in \mathbb{R}. \tag{19}$$

At the solution $\mathbf{d} = \mathbf{0}$, $y_i = \delta(N-1) > 0$ for all $\mathbf{x}_i \in I_1$ and from complementary slackness it follows that $a_{i-1} - a_i = \lambda$ for all $i = 1, \ldots, m_1$ (with $a_0 = 0$); thus, $a_{m_1} = -\lambda m_1$. Similarly, $y_i = \delta(N-2) > 0$ for all $\mathbf{x}_i \in I_2$. Therefore, from complementary slackness it follows that $a_{i-1} - a_i = \lambda$ for all $i = m_1 + 1, \ldots, m_1 + m_2$, thus leading to $a_{m_1+m_2} = -\lambda(m_1 + m_2)$.

With these results it is easy to see that $a_i = -i\lambda$ ($1 \leq i \leq M - m_N$) and $a_{i-1} - a_i \leq \lambda$ ($M - m_N < i \leq M$, $a_M = 0$), satisfy (16), (18) and (19). Plugging this solution into (17) and setting $\mathbf{z} = (a_{M-m_n+1}, \ldots, a_{M-1})^\top$ leads to (14). □

The above theorem shows that the trivial solution may occur if λ is not set properly (e.g., too low value) or if the data used to develop the model are inappropriate. In the absence of ties (indifferent alternatives), condition (14) simplifies to $\lambda \mathbf{P}^\top \mathbf{m} \leq \mathbf{e}$ which will always hold for every $\lambda \geq 0$ if $\mathbf{P} \leq \mathbf{0}$.

In general, the maximum value of λ that will lead to the trivial solution can be identified from the solution of the following simple LP:

$$\begin{aligned}
\max \quad & \theta_r = \lambda\delta \sum_{n=1}^{N} (N-n) m_n \\
\text{s.t.} \quad & \lambda \left[\mathbf{P}^\top \mathbf{m} + \sum_{n=1}^{N-1} \mathbf{I}_n^\top \left(\mathbf{m}'_n + \mathbf{e}_n \left(\sum_{t=1}^{n-1} m_t \right) \right) \right] - \mathbf{I}_N^\top \mathbf{z} \leq \mathbf{e}, \\
& -z_1 \leq \lambda (M - m_N + 1), \\
& z_{m_N - 1} \leq \lambda, \\
& z_{i-1} - z_i \leq \lambda, \quad i = 2, \ldots, m_N - 1, \\
& \lambda \geq 0, \qquad \mathbf{z} \in \mathbb{R}.
\end{aligned} \tag{20}$$

The objective function of this LP problem corresponds to the value of the objective function of (13) for the trivial solution $\mathbf{d} = \mathbf{0}$, whereas the constraints are the necessary and sufficient conditions imposed by theorem 1 for the existence of the trivial solution.

It should be noticed that (20) is always feasible because the trivial solution $\lambda = 0$ and $\mathbf{z} = \mathbf{0}$ is feasible. However, (20) can be unbounded. In this case irrespectively of the value used for λ in (13), the trivial solution will always be optimal. The unboudedness of (20) is defined solely from the first constraint that determines an upper bound for λ. Obviously, if

$$\mathbf{P}^\top \mathbf{m} + \sum_{n=1}^{N-1} \mathbf{I}_n^\top \left(\mathbf{m}'_n + \mathbf{e}_n \left(\sum_{t=1}^{n-1} m_t \right) \right) \leq \mathbf{0} \tag{21}$$

then λ can be set arbitrarily large without affecting feasibility. In the case of a complete ranking of the reference alternatives with no indifferences this reduces to $\mathbf{P}^\top \mathbf{m} \leq \mathbf{0}$. Furthermore, when $m_N = 1$ then $\mathbf{I}_N = 0$ and (20) can be solved explicitly with the maximum λ taken as the minimum of all the positive elements of the left-hand side of (21).

It should also be noticed that the solution of (20) depends solely on the data of the reference alternatives and the definition of the subintervals for the assessment of the piecewise marginal value functions. Therefore, θ_r^* can be considered as a measure of the "quality" of the subintervals that are defined. A high value for θ_r^* indicates that the specified subintervals provide poor information on the ranking of the reference alternatives and vice versa. This is further demonstrated from the following lemma.

Lemma 1 *Suppose that* $(\mathbf{d}, \mathbf{y} = \mathbf{0})$ *solves* (13). *Then, the resulting model that perfectly fits the ranking of the reference alternatives has* $\mathbf{e}^\top \mathbf{d} \geq \theta_r^*$.

