LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Carolin Strobl, Florian Wickelmaier and Achim Zeileis

Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning

Technical Report Number 54, 2009
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

Accounting for Individual Differences in Bradley-Terry Models by Means of Recursive Partitioning

Carolin Strobl
Ludwig-Maximilians-
Universität München

Florian Wickelmaier
Eberhard Karls
Universität Tübingen

Achim Zeileis
WU Wirtschafts-
universität Wien

Abstract

The preference scaling of a group of subjects may not be homogeneous, but different groups of subjects with certain characteristics may show different preference scalings, each of which can be derived from paired comparisons by means of the Bradley-Terry model. Usually, either different models are fit in predefined subsets of the sample, or the effects of subject covariates are explicitly specified in a parametric model. In both cases, categorical covariates can be employed directly to distinguish between the different groups, while numeric covariates are typically discretized prior to modeling.

Here, a semi-parametric approach for recursive partitioning of Bradley-Terry models is introduced as a means for identifying groups of subjects with homogeneous preference scalings in a data-driven way. In this approach, the covariates that – in main effects or interactions – distinguish between groups of subjects with different preference orderings, are detected automatically from the set of candidate covariates. One main advantage of this approach is that sensible partitions in numeric covariates are also detected automatically.

Keywords: Bradley-Terry model, subject covariates, recursive partitioning.

1. Introduction

The Bradley-Terry model (abbreviated as the "BT model" hereafter, Bradley and Terry 1952) is the most widely used means for deriving a latent preference scale from paired comparison data when no natural measuring scale is available. It has been applied in a variety of fields in psychology and related disciplines. Early applications and developments are summarized in an extensive bibliography compiled by Davidson and Farquhar (1976) containing more than 350 references. More recent applications include, for example, surveys on health care, educational, and political choice (see, e.g., McGuire and Davison 1991; Dittrich, Hatzinger, and Katzenbeisser 1998; Dittrich, Francis, Hatzinger, and Katzenbeisser 2006) as well as sensory studies on the evaluation of pain, sound, and taste (see, e.g., Matthews and Morris 1995; Choisel and Wickelmaier 2007; Oberfeld, Hecht, Allendorf, and Wickelmaier 2009).

In many applications it is reasonable to assume that the preference scaling of a group of subjects not only depends on characteristics of the stimuli to be judged by the subjects, but also on characteristics of the subjects themselves. It is common practice to fit separate BT models, e.g., for younger and older participants (McGuire and Davison 1991; Kissler and

Bäuml 2000). In more advanced approaches (such as Dittrich *et al.* 1998; Böckenholt 2001), the covariates are explicitly included in the model.

Here, we want to illustrate a new approach for incorporating subject covariates in BT models: The approach of model-based recursive partitioning, that is well established in the field of statistics and machine learning, can be applied intuitively to identifying groups of subjects that differ in their preference scalings.

The approach of model-based recursive partitioning in general, as well as the framework for treating the BT model with this approach, is introduced in the following section. Two application examples are presented to illustrate the usage and benefits of this new technique for incorporating subject covariates in BT models.

2. Methods

Model-based partitioning employs the same principle as the more widely known classification and regression trees (Breiman, Friedman, Olshen, and Stone 1984): The covariate space is recursively partitioned to distinguish between groups of subjects with different characteristics. In the following, we will briefly outline the rationale of recursive partitioning in general, before we provide the framework and technical details for model-based partitioning of the BT model.

2.1. A brief introduction to recursive partitioning

Following the principle of recursive partitioning, classification and regression trees produce a tree-structured partition of the covariate space, where, starting with the entire sample, subjects are divided into groups according to their values of selected covariates. The splitting rules represented by the tree are chosen such that the subjects within the resulting groups have similar values of the response variable, whereas their response values differ from the subjects in the other groups.

The illustrative example in Figure 1 shows a regression tree as applied to an artificial data set with response income and covariates gender and age. The tree detects three groups with different income levels: Subjects under the age of 30 (in the leftmost Node 2) have a low average income, male subjects over 30 (in the middle Node 4) have a high and females over 30 (in the rightmost Node 5) have a medium average income. (Note that the node numbers are only labels assigned recursively from left to right starting from the top node.)

In comparison to simple classification and regression trees, model-based partitioning does not aim at finding groups of subjects with different values of the response variable, but with different values of certain model parameters. Such parameters could be, e.g., intercepts and slopes in a linear regression or – as in our case – the worth parameters of the stimuli in a BT model, which may vary between groups of subjects.

An example for a BT-based tree is displayed in Figure 2. Here, the preference scales for the stimuli derived from the BT model vary between the groups of subjects represented by the terminal nodes: Subjects up to the age of 52 who anwered yes to question q2 (in the leftmost Node 3) clearly prefer the third stimulus over all other stimuli, while, e.g., subjects over 52 (in the rightmost Node 7) prefer all other stimuli over the second stimulus, etc.

In the remainder of this section, the construction of the tree is described in detail, serving as an illustration of the general method for estimation of BT tree models.

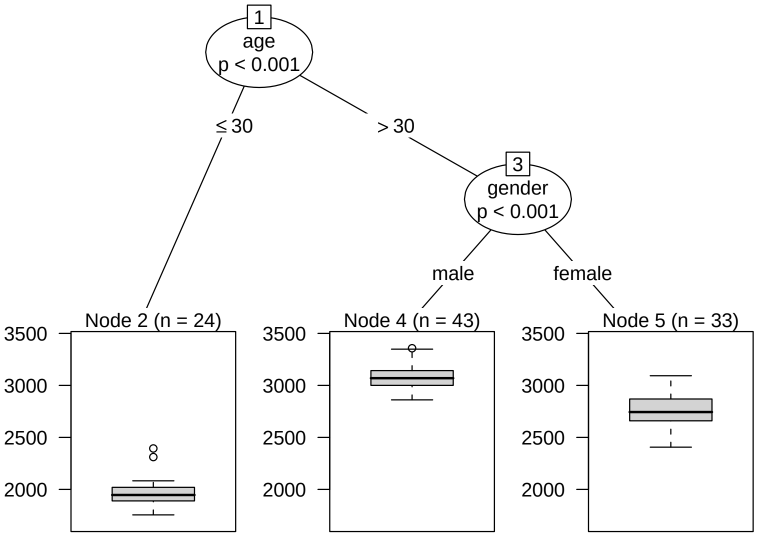

Figure 1: Exemplary regression tree indicating that the average income varies in groups defined by a combination of the covariates age and gender (artificial data).

