Entropy-optimal weight constraint elicitation with additive multi-attribute utility models

Gert van Valkenhoef[a,*], Tommi Tervonen[b]

[a] *Department of Epidemiology, University Medical Center Groningen, University of Groningen, The Netherlands*
[b] *Evidera Ltd, London, United Kingdom*

Abstract

We consider the elicitation of incomplete preference information for the additive utility model in terms of linear constraints on the weights. Eliciting incomplete preferences using holistic pair-wise judgments is convenient for the decision maker, but selecting the best pair-wise comparison is difficult. We propose a framework for comparing holistic preference elicitation questions based on their expected information gain, and introduce a procedure for approximating the optimal question. We extend the basic approach to generate reference alternatives that differ on only a few attributes, and to determine when further preference information is unlikely to reduce decision uncertainty. We present results from computational experiments that assess the performance of the procedure and assess the impact of limiting the number of attributes on which the reference alternatives differ. The tests show that the proposed method performs well, and when implemented in a decision support system it may substantially improve on-line elicitation using pair-wise comparisons.

Keywords: Multicriteria, Decision making/process, Probability, Information theory

1. Introduction

We consider multi-attribute decision problems where a set of alternatives indexed with $I = \{1, \ldots, m\}$ are evaluated based on their uncertain measurements on a set of attributes indexed with $J = \{1, \ldots, n\}$. Based on the evaluations, the Decision Maker (DM) either needs to rank the alternatives, or to choose a small subset (possibly of size one) of best alternatives. The alternatives' attribute measurements x^i_j, $\forall (i, j) \in I \times J$, have a joint density $f_X(x)$. We assume the DM's preferences are representable with an additive multi-attribute utility function u that consists, $\forall j \in J$, of partial utility functions $u_j : \mathbb{R} \to [0, 1]$ and their scaling factors (weights) w_j:

$$u(x^i, w) = \sum_{j \in J} w_j u_j(x^i_j) \ , \tag{1}$$

*Corresponding author
Email addresses: g.h.m.van.valkenhoef@rug.nl (Gert van Valkenhoef), tommi.tervonen@evidera.com (Tommi Tervonen)

where x^i is the vector of attribute measurements for alternative $i \in I$. The model (1) has two types of preference information that need to be elicited from the DM: partial utility functions u_j that simultaneously model the DM's risk attitude and the attractiveness of the attribute scale levels, and the weights w_j that express relative importance of the attribute scale swings, i.e. trade-offs [19].

There are various textbook methods for eliciting the partial utility functions and the weights. For example, the partial utility functions u_j can be obtained with the standard gamble method and the weights w_j with the swing weighting method [19]. However, research has shown that utility and weight elicitation are cognitively demanding and prone to behavioural biases [see e.g. 4, 10, 25, 41]. Various techniques have been developed to reduce these biases and to consequently make the additive model applicable in practical decision aiding settings. For example, instead of eliciting exact and complete weight information, preference information about the weights can be elicited in an incomplete format, which leads to having linear weight constraints [27, 29]. However, there is little evidence that eliciting such weight constraints directly would not be affected by the same behavioural biases that affect the direct methods for exact weight elicitation. Obtaining the constraints indirectly by asking the DM to provide preference information in the form of holistic pair-wise judgments over a given pair of reference alternatives with deterministic evaluations (e.g. x^1 is preferred over x^2 or vice versa) has been proposed as a practical technique for lowering the DM's cognitive burden [8, 15, 16].

On the other hand, instead of true utility functions, one can apply simpler value functions that do not model the DM's risk attitude but solely quantify the attractiveness of deterministic outcome levels. Uncertainty in the attribute measurements and incomplete information on the weights can then be analyzed to describe the possible decision outcomes. In the Stochastic Multicriteria Acceptability Analysis (SMAA) methodology [20, 21, 31] such analyses result in rank probabilities for each decision alternative. These are also known as rank acceptabilities and they are usually estimated through Monte Carlo simulation [32]. The SMAA approach has been succesfully applied for decision support in various domains [e.g. 11, 20, 34, 35, 38].

The usefulness of the simulation approach is related to the capability of the rank probabilities to discriminate the decision alternatives, which in turn depends on the linear weight constraints resulting from the answers to the pair-wise questions. The number of questions the DM needs to answer should be as small as possible to minimize the DM's cognitive load. This paper develops an entropy-based method for optimizing information gain from pair-wise holistic elicitation questions.

1.1. Background

Choosing the most appropriate pair of actual or fictitious reference alternatives for indirect preference elicitation is not a trivial task, and different approaches have been developed for tackling the problem. Rios-Insua and Mateos [28] defined conditions for reducing the efficient discrete set of alternatives after observing a single pair-wise choice, but they did not consider further which pair of alternatives to use in

the next elicitation iteration. The question-response process and the optimal choice sequence (policy) was analyzed by Holloway and White [12]. They assumed deterministic attribute measurements, a finite set of alternatives, and linear partial utility functions. Iyengar et al. [13] presented a heuristic for choosing the pair-wise question to ask so that the resulting weight space is cut approximately in half. However, none of these works took into account uncertain attribute measurements or the information theoretic basis of the decision; such an entropy-based method for problems with discrete outcome sets was introduced by Abbas [1].

Entropy quantifies the amount of uncertainty over a set of events [6]. In multi-attribute choice problems the events are the different alternatives that can obtain the first rank, that is, there are $|I|$ possible events. In ranking problems the events are the different possible rankings over the set of alternatives. Entropy allows us to compare the level of uncertainty associated with different imprecise preferences, and therefore select the elicitation question that maximizes the information gain. Entropy has been used in a related context for constructing a full joint outcome distribution when the marginal probabilities and pair-wise correlations are elicited from experts [3]. The objective of using entropy in specifying the outcome distribution is different: rather than eliciting preferences to minimize entropy so that the decision recommendation becomes more certain, it is used to select the maximum entropy (most uncertain) distribution compatible with the given information. We do not consider the construction of the full joint outcome distribution in this paper, and emphasize that our use of entropy as a metric of overall decision uncertainty is quite different from the maximum entropy approach to the selection of probability distributions.

1.2. Contribution

Our contribution in this paper is two-fold. First, we develop an entropy-based framework for reasoning about holistic preference elicitation questions in multi-attribute ranking problems with continuous measurement distributions. We consider additive utility models where the partial utility functions, apart from the scaling factors (weights) are elicited beforehand. We extend the previous work of Abbas [1] by considering a discrete set of decision alternatives with uncertain attribute measurements modeled through a joint probability distribution. Thus, the set of possible outcomes is uncountable as we do not integrate uncertainty into a utility measure, but the actual set of decision alternatives is finite. We believe the current work is the first to consider this problem setting, and therefore the previously mentioned approaches are not applicable. Neither are the various questioning methods developed for interactive multi-objective optimization [see e.g. 30]. The second contribution is the development of a greedy technique that chooses the myopically optimal question in each elicitation iteration. We numerically investigate the procedure's performance with artificial problems, and present an application of the technique on benefit-risk analysis of anti-thrombolytic drugs.

The remainder of this paper is structured as follows. In Section 2, we describe the theory of entropy optimal weight constraint elicitation and the specific case of pair-wise questions. Extensions that consider

restricted sets of reference alternatives and a stopping criterion are discussed in Section 3. The computational experiments and their results are presented in Section 4, and an example is analyzed in Section 5. Section 6 ends the paper with a discussion.

2. Entropy-optimal weight elicitation

In this section, we first outline our general framework for entropy optimal weight constraint elicitation. We then show how the relevant quantities can be approximated using Monte Carlo simulation, and how the framework can be applied to find approximately entropy-optimal pair-wise elicitation questions. In what follows, we will refer to riskless utility (value) when talking about utility, although the developed theory mostly also applies to true utility functions that model the DM's risk attitude. Incomplete information on the weights is represented through a density $f_W(w)$ that is non-zero only within the $(n-1)$-simplex

$$W_n = \left\{ w \in \mathbb{R}^n | w \geq 0, \sum_{j \in J} w_j = 1 \right\}. \tag{2}$$

The ranges of u_j in (1) are defined through hypothetical alternatives x^{worst} and x^{best} that have all the measurements at the worst and the best levels in x, respectively, and thus $\forall j \in J : u_j(x_j^{\text{worst}}) = 0, u_j(x_j^{\text{best}}) = 1$. In case x is not bounded, the worst and best values can be chosen e.g. as hull of the 95% confidence intervals [35]. We consider the case of a single DM, and model incomplete information about her preferences with a density $f_W(w)$ that is uniform within a *feasible weight space* W' formed by restricting W_n with linear constraints.

Starting from the current feasible weight space $W' \subseteq W_n$, an elicitation question Q with an answer set $A(Q)$ will reduce the feasible weight space to a sub-region $W'' \in A(Q)$. For example, when $n = 2$, the first question could be which one of two alternatives with partial utility vectors $u^1 = (1, 0)$, $u^2 = (0, 1)$ the DM (weakly) prefers. Possible answers to this question are $A(Q) = \big\{\{w \in W_2 | w_1 \geq w_2\}, \{w \in W_2 | w_2 \geq w_1\}\big\}$. In general, the answer set will be complete ($\bigcup_{W'' \in A(Q)} W'' = W'$), but the answers do not need to be disjoint. The answer set could be uncountable, e.g. when the DM is asked to pick a single number from a real valued scale (contingent valuation), but we consider only countable answer sets in this paper (conjoint valuation).

