Working Paper Series, N. 8, June 2011

Department of Statistical Sciences
University of Padua
Italy

Models for paired comparison data: a review with emphasis on dependent data

Manuela Cattelan
Department of Statistical Sciences
University of Padua
Italy

Abstract: Thurstonian and Bradley-Terry models are the most commonly applied models in the analysis of paired comparison data. Since their introduction, numerous developments of those models have been proposed in different areas. This paper provides an updated overview of these extensions, including how to account for object- and subject-specific covariates and how to deal with ordinal paired comparison data. Special emphasis is given to models for dependent comparisons. Although these models are more realistic, their use is complicated by numerical difficulties. We therefore concentrate on implementation issues. In particular, a pairwise likelihood approach is explored for models for dependent paired comparison data and a simulation study is carried out to compare the performance of maximum pairwise likelihood with other methods, such as limited information estimation. The methodology is illustrated throughout using a real data set about university paired comparisons performed by students.

Keywords: Bradley-Terry model, limited information estimation, paired comparisons, pairwise likelihood, Thurstonian models.

UNIVERSITÀ
DEGLI STUDI
DI PADOVA
DIPARTIMENTO
DI SCIENZE
STATISTICHE

Final version (2011-06-14)

Contents

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Manuela Cattelan
tel: +39 049 827 4124
manuela.cattelan@unipd.it

Models for paired comparison data: a review with emphasis on dependent data

Manuela Cattelan
Department of Statistical Sciences
University of Padua
Italy

Abstract: Thurstonian and Bradley-Terry models are the most commonly applied models in the analysis of paired comparison data. Since their introduction, numerous developments of those models have been proposed in different areas. This paper provides an updated overview of these extensions, including how to account for object- and subject-specific covariates and how to deal with ordinal paired comparison data. Special emphasis is given to models for dependent comparisons. Although these models are more realistic, their use is complicated by numerical difficulties. We therefore concentrate on implementation issues. In particular, a pairwise likelihood approach is explored for models for dependent paired comparison data and a simulation study is carried out to compare the performance of maximum pairwise likelihood with other methods, such as limited information estimation. The methodology is illustrated throughout using a real data set about university paired comparisons performed by students.

Keywords: Bradley-Terry model, limited information estimation, paired comparisons, pairwise likelihood, Thurstonian models.

1 Introduction

Paired comparison data originate from the comparison of objects in couples. This type of data arises in numerous contexts, especially when the judgement of a person is involved. Indeed, it is easier for people to compare pairs of objects than ranking a list of items. There are other situations that may be regarded as comparisons from which a winner and a loser can be identified without the presence of a judge. Both these instances can be analysed by the techniques described in this paper.

Since paired comparison data occur in various fields, the literature about their analysis is spread over numerous disciplines which use different terminologies. In fact, the objects involved in the paired comparisons can be beverages, carbon typewriter ribbons, lotteries, players, moral values, physical stimuli and many more. Here, the elements that are compared are called objects or sometimes stimuli. The paired comparisons can be performed by a person, an agent, a consumer, a judge, etc., so the terms subject or judge will be employed to denote the person that makes the choice, whenever there is one.

The bibliography by Davidson and Farquhar (1976), which includes more than

350 papers related to paired comparison data, testifies the widespread interest for this type of data. This interest is still present, especially in the psychometric and the statistical literature in which various developments and extensions of models for paired comparison data have been proposed. These extensions are most often based on the Thurstone (1927) and the Bradley-Terry (Bradley and Terry, 1952) models. The paper focuses on recent extensions of the two models, especially those subsequent to the review by Bradley (1976) and the monograph by David (1988), including in particular the work that has been done in the statistical and the psychometric literature.

The two classical models for the analysis of paired comparison data are presented in Section 2, while extensions for ordinal paired comparison data are reviewed in Section 3. Section 4 surveys how explanatory variables can be included in the model. Section 5 reviews models that include dependence among the observations and outlines the inferential problems related to such an extension. Here, a pairwise likelihood approach is proposed to estimate these models and a simulation study is performed in order to compare the estimates produced by maximum likelihood, limited information estimation and pairwise likelihood. Section 6 reviews existing `R` (`R` Development Core Team, 2011) packages for the statistical analysis of paired comparison data and Section 7 concludes.

2 Linear models

Let Y_{sij} denote the random variable associated with the result of the paired comparison between object i and j, $j > i = 1, \dots, n$, made by subject $s = 1, \dots, S$ and let $\boldsymbol{Y}_s = (Y_{s\,12}, \dots, Y_{s\,n-1\,n})$ be the vector of the results of all paired comparisons made by subject s. When $s = 1$ or the difference between judges is not accounted for in the model, then the subscript s will be dropped. If all paired comparisons are performed, they number $N = n(n-1)/2$ when there is just one judge and $SN = Sn(n-1)/2$ in a multiple judgement sampling scheme, that is when all paired comparisons are made by all S subjects.

Let $\mu_i \in \mathbf{R}$, $i = 1, \dots, n$, denote the true worth of the objects. Traditional models were developed assuming only two possible outcomes of the comparisons, so Y_{ij} is a binary random variable and π_{ij}, the probability that object i is preferred to object j, depends on the difference between the worth of the two objects

$$\pi_{ij} = F(\mu_i - \mu_j), \tag{1}$$

where F is a symmetric distribution function. Such models are called linear models by David (1988). When F is the normal cumulative distribution function, formula (1) defines the Thurstone (1927) model, while if F is the logistic cumulative distribution function, then the Bradley-Terry model (Bradley and Terry, 1952) is recovered. Other specifications are possible, for example Stern (1990) suggests to use a gamma distribution. The Thurstone model is also known as the Thurstone-Mosteller model since Mosteller (1951) presented some inferential techniques for the model, while the Bradley-Terry model was independently proposed also by Zermelo (1929) and Ford (1957). Indeed, an intuitive justification of the Bradley-Terry model is that, when only two outcomes are possible, the probability that i is preferred to j is $\pi_{ij} = \pi_i/(\pi_i + \pi_j)$, where $\pi_i = \exp \mu_i$. Model (1) is called unstructured model and the aim of the analysis is to make inference on the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ of worth parameters which can be used to determine a final ranking of all the objects compared. Note that the specification of model (1) through all the pairwise differences $\mu_i - \mu_j$ implies that a constraint is needed in order to identify the parameters. Various constraints can be specified, for example the sum constraint, $\sum_{i=1}^{n} \mu_i = 0$, or the reference object constraint, $\mu_i = 0$ for one object $i \in \{1, \ldots, n\}$, can be used.