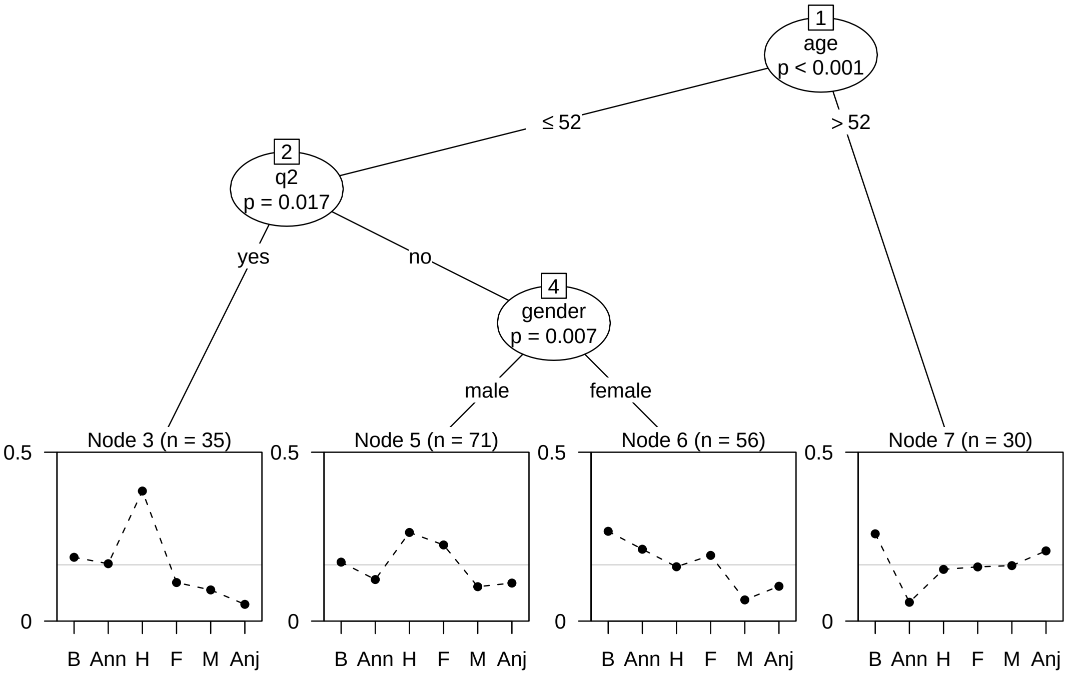

Figure 2: Partitioned paired comparison model for the Germany's Next Topmodel 2007 data, indicating that judged attractiveness varies in groups defined by combinations of the covariates age, q2 and gender (B: Barbara, Ann: Anni, H: Hana, F: Fiona, M: Mandy, Anj: Anja).

The data underlying the tree displayed in Figure 2 were collected in a study at the Department of Psychology, Universität Tübingen: $n = 192$ subjects were interviewed and asked to judge the attractiveness of the candidates of the second season of "Germany's Next Topmodel", which aired March–May 2007. Germany's Next Topmodel is a casting show for topmodel-hopefuls on German television – an adaptation of the corresponding US show – hosted by senior-topmodel Heidi Klum (see, e.g., Wikipedia 2009).

Based on photos taken at the beginning of the season (see Appendix A), the participants of the study judged the attractiveness of the $k = 6$ contestants (Barbara Meier, Anni Wendler, Hana Nitsche, Fiona Erdmann, Mandy Graff and Anja Platzer, listed here in decreasing order, i.e., starting with the winner of the show, Barbara Meier) in a forced choice experiment. Additionally, several subject-specific covariates about the raters are available: gender, age, and the answers (yes/no) to the following three questions:

(q1) Do you know the women on the photos? Do you know the TV show Germany's Next Topmodel?

(q2) Did you watch the latest season of Germany's Next Topmodel regularly?

(q3) Have you seen the final of the latest season of Germany's Next Topmodel? Do you know who won the latest season of Germany's Next Topmodel?

where for questions (q1) and (q3) a positive answer to at least one of the subquestions resulted in a positive overall answer. The sample was stratified with respect to gender and age (younger versus older than 30 years) with an equal number of subjects in each group. Overall, the sample contained 96 female and 96 male raters between the age of 15 and 77.

As explained in detail in the remainder of this section, the recursively partitioned BT model displayed in Figure 2 was generated by means of a simple algorithm consisting of the following consecutive steps:

1. Fit a BT model to the paired comparisons of all subjects in the current (sub-)sample, starting with the full sample.
2. Assess the stability of the BT model parameters with respect to each available covariate.
3. If there is significant instability, split the sample along the covariate with the strongest instability and use the cutpoint with the highest improvement of the model fit.
4. Repeat steps 1–3 recursively in the resulting subsamples until there are no more significant instabilities (or the subsample is too small).

We will now go through each of the steps of this algorithm, provide the statistical tools and explain how they were used to generate the model-based partition of the BT model depicted in Figure 2.

2.2. Fitting Bradley-Terry models

To fix notation, we consider paired comparison models with possible ties (see e.g., Critchlow and Fligner 1991, Section 4): Each comparison of two stimuli can result in (1) the first stimulus being preferred, (2) the second stimulus being preferred or (3) the subject being undecided between the two stimuli (i.e., a tie). The common forced choice experiments, where ties are prohibited by the experimental design, can be considered as a special case of this more general view.

