Available at
www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

ELSEVIER

Artificial Intelligence 146 (2003) 149–173

Artificial Intelligence

www.elsevier.com/locate/artint

Similarity of personal preferences: Theoretical foundations and empirical analysis

Vu Ha [a,*], Peter Haddawy [b]

[a] *Honeywell Labs, MN65-2600, 3660 Technology Drive, Minneapolis, MN 55418, USA*
[b] *Asian Institute of Technology, PO Box 4, Klong Luang, Pathumthani 12120, Thailand*

Received 4 February 2002

Abstract

We study the problem of defining similarity measures on preferences from a decision-theoretic point of view. We propose a similarity measure, called *probabilistic distance*, that originates from the Kendall's tau function, a well-known concept in the statistical literature. We compare this measure to other existing similarity measures on preferences. The key advantage of this measure is its extensibility to accommodate partial preferences and uncertainty. We develop efficient methods to compute this measure, exactly or approximately, under all circumstances. These methods make use of recent advances in the area of Markov chain Monte Carlo simulation. We discuss two applications of the probabilistic distance: in the construction of the Decision-Theoretic Video Advisor (DIVA), and in robustness analysis of a theory refinement technique for preference elicitation.

Keywords: Similarity measures on preferences; Preference elicitation; Decision theory; Case-based reasoning

1. Introduction

Imagine that Ms. Xaviera (let's call her X), an avid cineast, is watching the Ebert & Roeper show for their reviews of *Kiss of the Dragon*, a recently released movie. Ebert gives it a thumb-up, but Roeper a thumb-down. Who should X listen to in deciding whether to see the movie? While both are great film critics with whom she agrees most of the time, recently X tends to agree more with Ebert, and thus decides to go out and see the movie. X's preference is more "similar" to Ebert's than to Roeper's.

* Corresponding author.
E-mail addresses: vu.ha@honeywell.com (V. Ha), haddawy@cs.ait.th (P. Haddawy).

0004-3702/03/$ – see front matter
doi:10.1016/S0004-3702(03)00013-4

This approach to decision making is ubiquitous in our everyday life. We listen to advice, judgment, recommendations from people with whom we share common interests, tastes. Starting from this simple observation, the fledgling area of collaborative filtering attempts to build systems that recommend items of interest (e.g., movies, music, books, news articles) to people in a virtual community. Each user in the community rates various alternatives according to a numeric scale. The filtering system correlates the ratings in order to determine which users' ratings are most similar to each other. Finally, the system predicts how well users will like new items based on ratings from similar users. But exactly how should preference similarity be modeled? Perhaps subconsciously, Xaviera has a measure of similarity between her cinematic taste and those of Ebert and Roeper. How should we capture this measure?

Until now, similarity measures used in existing recommender systems have been defined on rather sketchy models of preference. For example, Amazon.com has a feature called *Customers who Bought* that suggests to a user browsing a particular item A a list of items bought by users that bought A. The assumption here is that people who express some form of interest in a common item (by buying, or just browsing) may have common interest in other items as well. In other examples, systems such as the GROUPLENS Usenet article recommender [26], the RINGO music recommender [27], and the BELLCORE video recommender [14], use a similarity measure that is defined on explicit numeric ratings that people assign to items. While numeric ratings provide a more accurate model to capture preferences than the naive *Customers who Bought*, they are still not expressive enough. For example, if a person ranks a movie as a 5 on a 1-5 scale and later decides he likes some other movies better, there is no way to accommodate this, without possibly having to reassign ratings to many movies. A similarity measure based on a numeric rating scheme is unable to accommodate even a simple pairwise preferential statement such as "Xaviera likes *Black Orpheus* more than *Lambada, the Forbidden Dance*".

So despite its intuitiveness, the concept of preference similarity is hitherto poorly understood, especially if aspects of preferences such as trade-offs among competing objectives, uncertainty, and attitude toward risk are also involved. Representing and reasoning with these complex aspects of preferences are the realm of decision theory. In this paper, we study the problem of defining, analyzing, and computing preference similarity measures using decision theory as our backdrop. Two aspects of preference similarity measures that mainly concern us here are (1) extensibility to accommodate *partial* preferences and (2) amenability to efficient computation. While the second aspect is self-evidently important, the first aspect is important because most of the time, the available preference information is incomplete.

The focus of this paper is a similarity measure, called *probabilistic distance*,[1] that has its roots in the Kendall's tau function, a well-known concept in the statistical literature. We define the probabilistic distance in Section 3. In a nutshell, the probabilistic distance between Xaviera's cinematic taste and that of Ebert is the *probability* that they disagree on a "randomly chosen" pair of movies (i.e., Xaviera prefers, say, the first to the second,

[1] Technically, distance measures should be referred to as *dissimilarity* measures, in reverse scale against similarity measures. For the sake of simplicity, we use distance measures and similarity measures interchangeably whenever no misunderstanding is foreseen.

while Ebert prefers the second to the first). Because this measure depends only on the fundamental concept of preference (something is *preferred* to something else), it is flexible enough to accommodate almost all aspects of preference. In the case when uncertainty is involved and the preferential information is incomplete, the probabilistic distance is to our knowledge the first measure of similarity on preferences.

Because the question of whether uncertainty is involved plays a significant role in our analysis, we organize our discussion accordingly. Section 4 is devoted to the case of *certainty*, and Section 5 to the case of *uncertainty*. In each case, we provide a comprehensive analytical comparison of the probabilistic distance versus existing similarity measures on preferences, whenever there is one. We provide efficient algorithms to compute the probabilistic distance, exactly or approximately, in all cases. The main ingredient of these algorithms is a Monte Carlo estimation method that makes use of a number of recent advances on sampling with rapidly mixing Markov chains. In Section 6, we describe two applications of the probabilistic distance. The first application is the Decision-Theoretic Video Advisor (DIVA), a recommender system that implements a case-based approach to preference elicitation [24]. Empirical analysis with DIVA shows that the probabilistic distance provides better recommendations than the popular Pearson's correlation coefficient. The second application of the probabilistic distance is its use to measure the degree of violating a domain theory for preference elicitation. We conclude with a summary in Section 7.

2. Preliminary

We start our technical discussion by introducing the necessary background on orders, partial orders, value functions, utility functions, and utility theory. We will occasionally use the terms *decision alternative* and *decision consequence* interchangeably, as we are mainly interested in the consequence of a decision.

2.1. Complete preference orders

A *preference order* $\prec$ on a set of decision consequences $\mathcal{D}$ is a weak order, i.e., an asymmetric ($a \prec b \Rightarrow b \not\prec a$), negatively transitive ($a \not\prec b, b \not\prec c \Rightarrow a \not\prec c$) binary relation on $\mathcal{D}$. For $a, b \in \mathcal{D}$, $a \prec b$ indicates that the decision maker *prefers* b to a. When neither of the two consequences is preferred to the other ($a \not\prec b, b \not\prec a$), we say that the decision maker is *indifferent* between them and denote this relation by $a \sim b$. We also use the notation $a \precsim b$ to denote that the decision maker prefers b to a or is indifferent between them. If for all $a, b \in \mathcal{D}$, either $a \prec b$ or $b \prec a$, we say that $(\mathcal{D}, \prec)$ is a *strict order*. An important technique that is often used in association with preference orders is the use of *consistent functions* that capture preference orders. A real-valued function $f : \mathcal{D} \to \Re$ is said to be *consistent* with a preference order $\prec$ on $\mathcal{D}$ if for all $a, b \in \mathcal{D}$, $a \prec b \Leftrightarrow f(a) < f(b)$. Any real-valued function $f : \mathcal{D} \to \Re$ induces a preference order $\prec_f$ according to the above $\Leftrightarrow$.

When the decision consequences are certain, we call them *outcomes*, and denote the set of outcomes by Ω (thus $\mathcal{D} = \Omega$). We will assume throughout the paper that Ω is finite

and $\Omega = \{1, 2, \ldots, n\}$. It can be proven that for any preference order $\prec$ over Ω there exists a function v, called a *value function*, that is consistent with $\prec$. Furthermore, if $(\Omega, \prec)$ is a strict order, then a value function consistent with $\prec$ is the permutation $\pi : \Omega \to \Omega$ that satisfies that $i \prec j \Leftrightarrow \pi(i) < \pi(j), \forall i, j \in \Omega$.[2] We will refer to this permutation as the *canonical value function* for the strict preference order $\prec$. In many applications, decision outcomes are assigned non-negative numbers called *ratings* (for example, movie ratings, news article ratings, etc). These ratings can be viewed as a value function if we assume that higher ratings are preferred to lower ratings, and equal ratings are equally preferred.

When the decision consequences are uncertain, they can be modeled by probability distributions over outcomes and are called *prospects*. We denote the set of all prospects, which is the set of all probability distributions over Ω by $\mathcal{S}$. The central result of von Neumann–Morgenstern utility theory is a representation theorem that identifies a set of conditions guaranteeing the existence of a function consistent with the preference of a decision maker [35]. This theorem states that if the preference order of a decision maker satisfies a few "rational" properties, then there exists a real-valued function, called a *utility function* $u : \Omega \to \Re$, over outcomes such that $p \prec q \Leftrightarrow \langle p, u\rangle < \langle q, u\rangle$. Here $\langle p, u\rangle$, the inner product of the probability vector p and the utility vector u, is the expected value of function u with respect to the distribution p: $\langle p, u\rangle = E_p[u]$. It is often convenient to extend u, by means of expectation, to a function $u : \mathcal{S} \to \Re$ that maps a prospect $p \in \mathcal{S}$ to $\langle p, u\rangle$. This function is clearly consistent with the preference order $(\mathcal{S}, \prec)$. In this paper, we work only with preference orders that satisfy the above rational properties.

