Noname manuscript No.
(will be inserted by the editor)

Modeling and Validating Stakeholder Preferences with Probabilistic Inversion

R.E.J Neslo, R.M. Cooke

6-9-2010

Abstract Despite the vast number of models that have been developed for analyzing stakeholders' preferences, it is difficult to find any true out-of-sample validation for these models. Based on the theory of rational preference, utilities are specific to the individual. Unlike subjective probability, there is no mechanism for changing utilities on the basis of observation, and no operation for getting people's utilities to converge. The proper goal of stakeholder preference modeling must therefore be the characterization of a population of stakeholders via a distribution over utility functions. Drawing on the theory of discrete choice and random utility theory, we apply probabilistic inversion methods to derive a distribution over utility functions. The utility functions may either attach to the choice alternatives directly, or may be functions of physical attributes. Because the utilities are inferred from discrete choice data, out of sample validation is enabled by splitting the data into a test set used to fit the model and a validation set. These techniques are illustrated using discrete choice data for valuation of health states

Keywords Out-of-sample Validation, Discrete Choice, Random Utility, Probabilistic Inversion, valuation of health states

R.E.J Neslo
Department of Mathematics, Delft University of Technology
Tel.: +31-1527883997
Fax: +31-1527851021
E-mail: r.e.j.neslo@tudelft.nl

R.M.Cooke
Resources for the future and Department of Mathematics, Delft University of Technology
appearing in Applied Stochastic Models in Business and Industry, 2011

1 Introduction

Savage [21] formulated axioms for rational preference of an individual, and showed that the preferences of a rational agent can be represented as expected utility, where the individual's subjective probability over possible states of the world is unique and the utility function over consequences is affine unique. The representation of a preference relation by a utility function implies that if an individual assigns higher (expected) utility to choice alternative a than to alternative b, then the individual prefers, selects or orders an alternative a above b.

Whereas the theory of subjective probability has flourished; it is fair to say that the modeling of utility has lagged behind. Without reviewing the activity in this area, suffice to say that, in our opinion, there are two major causes for this. First, modelling techniques like multi attribute utility theory (MAUT), multi criteria decision making (MCDM), analytical hierarchy process (AHP), etc. focus on capturing "the" utility function over choice alternatives. If an individual's utility function conformed to additional (rather severe) constraints, such a representation might be possible at an individual level. However, there is no reason to believe that "a" utility function exists for groups of individuals; the search for such is a fool's errand. Based on the theory of rational decision, the proper goal of utility modeling should be to capture the distribution over utility functions characterizing a group. Second, and perhaps not wholly unrelated, there has been a near total absence of attempts to validate the utility models produced by these various methods. As such, the field of applied utility theory remains parochial. There is a wealth of literature demonstrating that stakeholders often violate the axioms of rational preference. This does not threaten the normative status of the theory any more than the prevalence of invalid inferences imperils logic. However, it does lend urgency to the issue of validation.

The prospects for utility theory are brighter within the literature of discrete choice and random utility theory. Thurstone [25] pioneered this field with his celebrated law of comparative judgment. Assuming that utilities are normally distributed over a population of stakeholders, he fits the parameters of this distribution, under various correlation assumptions, using discrete choice data from pairwise comparisons. Later the Logit model was derived by Luce [14], [20], [15] under one of the consequences of his choice axioms namely the *Independence of Irrelevant Alternatives*(IIA). McFadden[17] also derived the Logit model under the random utility maximization principle. The Logit models assume that the error terms are generalized extreme value (GEV) distributed with mean zero and some constrained covariance matrix. Unhappy with the IIA assumption, McFadden and Train [18] picked up the thread of random utility maximization and extended the standard version of the Logit and Probit model, to deal with the limitations they pose. Whereas goodness of fit tests have been developed for many random utility models, true out-of-sample validation is not part of standard operating procedure.

The problem of inferring a distribution over utility functions from discrete choice data is a problem of probabilistic inversion. Theory tells us that each (rational) stakeholder has a utility function; if we knew these utility functions for a group of stakeholders we could predict the distribution of responses in discrete choice situations. We observe the distribution of responses and wish to infer the distribution over utility functions. Even more, we wish to model these utility functions as functions of physical attributes of the choice alternatives, and we wish to infer rather than impose dependence relations between utilities. This program is quite feasible, albeit that the techniques for solving probabilistic inversion problems are new to this field. A few applications are in press, or have been published [19], [2], [24], [12].

This paper introduces probabilistic inversion methods within the random utility community. The next section provides an intuitive introduction, followed in section 3 by an example on stakeholders' preferences and probabilistic inversion. An recent application on valuing health states is used to illustrate the methods. The application was designed and carried out by Dr. V. Flari [1]

[1] Dr. V.A. Flari, Policy and Regulation Programme B, Food and Environment Research Agency, Sand Hutton, York, YO41 1LZ

2 Intuitive Motivation

The goal of stakeholder preference modeling is to derive a distribution over the utility functions on a set of $\mathcal{A}$ of choice alternatives characterizing a population $\mathcal{S}$ of stakeholders. To motivate the approach we consider the health state valuation problem. Currently, health states are described and valued using EQ-5D[2]. EQ-5D is a standardized measure of health states developed by the EuroQol Group in order to provide a simple, generic measure of health for clinical and economic appraisal. Each health state is characterized by five criteria (mobility, self-care, usual activities, pain-discomfort, and anxiety-depression) which are measurable quantities that increase in a monotonic scale taking values one, two, three. An extended version (i.e. EQ-5D+C; see Table A.1 in the Appendix) of the system was introduced by Stouthard[23], which we will use. These different health states are possible outcomes of therapeutic procedures, and their valuation is critical in deciding which procedures to support and supply.

With 6 criteria taking 3 possible values, there are $3^6 = 729$ possible health states. The most direct approach would be to ask a group of stakeholders randomly chosen from the target population, to state their utility values for each health state. To estimate the distribution of utility functions, each stakeholder should apply a standardized utility scale with common zero and unit. The assessment burden for the stakeholders would be forbidding.

Although the distribution over utilities of these health states is the immediate goal; we want also to model the utility of health states in terms of the attribute scores. Existing approaches will often ask stakeholders to value the *attributes* and the *attribute scores*. This however is problematic for a number of reasons: (1) Whereas we choose health states in choosing a therapeutic procedure, we don't choose attributes as such. (2) The value attached to one attribute (eg "mobility") will depend on the whole set of attributes, as we must know what exactly falls under "usual activities" and "self care" to avoid double counting. Of course, valuing mobility score 3 versus mobility score 2 assumes that these values are unaffected by the values of other attributes. (3) It is unclear how the resulting distribution over health state utilities would be validated.

By adopting a simple model of the utility of a health state in terms of its attribute scores, we can simultaneously lighten the assessment burden and enable validation of the model. The score (utility) of health state i for subject s is modeled as:

$$u_s(a_i) = \sum_{j=1}^{6} \omega_{s,j} \times c_{i,j}; \; \sum_{j=1}^{6} \omega_{s,j} = 1; \; \omega_{s,j} > 0. \tag{2.1}$$

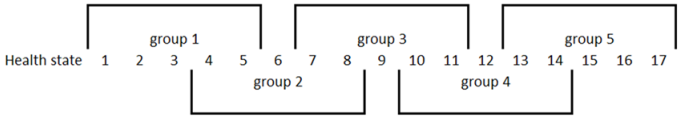
where $\omega_{s,j}$ is the weight for attribute j for subject s and $c_{i,j}$ is the score of health state i on attribute j. If this model is adequate, the distribution of utility functions over the set of stakeholders may be captured as a distribution over attribute weights $(\omega_1, ...\omega_6)$. If the model is not adequate, a better model must be sought. Instead of asking each stakeholder for his/her weight vector, we will ask them to rank order subsets of the 729 health states, and look for a distribution over weight vectors which recovers the pattern of rankings. Given a single ranking of a set of health states, we could in principle recover the set of weight vectors which would yield this ranking under model (2.1). For a set of rankings we could take the union of the corresponding sets of weight vectors. The uniform distribution over this set could be taken as our distribution over utility functions. Although this is a feasible approach, it is not the one we adopt. To understand why, we must discuss the discrete choice format.

