Statistical Modelling 2015; **15(6)**: 526–547

Subject-specific Bradley–Terry–Luce models with implicit variable selection

Giuseppe Casalicchio[1], Gerhard Tutz[1] and Gunther Schauberger[1]
[1]Department of Statistics, Ludwig-Maximilians-University Munich, Germany

Abstract: The Bradley–Terry–Luce (BTL) model for paired comparison data is able to obtain a ranking of the objects that are compared pairwise by subjects. The task of each subject is to make preference decisions in favour of one of the objects. This decision is binary when subjects prefer either the first object or the second object, but can also be ordinal when subjects make their decisions on more than two preference categories. Since subject-specific covariates, which reflect characteristics of the subject, may affect the preference decision, it is essential to incorporate subject-specific covariates into the model. However, the inclusion of subject-specific covariates yields a model that contains many parameters and thus estimation becomes challenging. To overcome this problem, we propose a procedure that is able to select and estimate only relevant variables.

Key words: boosting; Bradley–Terry–Luce model; paired comparison; subject-specific covariate; variable selection

Received June 2014; **revised** December 2014; **accepted** January 2015

1 Introduction

In paired comparisons, several objects are compared pairwise to obtain an overall preference ranking of the objects. In many application fields, such as, marketing research or psychometric experiments, objects are presented in a pairwise manner to judges, or, as they are called here, subjects. Their task for each comparison is to make a preference decision in favour of one of the presented objects according to specific subjective criteria, for example, the fragrance of perfumes when two perfumes are the objects being compared. Subjects can typically choose to prefer either the first or the second object, so that the response is represented by a binary variable. Alternatively, the subjects can make their preference decisions on more than two preference categories, such as, preferring the first object strongly or weakly, preferring neither of the objects, or preferring the second object strongly or weakly. This procedure yields an ordinal response that allows for a more precise preference ranking of the objects because it uses additional information about how strongly an object is preferred.

One of the most widely used models for paired comparisons with a binary response is the model suggested by Bradley and Terry (1952). It is closely related

Address for correspondence: Giuseppe Casalicchio, Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstrasse 33, D–80539 Munich, Germany.
E-mail: giuseppe.casalicchio@stat.uni-muenchen.de

 10.1177/1471082X15571817

to the choice axiom of Luce (1959) with the restriction of choices to between two objects. Thus, the model is also known as Bradley–Terry–Luce (BTL) model. Several extensions have been proposed in the literature to allow for responses with more than two preference categories. The former approaches allowed only a third category and were proposed by Rao and Kupper (1967) and Davidson (1970). Later, ordinal BTL models that allow for any number of ordered response categories were considered by Tutz (1986), Agresti (1992), Dittrich *et al.* (2004) and Dittrich *et al.* (2007). Ordinal regression models, in particular the cumulative logit model and the adjacent categories logit model, are in common use when an ordinal response is present. The models presented in this paper assume independence among all observations, so that pairwise ratings from the same subject are modelled as independent. However, in some applications, including a dependence structure is more realistic. In this context, Böckenholt (2001) introduced dependencies among observations by including subject- and object-specific random components.

The main objective of this article is to develop a method that allows for variable selection in a model that also contains subject-specific covariates. These covariates are characteristics of the subject and are assumed to affect the decision of a subject. Selection of relevant subject-specific covariates is very important because subject-specific BTL models typically contain a large number of parameters, and the estimation of all these parameters becomes challenging. Therefore, we focus on detecting important characteristics that determine the preference of objects.

2 Paired comparison models

2.1 The binary BTL model

Let M be the number of objects that are being compared and let the pair (r, s) refer to the comparison of object r and object s. The response connected to this comparison is denoted by Y_{rs}, where $Y_{rs} = 1$ indicates the preference for object r and $Y_{rs} = 2$ indicates the preference for object s. Let $\pi_k^{(r,s)} := \mathbb{P}(Y_{rs} = k|(r, s))$, with $k = 1, 2$ and $\pi_1^{(r,s)} + \pi_2^{(r,s)} = 1$ denote the probability of whether object r (for $k = 1$) or object s (for $k = 2$) is preferred. The binary BTL model can then be written as a logistic model

$$\begin{aligned} \log\left(\pi_1^{(r,s)} \Big/ \left(1 - \pi_1^{(r,s)}\right)\right) &= \gamma_r - \gamma_s \\ &= x_1^{(r,s)}\gamma_1 + \ldots + x_{M-1}^{(r,s)}\gamma_{M-1} \\ &= \left(\mathbf{x}^{(r,s)}\right)^\top \boldsymbol{\gamma}, \end{aligned} \tag{2.1}$$

where the components of the vector $\left(\mathbf{x}^{(r,s)}\right)^\top = \left(x_1^{(r,s)}, \ldots, x_{M-1}^{(r,s)}\right)$ are defined as

$$x_m^{(r,s)} = \begin{cases} +1 \text{ if } m = r \\ -1 \text{ if } m = s \\ \ 0 \ \text{ otherwise.} \end{cases}$$

Given the pair (r, s), the vector $\left(\mathbf{x}^{(r,s)}\right)^\top = (0, \ldots, 1, 0, \ldots, -1, 0 \ldots, 0)$ has a 1 on the rth position and a -1 on the sth position. The vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{M-1})^\top$ contains all object parameters that are to be estimated, where each parameter γ_m reflects the worth of object m. For model identifiability, we use the restriction $\gamma_M = 0$, that is, object M is considered as reference object. The ranking of objects is based on estimated object parameters, where estimation is carried out by maximum likelihood estimation for ordinary logistic models (Turner and Firth, 2012).

2.2 Ordinal BTL models

In the binary BTL model, the subjects can prefer the first object or the second object of a comparison. In more general settings, the subjects can make their preference decisions on more than two preference categories. In this case, the binary response is extended to a symmetric ordinal response $Y_{rs} \in \{1, \ldots, K\}$, where K denotes an arbitrary number of response categories. In our notation, lower response categories indicate the preference for object r. Thus, $Y_{rs} = 1$ is the most favourable response category for object r and $Y_{rs} = K$ is the most favourable response category for object s, or, equivalently, the least favourable response category for object r. To ensure that the comparison of the pair (r, s) yields the same result as the comparison of the pair (s, r), the ordinal response is assumed to be symmetric, that is, $Y_{rs} = k \Leftrightarrow Y_{sr} = K - k + 1$ and therefore $\pi_k^{(r,s)} = \pi_{K-k+1}^{(s,r)}$ (Agresti, 1992).

2.2.1 The cumulative model

The cumulative BTL model for ordinal responses uses the cumulative probabilities $\mathbb{P}(Y_{rs} \leq k|(r, s)) = \pi_1^{(r,s)} + \ldots + \pi_k^{(r,s)}$ and $\mathbb{P}(Y_{rs} > k|(r, s)) = \pi_{k+1}^{(r,s)} + \ldots + \pi_K^{(r,s)}$, with $\mathbb{P}(Y_{rs} \leq k|(r, s)) + \mathbb{P}(Y_{rs} > k|(r, s)) = \pi_1^{(r,s)} + \ldots + \pi_K^{(r,s)} = 1$. The model can be formulated as a cumulative logit model (McCullagh, 1980) with the link function $g = (g_1, \ldots, g_{K-1})^\top$, where

$$g_k\left(\pi_1^{(r,s)}, \ldots, \pi_K^{(r,s)}\right) = \log\left(\frac{\pi_1^{(r,s)} + \ldots + \pi_k^{(r,s)}}{1 - \left(\pi_1^{(r,s)} + \ldots + \pi_k^{(r,s)}\right)}\right)$$

links the considered probabilities to the linear predictor $\eta_k^{(r,s)}$, so that

$$\begin{aligned} g_k\left(\pi_1^{(r,s)}, \ldots, \pi_K^{(r,s)}\right) = \eta_k^{(r,s)} &= \theta_k + (\gamma_r - \gamma_s) \\ &= \theta_k + (x_1^{(r,s)}\gamma_1 + \ldots + x_{M-1}^{(r,s)}\gamma_{M-1}) \\ &= \theta_k + \left(\mathbf{x}^{(r,s)}\right)^\top \boldsymbol{\gamma}, \end{aligned} \tag{2.2}$$

for all $k = 1, \ldots, K-1$ and with the threshold parameters $-\infty = \theta_0 < \theta_1 < \ldots < \theta_{K-1} < \theta_K = \infty$. Because of the symmetric response, the thresholds need to be restricted as follows (see Tutz, 1986):

(1) If K is odd

$$\theta_k = -\theta_{K-k} \quad \text{for} \quad k = 1, \ldots, \tfrac{K-1}{2} = \left\lfloor \tfrac{K-1}{2} \right\rfloor. \tag{2.3}$$

(2) If K is even

$$\theta_{K/2} = 0 \quad \text{and} \quad \theta_{K/2-k} = -\theta_{K/2+k} \quad \text{for} \quad k = 1, \ldots, \tfrac{K}{2} - 1 = \left\lfloor \tfrac{K-1}{2} \right\rfloor. \tag{2.4}$$

It can be seen that only $q := \left\lfloor \frac{K-1}{2} \right\rfloor$ threshold parameters need to be estimated; the other ones are determined by the symmetry constraints (2.3) and (2.4).