Two value (or utility) functions that induce identical preference orders are said to be *strategically equivalent*. (Note that strategic equivalence is an equivalence relation, denoted $\simeq$.) Otherwise, they are said to be *strategically different*.

2.2. Partial preference orders

How should one represent partial preferences? In most cases, partial preferences are obtained via incomplete preference elicitation. For example, we may have determined that the utility function of a person has a certain parametric form (e.g., multiplicative form), but have yet to assess some parameters (e.g., a scaling coefficient, or a sub-utility function of a multi-attribute utility function). In the case of certainty, we also encounter partial preferences when a person assigns ratings or pairwise preferences to a *subset* of the set of decision outcomes.

For the most generality, we assume that a partial preference order $\prec$ is a binary relation on the set $\mathcal{D}$ of decision consequences. Furthermore, it is reasonable to assume that this binary relation is asymmetric: if we know that a person prefers a to b, then it is not the case that he prefers b to a. We may also assume transitivity: if he prefers a to b, and b to c, then he prefers a to c. In the theory of orders, an asymmetric, transitive binary relation is a called a *partial order*, or a *poset*. In this framework, we thus represent partial preferences

[2] We abuse notation a little bit here. In the proposition $i \prec j$, i and j are decision outcomes, while in the proposition $\pi(i) < \pi(j)$, $\pi(i)$ and $\pi(j)$ are integer numbers.

using partial orders.[3] We have a slightly different concept of consistent functions for partial orders. A real-valued function $f : \mathcal{D} \to \Re$ over the decision consequences is said to be *consistent* with a partial preference order $\prec$ if for any decision consequences a, b, $a \prec b \Rightarrow f(a) < f(b)$ and $a \sim b \Rightarrow f(a) = f(b)$. The set of all functions that are consistent with $\prec$ is denoted as $C(\prec)$. Intuitively, consistent functions capture all information contained in the partial orders, and they might contain more than that. Consequently, functions that are consistent with a partial preference order $\prec$ may be strategically different, as they induce weak orders that are different extensions of $\prec$. There is however a one-to-one correspondence between the weak order extensions of $\prec$ and the equivalence classes of $(C(\prec), \simeq)$.

3. The probabilistic distance on preference orders

3.1. Probabilistic distance on complete preferences

We first formally define the probabilistic distance on complete preference orders. Let the *conflict indicator function* $c_{\prec_1,\prec_2} : \mathcal{D}^2 \to \{0, 1\}$ be defined as follows:

$$c_{\prec_1,\prec_2}(a, b) := \begin{cases} 1 & \text{if } (a \precsim_1 b \wedge b \prec_2 a) \vee (a \prec_1 b \wedge b \precsim_2 a) \\ & \quad \vee (a \precsim_2 b \wedge b \prec_1 a) \vee (a \prec_2 b \wedge b \precsim_1 a), \\ 0 & \text{otherwise.} \end{cases}$$

The probabilistic distance is defined as:

$$\delta(\prec_1, \prec_2) := \Pr\big(c_{\prec_1,\prec_2}(a, b) = 1\big) = E\big[c_{\prec_1,\prec_2}(a, b)\big]. \tag{1}$$

Here the expectation is taken with respect to a and b, which are two independent identically distributed uniform random variables on $\mathcal{D}$. The probabilistic distance on complete preference orders satisfies the following important property.

Theorem 1. *In the case of certainty, the probabilistic distance on the set of weak orders on Ω is a metric with range* $[0, 1]$. *In the case of uncertainty, the probabilistic distance on the set of "rational" weak orders on $\mathcal{S}$ is a metric with range* $[0, 1]$.

Proof. Recall that the conditions for being a metric are:

(1) *Reflexivity*. $\delta(a, b) \geqslant 0$, "=" iff $a = b$.
(2) *Symmetry*. $\delta(a, b) = \delta(b, a)$.
(3) *Triangle Inequality*. $\delta(a, b) + \delta(b, c) \geqslant \delta(a, c)$.

[3] Note that the difference between the definition of complete preference order and that of partial preference order is the difference between negative transitivity and transitivity. Given asymmetry, transitivity is weaker than negative transitivity, i.e., the latter implies the former. This "weakness" reflects the incompleteness of our information about the person's preference.

It is evident that the probabilistic distance only takes values between 0 and 1,[4] the distance between two identical orders is zero, and zero distance implies two identical weak orders. The symmetry of the distance function trivially follows from the symmetry of the conflict function. Finally, to prove the triangle inequality, we note that for all weak orders $\prec_i$, $i = 1, 2, 3$, and alternatives a, b, $c_{\prec_1,\prec_3}(a, b) = 1$ implies either $c_{\prec_1,\prec_2}(a, b) = 1$ or $c_{\prec_2,\prec_3}(a, b) = 1$, and for all events X, Y, $\Pr(X \vee Y) \leqslant \Pr(X) + \Pr(Y)$. □

3.2. Probabilistic distance on partial preferences

We can extend the definition of probabilistic distance to partial orders in the following way. Let $\prec_1$ and $\prec_2$ be two partial orders with corresponding sets of weak order extensions E_1 and E_2. Recall that E_i can be viewed as a set of strategically different value/utility functions f_i consistent with $\prec_i$, for $i = 1, 2$. These functions form a one-to-one correspondence with the weak order extensions of $\prec_i$ (note that in the uncertainty case, the correspondence is with only extensions that satisfy the "rational properties" required for the existence of a utility function). We define the probabilistic distance $\delta(\prec_1, \prec_2)$ to be the average of the probabilistic distance between pairs of extensions of $\prec_1$ and $\prec_2$, respectively. Formally,

$$\delta(\prec_1, \prec_2) = E\big[\delta(\prec_{f_1}, \prec_{f_2})\big] = E\big[E\big[c_{\prec_{f_1},\prec_{f_2}}(a, b)\big]\big],$$

where f_i are uniform random variables on E_i, $i = 1, 2$, and a and b are uniform random variables on $\mathcal{D}$. Note that this distance is *not* a metric on the set of partial orders, since the distance between two identical partial orders that are not complete orders is always positive (which violates the "indistinguishability of identicals" metric requirement). This, however, is desirable if the two orders represent the preferences of two different users, since the complete preference orders for the two may actually differ. For example, if we know nothing about preferences of two people (i.e., their preferences orders are the vacuous partial order), we would not want to say that their preferences are the same.

The reader may have noticed that the uniform distribution is used both in defining the alternative variables (a, b), and the weak order extension variables (f_i). The short rationale behind this is that we want to assume as little as possible about the problem at hand, while still keeping it analytically tractable. The long rationale is the following. For alternative variables (a and b), in a general setting one would not want to define a similarity measure on *preference structures* that relies on some bias in the distributions of *decision outcomes*. For weak order extension variables (f_i), it is sensible and technically easier to convert any prior information about the preference orders $\prec_i$ (e.g., utility independence, risk attitudes, etc.) into *constrains* on the sets of extensions E_i, rather than into *probability distributions* on the set of extensions. This point is discussed in more detail in Section 5.2.

On the other hand, in eliciting partial preference information, it is quite likely that there is bias in what we have elicited information about, and that bias is induced by non-uniform distributions of decision outcomes. For example, with movies, people are more likely to express preferences about movies that they have some affinity for, since those are the ones

[4] The exact range of the probabilistic distance can be a sub-interval of [0 1], in which case we may want to scale it properly to [0 1].

they were more likely to have watched. Thus, we acknowledge that a similarity measure on two people's preferences seems like it should perhaps take this into account. We leave this question open for further research.

4. The case of certainty

4.1. The probabilistic distance on complete preferences

When the decision problem does not involve uncertainty, the distance $\delta(\prec_1, \prec_2)$ can be computed by averaging the conflict function $c_{\prec_1,\prec_2}(i, j)$ over all n^2 pairs $(i, j) \in \Omega^2$. In the case when both $\prec_1$ and $\prec_2$ are strict orders with corresponding canonical value functions π_1, π_2, we have:

$$\delta(\prec_1, \prec_2) = \frac{\tau(\pi_1, \pi_2)}{n^2},$$

where $\tau(\pi_1, \pi_2)$ is *Kendall's tau function* [20] which simply returns the number of conflicts. We can divide the Kendall's tau function by $n(n-1)$ instead of n^2 in order to properly scale the probabilistic distance to the range of [0,1]. According to that scale, the distance between a strict order and its complete reverse is 1.[5]

Example 1. Suppose that there are 3 decision outcomes ($\Omega = \{1, 2, 3\}$) and $1 \prec_1 2 \prec_1 3$ and $2 \prec_2 3 \prec_2 1$. Then $\delta(\prec_1, \prec_2) = 4/9$, since $\{(1, 3), (3, 1), (2, 3), (3, 2)\}$ are the four pairs of outcomes that cause conflict between $\prec_1$ and $\prec_2$.