If it is of interest to make inference on contrasts between the worth parameters, for example for testing $H_0 : \mu_i = \mu_j$ by means of the test statistic $(\hat{\mu}_i - \hat{\mu}_j)/(\text{var}(\hat{\mu}_i - \hat{\mu}_j))^{1/2}$, where $\hat{\mu}_i$ is the maximum likelihood estimator of μ_i, the whole covariance matrix of the worth parameters is needed. However, it is very inconvenient to report that matrix and a useful alternative may be to report quasi-standard errors (Firth and Menezes, 2004) instead of the usual standard errors since they allow approximate inference on any of the contrasts. In fact, quasi-standard errors can be employed for making inference on the differences of worth parameters as if they were independent.

Interval estimation may present some problems, too. In fact, when the reference object constrained is employed, there is no standard error for the fixed parameter. The problem may be overcome by means of quasi-variances that allow the computation of quasi-confidence intervals for all parameters.

Example. A program supported by the European Union offers an international degree in Economics and Management. Twelve universities take part in this program, and in order to receive a degree, a student in the program must spend a semester in another university joining the program. Usually, some universities receive more preferences than others and this may cause organisational problems. A study was carried out among 303 students of the Vienna University of Economics who were asked in which university they would prefer to spend the period abroad between six universities situated in Barcelona, London, Milan, Paris, St. Gallen and Stockholm compared pairwise. This example will be used throughout the paper with an illustrative purpose. For an exhaustive analysis of the data refer to Dittrich *et al.* (1998) and Dittrich *et al.* (2001). The data set is available in both the `prefmod` (Hatzinger, 2010) and the `BradleyTerry2` (Turner and Firth, 2010a) `R` packages, see Section 6. Table 1 reports the aggregated data on the 15 paired comparisons. For example, the first row shows that in the paired comparison between London and Paris 186 students prefer London, 91 students prefer Paris and 26 students do not have a preference between the two universities. Moreover, 91 students unintentionally overlooked the comparison between Paris and Milan which has only 212 answers. Table 2 shows the estimate of the worth parameters for the six universities using the Thurstone model and adding half of the number of no preferences to each university in the paired comparison. In Section 3 a more proper way to handle no preference data will be discussed. Here, the reference object constraint is used and the worth of Stockholm is set to zero. The estimates show that Stockholm is the least preferred university, while London is the most preferred one followed by Paris, Barcelona, St. Gallen and Milan. The estimated probability that London is preferred to Paris is $\Phi(0.982 - 0.561) = 0.66$, where Φ denotes the cumulative distribution function of a standard normal random variable. If it is of interest to test whether

Table 1: Universities paired comparison data. `1` and `2` refer to the number of choices in favour of the university in the fist and the second column, respectively, while `X` denotes the number of no preferences expressed.

		1	X	2
London	Paris	186	26	91
London	Milan	221	26	56
Paris	Milan	121	32	59
London	St. Gallen	208	22	73
Paris	St. Gallen	165	19	119
Milan	St. Gallen	135	28	140
London	Barcelona	217	19	67
Paris	Barcelona	157	37	109
Milan	Barcelona	104	67	132
St. Gallen	Barcelona	144	25	134
London	Stockholm	250	19	34
Paris	Stockholm	203	30	70
Milan	Stockholm	157	46	100
St. Gallen	Stockholm	155	50	98
Barcelona	Stockholm	172	41	90

the worth of St. Gallen is significantly higher than the worth of Milan, the standard error of the difference between these two worth parameters can be approximated by means of the quasi-standard errors as $(0.030^2 + 0.031^2)^{1/2} = 0.043$. The value of the test statistic is $(0.325 - 0.240)/0.043 = 1.98$, which yields a p-value of 0.02, hence the hypothesis of equal worth parameters between St. Gallen and Milan is not supported by the data.

Table 2: Estimates (`Est.`), standard errors (`S.E.`) and quasi-standard errors (`Q.S.E.`) of the universities worth parameters employing a Thurstone model.

	Est.	S.E.	Q.S.E.
Barcelona	0.333	0.043	0.030
London	0.982	0.045	0.033
Milan	0.240	0.044	0.031
Paris	0.561	0.044	0.031
St. Gallen	0.325	0.043	0.030
Stockholm	0	-	0.031

2.1 Applications

There are many different areas in which paired comparison data arise. Here, only the most recent applications are considered, further references can be found in Bradley (1976), Davidson and Farquhar (1976) and David (1988).

When human perceptions are involved, it is easier to perform paired comparisons than ranking all the objects at once. Ellermeir *et al.* (2004) and Choiser and Wickelmaier (2007) analyse pairwise evaluations of sounds, while Bäuml (1994) and Kissler and Bäuml (2000) present applications involving facial attractiveness. Duineveld *et al.* (2000) employ the Bradley-Terry model to analyse consumer preference data on orange soft drinks, while Francis *et al.* (2002) transform partial rank data into paired comparison data to study the value orientation of people in different European countries. In Mazzucchi *et al.* (2008) the Bradley-Terry model is applied to a reliability problem. A panel of wiring experts is asked to state which is the riskier one between different scenarios compared pairwise in order to determine the probability of wire failure as a function of influencing factors in an aircraft environment. Jerome *et al.* (2009) show that a questionnaire based on paired comparison data is a suitable methodology to capture user voice as regards sexual health services. Finally, Maydeu-Olivares and Böckenholt (2008) list 10 reasons to use Thurstone's method for modelling subjective health outcomes including the ease for respondents, the existence of extensions for modelling inconsistent choices and for including covariates and the possibility to discover the determinants of the valuations.

There are also many instances in which paired comparisons arise even without the presence of a judge, as in sport data. For example, Joe (1990) and Henery (1992) employ the Bradley-Terry model and the Thurstone model, respectively, to rank chess players. Applications to tennis data are shown in Agresti (2002) and McHale and Morton (2011) while dynamic extensions for this type of data have been proposed by Barry and Hartigan (1993), Fahrmeir and Tutz (1994), Knorr-Held (2000) and Cattelan *et al.* (2010) whereas an extension for continuous data is presented in Stern (2011). Stigler (1994) uses the Bradley-Terry model for ranking scientific journals and the same model is exploited in genetics by Sham and Curtis (1995). Many applications of the Bradley-Terry model can be found also in zoological data in order to investigate aspects of animal behaviour (Stuart-Fox *et al.*, 2006; Whiting *et al.*, 2006; Head *et al.*, 2008).