2.2.2 The adjacent categories model

Another model that also allows for ordinal responses is the adjacent categories BTL model. It is based on the adjacent categories logit model using the link function $g = (g_1, \ldots, g_{K-1})^\top$, with

$$g_k\left(\pi_1^{(r,s)}, \ldots, \pi_K^{(r,s)}\right) = \log\left(\frac{\pi_k^{(r,s)}}{\pi_{k+1}^{(r,s)}}\right).$$

The model is then defined by

$$g_k\left(\pi_1^{(r,s)}, \ldots, \pi_K^{(r,s)}\right) = \eta_k^{(r,s)} = \theta_k + \left(\mathbf{x}^{(r,s)}\right)^\top \boldsymbol{\gamma}, \qquad k = 1, \ldots, K-1. \tag{2.5}$$

Here, the same restrictions (2.3) and (2.4) are valid and allow for symmetric thresholds (see Agresti, 1992). The adjacent categories BTL model can also be represented as a log-linear model, which has been extensively discussed in the literature (see Agresti, 1992; Dittrich *et al.*, 1998, 2004, 2007). When the response consists of $K = 2$ categories, the binary BTL model (2.1) is a special case of the cumulative BTL model (2.2) and the adjacent categories BTL model (2.5). In this case, the cumulative BTL model and the adjacent categories BTL model are equivalent.

It should be noted that the cumulative model and the adjacent categories model for ordinal responses use the same set of linear predictors

$$\eta_k^{(r,s)} = \theta_k + \left(\mathbf{x}^{(r,s)}\right)^\top \boldsymbol{\gamma}, \qquad k = 1, \ldots, q. \tag{2.6}$$

Therefore, the inclusion of subject-specific covariates for both models is obtained by modifying the same linear predictors.

3 Including subject-specific covariates

A very restrictive assumption for the models considered in Section 2.2 is that the worth of an object is the same for all subjects. Therefore, the ranking of the objects is the same for all subjects. However, in most applications it is to be expected that the preference for an object and thus the ranking of all objects depends on characteristics of the subject that makes the preference decision (Dittrich *et al.*, 1998; Francis *et al.*, 2002). To explicitly model this heterogeneity, we introduce subject-specific covariates $x_{i,1}, \ldots, x_{i,P}$, which can be both categorical or continuous characteristics of the subject. In this case, P reflects the number of characteristics and i refers to a single subject. The object parameters are now assumed to be determined by these characteristics, such that the object parameters vary across subjects. The resulting object parameter for subject i is then specified by

$$\gamma_{i,m} = \gamma_m + \sum_{p=1}^{P} x_{i,p}\gamma_{m,p},$$

where γ_m is a parameter for object m that is independent of the characteristics of the subject, and $\gamma_{m,p}$ is a modifying effect for object m depending on the pth subject-specific covariate. This means that $\gamma_{m,p}$ is a subject–object interaction parameter. To ensure model identifiability, we constraint the subject–object interaction parameters by setting $\gamma_{M,p} = 0$, for all $p = 1, \ldots, P$ (Francis *et al.*, 2002). Assuming that there are $i = 1, \ldots, I$ subjects, the linear predictor $\eta_k^{(r,s)}$ from equation (2.6) can be replaced by a more flexible linear predictor $\eta_k^{(r,s),i}$ that also considers subject-specific covariates and has the form

$$\begin{aligned}
\eta_k^{(r,s),i} = \theta_k + (\gamma_{i,r} - \gamma_{i,s}) &= \theta_k + (\gamma_r - \gamma_s) + \sum_{p=1}^{P} x_{i,p}(\gamma_{r,p} - \gamma_{s,p}) \\
&= \theta_k + \sum_{m=1}^{M-1} x_m^{(r,s)}\gamma_m + \sum_{p=1}^{P}\sum_{m=1}^{M-1} x_{i,p}x_m^{(r,s)}\gamma_{m,p} \\
&= \theta_k + \left(\mathbf{x}^{(r,s)}\right)^\top \boldsymbol{\gamma} + \sum_{p=1}^{P} \left(\mathbf{x}_p^{(r,s),i}\right)^\top \boldsymbol{\gamma}_p.
\end{aligned}$$

Here, the vector $\left(\mathbf{x}_p^{(r,s),i}\right)^\top = \left(x_{i,p}x_1^{(r,s)}, \ldots, x_{i,p}x_{M-1}^{(r,s)}\right)$ contains all subject–object interactions that belong to the pth subject-specific covariate and $\boldsymbol{\gamma}_p^\top = (\gamma_{1,p}, \ldots, \gamma_{M-1,p})$ is the corresponding vector of subject–object (interaction) parameters, such that $\gamma_{m,p}$ refers to the mth object and the pth subject-specific covariate.

In the absence of subject-specific covariates, that is, $x_{i,p} = 0$, for all i, p, one obtains an ordinal BTL model as described previously. The model that considers subject-specific covariates is more flexible and accounts for heterogeneity between subjects but suffers from the large number of parameters. This is because one has to

estimate $M-1$ subject–object parameters, namely, $\boldsymbol{\gamma}_p^\top = (\gamma_{1,p}, \ldots, \gamma_{M-1,p})$, for each additional subject-specific covariate (that is, when p increases by 1). To overcome this problem, we present a component-wise boosting algorithm that implicitly selects the influential variables.

4 Boosting

4.1 Basic concept of boosting

Before introducing the boosting algorithm for ordinal BTL models, we will first illustrate the concept of boosting on a linear regression model and then proceed to boosting for ordinal BTL models.

Assume the linear regression model with C covariates and J observations. In matrix representation, the model can be described as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

with the vectors $\mathbf{y} = (y_1, \ldots, y_J)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_J)^\top$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_C)^\top$ and the design matrix $\mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_C]$ that contains the components $\mathbf{x}_c = (x_{1,c}, \ldots, x_{J,c})^\top$ for $c = 0, 1, \ldots, C$. The structure of a generalized linear model (GLM) is determined by $\mu_j = \mathbb{E}(y_j|\mathbf{x}_j) = h(\eta_j)$, where h is a known response function and $g = h^{-1}$ is the link function that links the conditional expectation $\mathbb{E}(y_j|\mathbf{x}_j)$ to the linear predictor $\eta_j = \sum_{c=0}^{C} x_{j,c}\beta_c$. Using the vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)^\top$ and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_J)^\top$, we can also write $\boldsymbol{\mu} = h(\boldsymbol{\eta})$ and $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$. The linear model above can be incorporated within the framework of GLMs by using the identity function as the link function, that is, $\mu_j = g(\mu_j) = \eta_j = h(\eta_j)$. For a large number of covariates C, it is of interest to estimate the model using only influential variables. There are several methods that perform well in such high-dimensional settings while offering a sparse model; one of these methods is boosting (see Bühlmann, 2006). Boosting has its origin in the machine learning community (see Schapire, 1990; Freund, 1995, 1997) and can be formulated within the framework of statistical modelling as iteratively fitting parts of the model using the residuals of the previous iteration as the response of the current iteration (Friedman *et al.*, 2000; Friedman, 2001). The basic concept of boosting is to use a set of so-called base learners $f(\cdot)$ and combine them to gain a strong learner. A base learner can be a function of a set of covariates or even a single covariate. In this article, we will consider only linear base learners that are able to express a linear effect of the considered covariate(s). In the linear model, a single linear base learner for the cth component is, for example, $f(\mathbf{x}_c, \beta_c) = \mathbf{x}_c\beta_c$.