Another popular metric on the set of permutations of $\{1, 2, \ldots, n\}$ is *Spearman's rank order correlation coefficient*, or *Spearman's rho* [29]:

$$\rho(\pi_1, \pi_2) = \frac{\sum_{i=1}^{n}(\pi_1(i) - \frac{n+1}{2})(\pi_2(i) - \frac{n+1}{2})}{\frac{n^2-1}{12}}. \tag{2}$$

Here, $(n+1)/2$ is the mean, and $\sqrt{(n^2-1)/12}$ is the standard deviation of both π_1 and π_2. Spearman's rho is often computed using the following form:

$$\rho(\pi_1, \pi_2) = 1 - \frac{6R^2(\pi_1, \pi_2)}{n^3 - n},$$

where $R(\pi_1, \pi_2) = (\sum_{i=1}^{n}(\pi_1(i) - \pi_2(i))^2)^{1/2}$ is the *Euclidean distance* between the two vector π_1 and π_2. Spearman's rho and Kendall's tau have been studied extensively in the statistical literature. Other commonly used metrics include:

[5] In Eq. (1), we can define δ by taking the average with $a \neq b$ and thus avoid having to rescale Kendall's tau function. But we now have one uniform, joint probability distribution over all pairs (a, b) where $a \neq b$, instead of two uniform distributions. This approach does not have an elegant generalization to the case of uncertainty, where there are infinitely many alternatives. In this case, it appears that we need to revert back to the definition based on two independent, uniformly distributed random variables. For this reason, we prefers the definition based on individual uniform distributions for the sake of uniformity.

– *Spearman's footrule* [30]:

$$F(\pi_1, \pi_2) = \tfrac{1}{2} \sum_{i=1}^{n} \big|\pi_1(i) - \pi_2(i)\big|. \tag{3}$$

– *Ulam's distance* [33]:

$$U(\pi_1, \pi_2) = n\text{—the max number of items ranked in the same order by } \pi_1 \text{ and } \pi_2.$$

– *Kemeny's distance* [19]:

$$K(\pi_1, \pi_2) = \text{the min number of pairwise inversions to obtain } \pi_2 \text{ from } \pi_1.$$

Ulam's distance is used in DNA research to measure the distance between two strings of molecules. Kemeny's distance is used to define *Kemeny ranking*, which is a ranking that aggregates a set of rankings in such a way that minimizes the total Kemeny's distances from the members of that set. Kemeny's ranking is often used in social choice theory as the best compromise between the possibly conflicting views of a set of judges (e.g., judges in figure skating). See Critchlow [7] for a discussion of these metrics from a statistical point of view.

These metrics can be used as distance measures on strict orders. But since their definitions are all based on the canonical value functions of *strict* orders, it is not straightforward to extend these metrics to define similarity measures on *weak* orders. Furthermore, while it is possible to define distance measures between two weak orders $\prec_1$ and $\prec_2$ based on Spearman's rho (see Eq. (2)) or Spearman's footrule (see Eq. (3)) using some functions π_1 and π_2 that are consistent with $\prec_1$ and $\prec_2$, respectively, the resulting measures will apparently be sensitive to the choice of those functions. For example, in the equation for Spearman's rho (Eq. (2)), $\pi_1(i)$ and $\pi_2(i)$ may be replaced with ratings that the two persons assign to item i, and $(n+1)/2$ may be replaced with corresponding means $\overline{\pi}_1, \overline{\pi}_2$ and $\sqrt{(n^2-1)/12}$ with the corresponding variances σ_1, σ_2 of these ratings. The resulting p is the well-known *Pearson's correlation coefficient*.

$$p(\pi_1, \pi_2) = \frac{\sum_{i=1}^{n} (\pi_1(i) - \overline{\pi}_1)(\pi_2(i) - \overline{\pi}_2)}{\sigma_1 \sigma_2}.$$

The Pearson correlation coefficient measures the degree to which a *linear* relationship exists between two variables (or in this case, two sets of ratings), and thus may be unsuitable as a similarity measure between two variables having a close but non-linear relationship. In addition, the Pearson correlation coefficient does not meet the "indistinguishability of identicals" and the "triangle inequality" metric properties. Many researchers have argued against the routine use of similarity measures that do not meet the requirements of a metric [6,15]. Nevertheless, the Pearson's correlation coefficient and several of its variants have been used as similarity measures quite extensively in many fields. Within AI, notable recommender systems that make use of the Pearson correlation coefficient include the GROUPLENS collaborative filtering system [21,26], the RINGO music recommender [27], and the BELLCORE video recommender [14].

4.2. The probabilistic distance on partial preferences

Let $\prec_1, \prec_2$ be two partial orders with corresponding sets of weak order extensions E_1, E_2. Recall that the probabilistic distance is defined as:

$$\delta(\prec_1, \prec_2) = E\big[\delta(\prec_{f_1}, \prec_{f_2})\big] = E\big[E\big[c_{\prec_{f_1},\prec_{f_2}}(i,j)\big]\big].$$

A simplistic approach to compute this quantity would be to evaluate the conflict function c for all possible 4-tuples $\{(f_1, f_2, i, j) \mid f_1 \in E_1, f_2 \in E_2, i, j \in \Omega\}$ and compute the average. This however is impractical because the number of weak order extensions of a partial order can be exponentially large (the number of strict order extensions of a vacuous partial order—a partial order in which everything is incomparable with everything else—is $n!$). In fact, the much easier problem of *counting linear extensions*[6] of finite posets was shown to be # P-complete[7] [3].

So the problem of computing the probabilistic distance becomes an issue for the first time. Given the hardness of counting and generating linear extensions, we turn to approximation techniques to estimate $\delta_p(\prec_1, \prec_2)$. For example, we can use the Monte Carlo simulation method to estimate $\delta_p(\prec_1, \prec_2)$, provided that we have an efficient algorithm to generate f_i uniformly randomly from E_i. It turns out that counting (approximately) and generating (uniformly randomly) elements of large combinatorial sets are two closely related problems. In fact, Sinclair [28] showed that an efficient algorithm for one problem can be used to construct an efficient algorithm for the other, provided the combinatorial sets have a certain structural property called *self-reducibility*. The set of linear extensions of a poset has this property and, not surprisingly, a number of algorithms for generating (almost) uniformly randomly linear extensions of posets have been developed [4,18]. These algorithms are all randomized algorithms based on the Markov chain Monte Carlo technique.[8] In Appendix A we describe the best known algorithm, due to Bubley and Dyer [4] that has a running time of $O(n^3 \log n\varepsilon^{-1})$, where n is the poset's cardinality, and ε is the desired accuracy.

Now with the help of the routine that almost uniformly randomly generates linear extensions of a poset, we can estimate $\delta_p(\prec_1, \prec_2)$ by randomly generating $f_{ij} \in E_i$ $(i = 1, 2;$ $j = 1, \ldots, k)$, computing $\delta_p(f_{1j}, f_{2j})$, $j = 1, \ldots, k$, and taking the sample mean $\hat{\delta}_p = \frac{1}{k}\sum_{j=1}^{k} \delta_p(f_{1j}, f_{2j})$. This sample mean is an unbiased estimator[9] of $\mathrm{E}Y = \delta_p(\prec_1, \prec_2)$

[6] This is a fundamental problem in the theory of ordered sets with applications in computer science (sorting) and social sciences.

[7] The complexity class #P, introduced by Valiant [34], consists of all counting problems whose solutions are the number of accepting states of some non-deterministic polynomial-time Turing Machine. A counting problem is *#P-complete* if the problem of counting the number of satisfying assignments to a 3-SAT problem can be reduced to it in polynomial time. #P-complete problems, which are analog counting counterparts of NP-complete problems, are considered very difficult, especially in the view of Toda's results [31], which implies that one call to a #P-complete oracle suffices to solve any problem in the polynomial hierarchy in deterministic polynomial time.

[8] See [16] for a recent survey of this method.

[9] Strictly speaking, $\hat{\delta}_p$ is *not* an unbiased estimator for δ_p, since the routine only generates *almost* uniform linear extensions. The incurred bias is insignificant and often simply ignored in Markov chain Monte Carlo analysis.

with variance $(\operatorname{Var} Y)/k$. We can derive a confidence interval for δ_p as follows. Let t be the ratio of Y's variance and square of its expectation: $t = \operatorname{Var} Y/(\mathrm{E}Y)^2$, a non-negative quantity that can usually be bounded above by τ, which is polynomial in terms of n, the input size. Thus $\operatorname{Var} Y \leqslant \tau(\mathrm{E}Y)^2$ and $\operatorname{Var}\hat{\delta}_p \leqslant (\tau(\mathrm{E}Y)^2)/k = (\tau\delta_p^2)/k$. For any positive number c, Chebysev's inequality states that:

$$\Pr\bigl((\hat{\delta}_p - \delta_p)^2 > c \operatorname{Var}\hat{\delta}_p\bigr) \leqslant 1/c,$$

and thus:

$$\Pr\bigl((\hat{\delta}_p - \delta_p)^2 > c\tau\delta_p^2/k\bigr) \leqslant 1/c,$$

or equivalently:

$$\Pr\bigl(\bigl(1 - \sqrt{c\tau/k}\bigr)\delta_p \leqslant \hat{\delta}_p \leqslant \bigl(1 + \sqrt{c\tau/k}\bigr)\delta_p\bigr) \geqslant 1 - 1/c.$$

As a consequence, if we want our estimator $\hat{\delta}_p$ to be within a multiplicative factor of $1 + \varepsilon$ of δ_p with probability of at least $1 - 1/c$, it is sufficient to take a sample of size $k = \lceil 4c\tau/\varepsilon^2 \rceil$.