In a notation similar to Critchlow and Fligner (1991), we consider $i = 1, \dots, n$ subjects who judge all unordered pairs of $j = 1, \dots, k$ stimuli. Thus, each subject performs $k^* = k \cdot (k-1)/2$ comparisons – each resulting in a choice for an answer c in $1, 2, 3$. According to the Davidson (1970) extension of the BT model, the three possible outcomes have probabilities:

$$\begin{aligned} p_{jj'1} &= \frac{\pi_j}{\pi_j + \pi_{j'} + \nu\sqrt{\pi_j \pi_{j'}}}, \\ p_{jj'2} &= \frac{\pi_{j'}}{\pi_j + \pi_{j'} + \nu\sqrt{\pi_j \pi_{j'}}}, \\ p_{jj'3} &= \frac{\nu\sqrt{\pi_j \pi_{j'}}}{\pi_j + \pi_{j'} + \nu\sqrt{\pi_j \pi_{j'}}}, \end{aligned}$$

where $\pi_j \geq 0$ $(j = 1, \dots, k)$ are stimulus-specific parameters, also called *worth parameters* or *merits*, and $\nu \geq 0$ is a discrimination constant governing the probability of ties.

This formulation of the model is easy to interpret but has two drawbacks when it comes to parameter estimation: it is over-identified (multiplication of all π_j with a constant does not change the probabilities) and the parameters are constrained to be non-negative. Hence, for parameter estimation one parameter is typically kept fixed and all others are considered on a log-scale yielding the k-dimensional parameter $\theta = (\log(\pi_1), \dots, \log(\pi_{k-1}), \log(\nu))^\top$. Without loss of generality $\log(\pi_k)$ is fixed at zero; equivalently, the sum of the worth parameters can be constrained to 1. This latter view will be adopted for reporting the π_j in our empirical results.

Note that the classical BT model for forced choice experiments without ties follows as the simple special case when $\nu = 0$ and thus $p_{jj'3} = 0$. Consequently, the parameter θ is only $k-1$-dimensional for the BT model.

Given $i = 1, \dots, n$ observations $y_i \in \{1, 2, 3\}^{k^*}$, i.e., each y_i containing the k^* comparisons with outcome $c = 1, 2, 3$, the joint log-likelihood is given by:

$$\begin{aligned} \log L(\theta \,|\, y_1, \dots, y_n) &= \sum_{i=1}^{n} \sum_{j<j'} \sum_{c=1}^{3} I(y_{i,jj'} = c) \log(p_{jj'c}) \\ &= \sum_{i=1}^{n} \Psi(y_i, \theta), \end{aligned}$$

where $\Psi(y_i, \theta)$ denotes the likelihood contribution of the i-th observation and $I(\cdot)$ is the indicator function. The parameter estimates $\hat{\theta}$ can then be obtained by maximum likelihood (ML) estimation:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \Psi(y_i, \theta).$$

Typically, the ML estimate is not derived by direct maximization of the multinomial likelihood above but instead by fitting a surrogate log-linear Poisson model for the aggregated frequencies

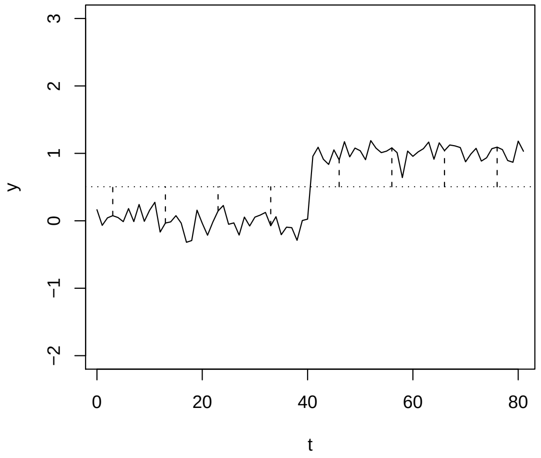

Figure 3: Structural change in mean stock return over time (artificial data). The dotted line indicates the overall mean, the dashed lines conveys that deviations from the overall mean are negative before the structural change and positive afterwards.

$n_{jj'c} = \sum_{i=1}^{n} I(y_{i,jj'} = c)$. This can be easily performed in many statistical software packages, see Critchlow and Fligner (1991) for more details.

2.3. Assessing parameter instability in Bradley-Terry models

After fitting the BT model, the next question is whether the set of parameters θ (i.e., the preference scale) is really the same for all n subjects, or whether there are subsamples with differing sets of parameters. As our aim is to capture potential effects of the availabe covariates, it should be formally tested whether there are parameter instabilities over one of the covariates. This step will be repeated reccursively within the newly created subsamples. However, in order to keep the notation simple, we only use the full-sample notation in the following.

An intuitive case of parameter instability (also termed structural change or structural break in the literature) is a change in the level of stock returns over time, as illustrated in the simplified artificial example in Figure 3. Technically speaking, what is depicted in Figure 3 is the change in the parameter “mean return” over the order implied by the variable “time” – and from this understanding, it is only a small step to describing the change in any kind of parameter over the range of any variable: The values of the model parameter of interest can be ordered with respect to each candidate variable, and the significance of the structural change over the range of this variable can be tested statistically.

Various approaches are conceivable for this objective. A particularly convenient one – due to its generality and ease of computation – is the usage of fluctuation tests (Zeileis and Hornik 2007) as adopted in the model-based recursive partitioning framework of Zeileis, Hothorn, and Hornik (2008). The idea of this class of tests is to compute subject-wise model deviations that should fluctuate randomly around zero under the null hypothesis of parameter stability.

In our example in Figure 3, under the null hypothesis of parameter stability the overall mean (dotted line) should hold over the entire time range. Accordingly, the deviations from the overall mean (dashed lines) should not show any systematic variation under the null hypothesis, while under the alternative of a structural break, we would expect the deviations to differ systematically from zero before and after the cutpoint, as is actually the case in Figure 3.