Tutz and Binder (2006) introduced likelihood-based boosting for GLMs. Instead of the iteratively refitting of residuals, they used the linear predictor $\boldsymbol{\eta}^{(b-1)}$ of the previous iteration $(b-1)$ as an offset that adds up to the linear predictor $\boldsymbol{\eta}^{(b)}$ of the current iteration b. We describe a component-wise boosting algorithm, where component-wise means that in each iteration only a single component is considered at a time.

Component-wise Boosting

Step 1: Initialization

Fit the intercept model $\boldsymbol{\mu}^{(0)} = h(\boldsymbol{\eta}^{(0)}) = h(\mathbf{x}_0\beta_0)$, with $\mathbf{x}_0 = (x_{1,0}, \ldots, x_{J,0})^\top = (1, \ldots, 1)^\top$ by maximizing the likelihood to obtain the estimated intercept $\hat{\beta}_0$. Initialize

$$\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_C)^\top = (\hat{\beta}_0, 0, \ldots, 0)^\top \quad \text{and} \quad \hat{\boldsymbol{\eta}}^{(0)} = \mathbf{x}_0\hat{\beta}_0.$$

Step 2: For each boosting iteration $b = 1, 2, \ldots, b_{stop}$

(a) *Estimation:* Fit the models

$$\boldsymbol{\mu} = h(\hat{\boldsymbol{\eta}}^{(b-1)} + f(\mathbf{x}_c, \beta_c))$$

for all components $c = 0, \ldots, C$ by one-step Fisher scoring, where $\hat{\boldsymbol{\eta}}^{(b-1)}$ is used as an offset. This yields the one-step Fisher scoring estimates $\hat{\beta}_c$ for the corresponding cth component. In general, the Fisher scoring algorithm is an iterative estimation procedure and uses the Fisher-matrix and the score function (for more details, see Tutz and Binder, 2006).

(b) *Selection:* From the models fitted in (a), choose the component c^* that yields the best fit with respect to some criteria (e.g., the model with the lowest deviance, Akaike information criterion (AIC) or Bayesian information criterion (BIC)) and set

$$\hat{\boldsymbol{\beta}}_{c^*} = (0, \ldots, 0, \hat{\beta}_{c^*}, 0, \ldots, 0)^\top.$$

(c) *Update:*

$$\hat{\boldsymbol{\eta}}^{(b)} = \hat{\boldsymbol{\eta}}^{(b-1)} + \mathbf{X}\hat{\boldsymbol{\beta}}_{c^*}, \text{ with } \mathbf{X} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_C]$$
$$\hat{\boldsymbol{\beta}}^{(b)} = \hat{\boldsymbol{\beta}}^{(b-1)} + \hat{\boldsymbol{\beta}}_{c^*}$$

4.2 Boosting ordinal BTL models

The boosting algorithm described in this section is a component-wise boosting procedure based on the *pomBoost* algorithm (Zahid and Tutz, 2013), which has been developed for the cumulative logit model. It allows for ordinal responses and can be extended to the adjacent categories logit model by modifying the link function $g = (g_1, \ldots, g_{K-1})^\top$ to obtain the adjacent categories logit model (see Section 2.2). We will use a specific feature of the *pomBoost* algorithm, which allows to split the parameters into two groups to distinguish between *obligatory* parameters that have

to be included in the model and *optional* parameters that might be of relevance and where a variable selection is applied. In our context, the object parameters $\boldsymbol{\gamma}^\top = (\gamma_1, \ldots, \gamma_{M-1})$ are considered as obligatory and the subject-specific parameters $\boldsymbol{\gamma}_p^\top = (\gamma_{1,p}, \ldots, \gamma_{M-1,p})$, for all $p = 1, \ldots, P$ are considered as optional. To be more flexible, one can divide the P optional subject-specific covariates into two disjoint sets V_{grouped} and V_{single}, with $V_{\text{grouped}} \cup V_{\text{single}} = \{1, \ldots, P\}$. This will allow selection of two different types of covariates, namely,

1. V_{grouped}: subject-specific covariates, for which all the associated subject–object interactions should be selected *simultaneously* and
2. V_{single}: subject-specific covariates, for which all associated subject–object interactions should be selected *separately*.

The reason for distinguishing these two groups is that for some subject-specific covariates, it might be interesting for the practitioner to have the possibility to select the respective subject–object interactions *simultaneously* or *separately* (we refer to the application section for an illustrating example). The subject–object interaction parameters that are associated with the pth subject-specific covariate are given by the set of parameters $\gamma_{1,p}, \ldots, \gamma_{M-1,p}$. If the subject-specific covariate is considered as being from V_{grouped}, the parameters are included as a set or left out, which yields variable selection in terms of the subject-specific covariates. However, if the subject-specific covariate is considered as being from V_{single}, only single parameters from the set are included or left out, which yields selection of subject–object interactions.

The boosting algorithm below uses the design matrix for paired comparisons $\mathbf{X} = [\mathbf{Q}, \mathbf{X}_0, \mathbf{X}_1, \ldots, \mathbf{X}_P]$, where $\mathbf{Q}$ contains the components for the thresholds, $\mathbf{X}_0$ contains the components for the objects and $\mathbf{X}_1, \ldots, \mathbf{X}_P$ contain the components for the subject-specific covariates (for more details on the multivariate structure, see Appendix A). Specifically, all matrices $\mathbf{X}_p$, with $p = 1, \ldots, P$, have $M - 1$ columns, where each column is denoted by $\mathbf{x}_{p,m}$ and reflects a single subject–object interaction between the pth subject-specific covariate and object m. Thus, these matrices have the form $\mathbf{X}_p = \left[\mathbf{x}_{p,1}, \ldots, \mathbf{x}_{p,M-1}\right]$. The algorithm proposed is the following:

BTLboost

Step 1: Initialization

Fit the intercept model $\boldsymbol{\mu}^{(0)} = h(\boldsymbol{\eta}^{(0)}) = h(\mathbf{Q}\boldsymbol{\theta})$ by maximizing the likelihood in order to obtain estimates for the threshold parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)^\top$. Initialize

$$\hat{\boldsymbol{\beta}}^{(0)} = (\hat{\boldsymbol{\theta}}^\top, \hat{\boldsymbol{\gamma}}^\top, \hat{\boldsymbol{\gamma}}_1^\top, \ldots, \hat{\boldsymbol{\gamma}}_P^\top) = (\hat{\theta}_1, \ldots, \hat{\theta}_q, 0, \ldots, 0)^\top \quad \text{and} \quad \hat{\boldsymbol{\eta}}^{(0)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(0)}.$$

Step 2: For each boosting iteration $b = 1, 2, \ldots, b_{\text{stop}}$

Step 2.1: Update of threshold and object parameters

1. *Estimation:* Fit the model $\boldsymbol{\mu} = h\left(\hat{\boldsymbol{\eta}}^{(b-1)} + \mathbf{Q}\boldsymbol{\theta} + \mathbf{X}_0\boldsymbol{\gamma}\right)$ by one-step Fisher scoring, where $\hat{\boldsymbol{\eta}}^{(b-1)}$ is used as an offset. One obtains the estimates $\hat{\boldsymbol{\theta}}^\top = (\hat{\theta}_1, \ldots, \hat{\theta}_q)$, $\hat{\boldsymbol{\gamma}}^\top = (\hat{\gamma}_1, \ldots, \hat{\gamma}_{M-1})$ and defines
$$\hat{\boldsymbol{\beta}}_0^\top = (\hat{\boldsymbol{\theta}}^\top, \hat{\boldsymbol{\gamma}}^\top, \hat{\boldsymbol{\gamma}}_1^\top, \ldots, \hat{\boldsymbol{\gamma}}_P^\top) = (\hat{\theta}_1, \ldots, \hat{\theta}_q, \hat{\gamma}_1, \ldots, \hat{\gamma}_{M-1}, 0, \ldots, 0).$$
2. *Update:*
$$\hat{\boldsymbol{\eta}}^{(b)} = \hat{\boldsymbol{\eta}}^{(b-1)} + \mathbf{X}\hat{\boldsymbol{\beta}}_0$$
$$\hat{\boldsymbol{\beta}}^{(b)} = \hat{\boldsymbol{\beta}}^{(b-1)} + \hat{\boldsymbol{\beta}}_0$$

Step 2.2: Update of parameters for the optional covariates

(a) *Estimation:* For each subject-specific covariate $p = 1, \ldots, P$, fit the models
$$\boldsymbol{\mu} = \begin{cases} h(\hat{\boldsymbol{\eta}}^{(b)} + \mathbf{X}_p\boldsymbol{\gamma}_p) & \text{if } p \in V_{\text{grouped}} \\ h(\hat{\boldsymbol{\eta}}^{(b)} + \mathbf{x}_{p,m}\gamma_{m,p}), \quad \forall m \in \{1, \ldots, M-1\} & \text{if } p \in V_{\text{single}} \end{cases},$$
where $\hat{\boldsymbol{\eta}}^{(b)}$ is used as offset. The estimated parameters $\hat{\boldsymbol{\gamma}}_p$ and $\hat{\gamma}_{m,p}$ are obtained by one-step Fisher scoring.