4.3. Further notes on distance measures on partial orders

To our knowledge, there is no general theory that addresses the problem of defining distance measures on partial orders. The closest to such theory we found is Critchlow's monograph [7]. Recall that on the strict orders on a finite set, there are a number of well-studied metrics such as Spearman's rho, Spearman's footrule, Kendall's tau, Ulam's distance, Hamming distance, and Cayley's distance. In [7], strict orders are referred to as *fully ranked data*, since they fully rank a set of items of interest. Critchlow extended the aforementioned six metrics to *partially ranked data*. Here, partially ranked data refer to certain *special cases* of partial orders. For example, they may correspond to the case when a person lists his first through kth choices, where $k < n$ (n is the number of decision outcomes). Note that if fully ranked data are identified with the elements of the permutation group (or, in layman's term, the permutations of Ω), then the above partially ranked data can be identified with points in a *coset space* of the permutation group. As a consequence, this special case of partial orders facilitates several group-theoretic techniques and thus makes the extended six metrics more amenable to analysis and computation.

Without a general theory of metrics on partial orders, researchers often extend metrics such as the Pearson correlation coefficient to partial orders in some simple way. For example, consider the GROUPLENS collaborative filtering system. Each user of the system has rated a set of news articles, and different users have rated different sets of articles. The similarity weight between two users is taken to be the Pearson correlation coefficient between the ratings over the articles both have rated (i.e., the intersection of the two rating sets). This solution could be unsatisfactory because it is insensitive to the number of articles rated by both users. Two users having the same rating on the only one article they have both rated would be maximally correlated, while their preferences may conceivably be quite different. This intuition is confirmed by the experiments with DIVA [24] (see Section 6.1 for a description of DIVA and summary of the findings). Recent research on GROUPLENS acknowledged this difficulty and proposed a *significance weighting* scheme to account

for the size s of the intersection set [13]. Basically, the Pearson correlation coefficient is multiplied with a significance weight of $s/50$ if $s < 50$. The modified similarity measure was empirically shown to provide more accurate recommendation [13]. Another approach to address the problem of small intersection was proposed by Breese et al. [2]. In this approach, the correlation is computed over the *union*, instead of the intersection of the two sets of ratings. This is made possible by assigning some default ratings to items that are in the union but not in the intersection (i.e., the symmetric difference of the two sets).

5. The case of uncertainty

Our discussion thus far has dealt with similarity of preferences in the case of certainty. The area of collaborative filtering is rich with examples of implemented systems that make use of such similarity measures, e.g., GROUPLENS [26] and the DIVA video recommender [24]. But many potential applications require the ability to define similarity of preference under uncertainty. Examples include medical decision making, financial decision making, and travel planning. (An effort along this line can be found in the work of Chajewska et al. [5], who cast the problem of eliciting utilities as a classification problem.) In order to serve as a basis for developing applications, it is important to have a measure of preference similarity that is both intuitive and amenable to efficient computation. The following example highlights some of the difficulties with defining appropriate similarity measures on preferences, when "uncertainty aspects" such as attitude toward risk have to be considered.

Miyamoto and Eraker [23] described a psychology experiment with 44 undergraduate students at the University of Michigan. The experiment is designed to test several assumptions about people's preferences and attitudes toward risks with regard to survival duration. The subjects were asked to assign certainty equivalences (CE) to a total of 42 standard gamble questions (SGQ) involving duration of survival. Below is a typical question:

> For any non-negative number n, let n be the event that you will live exactly n more years in good health, and then have a sudden and relatively painless death. Let $(m, .5, n)$, $0 \leqslant m < n$, be a lottery of 50% chance for m and 50% chance for n. What is the number p for which you regard $(m, .5, n)$ and p as equivalent (denoted $(m, .5, n) \sim p$)?

Suppose that u denotes the utility function of a subject. Each answer of the form $(m, .5, n) \sim p$ translates into the following constraint on u: $u(m) + u(n) = 2u(p)$. Thus for each subject, we have a set of 42 constraints on his/her utility function u. Given two subjects with utility functions u and u', how should we define a distance measure between u and u'? A simplistic approach may use some well-known statistical measures such as Spearman's footrule, Ulam distance, or various correlation coefficients. The problem with this approach is similar to the problem that plagues these statistical measures in the case of certainty. Specifically, such a distance measure can be defined only on (and thus sensitive to the size of) the intersection of the two sets of CE questions constraining u and u'. In addition, it is conceptually difficult to see how well, compared to each other, these

statistical measures can capture intricate aspects of preferences such as attitude toward risk. Another possible approach is to completely determine u and u' (using methods such as interpolation, curve-fitting, or parameter estimation), and compute the distance between two completely specified utility functions. We believe that because of the strong assumptions required to determine the complete utility functions, the suitability of this approach can only be determined on a case-by-case basis. As we shall show in this section, the probabilistic distance provides a principled solution for this problem that can be used in a wide range of other problems as well.

5.1. The probabilistic distance on complete preferences

Let $\prec_1$ and $\prec_2$ be two preference orders on the set $\mathcal{S}$ of prospects. The probabilistic distance is defined as:

$$\delta(\prec_1, \prec_2) = E\big[c_{\prec_1,\prec_2}(p,q)\big] = \int_{\mathcal{D}}\int_{\mathcal{D}} c_{\prec_1,\prec_2}(p,q)\,\partial p\,\partial q. \tag{4}$$

where p and q are two independent identically distributed uniform random variables on the set $\mathcal{D}$ of decision consequences. Note that the set $\mathcal{D}$ of decision consequences is a simplex in a multi-dimensional Euclidean space. A uniform random variable on $\mathcal{D}$ can be defined using the standard method of measure and probability theory.

Example 2. Let $\Omega = \{1,2,3\}$ and $\prec_\alpha, 0 \leqslant \alpha \leqslant 1$ be preference orders on the set $\mathcal{S}$ of all prospects over Ω with corresponding utility functions $u_\alpha = (0,\alpha,1)$ (i.e., $u_\alpha(1) = 0$, $u_\alpha(2) = \alpha$, and $u_\alpha(3) = 1$). Then for $0 \leqslant \beta \leqslant \alpha \leqslant 1$[10]

$$\delta(\prec_\alpha, \prec_\beta) = d(\alpha,\beta) := \frac{\alpha-\beta}{3}.$$

Let $\beta = 0$. Observe that $d(0,0) = 0$, $d(.5,0) = 1/6$, $d(1,0) = 1/3$. As α increases, the utility function u_α becomes less and less similar to u_0. This is reflected in the monotonic increase of $d(\alpha,0) = \alpha/3$, which tends to $1/3$ as α tends to 1.

When $\mathcal{D} = \mathcal{S}$, i.e., the set of decision consequences is the same as the set of all prospects, computing the above integral amounts to computing the volume of a polytope in the $(2n-2)$-dimension space (both p and q have $n-1$ coordinates that can vary). While computing the exact volume of a polytope in general is computationally complex [1,10], there is a simple Monte Carlo approximation algorithm for this particular problem. This algorithm works by sampling p_i, $i = 1,2,\ldots,k$, and q_i, $i = 1,2,\ldots,k$, according to the uniform distribution on $\mathcal{S}$, and taking the average $\bar{c} = \frac{1}{k}\sum_{i=1}^{k} c_{\prec_1,\prec_2}(p_i,q_i)$. With a sufficiently big sample size k, the sample mean $\bar{c}$ can approximate $\delta(\prec_1,\prec_2)$ with arbitrary degree of precision, according to the Weak Law of Large Numbers. Sampling p_i and q_i according to the uniform distribution on $\mathcal{S}$ is basically the well-studied problem of *random*

[10] The computation of this distance is straightforward with the help of the Maple® symbolic algebra package, albeit rather tedious.

Table 1
Algorithm for uniform sampling on $\mathcal{S}$

1. Generate $n-1$ numbers x_i, $i = 1, 2, \ldots, n-1$, according to $n-1$ independent uniform random variables on $[0, 1]$.

2. Sort x_i's: $0 \leqslant x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n-1)} \leqslant 1$. This is the *order statistics* of the sample. Let $x_{(0)} = 0$ and $x_{(n)} = 1$.

3. Let $p_i = x_{(i)} - x_{(i-1)}$, $i = 1, 2, \ldots, n$. (p_i's are called the *spacings* of the sample.) Return $(p_1, p_2, \ldots, p_n)$.

division of the unit interval and can be performed using the algorithm in Table 1. This algorithm has a time complexity of $O(n \log n)$. See [8,25] for more details.