729 health states is much too many for stakeholders to rank order. A *discrete choice format* is needed render the assessment burden bearable. Best practice suggest that stakeholders can rank at most 7 items at a time. The most popular discrete choice format is simple pairwise comparisons: subjects are presented with all pairs of choice alternatives and asked to choose one of the two offered. This is obviously infeasible for 729 alternatives, as there are 265,356 pairs. An economical

[2] Further information can be found in http://www.euroqol.org

discrete choice format is acceptable if it enables validation of the model on which it is based. For the present study we selected 17 non-dominated health states[3]. These 17 health states are broken into 5 overlapping groups of 5, where stakeholders then rank the health states in each group. The overlap structure is shown in Figure (2.1)

Fig. 2.1: Overlap structure



Some pairs of states occur in two groups of 5, which enables us to screen stakeholders for consistency. A stakeholder who ranks health state i above health state j in group k is called inconsistent if he ranks health state j above health state i in group $k+1$. With this discrete choice format we gather data which we then use to infer a distribution over the utility values. One of the goals in choosing a discrete choice format is to enable the analyst to identify inconsistent experts. Removing inconsistent experts can produce better results. Depending on the format, it may be unrealistic to expect perfect consistency. In pairwise comparisons, for example, it is usually sufficient that the number of inconsistencies (circular triads) is small enough to reject the hypothesis that a subject is choosing his/her preferences at random. Thus a solution technique for finding a distribution over utility vectors must be able to deal with inconsistency: even if there is no distribution over utility vectors which *exactly* reproduces the stakeholders' responses, it may be possible to minimize the lack of fit and thus arrive at a reasonably well validated model. Choosing a discrete choice format mixes science and craft, and as this approach to utility quantification is relatively recent, the craft is still evolving.

The necessity of dealing with inconsistencies drives our choice of solution technique. If the problem is feasible, the solution algorithm should converge to a unique solution, if the problem is not feasible, then it should converge to a unique distribution which minimizes lack of fit in some appropriate sense. We proceed as follows. For each group we define a square *preference ranking matrix* whose i, $j-th$ entry gives the proportion of stakeholders who ranked health state i in the $j-th$ position, in that group. The task is to find a distribution over the weights that reproduce the preference ranking matrices. Initially a distribution is chosen over the weights from which a large sample is drawn. For each sample weight vector we can compute how a stakeholder with that weight vector and model (2.1) would rank the health states in each 5-group. The entire sample will lead to 5 preference ranking matrices which will not agree with those from our stakeholder data. Iterative re-weighting schemes will then assign differential weights to the original sample such that if we resample this distribution using the weights, the resulting preference ranking matrices agree - as much as possible - with those of the stakeholders. If the problem is feasible then perfect agreement is possible and these techniques converge quickly. The solution is unique and is minimally informative with respect to the initial distribution over the weights. In case the problem is infeasible these techniques guarantee a solution that is the least infeasible in an appropriate sense. Solution techniques are explained in section 4.

Having obtained a solution using the solution techniques, we validate the solution by recovering the entries of the five preference ranking matrices using just a subset of the entries. In other words, we solve the model using a subset of entries and see if this model successfully predicts the remaining entries. We call this procedure out-of-sample validation. In the following sections we give detailed information about the discrete choice format, solution techniques and out-of-sample validation.

[3] The attribute scores increase in severity: pain score 3 is worse than pain score 1, etc. Health state i dominates health state j if $i's$ score on each criteria is greater than $j's$.

3 Stakeholders' Preference and Probabilistic Inversion

Fig. 3.1: Response to D_1 and $D_2 : u_s = (u_s(a_1), u_s(a_2), u_s(a_3), u_s(a_4))$, $G_1(u_s) = a_2, G_2(u_s) = a_4$

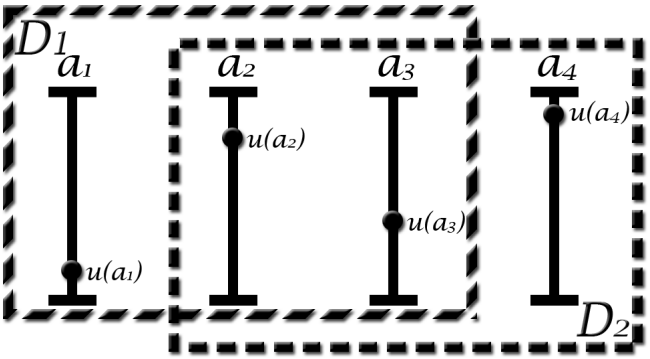


Figure (3.1) illustrates a mental projection of a stakeholder s faced with two groups of choice alternatives, D_1 and D_2. In this simple example, the stakeholder has to choose one most preferred alternative from each set. With the utilities as shown, s prefers alternative a_2 from subset D_1 and alternative a_4 from subset D_2. Section A.2 in the appendix gives a general formulation of the problem pictured in Figure (3.1). Another stakeholder might arrange the four alternatives differently on the utility scale, leading to other responses. In practice we can't observe the utility values that stakeholders assign to the alternatives, but we do observe their preferences in terms of responses to the subsets presented. Suppose that a set of stakeholders induce marginal distributions Q shown in Table (3.1.

Table 3.1: Marginal distribution Q over the responses

	a_1	a_2	a_3	a_4
D_1	0.5	0.3	0.2	N/A
D_2	N/A	0.25	0.5	0.25

If we knew the utility function of a stakeholder $s \in \mathcal{S}$, then we could obviously predict with certainty how s would respond to a discrete choice problem. Equivalently, if we are given a vector of utility values over the choice alternatives, we can uniquely determine how a stakeholder with that utility would respond in any discrete choice problem. Our problem is to infer a distribution over (standardized) utility functions given a set of responses from stakeholder population $\mathcal{S}$. We solve this problem using a technique called "probabilistic inversion". Probabilistic inversion (PI) is similar to ordinary inversion. In ordinary inversion there are quantities x, y, and a function g that maps x to y. The quantity y is observed and the task then is to find an x^* such that $y = g(x^*)$. In the probabilistic setup, the quantities x, y are random vectors instead of numbers. There are two formulations to solve a problem of probabilistic inversion namely, the measure theoretic approach and the random variable approach [10] from which we choose the latter.

Definition 1 Let $\mathbf{X}, \mathbf{Y}$ be random vectors taking values in $\mathbb{R}^N$ and $\mathbb{R}^M$ respectively. Further let $G : \mathbb{R}^N \to \mathbb{R}^M$ be a measurable function. $\mathbf{X}$ is called a probabilistic inverse of G at $\mathbf{Y}$ if $G(\mathbf{X}) \sim \mathbf{Y}$,where $\sim$ means "*has the same distribution as*". If $\mathbf{C}$ is a set of random vectors taking values in $\mathbb{R}^M$, then $\mathbf{X}$ is an element of the probabilistic inverse of G at $\mathbf{C}$ if $G(\mathbf{X}) \in \mathbf{C}$.

There are two main algorithms to carry out PI, namely IPF(Iterative Proportional Fitting)[11],[22],[7] and PARFUM (PARameter Fitting for Uncertainty Models)[1],[5],[13]. IPF was first described by Kruithof [11] and later rediscovered by Deming and Stephan [4], and many others. Csiszar [3] proved the convergence of IPF in case of a feasible problem. He shows that if the IPF algorithm converges, then it converges to the unique distribution (called the I-projection) which is minimally informative relative of the starting distribution, within the set of distributions satisfying the marginal constraints. PARFUM was introduced and studied by Cooke [1] . If the problem is feasible, PARFUM converges to a solution which is distinct from the IPF solution. Unlike IPF, PARFUM always converges, and it converges to a solution which minimizes a suitable information functional [5]. The convergence of PARFUM (and its canonical variations) was proved by Matus[16] but has not yet been published. When the problem is feasible IPF is preferred, because of its fast convergence. PARFUM is used when the problem is infeasible, because it insures a solution such that $I\left(G\left(\mathbf{X}\right)|\mathbf{Y}\right)$ is minimal.

The idea now is to infer a distribution over the utility values given the marginal distributions Q over the responses using PI. We denote the distribution over the utility values by $\mathcal{P}$. For the diffuse initial distribution we take u_1, u_2, u_3, u_4 to be independent and uniformly distributed on $[0, 1]$. This distribution would then yield the following distribution over responses see Table 3.2. It does not comply with (3.1). Note that we could use other distributions as the initial distribution.