(b) *Selection:* From the models fitted in (a), choose the model that maximally improves the fit with respect to some criteria (e.g., deviance, AIC or BIC) and set $\mathbf{X}_{\text{best}} = \mathbf{x}_{p^*,m^*}$ if the (p^*, m^*)-th subject–object interaction yields the best fit or $\mathbf{X}_{\text{best}} = \mathbf{X}_{p^*}$ if the set of subject–object interactions associated with the p^*-th subject-specific covariate yields the best fit. This yields the estimated parameter vector
$$\hat{\boldsymbol{\beta}}_{\text{best}}^\top = (\hat{\boldsymbol{\theta}}^\top, \hat{\boldsymbol{\gamma}}^\top, \hat{\boldsymbol{\gamma}}_1^\top, \ldots, \hat{\boldsymbol{\gamma}}_P^\top) = \begin{cases} (0, \ldots, 0, \hat{\gamma}_{p^*,m^*}, 0, \ldots, 0) & \text{if } \mathbf{X}_{\text{best}} = \mathbf{x}_{p^*,m^*} \\ (0, \ldots, 0, \hat{\boldsymbol{\gamma}}_{p^*}^\top, 0, \ldots, 0) & \text{if } \mathbf{X}_{\text{best}} = \mathbf{X}_{p^*} \end{cases}.$$

(c) *Update:*
$$\hat{\boldsymbol{\eta}}^{(b)} = \hat{\boldsymbol{\eta}}^{(b)} + \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{best}}$$
$$\hat{\boldsymbol{\beta}}^{(b)} = \hat{\boldsymbol{\beta}}^{(b)} + \hat{\boldsymbol{\beta}}_{\text{best}}$$

Within each boosting iteration, the algorithm switches between two stages. The first stage (*Step* 2.1) considers only the object parameters along with the threshold parameters for an update, so that these parameters will always be part of the

paired comparison model. In the second stage (*Step* 2.2), we consider the subject-specific covariates that might enter into the paired comparison model. If the associated subject–object interactions are to be selected *simultaneously* ($p \in V_{\text{grouped}}$), the full subject-specific parameter vector $\boldsymbol{\gamma}_p$ is updated. If they are to be selected *separately* ($p \in V_{\text{single}}$), only a single subject–object interaction parameter $\gamma_{p,m}$ from the full subject-specific parameter $\boldsymbol{\gamma}_p^\top = (\gamma_{1,p}, \ldots, \gamma_{M-1,p})$ is updated within each boosting iteration. This procedure is repeated until a predefined number of boosting iterations b_{stop} is reached. In the next section, we discuss how the optimal number of boosting iterations can be obtained.

4.3 Stopping criteria

The number of boosting iterations b_{stop} is the main tuning parameter in boosting. Typically, a sufficiently high number of iterations is chosen; the optimal number of iterations b^*_{stop} is then determined afterwards. Common choices for determining b^*_{stop} are either information criteria, such as, the AIC or BIC, or cross-validation. In this paper, we consider only the two information criteria mentioned earlier, which are known measures for the trade-off between goodness-of-fit and model complexity. The computation of both criteria is based on the deviance $\text{Dev}(\hat{\boldsymbol{\eta}}^{(b)})$ and the degrees of freedom $\text{df}(b)$ after b boosting iterations.

The AIC and BIC after b iterations are given by

$$\text{AIC}(b) = \text{Dev}(\hat{\boldsymbol{\eta}}^{(b)}) + 2 \cdot \text{df}(b)$$

and

$$\text{BIC}(b) = \text{Dev}(\hat{\boldsymbol{\eta}}^{(b)}) + \log(n) \cdot \text{df}(b),$$

where n is the number of observations, or, in our case, the number of comparisons that are made by all subjects, namely, $n = \binom{M}{2} \cdot I$. The penalty term, which is 2 for the AIC and $\log(n)$ for the BIC, controls the model complexity, so that, in general, higher values for the penalty term result in sparser models.

Since the boosting algorithm is an iterative procedure, the true degrees of freedom $\text{df}(b)$ after b boosting iterations are unknown and need to be approximated. In the literature, it is often suggested to use the trace of the hat matrix after b boosting iterations as an approximation for the degrees of freedom (Tutz and Binder, 2006; Bühlmann and Hothorn, 2007a). Thus, a possible approximation of the degrees of freedom $\text{df}(b)$ is given by

$$\text{df}_{\text{trace}}(b) = \text{trace}(\mathbf{H}_b),$$

where $\mathbf{H}_b$ is the approximate hat matrix after b iterations. A detailed derivation of the formula for this hat matrix can be found in Zahid and Tutz (2013).

An alternative, computationally simpler method proposed by Bühlmann and Hothorn (2007b) uses the size of the active set, which corresponds to the number

of non-zero parameters in the bth boosting iteration, in order to approximate the degrees of freedom df(b). In this case, the number of non-zero coefficients until the bth boosting iteration is given by

$$\mathrm{df}_{\mathrm{actset}}(b) = q + (M-1) + \sum_{m,p} I(\hat{\gamma}_{m,p}^{(b)} \neq 0), \tag{4.1}$$

where q is the number of threshold parameters, $M-1$ is the number of object parameters, $\hat{\gamma}_{m,p}^{(b)}$ are all non-zero subject–object parameters of the bth iteration, and $I(\cdot)$ is the indicator function, such that $I(\cdot) = 1$ if the expression $(\cdot)$ is true, else $I(\cdot) = 0$. To limit the computational burden, we use this more convenient method to approximate the degrees of freedom in the simulation and application section.

The optimal stopping iteration b^*_{stop} is chosen from among the iterations $b = 1, \ldots, b_{\mathrm{stop}}$ as the one yielding the best (=lowest) AIC or BIC, respectively. Thus, one has to compute the AIC or BIC for all iterations. The optimal stopping iteration is then determined afterwards using

$$b^*_{\mathrm{stop}} = b \quad \Leftrightarrow \quad \mathrm{AIC}(b^*_{\mathrm{stop}}) = \min_{b=1,\ldots,b_{\mathrm{stop}}} \mathrm{AIC}(b),$$

when the AIC is chosen as stopping criterion or

$$b^*_{\mathrm{stop}} = b \quad \Leftrightarrow \quad \mathrm{BIC}(b^*_{\mathrm{stop}}) = \min_{b=1,\ldots,b_{\mathrm{stop}}} \mathrm{BIC}(b),$$

when the BIC is chosen as stopping criterion, respectively.

5 Simulation

5.1 Simulation set-up

We investigate the performance of the boosting algorithm for different simulation settings with 100 simulations for each setting. For the simulated data, we generate 20 subject-specific covariates denoted by $p = 1, \ldots, 20$ from the following distributions $X_1, \ldots, X_{10} \sim B(1, 0.5)$ and $X_{11}, \ldots, X_{20} \sim N(0, 1)$ and use $M = 6$ objects that have to be compared by $I = 200$ subjects. We use $K = 3$ response categories, so that each simulated subject has the possibility to prefer the first object, neither of the objects or the second object. In all settings, the ordinal response is computed by assuming an underlying cumulative BTL model with the linear predictor

$$\eta_k^{(r,s),i} = \theta_k + \sum_{m=1}^{M-1} x_m^{(r,s)} \gamma_m + \sum_{p=1}^{P} \sum_{m=1}^{M-1} x_{i,p} x_m^{(r,s)} \gamma_{m,p}, \quad k = 1, 2,$$

using the threshold parameters $\theta_1 = -\theta_2 = -0.8$ and the object parameters $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{M-1})^\top = (1.5, 1.1, 0.7, -0.7, -1)^\top$. The settings are organized as follows.

