Analyzing paired-comparison data in R using probabilistic choice models

Florian Wickelmaier

Eberhard Karls Universität Tübingen

The R User Conference, August 12-14, 2008

Overview

Probabilistic choice models

Goal: Scaling of psychological attributes

Procedure:

Participants are not asked to provide a numerical judgment (e. g., on a rating scale), but their behavior in a choice situation is observed. Scaling follows from modeling the data.

- Psychological theory of decision making
- Easy task for participants: pairwise comparison between alternatives, avoiding "scale usage heterogeneity"
- Measurement-theoretical foundation: testable conditions for numerical representation, unique scale level

Probabilistic choice models: applications

Main areas of application: consumer research, opinion surveys, sensory evaluation, psychophysical scaling

- Decision between insurance packages (McGuire & Davison, 1991, $N = 14000$)
- Political choice (Tversky & Sattath, 1979)
- Ranking of universities (Dittrich et al., 1998)
- Experimental perception research:
  - Measurement of pain (Matthews & Morris, 1995)
  - Taste, food quality (Bradley & Terry, 1952; Lukas, 1991; Duineveld et al., 1999)
  - Facial attractiveness (Bäuml, 1994)
  - Unpleasantness of environmental sounds (Ellermeier et al., 2004; Zimmer et al., 2004)
  - Sound quality of reproduction systems (Choisel & Wickelmaier, 2007)

Choice models (1): Bradley-Terry-Luce (BTL) model

Choice of an alternative $(x, y, \dots)$ is probabilistic and depends on the weight (strength) of the alternative $(u(x), u(y), \dots)$

BTL model equations:

$$P_{xy} = \frac{u(x)}{u(x) + u(y)} = \frac{1}{1 + \frac{k \cdot u(y)}{k \cdot u(x)}}$$

- P_{xy}: probability of choosing alternative x over y in a paired comparison
- $u(\cdot)$: ratio scale of the stimuli
- BTL model very parsimonious: only $n - 1$ free parameters, $n =$ number of stimuli
- BTL imposes strong restrictions on the choice probabilities

Independence of irrelevant alternatives (IIA)

Choice between two options is independent of the context provided by the choice set

$$\frac{P(x,\{x,y\})}{P(y,\{x,y\})} = \frac{P(x,\{x,y,z\})}{P(y,\{x,y,z\})}$$

Problem: similarity between groups of stimuli may cause IIA to fail (Debreu, 1960; Rumelhart & Greeno, 1971; Zimmer et al., 2004; Choisel & Wickelmaier, 2007)

Consequence of IIA: strong stochastic transitivity

$$P_{xy} \geq 0.5, P_{yz} \geq 0.5 \Rightarrow P_{xz} \geq \max\{P_{xy}, P_{yz}\}$$

Choice models (2): "Elimination by aspects" (EBA)
(Tversky, 1972)

Alternatives (stimuli) are characterized by various features (aspects)

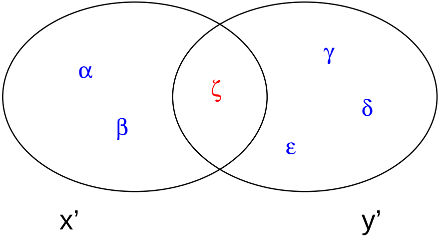
Choice is based on a hidden (sequential) elimination process:

- Aspects are chosen with a probability proportional to their weight (strength)
- Stimuli without the desired aspects are eliminated from the set of alternatives, until only one stimulus remains
- **Only the discriminating aspects influence the decision**

$\rightarrow$ EBA model does not require context independence (IIA)
$\rightarrow$ Bradley-Terry-Luce (BTL) model is a special case of EBA

Elimination by aspects (EBA): model equations

Stimuli $x, y, \ldots$ characterized by a set of aspects $x', y', \ldots$



Probability of choosing x over y:

$$P_{xy} = \frac{\sum_{\alpha \in x' \setminus y'} u(\alpha)}{\sum_{\alpha \in x' \setminus y'} u(\alpha) + \sum_{\beta \in y' \setminus x'} u(\beta)}$$