Table 3.2: Marginal distribution Q over the responses given uniform distribution

	a_1	a_2	a_3	a_4
D_1	1/3	1/3	1/3	N/A
D_2	N/A	1/3	1/3	1/3

With either IPF or PARFUM it is possible to find a distribution $\mathcal{P}$ over the utility values u_1, u_2, u_3, u_4 such that this distribution complies with the distribution over the responses Q. As mentioned, IPF and PARFUM are sample based algorithms for obtaining a probabilistic inverse. These algorithms re-weight a starting distribution to get a distribution satisfying the constraints (3.1). We first draw a number of samples from the starting distributions and then compute the responses to D_1, D_2 for each sampled utility vector. These constitute the values of the functions G_1, G_2. For demonstration purposes we use ten samples, but a far greater number of samples are needed in real applications. Table (3.3) shows ten samples for u_1, u_2, u_3, u_4 from a uniform distribution together with the computed outputs G_1, G_2.

Table 3.3: Ten input and output samples

Sample	u_1	u_2	u_3	u_4	G_1	G_2
1	0.6047	0.13987	0.8202	0.39849	a_3	a_3
2	0.20205	0.34152	0.47065	0.88651	a_3	a_4
3	0.11747	0.78388	0.81113	0.36028	a_3	a_3
4	0.1898	0.25268	0.04756	0.84957	a_2	a_4
5	0.86156	0.89332	0.20379	0.25159	a_2	a_2
6	0.14059	0.09522	0.30707	0.28061	a_3	a_3
7	0.29232	0.20926	0.73012	0.36256	a_3	a_3
8	0.87431	0.65964	0.42908	0.31673	a_1	a_2
9	0.36005	0.08888	0.12888	0.03023	a_1	a_3
10	0.31374	0.82145	0.00599	0.59636	a_2	a_2

The samples $\left(u_1\left(l\right), u_2\left(l\right), u_3\left(l\right), u_4\left(l\right), G_1\left(l\right), G_2\left(l\right)\right)$, on the l-th sample (l-th virtual stakeholder), are drawn from the starting joint distribution $\mathcal{P}^0$. Each sample has probability 0.1 under $\mathcal{P}^0$. The successive joint distributions obtained after m-th iterate of IPF or PARFUM will be denoted by $\mathcal{P}^m$.

Next we compute the marginal distributions for outputs of G_1 and G_2
$\mathcal{P}^0(G_1 = a_1), \mathcal{P}^0(G_1 = a_2), \mathcal{P}^0(G_1 = a_3)$,
$\mathcal{P}^0(G_2 = a_2), \mathcal{P}^0(G_2 = a_3), \mathcal{P}^0(G_2 = a_4)$
which are presented in table (3.4). Evidently these probabilities do not comply with the target probabilities in (3.1).

Table 3.4: Marginal distribution Q over the responses given $\mathcal{P}^0$

	a_1	a_2	a_3	a_4
D_1	0.2	0.3	0.5	N/A
D_2	N/A	0.3	0.5	0.2

An I-projection is an operation that adjusts a joint distribution such that it meets a given marginal constraint. The IPF procedure successively I-projects onto each margin, and repeats until convergence is reached. If the problem is feasible IPF converges to a $\mathcal{P}^*$ which satisfies all constraints and is minimally informative with respect to the starting distribution $\mathcal{P}0$. PARFUM on the other hands averages the I-projections of each margin to obtain the next iterate. The I-projection of a distribution $\mathcal{P}^0$ onto the margins of G_1 is defined as follows, where $Q_{k,i}$ is the (k,i) entry of Table (3.1)

$$I_{G_1}\left(\mathcal{P}^0\right) = I_{G_1=a_3}\left(I_{G_1=a_2}\left(I_{G_1=a_1}\left(\mathcal{P}^0\right)\right)\right) \tag{3.1}$$

with

$$I_{G_k=a_i}\left(\mathcal{P}_l^0\right) = \begin{cases} \mathcal{P}_l^0 * \frac{Q_{k,i}}{\mathcal{P}^0(G_k=a_i)} & , G_k(l) = a_i \\ \mathcal{P}_l^0 & , G_k(l) \neq a_i \end{cases} \tag{3.2}$$

The I-projection of $\mathcal{P}^0$ onto the margins of G_2 is computed in the same way as G_1, but $\mathcal{P}^0$ replaced by $I_{G_1}\left(\mathcal{P}_l^0\right)$. Table (3.5) shows how an I-projection of $\mathcal{P}^0$ onto the margins of G_1 is computed. Note that only the third column are weights that sum to 1, as (3.1) requires cycling through all values of G_i.

Table 3.5: I-projection of the margin of G_1; $I_{1,2} \circ I_{1,1}$ denotes $I_{G_1=a_2} \circ I_{G_1=a_1}\left(\mathcal{P}^0\right)$, etc.

$I_{G_1=a_1}(\mathcal{P}^0)$	$I_{1,2} \circ I_{1,1}$	$I_{1,3} \circ I_{1,2} \circ I_{1,1}$
0.1	0.1	$\frac{0.2}{0.5} * 0.1 = 0.04$
0.1	0.1	$\frac{0.2}{0.5} * 0.1 = 0.04$
0.1	0.1	$\frac{0.2}{0.5} * 0.1 = 0.04$
0.1	$\frac{0.3}{0.3} * 0.1 = 0.1$	0.1
0.1	$\frac{0.3}{0.3} * 0.1 = 0.1$	0.1
0.1	0.1	$\frac{0.2}{0.5} * 0.1 = 0.04$
0.1	0.1	$\frac{0.2}{0.5} * 0.1 = 0.04$
$\frac{0.5}{0.2} * 0.1 = 0.25$	0.25	0.25
$\frac{0.5}{0.2} * 0.1 = 0.25$	0.25	0.25
0.1	$\frac{0.3}{0.3} * 0.1 = 0.1$	0.1

The I-projections onto G_1, G_2 are respectively equal to $(0.04, 0.04, 0.04, 0.1, 0.1, 0.04, 0.04, 0.25, 0.25, 0.1)$ and $(0.0488, 0.0714, 0.0488, 0.1786, 0.0556, 0.0488, 0.0488, 0.1389, 0.3049, 0.0556)$. The next joint distribution $\mathcal{P}^1$ is equal to last I-projection

$$\mathcal{P}^1 = (0.0488, 0.0714, 0.0488, 0.1786, 0.0556, 0.0488, 0.0488, 0.1389, 0.3049, 0.0556) \tag{3.3}$$

Table 3.6: Marginal distributions G_1, G_2

Step m	$G_1 = a_1$	$G_1 = a_2$	$G_1 = a_3$	$G_2 = a_2$	$G_2 = a_3$	$G_2 = a_4$
1	0.4437669	0.2896825	0.2665505	0.25	0.5	0.25
5	0.4998888	0.3001261	0.1999851	0.25	0.5	0.25
10	0.4999959	0.3000048	0.1999993	0.25	0.5	0.25
15	0.4999998	0.3000002	0.2	0.25	0.5	0.25
17	0.5	0.3	0.2	0.25	0.5	0.25

The marginal distributions for G_1, G_2 given several IPF iteration steps are presented in Table 3.6. After 17 iterations IPF has converged. Let $\mathcal{P}^{IPF}$ be the solution of the IPF procedure. $\mathcal{P}^{IPF}$ is then equal to

$$\mathcal{P}^{IPF} = (0.03625,\ 0.055,\ 0.03625,\ 0.195,\ 0.0525,\ 0.03625,\ 0.03625,\ 0.145,\ 0.355,\ 0.0525) \tag{3.4}$$

For PARFUM, the next iterate is computed by first I-projecting onto the margins of G_1, G_2 and then averaging the result. The next iterate $\mathcal{P}^1$ is equal to

$$\mathcal{P}^1 = (0.07, 0.0825, 0.07, 0.1125, 0.0917, 0.07, 0.07, 0.1667, 0.175, 0.0917) \tag{3.5}$$

The marginal distributions for G_1, G_2 given several PARFUM iteration steps are presented in Table 3.7.