Depending on the problem structure, the choice of the elicitation question can have a large impact on how well the decision alternatives are discriminated with the answer. The question should maximize the expected information gain, which we define by taking into account the objective of applying the decision model, e.g. to rank the alternatives. Figure 1 illustrates this in a simple problem with two alternatives A and B and two attributes. The partial utilities are distributed as $u_1(x_1^A) \sim \mathcal{U}(0, 0.3)$, $u_2(x_2^A) \sim \mathcal{U}(0.7, 1)$, $u_1(x_1^B) \sim \mathcal{U}(0.3, 0.6)$, and $u_2(x_2^B) \sim \mathcal{U}(0.2, 0.4)$. In this case, the optimal cut of the weight-space in two half-spaces lies at $w_1 \approx 0.62$ and $w_2 \approx 0.38$ rather than at the half-volume cut ($w_1 = w_2 = 0.5$).

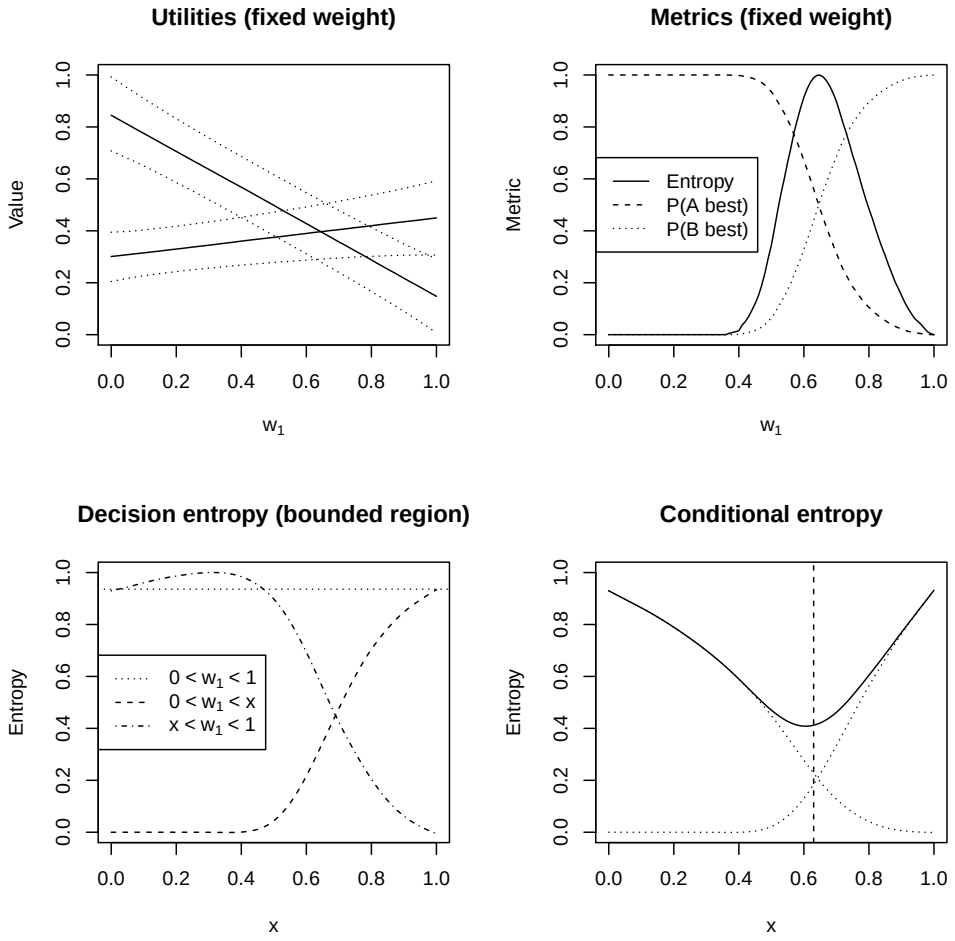

Figure 1: Example 2-attribute choice problem with two alternatives, A and B, where the half-volume cut at $w_1 = w_2 = 0.5$ is not equal to the entropy-optimal cut at $w_1 \approx 0.62$, $w_2 \approx 0.38$. The top-left panel shows the median utilities (solid lines) and the 2.5% and 97.5% percentile utilities (dotted lines) the alternatives obtain; the utility of A decreases as a function of w_1. The top-right panel shows how the probabilities of each alternative being the best and the overall decision entropy depend on w_1. The bottom-left panel shows the entropy (remaining uncertainty) when the weight space is restricted in different ways. Finally, the entropy conditional on the DM's answer is shown in the bottom-right panel. The contributions of the two answers are illustrated with dotted lines, and the optimal weight space cut (corresponding to the entropy-optimal pair-wise question) with a dashed vertical line.

We use the entropy [6] $H_{W'}(Y)$ to define information for a given weight space W' over the space of possible outcomes Y:

$$H_{W'}(Y) = -\sum_{y \in Y} p(y|w \in W') \log p(y|w \in W') \ . \tag{3}$$

When the objective is to rank the alternatives, Y contains all $m!$ possible rankings, the maximum possible entropy is $(m^2 + m)/2$, and a single ranking can be represented through a rank vector (e.g. for $m = 3$, y could be $[1, 3, 2]$). The maximum entropy corresponds to all $m!$ rankings being equally likely, whereas zero entropy corresponds to the ranking being known exactly. Then, $p(y|w \in W')$ is the share of utilities that result in ranking y, given the set of weights W':

$$p(y|w \in W') = \int_{w \in W_n} \int_{x \in \mathbb{R}^{m \times n}} f_W(w|w \in W') f_X(x) p(y|x, w) \, \mathrm{d}x \, \mathrm{d}w \ , \tag{4}$$

where

$$p(y|x, w) = \begin{cases} 1 & \text{if } y_i = 1 + \sum_{k \in I} \rho(u(x^k, w) > u(x^i, w)), \ \forall i \in I, \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

and ρ is the Boolean indicator function. Definitions of the outcome space Y and the probability mass $p(y|w \in W')$ depend on the decision aiding objective, and they can be defined similarly for choice and sorting problems. Now, given a countable answer set $A(Q)$ and suitable definitions of Y and $p(y|w \in W')$, the entropy conditional on the answer is [6]:

$$H(Y|A(Q)) = \sum_{W'' \in A(Q)} p(W'') H_{W''}(Y) \tag{6}$$

$$= -\sum_{W'' \in A(Q_p)} p(W'') \sum_{y \in Y} p(y|w \in W'') \log p(y|w \in W'') \ . \tag{7}$$

For a set of possible elicitation questions $\mathbf{Q}$, the entropy-optimal elicitation question $Q^* \in \mathbf{Q}$ is the one that minimizes $H(Y|A(Q))$:

$$Q^* = \arg\min_{Q \in \mathbf{Q}} H(Y|A(Q)) \tag{8}$$

2.1. Estimating question entropy

In order to determine the question entropy $H(Y|A(Q))$, we need to evaluate both the probabilities of all outcomes $y \in Y$, $p(y|w \in W'')$, and the probabilities $p(W'')$ associated with each $W'' \in A(Q)$. The $p(y|w \in W'')$ are not known in most applications and their exact computation is in general intractable as they depend on both f_W and f_X. Usually f_X is defined as a joint distribution for which efficient sampling algorithms exist. If f_W is uniform, the Hit-And-Run (HAR) Markov Chain Monte Carlo (MCMC) algorithm can be used to sample uniformly from W'' with $O^*(n^3)$ complexity [36, 39]. 10^4 Monte Carlo iterations provide an accuracy for rank probabilities that is sufficient in most practical applications [32]. A similar sampling approach can be used for estimating the probabilities of all possible rankings. However, given the

limited number of Monte Carlo iterations, some of the possible rankings $y \in Y$ might not be encountered in the simulations. Their corresponding best estimates of $p(y|w \in W') \log p(y|w \in W')$ are 0.

To estimate the answer probabilities $p(W'')$, we can either use the density f_W: $p(W'') = \int_{w \in W''} f_W(w|w \in W')\mathrm{d}w$, or assume the answers are equiprobable. As the set of weights corresponding to the DM preferences is often a small subregion of W_n, and the meaning of the weights depends on the attribute scale ranges, both ways of defining $p(W'')$ seem reasonable. That is, the density-based probabilities bias convergence towards decreasing the size of W' faster, whereas equiprobable $p(W'')$ are consistent with the behavioural foundations of pair-wise choices (i.e. there is no *a priori* reason to assume that one of the answers would be more probable). However, when answers are assumed to be equiprobable, the procedure tends to select extreme questions where one of the answers corresponds to a tiny region of the feasible weight space with a very low entropy. Early computational tests showed that the procedure will then fail to converge on an assumed "true" weight vector. This is not surprising given that the entropy measure assumes that a probability distribution over W exists, and equiprobable answers do not correspond to any valid probability distribution. Therefore, density based probabilities are used in the remainder of the paper.

2.2. Pair-wise judgments and the optimal separating hyperplane

Pair-wise preference elicitation consists of asking the decision maker to compare two deterministically evaluated reference alternatives x^1 and x^2. When x^1 is (weakly) preferred to x^2, this implies that $u(x^1, w) \geq u(x^2, w)$. Such statements cut the weight space along a hyperplane $\{w \mid a \cdot (w - w^*) = 0\}$ characterized by a point w^* and a normal vector a. The set of candidate questions is then

$$\mathbf{Q} = \{\{w \mid a \cdot (w - w^*) = 0\} \mid w^* \in \mathrm{int}(W'),\ a \in (\mathbb{R}^n - 0)\}, \tag{9}$$

where $\mathrm{int}(W')$ is the interior of W'.

The optimal hyperplane is the one that minimizes $H(Y|A(Q))$. This can be approximated by sampling a sufficiently large number of candidate hyperplanes and computing the question entropies for each of them. Candidate hyperplanes can be sampled by drawing $n-1$ points uniformly from the boundary of W', with the restriction that they do not all lie in a single face. Then, the normal vector a is the points' null space and w^* an arbitrary convex combination of the points. Sampling from the boundary can be achieved with the shake and bake algorithm [5], a variant of HAR. Because this is an MCMC algorithm, the samples must be thinned to reduce autocorrelation. That is, when the thinning factor is δ, only every δ-th sample is actually used. We used thinning factors that have been shown to be sufficient for HAR [36]. These depend on the polytope's dimension n as $O^*(n^3)$ [23]. In addition, we use a rejection technique to ensure not all points lie on the same face, and obtain additional samples from the Markov chain if necessary. This does not increase the computation time significantly because the probability of all samples lying in the same face decreases with the problem dimensionality.