A general measure of deviation for likelihood-based models is the subject-wise *score function* or *estimating function*: the derivative of the likelihood contributions w.r.t. the parameter vector. For the BT model, these are given by:

$$\psi(y_i, \theta) \;=\; \frac{\partial \Psi(y_i, \theta)}{\partial \theta} \;=\; \sum_{j<j'} \sum_{c=1}^{3} I(y_{i,jj'} = c) \frac{\partial \log(p_{jj'c})}{\partial \theta}.$$

Thus, for computing the estimating functions for the parameters θ, the gradients of the log-probabilities $\log(p_{jj'c})$, $c = 1, 2, 3$, just need to be aggregated suitably. These can be shown to be:

$$\frac{\partial \log(p_{jj'1})}{\partial \theta_h} \;=\; \left\{ \begin{array}{lllrl} 1 & -p_{jj'1} & & -0.5\; p_{jj'3} & h = j \\ & & -p_{jj'2} & -0.5\; p_{jj'3} & h = j' \\ & & & - \quad p_{jj'3} & h = k \\ 0 & & & & \text{otherwise} \end{array} \right.$$

$$\frac{\partial \log(p_{jj'2})}{\partial \theta_h} \;=\; \left\{ \begin{array}{lllrl} & -p_{jj'1} & & -0.5\; p_{jj'3} & h = j \\ 1 & & -p_{jj'2} & -0.5\; p_{jj'3} & h = j' \\ & & & - \quad p_{jj'3} & h = k \\ 0 & & & & \text{otherwise} \end{array} \right.$$

$$\frac{\partial \log(p_{jj'3})}{\partial \theta_h} \;=\; \left\{ \begin{array}{lllrl} 0.5 & -p_{jj'1} & & -0.5\; p_{jj'3} & h = j \\ 0.5 & & -p_{jj'2} & -0.5\; p_{jj'3} & h = j' \\ 1 & & & - \quad p_{jj'3} & h = k \\ 0 & & & & \text{otherwise} \end{array} \right.$$

With this notion of model deviation available, it can be assessed wether systematic deviations occur along one of the m covariates: $x_{i\ell}$ $(i = 1, \dots, n,\ \ell = 1, \dots, m)$. To do so, the deviations are cumulatively aggregated along each of the m covariates:

$$W_\ell(t) \;=\; \hat{V}^{-1/2} n^{-1/2} \sum_{i=1}^{\lfloor n \cdot t \rfloor} \psi(y_{(i|\ell)}, \hat{\theta}) \qquad (0 \le t \le 1),$$

where the index $(i|\ell)$ denotes the i-th ordered observation w.r.t. the ℓ-th covariate, $\lfloor \cdot \rfloor$ denotes the integer part, and $\hat{V} = \sum_{i=1}^{n} \psi(y_i, \hat{\theta}) \psi(y_i, \hat{\theta})^\top$ is the outer-product-of-gradients estimate of the covariance matrix.

The cumulative aggregation is used here to incorporate the order of the individual deviations w.r.t. the considered covariate: The $i = 1, \dots, n$ individual deviations are ordered with respect to the ℓ-th covariate, and aggregated up to the $\lfloor n \cdot t \rfloor$-th element in each step. When $W_\ell(t)$ is considered as a function of the fraction t of the sample size, the null-model with no structural change corresponds to the path of a random process with constant zero mean.

The advantage of this approach is that the model does not have to be reestimated for all subsamples, because the individual deviations remain the same and only their order (and the corresponding path of $W_\ell(t)$) needs to be adjusted for evaluating the different covariates.

Under the null hypothesis of parameter stability, the cumulative sum process $W_\ell(\cdot)$ can be shown to converge to a k-dimensional Brownian bridge (Zeileis and Hornik 2007), which can be used as the basis for statistical inference. To capture systematic deviations in $W_\ell(\cdot)$ different test statistics can be used depending on wether the ℓ-th covariate is a numeric or a categorical variable. If it is numeric, Zeileis *et al.* (2008) point out that a natural test statistic is:

$$S_\ell \quad = \quad \max_{i=\underline{\imath},\dots,\overline{\imath}} \left(\frac{i}{n} \cdot \frac{n-i}{n} \right)^{-1} \left| \left| W_\ell \left(\frac{i}{n} \right) \right| \right|_2^2 .$$

This can be interpreted as the maximum Lagrange-multiplier statistic (also known as score statistic) for a single shift alternative over all conceivable cutpoints in $[\underline{\imath}, \overline{\imath}]$. The limiting distribution is the supremum of a tied-down Bessel process from which p-values can be computed (see Zeileis *et al.* 2008, for details).

If, on the other hand, the ℓ-th covariate is categorical (with $q = 1, \dots, Q$ categories, say), it is more natural to use the following test statistic:

$$S_\ell \quad = \quad \sum_{q=1}^{Q} n \left(\sum_{i=1}^{n} I(x_{i\ell} = q) \right)^{-1} \left| \left| \Delta_q W_\ell \left(\frac{i}{n} \right) \right| \right|_2^2 ,$$

where Δ_q is the increment within the q-th category. This test statistic is invariant to reordering of the Q categories and the subjects within each category. The test statistic captures the instability over the Q subsamples. Its limiting distribution is χ^2 with $(Q-1) \cdot k$ degrees of freedom from which p-values can be computed.

Although the technical details of this testing procedure are somewhat challenging, the results are easy to interpret: Parameter instability test statistics S_ℓ ($\ell = 1, \dots, m$) with associated p-values (corrected for multiple testing) are provided for each candidate variable. The variable with the smallest p-value is then used for determining the subsamples in the current step of the recursive partitioning algorithm – unless all p-values exceed the significance level (commonly 5%), indicating that there is no (more) significant parameter instability and thus no need for partitioning.

For our example, the parameter instability test statistics and p-values of each candidate splitting variable in the full sample are displayed in Table 1. Accordingly, the variable age associated with the smallest p-value is used for the first split in Figure 2. The choice of the cutpoint within the chosen splitting variable is discussed in the next section. In the resulting subsamples, splitting continues recursively, until no more significant parameter instability is detected or until the number of observations is a subsample falls below a given threshold.

	gender	age	q1	q2	q3
statistic	17.0880	32.3566	12.6320	19.8392	6.7586
p-value	0.0217	0.0008	0.1283	0.0067	0.7452

Table 1: Parameter instability test statistics $S_1, \dots, S_5$ and corresponding p-values for the full-sample BT model for the Germany's Next Topmodel 2007 data.