Table 3.7: Marginal distributions G_1, G_2

Step m	$G_1 = a_1$	$G_1 = a_2$	$G_1 = a_3$	$G_2 = a_2$	$G_2 = a_3$	$G_2 = a_4$
1	0.3416667	0.2958333	0.3625	0.35	0.455	0.195
5	0.4744914	0.2919279	0.2335807	0.2847194	0.4803188	0.2349619
25	0.5000717	0.2999113	0.200017	0.2500172	0.4999107	0.2500721
50	0.5000018	0.2999979	0.2000003	0.2500003	0.499998	0.2500017
75	0.5	0.3	0.2	0.25	0.5	0.25

Finally after 75 iterations PARFUM converges. The PARFUM solution is then equal to

$$\mathcal{P}^{PARFUM} = (0.0363,\ 0.0547,\ 0.0363,\ 0.1953,\ 0.0524,\ 0.0363,\ 0.0363,\ 0.1453,\ 0.3547,\ 0.0524) \tag{3.6}$$

As expected IPF converges faster to a solution than PARFUM. Note that the sample weights obtained from IPF (3.4) are slightly different those obtained using PARFUM (3.6). The utilities of the choice alternatives under IPF and PARFUM are also close (see Tables 3.8,3.9). The resulting joint distribution over the choice alternatives induces correlations, as shown in Table 3.10 for IPF.[4] Thus a stakeholder who values a_2 highly is also likely to value a_1 highly, but little can be said about his/her valuation of a_4. More information on iterative methods for probabilistic inversion may be found in [13]

Table 3.8: Means and Standard Deviations of Utility Values using IPF

	a_1	a_2	a_3	a_4
Mean	0.40629	0.32981	0.25088	0.36642
Standard Deviation	0.25053	0.27967	0.23545	0.32350

[4] The correlation matrix 3.10 is not semi-positive definite due to rounding and the small number of samples used.

Table 3.9: Means and Standard Deviations of Utility Values using PARFUM

	a_1	a_2	a_3	a_4
Mean	0.4063	0.3298	0.2510	0.3665
Standard Deviation	0.2506	0.2796	0.2356	0.3234

Table 3.10: Correlation coefficients of the Utility Values using IPF

	a_1	a_2	a_3	a_4
a_1	1.00	0.78	0.24	0.23
a_2	0.78	1.00	0.69	0.16
a_3	0.24	0.69	1.00	0.60
a_4	0.23	0.16	0.60	1.00

3.1 Model Validation with IPF and PARFUM

In many applications we do not simply want utility values for alternatives, we want to model the utility values as functions of underlying physical variables. The paradigm case is Multi Attribute Utility Theory (MAUT) where utility is expressed as a weighted combination of physical attributes, such as price, weight, reliability, maintainability, etc. The reservations regarding standard utility modeling expressed in the introduction can be dispelled to some extent within the probabilistic inversion approach.

First, we may assume that the utility of stakeholder s for alternative a_i, $u_s(a_i)$ can be expressed as some function

$$u_s\left(a_i\right) = \Phi\left(c_i, \omega_s\right) \tag{3.7}$$

where c_i is a vector of 'criteria scores' which depend on i but not on s, and ω_s is a vector of parameters which depend on s but not on i. The population of stakeholders $\mathcal{S}$ would be described by a distribution over ω_s. The most familiar form is the standard $MAUT$ expression:

$$u_s(a_i) = \sum_{j=1}^{M} \omega_{s,j} \times c_{i,j}. \tag{3.8}$$

$MAUT$ assumes that the ω are normalized weights. The solution algorithms using IPF and PARFUM procede exactly as before: we begin with a diffuse starting distribution over ω from which a large number of samples are drawn. Using (3.8) we compute the joint distribution of utility values associated with the starting distribution. IPF and/or PARFUM are applied to re-weight the starting distribution to comply with the discrete choice data. We hasten to add that a wide variety of functional forms would be tractable. For example, we might add quadratic and interaction terms of arbitrary order to (3.8) without compromising solvability. Constraints on the parameters ω, (non-negativity, normalization) can also be imposed. Sampling the resulting joint distribution of ω we obtain the joint distribution of utilities for all alternatives, characteristic for $\mathcal{S}$. Note that no assumption regarding the dependence of utility values across $\mathcal{S}$ is imposed; rather, dependencies emerge from the fitting algorithms themselves. Of course the simple linear form (3.8) has distinct advantages; because of the linearity of expectation, the expected utility of an alternative can be simply computed by plugging in the expected values of ω.

To exploit the modeling freedom afforded by probabilistic inversion, and to conform with sound science, it is essential to evaluate and validate model forms. As mentioned in the introduction, the near absence of validation has rendered the field of utility theory more confessional than scientific. To effectuate validation, we use part of the discrete choice data to derive the distributions over the parameters ω, or indeed over the utilities $u_s(a_i)$. For example we used the ranks that received

more than 30% of the vote to infer a distributions over the parameters ω using IPF. Then we use these to predict the empirical distribution of responses in the remaining part of the discrete choice data. This is true out-of-sample validation, and should be recognized as an essential part of any application. The application discussed in the next section illustrates these ideas. There are many issues to be addressed with regard to utility model validation; the approach presented below is not the last word; it is intended to introduce this subject, not close it.

4 Application: Valuing Health States

As mentioned in section 2, this valuation study uses the extended version EQ-5D+C (see table A.1 in the Appendix) introduced by Stouthard[23].

Although the EuroQol descriptive system is non disease specific, health states can be associated with diseases and/or disease stages via the use of questionnaires filled by patients, proxies and/or physicians. Due to human variability, as well as variability in symptoms and stages of a disease, a particular disease could be associated with a number of health states[6]. The number of health states used is 17 The health states were the scenarios that stakeholders participating in the current study judged. The number of participating stakeholders was 19. A great majority (17/19) of stakeholders, represented a panel of "non-health care professionals" defined as people with academic background but with no health care professional experience. This did not exclude any health care personal experience, however they were asked to declare this in advance.

The scenarios where presented in five groups of five. Figure (A.1) in the appendix illustrate the questions asked to rank the scenarios per group. Scenarios overlapped among the groups; the last two in each group were repeated as the first two in the consecutive group (see Figure 2.1). This design ensured that we could test stakeholders for consistency in their results and to test if we get similar weights for the criteria in each 5 group. The 17 scenarios are non-dominated.

4.1 The Model

We used two models in the analysis. The first is the so called *unmodelled scores* where $u_s(a_i) = u_i$ as in the IPF and PARFUM example of section 3, taking values between zero and one. The second is the linear model formulated by equation 3.8. An overview of the health criteria is illustrated by Figure (A.2) in the appendix. The criteria took values 1, 2, 3 and we assumed that a higher values are always worse. As a result, the health scores will take values between -1 and -3.

The *unmodelled scores* can be seen as a composition of both observed and unobserved criteria and will be used as a benchmark. If we can fit the unmodelled scores, but fail to fit linear model then we know the lack of fit is due to the linear model.

The number of health states used is 17 so $\mathcal{A} = \{a_1, \ldots, a_{17}\}$. The five groups of five lead to the following the discrete choice problem.

$$\begin{aligned}
\mathcal{D} &= \{D_1, \ldots, D_5\} \\
D_1 &= \{a_1, a_2, a_3, a_4, a_5\} \\
D_2 &= \{a_4, a_5, a_6, a_7, a_8\} \\
D_3 &= \{a_7, a_8, a_9, a_{10}, a_{11}\} \\
D_4 &= \{a_{10}, a_{11}, a_{12}, a_{13}, a_{14}\} \\
D_5 &= \{a_{13}, a_{14}, a_{15}, a_{16}, a_{17}\}
\end{aligned} \tag{4.1}$$

Stakeholders were asked to rank the health states within each group. The task was to find a distribution over the weights $(\omega_1, \ldots, \omega_6)$ which reproduces the distributions over the rankings.

$$\mathcal{P}\left(\omega_1, \ldots, \omega_6 \,|\, u(a_{k_i}) \text{ is j-th ranked in } u(D_k)\right) = \frac{\#\{s \in \mathcal{S} \,|\, s \text{ ranks } a_{k_i} \text{ in j-th position in } D_k\}}{\#\mathcal{S}} \tag{4.2}$$

If an stakeholder ranks health state a_i above a_j in group k then we consider an stakeholder to be inconsistent if he ranks a_j above a_i in group $k+1$.

4.2 Results and Validation

4.2.1 Model Adequacy

Of the 19 stakeholders who participated in the study, 13 consistently ranked all of the health states. Using Kendall's coefficient of concordance W[9],[8] we also looked if the stakeholders rankings are random in each group. If W is one, then each stakeholder has assigned the same order to the choice alternatives. If W is zero, then there is no overall agreement among the stakeholders, and their responses may be regarded as essentially random. Intermediate values of W indicate a greater or lesser degree of agreement among the various responses.