Although the accuracy of the entropy estimate does not depend on the number of attributes, the distance between the best found hyperplane estimate and the optimal one grows with the problem dimensionality (number of attributes). Also, the gradient of $H(Y|A(Q))$ is not easy to derive from the scalar field $H(Y|w)$. Therefore, the optimization problem of finding the entropy-minimizing hyperplane (8) is hard and on higher dimensionality problems heuristic approaches might have to be used instead of the sampling technique described above.

2.3. Forming a pair-wise elicitation question

After the question $Q = \{w \mid a \cdot (w - w^*) = 0\}$ has been selected, the reference alternatives x^1 and x^2 need to be constructed to let the DM decide which half space corresponds to her preferences. The answer set is given by:

$$A(Q) = \{W_A, W_B\}, \text{ where} \tag{10}$$

$$W_A = \{w \in W' \mid a \cdot (w - w^*) \leq 0\} \text{ and} \tag{11}$$

$$W_B = \{w \in W' \mid a \cdot (w - w^*) \geq 0\} \ . \tag{12}$$

The reference alternatives' utilities are equal, $u(x^1, w) = u(x^2, w)$, for all weight vectors w on the hyperplane, i.e. all $w \in Q = W_A \cap W_B$. We show how to construct reference alternatives that satisfy this constraint, and that are also maximally separated in the partial utility space. The reference alternatives are then representative of the half spaces in the sense that $u(x^2, w) \geq u(x^1, w) \Leftrightarrow w \in W_A$. The alternatives can be generated by setting the differences of the alternatives' partial utilities as $u_j(x_j^1) - u_j(x_j^2) = c(a_j - a \cdot w^*)$ (proof in Appendix), where the scaling constant c can be set as $c = 1/\max_{j \in J} |a_j - a \cdot w^*|$. The partial utilities of the first reference alternative are then:

$$u_j(x_j^1) = \begin{cases} 1 \text{ if } u_j(x_j^1) - u_j(x_j^2) > 0 \\ 0 \text{ otherwise } . \end{cases} \tag{13}$$

The partial utilities of $u_j(x_j^2)$ are trivially constructed with $u_j(x_j^2) = u_j(x_j^1) - (u_j(x_j^1) - u_j(x_j^2))$. The actual attribute values are obtained by taking an inverse of the partial utility functions, which is uniquely defined if they are strictly monotonic.

For example, consider a question characterized by $a = (1, 0)^T$ and $w^* = (0.1, 0.9)^T$. Then, $a \cdot w^* = 0.1$. By applying the above procedure we obtain $u_1(x_1^1) - u_1(x_1^2) = 0.9$ and $u_2(x_2^1) - u_2(x_2^2) = -0.1$. The reference alternatives are then $x^1 = \big(u_1^{-1}(1), u_2^{-1}(0)\big) = (x_1^{\text{best}}, x_2^{\text{worst}})$, $x^2 = \big(u_1^{-1}(0.1), u_2^{-1}(0.1)\big)$.

3. Extensions

We extend our core procedure in two directions: (1) pair-wise judgments where the reference alternatives differ at most on k attributes, and (2) determining when the elicitation procedure can be terminated because further preferences are unlikely to reduce decision uncertainty.

3.1. Constraining the set of reference alternatives

Pair-wise judgments are simpler than the full decision problem in two main ways: the DM is asked to compare only two alternatives at a time, and these alternatives have deterministic rather than stochastic evaluations. However, in higher dimensionality decision problems, general pair-wise judgments may be difficult to answer because the constructed reference alternatives often differ on all the attributes [see e.g. 2, Section 5.4]. The decision maker is then likely to focus on only the top-k most salient attributes, ignoring the other ones and potentially introducing bias. Therefore, in problems with a large number of attributes, it can be useful to generate reference alternatives that differ only on $k < n$ attributes. To do this, we note that the simplex vertices corresponding to the $n - k$ attributes on which the reference alternatives do not differ lie on the separating hyperplane (proof in Appendix). The remaining $k - 1$ points that define the hyperplane can be sampled from the boundary of the feasible weight space and, similar to the unrestricted case, candidate planes for which both the $k - 1$ sampled points and the $n - k$ vertices all lie on the same (extended) face of the polytope should be rejected.

3.2. Stopping criterion

With pair-wise judgments the elicitation procedure can in principle be continued indefinitely and when the alternatives' evaluations are stochastic, there can be a considerable amount of residual uncertainty even when an exact weight vector is elicited. Therefore, it is important to assess whether the decision uncertainty under the current preference information is due to imprecision of the preferences or the uncertainty of the alternatives' evaluations. This is appropriately quantified using the mutual information [6] of the decision metric and the weight vector:

$$I_{W'}(Y;\ w) = H_{W'}(Y) - H_{W'}(Y|w)$$

Mutual information signifies the number of bits of uncertainty in Y that are due to the uncertainty in w. This measure can be normalized to an uncertainty coefficient

$$R = I_{W'}(Y;\ w)/H_{W'}(Y)\ \ ,$$

where $R = 1$ indicates that further preference information can potentially eliminate all decision uncertainty and $R = 0$ indicates that further preference information would not reduce uncertainty at all. Uncertainty coefficients are widely used to judge the strength of association between stochastic variables, e.g. in machine learning [e.g. 9, 18] and the life sciences [e.g. 14, 42]. As a formal stopping criterion, we can set a threshold δ, and terminate the elicitation process when $R \leq \delta$. Choosing a suitable value for δ is difficult, and may depend on the specifics of the decision problem at hand, but values between 0.1 and 0.3 seem reasonable.

This absolute stopping criterion has two potential drawbacks: it requires estimating $H_{W'}(Y|w)$ for a sufficiently large sample of $w \in W'$, and while it is an estimate of how much the entropy can be improved, there is no indication of how many questions would be required to get there. Therefore it may be useful, and

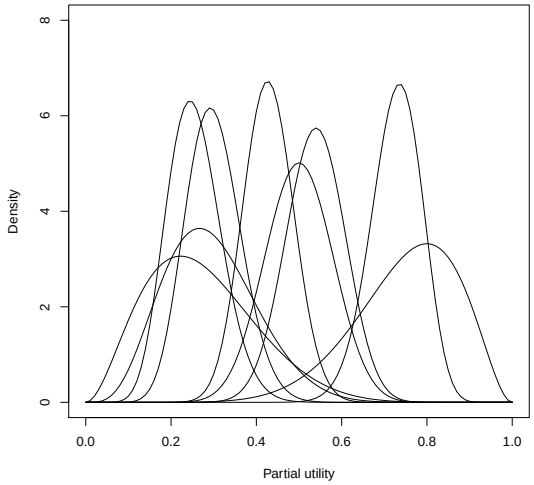

Figure 2: Pre-defined partial utility distributions (Beta density functions) for the random test problems.

more feasible, to plot a time series of the question entropies $H(Y|A(Q))$ and the answer entropies $H_{W''}(Y)$ after each elicitation question, and to compare these with the entropy of the newly proposed question, as well as the minimum and maximum potential answer entropies.

4. Computational experiments

We performed computational experiments to assess (i) convergence of the procedure, i.e. the amount of samples (planes) required for a sufficient estimation accuracy, (ii) running times of different parts of the computation, (iii) efficiency of the entropy-optimal question choice against random questions and ones that maximize the half-space volume ratio, and (iv) the effect of restricting the number of attributes the reference alternatives can differ on. The code for all experiments is freely available online [33].

The computational tests used randomly generated problems with $m = 8$ alternatives and $n = 15$ attributes. For each attribute, the alternatives' partial utility distributions were randomly drawn (without replacement) from the nine Beta distributions presented in Figure 2. Problems with fewer alternatives or attributes were generated by omitting the remaining alternatives or attributes from the randomly generated problem.

4.1. Convergence on the best question

We assessed convergence of the entropy metric for the number of attributes $n \in \{3, \ldots, 15\}$ in a problem with $m = 8$ alternatives. For each dimension, we sampled 10^6 planes and computed entropies with different amounts of planes $(1000, 2000, \ldots, 50000)$, so that at least 20 different sets of planes were available for each