$x' \setminus y'$: aspects belonging to x, but not to y
$u(\cdot)$: ratio scale of the aspects
Scale value of x equals the sum of the characterizing aspect values

Example:

$x' = \{\alpha, \beta, \zeta\}$, $y' = \{\gamma, \delta, \varepsilon, \zeta\}$ $\rightsquigarrow$ $P_{xy} = \frac{u(\alpha)+u(\beta)}{u(\alpha)+u(\beta)+u(\gamma)+u(\delta)+u(\varepsilon)}$

The eba package

- Provides functionality for fitting and testing probabilistic choice models: Bradley-Terry-Luce, elimination by aspects, preference tree, Thurstone-Mosteller

- Key functions

`strans`	Counting stochastic transitivity violations
`eba`	Fitting and testing EBA models
`summary, anova` `plot, residuals`	Extractor functions
`group.test`	Comparing samples of subjects
`eba.order`	Testing within-pair order effects

- Manual

  Wickelmaier, F. & Schmid, C. (2004). A Matlab function to estimate choice-model parameters from paired-comparison data. *Behavior Research Methods, Instruments, & Computers*, **36**, 29–40.

Survey: perceived health risk of drugs

- $N = 192$ stratified by sex and age, 48 in each subgroup
- Task: Which of the two drugs do you judge to be more dangerous for your health?
- Drugs

Alcohol	Tobacco
Cannabis	Ecstasy
Heroine	Cocaine

- Each participant did all $6 \cdot 5/2 = 15$ pairwise comparisons.
- Analyses performed separately in the four subgroups

Descriptive statistics

Aggregate judgments (male participants, younger than 30)

	Alc	Tob	Can	Ecs	Her	Coc
Alc	0	28	35	10	4	7
Tob	20	0	18	2	0	3
Can	13	30	0	3	1	0
Ecs	38	46	45	0	1	17
Her	44	48	47	47	0	44
Coc	41	45	48	31	4	0

Probability of choosing x over y:

$$\hat{P}_{xy} = \frac{N_x}{N_x + N_y}$$

Example:

$$\hat{P}_{Alc,Tob} = \frac{28}{28 + 20} = 0.58$$

Counting the number of transitivity violations

```
strans(dat)
         violations  error.ratio  mean.dev  max.dev
weak              0         0.00    0.0000   0.0000
moderate          1         0.05    0.0417   0.0417
strong            5         0.25    0.0625   0.1458
---
Number of Tests:  20
```

BTL model

Fitting a BTL model using the `eba()` function

```
btl <- eba(dat)
```

Obtaining summary statistics and model tests

```
summary(btl)
```

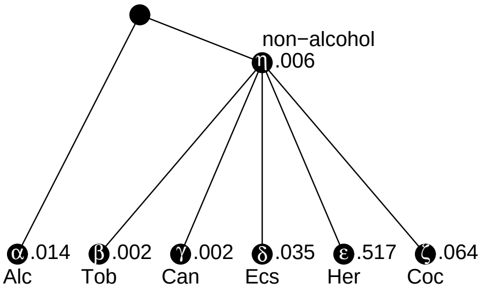
```
...
Model tests:
          Df1 Df2  logLik1 logLik2 Deviance Pr(>|Chi|)
EBA         5  15   -34.09  -21.62    24.94    0.00546 **
Effect      0   5  -284.57  -34.09   500.97    < 2e-16 ***
Imbalance   1  15   -42.84  -42.84     0.00    1.00000

AIC:  78.181
Pearson Chi2: 28.09
```

The BTL model does not describe the data adequately ($G^2(10) = 24.94$, $p < .001$).