Let $R(a_i, s)$ be the rank given to health state a_i by stakeholder s. Then the sum of ranks given to a_i is

$$R(a_i) = \sum_e R(a_i, e) \tag{4.3}$$

The mean value of these sum ranks is equal to

$$\bar{R} = \frac{1}{2} m(n+1) \tag{4.4}$$

with m the number of stakeholders and n the number of health states. The sum of squared deviations is defined as

$$S = \sum_{i=1}^{n} \left(R(a_i) - \bar{R}\right)^2 \tag{4.5}$$

The coefficient of concordance is then equal to

$$W = \frac{12S}{m^2(n^3 - n)} \tag{4.6}$$

The null hypothesis that stakeholders choose ranks at random can be tested in terms of the values for S given n and m. Friedman[8] derived a Table which contains the critical values $*$ of S at 5% significance level, for n between 3 and 7 and m between 3 and 20. For each group we computed the values of S and W, shown in Table (4.1).

Table 4.1: Values of S and W for each group

	Group 1	Group 2	Group 3	Group 4	Group 5
S	1166	1854	1744	1282	1597
W	0.323	0.514	0.483	0.355	0.443

Friedman's Table doesn't contain the critical values for S given $m = 19$; we therefore used the values for $m = 20$. The null hypothesis would be rejected at the 5% level for $m = 20$ if $S > 468.5$. With this criterion the null hypothesis is rejected for all groups

We fitted both the rank data of all stakeholders and the rank data of the consistent stakeholders to the MCDM model (3.8). Each point is an alternative-rank in each of the sets D_k. Each alternative could possibly be ranked in any of the five positions; yielding $5 \times 5 \times 5$ points; however the actual number of points is smaller as the stakeholders confined some alternatives to a smaller number of ranks, and zero probabilities are not plotted. The observed probabilities of rankings are on the horizontal axis, and the probabilities recovered by the linear model of the utilities are on the vertical axis. We used linear regression as goodness of fit measure for our method. The slope tells us how accurate the predictions are on average. And the coefficient of determination tells us how much of the variation of the observed frequencies is explained by the predicted frequencies. Both fits have a accuracy of around 95% and a high coefficient of determination ($R^2 = 0.896$), ($R^2 = 0.964$) which suggests that the preferences of the stakeholders are consistent with the linear model (3.8). We continue the analysis using the rankings of the consistent stakeholders.

4.2.2 Criteria Weights

We fitted the model to each health state group separately to see how it affects the weights for the criteria. Figures (4.1), (4.2), (4.3) illustrate the low, mean and high values for the weights. The low and high values are computed by respectively subtracting and adding one standard deviation to the means of the weights.

Criterion *Pain Discomfort* seems to be the most important factor followed by *Cognitive Functioning*, which values seems to be stable for all groups. However if we look at the weights from fitting the rank data in each group we notice that criterion *Pain Discomfort* no longer is the most important factor, but *Self Care*. We could also read the most important criteria from the cumulative distribution over the weights see Figure (4.4). The cumulative distribution over the weights is computed from the distribution over the weights obtained from the distribution over the rankings. The most important criterion is represented by the right most distribution.

Fig. 4.1: Low Values Weights Obtained from Fitting Experts' Rankings

Group	Mobility	Self Care	Usual Activities	Pain Discomfort	Anxiety Depression	Cognitive Functioning
1	3.62%	7.22%	3.59%	6.82%	1.80%	13.67%
2	3.26%	2.93%	0.28%	22.20%	10.28%	6.59%
3	6.49%	7.56%	6.07%	9.99%	1.65%	7.77%
4	6.23%	5.69%	2.37%	13.75%	0.24%	12.19%
5	1.60%	5.33%	2.77%	11.24%	3.37%	11.61%

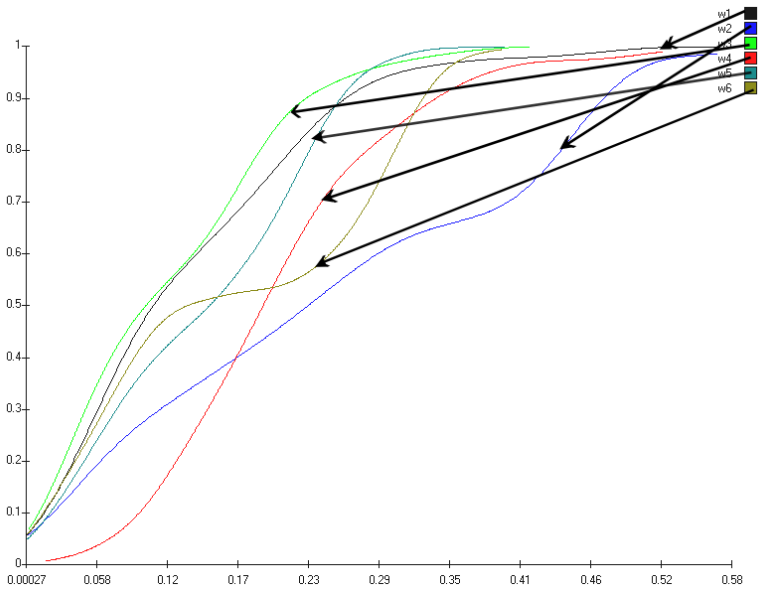
Fig. 4.2: Mean Values Weights Obtained from Fitting Experts' Rankings

Group	Mobility	Self Care	Usual Activities	Pain Discomfort	Anxiety Depression	Cognitive Functioning
1	13.69%	18.08%	10.78%	22.53%	10.63%	24.79%
2	10.91%	11.11%	7.25%	33.10%	19.10%	18.54%
3	15.57%	17.80%	15.94%	20.08%	9.89%	20.72%
4	16.71%	15.51%	11.54%	23.96%	9.50%	22.78%
5	10.24%	14.44%	10.82%	28.27%	13.02%	23.94%

Fig. 4.3: High Values Weights Obtained from Fitting Experts' Rankings

Group	Mobility	Self Care	Usual Activities	Pain Discomfort	Anxiety Depression	Cognitive Functioning
1	22.76%	28.95%	17.97%	38.24%	19.45%	35.92%
2	18.55%	19.29%	14.22%	44.00%	27.91%	30.49%
3	24.56%	28.05%	25.80%	30.17%	18.13%	33.67%
4	27.18%	25.34%	20.72%	34.17%	18.76%	33.37%
5	19.24%	23.54%	18.86%	45.30%	22.66%	34.47%

Fig. 4.4: Cumulative Distribution of the Weights



We say that the stakeholder weights have interactions if, for example, knowledge that a stakeholder assigns high weight to the *Self Care* criterion gives significant information regarding weights for other criteria. Detailed analysis of interactions is not undertaken, but the correlation matrix presented in Table (4.2) suggests that the *mobility - anxiety depression; self care - cognitive functioning; self care - usual activities; pain discomfort - anxiety depression* interactions are rather strong. The requirement that the weights sum to one imposes an overall negative correlation.

Table 4.2: Correlation coefficients of the weights.

	Mobility	Self Care	Usual Act.	Pain Disc.	Anxiety Depr.	Cognitive Func.
Mobility	1	-0.2812	0.1018	0.1044	-0.4669	-0.1695
Self Care	-0.2812	1	-0.5247	-0.3103	0.1577	-0.5574
Usual Act.	0.1018	-0.5247	1	-0.0104	-0.0146	-0.0362
Pain Disc.	0.1044	-0.3103	-0.0104	1	-0.4643	-0.1107
Anxiety Depr.	-0.4669	0.1577	-0.0146	-0.4643	1	-0.1880
Cognitive Func.	-0.1695	-0.5574	-0.0362	-0.1107	-0.1880	1

4.2.3 Health State Scores

Fitting the MCDM model not only gives us statistics about the disability weights, but also statistics about the health scores. In the introduction we mentioned that utility is affine unique, which we

have used to transform the health scores to the zero one interval. Initially the health scores given the MCDM model vary between -3 and -1, which have no tangible meaning. However if we standardize these scores we can think of them as the relative impact on health see Figure (A.3).

4.2.4 Unmodelled Scores

Fitting the unmodelled scores yields a near perfect fit ($R^2 = 0.999$). This means that we can assign a distribution over the utilities for the 17 health states, such that the probabilities of observing each health state at each rank in each D_k are predicted nearly perfectly. This merely says that our population of stakeholders can be modeled as rational in the sense of Savage.

