UNIBALANCE

Roger M Cooke
ShortCourse on Expert Judgment
National Aerospace Institute
April 15,16, 2008

Introduction

Unibalance is a stand alone program that models preferences / values / utilities of a group of experts or stakeholders. It takes paired comparison input from E experts / stakeholders and calculates scale values for the N objects compared according to three models:

1) Bradley Terry /Negative Exponential Life (NEL) model
2) Thurstone models B and C
3) Probabilistic Inversion Models based in Iterative Proportional Fitting (IPF) and PARFUM

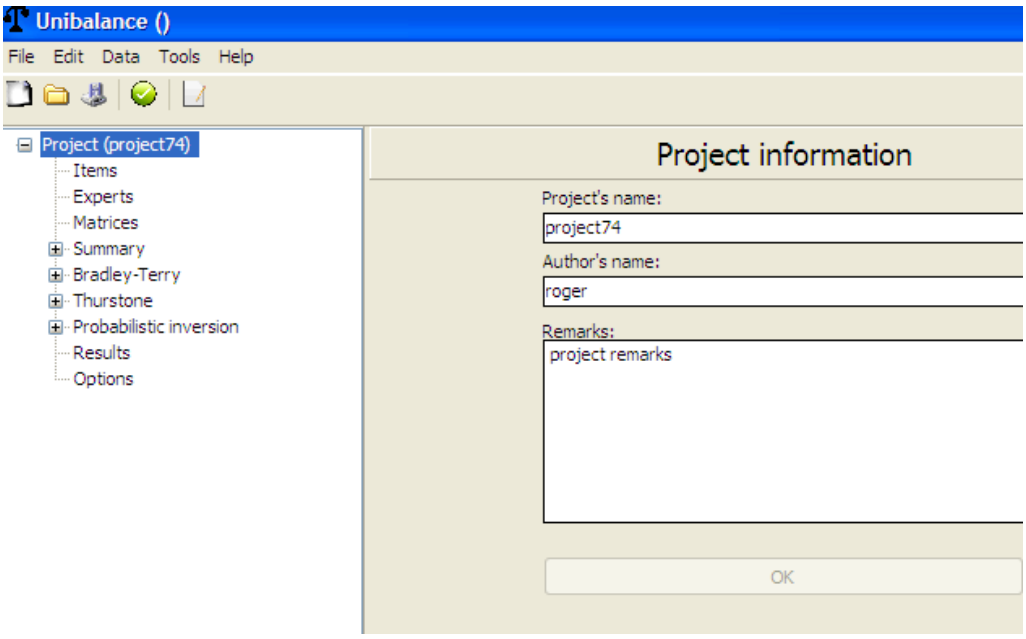
Although UNIBALANCE stands alone, it can link to EXCEL, to the graphics package UNIGRAPH and the sensitivity analysis program UNISENS.

Note:
It is recommended that each expert/stakeholder express strict preference between each pair of items. Equality of preference, and absence of preference may be used and processed; however, coefficients of consistence, agreement, and concordance will not work, nor will statistical tests based on these coefficients. The model processing will work.