Proof Follows by contradiction. Solving (13) with $\lambda = \lambda^*$, where λ^* is taken from the solution of (20), leads to an optimal solution $(\mathbf{d} = \mathbf{0}, \mathbf{y})$ with objective function value equal to θ_r^*. If the solution $(\mathbf{d}, \mathbf{y} = \mathbf{0})$ is feasible to (13) for the selected λ, and $\mathbf{e}^\top \mathbf{d} < \theta_r^*$, then $(\mathbf{d} = \mathbf{0}, \mathbf{y})$ cannot be optimal, thus leading to a contradiction. □

According to this lemma, the higher the value of θ_r^* the larger the value of $\mathbf{e}^\top \mathbf{d}$ should be to get a model that perfectly fits the data. This would require to use a high value for λ. In the extreme case where (20) is unbounded (i.e., $\theta_r^* = \infty$), it is not possible to develop such a model.

3.3 Classification

The analysis of the previous section can also be extended in the context of classification problems. Denoting by $\mathbf{W}_n$ the matrix that consists of all $\mathbf{w}_i, \forall \mathbf{x}_i \in c_n$, the linear programming problem (10) is transformed as follows:

$$\begin{aligned}
\min \quad & \mathbf{e}^\top \mathbf{d} + \sum_{n=1}^{N-1} \lambda_n \mathbf{e}_n^\top \mathbf{y}_n^+ + \sum_{n=2}^{N} \lambda_n \mathbf{e}_n^\top \mathbf{y}_n^- \\
\text{s.t.} \quad & \mathbf{W}_n \mathbf{d} - \mathbf{e}_n h_n + \mathbf{y}_n^+ \geq \mathbf{e}_n \delta, && \forall n = 1, \ldots, N-1, \\
& \mathbf{W}_n \mathbf{d} - \mathbf{e}_n h_{n-1} - \mathbf{y}_n^- \leq -\mathbf{e}_n \delta, && \forall n = 2, \ldots, N, \\
& h_n - h_{n+1} \geq s, && \forall n = 1, \ldots, N-2, \\
& \mathbf{d}, \mathbf{h}, \mathbf{y}^+, \mathbf{y}^- \geq \mathbf{0}.
\end{aligned} \tag{22}$$

The remarks made earlier in the ranking case, on the role of δ as a scaling parameter as well as the relationship between problems (10) and (22) also apply in the classification context.

Denoting by m_n the number of reference alternatives in class c_n, the following theorem characterizes the cases under which the trivial solution $\mathbf{d} = \mathbf{0}$ is optimal for (22).

Theorem 2 *Let* $\sum_{n=1}^{N-1} \lambda_n m_n > \lambda_N m_N$. *Then, the trivial solution* $\mathbf{d} = \mathbf{0}$ *is optimal to* (22) *if and only if*:

$$\sum_{n=1}^{N-1} \lambda_n \mathbf{W}_n^\top \mathbf{e}_n - \lambda_N \mathbf{W}_N^\top \mathbf{e}_N \leq \mathbf{e}. \tag{23}$$

Proof First note that the condition $\sum_{n=1}^{N-1} \lambda_n m_n > \lambda_N m_N$ does not lead to any loss of generality, because of the symmetry in the definition of the classes. That is, reversing the ordering of the classes and the signs of the criteria does not alter the characteristics of the problem [1].

Let $\mathbf{b}_n$ $(n = 1, \ldots, N-1)$ be the dual variables associated to the first set of constraints in (22). Similarly, let $\mathbf{a}_n$ $(n = 2, \ldots, N)$ and $\mathbf{z}_n$ $(n = 1, \ldots, N-2)$ denote the dual variables associated with the second and the third sets of constraints in (22).

The dual of (22) is then expressed as

$$\begin{aligned}
\max \quad & \delta \sum_{n=1}^{N-1} \mathbf{e}_n^\top \mathbf{b}_n + \delta \sum_{n=2}^{N} \mathbf{e}_n^\top \mathbf{a}_n + s\mathbf{e}_0^\top \mathbf{z} \\
\text{s.t.} \quad & \sum_{n=1}^{N-1} \mathbf{W}_n^\top \mathbf{b}_n - \sum_{n=2}^{N} \mathbf{W}_n^\top \mathbf{a}_n \leq \mathbf{e}, \\
& \mathbf{e}_1^\top \mathbf{b}_1 - \mathbf{e}_2^\top \mathbf{a}_2 - z_1 \geq 0, \\
& \mathbf{e}_n^\top \mathbf{b}_n - \mathbf{e}_{n+1}^\top \mathbf{a}_{n+1} + z_{n-1} - z_n \geq 0, \quad \forall n = 2, \ldots, N-2, \\
& \mathbf{e}_{N-1}^\top \mathbf{b}_{N-1} - \mathbf{e}_N^\top \mathbf{a}_N + z_{N-2} \geq 0 \\
& \mathbf{b}_n \leq \lambda_n \mathbf{e}_n, \quad \forall n = 1, \ldots, N-1, \\
& \mathbf{a}_n \leq \lambda_n \mathbf{e}_n, \quad \forall n = 2, \ldots, N, \\
& \mathbf{b}, \mathbf{a}, \mathbf{z} \geq \mathbf{0}.
\end{aligned} \tag{24}$$