EBA model with one additional aspect – EBA1

Model structure

$$A_1 = \{\{\alpha\}, \{\beta, \eta\}, \{\gamma, \eta\}, \{\delta, \eta\}, \{\varepsilon, \eta\}, \{\zeta, \eta\}\}$$



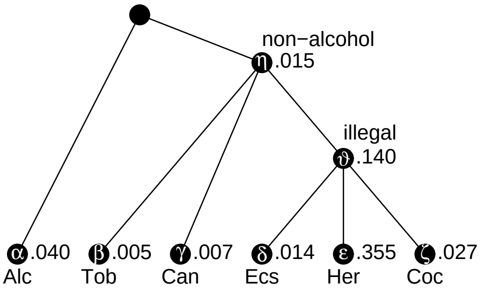

```
A1 <- list(c(1), c(2,7), c(3,7), c(4,7), c(5,7), c(6,7))
eba1 <- eba(dat, A1)
```

Non-alcohol drugs share a feature that affects decision when comparing them with alcohol.

EBA model with two additional aspects – EBA2

Model structure

$$A_2 = \{\{\alpha\}, \{\beta, \eta\}, \{\gamma, \eta\}, \{\delta, \eta, \vartheta\}, \{\varepsilon, \eta, \vartheta\}, \{\zeta, \eta, \vartheta\}\}$$




```
A2 <- list(c(1),c(2,7),c(3,7),c(4,7,8),c(5,7,8),c(6,7,8))
eba2 <- eba(dat, A2)
```

Three of the non-alcohol drugs share a feature that comes into play only when comparing them with the other drugs.

Model selection

Nested models can be compared using likelihood ratio tests.

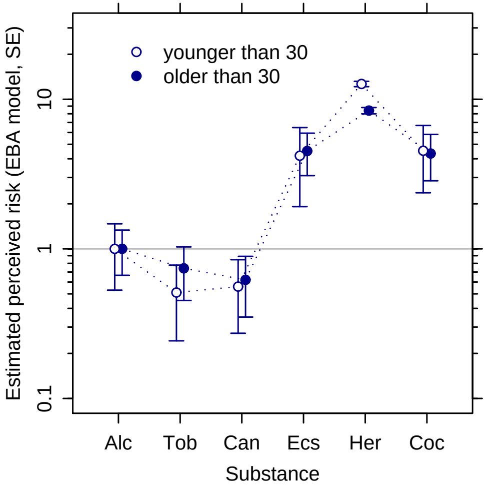
```
anova(btl, eba1, eba2)
  Model Resid. df Resid. Dev    Test Df LR stat.   Pr(Chi)
1   btl        10   24.94225         NA       NA        NA
2  eba1         9   17.54611 1 vs 2  1 7.396143 0.006536
3  eba2         8   11.45401 2 vs 3  1 6.092099 0.013579
```

Non-nested models may be selected based on information criteria.

```
AIC(btl, eba1, eba2)
     df      AIC
btl   5 78.18143
eba1  6 72.78528
eba2  7 68.69318
```

Conclusion: The elimination-by-aspects model with two extra parameters (eba2) fits the data best.

Scales derived from EBA model



- Younger males judge heroine to be about 13 times as dangerous as alcohol.
- Older males judge heroine to be only about 8 times as dangerous as alcohol.

Comparing subsamples

Is the same scaling valid in several groups?

Comparing male participants younger and older than 30 years

```
males <- array(c(young, old), c(6,6,2))

group.test(males, A2)
          Df1 Df2  logLik1  logLik2 Deviance Pr(>|Chi|)
EBA.g      14  30   -60.49   -48.94    23.09   0.111307
Group       7  14   -74.08   -60.49    27.18   0.000309 ***
Effect      0   7  -490.56   -74.08   832.96    < 2e-16 ***
Imbalance   1  30   -85.69   -85.69     0.00   1.000000
```

The scales of perceived health risk are significantly different ($G^2(7) = 27.18, p = .0003$) in the two groups.