Data description, getting started

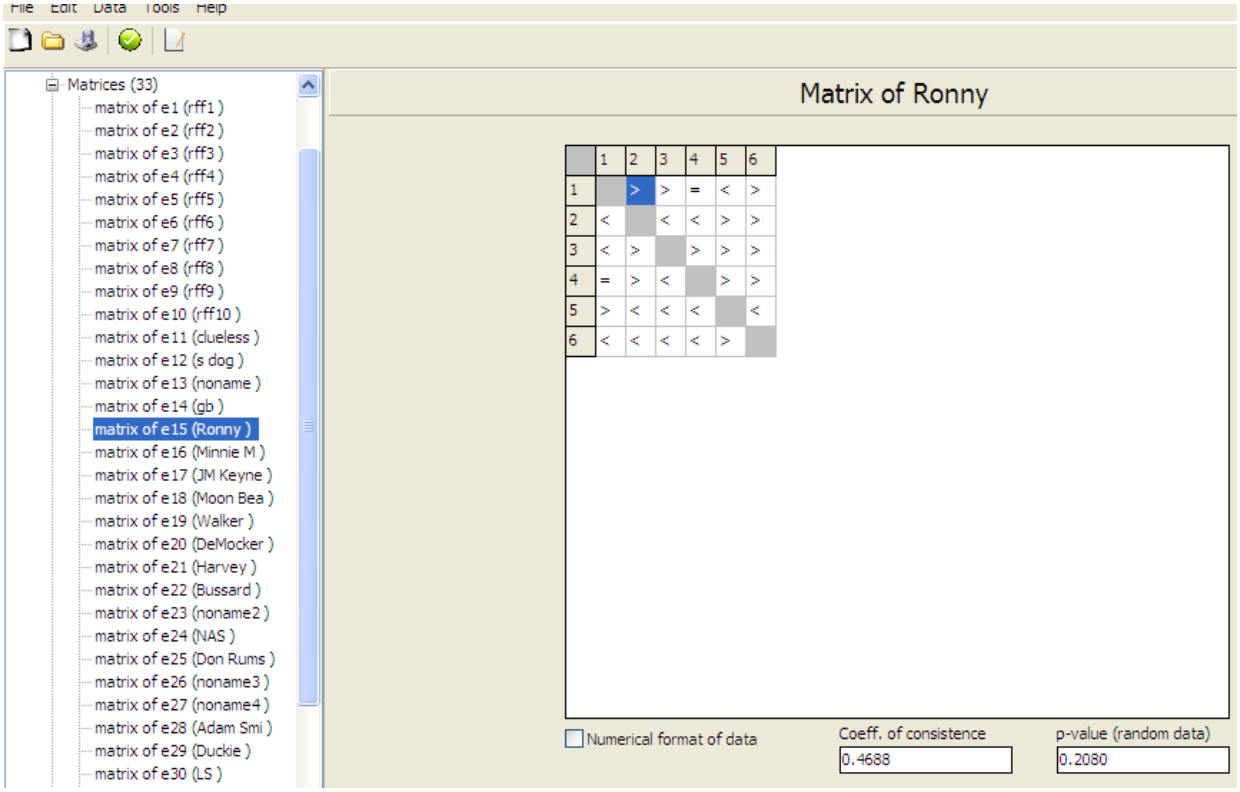
Each of E experts compare N objects pairwise, indicating which of each pair they prefer. The features discussed in this section are independent of model type, and reflect data entry and preliminary analysis. UNIBALANCE files have the extension *.XML. The demo file *Policies.xml* supplied with the program is used to illustrate features.

On start up, the basic menu structure is shown on the left panel:



1) Assign the project a name and enter an appropriate description.
2) Enter the item names by Highlighting Items. If true values for some items are know, they may be entered as well. These can be used to calibrate the Bradley Terry and Thurstone models. When names are entered, a "+" appears by the Item menu, indicating that this can be expanded.
3) Enter the expert names and any description to be used in the report, by highlighting Experts. A "+" appears by the Expert and Matrices fields, indicating that these can be expanded.
4) Expand the Matrices field. Double click on an expert's matrix. For each cell (row,col), enter
    a. > if the row item is strictly preferred to the column item
    b. < if the column item is strictly preferred to the row item
    c. = if the preference between row and column items is equal
    d. (empty) if the preference is not filled in.

"=" and " " should be avoided if possible. The screenshot below shows the matrix of expert 15, Ronny:



The p-value at which we would reject the hypothesis that Ronny filled in his preferences at random, is 0.2080, too high for rejection. This test is based on the number of circular triads (see glossary below) which one would expect if the preferences were indeed at random. The coefficient of consistence measures the degree of consistency (that is, absence of circular triads, see glossary below). The value 1 is perfectly consistent, and zero is maximally inconsistent. It must be stressed that these measures are derived under the assumption that all preferences are strict. When equal or void preferences are admitted, they should be viewed as indicative.

Brief model descriptions

When the data has been entered, the models can be analyzed.