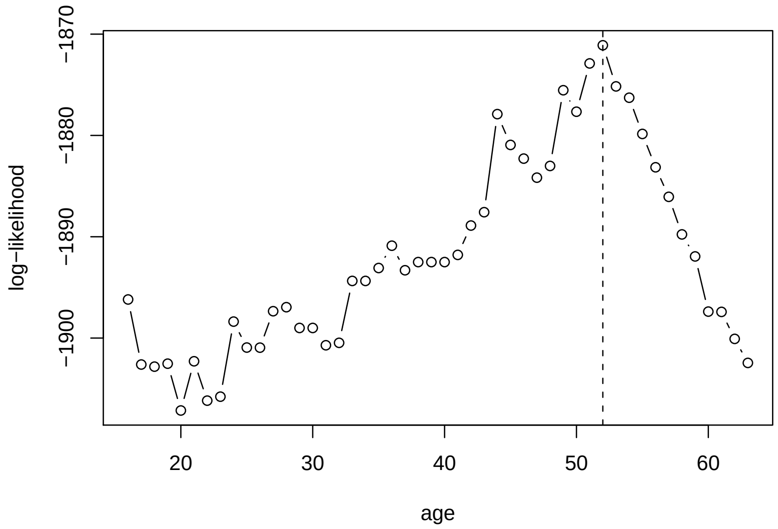

Figure 4: Log-likelihood of partitioned BT model for the first split in the covariate age.

2.4. Cutpoint selection in Bradley-Terry models

After the ℓ-th covariate was chosen for splitting, the optimal cutpoint within this variable is selected by maximizing the partitioned likelihood (i.e., the sum of the likelihoods for the observations before and after the cutpoint) over all candidate cutpoints.

More formally, for a numeric splitting variable, we can define the subsamples $L(\xi) = \{i \,|\, x_{i\ell} \leq \xi\}$ and $R(\xi) = \{i \,|\, x_{i\ell} > \xi\}$ on the left and right, respectively, of some cutpoint ξ. For both subsamples, the parameters $\hat{\theta}^{(L)}$ and $\hat{\theta}^{(R)}$ can be estimated as described above. To determine the optimal cutpoint ξ, the partitioned likelihood

$$\sum_{i \in L(\xi)} \Psi\left(y_i, \hat{\theta}^{(L)}\right) + \sum_{i \in R(\xi)} \Psi\left(y_i, \hat{\theta}^{(R)}\right)$$

is maximized over all candidate cutpoints ξ (typically requiring a certain minimal subsample size) as illustrated in Figure 4.

While this approach can be applied to numeric and ordered covariates, it is inappropriate for unordered categorical covariates. Instead, the Q categories of an unordered categorical covariate can be split into any two groups. From all these candidate binary partitions, again the one with the maximal partitioned likelihood is chosen. (Note that, in principle, the partitioning idea is not limited to binary splits – however, binary splits are typically found convenient in practice. See Zeileis *et al.* (2008) for strategies to compute multi-way splits.)

For our topmodel data example, Figure 4 depicts the partitioned log-likelihood for all candidate cutpoints within the range of the numeric covariate age that was selected for the first split. The maximum is achieved for the cutpoint $\xi = 52$. Accordingly, the sample is split into two subsamples with age ≤ 52 and age > 52, as displayed in Figure 2. Within the subsamples, splitting is again repeated recursively. However, for the following splitting variables q2 and gender no cutpoint selection is necessary because there are only two subsamples asociated with the two categories of both variables. Thus, the cutpoint is already determined by the selection of these covariates for splitting.

This concludes the discussion of the recursive partitioning procedure for BT models: The four basic steps – (1) BT model estimation, (2) parameter instability tests for splitting variable selection, (3) maximization of the segmented likelihood for cutpoint selection, (4) sample splitting – are repeated recursively until there are no more significant instabilities or the subsample size is too small to consider further splitting. Note that the significance level and minimal subsample size required for further splitting need to be defined by the researcher. While in most cases the common significance level of 5% will be appropriate, lower values should be chosen when the overall sample size is very large in order to avoid growing too complex trees that may induce overfitting. The minimal subsample-size, on the other hand, should be chosen such as to provide a sufficient basis for parameter estimation in each subsample, and should thus be increased when the number of stimuli becomes large.

For the application examples presented so far and in the following, that have moderate sample sizes and numbers of stimuli, a significance level of 5% and a minimal subsample size of 5 subjects was employed.

3. Application examples

For the illustrative topmodel example already presented in the previous section, a more thorough discussion is provided here to highlight the straightforward interpretability of the tree-structured model. Additionally, the BT tree method is applied to a more well-known data set from the field of education: Following Dittrich *et al.* (1998) and Böckenholt (2001), we investigate which covariates influence the choice of university for a study abroad among business students.

3.1. Germany's Next Topmodel 2007 data

The model-based partitioning procedure for the topmodel data was outlined in the previous section, with the resulting tree displayed in Figure 2. This data set is particularly useful for illustrating the BT tree method because, in addition to binary covariates with only a single potential cutpoint, it also contains the numeric covariate age. As emphasized above, one important advantage of model-based partitioning for including subject-covariate information in BT models is that such a numeric covariate does not need to be discretized in advance, but can be directly included in the analysis, where an appropriate cutpoint is selected in a data driven way.

In addition to the graphical representation of the partitioned model in Figure 2, the results can also be summarized by reporting the worth-parameter estimates (scaled to sum to unity) in each subsample, as in Table 2. These show that the rating of those subjects up to age 52 who watched the show on a regular basis (Node 3) essentially conforms with the assessment of the jury – except for the rating of the candidate Hana, who was judged by viewers of the show to be about twice as attractive as Barbara, the actual winner. This extreme preference for Hana cannot be found in any of the groups who did not watch the show on a regular basis. Of the subjects up to age 52, who did not watch the show on a regular basis, males (Node 5) have preferences for Hana and Fiona, while females (Node 6) rank Barbara highest, followed

	Barbara	Anni	Hana	Fiona	Mandy	Anja
Node 3	0.19	0.17	0.39	0.11	0.09	0.05
Node 5	0.17	0.12	0.26	0.23	0.10	0.11
Node 6	0.27	0.21	0.16	0.19	0.06	0.10
Node 7	0.26	0.06	0.15	0.16	0.16	0.21

Table 2: Estimates of worth parameters in terminal nodes from the partitioned paired comparison model for the Germany's Next Topmodel 2007 data.

by Anni and Fiona.