3 Ordinal paired comparisons

Traditional models for the analysis of paired comparison data introduced in Section 2 where developed assuming only two outcomes. Extensions for the case of three possible outcomes, in which a no preference judgement can be expressed or a tie can occur are proposed by Glenn and David (1960) for the Thurstone-Mosteller model and by Rao and Kupper (1967) and Davidson (1970) for the Bradley-Terry model. Subsequently, the situation in which subjects are requested to express a degree of preference has been considered (Agresti, 1992). Suppose that objects i and j are compared and the subject can express strong preference for i over j, mild preference for i, no preference, mild preference for j over i or strong preference for j. If H

denotes the number of grades of the scale, then in this example $H = 5$.

Let $Y_{ij} = 1, \ldots, H$, where 1 denotes the least favourable response for i and H is the most favourable response for i. Agresti (1992) shows how cumulative link and the adjacent categories models for the analysis of ordinal data can be adapted to ordinal paired comparison data. The cumulative link models exploit the latent random variable representation. Let Z_{ij} be an underlying continuous random variable and let $\tau_1 < \tau_2 < \ldots < \tau_{H-1}$ denote thresholds such that $Y_{ij} = h$ when $\tau_{h-1} < Z_{ij} \leq \tau_h$. Then,

$$\mathrm{pr}(Y_{ij} \leq y_{ij}) = F(\tau_{y_{ij}} - \mu_i + \mu_j), \tag{2}$$

where $-\infty = \tau_0 < \tau_1 < \ldots < \tau_{H-1} < \tau_H = \infty$ and F is the cumulative distribution function of the latent variable Z_{ij}. F can be either the logistic or the normal distribution function leading to the cumulative logit or the cumulative probit model, respectively. The symmetry of the model imposes that $\tau_h = -\tau_{H-h}$, $h = 1, \ldots, H$ and $\tau_{H/2} = 0$ when H is even. When $H = 3$ there are two threshold parameters τ_1 and τ_2 such that $\tau_1 = -\tau_2$ and model (2) corresponds to the extension of the Bradley-Terry model introduced by Rao and Kupper (1967) when a logit link is considered and the extension of the Thurstone model by Glenn and David (1960) when the probit link is employed.

An alternative model proposed by Agresti (1992) is the adjacent categories model. In this case the link is applied to adjacent response probabilities, rather than cumulative probabilities:

$$\log \left[\frac{\mathrm{pr}(Y_{ij} = h)}{\mathrm{pr}(Y_{ij} = h + 1)} \right] = \tau_h - \mu_i + \mu_j, \tag{3}$$

where the same symmetry constraints of the threshold parameters as described for the cumulative link models must be satisfied. Model (3) reduces to the Bradley-Terry model when only 2 categories are allowed and to the model proposed by Davidson (1970) when 3 categories are allowed. According to Agresti (1992), model (3) is simpler to interpret than cumulative link models since the odds ratio refers to a given outcome instead of referring to groupings of outcomes. The adjacent categories model, as well as the Bradley-Terry model, have also a log-linear representation (Dittrich *et al.*, 2004).

An application of the adjacent categories model to market data is illustrated in Böckenholt and Dillon (1997b). In the experiment consumers are asked to allocate 11 chips between two products of different brands. The number of chips allocated to each product can be analysed by means of models for ordinal paired comparison data. Böckenholt and Dillon (1997a) discuss the problem of the bias that may be present in the data due to the different allocation procedures adopted by different people. In fact, people may tend to prefer one of the two products either strongly or very mildly, thus using actually only a subset of all the possible categories. Hence, judges may not only have different preferences for the objects, but also vary in the way they use the response scale.

Example. In the paired comparisons of universities, students were allowed to express no preference between two universities. Therefore, the data should be analysed by means of a model for ordinal data. Table 3 shows the estimates of a

Table 3: Estimates (`Est.`), standard errors (`S.E.`) and quasi-standard errors (`Q.S.E.`) of the universities worth parameters employing a cumulative extension of the Thurstone model.

	Est.	S.E.	Q.S.E.
Barcelona	0.332	0.041	0.028
London	0.998	0.043	0.031
Milan	0.241	0.041	0.029
Paris	0.566	0.042	0.030
St. Gallen	0.324	0.040	0.028
Stockholm	0	-	0.029
τ_2	0.153	0.007	-

cumulative probit extension of the Thurstone model for the university data. In this particular case, the estimates of the worth parameters and their standard errors are very similar to those of Table 2 and the ranking of universities remains the same, but in general, especially when the number of no preferences is large, results can be different. Moreover, in this case we can estimate the probability of no preference between London and Paris which is $\Phi(0.153 - 0.998 + 0.566) - \Phi(-0.153 - 0.998 + 0.566) = 0.11$ and the estimated probability that London is preferred to Paris reduces to $1 - \Phi(0.153 - 0.998 + 0.566) = 0.61$, hence the estimated probability that Paris is preferred to London is 0.28.

4 Explanatory variables

It is often of interest to analyse whether some explanatory variables have an impact on the results of the paired comparisons. Besides the characteristics of the objects, some subject-specific covariates may influence the result of the comparisons.

4.1 Object-specific covariates

In the statistical literature, there has been an early interest in the analysis of object-specific covariates and the detection of which features may influence the outcome of a comparison. Let $\boldsymbol{x}_i = (x_{i1}, \dots, x_{iP})'$ be a vector of P explanatory variables related to object i and $\boldsymbol{\beta}$ be a P-dimensional parameter vector. Then, in the context of the Bradley-Terry model, Springall (1973) proposes to describe the worth parameters as the linear combination

$$\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}, \qquad i = 1, \dots, n. \tag{4}$$

A paired comparisons model with explanatory variables is called structured model. The same extension can be applied to Thurstonian models. Note that since only the differences $\mu_i - \mu_j = (\boldsymbol{x}_i - \boldsymbol{x}_j)'\boldsymbol{\beta}$ enter the linear predictor, an intercept cannot be identified.

Dittrich *et al.* (1998) consider the log-linear representation of the Bradley-Terry model introduced by Sinclair (1982) in case multiple paired comparisons are made by multiple subjects and show how object-specific covariates can be included in this type of specification.