The *unmodelled scores* presented by Figure (A.4) are assumed to take values between -1 and 0. The variance of the *unmodelled scores* seems to be higher than the variance of the health scores from the linear model (3.8). Note that the ranks of the health states have slightly changed with respect to the scores from the linear model.

4.2.5 Validation

Finally we validated the results by fitting the model to the ranks that got more than 30% of the votes from the stakeholders, and then predicted the remaining ranks using the fitted model. At first glance, from Figure (4.5) the prediction doesn't look good at all; but if we zoom in on the ranks that got less than 30% and take the average of these predicted ranks we get a fit that is not so bad, see Figure (4.6). However the *unmodelled scores* seems to give better prediction both overall and on the unpopular ranks, see Figures (4.7) and (4.8).

The out-of-sample validation is not overwhelming; the distribution over weights derived from the "popular rankings" (i.e. those rankings assigned by more than 30% of the stakeholders, did not accurately predict the probabilities for the "unpopular rankings" (i.e. rankings assigned by less than 30% of the stakeholders). Nonetheless, the average predictions do align with the observed predictions. Without undertaking an in-depth analysis, it seems that the unpopular ranks concerned primarily the lowest ranked health states, and one possible explanation for these results is that the preference judgments for the 'less important' health states were less discerning.

Fig. 4.5: Prediction of Rank Percentages Using Rank Percentages Greater Than 30%

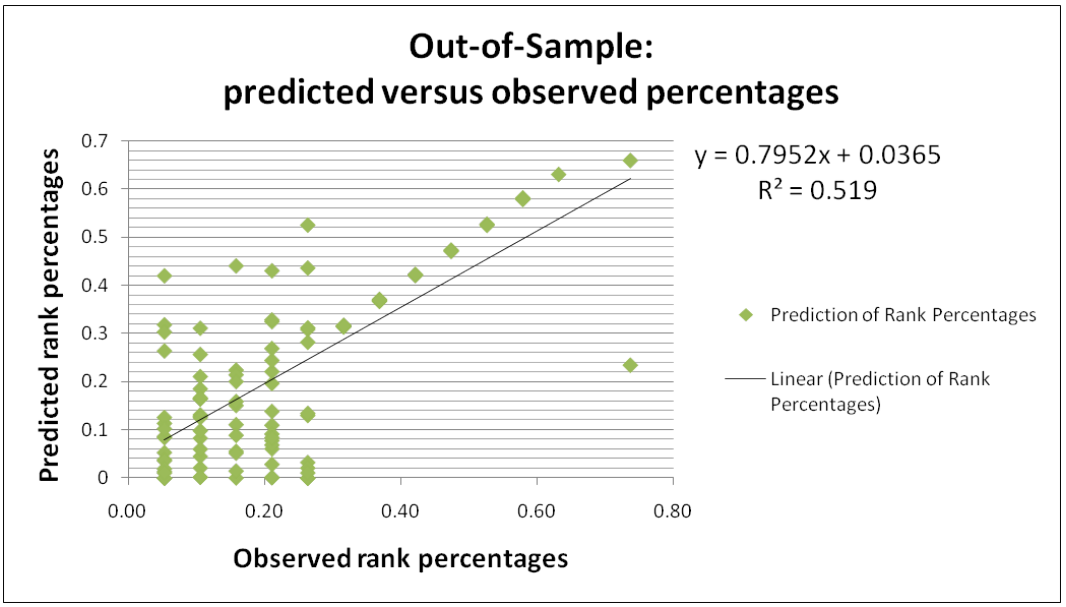


Fig. 4.6: Average Prediction of Rank Percentages Less Than 30%

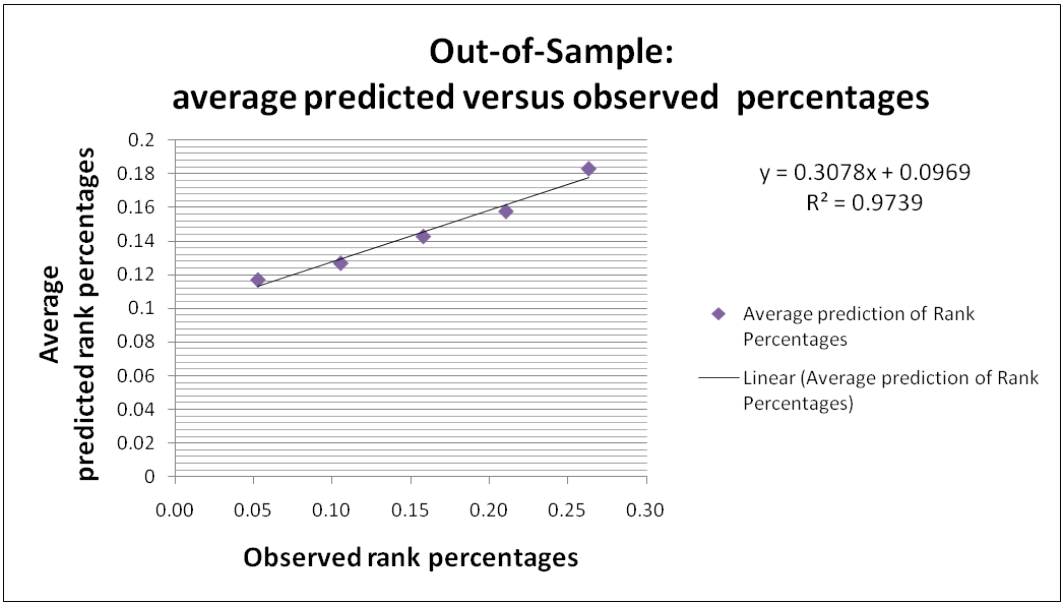


Fig. 4.7: Prediction of Rank Percentages Using Rank Percentages Greater Than 30%

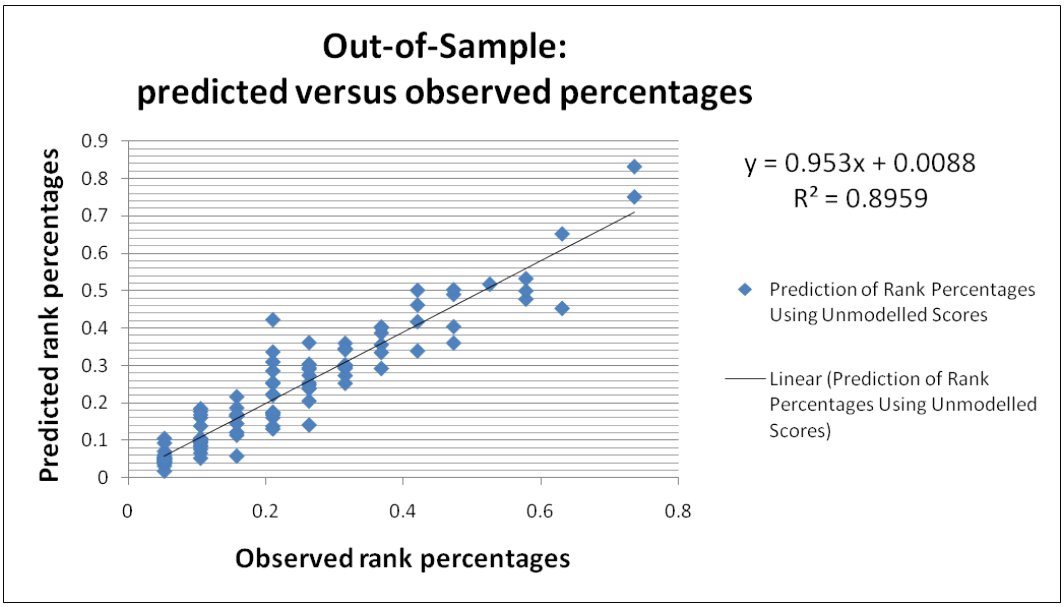
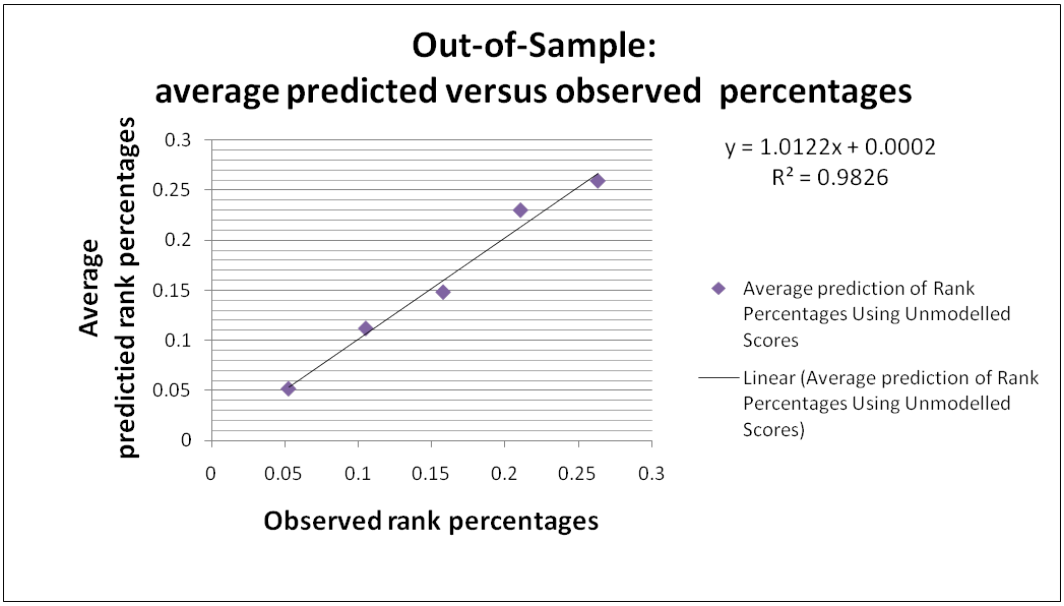