Interestingly, the preferences for older participants (Node 7) are completely different from all other groups: Unlike the other goups, subjects over 52 judged Anja to be almost as attractive as Barbara, while they strongly dislike Anni (her attractiveness scale value is only about 20% of Barbara's). In addition to that, this group shows almost no discrimination between Hana, Fiona and Mandy.

While the latter finding supports the common perception that the ideal of beauty varies between generations, the fact that those subjects who regularly watched the show have such a strong preference for one candidate may indicate that the candidates' personality, rather than their physical appearance, can be crucial for the audience's appreciation of candidates in casting shows.

3.2. CEMS university choice data

Students of the WU Wirtschaftsuniversität Wien can spend part of their study abroad, visiting one of currently 17 CEMS (Community of European Management Schools and International

	>	=	<	NA's
London : Paris	186	26	91	0
London : Milano	221	26	56	0
Paris : Milano	121	32	59	91
London : St Gallen	208	22	73	0
Paris : St Gallen	165	19	119	0
Milano : St Gallen	135	28	140	0
London : Barcelona	217	19	67	0
Paris : Barcelona	157	37	109	0
Milano : Barcelona	104	67	132	0
St Gallen : Barcelona	144	25	134	0
London : Stockholm	250	19	34	0
Paris : Stockholm	203	30	70	0
Milano : Stockholm	157	46	100	0
St Gallen : Stockholm	155	50	98	0
Barcelona : Stockholm	172	41	90	0

Table 3: Observed frequencies of comparisons for the CEMS university choice data.

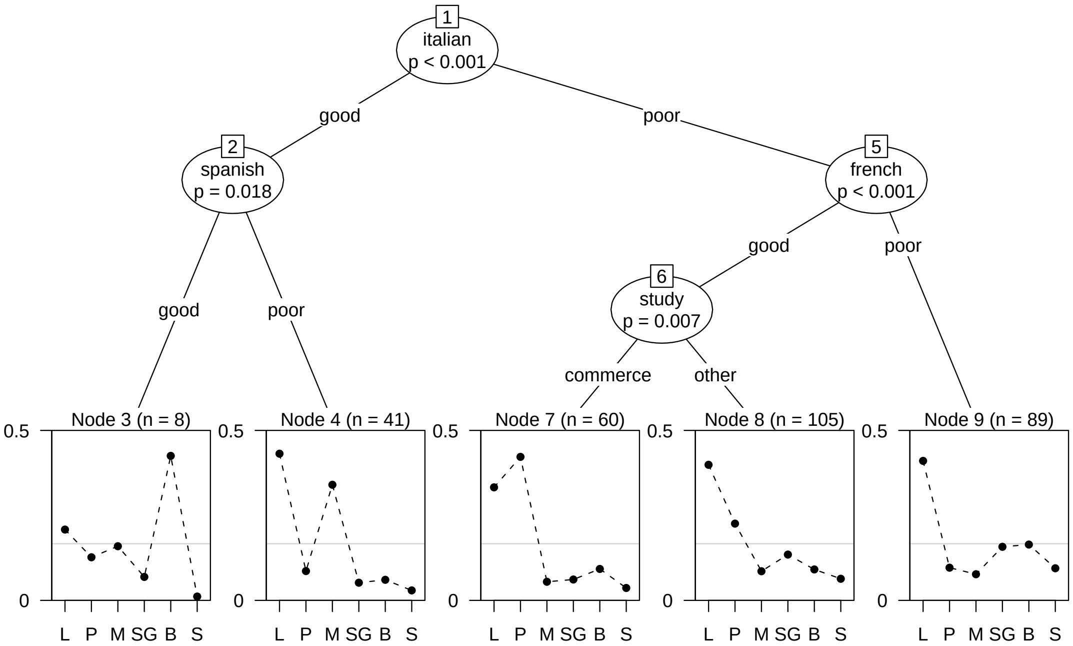

Figure 5: Partitioned paired comparison model for the CEMS university choice data (L: London, P: Paris, M: Milan, SG: St. Gallen, B: Barcelona, S: Stockholm).

	London	Paris	Milano	St Gallen	Barcelona	Stockholm
Node 3	0.21	0.13	0.16	0.07	0.43	0.01
Node 4	0.43	0.09	0.34	0.05	0.06	0.03
Node 7	0.33	0.42	0.05	0.06	0.09	0.04
Node 8	0.40	0.23	0.09	0.13	0.09	0.06
Node 9	0.41	0.10	0.08	0.16	0.16	0.09

Table 4: Estimates of worth parameters in terminal nodes from the partitioned paired comparison model for the CEMS university choice data.

Companies) universities. Dittrich *et al.* (1998) conduct and analyze a survey of $n = 303$ first-year students to examine the student's preferences for $k = 6$ CEMS universities located in different European cities: London School of Economics, Haut Etudes Commercials (Paris), Università Commerciale Luigi Bocconi (Milan), Universität St. Gallen, Escueala Superior de Administracion y Direccion de Empresas (Barcelona) and Handelshögskolan i Stockholm. To identify reasons for the students' preferences, several subject-specific covariates have been assessed as well.

The covariates included in the analysis are gender, major field of study and indicators of whether the students work full time, aim for an international degree, and have good skills in French, Spanish, and Italian. The aggregated observed frequencies $n_{jj'c}$ for the $k^* = 15$ possible comparisons are listed in Table 3. Note that in this examples ties are possible if a subject is undecided between two stimuli. For 91 subjects, the comparison Paris:Milan has unintentionally been left out (Dittrich *et al.* 1998).

To assess the influence of the subject-specific covariates, the paired comparison model is recursively partitioned. Figure 5 shows the resulting tree. The corresponging worth-parameter estimates (scaled to sum to unity) in each of the subsamples are displayed in Table 4.