Setting 1: Only the subject-specific covariates $p = 1, 2, 11, 12$ are influential (parameter values can be obtained from Table 1), while all other subject-specific covariates are non-influential. For each subject-specific covariate, the associated subject–object interactions are selected simultaneously, that is, $V_{\text{grouped}} = \{1, \ldots, 20\}$ and $V_{\text{single}} = \emptyset$.

Table 1 Influential subject–object interaction parameters.

$\gamma_{m,p}$	$p = 1$	$p = 2$	$p = 11$	$p = 12$	$p = 3$	$p = 4$	$p = 13$	$p = 14$
$m = 1$	0.56	−0.39	−0.48	0.49	−0.48	0	0	0.45
$m = 2$	−0.66	0.29	−0.44	−0.48	0	−0.48	−0.32	0
$m = 3$	−0.58	0.34	−0.43	0.44	0	0.44	−0.38	0
$m = 4$	−0.68	0.31	0.5	0.46	0.5	0.46	−0.35	0
$m = 5$	0.69	−0.22	0.43	0.3	0.43	0.3	0	0.4

Source: Authors' computation.

Setting 2: Only the subject-specific covariates $p = 3, 4, 13, 14$ are influential (parameter values can be obtained from Table 1). For each subject-specific covariate, the associated subject–object interactions are selected separately, that is, $V_{\text{grouped}} = \emptyset$ and $V_{\text{single}} = \{1, \ldots, 20\}$.

Setting 3: In this setting, we consider a miss-specified model. That is, after computing the ordinal response using the influential subject-specific covariates $p = 1, 2, 3, 4, 11, 12, 13, 14$ with the associated parameter values from Table 1, some influential subject-specific covariates were removed from the simulated data set:

(a) The covariates $p = 3, 4, 13, 14$ were removed and the boosting algorithm was applied as in Setting 1, that is, the subject–object interactions for each subject-specific covariate were selected simultaneously, so that all considered covariates are from V_{grouped}.

(b) The covariates $p = 1, 2, 11, 12$ were removed and the boosting algorithm was applied as in Setting 2, that is, the subject–object interactions for each subject-specific covariate were selected separately, so that all considered covariates are from V_{single}.

Setting 4: Same as Setting 1, except that only one subject-specific covariate, $p = 1$, is considered as influential.

Setting 5: Same as Setting 2, except that only one subject-specific covariate, $p = 3$, is considered as influential.

As variable selection criterion in each boosting iteration, we use the deviance, which often yields the fully saturated model when the stopping iteration is sufficiently high, i.e., $b_{\text{stop}} \to \infty$ (see Bühlmann and Yu, 2003). Therefore, there is a need for

determining the optimal number of iterations b^*_{stop} using a stopping criterion that chooses the best model among all iterations. In the next section, we compare the results when the AIC and BIC are used as stopping criteria. The degrees of freedoms for the computation of the AIC and BIC are approximated as in equation (4.1).

5.2 Results

To investigate the performance of the algorithm, we compute the hit rate

$$\text{HR} = \frac{\sum_{p=1}^{P}\sum_{m=1}^{M-1} I(\gamma_{m,p} \neq 0) \cdot I(\hat{\gamma}_{m,p} \neq 0)}{\sum_{p=1}^{P}\sum_{m=1}^{M-1} I(\gamma_{m,p} \neq 0)},$$

which represents the percentage of correctly identified influential subject–object interactions, and the false alarm rate

$$\text{FAR} = \frac{\sum_{p=1}^{P}\sum_{m=1}^{M-1} I(\gamma_{m,p} = 0) \cdot I(\hat{\gamma}_{m,p} \neq 0)}{\sum_{p=1}^{P}\sum_{m=1}^{M-1} I(\gamma_{m,p} = 0)},$$

which represents the percentage of non-influential subject–object interactions that are mistakenly identified as influential subject–object interactions. The closer the hit rate to 1 and the closer the false alarm rate to 0, the better, because then the model contains many influential subject–object interactions and few non-influential subject–object interactions at the same time.

Table 2 shows the averaged number of boosting iterations, as well as the averaged hit rates and false alarm rates over all 100 simulations in each setting. As the model complexity is supposed to be penalized stronger by the BIC, one sees an earlier stopping of the boosting algorithm when using the BIC instead of the AIC. Except for Setting 2 and Setting 3 (b), the hit rates for the AIC and BIC are very similar and close to 1, that is, the boosting algorithm was always able to identify all (or almost all when using the BIC in Setting 5) influential covariates for Settings 1, 3 (a), 4 and 5. However, at the same time, using the BIC yields a much lower false alarm rate suggesting that using the BIC is more appropriate. In Setting 2, we have a slightly higher hit rate when the AIC is used (the difference in the hit rates is $0.9967 - 0.9592 = 0.0375$). However, the false alarm rate is almost five times higher when the AIC is used instead of the BIC (0.2572 compared to 0.0517). Using the AIC in Setting 3 (a) yields a higher hit rate, where the difference in the hit rates is $0.975 - 0.7858 = 0.1892$. At the same time, the false alarm rate is much worse (0.3860 for the AIC and 0.0615 for the BIC). Thus, the question of whether the AIC or BIC should be used depends on

Table 2 Average of the HR, the FAR and the number of boosting iterations.

Setting	Stopping criterion	Stopping iteration	HR	FAR
Setting 1	AIC	68	1.0000	0.1831
	BIC	47	1.0000	0.0119
Setting 2	AIC	103	0.9967	0.2572
	BIC	44	0.9592	0.0517
Setting 3 (a)	AIC	69	1.0000	0.3042
	BIC	40	1.0000	0.0167
Setting 3 (b)	AIC	115	0.9750	0.3860
	BIC	30	0.7858	0.0615
Setting 4	AIC	29	1.0000	0.1747
	BIC	17	1.0000	0.0200
Setting 5	AIC	39	1.0000	0.1507
	BIC	16	0.9967	0.0273

Source: Authors' computation.

whether one is interested in identifying most of the influential subject–object interactions (then the AIC might be more appropriate) or in identifying many influential subject–object interactions and few non-influential subject–object interactions at the same time (then the BIC might be more appropriate). In all settings, the false alarm rate is much higher when the AIC is used instead of the BIC. This suggests that more non-influential subject–object interactions are included in the final model determined by the AIC.

To measure the discrepancy of the parameter estimates with the true parameter values, we use the Mean squared error (MSE). Before computing the MSE for the final model, a final refitting step using the selected subject–object interaction parameters is done. Figure 1 displays the MSE for all simulations and suggests that for the selected subject–object interaction parameters the BIC performs better than the AIC because of smaller MSEs in all settings. It can also be seen that in Setting 3, where a miss-specified model is considered, the MSEs are much larger for the object parameters than for the selected subject–object interaction parameters as compared to the other settings.

Figure 2 illustrates an exemplary coefficient build-up of the estimated parameters for one out of the 100 simulations of Setting 1 and Setting 2. Within each boosting iteration of Setting 1 (left figure), the set of subject–object interaction parameters associated with a single subject-specific covariate are updated. Therefore, the norm of a single subject-specific parameter vector $||\hat{\boldsymbol{\gamma}}_p^\top||$ for each of the $p = 1, \ldots, 20$ covariates is plotted against the norm of the parameter vector containing all subject–object parameters $||(\hat{\boldsymbol{\gamma}}_1^\top, \ldots, \hat{\boldsymbol{\gamma}}_{20}^\top)^\top||$. Conversely, in Setting 2 (right figure) only the parameter of a single subject–object interaction is updated within each boosting iteration. Therefore, the estimates for a single subject–object interaction parameter $\hat{\gamma}_{m,p}$ are plotted against the norm of the parameter vector containing all subject–object parameters $||(\hat{\boldsymbol{\gamma}}_1^\top, \ldots, \hat{\boldsymbol{\gamma}}_{20}^\top)^\top||$. The coefficient paths show that the non-influential subject–object