Spline representations for the covariates in the context of paired comparison data are proposed by De Soete and Winsberg (1993). A Thurstone model is considered in which for each object a small set of physical measurements is available and the case in which $\mu_i = f(\boldsymbol{x}_i)$, where f is a P-variate function, is discussed. De Soete and Winsberg (1993) analyse both the case of additive univariate spline models, in which there is a separate spline transformation g for each dimension, i.e. $f(\boldsymbol{x}_i) = \sum_{p=1}^{P} g_p(x_{ip})$, and the case of a multivariate spline model in which $f(\boldsymbol{x}_i)$ is assumed to be a general multivariate spline function of $\boldsymbol{x}_i$. However, the Authors conclude that large data sets may be necessary to estimate nonlinearities reliably.

Example. It is of interest to check whether some particular features of the universities influence the preferences of students. The universities specialise in different subjects, specifically the two universities in London and Milan specialise in economics, those in Paris and Barcelona specialise in management science and the remaining two in finance. Probably, some subjects are more popular and universities that specialise in those subjects will receive more preferences. Another element that may influence the choice is the location of the university, on this respect universities can be divided into two groups: universities in a Latin country (Italy, France or Spain) and universities in other countries. Table 4 shows the estimates of a Thurstonian model which includes covariates as described in formula (4). The reference category is a university that specialises in finance and is not in a Latin country. It is evident that universities that specialise in management are preferred, followed by those that specialise in economics. Finally, universities in northern countries are preferred to universities in Latin countries. Consider the universities in London and Paris. The former specialises in economics and is not in a Latin country while the latter specialises in management science and is in a Latin country. In this structured model, the estimated probability that London is preferred to Paris is $1 - \Phi(0.150 - (0.827 - 1.022 + 0.743)) = 0.66$, the estimated probability of no preference is $\Phi(0.150 - (0.827 - 1.022 + 0.743)) - \Phi(-0.150 - (0.827 - 1.022 + 0.743)) = 0.10$ and the estimated probability of a loss for London is 0.24.

Table 4: Estimates (`Est.`) and standard errors (`S.E.`) of a Thurstonian model for university data including object-specific covariates.

	Est.	S.E.
Economics	0.827	0.038
Management	1.022	0.052
Latin country	-0.743	0.043
τ_2	0.150	0.007

4.2 Subject-specific covariates

The results of the comparisons can be influenced also by characteristics of the subject that performs the paired comparison, hence it may be of interest to take into account the individual differences in order to understand how they affect the preference for an item over another.

In the log-linear representation of the Bradley-Terry model, Dittrich *et al.* (1998) show how to include categorical subject specific covariates, while Francis *et al.* (2002) tackle the problem of continuous subject-specific covariates.

When modelling the data, subject-specific covariates can be included in different ways. For example, Dillon *et al.* (1993) consider a marketing application in which consumers are divided in latent classes. Consumers inside the same class share the same worth parameters for the objects, while subjects in different classes have different worth parameters. In this case the probability of belonging to a certain class is a function of the subject explanatory variables, while subjects belonging to the same class have preferences for objects that follow an unstructured Bradley-Terry model.

A semiparametric approach which accounts for subject-specific covariates is proposed by Strobl *et al.* (2011) who suggest a methodology to partition recursively the subjects that perform the paired comparisons on the basis of their covariates. After the subjects have been split, an unstructured Bradley-Terry model is estimated for each of the homogeneous subsamples. This semiparametric method does not require to specify a functional form for the covariates and allows to identify groups of subjects with covariates that are structurally different for which different sets of preferences are recovered.

Example. Some features of the students that performed the universities paired comparisons were collected. In particular, it is known whether students have good knowledge of English, Italian, Spanish and French and which is the main topic of their studies. It is conceivable that, for example, students with a good knowledge of Spanish are more inclined to prefer the university in Barcelona. It is also reasonable to expect that students whose main discipline of study is commerce may prefer a university that specialises in management. Table 5 shows the estimates of a model including some subject-specific covariates. The notation `French:Paris` means that the model includes a dummy variable which is equal to 1 only when subjects with a good knowledge of French make a paired comparison including the university in Paris. The good knowledge of a foreign language induces students to choose the university situated in the country where that foreign language is spoken. Consider a student with a good knowledge of both English and French and whose main discipline of study is management, then the estimated probability that this student prefers London to Paris is $1-\Phi(0.160-(0.141+0.757-0.652-0.789+0.835-0.238)) = 0.46$ while the estimated probabilities of no preference and preference for Paris are 0.13 and 0.41, respectively. If this student's main discipline of study was not management, which is the subject in which Paris specialises, then the above estimated probabilities of preferring London, no preference and preferring Paris would become 0.55, 0.12 and 0.33, respectively.

The same data set is studied in Strobl *et al.* (2011) as an application for their

Table 5: Estimates (`Est.`) and standard errors (`S.E.`) of universities data with subject- and object-specific covariates.

	Est.	S.E.
Economics	0.757	0.066
Management	0.789	0.080
Latin country	-0.835	0.071
Discipline:Management	0.238	0.054
English:London	0.141	0.075
French:Paris	0.652	0.049
Italian:Milan	1.004	0.094
Spanish:Barcelona	0.831	0.095
τ_2	0.160	0.007

recursive partitioning methodology. Further subject-specific explanatory variables are available: the gender, the indicator of whether a student works full time and the aim for an international degree. Strobl *et al.* (2011) consider all the available covariates and find that students can be divided into subsamples sharing the same worth parameters only on the basis of the knowledge of Italian, Spanish and French and the main discipline of study.

5 Models for dependent data

The models presented so far are estimated assuming independence among all observations. The inclusion of a dependence structure is not only more realistic, but also has an impact on the transitivity properties of the model. Intransitive choices occur when object i is preferred to j and object j is preferred to k, but in the paired comparison between i and k, the latter is preferred. These are also called circular triads. Paired comparison models can present different transitivity properties. Assume that $\pi_{ij} \geq 0.5$ and $\pi_{jk} \geq 0.5$, then a model satisfies

- weak stochastic transitivity if $\pi_{ik} \geq 0.5$;
- moderate stochastic transitivity if $\pi_{ik} \geq \min(\pi_{ij}, \pi_{jk})$;
- strong stochastic transitivity if $\pi_{ik} \geq \max(\pi_{ij}, \pi_{jk})$.