The results show that the preference scaling of the universities highly depends on the subject covariates: While for the majority of students London is the most appealing option, students with good skills in Italian and Spanish (Node 3) have the strongest preference for Barcelona (more than twice as strong as for London), students with good skills in Italian but not in Spanish (Node 4) have a preference for Milan that is almost as pronounced as that for London, and for students with poor skills in Italian but good skills in French, the preference depends on the students' main discipline of study: Those students with an emphasis on commerce (Node 7) have a preference for Paris, that has a high reputation in this field, while the remaining students share the preference for London, that is most likely due to the fact that all Austrian university students have been exposed to several years of English language training.

Interestingly, Dittrich *et al.* (1998) point out that the low preference for Stockholm throughout the entire sample is most likely due to the fact that most students believe that lectures at Handelshögskolan i Stockholm were held in Swedish – while in fact they are being held in English, too.

Our results illustrate that the model-based partitioning approach for incorporating subject-covariates in BT models is convenient for identifying groups of subjects with common preference scales. All covariates found relevant for partitioning here were also included in the model of Dittrich *et al.* (1998). However, the visual representation as a tree makes the fitted models more accessible and intuitive to interpret compared to the parametric approaches of Dittrich *et al.* (1998) and Böckenholt (2001).

4. Summary and outlook

Model-based recursive partitioning is a flexible semi-parametric method adopted from machine learning, that is extended to BT models for identifying groups of subjects with different latent preference scales. The method employs splits in different covariates for partitioning the subjects, relying on the well-established statistical inference framework of fluctuation tests for detecting structural change points. Advantages of the resulting BT trees for paired comparison data are that (1) they are easy to interpret by means of visualization, (2) numeric covariates do not need to be discretized in advance, but suitable cutpoints are detected in a data-driven way, (3) from a potentially large number of candidate covariates those that correspond to a significant change in the model parameters are automatically detected and (4) interactions between covariates are also included in the same way.

Future work will aim at expanding applications of model-based partitioning in psychometrics to cover extensions of the BT model including observed stimulus-covariates (Dittrich *et al.* 1998) and latent characteristics of the stimuli as in the elimination by aspects (EBA) model (Tversky 1972), as well as the Rasch model (Rasch 1960) and its extensions.

Computational details

Our results were obtained using R 2.6.2 (R Development Core Team 2009) using the package **psychotree** 0.1-0 (Zeileis, Strobl, and Wickelmaier 2009) which implements BT trees as introduced in this manuscript. It relies on packages **party** 0.9-992 (Hothorn, Hornik, Strobl, and Zeileis 2009) and **prefmod2** 0.1-0 (Dittrich, Hatzinger, Strobl, Wickelmaier, and Zeileis 2009) for recursive partitioning and estimation of the BT model, respectively. The latter package contains the data for the topmodel and the university choice examples. R itself and all packages used are freely available under the terms of the General Public License from the Comprehensive R Archive Network at `http://CRAN.R-project.org/` and the R-Forge site at `http://R-Forge.R-project.org/`, respectively. Code for replicating our analysis is available in the **psychotree** package via `example("bttree", package = "psychotree")`.

References

Böckenholt U (2001). "Thresholds and Intransitivities in Pairwise Judgments: A Multilevel Analysis." *Journal of Educational and Behavioral Statistics*, **26**(3), 269–282.

Bradley RA, Terry ME (1952). "Rank Analysis of Incomplete Block Designs. I. The Method of Paired Comparisons." *Biometrika*, **39**(3/4), 324–345.

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984). *Classification and Regression Trees*. Chapman and Hall, New York.

Choisel S, Wickelmaier F (2007). "Evaluation of Multichannel Reproduced Sound: Scaling Auditory Attributes Underlying Listener Preference." *Journal of the Acoustical Society of America*, **121**(1), 388–400.

Critchlow DE, Fligner MA (1991). "Paired Comparison, Triple Comparison, and Ranking Experiments as Generalized Linear Models, and their Implementation on GLIM." *Psychometrika*, **56**(3), 517–533.

Davidson RR (1970). "On Extending the Bradley-Terry Model to Accomodate Ties in Paired Comparison Experiments." *Journal of the American Statistical Association*, **65**(329), 317–328.

Davidson RR, Farquhar PH (1976). "A Bibliography on the Method of Paired Comparisons." *Biometrics*, **32**(2), 241–252.

Dittrich R, Francis B, Hatzinger R, Katzenbeisser W (2006). "Modelling Dependency in Multivariate Paired Comparisons: A Log-Linear Approach." *Mathematical Social Sciences*, **52**(2), 197–209.

Dittrich R, Hatzinger R, Katzenbeisser W (1998). "Modelling the Effect of Subject-Specific Covariates in Paired Comparison Studies with an Application to University Rankings." *Journal of the Royal Statistical Society C*, **47**(4), 511–525.

Dittrich R, Hatzinger R, Strobl C, Wickelmaier F, Zeileis A (2009). ***prefmod2**: Paired Comparison Models for Preferences.* R package version 0.1-0/r39, URL `http://R-Forge.R-project.org/projects/prefmod/`.

Hothorn T, Hornik K, Strobl C, Zeileis A (2009). ***party**: A Laboratory for Recursive Partytioning.* R package version 0.9-995, URL http://CRAN.R-project.org/package=party.

Kissler J, Bäuml KH (2000). "Effects of the Beholder's Age on the Perception of Facial Attractiveness." *Acta Psychologica*, **104**(2), 145–166.

Matthews JNS, Morris KP (1995). "An Application of Bradley-Terry-Type Models to the Measurement of Pain." *Journal of the Royal Statistical Society C*, **44**(2), 243–255.

McGuire DP, Davison ML (1991). "Testing Group Differences in Paired Comparisons Data." *Psychological Bulletin*, **110**(1), 171–182.

Oberfeld D, Hecht H, Allendorf U, Wickelmaier F (2009). "Ambient Lighting Modifies the Flavor of Wine." *Journal of Sensory Studies*. In press.

Rasch G (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* The University of Chicago Press, Chicago. Reprint 1980.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Tversky A (1972). "Elimination by Aspects: A Theory of Choice." *Psychological Review*, **79**(4), 281–299.