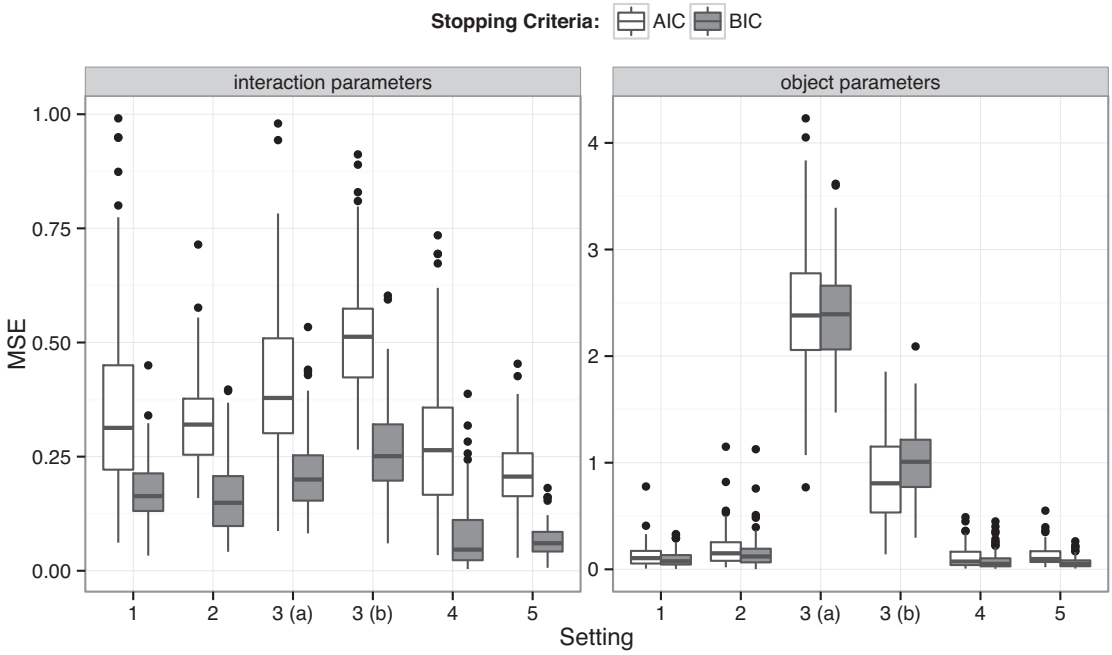

Figure 1 Box plots of MSEs for object and subject–object interaction parameters.
Source: Authors' computation.

interaction parameters (dashed gray lines) enter the model in higher boosting iterations and their parameter estimates are small because they are either selected and updated in higher boosting iterations, or they are updated only a few number of times. Since the AIC determines a higher optimal boosting iteration than the BIC, the figure confirms that the final model based on the AIC criterion contains more non-influential subject–object interactions than the final model based on the BIC criterion.

6 CEMS data

To illustrate how the variable selection method works, we use the Community of European management schools (CEMS) data from Dittrich *et al.* (1998), which was collected in a survey of 303 students of the Vienna University of Economics. The aim of the study was to investigate the preferences of students for studying at least one semester abroad in one of six different universities (London, Paris, Barcelona, St.Gall, Milan and Stockholm) and to establish an overall ranking of these universities. For each of the $\binom{M}{2} = \binom{6}{2} = 15$ comparisons of universities, the students could either prefer the first university, none of both universities or the second university. Additionally, the data contains $P = 8$ subject-specific different characteristics of the students (subject-specific covariates). An overview is given in Table 3.

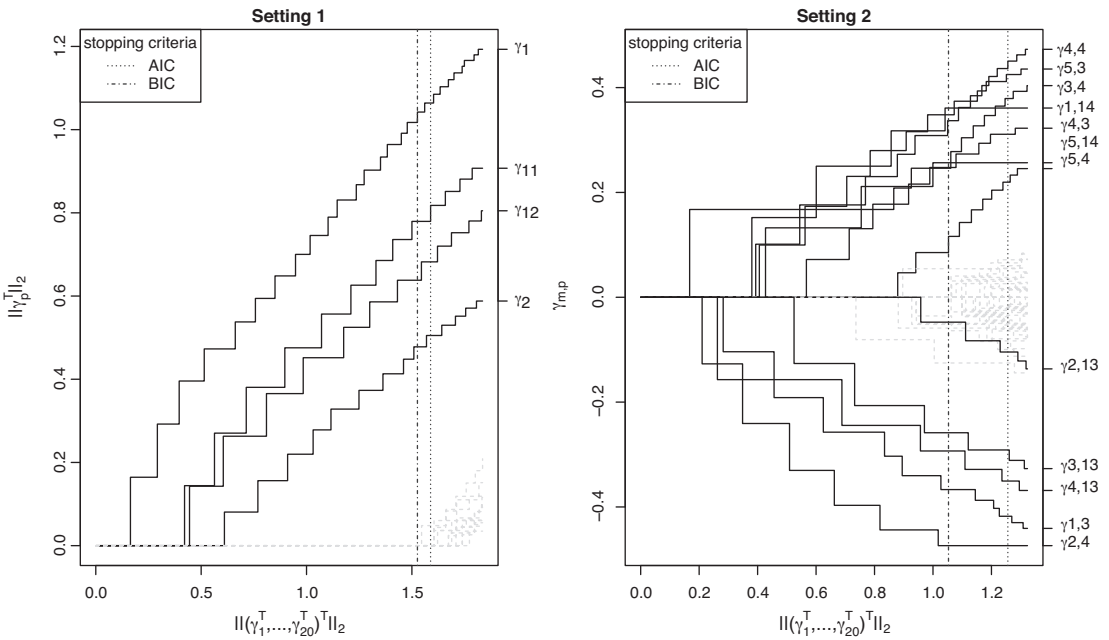

Figure 2 Exemplary coefficient build-ups for Setting 1 and Setting 2. The dashed gray lines are the paths for the non-influential subject-specific parameters (left figure) and for the non-influential subject–object interactions (right figure). The vertical lines indicate the optimal boosting iteration when either the AIC or the BIC is used as stopping criterion.
Source: Authors' computation.

A model that includes all these subject-specific covariates has a total number of 40 subject–object interaction parameters. To identify only influential subject-specific covariates, we apply the *BTLboost* algorithm with two different settings.

In the first setting, we use $V_{\text{grouped}} = \{\text{STUD}, \text{WOR}, \text{DEG}, \text{SEX}\}$ and $V_{\text{single}} = \{\text{ENG}, \text{FRA}, \text{SPA}, \text{ITA}\}$. Therefore, subject-specific covariates containing information about the knowledge of a specific language are included in the set V_{single} and are

Table 3 Description of subject-specific covariates.

Covariate	Description	Coding
STUD	Main discipline of study	0 = other, 1 = commerce
ENG	Knowledge of English	0 = good, 1 = poor
FRA	Knowledge of French	0 = good, 1 = poor
SPA	Knowledge of Spanish	0 = good, 1 = poor
ITA	Knowledge of Italian	0 = good, 1 = poor
WOR	Full-time employment while studying	0 = no, 1 = yes
DEG	Intention to take an international degree	0 = no, 1 = yes
SEX	Gender	0 = female, 1 = male

Source: Authors' computation.

selected interaction-wise, so that the final model does not necessarily contain all subject–object interactions that are associated to a specific language. In contrast to this, the set V_{grouped} contains subject-specific covariates that are selected along with all associated subject–object interactions. This distinction was chosen because language-specific covariates may interact stronger with one university. For example, students with poor knowledge of Italian may have a lower tendency to prefer the university of Milan.

In the second setting, we use $V_{\text{single}} = \{\text{STUD, WOR, DEG, SEX, ENG, FRA, SPA, ITA}\}$ and $V_{\text{grouped}} = \emptyset$. Thus, for each subject-specific covariate only the associated subject–object interactions are considered. For both settings, we apply the *BTLboost* algorithm to the cumulative BTL model using the AIC and BIC as stopping criteria. The estimated parameters after a final refitting step with bootstrapped standard errors can be found in Table 4 and Table 5. Because the BIC performed better in the simulation study, both tables show only the estimated parameters when using the BIC as stopping criterion. The optimal number of iterations when using the BIC was 34 in the first setting and 30 in the second setting.

The results show that the ranking of objects (here universities), which is based on ordering the estimates for $\hat{\boldsymbol{\gamma}}$, is the same for both settings. Milan has the greatest estimate (e.g., $\hat{\gamma}_{\text{Milan}} = 1.978$ from Table 4) and thus the most preferred university is Milan, followed by London, Barcelona, Paris, St.Gall and the reference university Stockholm, for which $\gamma_{\text{Stockholm}} = 0$. However, the ranking changes for different values of the subject-specific covariates. For example, if we consider students who are employed full-time while studying (WOR = 1), the ranking is based on $\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{\gamma}}_{\text{WOR}}$ yielding a different university ranking for those students.