The Bradley-Terry and Thurstone model as presented so far satisfy strong stochastic transitivity. This property may be desirable sometimes, for example when asking wiring experts which is the riskier situation between different scenarios in an aircraft environment. In this case it is desirable that choices are consistent, so Mazzucchi *et al.* (2008) use transitivity to check the level of reliability of experts. However, in some situations choices can be systematically intransitive, for example when the same objects have more than one aspect of interest and different aspects prevail in different comparisons.

Causeur and Husson (2005) propose a 2-dimensional Bradley-Terry model in which the worth parameter of each object is bidimensional and can thus be represented on a plane. In fact, the traditional Bradley-Terry model provides a linear score for all the objects compared and hence may not be appropriate when worth are not transitively related, while a bidimensional worth parameter gives more insights into the relations among objects. A further multidimensional extension is proposed by Usami (2010). However, this methodology does not provide a final ranking of all objects.

A different method that allows the inclusion in the model even of systematic intransitive comparisons while yielding a ranking of all the objects consists in modelling the dependence structure among comparisons. The development of inferential techniques for dependent data has recently allowed an investigation of models for dependent observations.

5.1 Multiple judgement sampling

The assumption of independence is questioned in the case of the multiple judgement sampling, that is when S people make all the N paired comparisons. It seems more realistic to assume that the comparisons made by the same person are dependent. This aspect has received much attention in the literature during the last decade.

5.1.1 Thurstonian models

The original model proposed by Thurstone (1927) includes correlation among the observations. The model was developed for analysing sensorial discrimination and assumes that the stimuli $\boldsymbol{T} = (T_1, \dots, T_n)'$ compared in a paired comparison experiment follow a normal distribution with mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ and variance $\boldsymbol{\Sigma}_T$, $\boldsymbol{T} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}_T)$. Hence, the single realisation t_i of the stimulus T_i can vary and the result of the paired comparison between the same two stimuli can be different in different occasions. Assume that only either a preference for i or a preference for j can be expressed, then in a paired comparison when $T_i > T_j$ object i is preferred, or alternatively, when the latent random variable $Z_{ij} = T_i - T_j$ is positive, a win for i is observed, otherwise a win for j occurs. The most general model assumes an unrestricted, non-diagonal $\boldsymbol{\Sigma}_T$. However, this complicates noticeably the estimation procedures and in many applications estimation is carried out assuming that the observations are independent.

Psychometricians are interested in understanding the relations between stimuli, hence they are primarily interested in the unstructured and unrestricted Thurstone model which is the model without any constraint on the covariance matrix but those necessary for identification. Indeed, the original model is over-parametrised and some restrictions are needed. Thurstone (1927) proposes different specifications of the covariance matrix $\boldsymbol{\Sigma}_T$, among them the Case III, which assumes that $\boldsymbol{\Sigma}_T = \mathrm{diag}(\sigma_1^2, \dots, \sigma_n^2)$, and Case V, which assumes a homogeneous $\boldsymbol{\Sigma}_T = \sigma^2 \mathbf{I}_n$, where $\mathbf{I}_n$ denotes the identity matrix of dimension n. However, it is not possible to distinguish between Case V and an unrestricted $\boldsymbol{\Sigma}_T$ since every positive definite matrix of the form $\boldsymbol{\Sigma}_T + \boldsymbol{d}\mathbf{1}' + \mathbf{1}\boldsymbol{d}'$ where $\mathbf{1}$ is a vector of n ones and $\boldsymbol{d}$ is an n-dimensional vector of constants, produces the same probability of the outcomes (Tsai, 2000). This means that it is not possible to distinguish, in terms of likelihood, between a model with 0 correlation between all stimuli and a model with correlation d between all stimuli. The advantage is that while Case V satisfies strong stochastic transitivity, an unrestricted Thurstonian model satisfies only moderate stochastic transitivity (Takane, 1989).

Subsequent developments of the original model by Thurstone first address the issue of accounting for within and between judges variability (Takane, 1989; Böckenholt and Tsai, 2001). Then, an extension that allows for different worth of the items in different comparisons and systematic intransitive behaviour, which consists in choices by judges that systematically produce circular triads, is proposed by Tsai and Böckenholt (2006) and Tsai and Böckenholt (2008). Note that in these models the evaluations made by different judges are assumed independent while those made by the same person are correlated.

Takane (1989) extends the Thurstone model including a vector of pair specific errors, which seems appropriate in the context of multiple judgement sampling. Let $\boldsymbol{Z}_s = (Z_{s\,12}, \dots, Z_{s\,n-1\,n})'$ be the vector of all latent continuous random variables pertaining to subject s, then

$$\boldsymbol{Z}_s = \boldsymbol{A}\boldsymbol{T} + \boldsymbol{e}_s, \tag{5}$$

where $\boldsymbol{e}_s = (e_{s12}, e_{s13}, \dots, e_{sn-1\,n})'$ is the vector of pair-specific errors which has zero mean, covariance $\boldsymbol{\Omega}$ and is independent of $\boldsymbol{T}$ and of $\boldsymbol{e}_{s'}$ for another subject $s' \neq s$ and $\boldsymbol{A}$ is the design matrix of paired comparisons whose rows identify the paired comparisons and columns correspond to the objects. For example, if $n = 4$ the paired comparisons are $(1,2), (1,3), (1,4), (2,3), (2,4)$ and $(3,4)$

$$\boldsymbol{A} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

A similar model is employed by Böckenholt and Tsai (2001), who assume that $\boldsymbol{\epsilon}_s \sim N(\boldsymbol{0}, \omega^2 \mathbf{I}_N)$ and consider the inclusion of covariates in the model. The identification restrictions needed to estimate this model are discussed in Maydeu-Olivares (2001), Maydeu-Olivares (2003), Tsai and Böckenholt (2002) and Tsai (2003). The problem arises because binary data allow to identify only the correlation matrix corresponding to the covariance matrix of the latent variable $\boldsymbol{Z}_s$. In this case $\boldsymbol{\Sigma}_Z = \text{Cov}(\boldsymbol{Z}_s) = \boldsymbol{A}\boldsymbol{\Sigma}_T\boldsymbol{A}' + \boldsymbol{\Omega}$, where $\boldsymbol{\Sigma}_T$ is an unrestricted covariance matrix. As demonstrated by Tsai (2003), $n+2$ constraints are needed in order to identify a model with such a covariance matrix, including the constraint on the worth parameters. As for the mean parameters, many different constraints of the covariance matrix are possible. For example, Maydeu-Olivares (2003) sets all the diagonal elements of $\boldsymbol{\Sigma}_T$ equal to 1 and one of the diagonal elements of $\boldsymbol{\Omega}$ to 1.