First note that any feasible solution of (22) satisfies $h_1 > h_2 > \cdots > h_{N-2} > \delta > 0$ because $h_n - h_{n+1} \geq s \geq \delta > 0$. Therefore, for $\mathbf{d} = \mathbf{0}$ it holds that $\mathbf{y}_n^+ = \max\{\mathbf{0}, \mathbf{e}_n(\delta + h_n)\} = \mathbf{e}_n(\delta + h_n) > \mathbf{0}$ (for all $n = 1, \ldots, N-1$). By complementary slackness, it follows that $\mathbf{b}_n = \lambda_n \mathbf{e}_n$. Similarly, $\mathbf{y}_n^- = \max\{\mathbf{0}, \mathbf{e}_n(\delta - h_{n-1})\} = \mathbf{0}$, for all $n = 2, \ldots, N-1$. Thus, $-\mathbf{e}_n h_{n-1} - \mathbf{y}_n^- < -\mathbf{e}_n \delta$ and by complementary slackness it follows that $\mathbf{a}_n = \mathbf{0}$, for all $n = 2, \ldots, N-1$. Similarly, since $h_1, \ldots, h_{N-2} > 0$, it follows that:

$$\begin{aligned}
& \mathbf{e}_1^\top \mathbf{b}_1 - \mathbf{e}_2^\top \mathbf{a}_2 - z_1 = 0, \\
& \mathbf{e}_n^\top \mathbf{b}_n - \mathbf{e}_{n+1}^\top \mathbf{a}_{n+1} + z_{n-1} - z_n = 0, \quad \forall n \in [2, N-2].
\end{aligned} \tag{25}$$

Now suppose that there exists some $l \in [1, N-2]$ for which $h_l - h_{l+1} > s$. This would imply that $z_l = 0$, and since $\mathbf{a}_n = \mathbf{0}$, it leads to $\mathbf{e}_l^\top \mathbf{u}_l + z_{l-1} = 0$. However, $\mathbf{b}_l = \mathbf{e}_l \lambda_l > \mathbf{0}$ and $z_{l-1} \geq 0$ lead to a contradiction. Thus, $h_n - h_{n+1} = s$ for all $n = 1, \ldots, N-2$.

With these results it is easy to show that for $\mathbf{d} = \mathbf{0}$ the optimal value for the objective function of (22) is:

$$s \sum_{n=1}^{N-1} (N-n-1)\lambda_n m_n + (\delta + h_{N-1}) \sum_{n=1}^{N-1} \lambda_n m_n + \lambda_N m_N \max\{0, \delta - h_{N-1}\} \tag{26}$$

From this expression it follows that if $\sum_{n=1}^{N-1} \lambda_n m_n > \lambda_N m_N$, then $h_{N-1} = 0$ and the optimal value for the objective function of (22) is:

$$s \sum_{n=1}^{N-1} (N-n-1)\lambda_n m_n + \delta \sum_{n=1}^{N} \lambda_n m_n. \tag{27}$$

In this case $\mathbf{y}_N^- = \mathbf{e}_N \delta > \mathbf{0}$ and by complementary slackness it follows that $\mathbf{a}_N = \mathbf{e}_N \lambda_N$. Therefore, for the dual solution $\mathbf{b}_n = \lambda_n \mathbf{e}_n$ $(n = 1, \ldots, N-1)$, $\mathbf{a}_n = \mathbf{0}$

$(n = 2, \ldots, N-1)$, $\mathbf{a}_N = \lambda_N \mathbf{e}_N$, $z_n = \sum_{l=1}^{n} \lambda_l m_l$ $(n = 1, \ldots, N-2)$ the values for the objective functions for the primal and dual coincide. Therefore, the trivial solution $\mathbf{d} = \mathbf{0}$ is optimal if and only if:

$$\sum_{n=1}^{N-1} \lambda_n \mathbf{W}_n^{\top} \mathbf{e}_n - \lambda_N \mathbf{W}_N^{\top} \mathbf{e}_N \leq \mathbf{e}. \tag{28}$$

□

Similarly to the ordinal regression case, the result of Theorem 2 can be further explored to analyze the role of the scaling vector $\mathbf{\Lambda} = (\lambda_1, \ldots, \lambda_N)$ in connection to the data characteristics. However, in the classification case different scaling constants are used for the error variables of each class, instead of a single constant that applied to all the errors in the ordinal regression case. Therefore, the existence of a trivial solution to (22) depends both on the absolute values of the different λ's used for each class as well as on the relationship of these scaling constants for different classes. In order to take into account this issue in the analysis of the role of the scaling vector $\mathbf{\Lambda}$ in connection to the data characteristics, we introduce the concept of the trivially equivalent (TE) model as follows.