Bradley Terry
It is assumed that each item i has a value V(i) to be estimated from the paired comparison data. The percentage of experts %(i,j) who prefer item i to item j, estimates of the conditional probability of drawing an expert from the population of experts who prefers i to j, given that (s)he must choose between i and j. It is assumed that this probability is equal to V(i)/[V(i)+V(j)]. The model is solved by maximum likelihood. Assuming that each expert for each comparison is drawn independently from the expert population, the lielihood of the data is

$$\Pi_{\text{all pairs (i,j)}} \; [V(i)/(V(i)+V(j))]^{E\times\%(ij)} \times [V(j)/(V(i)+V(j))]^{E\times\%(j,i)}.$$

The circumstance that each expert assess all pairs of items approximates the will approximately satisfy this assumption as the number of experts gets large. Values are sought which make the above expression as large as possible. Evidently, a solution V(1)...V(N) is determined only up to a positive multiplicative constant. If one of the V(i) is known, the other values are uniquely determined.

Negative exponential lifetime model

This model is computationally identical to the Bradely Terry model, though the interpretation is different. It is assumed that the item i is a component with exponential life distribution, and with failure rate λ_i, i = 1,...N. For all pairs of objects, each subject is asked to determine which component has the greater probability of failing per unit time. In answering each such question, it is assumed that each subject performs a "mental experiment", observing which of two independent exponential components with failure rates λ_i, λ_j fails first. Under these assumptions the probability p_{ij} that i fails before j is given by:

$$p_{ij} = \frac{\lambda_j}{\lambda_i + \lambda_j} \quad .$$

The percentage of experts preferring i to j is taken to estimate p_{ij}. The above system of equations is overdetermined, and the principle of maximal likelihood may be used to solve for the λ_i's. Multiplying each failure rate in the above equation by a constant would leave the equations invariant, hence the resulting λ_i's are determined only up to a constant. If one of the λ_i is known, (reference value) the others may be determined from the maximum likelihood solution.

Thurstone models

It is assumed that a subject comparing two objects i and j effectively samples two normal random variables X_i and X_j with means and standard deviations μ_i, σ_i, μ_j, σ_j. Item i is preferred to j if and only if $X_i - X_j > 0$. It is further assumed that all experts sample the same normal variables. The model is sometimes explained as follows: each item i has a "true value" μ_i; each expert perceives this value with an "error" σ_i.

The variable $X_{ij} = X_i - X_j$ is normal with mean m_{ij} and standard deviation s_{ij}:

$$\mu_{ij} = \mu_i - \mu_j$$

$$\sigma_{ij} = \sqrt{(\sigma_i^2 + \sigma_j^2 - 2\rho_{ij}\sigma_i\sigma_j)}$$

where ρ_{ij} is the correlation between X_i and X_j. The percentage %(ij) of experts preferring i to j estimates the probability that X_{ij} is positive, or equivalently, that

$$(X_{ij} - \mu_{ij})/\sigma_{ij} > -\mu_{ij}/\sigma_{ij}.$$

Letting Φ denote the cumulative standard normal distribution function, and let Φ^{-1} denote its inverse. We derive Thurstone's law of comparative judgments:

$$\Phi^{-1}(\%(ij)) \sim \mu_{ij}/\sigma_{ij} = (\mu_i - \mu_j)/\sigma_{ij}.$$

Note that the law of comparative judgments is invariant under positive affine transformations of all variables. That is, if for each i, X_i is replaced by a positive affine transform $X_i' = aX_i + b$, $a > 0$, then $\mu_{ij}'/\sigma_{ij}' = \mu_{ij}/\sigma_{ij}$. This means that paired comparison data can determine the values of the μ_i only up to a positive affine transformation, and the values of σ_i only up to a linear transformation.

Considered as a system of equations, the law of comparative judgments contains more unknowns than knowns, and hence cannot be solved. However, under various ssumptions it is possible to determine the means mi up to a positive affine transformation, i.e. up to a choice of zero and unit.