Takane (1989) proposes a factor model for $\boldsymbol{\Sigma}_T$ in order to overcome the over-parameterisation problem. A factor model assumes that $\boldsymbol{\Sigma}_T = \boldsymbol{X}\boldsymbol{X}' + \boldsymbol{\Psi}$, where $\boldsymbol{X}$ is an $n \times r$ matrix which represents the loadings of vector $\boldsymbol{T}$ and $\boldsymbol{\Psi}$ is a diagonal matrix. The more general analysis of covariance structure proposed by Takane (1989) can accommodate both the wandering vector (De Soete and Carroll, 1983; Carroll and De Soete, 1991) and the wandering ideal point (De Soete *et al.*, 1989) models. The former model assumes that each subject is represented by a vector emanating from the origin whose termini follow a multivariate normal distribution and the objects are represented by points in an r dimensional space. In each paired comparison, a subject samples randomly a vector and the object whose orthogonal projection on the vector is larger is preferred. The wandering ideal point model assumes that each subject is represented by a normally distributed random point. Each time the subject has to perform a paired comparison, a random point is sampled and the object with smaller Euclidean distance from the sampled point is preferred. The wandering vector and wandering ideal point models do not impose the number of dimensions which is determined from the data alone. So they are powerful models to analyse human choice behaviour and inferring perceptual dimensions, but also to analyse the behaviour of organisms at different levels of evolution.

A further extension of model (5) is proposed in Tsai and Böckenholt (2008) who unify Tsai and Böckenholt (2006) with Takane (1989) to obtain a general class of models that can account simultaneously for transitive choice behaviour and systematic deviations from it. In this case the latent variable is

$$\boldsymbol{Z}_s = \boldsymbol{A}\boldsymbol{T} + \boldsymbol{B}\boldsymbol{V}_s, \tag{6}$$

where $\boldsymbol{V}_s = (V_{s1(2)}, V_{s1(3)}, \ldots, V_{s2(1)}, V_{s2(3)}, \ldots, V_{sn\,(n-1)})'$ is a vector of zero mean random effects which capture the random variation in judging an object when compared to another specific object and $\boldsymbol{B}$ is a matrix with rows corresponding to the paired comparisons and columns corresponding to the elements of $\boldsymbol{V}_s$, so for example if $n = 3$, $\boldsymbol{V}_s = (V_{s1(2)}, V_{s1(3)}, V_{s2(1)}, V_{s2(3)}, V_{s3(1)}, V_{s3(2)})'$ and

$$\boldsymbol{B} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix}.$$

It is assumed that $\boldsymbol{V}_s$, the within-judge variability, is normally distributed with mean 0 and covariance $\boldsymbol{\Sigma}_V$ so that $\boldsymbol{Z}_s \sim N(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}_T\boldsymbol{A}' + \boldsymbol{B}\boldsymbol{\Sigma}_V\boldsymbol{B}')$. Particular care is needed when specifying the structure of $\boldsymbol{\Sigma}_V$ to not incur in identification issues.

5.1.2 Models with logit link

The dependence between evaluations made by the same judge has been introduced also in models employing logit link functions.

Lancaster and Quade (1983) consider multiple judgements by the same person and introduce correlation in the Bradley-Terry model assuming that the worth parameters are random variables. Let π_{ij} denote the probability that i is preferred to j in a two categorical paired comparison experiment. Lancaster and Quade (1983) assume that each π_{ij} is a random variable following a Beta distribution with shape parameters a_{ij} and b_{ij}. The Bradley-Terry model is imposed on the means of the Beta distributions, that is $\mathrm{E}(\pi_{ij}) = a_{ij}/(a_{ij} + b_{ij}) = \pi_i/(\pi_i + \pi_j)$. Hence, the average preference probabilities of the population of judges will satisfy the Bradley-Terry model even though each judge may not. However, the proposed model introduces correlation only between comparisons of the same judge on the same pair of objects, while the other comparisons remain independent.

Matthews and Morris (1995) extend on Lancaster and Quade (1983) considering a model with three possible response categories. They assume that probabilities of preference for one of the objects or ties follow a Dirichlet distribution.

Böckenholt and Dillon (1997a) state that the approach by Agresti (1992) which allows for within-judge dependencies by fitting an ordinal extension of the Bradley-Terry model to the marginal paired comparison distribution can be useful only when individual preference differences are of secondary importance in the data analysis. Böckenholt and Dillon (1997a) consider an H categorical response model and, since the adjacent categories model reduces to the Bradley-Terry model in case of two response categories, the Authors propose the log-odds ratio as the relevant association measure. For example, in case of two categorical responses the 2×2 subtables formed by the possible outcomes of two paired comparisons can be considered. In case of H categorical responses the odds-ratios are

$$\gamma_{i,kj} = \ln \left(\frac{\mathrm{pr}(Y_{ik} = h, Y_{ij} = h)\mathrm{pr}(Y_{ki} = h+1, Y_{ji} = h+1)}{\mathrm{pr}(Y_{ki} = h, Y_{ij} = h+1)\mathrm{pr}(Y_{ik} = h+1, Y_{ij} = h)} \right).$$

Böckenholt (2001) considers the case in which the worth of object i for subject s is

$$\mu_{si} = \mu_i + \sum_{p=1}^{P} \beta_{ip} x_{sip} + U_{si},$$

where U_{si} is a random component and $\boldsymbol{x}_{si}$ is a vector of P subject-specific (and possibly item specific) covariates. Böckenholt (2001) employs a logit link function and assumes that $\boldsymbol{U}_s = (U_{s1}, \dots U_{sn})'$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_U$.

In the log-linear representation of the Bradley-Terry model, Dittrich *et al.* (2002) include further parameters that account for dependence between choices involving a common object.