Definition 1 A model defined from the solution of (22) for some $\mathbf{\Lambda}$, with optimal objective function value F^*, is TE if there exists $\mathbf{\Lambda}'$ satisfying the conditions of Theorem 2 leading to a trivial solution with objective function value $F_{triv} = F^*$.

A TE model does not necessarily lead to the same classification result as the trivial solution. However, since during model development the model selection criterion is the objective function of (22), the existence of a trivially equivalent model indicates that it is possible to select some values for the λ's that will produce a trivial model which will perform equally well to the TE model (in terms of the objective function of (22)).

The existence of a TE model can be identified through the solution of the following simple linear program:

$$\begin{aligned} \max \quad & \theta_c = s \sum_{n=1}^{N-1} (N-n-1)\lambda_n m_n + \delta \sum_{n=1}^{N} \lambda_n m_n \\ \text{s.t.} \quad & \sum_{n=1}^{N-1} \lambda_n \mathbf{W}_n^{\top} \mathbf{e}_n - \lambda_N \mathbf{W}_N^{\top} \mathbf{e}_N \leq \mathbf{e}, \\ & \sum_{n=1}^{N-1} \lambda_n m_n - \lambda_N m_N \geq 0, \\ & \lambda_n \geq 0, \quad \forall n = 1, \ldots, N. \end{aligned} \tag{29}$$

The objective function of this problem is the value of the objective function of (22) for the trivial solution $\mathbf{d} = \mathbf{0}$, whereas the constraints are the necessary and sufficient conditions imposed by Theorem 2 for the existence of the trivial solution. Suppose that (22) is solved using a scaling vector $\mathbf{\Lambda}$ leading to the development of a model

with objective function value F^* for (22). If $F^* \leq \theta_c^*$, where θ_c^* is the optimal objective function value of (29), then the model is TE because there exists a scaling vector $\mathbf{\Lambda}'$ that solves (29) with the additional constraint:

$$s \sum_{n=1}^{N-1} (N-n-1)\lambda_n m_n + \delta \sum_{n=1}^{N} \lambda_n m_n = F^*. \tag{30}$$

Note that (29) can be unbounded, in which case, irrespective of the selected scaling vector used in (22), the resulting model will always be TE. The following theorem characterizes the conditions under which this is possible, in terms of the class averages $\overline{\mathbf{W}}_1, \ldots, \overline{\mathbf{W}}_N$.

Theorem 3 *If there does not exist* $\mathbf{z} \geq \mathbf{0}$, *such that* $(\overline{\mathbf{W}}_n - \overline{\mathbf{W}}_N)\mathbf{z} \geq s(N-n-1)+2\delta$, *for all* $n = 1, \ldots, N-1$, *then irrespective of the selected scaling vector used in* (22), *the resulting model is always TE.*

Proof From the above discussion, it is clear that it suffices to characterize the conditions under which (29) is unbounded. Note that (29) is always feasible because $\mathbf{\Lambda} = \mathbf{0}$ satisfies all the constraints.

Denoting by $\mathbf{z}$ and α the dual variables for each set of constraints of (29), the optimality conditions are written as follows:

$$\mathbf{e}^\top \mathbf{z} = s \sum_{n=1}^{N-1} (N-n-1)\lambda_n m_n + \delta \sum_{n=1}^{N} \lambda_n m_n, \tag{31}$$

$$\mathbf{e}_n^\top \mathbf{W}_n \mathbf{z} - m_n \alpha \geq s(N-n-1)m_n + \delta m_n, \quad \forall n = 1, \ldots, N-1, \tag{32}$$

$$-\mathbf{e}_N^\top \mathbf{W}_N \mathbf{z} + m_N \alpha \geq \delta m_N, \tag{33}$$

$$\mathbf{z}, \alpha \geq 0. \tag{34}$$

The inequalities (32–33) define lower and upper bounds for α:

$$\alpha \leq \overline{\mathbf{W}}_n \mathbf{z} - s(N-n-1) - \delta, \quad \forall n = 1, \ldots, N-1, \tag{35}$$

$$\alpha \geq \overline{\mathbf{W}}_N \mathbf{z} + \delta. \tag{36}$$

Therefore, (29) will be unbounded if and only if there does not exist $\mathbf{z} \geq \mathbf{0}$ such that:

$$(\overline{\mathbf{W}}_n - \overline{\mathbf{W}}_N)\mathbf{z} \geq s(N-n-1) + 2\delta, \quad \forall n = 1, \ldots, N-1. \tag{37}$$

□

A direct consequence of the above theorem is that if there exists a class c_n with $\overline{\mathbf{W}}_n - \overline{\mathbf{W}}_N \leq \mathbf{0}$, then any model taken from the solution of (22) will always be TE, irrespective of the scaling factors $\mathbf{\Lambda}$.