The estimates $\hat{\gamma}_{\text{London,ENG}}$, $\hat{\gamma}_{\text{Paris,FRA}}$, $\hat{\gamma}_{\text{Milan,ITA}}$ and $\hat{\gamma}_{\text{Barcelona,SPA}}$ are all negative valued, and therefore they indicate that students with poor knowledge of English, French, Italian and Spanish have a lower tendency to prefer the universities in London, Paris, Milan and Barcelona, respectively.

Table 4 Estimates for selected subject–object interaction and object parameters after a final refitting step for Setting 1. The figures in brackets reflect standard error estimates based on 1000 bootstrapped samples.

Setting 1 (DEG, SEX, STUD, WOR grouped)									
	$\hat{\boldsymbol{\gamma}}_{\text{DEG}}$	$\hat{\boldsymbol{\gamma}}_{\text{SEX}}$	$\hat{\boldsymbol{\gamma}}_{\text{STUD}}$	$\hat{\boldsymbol{\gamma}}_{\text{WOR}}$	$\hat{\boldsymbol{\gamma}}_{\text{ITA}}$	$\hat{\boldsymbol{\gamma}}_{\text{ENG}}$	$\hat{\boldsymbol{\gamma}}_{\text{SPA}}$	$\hat{\boldsymbol{\gamma}}_{\text{FRA}}$	$\hat{\boldsymbol{\gamma}}$
Milano	−0.076	**−0.447**	0.033	1.141	−1.613	0	0	0	1.978
	(0.147)	(0.139)	(0.152)	(0.39)	(0.187)	(0.064)	(0.047)	(0.076)	(0.231)
London	−0.259	−0.344	0.285	0.447	0	−0.258	0	0	1.959
	(0.161)	(0.149)	(0.166)	(0.404)	(0.051)	(0.087)	(0.029)	(0.087)	(0.261)
Barcelona	−0.11	−0.315	0.105	1.205	0	−0.366	−1.421	0.248	1.938
	(0.142)	(0.13)	(0.149)	(0.422)	(0.02)	(0.119)	(0.182)	(0.101)	(0.221)
Paris	−0.036	−0.323	**0.824**	**1.544**	0	0	0	−1.192	1.246
	(0.152)	(0.143)	(0.165)	(0.436)	(0.078)	(0.097)	(0.044)	(0.156)	(0.313)
St.Gall	**0.399**	−0.058	−0.293	0.002	0	0.178	0	0	0.494
	(0.146)	(0.137)	(0.144)	(0.383)	(0.024)	(0.094)	(0.042)	(0.071)	(0.187)
$\hat{\theta}_1 = -\hat{\theta}_2 = -0.276$ (0.013)									

Source: Authors' computation.

Table 5 Estimates for selected subject–object interactions and object parameters after a final refitting step for Setting 2. The figures in brackets reflect standard error estimates based on 1000 bootstrapped samples.

	Setting 2 (single)								
	$\hat{\gamma}_{DEG}$	$\hat{\gamma}_{SEX}$	$\hat{\gamma}_{STUD}$	$\hat{\gamma}_{WOR}$	$\hat{\gamma}_{ITA}$	$\hat{\gamma}_{ENG}$	$\hat{\gamma}_{SPA}$	$\hat{\gamma}_{FRA}$	$\hat{\gamma}$
Milano	0	−0.256	0	0.999	−1.576	0	0	0	1.832
	(0.057)	(0.088)	(0.06)	(0.228)	(0.174)	(0.044)	(0)	(0.06)	(0.195)
London	0	0	0	0	0	−0.283	0	0	1.81
	(0.055)	(0.039)	(0.041)	(0.237)	(0.031)	(0.07)	(0.009)	(0.069)	(0.215)
Barcelona	0	0	0	1.041	0	−0.366	−1.424	0.204	1.794
	(0.041)	(0.025)	(0.046)	(0.26)	(0.009)	(0.105)	(0.173)	(0.07)	(0.19)
Paris	0	0	0.742	1.356	0	0	0	−1.235	1.114
	(0.05)	(0.041)	(0.14)	(0.34)	(0.038)	(0.073)	(0.025)	(0.148)	(0.217)
St.Gall	0.493	0	−0.391	0	0	0.175	0	0	0.463
	(0.116)	(0.052)	(0.118)	(0.22)	(0.008)	(0.082)	(0.005)	(0.054)	(0.143)
$\hat{\theta}_1 = -\hat{\theta}_2 = -0.275$ (0.013)									

Source: Authors' computation.

The main difference between Setting 1 and Setting 2 is that the subject–object interactions associated with the subject-specific covariates STUD, WOR, DEG and SEX were selected simultaneously in Setting 1 and separately in Setting 2. Nevertheless, all subject–object interactions that are highlighted in boldface in Table 4 were also identified in Setting 2. These subject–object interactions are the ones with the largest absolute value within the associated subject-specific parameter. Thus, the model from Setting 2 was able to identify the subject–object interactions from Setting 1 with the strongest effect.

7 Concluding remarks

Boosting is a technique that addresses the estimation problems in high-dimensional settings and provides variable selection when using component-wise boosting. In this article, we proposed a new estimation procedure for ordinal BTL models based on this technique. The simulation results showed that the method performs well concerning the identification of influential covariates. In the application section, the selected subject–object interactions are similar to those from the final model of Dittrich *et al.* (1998), although they used a different model and a variable selection approach based on forward selection and backward elimination. In the algorithm proposed here, the variable selection is carried out during the fitting process. Thus, it can also be applied in cases where the maximum likelihood estimate does not exist, for example, when the data contains more subject-specific covariates than observations.

The model estimation for ordinal BTL models is computed with the ordBTL package (Casalicchio, 2013), which is implemented in the statistical software R (R Core Team, 2013). The package is able to fit models with any number of response categories and implements the *BTLboost* algorithm. The data used in Section 6 is also available

in the package. Other packages for model estimation, which can handle responses up to 3 categories but have no built-in variable selection procedure, are prefmod (Hatzinger and Dittrich, 2012) and BradleyTerry2 (Turner and Firth, 2012).

Acknowledgement

Financial support from LMUexcellent is gratefully acknowledged.

Appendix A: Matrix representation

The representation used here is that of a multivariate generalized linear model, which has, in terms of paired comparisons, the following basic form

$$g\left(\boldsymbol{\pi}^{(r,s),i}\right) = \boldsymbol{\eta}^{(r,s),i} = \mathbf{X}^{(r,s),i}\boldsymbol{\beta}, \tag{A.1}$$

where g is a $(K-1)$-dimensional link function, $\boldsymbol{\beta}$ is the vector of coefficients, $\boldsymbol{\pi}^{(r,s),i} = \left(\pi_1^{(r,s),i}, \ldots, \pi_K^{(r,s),i}\right)^\top$ is the vector of response probabilities, $\boldsymbol{\eta}^{(r,s),i} = \left(\eta_1^{(r,s),i}, \ldots, \eta_{K-1}^{(r,s),i}\right)^\top$ is the subject-specific linear predictor, and $\mathbf{X}^{(r,s),i}$ is the design matrix for the pair (r, s) and subject i. If the design matrix and the link function g is specified, the Fisher scoring algorithm for multivariate maximum likelihood estimation can be used to obtain the parameter estimates (see Fahrmeir and Tutz, 2001; Tutz, 2012).

Before describing the design matrix in more detail, we first let $\boldsymbol{\theta}^\top = (\theta_1, \ldots, \theta_q)$, with $q = \left\lfloor \frac{K-1}{2} \right\rfloor$, denote a vector containing all threshold parameters that have to be estimated and $\tilde{\boldsymbol{\theta}}^\top = (\theta_1, \ldots, \theta_{K-1})$ denote a vector containing all threshold parameters in the model, including those that are restricted by the symmetry constraints from equations (2.3) and (2.4). For each pair (r, s) and each subject i, we define the matrix $\mathbf{Q}^{(r,s),i}$ of dimension $(K-1) \times q$, such that

$$\mathbf{Q}^{(r,s),i}\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}} \tag{A.2}$$

satisfies the symmetry constraints and ensures that the thresholds will be symmetric. Thus, the matrix can be seen as a so-called constraint matrix (for more details, see Yee, 2010).