Fig. 4.8: Average Prediction of Rank Percentages Less Than 30%



5 Conclusion

Probabilistic inversion methods can be used to infer a distribution over utility functions based on discrete choice data. This type of application is rather new and more experience in real applications is needed. In particular, questions regarding the optimal format of the discrete choice data, the best approach to out-of-sample validation are still largely open. Equally important is how best to model utilities in terms of physical attributes. It is hoped that out-of-sample validation will eventually yield utility models with a solid scientific foundation.

References

1. Roger M. Cooke. Parameter fitting for uncertain models: modeling uncertainty in small models. *Reliability engineering & systems safety*, 44(1):89–102, 1994.
2. Roger M. Cooke. Obtaining distributions from groups for decisions under uncertainty. In Knut Samset Terry M. Williams and Kjell J. Sunnevag, editors, *Making essential choices with scant information: frond-end decisions making in major projects*, pages 257–274. Palgrave Macmillan, 2009.
3. Imre Csiszar. divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 2 1975.
4. Edwards W. Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11(4):427–444, 1940.
5. C. DU, D. Kurowicka, and R.M. Cooke. Techniques for generic probabilistic inversion. *Computational Statistics and Data Analysis*, 50(5):1164–1187, 2006.
6. J.J. Ellis, K.A. Eagle, E.M. Kline-Rogers, and S.R. Erickson. Validation of the eq-5d in patients with a history of acute coronary syndrome. *Current medical research and opinion*, 21(8):1209–16, 8 2005.
7. Stephen E. Fienberg. An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907–917, 1970.
8. Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.
9. M. G. Kendall and B. Babington Smith. The problem of m rankings. *The Annals of Mathematical Statistics*, 10(3):275–287, 1939.
10. B.C.P Kraan and Bedford. T.J. Probabilistic inversion of expert judgements in the quantification of model uncertainty. *Management Science*, 51(6):995–1006, 2005.
11. J. Kruithof. Telefoonverkeersrekening. *De Ingenieur*, 52(8):15–25, 1937.
12. D. Kurowicka, C. Bucura, A Havelaar, and R.M. Cooke. Probabilistic inversion in priority setting of emerging zoonoses. *Risk Analysis*, 2010.

13. Dorota Kurowicka and Roger M. Cooke. *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley, 2006.
14. R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.
15. Jacob Marschak. Binary choice constraints on random utility indicators. Cowles Foundation Discussion Papers 74, Cowles Foundation, Yale University, 1959.
16. F. Matus. On iterated averages of i-projections. bielefeld: Statistiek und informatik, university. Not published, 2007.
17. Daniel McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, pages 198–272, 1974.
18. Daniel McFadden and Kenneth Train. Mixed mnl models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, 5 200.
19. Rabin E.J. Neslo, F. Micheli, C. V. Kappel, K. A. Selkoe, Benjamin S. Halpern, and Roger M. Cooke. Modeling stakeholder preferences with probabilistic inversion application to prioritizing marine ecosystem vulnerabilities. In *Real-Time and Deliberative Decision Making*, NATO Science for Peace and Security Series C: Environmental Security, pages 265–284. Springer Netherlands, 2008.
20. M. E. Terry R. Bradley. Rank analysis of incomplete block designs: I. the method of paired comparisons., 1952.
21. L. Savage. *The Foundations of Statistics*. John Wiley & Sons., 1954.
22. F.F. Stephan. An iterative method of adjusting sample frequency tables when expected marginal totals are known. *Ann. Math. Stat.*, 13:166–178, 1942.
23. Marlies E.A. Stouthard, Essink-Bot Marie Louise., and Bonsel Gouke J. Disability weights for diseases. a modified protocol and results for a western european region. *European Journal of Public Health*, 10(1):24–30, 2000.
24. Sarah Teck, Benjamin Halpern, Carrie Kappel, Fiorenza Micheli, Kimberly Selkoe, Caitlin Crain, Rebecca Martone, Christine Shearer, Joe Arvai, Baruch Fischhoff, Grant Murray, Rabin Neslo, and Roger Cooke. Using expert judgment to estimate marine ecosystem vulnerability in the california current. *Ecological Applications*, 0(0), 2010.
25. L.L. Thurstone. A law of comparative judgement. *Psychological Review*, 34:278–286, 1927.

A Appendix

A.1 Figures and Tables

Fig. A.1: Questions group 1

GROUP I

Health States	Description: Mobility	Self Care	Usual Activities	Pain Discomfort	Anxiety Depression	Cognitive Functioning	Your Ranking
HS "R"	3	2	2	1	2	1	
HS "T"	2	1	2	2	3	1	
HS "A"	2	1	3	2	2	1	
HS "O"	1	1	1	2	3	2	
HS "C"	2	2	3	1	3	2	

Please indicate the following:

When you rank a Health State as 1 it means for you that it is:

The most severe Health State of the 5	☐
The less severe Health State of the 5	☐

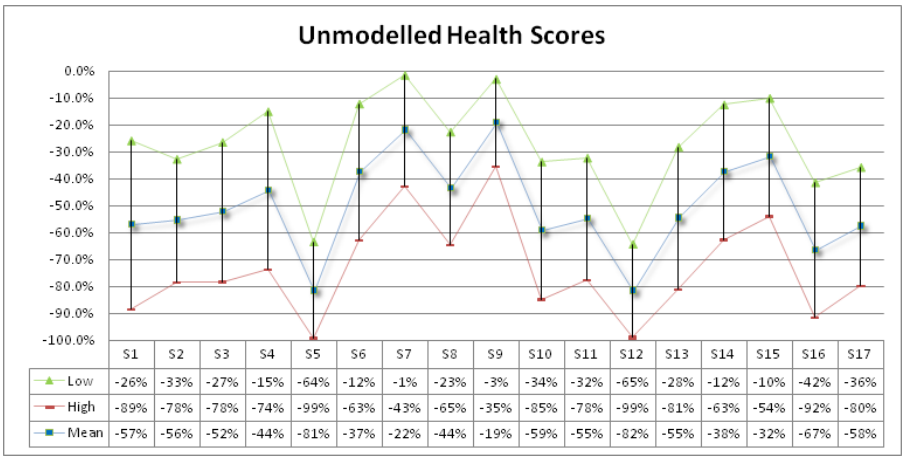
Table A.1: The extended, six dimensional version of the original EuroQol descriptive system, i.e. EQ-5D+C

Value	Mobility	Self Care	Usual Activities	Pain Discomfort	Anxiety Depression	Cognitive Functioning
1	No problems in walking about	No problems with self care	No problems with performing usual activities (e.g. work, study, housework)	No pain or discomfort	Not anxious or depressed	No problems in cognitive functioning
2	Some problems in walking about	Some problems washing or dressing self	Some problems with performing usual activities	Moderate pain or discomfort	Moderately anxious or depressed	Moderate problems in cognitive functioning
3	Confined to bed	Unable to wash or dress self	Unable to perform usual activities	Extreme pain or discomfort	Extremely anxious or depressed	Severe problems in cognitive functioning