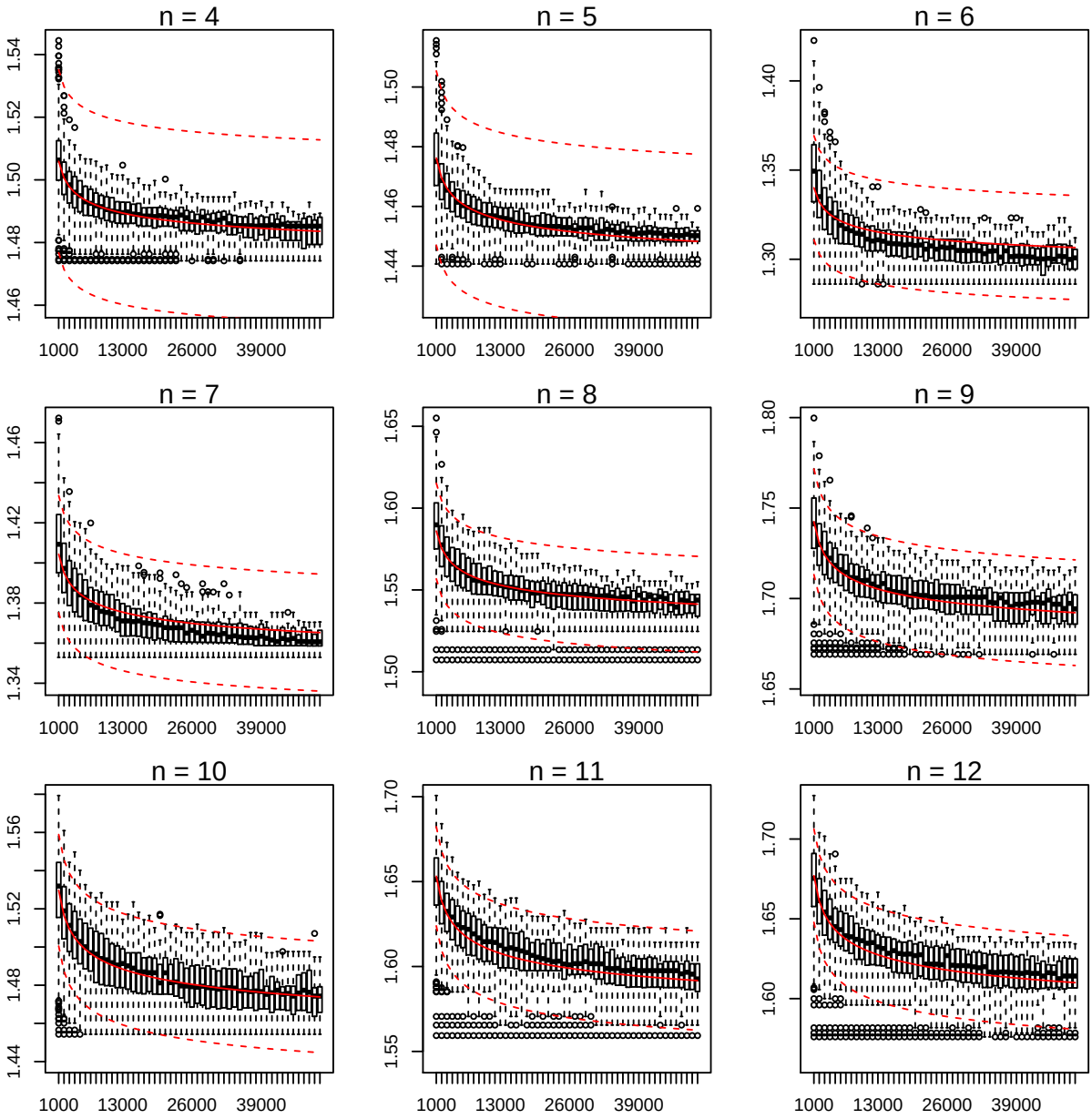

Figure 3: Boxplot of question entropies for the test problems with $n \in \{4, \ldots, 12\}$. The number of planes sampled are depicted on the X-axis, whereas the Y-axis gives the entropy. The red solid and dashed lines are the fit and confidence interval of a joint regression analysis that approximates the rate of convergence to the optimal question.

test. The results for $n \in \{4, \ldots, 12\}$ are illustrated in Figure 3, which shows a box plot of the question entropy after N iterations as well as a fitted regression line described below.

Although we do not know the theoretical rate at which the sampled planes converge on the optimal one, these data nevertheless allowed us to to estimate the practical rate of convergence. Each of the problem instances has an unknown optimal question entropy, and because for each dimension n we used only a single problem, we will refer to this optimum as a_n. We explored a number of functional forms for the rate at which the entropy converges to the optimum by plotting N^{-1}, N^{-2}, $N^{-\frac{1}{2}}$, etc. on the x-axis. Of the candidates explored, $N^{-\frac{1}{4}}$ and $N^{-\frac{1}{5}}$ resulted in the most linear graphs. We then performed a regression of the form

$$y = a_n + bnf(N)$$

where $f(N) = N^{-\frac{1}{4}}$ and $f(N) = N^{-\frac{1}{5}}$ were tried, and the latter was found to have the best fit. The regression was highly significant ($p < 10^{-15}$) and explained nearly all variance ($R^2 > 0.999$). Therefore, it appears that the required number of planes scales as n^5. The estimate for b was 0.041 ± 0.0002, meaning that for $n = 12$, we would need $N \approx 3000$ iterations to get within 0.10 of the optimal entropy, and $N \approx 9000$ iterations for an error of 0.08. We conclude that sampling 10^4 planes is likely to result in questions that are sufficiently close to optimal in most practical problems with less than 15 attributes, and adopt this as the default sample size in the remaining computational tests.

4.2. Running time

In order to assess the actual running time of the procedure, we ran it with $m = 8$ alternatives and varied the number of attributes within $n \in \{3, \ldots, 10\}$. We performed 20 test runs for every value of n. The entropy-optimal estimation procedure has three phases in each elicitation iteration: sampling the 10^4 planes, estimating the optimal plane, and computing the stopping criterion (R). Individual test instances were run on the Dutch national Lisa cluster, in a non-parallel fashion, on single cores that ranged in speeds from 1.80 GHz to 2.60 GHz.

The average times per phase are illustrated in Figure 4. They show that even for moderately small size problems ($n = 5$), computing the three phases takes approximately an hour. Most of the time is taken by estimating the question entropies, which we did to high accuracy: for each of the 10^4 candidate planes, we sampled 10^4 weight vectors from both half-spaces. Therefore, this step is likely to be amenable to optimization. Furthermore, except for the sampling of the planes, each step of our procedure is trivial to parallelize, and even larger problems ($n = 15$) took less than 15 minutes to complete when run in parallel over the cluster.

4.3. Comparison of the entropy-optimal choice against random questions and equal weight space cuts

As alternative procedures for choosing pair-wise questions in our problem setting do not yet exist, we assessed performance of the proposed approach against randomly chosen questions, and questions that are

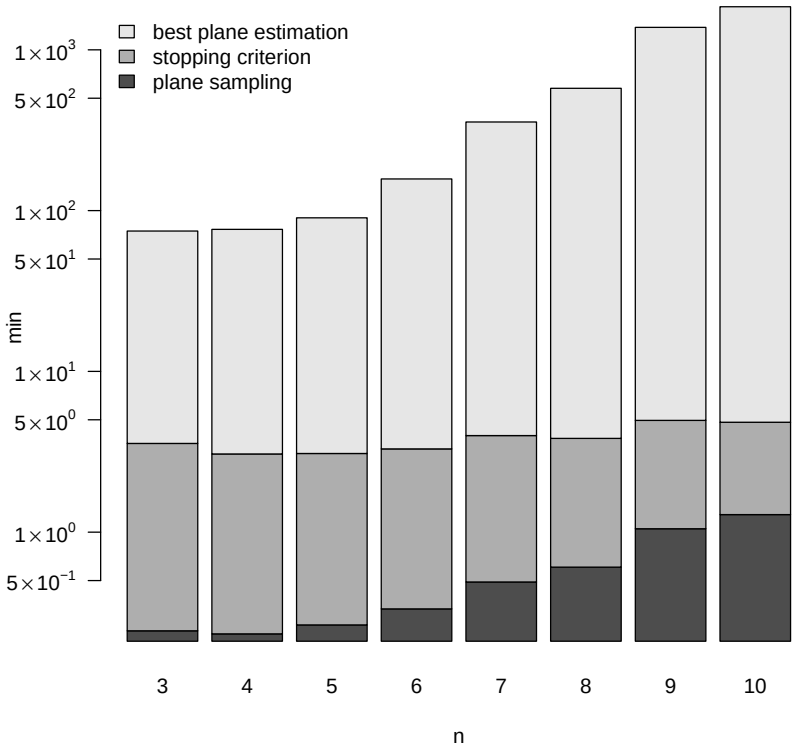

Figure 4: Average times per phase in the computation in minutes, for $m = 10$ and different $n \in \{3, \ldots, 10\}$. The reported times are averages of 20 test runs with different seeds. The Y-axis is logarithmic.

chosen by maximizing the ratio of the volumes of half-spaces corresponding to the two answers. Random questions provide a lower bound benchmark for the given problem setting. For each of the three different plane evaluation approaches, we performed 20 tests with distinct random number generator seeds and $n \in \{3, 6, 9\}$. In each different seed-n combination, we sampled a random test problem and a random "true" weight vector from the $n-1$ simplex. Then, we ran the optimization for 10 iterations, in which the answers corresponding to the "true" preferences were chosen. Average entropies per iteration and the distribution of entropies for the $n = 9$ case are presented in Figure 5. The results indicate that the entropy-optimal choice outperforms the two other ones, and that the differences depend on the specific problem, where problems with more attributes may result in a larger advantage for the entropy-based procedure. As is to be expected, more questions tend to be required in problems with more attributes.

4.4. Effect of restricting the number of attributes the reference alternatives can differ on (k)

The extension presented in Section 3.1 could have a large effect on the optimality of the selected questions, because reducing the number of attributes the alternatives can differ on, k, causes the space of candidate planes to decrease in dimension. To assess the effect, we performed 20 tests with different random number generator seeds, $n = 10$ attributes, randomly chosen "true" preferences (as in Section 4.3) and varied $k \in \{3, \ldots, 9\}$. The results are illustrated in Figure 6. The uncertainty coefficient, which represents residual uncertainty due to the preferences, behaves remarkably similarly for different values of k. These results indicate that, at least within our fixed test problems and with random preferences, limiting k has a little effect on the efficiency of the procedure.