To obtain the general structure of $\mathbf{Q}^{(r,s),i}$, we use the null vector $\mathbf{0}_q^\top = (0, \ldots, 0)$ of length q and the matrices

$$\mathbf{I}_{q\times q} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} \text{ and } \mathbf{J}_{q\times q} = \begin{pmatrix} 0 & \cdots & 0 & -1 \\ \vdots & \iddots & -1 & 0 \\ 0 & \iddots & \iddots & \vdots \\ -1 & 0 & \cdots & 0 \end{pmatrix}.$$

The matrix $\mathbf{Q}^{(r,s),i}$ satisfying equation (A.2) is a block matrix that has the same form for each pair (r, s) and each subject i, namely,

$$\mathbf{Q}^{(r,s),i} = \begin{bmatrix} \mathbf{I}_{q\times q} \\ \mathbf{0}_q^\top \\ \mathbf{J}_{q\times q} \end{bmatrix} \text{ if } K \text{ is even and } \mathbf{Q}^{(r,s),i} = \begin{bmatrix} \mathbf{I}_{q\times q} \\ \mathbf{J}_{q\times q} \end{bmatrix} \text{ if } K \text{ is odd.}$$

The subject-specific design matrix used in equation (A.1) is given by

$$\mathbf{X}^{(r,s),i} = \Big[\mathbf{Q}^{(r,s),i}, \underbrace{\mathbf{1}_{K-1} \otimes \left(\mathbf{x}^{(r,s)}\right)^\top}_{\mathbf{X}_0^{(r,s),i}}, \underbrace{\mathbf{1}_{K-1} \otimes \left(\mathbf{x}_1^{(r,s),i}\right)^\top}_{\mathbf{X}_1^{(r,s),i}}, \ldots, \underbrace{\mathbf{1}_{K-1} \otimes \left(\mathbf{x}_P^{(r,s),i}\right)^\top}_{\mathbf{X}_P^{(r,s),i}}\Big],$$

where the vector $\mathbf{1}_{K-1} = (1, \ldots, 1)^\top$ has the length $K - 1$.

In a complete paired comparison experiment, we have $\binom{M}{2}$ comparisons for each subject, where $\binom{M}{2} = \frac{M!}{2!(M-2)!}$ denotes the binomial coefficient representing the number of all distinct pairs when comparing M different objects, namely,

$$(1, 2), (1, 3), \ldots, (r, s), \ldots, (M - 1, M), \text{ for all } r < s.$$

The complete design matrix $\mathbf{X}$ contains information about all possible comparisons made by any subject and has therefore $I \cdot \binom{M}{2}$ rows. It can be written

as a block matrix of the form

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1,2),1} \\ \vdots \\ \mathbf{X}^{(M-1,M),1} \\ \vdots \\ \mathbf{X}^{(1,2),I} \\ \vdots \\ \mathbf{X}^{(M-1,M),I} \end{bmatrix} = \left[\begin{array}{c|c|ccc} \mathbf{Q}^{(1,2),1} & \mathbf{X}_0^{(1,2),1} & \mathbf{X}_1^{(1,2),1} & \dots & \mathbf{X}_P^{(1,2),1} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{Q}^{(M-1,M),1} & \mathbf{X}_0^{(M-1,M),1} & \mathbf{X}_1^{(M-1,M)} & \dots & \mathbf{X}_P^{(M-1,M),1} \\ \vdots & \vdots & & \vdots & \\ \mathbf{Q}^{(1,2),I} & \mathbf{X}_0^{(1,2),I} & \mathbf{X}_1^{(1,2),I} & \dots & \mathbf{X}_P^{(1,2),I} \\ \vdots & \vdots & \vdots & & \vdots \\ \underbrace{\mathbf{Q}^{(M-1,M),I}}_{\mathbf{Q}} & \underbrace{\mathbf{X}_0^{(M-1,M),I}}_{\mathbf{X}_0} & \underbrace{\mathbf{X}_1^{(M-1,M),I}}_{\mathbf{X}_1} & \dots & \underbrace{\mathbf{X}_P^{(M-1,M),I}}_{\mathbf{X}_P} \end{array}\right],$$

where $\mathbf{Q}$ is the matrix for the thresholds $\boldsymbol{\theta}$, $\mathbf{X}_0$ is the matrix for the object parameters $\boldsymbol{\gamma}$ and $\mathbf{X}_1, \dots, \mathbf{X}_P$ are matrices for the subject-specific parameters $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_P$.

The linear predictor $\boldsymbol{\eta} = \left(\boldsymbol{\eta}^{(1,2),1\top}, \dots, \boldsymbol{\eta}^{(M-1,M),1\top}, \dots, \boldsymbol{\eta}^{(1,2),I\top}, \dots, \boldsymbol{\eta}^{(M-1,M),I\top}\right)^\top$ contains all comparisons made by any subject and has the form $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, with the vector $\boldsymbol{\beta} = (\boldsymbol{\theta}^\top, \boldsymbol{\gamma}^\top, \boldsymbol{\gamma}_1^\top, \dots, \boldsymbol{\gamma}_P^\top)$ containing all parameters that have to be estimated, that is, all q threshold parameters, all $M-1$ object parameters and all $M-1$ subject–object interaction parameters for each of the P subject-specific covariates.

References

Agresti A (1992) Analysis of ordinal paired comparison data. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **41**, 287–97.

Böckenholt U (2001) Hierarchical modeling of paired comparison data. *Psychological Methods*, **6**, 49.

Bradley RA and Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, **39**, 324–45.

Bühlmann P (2006) Boosting for high-dimensional linear models. *The Annals of Statistics*, **34**, 559–83.

Bühlmann P and Hothorn T (2007a) Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 477–505.

Bühlmann P and Hothorn T (2007b) Rejoinder: Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, **22**, 516–22.

Bühlmann P and Yu B (2003) Boosting with the L_2-loss: Regression and classification. *Journal of the American Statistical Association*, **98**, 324–39.

Casalicchio G (2013) *ordBTL: Modelling comparison data with ordinal response*. R package version 0.8.

Davidson RR (1970) On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, **65**, 317–28.

Dittrich R, Hatzinger R and 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: Series C (Applied Statistics)*, **47**, 511–25.

Dittrich R, Hatzinger R and Katzenbeisser W (2004) A log-linear approach for modelling ordinal paired comparison data on motives to start a PhD programme. *Statistical Modelling*, **4**, 181–93.

Dittrich R, Francis B, Hatzinger R and Katzenbeisser W (2007) A paired comparison approach for the analysis of sets of Likert-scale responses. *Statistical Modelling*, **7**, 3–28.

Fahrmeir L and Tutz G (2001) *Multivariate statistical modelling based on generalized linear models*. New York: Springer.

Francis B, Dittrich R, Hatzinger R and Penn R (2002) Analysing partial ranks by using smoothed paired comparison methods: An investigation of value orientation in Europe. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **51**, 319–36.

Freund Y (1995) Boosting a weak learning algorithm by majority. *Information and Computation*, **121**, 256–85.

Freund Y (1997) A decision-theoretic generalization of on-line learning and an application to boosting? *Journal of Computer and System Sciences*, **55**, 119–39.

Friedman J, Hastie T and Tibshirani R (2000) Special invited paper. Additive ogistic regression: A statistical view of boosting. *The Annals of Statistics*, **28**, 337–74.

Friedma, JH (2001) Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, **29**, 1189–232.

Hatzinger R and Dittrich R (2012) prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings. *Journal of Statistical Software*, **48**, 1–31.

Luce RD (1959) *Individual choice behavior a theoretical analysis*. New York: Wiley.

McCullagh P (1980) Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **42**, 109–42.

R Core Team (2013) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao PV and Kupper LL (1967) Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, **62**, 194–204.

Schapire RE (1990) The strength of weak learnability. *Machine Learning*, **5**, 197–227.

Turner H and Firth D (2012) Bradley-Terry models in R: The BradleyTerry2 package. *Journal of Statistical Software*, **48**.

Tutz G (1986) Bradley-Terry-Luce models with an ordered response. *Journal of Mathematical Psychology*, **30**, 306–16.

Tutz G (2012) *Regression for categorical data*. Cambridge University Press.

Tutz G and Binder H (2006) Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, **62**, 961–71.

Yee TW (2010) The VGAM package for categorical data analysis. *Journal of Statistical Software*, **32**, 1–34.

Zahid FM and Tutz G (2013) Proportional odds models with high-dimensional data structure. *International Statistical Review*, **81**, 388–406.