Thurstone C

In this model it is assumed that $\sigma_{ij} = \sigma$ (i.e. does not depend on i and j). For all practical purposes this equivalent to assuming that $\rho_{ij} = 0$ and $\sigma_i = \sigma/\sqrt{2}$. Filling these into Thurstone's law, and recalling that the X_i are determined only up to an affine transformation, Thurstone's law becomes simply:

$$\mu_i - \mu_j = \Phi^{-1}(\%(ij))$$

A least squares routine is used to determine the m_i up to a positive affine transformation.

Thurstone B

Thurstone's model B is computed and displayed under "Results" but confidence intervals for this model are not supported.

In this model it is assumed that for all $i \neq j$.

$$\rho_{ij} = \rho,$$
$$|\sigma_i - \sigma_j| << \sigma_i.$$

Under these conditions we can derive

$\mu_i - \mu_j = \Phi^{-1}(\%(ij)) \times \sqrt{(1-\rho)} \times (\sigma_i + \sigma_j) / \sqrt{2}$.

Since the units of the X_i are arbitrary, we can write this as

$\mu_i - \mu_j = \Phi^{-1}(\%(ij))(\sigma_i + \sigma_j)$.

which is equivalent to assuming $\rho_{ij} = -1$ in Thurstones law of comparative judgments. With N objects, there are N CHOOSE 2 such equations with 2N unknowns. For N > 5 there are more equations than unknowns. We find a least squares solution for the m_i, unique up to a positive affine transformation.

Confidence intervals

Confidence intervals are simulated for the Thurstone C linear and the NEL models. The procedure is roughly as follows:

1. The model is solved using the expert data and reference values. This yields values for the model parameters m_i (Thurstone C) or λ_i (NEL).

2. Expert responses are simulated by appropriately sampling distributions with the model parameter values emerging from step 1.

3. The model is solved again, using the expert preference data emerging from step 2, and the initial reference values. The results are stored.

4. Steps 2 and 3 are repeated 500 times for each model and 90% central confidence bands are extracted and displayed.

Probabilistic Inversion

Probabilistic inversion denotes the operation of inverting a function at a (set of) distributions, and it constitutes a different approach to utility quantification than the Bradley Terry or Thurstone models. This model involves one very mild assumption, which, in combination with a choice of starting "non-informative" starting distribution, yields an estimate of the joint distribution of utility scores over the population of experts.

<u>Assumption:</u> *There are two alternatives, not necessarily included in the set of items, for which all experts agree that one of these is strictly preferred to the other and that all items are between these two in preference.*

That is, we can find two alternatives, say "very very good" and "very very bad" which everyone agrees are strictly better and worse respectively, than the items of study. Decision theory teaches that all utility functions are unique up to a positive affine

transformation, that is, up to a choice of 0 and unit. With the above assumption, we may assume that all experts' utilities are normalized to the [0, 1] interval.

Starting distribution: *The starting distribution is a sample from product uniform distribution* $[0,1]^N$.

The size of the sample for the starting distributions is controlled under the "Options" field for probabilistic inversion. With N items, we start by assuming that the population of experts' utilities are independently uniformly distributed on the [0, 1] interval for each item.

We now want to 'minimally perturb' the starting distribution so as to comply with the expert preferences. That is if $U_1,\ldots U_N$ is a vector of utilities drawn from the perturbed distribution, then for each i, j, the following constraint is satisfied:

the probability $U_i > U_j$ *= percentage %(i,j) of experts who preferred i to j.*

Two algorithms are available to accomplish this. **Iterative Proportional Fitting (IPF)** finds the maximum likelihood distribution satisfying the expert preference constraints, relative to the starting distribution *IF* the problem is feasible. The problem may not be feasible; that is, there may be no distribution on $[0, 1]^N$ satisfying the above constraint. In this case IPF does not converge.

Infeasibility is rather common in this context, as there are a very large number of constraints. In such cases we seek a distribution which is 'minimally infeasible'. This can be found with the PARFUM algorithm. If the problem is feasible, the PARFUM algorithm returns a solution which is close to, but not equal to, the IPF solution. If the problem is infeasible PARFUM converges to a distribution which 'distributes the pain' equally over the constraints, in an appropriate information sense. Parfum takes much longer to converge.