The probabilistic distance between two preference orders, defined this way, depends only on the orders. It can be computed given the two orders, or two utility functions that are consistent with the two orders. This definition can be useful when the two preference orders, or the two consistent utility functions are given, but little is known about the available decision alternatives. When we have more information about the decision alternatives and their consequences, it is desirable that we tailor the definition of the probabilistic distance to reflect this knowledge. So if the set $\mathcal{D}$ of decision consequences is finite and known, the probabilistic distance can be defined as:

$$\delta(\prec_1, \prec_2) = \frac{\sum_{(p,q)\in\mathcal{D}^2} c_{\prec_1,\prec_2}(p, q)}{|\mathcal{D}|^2}. \tag{5}$$

The computation of this (discrete) formula is obviously much simpler than the integral formula of Eq. (4), *provided that* we know the set of decision alternatives $\mathcal{D}$. In addition, it is a subtle issue to determine which decision alternatives to include in $\mathcal{D}$ in the above definition. We may also replace $|\mathcal{D}|^2$ with $|\mathcal{D}|(|\mathcal{D}| - 1)$ so that the distance δ scales to the range of $[0, 1]$.

5.2. The probabilistic distance on partial preferences

Let $\prec_1$ and $\prec_2$ be the partial preference orders of two persons, A_1 and A_2. Recall that the probabilistic distance $\delta(\prec_1, \prec_2)$ is defined as:

$$\delta(\prec_1, \prec_2) = E\big[\delta(\prec_{f_1}, \prec_{f_2})\big],$$

where f_1, f_2 are uniform random variables on E_1, E_2, the sets of weak order extensions of $\prec_1$, $\prec_2$, respectively. Exactly how should we interpret this definition? In the certainty case, this is easy since E_1 and E_2 are finite sets (a finite poset has only finitely many extensions) and we can just take the average of $\{\delta(\prec_{f_1}, \prec_{f_2}) \mid f_1 \in E_1, f_2 \in E_2\}$. But in the case of uncertainty, the set E_1 and E_2 are typically infinite. For example, consider a typical partial preference elicitation process. We may have determined that the utility function of A_1 is additive over two binary attributes $\{x_1, x_2\}$:

$$u(x) = k_1 u_1(x_1) + k_2 u_2(x_2), \quad k_1, k_2 \geqslant 0. \tag{6}$$

In addition, we have also elicited the sub-utility functions u_1, u_2. We have not, however, assessed the scaling constants (or trade-off coefficients) k_1, k_2. The set E_1 is thus the set of all utility functions of the form in Eq. (6), which is obviously infinite.

5.2.1. Partial utility functions as polyhedral cones

Defining the expectation of a quantity involving random variables over infinite, multi-dimensional domains often requires the language and formalism of measure theory. With a simplifying assumption, however, we can define the probabilistic distance δ using more elementary concepts. Note that since a utility function $u : \Omega \to \Re$ can be viewed as a point in the n-dimensional Euclidean space $\Re^n$: $u = (u(1), u(2), \ldots, u(n))$, we can (and will) talk about the sets E_1, E_2 of consistent utility functions as sets of points in $\Re^n$. The simplifying assumption we shall make regarding E_1, E_2 is that they are determined by linear, homogeneous inequalities. Formally, they are sets of the forms

$$\left\{\vec{u} \in \Re^n \mid A\vec{u} \leqslant \vec{0}\right\}, \tag{7}$$

where A is some $m \times n$ matrix of real numbers, and $\vec{0}$ is the $m \times 1$ zero vector. In geometric terms, such a set is the intersection of m *half-spaces*, each of which crosses the origin and having one of the rows of matrix A as its *outward normal vector*, and is called a *polyhedral cone*. Partial utility functions satisfying the above assumption encompass most of the common kinds of partial utility functions encountered in the practice of decision analysis. For example, a multi-linear utility function with known sub-utility functions and unknown scaling coefficients, a model studied in [11], satisfies this assumption. It is not difficult to see that the same is true for multiplicative and additive utility functions with known sub-utility functions and unknown scaling constants. Furthermore, a constraint on the partial preference order $\preceq$ of the form $p \preceq q$, for some $p, q \in \mathcal{S}$ would also translate to a homogeneous linear inequality: $\langle u, p - q\rangle \leqslant 0$.

The nice thing of having E_1 and E_2 as polyhedral cones is that in the defining formula of the probabilistic distance

$$\delta(\prec_1, \prec_2) = E\left[\delta(\prec_{f_1}, \prec_{f_2})\right] = \int_{E_1}\int_{E_2}\int_{\mathcal{D}}\int_{\mathcal{D}} c_{\prec_{f_1},\prec_{f_2}}(p, q)\,\partial f_1\,\partial f_2\,\partial p\,\partial q,$$

we can interpret the integral on the right hand side as the *volume* of a bounded polyhedral cone in some multi-dimensional Euclidean space. But more importantly, we can reduce the problem of computing the probabilistic distance on partially specified utility functions to the well-studied problem of computing the volume of polyhedral cones. (In fact, the problem of computing the probabilistic distance on partial orders in the certainty case can also be reduced to the volume-computing problem, using some elementary geometric arguments.)

5.2.2. Computing the volume of convex bodies

The problem of computing the volume of convex bodies has received considerable interest in the theoretical computer science community in the past fifteen years. Early results were negative for the prospect of finding an efficient deterministic algorithm [1]. But randomization techniques once again come to the rescue. The first work that uses

randomization to obtain a polynomial time algorithm for this problem is due to Dyer et al. [9]. A series of work followed and refined the algorithm of Dyer et al., substantially reducing its complexity [22]. These results are all based on various Markov chain-based sampling techniques that sample points from the convex body according to a nearly uniform distribution. The convex body is input to the algorithm by means of a *membership oracle*, i.e., a black box that answers whether a given point belongs to the convex body. Note that this requirement fits excellently with the assumption that the set E_1, E_2 are polyhedral cones determined by a set of homogeneous linear inequalities as in Eq. (7): we can check if a utility function $\vec{u}$ is consistent if $A\vec{u} \leqslant \vec{0}$ in time $O(m)$ (recall that m is the number of rows of A).

We now sketch out the main ideas behind the sampling algorithm. To sample uniformly from a convex body K, we perform a random walk on the points of K. Starting at an arbitrary point inside K, we move at each step to a uniformly selected random point in a ball of radius ε about the current point if this remains inside K. If the new point is outside K, we remain where we were. The size ε of the radius is typically $1/\sqrt{n}$. It follows from elementary Markov chain theory that the distribution of the point after t steps tends to the uniform distribution as t tends to infinity. The crucial issue is, how long to walk before the walking point becomes nearly uniformly distributed? There are two reasons for needing a long walk: we have to get to the "distant parts" of K, and we may get stuck in "corners", especially "sharp corners" of K. The first reason suggests that we choose a step-size that is large enough relative to the diameter of K, while the probability of the second can be reduced by choosing a small step-size. A number of advanced techniques have been developed to address this dilemma to ensure that the Markov chain settles quickly to a nearly uniform distribution (in technical terms, such a chain is called *rapidly mixing*). See Lovász et al. [22] for a comprehensive treatment of this topic.

While this Markov chain-based sampling algorithm was developed for the purpose of computing the volume of convex bodies (and thus can be used to compute the volume of the polyhedron that is $\delta(\prec_1, \prec_2)$), we can use it directly to perform a Monte Carlo estimation of the probabilistic distance on partial utility functions. Specifically, we can estimate $\delta(\prec_1, \prec_2)$ by sampling f_{ij}, $i = 1, 2$, $j = 1, 2, \ldots, k$, according to nearly uniform distributions on E_i, and taking the average $\bar{\delta} = \frac{1}{k}\sum_{i=1}^{k} \delta(f_{1j}, f_{2j})$. Again, the Weak Law of Large Numbers ensures that with a sufficiently large sample size k, the sample mean $\bar{\delta}$ can approximate $\delta(\prec_1, \prec_2)$ with arbitrary degree of precision.

5.3. An illustrative example

We illustrate the algorithm to compute the probabilistic distance on partially specified utility functions. The data we use are taken from the psychology experiment by Miyamoto and Eraker [23], as described at the beginning of Section 5. Out of the 44 subjects, 6 were dropped due to failure to complete the interview in the allocated time, or failure to understand the CE task. The effective sample size is thus 38. There are a total of 42 CE questions for each subject (see Table 2). Note that in this experiment, subjects are not asked about their preferences with respect to specific decision alternatives. As a consequence, it is not possible to define, based on this data set, a distance measure that requires the knowledge of the decision alternatives (Eq. (5)).

Table 2
Table describing the 42 standard gamble questions used in the psychology experiment by Miyamoto and Eraker [23]. X/Y denotes a 50/50 gamble between X and Y years of survival

Basic	Times 2	Times 3	Plus 10	Plus 20	Zero
1/10	2/20	3/30	11/20	21/30	0/32
2/10	4/20	6/30	12/20	22/30	0/36
3/10	6/20	9/30	13/20	23/30	
4/10	8/20	12/30	14/20	24/30	
1/12	2/24	3/36	11/22	21/32	
2/12	4/24	6/36	12/22	22/32	
3/12	6/24	9/36	13/22	23/32	
4/12	8/24	12/36	14/22	24/32	

5.3.1. Computing the probabilistic distance on the subjects for clustering

Since the survival duration in the CE questions ranges from 0 to 36, we scale the utility functions so that $u(0) = 0$ and $u(36) = 1$. The next step is to discretize the outcome space, which is discretizing the number of years of survival. This is necessary because our framework requires that the set of decision outcomes is finite. Because each subject gave four different answers (at four different time points) to each CE question, we take the average of the 4 answers as the CE. Because each answer is either an integer or an integer plus 0.5 (e.g., $(1, .5, 10) \sim 4.5$), we discretize the number of years of survival to the granularity of $1/8$,[11] resulting in $36 \times 8 + 1 = 289$ outcomes. We also assume that all subjects prefer longer survival to shorter survival: $u(i/8) \leqslant u((i+1)/8)$, $i = 0, 1, \ldots, 287$. Framed this way, the utility function u of each subject has a total of 288 inequality constraints and $42+2 = 44$ equality constraints. It is easy to see that these linear constraints determine a convex set of consistent utility functions.