Conclusions

- Pronounced differences between drugs w.r.t. perceived health risk
- Differences between male/female and younger/older participants
- Bradley-Terry-Luce model not valid in the male samples
- Elimination-by-aspects model with two additional parameters fits the data
- Elimination-by-aspects modeling is now easy to do using `eba()`

Thank you for your attention

florian.wickelmaier@uni-tuebingen.de

The 'eba' package   http://CRAN.r-project.org

References

Bäuml, K.-H. (1994). Upright versus upside-down faces: How interface attractiveness varies with orientation. *Perception & Psychophysics*, *56*, 163–172.

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*, 324–345.

Choisel, S. & Wickelmaier, F. (2007). Evaluation of multichannel reproduced sound: scaling auditory attributes underlying listener preference. *Journal of the Acoustical Society of America*, *121*, 388–400.

Debreu, G. (1960). Review of R. D. Luce's Individual choice behavior: A theoretical analysis. *American Economic Review*, *50*, 186–188.

Dittrich, R., Hatzinger, R., & Katzenbeisser (1998). Modelling the effect of subject-specific covariates in paired comparison studies withan application to university rankings. *Applied Statistics*, *47*, 511–525.

Duineveld, C. A. A., Arents, P., & King, B. M. (1999). Log-linear modelling of paired comparison data from consumer tests. *Food Quality and Preference*, *11*, 63–70.

Ellermeier, W., Mader, M., & Daniel, P. (2004). Scaling auditory unpleasantness according to the BTL model: Ratio-scale representation and psychoacoustical analysis. *Acta Acustica united with Acustica*, *90*, 101–107.

Lukas, J. (1991). BTL-Skalierung verschiedener Geschmacksqualitäten von Sekt (BTL scaling of different taste qualities of champagne). *Zeitschrift für experimentelle und angewandte Psychologie*, *38*, 605–619.

Matthews, J. N. S. & Morris, K. P. (1995). An application of Bradley-Terry-type models to the measurement of pain. *Applied Statistics*, *44*, 243–255.

McGuire, D. P. & Davison, M. L. (1991). Testing group differences in paired comparisons data. *Psychological Review*, *110*, 171–182.

Rumelhart, D. L. & Greeno, J. G. (1971). Similarity between stimuli: An experimental test of the Luce and Restle choice models. *Journal of Mathematical Psychology*, *8*, 370–381.

Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, *79*, 281–299.

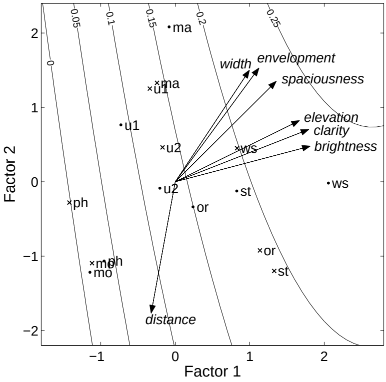
Tversky, A. & Sattath, S. (1979). Preference trees. *Psychological Review*, *86*, 542–573.

Wickelmaier, F. & Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, & Computers*, *36*, 29–40.

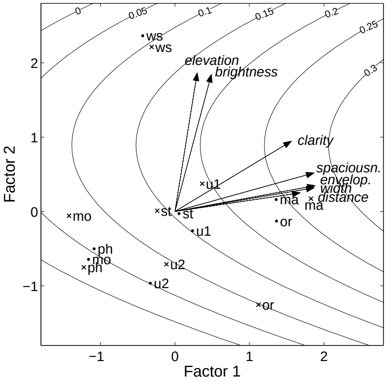
Zimmer, K., Ellermeier, W., & Schmid, C. (2004). Using probabilistic choice models to investigate auditory unpleasantness. *Acta Acustica united with Acustica*, *90*, 1019–1028.

Predicting preference from specific auditory attibutes
(Choisel & Wickelmaier, 2007, JASA)

Equal-preference contours for eight audio formats



Classical music



Pop music