For the 6 policies case, the "proportion matrix" showing the target preference probabilities (black) and the preference probabilities realized by IPF (gray) after 100 iterations. They are identical up to 4 decimals. This problem is feasible.

	1	2	3	4	5	6
1	0	0.7273 0.7273	0.7424 0.7424	0.5909 0.5909	0.7273 0.7273	0.8939 0.8939
2	0.2727 0.2727	0	0.5455 0.5455	0.4545 0.4545	0.6667 0.6667	0.9394 0.9394
3	0.2576 0.2576	0.4545 0.4545	0	0.3939 0.3939	0.6212 0.6212	0.8182 0.8182
4	0.4091 0.4091	0.5455 0.5455	0.6061 0.6061	0	0.6515 0.6515	0.8333 0.8333
5	0.2727 0.2727	0.3333 0.3333	0.3788 0.3788	0.3485 0.3485	0	0.7727 0.7727
6	0.1061 0.1061	0.0606 0.0606	0.1818 0.1818	0.1667 0.1667	0.2273 0.2273	0

The proportion matrix for PARFUM after 1000 iterations is shown below:

	1	2	3	4	5	6
1	0	0.7273 0.7272	0.7424 0.7424	0.5909 0.5910	0.7273 0.7272	0.8939 0.8940
2	0.2727 0.2728	0	0.5455 0.5454	0.4545 0.4547	0.6667 0.6665	0.9394 0.9394
3	0.2576 0.2576	0.4545 0.4546	0	0.3939 0.4447	0.6212 0.5707	0.8182 0.8182
4	0.4091 0.4090	0.5455 0.5453	0.6061 0.5553	0	0.6515 0.6518	0.8333 0.8583
5	0.2727 0.2728	0.3333 0.3335	0.3788 0.4293	0.3485 0.3482	0	0.7727 0.7323
6	0.1061 0.1060	0.0606 0.0606	0.1818 0.1818	0.1667 0.1417	0.2273 0.2677	0

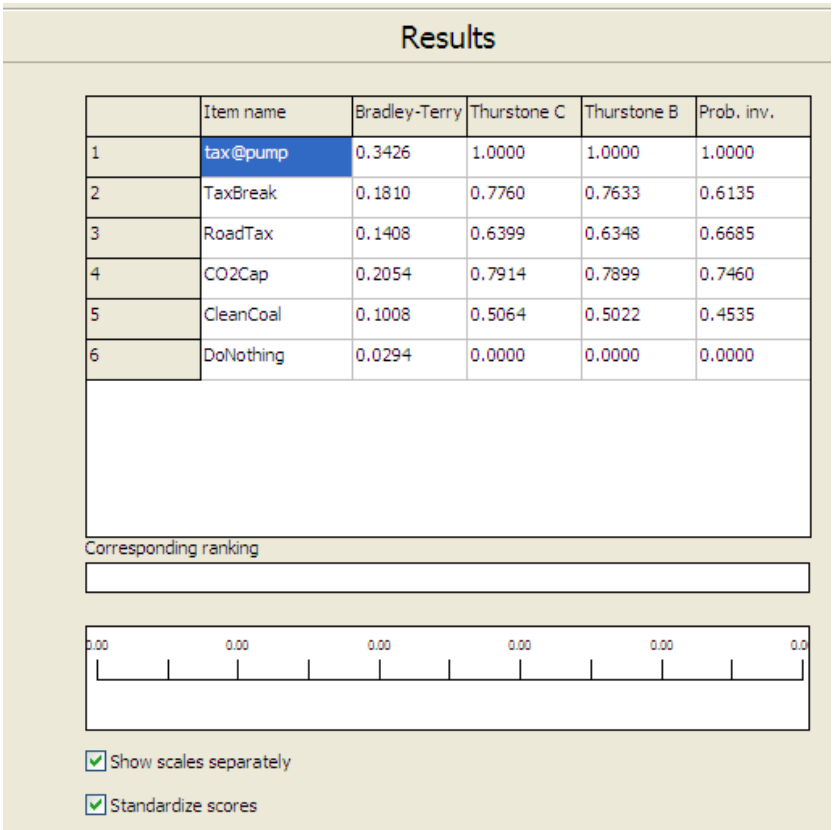
There are still substantial differences, the largest of which are highlighted. 5000 iterations are required to achieve the same accuracy as with 100 IPF iterations.