To find a starting point for the random walk, we need to find a consistent utility function, i.e., a feasible solution for the linear constraints. For this we use the linear programming facility LINPROG of Matlab® Optimization Toolbox, with some randomly generated target function. Interestingly, we found that out of the 38 subjects, only 3 provided consistent answers; the rest provided answers that lead to linear programs that are infeasible. This inconsistency can be attributed to the fact that the expected utility paradigm is normative but not descriptive [17]. An example of this school of thought is the approach called *subjective expected utility* (SEU) [32], according to which a CE statement $(m, .5, n) \sim p$ translates into the equation: $(1 - w(.5))u(m) + w(.5)u(n) = u(p)$. Here $0 < w(.5) < 1$ is the *probability distortion* for a .5 probability applying to the superior outcome. Note that in the standard expected utility paradigm, $w(.5) = .5$.

But even with more general utility models such as SEU, it is likely that subjects will have inconsistent preferences, due to *variations in subject responses*. Our approach is to stay within the standard expected utility paradigm and account for the inconsistency in some way. While the fact that random error in judgment exists is well-known, the question

[11] For example, if the four answers for the (1, .5, 10) question are (4, 4, 4, 4.5), then the average of these four answers is 4.125.

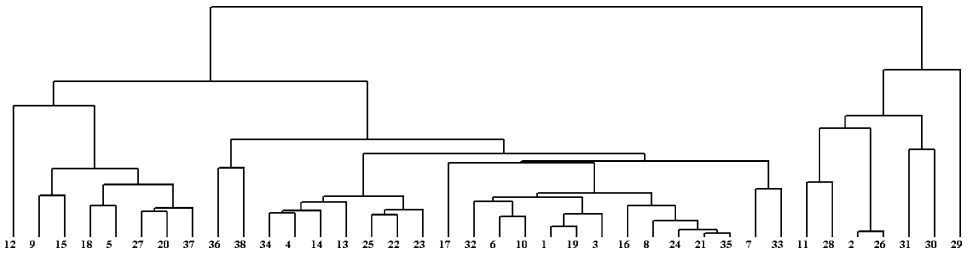

Fig. 1. Hierarchical cluster of the 38 subjects.

of how to deal with it remains open. For the purpose of our experiment, we take the following simple approach. We keep all of the 288 inequality constraints that capture the "longer survival is better" assumption. For each subject, from the set of the 42 equality constraints provided by the CE answers, we incrementally randomly add one at a time to LINPROG and keep doing this as long as a feasible solution exists. Note that due to differences between subjects' responses and the randomness of this method, different sets of CE answers may be taken into account for different subjects. Fortunately, this is not a problem for the probabilistic distance.

Now that a set of consistent CE answers is selected for each subject, we simultaneously start 38 random walks from 38 consistent utility functions, one for each subject. The radius ε of the ball is initialized to 0.001. At each iteration, we generate a random point in each ball of radius ε. If the generated point is consistent with the constraints, we move to the new point and mark the iteration a *success*; otherwise we stay at the current location and call the iteration a *failure*. If two successes occur consecutively, we double the radius. If two failures occur consecutively, we halve the radius. We stop the random walk after 1000 iterations, at which point we obtain a random sample of consistent utility functions for the 38 subjects. We compute the distance between any two consistent utility functions and record the distances in a square dissimilarity matrix of size 38×38. This computation is performed by a routine that implements the algorithm in Table 1. We repeat the whole process for a total of 1000 times, updating the averages of the distances as we go. We can use this distance matrix, for example, to compute *clusters* of subjects. Fig. 1 shows the hierarchical cluster obtained using the ClustanGraphics® package. The method used was average-linkage.[12]

5.3.2. Relation to attitude toward risk

Observe that in this psychology experiment, since there is a single attribute—the duration of survival in number of years—the difference of one subject's preference from another's is in essence the difference in *attitudes toward risk* (*ATR*). It is thus interesting to see if there is some correlation between ATR and the probabilistic distance we just computed.

[12] All of the code was written in Java™ and the mathematical programming language of MatLab®. The computations were performed on an Athlon™@850Mhz system with 512MB RAM running Windows® 2000, and took about 30 minutes to finish.

How should we define and measure ATR? We take the following straightforward approach. Consider a SGQ $(l, .5, u)$ to which the answer from a subject X is ce, i.e., $(l, .5, u) \sim ce$, and $l < ce < u$. Define the proportional match of this SGQ as:

$$pm = \frac{ce - l}{u - l}.$$

Intuitively, subject X is risk-averse, risk-neutral, and risk-seeking respectively if $pm < .5$, $pm = .5$, and $pm > .5$ respectively. We thus may define the attitude toward risk of X as the sample average of pm, averaging over 42 SGQ's:

$$ATR(X) = \text{avg}\{pm \mid 42 \text{ answers}\}.$$

The ATR values for the 38 subjects range from .16 to .75, with an average of .448 and standard deviation of .126. We analyze the correlation between ATR and the probabilistic distance in the following way. We consider each ordered triple of subjects (X, Y, Z). We look at the probabilistic distance $\delta(X, Y)$ and $\delta(X, Z)$ to see which of the two subjects Y and Z is closer to subject X. We then look at the ATR of X, Y, and Z to see which of the two ATR's of subjects Y and Z is closer to the ATR of X (i.e., which of the two quantities $|ATR(Y) - ATR(X)|$ and $|ATR(Z) - ATR(X)|$ is smaller). If the answers in both instances match, we mark the ordered triple (X, Y, Z) as OK. We then compute the percentage of ordered triple of subjects that is marked OK out of all possible ordered triples. This percentage is in between 72% and 80%, depending on the number of iterations used in computing the probabilistic distance. This shows that there is a strong correlation between the probabilistic distance and the ATR, since the correlation between ATR and a random distance measure can be shown to be approximately 33%.

5.4. Related work on similarity measures on utility functions

The only existing similarity measure on preferences in the case of uncertainty that we are aware of is defined in Chajewska et al. [5]. This measure is also based on a finite set of decision alternatives. But in contrast to the probabilistic distance that is defined based on the *preference orders*, this measure is defined based on *consistent utility functions*. Specifically, let u_1 and u_2 be two utility functions, and the decision alternatives be $\{p_1, p_2, \ldots, p_m\}$, indexed in such a way that $i = \text{argmax}_{j:1 \leqslant j \leqslant m} \langle u_i, p_j \rangle$, $i = 1, 2$. This means that according to the utility function u_i, p_i is an optimal decision alternative, for $i = 1, 2$. The distance between u_1 and u_2 is defined as

$$d(u_1, u_2) = \frac{\langle u_1, p_1 \rangle - \langle u_1, p_2 \rangle + \langle u_2, p_2 \rangle - \langle u_2, p_1 \rangle}{2}.$$

The difference of the first two terms in the above numerator, $\langle u_1, p_1 \rangle - \langle u_1, p_2 \rangle$, is called the *utility loss* of u_1 with respect to u_2, and that of the last two terms is the *utility loss* of u_2 with respect to u_1. The utility loss of one utility function with respect to another is the difference of, or loss in expected utility by choosing a decision alternative that is optimal according to the latter instead of the former.

There are several issues with this definition of distance between utility functions. First, this is a similarity measure between utility functions. If it is to be used as a measure of similarity between preference orders, one must deal with the issue of choosing the

corresponding consistent utility functions. A standard solution is to scale the consistent functions to the range of [0, 1] (a sort of the equivalence of canonical value functions for strict orders in the definition of Spearman's rho in the case of certainty). The second issue is that since this measure focuses on optimal decision consequences, it can become vacuous if there is a clear optimal decision alternative, a "clear winner" among the competing candidates. Formally, suppose that p_1 is the optimal decision alternative according to both u_1 and u_2, then $d(u_1, u_2) = 0$. The implication here is that the two preference orders are maximally similar, while the truth beneath is that they only agree on choosing the "clear winner". On the other hand, depending on the intended use, this measure can still be useful—imagine that it is used only under circumstances when there is no "clear winner". The third, mainly technical issue is that, strictly speaking, this distance is not well-defined, since $\operatorname{argmax}_{j:\,1\leqslant j\leqslant m}\langle u_i, p_j\rangle$, $i = 1, 2$, are not well-defined. Suppose for example that in addition to p_1, p_3 is also an optimal decision alternative according to the utility function u_1. Replacing p_1 with p_3 in the defining formula for this distance may result in a different value, since $\langle u_2, p_3\rangle$ may be different from $\langle u_2, p_1\rangle$. But again, depending on the intended use of the distance measure, this variance may not play an important role. Finally, as Chajewska et al. noted, this distance measure is not a metric since it does not satisfy the triangle inequality.