Wikipedia (2009). "Germany's Next Topmodel – Wikipedia, The Free Encyclopedia." URL http://en.wikipedia.org/wiki/Germany's_Next_Topmodel, accessed 2009-02-06.

Zeileis A, Hornik K (2007). "Generalized M-Fluctuation Tests for Parameter Instability." *Statistica Neerlandica*, **61**(4), 488–508.

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**(2), 492–514.

Zeileis A, Strobl C, Wickelmaier F (2009). ***psychotree**: Recursive Partitioning Based on Psychometric Models.* R package version 0.1-0/r37, URL http://R-Forge.R-project.org/projects/prefmod/.

A. Stimuli for Germany's Next Topmodel 2007 data

The photos of the candidates of Germany's Next Topmodel 2007 that were presented to the participants of the study are displayed in Figure 6. They are arranged here in the order of their placement by the jury (left to right, top to bottom): Barbara Meier, Anni Wendler, Hana Nitsche, Fiona Erdmann, Mandy Graff, and Anja Platzer. All photos were obtained from the webpage of the hosting TV channel ProSieben (http://www.ProSieben.de/) where they were available when the show originally aired.

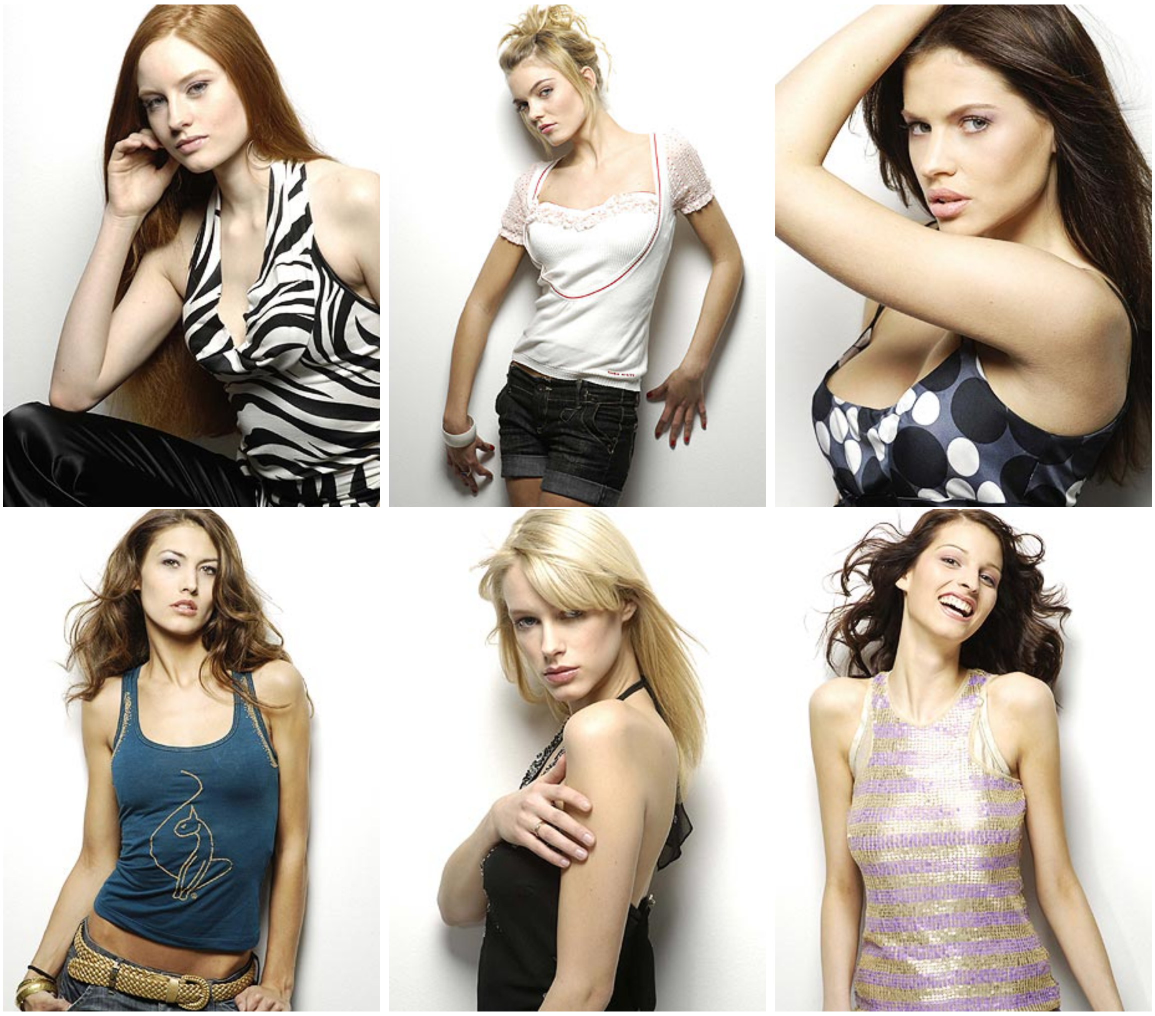

Figure 6: Photos presented to the participants of the Germany's Next Topmodel 2007 study, conducted by the Department of Psychology, Universität Tübingen.

Affiliation:

Carolin Strobl
Department of Statistics
Ludwig-Maximilians-Universität München
Ludwigstraße 33
DE-80539 München, Germany
E-mail: Carolin.Strobl@stat.uni-muenchen.de
URL: http://www.stat.uni-muenchen.de/~carolin/

Florian Wickelmaier
Department of Psychology
Eberhard Karls Universität Tübingen
Friedrichstraße 21
DE-72072 Tübingen, Germany
E-mail: Florian.Wickelmaier@uni-tuebingen.de
URL: http://homepages.uni-tuebingen.de/florian.wickelmaier/

Achim Zeileis
Department of Statistics and Mathematics
WU Wirtschaftsuniversität Wien
Augasse 2–6
AT-1090 Wien, Austria
E-mail: Achim.Zeileis@R-project.org
URL: http://statmath.wu.ac.at/~zeileis/