5. Example: benefit-risk analysis of anti-thrombolytics

We analyzed the benefit-risk profiles of two drugs for the prophylaxis of deep vein thrombosis (DVT) following major trauma. The analysis builds upon a previous benefit-risk analysis published in the medical literature [24], and is based on the results of a clinical trial comparing Heparin and Enoxaparin for both efficacy and safety [7]. The treatment benefits are assessed as the prevention of proximal and distal DVT, where proximal DVT is more often associated with the development of serious complications. The safety concern is that administering anticoagulants to trauma patients already at an elevated risk of bleeding might cause additional major bleeding episodes. The original trial data as well as the parameters and characteristics of the estimated beta distributions that serve as attribute measurements are given in Table 1. To construct the partial utility functions, we determined hulls of the 95% confidence intervals of the attribute measurements and set the endpoints of the measurement scales to values that enclose these hulls, similarly to Tervonen et al. [35]. We assumed the partial utility functions to be linear between these endpoints. For proximal DVT, the utility functions were defined for the range from 0.0 (best value) to 0.25 (worst value), for distal DVT from 0.15 (best value) to 0.4 (worst value), and for major bleeding from 0.0 (best value) to

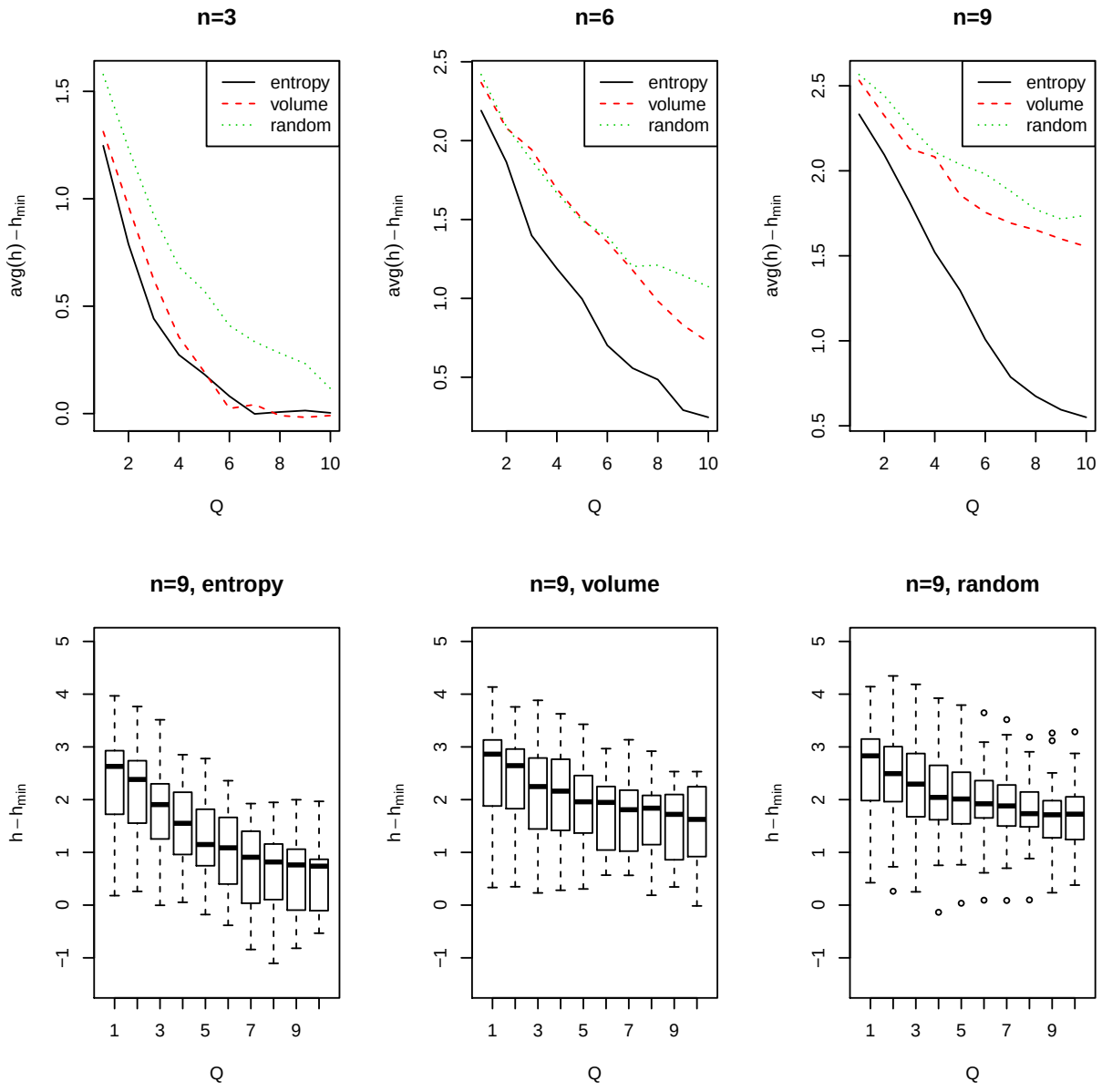

Figure 5: Top: average differences between the current- and minimum entropies (computed using the exact "true" preferences) for the randomly chosen, half-space volume ratio maximizing, and entropy-optimal plane choices for 10 questions, with 20 test instances per parameter combination. Bottom: boxplots of the same results for $n = 9$.

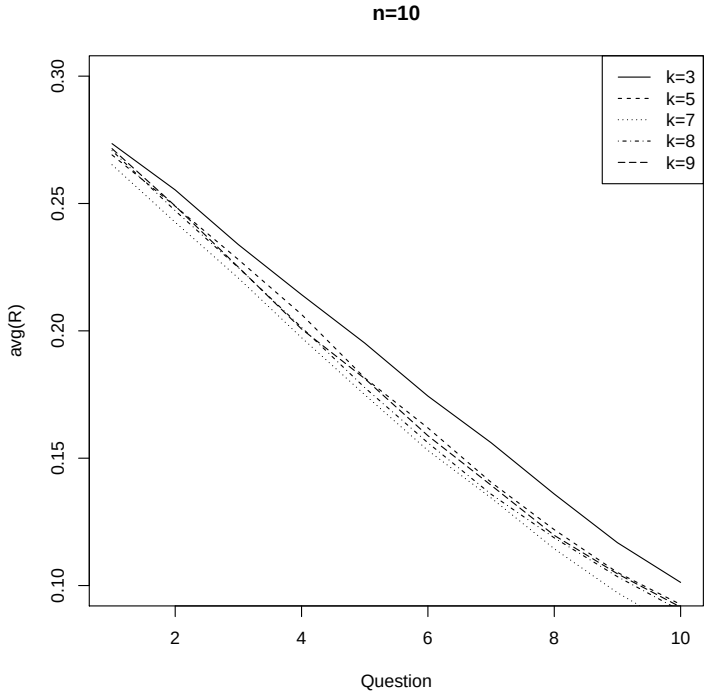

Figure 6: Average uncertainty coefficients (R) over 20 runs in which the procedure was run for 10 questions, for $n = 10$ and $k \in \{3, 5, \ldots, 9\}$.

Table 1: The original trial data given as number of events r and proportion of events r/n. Estimated beta distribution parameters α and β, and the characteristics of the estimated beta distribution are given as the median and the 0.025 and 0.975 quantiles.

	Data		Beta distribution		
Event	r	r/n	α	β	median (95% CI)
Heparin ($n = 136$)					
Proximal DVT	20	0.147	20	116	0.145 (0.093–0.211)
Distal DVT	40	0.294	40	96	0.293 (0.221–0.373)
Major bleeds	1	0.007	1	135	0.005 (0.000–0.027)
Enoxaparin ($n = 129$)					
Proximal DVT	8	0.062	8	121	0.060 (0.027–0.109)
Distal DVT	32	0.248	32	97	0.247 (0.178–0.326)
Major bleeds	5	0.038	5	124	0.036 (0.013–0.078)

0.1 (worst value). We label the weight vector (w_p, w_d, w_b), where w_p is the weight for proximal DVT, w_d for distal DVT and w_b for major bleeding. The measurement probability densities and the decision uncertainty under all possible weight vectors are illustrated in Figure 7.

In this setting, imprecise trade-off preference information was elicited from an expert decision maker (epidemiologist) in the area of thrombolytic drugs [37]. To computationally evaluate our approach, we assumed that the DM's true weight vector corresponds to the centroid of this weight space $(0.53, 0.17, 0.30)$, and simulated the question-answer procedure for 10 iterations. In each iteration, we sampled 10 000 random questions (hyperplanes) and chose from these the one with the lowest entropy. Then, the answer that included the true weight vector was chosen, and the procedure repeated. As the benefit-risk profiles were originally analysed using the SMAA-2 rank acceptability indices [20], we compute these for each answer. Note that in a 2-alternative problem it is sufficient to analyze convergence towards any single rank acceptability, and we chose to use the first rank acceptability of Enoxaparin.

Figure 8 illustrates results from two different tests cases that differ only in the random number generator seed. Although the first three estimated best questions are different (right panel), the question sequences show similar convergence. In both analyses a single answer provides good discrimination of the alternatives, and although the resulting W' are very different in the two cases, the rank acceptability indices are similar and would result in the same decision recommendation. In both cases the uncertainty coefficient R appears to be indicative for stopping the elicitation process: after the first answer $R < 0.3$ and after the second $R < 0.1$ in both cases, which indicates that asking more than two questions would be unnecessary. In both cases, the decision metric is indistinguishable from its true value after four questions.