The correlation matrix for the utilities of the items is shown below for the IPF solution (10000 iterations):

	1	2	3	4	5	6
1	1	0.0430	0.0651	0.0042	-0.0354	0.0031
2	0.0430	1	0.0143	0.0196	0.0652	0.2817
3	0.0651	0.0143	1	-0.0172	-0.0370	0.0479
4	0.0042	0.0196	-0.0172	1	-0.0285	0.0419
5	-0.0354	0.0652	-0.0370	-0.0285	1	0.0697
6	0.0031	0.2817	0.0479	0.0419	0.0697	1

Note that these correlations are not identical. If we draw a 'random expert' the utility values for items 2 and 6 are correlated with 0.2817, whereas the values for items 2 and 1 are practically uncorrelated.

The results of the various models can be compared from the Results field, as shown below:



For the Thurstone and PI models, the option of rescaling the utility values so that the highest is 1.000 and the lowest is 0 is available. These are called standardized scores, and

they facilitate model comparisons. Standardization is not meaningful for the Bradley Terry model, as these are not measured on an affine scale, but on a ratio scale.

From the "Tools → generate report" menu option, a report can be generated with all input and output.

Issues and Features

UNIBALANCE has a few warts and blemishes, which however are easily worked around

1. **Error messages:** these are placed on a balk which is only visible if run full screen
2. **file management:** the first time UNIBALANCE is run, it creates a subdirectory SIMULATION in the directory where UNIBALANCE lives. That is by default the locus of all output files. Results are stored in files named file.*. file.sam is an ascii file read by UNIGRAPH and UNISENS. File.prb is a list of probabilities for resampling file.sam. Normally you have no use fot this file. File.txt is a space separated ascii file which you can open with EXCEL.
3. **UNIGRAPH** is called by UNIBALANCE after the file location has been entered in the OPTIONS field. However, you will have to open file.sam manually from UNIGRAPH. Alternatively, open UNIGRAPH directly, not via UNIBALANCE, and it will remember the last visited directory.
4. **Entering Data:** items and experts cannot be deleted once entered. If you enter items and forget to hit "OK" you don't see the number if items indicated. If you go back and re-enter the items, the old ones re-appear as "item 1,.." and cant be eliminated. Start over. Similar remarks hold for experts.
5. **Preference matrix** in Summary field: the p-value refers to the coefficient of agreement. These coefficients should only apply when all preferences are strict.
6. **Can't open file:** if UNIGRAPH is open, you cannot write another file.sam.
7. **Tolerance:** stops the iterative algorithms when a prescribed absolute difference between successive iterates is achieved. When tolerance = 0, a non convergence message will always appear. Tolerance is given in multiples of 10^{-15}, but only 7 digits are shown in the corresponding fenster.

Glossary of Terms

Circular Triads
A *circular triad* obtains in the preferences of an expert for items i,j,k if $i > j$, $j > k$, and $k > i$. Too many circular triads in an expert's data is an indication that his preferences are being "drawn at random", or that the items are indistinguishable. Simple computations and tests are available when all preferences are strict. Letting a_i denote the number of times that item i is (strictly) preferred to some other item, the number of circular triads in the response data for a given expert is

$$CT = \frac{T(T^2-1)}{24} - \frac{C}{2}$$

Where C =

$$\sum_{i=1}^{T}\left[a_i - \frac{\sum_{i=1}^{T} a_i}{T} \right]^2$$

If the number of items T < 7, we obtain the p-value of the hypothesis that the expert specifies his preference randomly from the tables (e.g. Table10C from Kendall (1975)). For more than 7 items, Kendall showed that the transformed number of circular triads is χ^2 distributed with *d* degrees of freedom:

$$d = \frac{T(T-1)(T-2)}{(T-4)^2}$$

$$\chi^2 = \frac{8}{n-4}\left(\frac{1}{4}\binom{n}{3} - CT + \frac{1}{2}\right) + d$$

Based on these equations we are able to test the hypothesis that the expert specified his preferences randomly. We would like to *reject* this hypothesis. If the p value is above, say, 0.05, then the 'at-random' hypothesis would not be rejected. In this case we could consider dropping the expert.