5.2 Object-related dependencies

In the multiple judgement sampling the dependence among observations derives from repeated comparisons made by the same person, usually involving a common object. In case paired comparisons are not performed by a judge, the correlation may arise from the fact that the same object is involved in multiple paired comparisons. For example, when zoological or sport data are considered, it is realistic to assume that comparisons involving the same animal or the same player are correlated. In this perspective, Firth (2005) suggests to set $\mu_i = \boldsymbol{x}_i'\boldsymbol{\beta} + U_i$, where U_i is a zero mean object-specific random effect. This approach is investigated in Cattelan

(2009). This model has many elements in common with those proposed by psychometricians and poses similar inferential challenges. However, while in pshychometric applications n is not very large because it is unlikely that a person will make all the paired comparisons when $n > 10$, this will typically happen in sport tournaments or in paired comparison data about animal behaviour. Moreover, in the multiple judgement sampling scheme S independent replications of all the comparisons are available, but in other contexts this does not occur.

5.3 Inference

Thurstonian models are most commonly specified when dealing with dependent paired comparison data mainly for computational convenience, nevertheless inference in those models poses non-trivial problems. The first issue in the unstructured and unrestricted Thurstonian models proposed in the psychometric literature regards identifiability of the model itself which requires some constraints. Then, estimation of models for paired comparison dependent data presents numerical difficulties.

5.3.1 Estimation

In Thurstonian models estimation is problematic since the computation of the likelihood of the paired comparisons expressed by a judge requires the approximation of an integral of dimension N, the number of the paired comparisons. The latent variable $\boldsymbol{Z}_s$ can be standardised as $\boldsymbol{Z}_s^* = \boldsymbol{D}(\boldsymbol{Z}_s - \boldsymbol{A}\boldsymbol{\mu})$ where $\boldsymbol{D} = [\mathrm{Diag}(\boldsymbol{\Sigma}_Z)]^{1/2}$, and $\boldsymbol{\Sigma}_Z$ denotes the covariance matrix of $\boldsymbol{Z}_s$ expressed as in model (5) or in model (6). Then, $\boldsymbol{Z}_s^*$ follows a multivariate normal distribution with mean $\boldsymbol{0}$ and correlation matrix $\boldsymbol{\Sigma}_{Z^*} = \boldsymbol{D}\boldsymbol{\Sigma}_Z\boldsymbol{D}$. Object i is preferred to object j when $z_{sij}^* \geq \tau_{ij}^*$, where the vector of the thresholds is given by $\boldsymbol{\tau}^* = -\boldsymbol{D}\boldsymbol{A}\boldsymbol{\mu}$. Then, the probability of the observed results for the paired comparisons performed by the judge s in case of only two possible outcomes is

$$\mathcal{L}_s(\boldsymbol{\psi}; \boldsymbol{Y}_s) = \int_{R_{s12}} \cdots \int_{R_{sn-1\,n}} \phi_N(\boldsymbol{z}_s^*; \boldsymbol{\Sigma}_{Z^*}) \mathrm{d}\boldsymbol{z}_s^*, \tag{7}$$

where $\boldsymbol{\psi}$ denotes the model parameters, $\phi_N(\cdot; \boldsymbol{\Sigma}_{Z^*})$ denotes the density function of an N-dimensional normal random variable with mean $\boldsymbol{0}$ and correlation matrix $\boldsymbol{\Sigma}_{Z^*}$ and

$$R_{sij} = \begin{cases} (-\infty, \tau_{ij}^*) & \text{if } Y_{sij} = 1 \\ (\tau_{ij}^*, \infty) & \text{if } Y_{sij} = 2. \end{cases}$$

In case there is more than one judge, the likelihood is the product of the probability of the observations for each judge

$$\mathcal{L}(\boldsymbol{\psi}; \boldsymbol{Y}) = \prod_{s=1}^{S} \mathcal{L}_s(\boldsymbol{\psi}; \boldsymbol{Y}_s).$$

Note that the dimension of integral (7) is equal to $N = n\,(n-1)/2$, the number of paired comparisons, so its growth is quadratic with the increase in the number of objects.

There are different algorithms to approximate multivariate normal probabilities. The algorithm proposed by Genz and Bretz (2002) is based on quasi-Monte Carlo methods and Craig (2008) warns against the randomness of this method for likelihood evaluation. A deterministic approximation is developed by Miwa *et al.* (2003), but it is available only for integrals of dimension up to 20 since even for such a dimension its computation is very slow. Different methods have been proposed to overcome this problem, some of them are described below.

Maydeu-Olivares (2001), Maydeu-Olivares (2002) and Maydeu-Olivares and Böckenholt (2005) propose a limited information procedure for multiple judgement sampling which is performed in three stages. The first stage consists in estimating the threshold parameters exploiting the empirical univariate proportions of wins. In the second stage the elements of $\boldsymbol{\Sigma}_{Z^*}$, which are tetrachoric correlations, are estimated employing the bivariate proportions of wins. Finally, in the third stage the model parameters $\boldsymbol{\psi}$ are estimated by minimising the function

$$G = [\tilde{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\boldsymbol{\psi})]'\hat{\boldsymbol{W}}[\tilde{\boldsymbol{\kappa}} - \boldsymbol{\kappa}(\boldsymbol{\psi})],$$

where $\tilde{\boldsymbol{\kappa}}$ denotes the thresholds and tetrachoric correlations estimated in the first and second stages, $\boldsymbol{\kappa}(\boldsymbol{\psi})$ denotes the thresholds and tetrachoric correlations under the restrictions imposed on those parameters by the model parameters $\boldsymbol{\psi}$ and $\hat{\boldsymbol{W}}$ is a non negative definite matrix. Let $\boldsymbol{\Xi}$ denote the asymptotic covariance matrix of $\tilde{\boldsymbol{\kappa}}$. Then it is possible to use $\hat{\boldsymbol{W}} = \hat{\boldsymbol{\Xi}}^{-1}$, (Muthén, 1978), $\hat{\boldsymbol{W}} = \mathrm{diag}(\hat{\boldsymbol{\Xi}})^{-1}$ (Muthén *et al.*, 1997) or $\hat{\boldsymbol{W}} = \mathbf{I}$ (Muthén, 1993). The last two options seem more stable in data sets with a small number of objects (Maydeu-Olivares, 2001). This method is very fast and Maydeu-Olivares (2001) states that it may have an edge over full information methods because it uses only the one and two dimensional marginals of a large and sparse contingency table.