The observed behavior can be explained by careful interpretation of Figure 7 (bottom panel). The plot of the first rank acceptability of Enoxaparin shows that it is likely to be best with most of the feasible weight vectors, whereas Heparin is the best option only if Bleeding is given much weight. The true weight vector assigns a relatively high weight to proximal DVTs, and Enoxaparin has a high first-rank acceptability in a fairly large range around that vector. Therefore obtaining very precise weight information is not necessary,

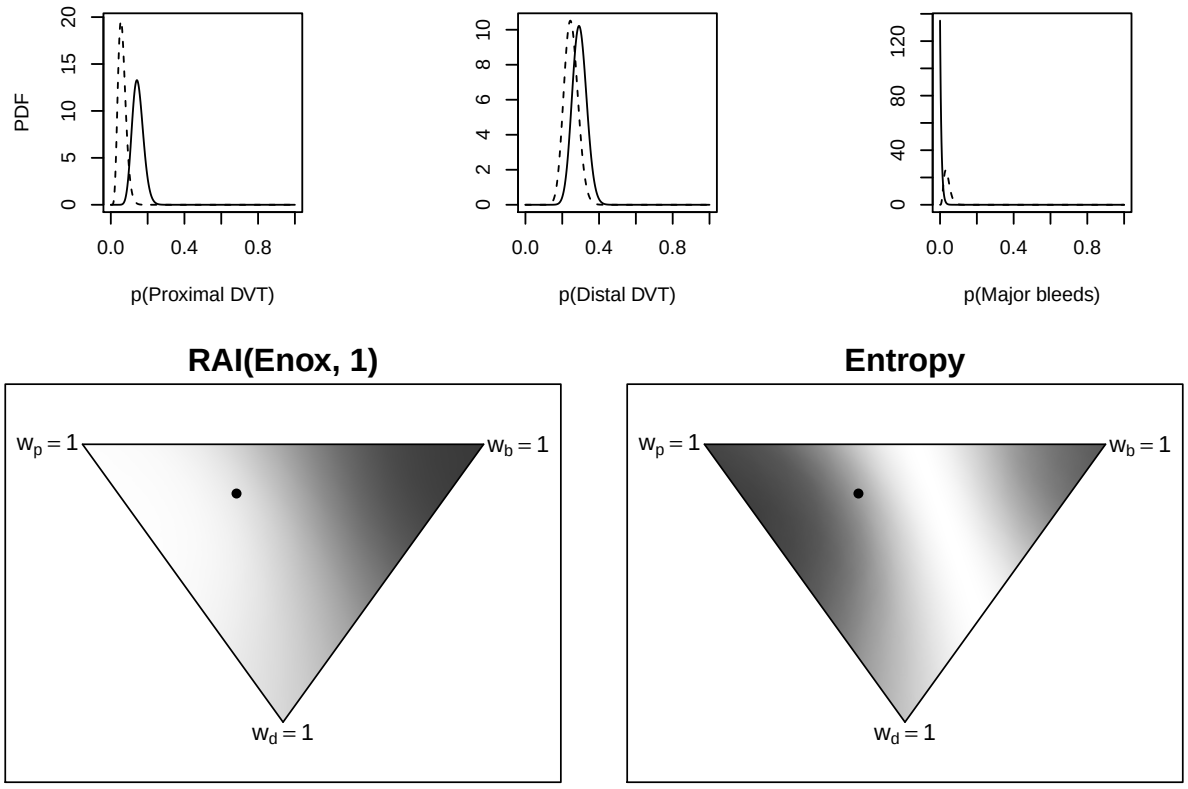

Figure 7: Characteristics of the anti-thrombolytics benefit-risk analysis. Top panels: probability densities for the three attributes used in the anti-thrombolytics benefit-risk analysis; solid lines for Heparin and dashed ones for Enoxaparin. Bottom panels: Enoxaparin first rank acceptabilities (left) and decision entropy (right) for each weight vector; lighter colors indicate higher values (white = 1, dark gray = 0); the black dot indicates the true weight vector.

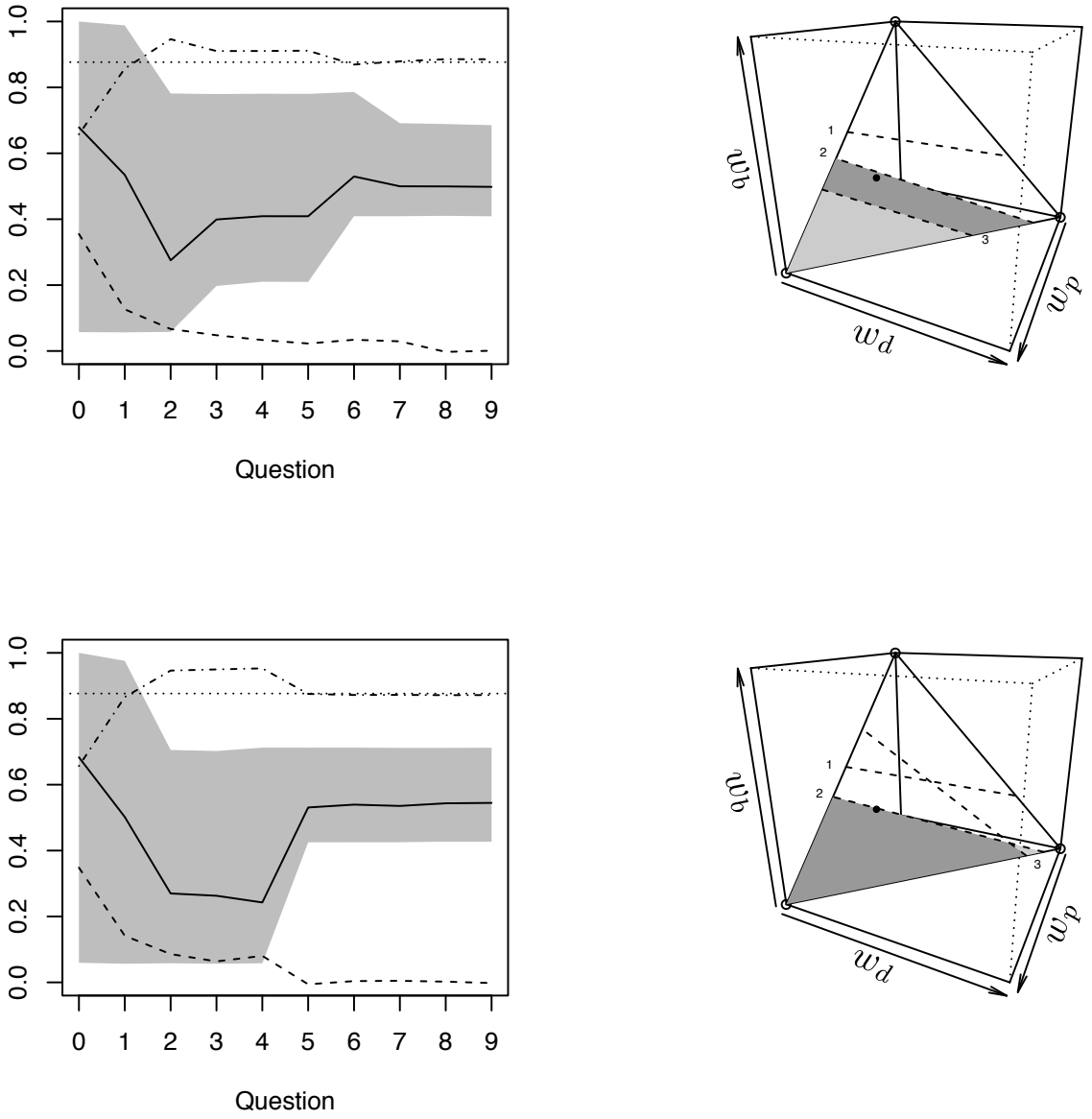

Figure 8: Results for two preference simulations of the anti-thrombolytics benefit-risk assessment that differ only in the random number generator seed. Left: first rank acceptability of Enoxaparin with the true preferences (dotted line, 0.88), first rank acceptability of Enoxaparin with current preferences (dash-dotted line), decision entropy (solid line), minimum and maximum decision entropy with exact preferences (gray area), and uncertainty coefficient (dashed line) after the number of questions indicated on the y-axis has been answered. Right: weight space restrictions imposed by the first three questions of the iterative process, with the dark gray area indicating the remaining weight space W' after the third question. The black dot indicates the (assumed) exact preferences of the decision maker.

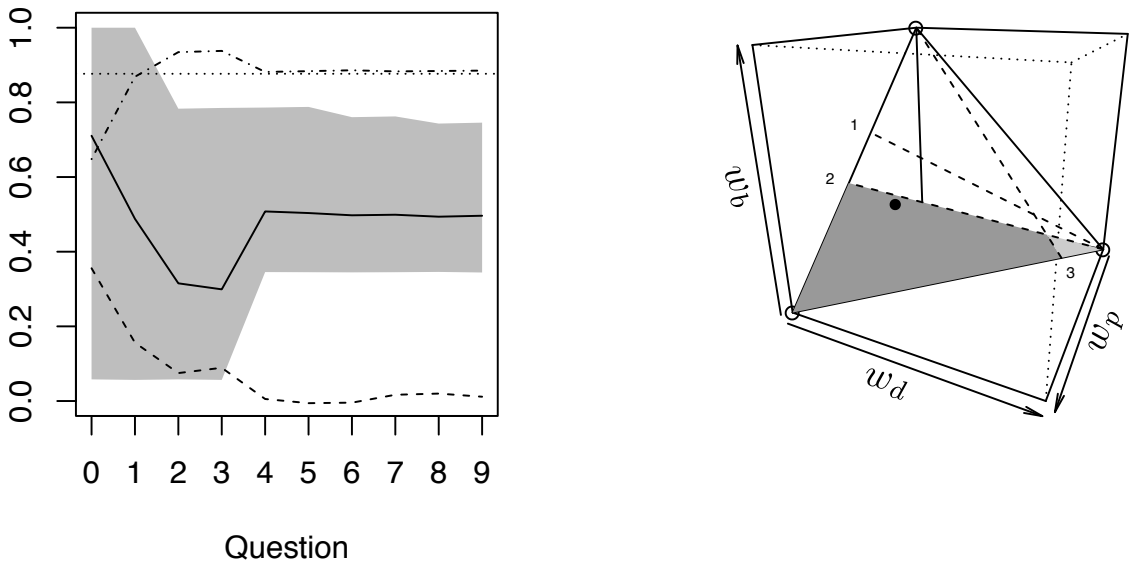

Figure 9: Results for the preference simulation of the anti-thrombolytics benefit-risk assessment with the amount of attributes on which the questions may differ set to $k = 2$. Compare to Figure 8.

as is confirmed by the left panel in Figure 8. In the entropy heat map (Figure 7, bottom right) there is a high entropy region approximately parallel to the $w_p + w_d = 1$ edge of the polytope. When the first question is chosen by minimizing entropy defined through the density-based answer probabilities, the weight space is partitioned into two low entropy regions by cutting through this region.