What other kinds of similarity measure can be defined on preferences in the uncertainty case? An immediate thought that comes to one's mind is to extend measures such as Pearson's coefficient, Spearman's footrule, etc. to accommodate utility functions. This generalization, however, has another difficulty beside the issue of scaling the utility functions. To see why, let us consider how we might define Spearman's footrule on utility functions:

$$F(u_1, u_2) = \sum_{i=1}^{n} \big|u_1(i) - u_2(i)\big|. \tag{8}$$

As Chajewska et al. noted, this approach gives all outcomes equal weight: a difference in utility for a highly probable outcome contributes the same to F as the same difference in utility for a highly improbable outcome. In other words, generalizing similarity measures to utility functions this way cannot account for any *a priori* knowledge about the resulting decision consequences.

It is interesting to see what happens to the Pearson correlation coefficient on utility functions. Since strategically equivalent utility functions are positive linear transformations of each other, they are maximally similar (or perfectly correlated) according to the Pearson's measure.

6. Applications

6.1. The decision-theoretic video advisor

Nguyen and Haddawy [24] describe the Decision-Theoretic Video Advisor (DIVA), a collaborative filtering system that uses the probabilistic distance measure to determine similarity of user preferences over movies. DIVA represents user preferences as complete

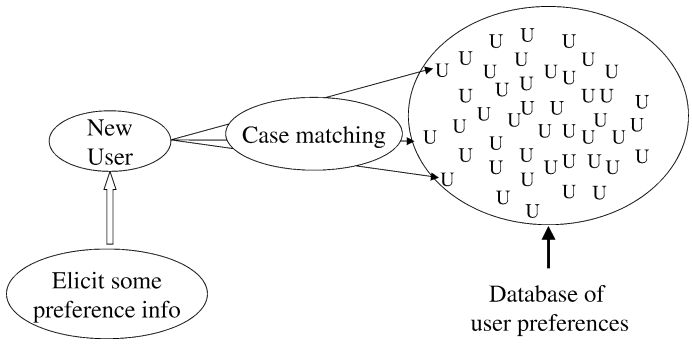

Fig. 2. The case-based approach to preference elicitation.

or partial orders over the set of all movies in its database. It maintains a population of users with their preferences over movies partially or completely specified. When a new user A comes to the system, it elicits some preference information from A and then determines which user in the population has a preference structure most similar to that of A. It then uses that preference structure to supplement A's directly expressed preferences, resulting in a complete order over movies. A recommendation list of the top N ranked movies is then displayed to A. If an old user accesses the system, DIVA simply retrieves that users's stored preference structure. Fig. 2 provides a schematic view of this approach.

In DIVA, partial preference orders of users consist of pair-wise preferences, which are obtained by having the user classify movies into like, ok, and dislike categories. DIVA also has an interface where the user can provide feedback on recommended movies. The feedback options are: (1) I have seen this movie already and liked it, (2) I have seen this movie already and did not like it, and (3) I have not seen this movie, but know that I would not like to see it (interpreted as dislike). Pairwise preferences are obtained by inferring that every movie classified as like is preferred to those classified as ok or dislike and, similarly, every movie classified as ok is preferred to every one classified as dislike.

The probabilistic distance is computed by using an implementation of the sampling algorithm by Bubley and Dyer (see Appendix A). First the similarity between the active user and each user in the population is computed. Then the most similar user is selected and the sampled linear extension of the active user most similar to that user is chosen as the complete representation of the active user's preferences. In this way the sampling process is used to both compute similarity and to generate predicted preferences.

Nguyen and Haddawy compared the performance of DIVA with that of the GroupLens collaborative filtering algorithm [26] using the EachMovie database, courtesy of the Digital Equipment Corporation. The GroupLens algorithm works with preferences represented as numeric ratings and uses Pearson's correlation coefficient to compute similarity among users. It predicts user preferences by computing a weighted sum of the correlation coefficient and the ratings of positively and negatively correlated users. The results of the comparison showed that the algorithm used in DIVA outperformed the GroupLens algorithm in terms of both precision and recall.

6.2. Preference elicitation via theory refinement

While originating from our work on case-based preference elicitation, the notion of probabilistic distance also found application in a different area: theory refinement for preference elicitation [12]. To reduce the complexity of preference elicitation, traditional approaches from Decision Analysis and Multiple Criteria Decision Making make assumptions concerning the structure of preferences (e.g., monotonicity or independence) and then perform elicitation within the constraints of those assumptions. But inaccurate assumptions can result in inaccurate elicitation. Nevertheless, assumptions can be a useful guide if they at least approximately apply to some large segment of the population. Ideally we need a method of using assumptions to guide but not constrain the elicitation process. This kind of functionality is provided by theory refinement techniques. The basic idea behind theory refinement is that we can start with a domain theory that may be approximate and incomplete and then correct for inaccuracies and incompleteness by training on examples. If the domain theory is at least approximately correct, we can learn faster with it than without it.

Haddawy et al. [12], explore the use of one particular theory refinement technique, Knowledge-Based Artificial Neural Networks (KBANN), to learn user preferences. In the case of certainty, they describe the problem of choosing a flight, where it is reasonable to make several assumptions about preferential independence and monotonicity. They then show how to represent these assumptions as Horn-clause theories that can be encoded in a KBANN network. This KBANN network can be trained to learn fine-grained preference structures from a variety of preferential data, including numeric ratings and simple binary classification. One of the main hypothesized advantages of the KBANN technique is its robustness to noise: the domain theory only needs to be approximately correct for KBANN to be useful. To evaluate this hypothesis in the flight selection domain, Haddawy et al. examine the performance of KBANN in learning preferences using examples generated from a number of value functions that violate the independence assumptions to various degrees. It is expected that the more a value function violates the domain theory, the worse the performance. The hypothesis is confirmed if this decrease in performance does not occur in a precipitous manner.

The main issue to be addressed in this robustness analysis is the definition of the *degree of violating the independence assumptions*, or *DOVI*. Haddawy et al. define this measure in the following way. Given a domain theory $\mathcal{D}$ and a value function u, note that $\mathcal{D}$ can be viewed as a set of value functions that satisfy $\mathcal{D}$. The DOVI measure of u violating $\mathcal{D}$ is defined to be the distance between u and the member of $\mathcal{D}$ that is closest to u:

$$V_{\mathcal{D}}(u) = \min_{f \in \mathcal{D}} \delta(f, u),$$

where δ is the probabilistic distance between value functions. Imagine this approach as an analogy to the definition of a distance from a point to a set of points in Euclidean geometry. This definition of degree of violating the domain theory is *semantic*, a departure from existing *syntactic* approaches used in robustness analysis of theory refinement techniques.

Fig. 3 shows the results of Haddawy et al.'s robustness analysis using seven preference orders with DOVI varying from 0.05 to 0.92. They analyze the performance of KBANN on these seven preference orders, using a training set size of 30, 50, 100, and 150, and

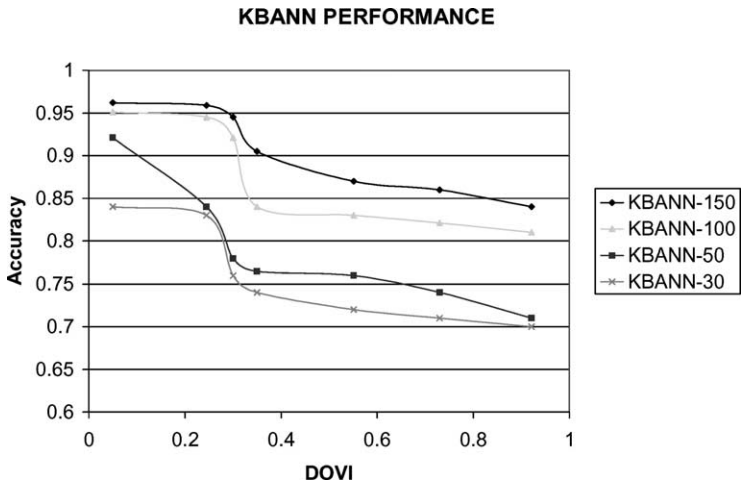

Fig. 3. Performance of KBANN for value functions of various degrees of violation of the independence domain theory. The number of training examples are 30, 50, 100, and 150. The size of the set of test examples is kept constant at 50.

keeping the test set size constant at 50. Notice that for any given number of training examples (30, 50, 100, 150), the performance of KBANN decreases as the DOVI increases. KBANN's performance when the DOVI is 0.24 is still very close to when the DOVI is 0.05. This implies that KBANN is fairly robust to some small amount of noise in the domain theory. The sharpest performance decrease occurs when the DOVI hits the range of 0.3 to 0.35. The fewer examples we use to train KBANN, the sharper this performance decrease. Fewer training examples means that KBANN is relying more on the (inaccurate) domain theory. In addition to confirming the robustness hypothesis, this experiment also justifies the probabilistic distance as a proper measure of distance on preferences: if it were not, the performance curves of KBANN would have been much more random.

7. Summary

In this paper, we study the problem of similarity measures on preferences from a decision-theoretic point of view. The focus of our investigation is the probabilistic distance, a measure of similarity among people's preferences that has its roots in the Kendall's tau function. Below are our findings:

- *The probabilistic distance is theoretically attractive.* It is a metric on complete preference orders, regardless of whether the decision problem involves uncertainty or not (Theorem 1).
- *It can be computed efficiently* (*exactly or approximately*). In the case of certainty, the probabilistic distance on partial preferences can be approximated using a randomized algorithm that samples uniformly randomly from the set of linear extensions of a

partial order (Section 4.2). Under uncertainty, the problem is innately harder, because of the complexity introduced by probabilities and utilities. We have shown that with the reasonable assumption that the set of consistent utility functions is linearly bounded, computing the probabilistic distance can be reduced to the well-studied problem of computing the volumes of convex bodies for which efficient approximate algorithms exist (Section 5.2).