Coefficient of consistence

We can also introduce the *coefficient of consistence* ζ. It reaches the maximum value 1 if there are no inconsistencies in the data, which implies that the output of comparison process might be expressed as a ranking. The value decreases if the number of circular triads increases. The lowest value of ζ is 0.

$$\zeta = 1 - \frac{24 \cdot CT}{T(T^2-1)}, \text{ if number of items is odd}$$

$$\zeta = 1 - \frac{24 \cdot CT}{T(T^2-4)}, \text{ if number of items is even}$$

Coefficient of agreement

Let E be the number of experts, N the number of object is to be compared, and a(ij) the number of experts who prefer object i to j. Define "A CHOOSE B"
as the binomial coefficient A!/(B!(A-B)!), and define

$\Delta = \Sigma_{i \neq j}$ (a(ij) CHOOSE 2)

The coefficient of agreement u is defined as

$$u = \frac{2\Delta}{(E \text{ CHOOSE } 2) * (N \text{ CHOOSE } 2)} - 1.$$

u attains its maximal value, 1, when there is complete agreement. u can be used to test the hypothesis that all agreements are due to chance (see Experts in Uncertainty chapter 14). The coefficient of agreement is not diminished for lack of consistency in the experts; if there are 3 items and all say 1 > 2 > 3 > 1, then u = 1.

Coefficient of concordance
Let R(i,e) denote the rank of object i for expert e, where the rank is determined by the number of times e prefers i to some other object. Define

$$R(i) = \Sigma_{e=1..E} R(i,e)$$

$$S = \Sigma_{i=1..N}[(R(i) - (1/N) \times \Sigma_{i=1..N} R(j)]^2$$

The coefficient of concordance w is defined as

$$w = \frac{S}{E^2*(N^3-N)/12}.$$

w attains its maximal value, 1, when there is complete agreement, and can be used to test the hypothesis that all agreements are due to chance (see Experts in Uncertainty, chapter 14). If for three items, all experts say 1 > 2 > 3 > 1, then S = 0.

Iterative proportional fitting
This algorithm finds a distribution which is minimally informative (and thus maximum likely) relative to a starting distribution and which satisfies finitely many marginal constraints. The algorithm works by cyclically adapting the starting distribution to each constraint. If a distribution satisfying the constraints exists, which is absolutely continuous with respect to the starting distribution, then IPF converges rapidly.

PARFUM
The PARFUM algorithm is a variant on the IPF algorithm. Instead of cycling through the constraints, a starting distribution is adapted to each constraint individually, and these distributions are then averaged to form the next iteration. PARFUM always converges, and if the problem is feasible, it converges to a solution which is close to, though not identical with, the IPF solution.