The solution of (29) depends solely on the data of the reference alternatives and the definition of the subintervals for the assessment of the piecewise marginal value functions. Therefore, similarly to the ordinal regression case, θ_c^* can be considered as

a measure of the "quality" of the subintervals that are defined. A high value for θ_c^* indicates that the specified subintervals provide poor discriminating ability and vise versa. This is further demonstrated from the following lemma.

Lemma 2 *Suppose that* $(\mathbf{d}, \mathbf{h}, \mathbf{y}^+ = \mathbf{y}^- = \mathbf{0})$ *solves* (22). *Then, the resulting model that perfectly fits the classification of the reference alternatives has* $\mathbf{e}^\top \mathbf{d} \geq \theta_c^*$.

Proof Follows by contradiction similarly to Lemma 1 for the ranking case. □

Therefore, similarly to the ranking case, the higher the value of θ_c^* the larger the value of $\mathbf{e}^\top \mathbf{d}$ should be to get a model that perfectly fits the data. This would require to use high values for the scaling factors in $\mathbf{\Lambda}$. In the extreme case where (29) is unbounded (i.e., $\theta_c^* = \infty$), it is not possible to develop such a model.

4 Computational results

4.1 Experimental design

The performance of the new formulations developed for the ranking and classification cases is assessed through an experimental analysis.

The analysis considers both numerical data as well as qualitative data. The former are generated from the continuous uniform distribution in [0, 1]. The qualitative data are modelled in a five-point scale (1 to 5) similarly to a Likert scale, which is widely used in questionnaire design. The generation of qualitative data is based on the discrete uniform distribution.

In all cases two pools of data (data pools 1 and 2), each consisting of 5,000 alternatives in $\mathbb{R}^K$ are randomly generated. Data pool 1 is used to construct the reference set, whereas data pool 2 is used to construct a validation set.

Three scenarios are considered for the number of criteria K, ranging from a small set of criteria ($K = 5$) up to a larger set of criteria ($K = 15$). An intermediate case ($K = 10$) is also considered in the analysis.

The alternatives in each data pool are evaluated using an additive value function of the form (1), where the trade-off constants $p_1, \ldots, p_K$ are taken as uniformly distributed random variables. Each marginal value function $v_k(x_k)$ is modelled as:

$$v_k(x_k) = \frac{1 - \exp(x_k \gamma_k)}{1 - \exp(\gamma_k)}.$$

Assuming that the data are normalized in [0, 1], this is a general class of concave, convex or linear functions ranging in [0, 1]. The shape parameter $\gamma_k \neq 0$ defines the type of the function. Negative values for γ_k define a concave marginal value function, whereas positive values define a convex function. Finally, values of γ_k close to zero define an almost linear function. In this analysis the shape parameter γ_k for the marginal value function of each criterion x_k is taken as a uniformly distributed random variable in $[-8, 8]$.

The additive value function which is randomly constructed with the above process provides the global values for the alternatives in each data pool. These global values

are used to rank the alternatives from the best to the worst or to classify them in predefined classes.

In the classification case, two-class and three-class problems are considered. The cut-off point in the two-class case is defined as the median of the global values of the alternatives in data pool 1. In the three-class case the two cut-off points, which are required to classify the alternatives, are defined from the 25% and the 75% percentiles of the global values of the alternatives in data pool 1.

The evaluation of the alternatives in the two data pools is subject to inconsistencies (noise). In the ranking case the inconsistencies are imposed through the introduction of normally distributed random noise in the actual global values of the alternatives. This random noise has zero mean and standard deviation of $\sigma/6$, $\sigma/3$ and $\sigma/2$, each corresponding to increased inconsistencies in the evaluation and ranking of the alternatives, where σ denotes the standard deviation of the scores of the alternatives in the two data pools. In the classification case, the imposed inconsistencies involve random perturbations of the classification of the alternatives. Three inconsistency levels (5%, 15% and 25%) are considered, each corresponding to the number of alternatives whose classification is perturbed.