In the case of certainty, the probabilistic distance theoretically appears to be better suited for use in recommendation systems than the predominant Pearson correlation coefficient measure. This is confirmed by the experiments with the Decision-Theoretic Video Advisor (Section 6.1). In the case of uncertainty, the probabilistic distance is the first similarity measure that is defined on partial preference orders. Because of its reliance on orders instead of utilities, the probabilistic distance can be defined and computed in a wide range of situations (Section 5).

In addition, we believe that the probabilistic distance has the potential to find applications beyond the context of case-based preference elicitation, since it is in its most general form a *distance measure on partial orders*—a topic that has not received adequate treatment. A manifestation of this statement is the use of the probabilistic distance in the robustness analysis of the KBANN network for user preference modeling (Section 6.2). In the context of group decision making, the probabilistic distance may also be used to aggregate preferences of multiple users into a single preference structure, thus playing the same role in the uncertainty case that the Kemeny distance plays in aggregating multiple rankings in the certainty case.

Acknowledgements

We thank John Miyamoto for providing the utility database described in Section 5, Hien Nguyen for her work on DIVA (Section 6.1), and Benjamin Geisler for his contribution on KBANN-based preference elicitation (Section 6.2).

Appendix A. Random generation of linear extensions of a partial order

We describe below an algorithm, due to Bubley and Dyer [4], that almost uniformly randomly generates linear extensions of a partial order. The algorithm has running time of $O(n^3 \log n\varepsilon^{-1})$, where n is the number of the elements of the partial order, and ε is the desired accuracy, which means that the generated random linear extension has a probability distribution that is within a total variation distance[13] of ε from the uniform distribution. The running time required to obtain a certain precision ε is often called the *mixing time* of the Markov chain. A Markov chain with a mixing time polynomial with respect to the input

[13] The total variation distance between two discrete distributions P, Q over a finite sample space S resembles the Spearman's footrule, and is defined as $d_{TV}(P, Q) = \frac{1}{2}\sum_{s\in S} |P(s) - Q(s)|$.

size (which is the number of elements of the partial order in this case) and ε^{-1} is called *rapidly mixing*.

Suppose that the partial order $\prec$ has n elements, and $N = \{1, 2, \ldots, n\}$. We encode the orderings of these elements with the permutations of the elements of N, and the set of linear extensions of $\prec$ by a subset $\mathcal{LE}(\prec)$ of the set of all permutations of the elements of N.

For a given concave probability distribution f on $\{1, 2, \ldots, n-1\}$, define a Markov chain $\mathcal{M}_f = \{S_t\}_{t \geqslant 0}$ on $\mathcal{LE}(\prec)$ as follows. At any time point $t \geqslant 0$, toss a fair coin. If the coin lands head, then let $S_{t+1} = S_t$. If the coin lands tail, then choose an index $i \in \{1, 2, \ldots, n-1\}$ according to the distribution f. If the permutation obtained from S_t by switching the ith and $(i+1)$-st elements of S_t is also a linear extension of $\prec$, i.e., an element of $\mathcal{LE}(\prec)$, then let S_{t+1} be this new permutation. Otherwise, let $S_{t+1} = S_t$.

It is easily seen that $\mathcal{M}_f$ is ergodic with uniform stationary distribution. When f is the uniform distribution on $\{1, 2, \ldots, n-1\}$, $\mathcal{M}_f$ is the Karzanov–Kachiyan chain with mixing time $O(n^5 \log n + n^4 \log \varepsilon^{-1})$ [18]. Bubley and Dyer showed that if f is defined as $f(i) = i(n-i)/K$, where $K = (n^3 - n)/6$, then $\mathcal{M}_f$ has mixing time of $O(n^3 \log n \varepsilon^{-1})$.

References

[1] I. Barany, Z. Furedi, Computing the volume is difficult, in: Proc. 18th Annual Symposium on Theory of Computing, 1986, pp. 442–447.
[2] J. Breese, D. Heckerman, C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, in: Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence, 1998, pp. 43–52.
[3] G. Brightwell, P. Winkler, Counting linear extensions, Order 8 (3) (1991) 225–242.
[4] R. Bubley, M. Dyer, Faster random generation of linear extensions, in: Proc. Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA, 1998, pp. 175–186.
[5] U. Chajewska, L. Getoor, J. Norman, Y. Shahar, Utility elicitation as a classification problem, in: Uncertainty in Artificial Intelligence 14 (UAI '98), pp. 79–88.
[6] H. Clifford, W. Stephenson, An Introduction to Numerical Taxonomy, Academic Press, New York, 1975.
[7] D.E. Critchlow, Metric Methods for Analyzing Partially Ranked Data, in: Lecture Notes in Statistics, Springer, Berlin, 1980.
[8] H. David, Order Statistics, 2nd Edition, John Wiley, New York, 1981.
[9] M. Dyer, A. Frieze, R. Kannan, A random polynomial-time algorithm for approximating the volume of convex bodies, J. ACM 38 (1) (1991) 1–17.
[10] G. Elekes, A geometric inequality and the complexity of computing volume, Discrete Comput. Geom. 1 (1986) 289–292.
[11] V. Ha, P. Haddawy, A hybrid approach to reasoning with partial preference models, in: Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999, pp. 263–270.
[12] P. Haddawy, V. Ha, A. Restificar, B. Geisler, and J. Miyamoto, Preference elicitation via theory refinement, J. Machine Learning Res. (Special Issue on Fusion of Domain Knowledge with Data for Decision Support), to appear.
[13] J. Herlocker, J. Konstan, A. Borchers, J. Riedl, An algorithmic framework for performing collaborative filtering, in: Proc. 1999 Conference on Research and Development in Information Retrieval, 1999.
[14] W. Hill, L. Stead, M. Rosenstein, G. Furnas, Recommending and evaluating choices in a virtual community of use, in: Proc. ACM CHI'95 Conference on Human Factors in Computing Systems, 1995, pp. 194–201.
[15] N. Jardine, R. Sibson, The construction of hierarchic and non-hierarchic classifications, Comput. J. 11 (1971) 117–184.
[16] M. Jerrum, A. Sinclair, The Markov chain Monte Carlo method: An approach to approximate counting and integration, in: D. Hochbaum (Ed.), Approximations for NP-hard Problems, PWS Publishing, Boston, MA, 1996, pp. 482–520.

[17] D. Kahneman, A. Tversky, Prospect theory: An analysis of decisions under risk, Econometrica 47 (1979) 276–287.
[18] A. Karzanov, L. Kachiyan, On the conductance of order Markov chains, Order 8 (1) (1991) 7–15.
[19] J. Kemeny, Mathematics without numbers, Daedalus 88 (1959) 571–591.
[20] M.G. Kendall, Rank Correlation Method, 3rd Edition, Hafner Publishing Company, New York, 1962.
[21] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon, J. Reidl, Applying collaborative filtering to usenet news, Comm. ACM 40 (4) (1997) 77–87.
[22] L. Lovasz, R. Kannan, M. Simonovits, Random walks and an $o*(n^5)$ volume algorithm for convex bodies, Random Structures and Algorithms 11 (1997) 1–50.
[23] J.M. Miyamoto, S.A. Eraker, Parametric models of the utility of survival: Tests of axioms in a generic utility framework, Organizational Behavior and Human Decision Processes 44 (1989) 166–202.
[24] H. Nguyen, P. Haddawy, The decision-theoretic interactive video advisor, in: The Fifteenth Conference on Uncertainty in Artificial Intelligence, 1999.
[25] R. Pyke, Spacings, J. Roy. Statist. Soc. B 27 (1965) 395–436. Discussion: 437–449.
[26] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, J. Reidl, Grouplens: An open architecture for collaborative filtering of netnews, in: Proc. 1994 Computer Supported Cooperative Work Conference, New York, 1994, pp. 175–186.
[27] U. Shardandand, P. Maes, Social information filtering: Algorithms for automating "word of mouth", in: Proc. CHI-95, 1995.
[28] A. Sinclair, Algorithms for Random Generation and Counting: A Markov Chain Approach, Birkhauser, Boston, 1993.
[29] C. Spearman, The proof and measurement for association between two things, Amer. J. Psychology 15 (1904) 72–101.
[30] C. Spearman, Footrule for measuring correlation, British J. Psychology 2 (1906) 89–108.
[31] S. Toda, On the computational power of pp and +p, in: Proc. 30th IEEE Symposium on Foundations of Computer Science, 1989, pp. 514–519.
[32] A. Tversky, D. Kahneman, Utility theory and additivity analysis of risky choices, J. Experimental Psychology 75 (1967) 27–36.
[33] S.M. Ulam, Future applications of mathematics in the natural sciences, American Mathematical Heritages: Algebra and Applied Mathematics, Texas Tech. University, Mathematics Series 13 (1981) 101–114.
[34] L.G. Valiant, The complexity of computing the permanent, Theoret. Comput. Sci. 8 (1979) 410–421.
[35] J. von Neumann, O. Morgenstern, Theory of Games and Economic Behavior, Princeton University Press, Princeton, 1944.