To illustrate the impact of limiting the number of dimensions in which the generated reference alternatives differ, we re-analyze the anti-thrombolytics case. The alternatives' evaluations and the DM's true preferences are fixed at the same values, but we restrict the number of attributes on which the reference alternatives differ to $k = 2$. The results are presented in Figure 9. It shows that although the generated questions are qualitatively different (compare to Figure 8), the rank acceptabilities still rapidly converge on their true values.

6. Discussion

In this paper, we presented a framework for evaluating elicitation questions that result in linear constraints on the weight space of multi-attribute utility models. The framework enables prioritizing elicitation questions that result in large information gains, and also allows evaluating whether additional preference information can improve the certainty with which a decision can be made. Evaluation of the question information gain assumes probabilities for the question answers to be defined in some manner. Although one would normally not assume any answer to be more likely than another *a priori*, assigning equal probability to each answer results in evaluations that favor extreme questions. Thus, we recommended defining the answer probabilities based on the volumes of the corresponding weight spaces. Initially this is rather arbitrary

because interpretation of the weights depends entirely on how the attribute scales are defined. However, it does bias the procedure towards eliminating large parts of the weight space, and as the feasible weight space converges to represent the DM's preferences, the probability assignments will represent the answer probabilities more realistically.

We then proposed a greedy procedure for online preference elicitation, where an optimal pair-wise comparison of fictional reference alternatives is selected from a large set of randomly generated candidate questions. The approach is myopic, i.e. we do not look ahead beyond the next elicitation question, and do not construct an optimal questioning policy. Although in principle the procedure could be extended to look ahead a certain number of questions, it is already relatively expensive, and the computational cost increases exponentially with the look ahead depth. Therefore, we believe that in practice it would be difficult to do better than our greedy approach when the alternatives' evaluations are stochastic. With deterministic evaluations, one of the alternative methods described in the introduction is likely to be more appropriate.

The elicitation process should be terminated when additional questions do not allow further discrimination of the decision alternatives. For example, Figure 8 indicates that in the benefit-risk analysis of anti-thrombolytics there is no need to ask more than one or two questions because uncertainty due to the measurements then dominates uncertainty due to the preferences. In our computational experiments the uncertainty coefficient consistently appeared to be useful as a stopping condition for the questioning procedure. However, the uncertainty coefficient only indicates the degree to which uncertainty is due to imprecision of the preferences. This number may be difficult to interpret, and therefore it is challenging to agree on a suitable threshold appropriate for a single decision problem, let alone to propose a single threshold for all problems. Moreover, convergence to a sufficiently low entropy can be difficult to assess based on either the uncertainty coefficient or the question entropy. Convergence plots that incorporate both the uncertainty coefficient and the question entropy, as well as the relevant decision metrics, could be used to help the DM decide when the elicitation process should be terminated (e.g. as in Figure 8).

Our results show that our implementation of the questioning procedure is computationally quite expensive, though not unfeasible, especially given parallelization. Unfortunately this means that the present implementation can not be applied in cases where the DM is not willing to wait at least several minutes for the next question to be generated. However, our implementation is poorly optimized and in many cases we made decisions that minimized the variance of the computational tests at the cost of increasing running times. For example, to evaluate the question entropy we sample two sets of 10^4 points from each corresponding half space, and this sampling process dominates the time taken to find the optimal question. This step could be optimized by pre-sampling a large set of samples from W', and partitioning this set to obtain samples from each $W'' \in A(Q)$. Because questions where $p(W'') \approx 0$ for some W'' are unlikely to be optimal, any question that results in a number of samples below a threshold could simply be discarded. Also, we evaluated a large number of hyperplanes (10^4), and practical applications could reduce this number and still

obtain good questions. Heuristic optimization (e.g. simulated annealing or genetic algorithms) could further reduce the number of hyperplanes that need to be evaluated. It may also be possible to approximate the stopping criterion while evaluating fewer weight vectors, e.g. by exploiting the alternatives' central weights [20]. Future work should investigate more efficient algorithms and approximations for the entropy based framework proposed in this paper.

We focused on ranking problems with additive multi-attribute value models applied with simulation. We believe that our question choice procedure is also adaptable for other problem contexts, such as for selecting among a number of efficient project portfolios [22], and with different preference structures, such as non-additive utility models [26]. We considered only elicitation through pair-wise comparison of reference alternatives. Our procedure can be extended to prioritize different types of elicitation questions, such as rank-related requirements [17], but we found pair-wise comparisons to be the most natural application because the set of possible questions is uncountable and each question has only two distinct answers.

Our methods generate pair-wise comparison questions that are optimal from the perspective of information gain, but other aspects could be considered as well. In addition to technical considerations, we addressed the cognitive load of the decision making in two ways: by generating alternatives that are maximally separated in utility space, and by (optionally) generating alternatives that differ on only a few attributes. However, we did not seek to generate plausible alternatives, i.e., alternatives that do not violate real-world inter-attribute constraints and correlations. Empirical evidence suggests that more plausible alternatives may not always be better [40], possibly because plausible alternatives tend to be closer together in utility space. In principle, more plausible alternatives could be generated by incorporating additional constraints in the question selection procedure. However, if comparing plausible alternatives is a key concern, alternative elicitation approaches might be preferred.

Our framework highlights the two main sources of uncertainty encountered in normative multi-criteria decision analysis: uncertain attribute measurements and imprecise preference information. Obtaining precise preference information is not always possible, and DMs are in some cases unable or unwilling to provide additional preference statements, for example for political reasons [20]. Our approach allows to distinguish situations where additional preference statements in weights are useless for further discrimination of the alternatives. In such cases, if the results are not sufficiently accurate to arrive at a conclusion, efforts should instead aim to obtain more exact attribute measurements, or proceed to apply true utility functions that capture the DM's risk attitude.

Acknowledgement

The computational tests of this research were performed on the Dutch National LISA cluster, and supported by the Dutch National Science Foundation (NWO) grant MP-274-14.

Appendix A. Proofs

Theorem 1. *Let $a \cdot (w - w^*) = 0$ characterize a hyperplane and let $u(x^i, w)$ be defined as in (1). Choose a scalar $c \neq 0$ such that $\forall j \in J$, $-1 \leq c(a_j - a \cdot w^*) \leq 1$, and construct x^1 and x^2 such that $\forall j \in J$, $u_j(x_j^1) - u_j(x_j^2) = c(a_j - a \cdot w^*)$. Then $u(x^1, w) = u(x^2, w)$ if and only if $a \cdot (w - w^*) = 0$.*

Proof.

$$\begin{aligned} u(x^1, w) - u(x^2, w) &= \sum_{j=1}^{n} w_j u_j(x_j^1) - \sum_{j=1}^{n} w_j u_j(x_j^2) = \sum_{j=1}^{n} w_j [u_j(x_j^1) - u_j(x_j^2)] \\ &= \sum_{j=1}^{n} w_j c(a_j - a \cdot w^*) = c \sum_{j=1}^{n} [w_j a_j - w_j (a \cdot w^*)] = \\ &= ca \cdot w - c \sum_{j=1}^{n} w_j (a \cdot w^*) = ca \cdot (w - w^*) \end{aligned}$$

□

Theorem 2. *Let the reference alternatives x^1 and x^2 be defined so that for all $j \in J_= \subset J$, $u_j(x_j^1) - u_j(x_j^2) = 0$ and let $a \cdot (w - w^*) = 0$ characterize the corresponding hyperplane. Then for any $g \in J_=$ the vertex $w^g : w_h^g = I(h = g)$ lies on the hyperplane: $a \cdot (w^g - w^*) = 0$.*

Proof. Set $J_{\neq} = J - J_=$. Note that for $j \in J_=$ we have (due to the previous theorem):

$$u_j(x_j^1) - u_j(x_j^2) = 0 \Rightarrow a_j - a \cdot w^* = 0 \Rightarrow a_j = a \cdot w^* \ .$$

This is a set of linear constraints on the hyperplane's normal vector a. Hence,

$$a \cdot w^* = \sum_{j \in J} a_j w_j^* = \sum_{j \in J_{\neq}} a_j w_j^* + \sum_{j \in J_=} a_j w_j^* = \sum_{j \in J_{\neq}} a_j w_j^* + (a \cdot w^*) \sum_{j \in J_=} w_j^*$$

and, solving for $a \cdot w^*$:

$$a \cdot w^* = \frac{1}{1 - \sum_{j \in J_=} w_j^*} \sum_{j \in J_{\neq}} a_j w_j^* \ .$$

A basis for the space of compatible normals can be generated by, $\forall j \in J_{\neq}$, defining a^j as:

$$a_h^j = \begin{cases} 1 & \text{if } h = j \\ 0 & \text{if } h \neq j \ \& \ h \in J_{\neq} \\ a^j \cdot w^* & \text{if } h \in J_= \end{cases}$$

where

$$a^j \cdot w^* = \frac{w_j^*}{1 - \sum_{g \in J_=} w_g^*} \ .$$

Now, write the normal as $a = \sum_{j \in J_{\neq}} c_j a^j$, and

$$a \cdot w^g = a_g = \sum_{j \in J_{\neq}} c_j a_g^j = \sum_{j \in J_{\neq}} c_j (a^j \cdot w^*) = \left(\sum_{j \in J_{\neq}} c_j a^j \right) \cdot w^* = a \cdot w^*$$

□

References

[1] A. E. Abbas. Entropy methods for adaptive utility elicitation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 34(2):169–178, 2004. doi: 10.1109/TSMCA.2003.822269.