Once the alternatives in the two data pools are generated and evaluated (ranked or classified) with the above procedure, a reference and a validation set are constructed, each consisting of 1,500 alternatives, randomly selected from the corresponding data pools. In the classification case, the selection of the reference alternatives is performed such that all classes are of equal size in both the reference and the validation set. This assumption of balanced class sizes is not unrealistic. Of course, in many real-world situations, there is a considerable imbalance in the size of the classes. However, developing a model without considering this imbalance, is likely to lead to biased results towards the larger class (i.e., excellent performance for the larger class, but poor performance for the smaller one). Such a model is rather unlikely to be acceptable. To cope with this difficulty, the reference alternatives in each class are often weighted to ensure that each class contributes equally to the performance of the model during its development. Obviously, this weighting process implicitly assumes equal class sizes.

Overall, the design factors of the experimental analysis for the ordinal regression case include the number of criteria (5, 10, 15), the type of the criteria (continuous, discrete), as well as the inconsistency level (low, medium, high). In the classification case, the number of classes is also used as an additional design factor. For each combination of the design factors, 100 runs are performed. Overall, for all the 18 combinations of the three design factors in the ordinal regression case 1,800 reference sets are constructed, each matched to one validation set. Similarly, for all the 36 combinations of the four design factors in the classification case 3,600 reference sets are constructed, each matched to one validation set. The experimental analysis is performed in MATLAB R14, with CPLEX v9.0 as the linear programming solver.

4.2 Results

The results obtained from the application of the new formulations (13) and (22) are compared to the ones traditionally used in the PDA context by the UTA and the UTADIS methods (formulations (9) and (10), respectively).

The methods are applied with different settings regarding the definition of the subintervals, which are required to perform the piecewise linear modeling of the marginal value functions. For the number of subintervals three different specifications are used involving 2, 4 and 8 subintervals. Furthermore, two different approaches are used to define these subintervals. In the first approach, the size of the subintervals is defined such that approximately the same number of reference alternatives fall into each subinterval (henceforth this will be referred as equal volume). In the second approach, all subintervals are defined to have the same size. Of course, these settings apply only to the continuous data sets. For all discrete data sets, the number of subintervals is always equal to 4, each of size 1 (because the discrete data are measured in a 5-point scale).

Table 1 Summary of ordinal regression results

	Factors	Levels	UTA-R	UTA	p-values
Kendall's τ	Criteria	5	0.7657	0.7391	<0.01
		10	0.7652	0.7511	<0.01
		15	0.7606	0.7531	<0.01
	Data type	Continuous	0.7618	0.7410	<0.01
		Discrete	0.7701	0.7681	0.527
	Inconsistencies	Low	0.8725	0.8737	<0.01
		Medium	0.7602	0.7445	<0.01
		High	0.6588	0.6252	<0.01
	Subintervals	Equal size	0.7617	0.7405	<0.01
	method	Equal volume	0.7618	0.7415	<0.01
	Number of	2	0.7551	0.7509	0.155
	subintervals	4	0.7650	0.7429	<0.01
		8	0.7652	0.7291	<0.01
MAE	Criteria	5	2.1629	4.5707	<0.01
		10	1.2456	1.5987	<0.01
		15	0.8818	1.0260	<0.01
	Data type	Continuous	1.5225	2.6519	<0.01
		Discrete	1.1528	1.6383	<0.01
	Inconsistencies	Low	0.8183	0.8763	<0.01
		Medium	1.4431	2.4613	<0.01
		High	2.0288	3.8579	<0.01
	Subintervals	Equal size	1.5202	2.7629	<0.01
	method	Equal volume	1.5249	2.5410	<0.01
	Number of	2	1.7524	1.8609	<0.01
	subintervals	4	1.4178	2.2968	<0.01
		8	1.3974	3.7980	<0.01

In the ranking case, the performance of the developed models is assessed using the Kendall's τ coefficient to measure the concordance between the ranking of the validation alternatives that is obtained by the developed models and their actual ranking as defined by the true additive value function. In the classification case the performance of the models is assessed with the classification accuracy for the validation alternatives, as well as with the Gini coefficient (for the three-class case the Gini coefficient is estimated as using the extension proposed in [3]). In both cases, the mean absolute error between the actual trade-off coefficients of the criteria and the estimated ones is also considered.

The results of the experimental analysis for the ordinal regression case are summarized in Table 1. The results clearly show that in most cases the new formulation based on the regularization principle (UTA-R) provides significantly better results compared to the UTA method (the p-values of a t-test on the differences of the formulations are reported). The improvements of the new formulation are higher for smaller criteria sets, continuous data, larger number of subintervals as well as when the inconsistencies increase. In terms of the Kendall's τ coefficient, the traditional formulation used in the UTA method outperforms the new one only in the case of low inconsistencies. In two other cases (discrete data, small number of subintervals) the differences between the two formulations are not found statistically significant. As far as the mean absolute error (MAE) in the estimation of the actual criteria trade-off constants is concerned, the new formulation performs consistently better than the one of the UTA method in all cases. Similarly to the results for the Kendall's τ coefficients the improvements of the new formulation are higher for smaller criteria sets, continuous data, larger number of subintervals, and higher inconsistencies.