Fig. A.2: Criteria Values Per Health State

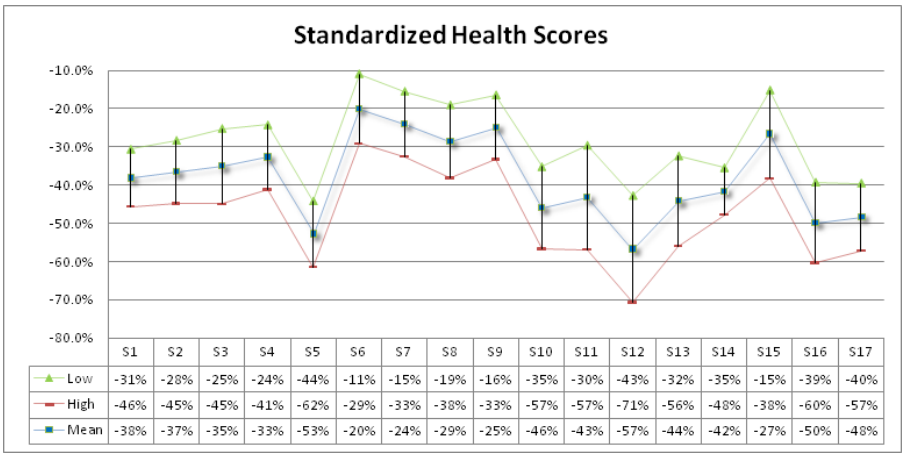
Health States	Description					
	Mobility	*Self Care*	*Usual Activities*	*Pain Discomfort*	*Anxiety Depression*	*Cognitive Functioning*
HS1	3	2	2	1	2	1
HS2	2	1	2	2	3	1
HS3	2	1	3	2	2	1
HS4	1	1	1	2	3	2
HS5	2	2	3	1	3	2
HS6	1	1	1	3	1	2
HS7	1	1	2	2	1	2
HS8	3	1	1	1	2	2
HS9	2	1	1	2	1	2
HS10	2	2	1	3	2	1
HS11	3	3	2	1	1	1
HS12	2	3	3	1	3	1
HS13	1	3	3	1	1	2
HS14	2	2	2	2	2	1
HS15	1	1	1	2	1	3
HS16	2	2	1	1	3	3
HS17	2	2	2	1	2	3

Fig. A.4: Unmodelled Health Scores



	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17
Low	-26%	-33%	-27%	-15%	-64%	-12%	-1%	-23%	-3%	-34%	-32%	-65%	-28%	-12%	-10%	-42%	-36%
High	-89%	-78%	-78%	-74%	-99%	-63%	-43%	-65%	-35%	-85%	-78%	-99%	-81%	-63%	-54%	-92%	-80%
Mean	-57%	-56%	-52%	-44%	-81%	-37%	-22%	-44%	-19%	-59%	-55%	-82%	-55%	-38%	-32%	-67%	-58%

Fig. A.3: Standardized Health Scores Obtained From Fitting The Model To The Discrete Choice Data



	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17
Low	-31%	-28%	-25%	-24%	-44%	-11%	-15%	-19%	-16%	-35%	-30%	-43%	-32%	-35%	-15%	-39%	-40%
High	-46%	-45%	-45%	-41%	-62%	-29%	-33%	-38%	-33%	-57%	-57%	-71%	-56%	-48%	-38%	-60%	-57%
Mean	-38%	-37%	-35%	-33%	-53%	-20%	-24%	-29%	-25%	-46%	-43%	-57%	-44%	-42%	-27%	-50%	-48%

A.2 Definitions

Discrete choice or random utility models describe and analyze the preferences of a group of stakeholders $\mathcal{S}$ for set of choice alternatives $\mathcal{A} = \{a_1, \ldots, a_N\}$. The preferences of each stakeholder are denoted in the customary way as $a_i \succeq a_j$. If necessary to distinguish $s \in (S)$, we will write $\succeq_s$.

Savage's [21] theory of rational decision ensures that the preferences of a single rational stakeholder can be expressed in terms of expected utility. Each $s \in \mathcal{S}$ may be assigned a utility function over choice alternatives that is unique up to a positive affine transformation, i.e. unique up to choice of zero and unit. If $u : \mathcal{A} \to \mathbb{R}$ is a utility function for a given stakeholder, then $cu + d, c > 0, d \in \mathbb{R}$, is also a utility function for this stakeholder. We will assume that our set of stakeholders have utility functions which can be assigned the same unit. This means that there are two consequences, say g and b, not necessarily belonging to the choice set $\mathcal{A}$, such that all stakeholders agree that g is strictly preferred to b. Each stakeholder would then choose respectively g and b as the unit and zero of his utility scale. We call such a set of stakeholders *orientable*. For an orientable set of rational stakeholders, the utilities over $\mathcal{A}$ may be represented as standardized $\mathcal{A}$-vectors taking values in $[0, 1]$, that is, as elements of $[0, 1]^{\mathcal{A}}$.

Definition A.2 $\mathcal{D}$ is called a discrete choice problem on $\mathcal{A}$ if

1. $\mathcal{A} = a_1, \ldots, a_N$ is a finite non-empty set of N choice alternatives
2. $\mathcal{D} = \{D_1, \ldots, D_K | D_i \subseteq \mathcal{A}, D_i \neq \emptyset, i = 1, \ldots, K\}$

A familiar type of a discrete choice problem is paired comparisons. Choice alternatives are presented in $\binom{N}{2}$ pairs from which stakeholders pick their preferred alternative.

The response of a stakeholder to a discrete choice problem may take many forms. For example, (s)he may choose a unique preferred alternative from each set $D_k \in \mathcal{D}$ (strict choice) or a set of non-dominated alternatives in $D_k \in \mathcal{D}$ (non-dominated choice) , or (s)he may order the elements of $D_k \in \mathcal{D}$, such that the response is a permutation $\pi \in D_k!$ (strict preference order), or (s)he may produce an ordered partition of elements D_k where the alternatives in each element of the partition are equivalent (non-dominated preference order). These are captured in

Definition A.3
- A *strict choice response* $r = (r_1, \ldots r_K)$ to discrete choice problem $\mathcal{D}$ is a set of mappings $r_k : \mathcal{S} \to D_k, k = 1 \ldots K$.
- A *non-dominated choice response* $r = (r_1, \ldots r_K)$ to discrete choice problem $\mathcal{D}$ is a set of mappings $r_k : \mathcal{S} \to 2_+^{D_k}$, with $2_+^{D_k} = 2^{D_k} \backslash \emptyset, k = 1 \ldots K$.
- A *strict preference order response* $r = (r_1, \ldots r_K)$ to discrete choice problem $\mathcal{D}$ is a set of mappings $r_k : \mathcal{S} \to D_k!$ where $D_k!$ is the set of permutations of $D_k, k = 1 \ldots K$.
- A *non-dominated preference order response* $r = (r_1, \ldots r_K)$ to discrete choice problem $\mathcal{D}$ is a set of mappings $r_k : \mathcal{S} \to \Pi_k$ where Π_k is the set of ordered partitions of $D_k, k = 1 \ldots K$.

The set of responses to $\mathcal{D}$ for all $s \in \mathcal{S}$ will be denoted by $r_{\mathcal{D}}$. Many other response forms are conceivable, but the above are the most straightforward. Note the difference between strict and non dominated choice. In the standard versions of random utility theory, when a stakeholder chooses element a_i from a set $\{a_1, \ldots a_n\}$, this is interpreted to mean that a_i is at least as good as the other elements. On this choice data alone we

cannot distinguish strict preference from equivalence in preference. While this might not be severe in modeling the preferences of one individual, for populations of stakeholders, such ambiguity can cause problems. Thus if 50% of stakeholders preferred a red bus to a blue bus, and 50% preferred the blue to red bus, this might either mean that everyone had strict preferences evenly divided over the population, or alternatively it might mean that everyone in the population was indifferent to the busses's color, and was choosing one color at random. Failure to distinguish these cases can cause problems in modeling the preferences of the population. These issues are important and have dominated a good deal of the discrete choice literature. Nonetheless, they are not the point of the present study. The tools that we develop for deriving a distribution over utility functions, given a distribution of responses to a discrete choice problem, apply equally well for strict and non-strict preference. However, allowing for equivalence in preference considerably complicates the notation, as can be inferred from Definition A.3. We therefore restrict attention in this study to strict preference.