[2] N. Argyris, A. Morton, and J. R. Figueira. Cut: A multicriteria approach for concavifiable preferences. *Operations Research*, 62(3):633–642, 2014. doi: 10.1287/opre.2014.1274.

[3] J. E. Bickel and J. E. Smith. Optimal sequential exploration: A binary learning model. *Decision Analysis*, 3(1):16–32, 2006. doi: 10.1287/deca.1050.0052.

[4] H. Bleichrodt, J. M. Abellan-Perpiñan, J. L. Pinto-Prades, and I. Mendez-Martinez. Resolving inconsistencies in utility measurement under risk: Tests of generalizations of expected utility. *Management Science*, 53(3):469–482, 2007. doi: 10.1287/mnsc.1060.0647.

[5] C. G. E. Boender, R. J. Caron, J. F. McDonald, A. H. G. R. Kan, H. E. Romeijn, R. L. Smith, J. Telgen, and A. C. F. Vorst. Shake-and-bake algorithms for generating uniform points on the boundary of bounded polyhedra. *Operations Research*, 39(6):945954, 1991. doi: 10.1287/opre.39.6.945.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, USA, 2nd edition, 2006. ISBN 978-0471241959.

[7] W. H. Geerts, R. M. Jay, K. I. Code, E. Chen, J. P. Szalai, E. A. Saibil, and P. A. Hamilton. A comparison of low-dose heparin with low-molecular-weight heparin as prophylaxis against venous thromboembolism after major trauma. *New England Journal of Medicine*, 335(10):701–707, 1996. doi: 10.1056/NEJM199609053351003.

[8] S. Greco, V. Mousseau, and R. Słowiński. Ordinal regression revisited: multiple criteria ranking using a set of additive value functions. *European Journal of Operational Research*, 191(2):415–435, 2008. doi: 10.1016/j.ejor.2007.08.013.

[9] M. A. Hall and L. A. Smith. Practical feature subset selection for machine learning. In *Proceedings of the 21st Australian Computer Science Conference*, pages 181–191, 1998.

[10] J. C. Hershey and P. J. H. Schoemaker. Probability versus certainty equivalence methods in utility measurement: Are they equivalent? *Management Science*, 31(10):1213–1231, 1985. doi: 10.1287/mnsc.31.10.1213.

[11] J. Hokkanen, R. Lahdelma, and P. Salminen. A multiple criteria decision model for analyzing and choosing among different development patterns for the Helsinki cargo harbor. *Socio-Economic Planning Sciences*, 33(1):1–23, 1999. doi: 10.1016/s0038-0121(98)00007-x.

[12] H. A. Holloway and C. C. White. Question selection for multi-attribute decision-aiding. *European Journal of Operational Research*, 148(3):525–533, 2003. doi: 10.1016/S0377-2217(02)00436-8.

[13] V. S. Iyengar, J. Lee, and M. Campbell. Q-Eval: evaluating multiple attribute items using queries. In *Proceedings of the 3rd ACM conference on electronic commerce*, pages 144–153, 2001. doi: 10.1145/501158.501174.

[14] R. Jelier, M. J. Schuemie, P.-J. Roes, E. M. van Mulligen, and J. A. Kors. Literature-based concept profiles for gene annotation: the issue of weighting. *International journal of medical informatics*, 77(5):354–362, 2008.

[15] M. Kadziński and T. Tervonen. Robust multi-criteria ranking with additive value models and holistic pair-wise preference statements. *European Journal of Operational Research*, 228(1):169–180, 2013. doi: 10.1016/j.ejor.2013.01.022.

[16] M. Kadziński and T. Tervonen. Stochastic ordinal regression for multiple criteria sorting problems. *Decision Support Systems*, 55(1):55–66, 2013. doi: 10.1016/j.dss.2012.12.030.

[17] M. Kadziński, S. Greco, and R. Słowiński. RUTA: A framework for assessing and selecting additive value functions on the basis of rank related requirements. *Omega*, 41(4):735–751, 2013. doi: 10.1016/j.omega.2012.10.002.

[18] M. Kamber, L. Winstone, W. Gong, S. Cheng, and J. Han. Generalization and decision tree induction: efficient classification in data mining. In *Proceedings of the Seventh International Workshop on Research Issues in Data Engineering*, pages 111–120. IEEE, 1997.

[19] R. Keeney and H. Raiffa. *Decisions with multiple objectives: preferences and value tradeoffs*. Wiley, New York, 1976.

[20] R. Lahdelma and P. Salminen. SMAA-2: Stochastic multicriteria acceptability analysis for group decision making. *Operations Research*, 49(3):444–454, 2001. doi: 10.1287/opre.49.3.444.11220.

[21] R. Lahdelma, J. Hokkanen, and P. Salminen. SMAA - stochastic multiobjective acceptability analysis. *European Journal of Operational Research*, 106(1):137–143, 1998. doi: 10.1016/S0377-2217(97)00163-X.

[22] J. Liesiö, P. Mild, and A. Salo. Preference programming for robust portfolio modeling and project selection. *European Journal of Operational Research*, 181(3):1488–1505, 2007. doi: 10.1016/j.ejor.2005.12.041.

[23] L. Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999. doi: 10.1007/s101079900093.

[24] L. D. Lynd and B. J. O'Brien. Advances in risk-benefit evaluation using probabilistic simulation methods: an application to the prophylaxis of deep vein thrombosis. *Journal of Clinical Epidemiology*, 57(8):795–803, 2004. doi: 10.1016/j.jclinepi.2003.12.012.

[25] M. McCord and R. de Neufville. lottery equivalents: Reduction of the certainty effect problem in utility assessment. *Management Science*, 32(1):56–60, 1986. doi: 10.1287/mnsc.32.1.56.

[26] L. V. Montiel and J. E. Bickel. A generalized sampling approach for multilinear utility functions given partial preference information. *Decision Analysis*, 11(3):147–170, 2014. doi: 10.1287/deca.2014.0296.

[27] J. Mustajoki and R. P. Hämäläinen. A preference programming approach to make the even swaps method even easier. *Decision Analysis*, 2(2):110–123, 2005. doi: 10.1287/deca.1050.0043.

[28] S. Rios-Insua and A. Mateos. The utility efficient set and its interactive reduction. *European Journal of Operational Research*, 105(3):581–593, 1998. doi: 10.1016/S0377-2217(97)00068-4.

[29] A. Salo and R. P. Hämäläinen. Preference programming through approximate ratio comparisons. *European Journal of Operational Research*, 82(3):458–475, 1995. doi: 10.1016/0377-2217(93)E0224-L.

[30] W. S. Shin and A. Ravindran. Interactive multi-objective optimization: Survey i – continuous case. *Computers & Operations Research*, 18(1):97–114, 1991. doi: 10.1016/0305-0548(91)90046-T.

[31] T. Tervonen and J. R. Figueira. A survey on stochastic multicriteria acceptability analysis methods. *Journal of Multi-Criteria Decision Analysis*, 15(1–2):1–14, 2008. doi: 10.1002/mcda.407.

[32] T. Tervonen and R. Lahdelma. Implementing stochastic multicriteria acceptability analysis. *European Journal of Operational Research*, 178(2):500–513, 2007. doi: 10.1016/j.ejor.2005.12.037.

[33] T. Tervonen and G. van Valkenhoef. Computational experiment code for 'entropy-optimal weight constraint elicitation with additive multi-attribute utility models'. ZENODO. URL: http://dx.doi.org/10.5281/zenodo.28589.

[34] T. Tervonen, H. Hakonen, and R. Lahdelma. Elevator planning with Stochastic Multicriteria Acceptability Analysis. *Omega*, 36(3):352–362, 2008. doi: 10.1016/j.omega.2006.04.017.

[35] T. Tervonen, G. van Valkenhoef, E. Buskens, H. L. Hillege, and D. Postmus. A stochastic multi-criteria model for evidence-based decision making in drug benefit-risk analysis. *Statistics in Medicine*, 30(12):1419–1428, 2011. doi: 10.1002/sim.4194.

[36] T. Tervonen, G. van Valkenhoef, N. Baştürk, and D. Postmus. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. *European Journal of Operational Research*, 224(3):552–559, 2013. doi: 10.1016/j.ejor.2012.08.026.

[37] G. van Valkenhoef. *Making better use of clinical trials: computational decision support methods for evidence-based drug benefit-risk assessment.* PhD thesis, Graduate School Medical Sciences, University of Groningen, Groningen, The Netherlands, 2012.

[38] G. van Valkenhoef, T. Tervonen, J. Zhao, B. de Brock, H. L. Hillege, and D. Postmus. Multi-criteria benefit-risk assessment using network meta-analysis. *Journal of Clinical Epidemiology*, 65(4):394–403, 2012. doi: 10.1016/j.jclinepi.2011.09.005.

[39] G. van Valkenhoef, T. Tervonen, and D. Postmus. Notes on 'hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis'. *European Journal of Operational Research*, 239(3):865–867, 2014. doi: 10.1016/j.ejor.2014.06.036.

[40] R. Vetschera, W. Weitzl, and E. Wolfsteiner. Implausible alternatives in eliciting multi-attribute value functions. *European Journal of Operational Research*, 234(1):221–230, 2014. doi: 10.1016/j.ejor.2013.09.016.

[41] M. Weber and K. Borcherding. Behavioral influences on weight judgments in multiattribute decision making. *European Journal of Operational Research*, 67(1):1–12, May 1993. doi: 10.1016/0377-2217(93)90318-h.

[42] X. Zhang. *Learning Biological Interactions from Multiple Data Sources.* PhD thesis, Tempe, AZ, USA, 2008. AAI3319412.