Table 2 provides some further results for the ordinal regression case, focusing on qualitative data. Similarly to the overall results, the new formulation generally provides higher values for the Kendall's τ coefficient, but in most cases the differences are not found significant. However, it is still evident that as the inconsistencies in

Table 2 Ordinal regression results for qualitative data

	Factors	Levels	UTA-R	UTA	p-values
Kendall's τ	Criteria	5	0.7711	0.7655	0.287
		10	0.7711	0.7710	0.971
		15	0.7679	0.7679	1.000
	Inconsistencies	Low	0.8801	0.8831	<0.01
		Medium	0.7658	0.7664	0.410
		High	0.6643	0.6549	<0.01
MAE	Criteria	5	1.7216	2.9965	<0.01
		10	0.9981	1.1230	<0.01
		15	0.7387	0.7954	0.016
	Inconsistencies	Low	0.5434	0.6225	0.119
		Medium	1.1761	1.5122	<0.01
		High	1.7389	2.7803	<0.01

Table 3 Summary of classification results

Factors	Levels	Accuracy			Gini index			MAE		
		UTADIS-R	UTADIS	p-values	UTADIS-R	UTADIS	p-values	UTADIS-R	UTADIS	p-values
Criteria	5	0.8038	0.7876	<0.01	0.7124	0.7004	<0.01	2.9940	4.2312	<0.01
	10	0.8021	0.7805	<0.01	0.7161	0.6982	<0.01	1.2149	1.7526	<0.01
	15	0.7941	0.7770	<0.01	0.7112	0.6959	<0.01	0.8824	1.1900	<0.01
Data type	Continuous	0.7999	0.7792	<0.01	0.7145	0.6964	<0.01	1.6903	2.4738	<0.01
	Discrete	0.8006	0.7962	0.165	0.7053	0.7087	0.527	1.7384	1.8959	0.028
Classes	2	0.8062	0.7898	<0.01	0.6791	0.6580	<0.01	2.1448	2.9945	<0.01
	3	0.7938	0.7735	<0.01	0.7473	0.7383	<0.01	1.2494	1.7880	<0.01
Inconsistencies	Low	0.9064	0.9048	<0.01	0.8997	0.8987	<0.01	1.1890	1.6163	<0.01
	Medium	0.7966	0.7762	<0.01	0.7112	0.6962	<0.01	1.6783	2.4553	<0.01
	High	0.6970	0.6640	<0.01	0.5287	0.4996	<0.01	2.2241	3.1022	<0.01
Subintervals method	Equal size	0.7999	0.7792	<0.01	0.7145	0.6963	<0.01	1.6891	2.4898	<0.01
	Equal volume	0.7999	0.7793	<0.01	0.7146	0.6945	<0.01	1.6914	2.4578	<0.01
Number of subintervals	2	0.8001	0.7849	<0.01	0.7154	0.7046	<0.01	1.6595	2.1891	<0.01
	4	0.8027	0.7812	<0.01	0.7159	0.6977	<0.01	1.5808	2.1261	<0.01
	8	0.7970	0.7717	<0.01	0.7124	0.6868	<0.01	1.8304	3.1063	<0.01

Table 4 Classification results for discrete data

	Factors	Levels	UTADIS-R	UTADIS	p-values
Accuracy	Criteria	5	0.7922	0.8022	0.068
		10	0.8077	0.7951	0.020
		15	0.8018	0.7913	0.050
	Classes	2	0.8054	0.8003	0.236
		3	0.7957	0.7921	0.430
	Inconsistencies	Low	0.9104	0.9181	<0.01
		Medium	0.7944	0.7910	<0.01
		High	0.6968	0.6795	<0.01
Gini index	Criteria	5	0.6848	0.7114	<0.01
		10	0.7158	0.7083	0.424
		15	0.7152	0.7065	0.351
	Classes	2	0.6676	0.6715	0.646
		3	0.7429	0.7459	0.640
	Inconsistencies	Low	0.8972	0.9034	<0.01
		Medium	0.6982	0.7072	<0.01
		High	0.5203	0.5156	0.244
MAE	Criteria	5	3.6119	3.2058	0.016
		10	0.9620	1.5083	<0.01
		15	0.6413	0.9737	<0.01
	Classes	2	2.3425	2.3523	0.935
		3	1.1343	1.4395	<0.01
	Inconsistencies	Low	1.0023	0.9416	0.434
		Medium	1.8395	1.9173	0.506
		High	2.3734	2.8289	<0.01

the data increase, the results of the new formulation are improved over the ones of the UTA method (in the case of high inconsistencies the difference between the two approaches are significant at the 1% level). Furthermore, despite the minor improvements in the Kendall's τ coefficient, the new formulation performs (in most cases) significantly better than the one of the UTA method as far as the estimation of the actual weights is concerned. Except for the case of low inconsistencies, in all other cases the average MAE of the new formulation is significantly lower compared to the one of the UTA method.