Selected Bibliography

1. Anderson, S.P., A. de Palma and J-F Thissen, 1996 Discrete Choice Theory of Product Differentiation, MIT Press, Cambridge.
2. Bradley, R. (1953) " Some statistical methods in taste testing and quality evaluation" *Biometrica*, vol. 9, 22-38.
3. Bradley, R. and Terry, M. (1952) "Rank analysis of incomplete block designs *Biometrica*, vol. 39, 324-345.
4. Cooke R. M. and Misiewicz, J. 2007 Discrete Choice with Probabilistic Inversion: Application to energy policy choice and wiring failure, *Mathematical Methods in Reliability.*
5. Csiszar I. 1975 I-divergence geometry of probability distributions and minimization problems. *Ann. of Probab.*, 3:146-158.
6. Deming, W.E., and Stephan, F.F. 1944. On a least squares adjustment to sample frequency tables when the expected marginal totals are known, Ann Math. Statist. 40, 11, 427-44.
7. Du, C. Kurowicka D. and Cooke R.M. 2006 Techniques for generic probabilistic inversion, Comp. Stat. & Data Analysis (50), 1164-1187.
8. Environmental Security in Harbors and Coastal Areas Likeov I,. Kiker, G.A. and Wenning, R.J. (eds) Springer, dordrecht, 2007
9. French, S. (1988) *Decision Theory; an Introduction to the Mathematics of rationality* Ellis Horwood, Chichester.
10. Halpern, B.S., K.A. Selkoe. F. Micheli, and C.V. Cappel. 2007. Evaluating and ranking global and regional threats to marine ecosystems. Conservation Biology 21:1301-1315
11. Ireland C.T. and Kullback. S. 1968 Contingency tables with given marginals. *Biometrika*, 55:179-188.
12. Kraan B.C.P. and Cooke. R.M. 2000 Processing expert judgements in accident consequence modeling. *Radiation Protection Dosimetry*, 90(3):311-315.
13. Kraan B.C.P. and Cooke. R.M. 2000 Uncertainty in compartmental models for hazardous materials - a case study. *J. of Hazardous Materials*, 71:253-268.
14. Kraan, B.C.P and Bedford. T.J. 2005 Probabilistic inversion of expert judgements in the quantification of model uncertainty. *Management Science*, 51(6):995-1006.
15. Kraan. B.C.P. 2002 *Probabilistic Inversion in Uncertainty Analysis and related topics*. PhD dissertation, TU Delft, Dept. Mathematics.
16. Kruithof J.. Telefoonverkeersrekening. *De Ingenieur*, 52(8):E15{E25, 1937.
17. Kullback S. 1959 *Information theory and Statistics*. John Wiley and Sons, New York.
18. Kullback S. 1968 Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39(4):1236{1243.
19. Kullback S. 1971 Marginal homegeneity of multidimensional contingency tables. *The Annals of Mathematical Statistics*, 42(2):594-606.
20. Kurowicka D. and Cooke R.M. 2006. *Uncertainty Analysis with High Dimensional Dependence Modelling*. Wiley.

21. Linkov, I., Sahay, S. Kiker, G. Bridges, T. Belluck, D. and Meyer, A. (2005) "Multi-criteria decision analysis; comprehensive decision analysis tool for risk management of contaminated sediments"
22. Luce R. D. and P. Suppes, 1965, Preference, Utility, and Subjective Probability. In Handbook of Mathematical Psychology, vol. 3, ed. R. D. Luce, R. Bush, and E. Calanter. New York, Wiley.
23. Luce, R. D. 1959, Individual Choice Behavior; A Theoretical Analysis. New York, Wiley.
24. M.E.A. (Millennium Ecosystem Assessment). 2005. Ecosystems and Human Well-Being: Synthesis Report. Washington, D.C.
25. Matus, F. 2007 On iterated averages of I-projections, Statistiek und Informatik, Universit"at Bielefeld, Bielefeld, Germany, matus@utia.cas.cz.
26. McFadden, D. (1974) "Conditional logit analysis of qualitative choice behavior" in Zarembka, P (ed) *Frontiers in Econometrics*, New York Academic Press, USA, 105-142.
27. Siikamäki, J. and D. F. Layton, 2007 Discrete Choice Survey Experiments: A Comparison Using Flexible Methods, *Journal of Environmental Economics and Management*, Vol. 53, pp. 127-139.
28. Thurstone, L. (1927) "A law of comparative judgment" *Pschyl. Rev*. vol. 34, 273-286.
29. Torgerson, W. (1958) *Theory and Methods of Scaling*, Wiley, New York.
30. Train K.E. 2003 "Discrete Choice Methods with Simulation" Cambridge University Press.
31. Train, K. 1998 Recreation Demand Models with Taste Differences over People, *Land Economics* 74: 230-39.
32. Vomlel,J 1999 Methods of Probabilistic Knowledge Integration Phd thesis Czech Technical University, Faculty of Electrical Engineering.