The results for the classification case are summarized in Table 3. In terms of the classification accuracy for the validation set, the new formulation (UTADIS-R) provides consistently better results compared to the UTADIS method. The differences between the two approaches are significant at the 1% level, except for the case of discrete data. The improvements are higher for larger number of subintervals, as well as in cases where the inconsistencies in the data increase. Similar results are also

observed for the Gini index, as well as for the MAE in the estimation of the actual trade-off coefficients of the criteria. Additional results for the classification case are given in Table 4 for the discrete data. In terms of generalizing ability (as measured by the classification accuracy and the Gini index), the differences between the new formulation and the one of the UTADIS method are now reduced. Nevertheless, it is still evident that its performance relative to the UTADIS method is improved as the inconsistencies in the data increase. Furthermore, its estimates for the actual trade-off coefficients are significantly better than the ones of the UTADIS method when the complexity of the problem increases (large number of criteria, multi-class problem, high inconsistencies).

5 Conclusions

Disaggregation techniques are often used in MCDM to build decision models and to extract preferential information from a set of global judgments made by decision makers. This study explored the introduction of the regularization principle, which is well known to statistical learning theory, in the disaggregation context of MCDM. Based on existing methods for ranking and classification problems (UTA and UTADIS methods), new formulations were proposed and their properties were analyzed in terms of the data characteristics. Preliminary computational results were also given demonstrating the increased performance of the new models over the existing ones. In particular, the new formulations were found more robust to inconsistencies, as well as to misspecification of some technical parameters of the models, such as the piecewise linear modeling of the marginal value functions. The improvements involved both the generalizing ability of the models as well as the estimation of the actual significance of the criteria.

Although the analysis focused on the development of decision models in the form of additive value functions, it can also be extended to consider other modeling forms, which are commonly used in MCDM, such as outranking relation models. Further experimental testing on real-world data would also help in assessing the performance of the new formulations. Finally, other loss functions can also be considered instead of the L_1 loss which was used in this analysis, in order to investigate the impact of the optimization criterion used during model development on the form and the performance of the final models.

References

1. Bouyssou, D., Marchant, T.: An axiomatic approach to noncompensatory sorting methods in MCDM, I: the case of two categories. Working paper, LAMSADE, Université Paris Dauphine (2004).
2. Doumpos, M., Zopounidis, C.: Multicriteria Decision Aid Classification Methods. Kluwer Academic, Dordrecht (2002)
3. Hand, D.J., Till, R.J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach. Learn. **45**, 171–186 (2001)
4. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York (2001)
5. Jacquet-Lagrèze, E., Siskos, Y.: Assessing a set of additive utility functions for multicriteria decision making: the UTA method. Eur. J. Oper. Res. **10**(2), 151–164 (1982)

6. Jacquet-Lagrèze, E., Siskos, Y.: Preference disaggregation: Twenty years of MCDA experience. Eur. J. Oper. Res. **130**(2), 233–245 (2001)
7. Keeney, R.L., Raiffa, H.: Decisions with Multiple Objectives: Preferences and Value Trade-offs. Cambridge University Press, Cambridge (1993)
8. Moody, J.E.: The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. In: Moody, J.E., Hanson, S.J., Lippmann, R.P. (eds.) Advances in Neural Information Processing Systems, pp. 847–854. MIT, Cambridge (1992)
9. Mousseau, V., Figueira, J., Naux, J.Ph.: Using assignment examples to infer weights for ELECTRE TRI method: Some experimental results. Eur. J. Oper. Res. **130**(2), 263–275 (2001)
10. Mousseau, V., Slowinski, R.: Inferring an ELECTRE-TRI model from assignment examples. J. Glob. Optim. **12**(2), 157–174 (1998)
11. Roy, B.: The outranking approach and the foundations of ELECTRE methods. Theory Decis. **31**(1), 49–73 (1991)
12. Schölkopf, B., Smola, A.: Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond. MIT, Cambridge (2002)
13. Zopounidis, C., Doumpos, M.: Multicriteria classification and sorting methods: a literature review. Eur. J. Oper. Res. **138**(2), 229–246 (2002)