A pairwise likelihood (Le Cessie and Van Houwelingen, 1994) approach proved useful for data with object-related dependencies (Cattelan, 2009), so it may be a valid alternative also in multiple judgement sampling. Pairwise likelihood is a special case of the broader class of composite likelihoods (Lindsay, 1988; Varin *et al.*, 2011). Pairwise likelihood for paired comparison models consists of the product of the marginal bivariate probabilities

$$\mathcal{L}^s_{\mathrm{pair}}(\boldsymbol{\psi}; \boldsymbol{Y}_s) = \prod_{i=1}^{n-2} \prod_{j=i+1}^{n-1} \prod_{k=i}^{n-1} \prod_{l=j+1}^{n} \mathrm{pr}(Y_{sij} = y_{sij}, Y_{skl} = y_{skl}).$$

The pairwise likelihood of all the observations is the product of the pairwise likelihoods relative to the single judges $\mathcal{L}_{\mathrm{pair}}(\boldsymbol{\psi}; \boldsymbol{Y}) = \prod_{s=1}^{S} \mathcal{L}^s_{\mathrm{pair}}(\boldsymbol{\psi}; \boldsymbol{Y}_s)$. The logarithm of the pairwise likelihood for subject s is $\ell^s_{\mathrm{pair}}(\boldsymbol{\psi}; \boldsymbol{Y}_s) = \log \mathcal{L}^s_{\mathrm{pair}}(\boldsymbol{\psi}; \boldsymbol{Y}_s)$ while the whole pairwise log-likelihood is $\ell_{\mathrm{pair}}(\boldsymbol{\psi}; \boldsymbol{Y}) = \sum_{s=1}^{S} \ell^s_{\mathrm{pair}}(\boldsymbol{\psi}; \boldsymbol{Y}_s)$. Under regularity conditions, the maximum pairwise likelihood estimator is consistent and asymptotically normally distributed with mean $\boldsymbol{\psi}$ and covariance matrix $\mathbf{G}(\boldsymbol{\psi}) = \mathbf{H}(\boldsymbol{\psi})^{-1}\mathbf{J}(\boldsymbol{\psi})\mathbf{H}(\boldsymbol{\psi})^{-1}$, where $\mathbf{J}(\boldsymbol{\psi}) = \mathrm{var}\ \{\nabla \ell_{\mathrm{pair}}(\boldsymbol{\psi}; \boldsymbol{Y})\}$ and $\mathbf{H}(\boldsymbol{\psi}) = E\left\{-\nabla^2 \ell_{\mathrm{pair}}(\boldsymbol{\psi}; \boldsymbol{Y})\right\}$, see Cox and Reid (2004). Pairwise likelihood reduces noticeably the computational effort since it requires only the computations of bivariate normal probabilities. Moreover, in the multiple judgement sampling scheme the standard errors can be computed straightforwardly by exploiting the independence between the observations of different judges. In fact, $\mathbf{H}(\boldsymbol{\psi})$ can be estimated by the Hessian matrix computed in the maximum pairwise likelihood estimate, while $\mathbf{J}(\boldsymbol{\psi})$ can be estimated through the cross-product $\sum_{s=1}^{S} \nabla \ell_{\text{pair}}^{s}(\hat{\boldsymbol{\psi}}; \boldsymbol{Y}_s) \nabla \ell_{\text{pair}}^{s}(\hat{\boldsymbol{\psi}}; \boldsymbol{Y}_s)'$.

5.3.2 Simulation studies

Simulation studies were performed in order to check the performance of the proposed pairwise likelihood approach. The results of the pairwise likelihood method are compared to the limited information estimation method proposed by Maydeu-Olivares and Böckenholt (2005) in two different settings employing model (5) and model (6). It is assumed that $n = 4$, hence also a full likelihood approach based on the algorithm by Miwa *et al.* (2003) can be used since the integral has dimension 6.

The first simulation setting is the same as that proposed in Maydeu-Olivares (2001), where the model $\boldsymbol{Z}_s = \boldsymbol{A}\boldsymbol{T} + \boldsymbol{e}_s$ is assumed with

$$\boldsymbol{\mu} = \begin{pmatrix} 0.5 \\ 0 \\ -0.5 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma}_T = \begin{pmatrix} 1 & & & \\ 0.8 & 1 & & \\ 0.7 & 0.6 & 1 & \\ 0.8 & 0.7 & 0.6 & 1 \end{pmatrix},$$

and the covariance matrix of $\boldsymbol{e}$ is $\boldsymbol{\Omega} = \omega^2 \mathbf{I}_4$. For identification purposes it is necessary to set the diagonal elements of $\boldsymbol{\Sigma}_T$ equal to 1, $\mu_4 = 0$ and $\omega^2 = 1$. Hence, in this case $\boldsymbol{\Sigma}_T$ is actually a correlation matrix. Table 6 shows the estimates and standard errors of 1,000 simulations assuming $S = 100$ judges. For limited information estimation the matrix $\hat{\boldsymbol{W}} = \mathbf{I}$ is employed. In this setting in which $\boldsymbol{\Sigma}_T$ is actually a correlation matrix, all the methods seem to perform comparably well.

Table 6: Empirical means and standard errors of 1,000 simulated estimates obtained by maximum likelihood (`ML`), limited information estimation (`LI`) and pairwise likelihood (`PL`) with $S = 100$.

		ML		LI		PL	
	True value	Est.	S.E.	Est.	S.E.	Est.	S.E.
μ_1	0.5	0.507	0.128	0.507	0.130	0.504	0.128
μ_2	0	0.008	0.128	0.009	0.123	0.008	0.124
μ_3	-0.5	-0.491	0.151	-0.495	0.152	-0.494	0.150
σ_{12}	0.8	0.791	0.128	0.783	0.131	0.790	0.130
σ_{13}	0.7	0.693	0.170	0.689	0.169	0.693	0.179
σ_{14}	0.8	0.781	0.132	0.775	0.133	0.780	0.142
σ_{23}	0.6	0.576	0.192	0.566	0.192	0.575	0.189
σ_{24}	0.7	0.670	0.164	0.665	0.163	0.670	0.189
σ_{34}	0.6	0.575	0.207	0.571	0.199	0.574	0.200

The second simulation setting considers model (6) proposed by Tsai and Böckenholt (2008). Since only differences are recoverable, the model can be defined using means and variances of the differences $T_i - T_n$ for $i = 1, \ldots, n-1$. The assumed worth parameters of these differences are $(-0.2, 1, -1.5)$ while the covariance matrix is

$$\begin{pmatrix} 1.5 & 1 & 1.3 \\ 1 & 4 & 2.5 \\ 1.3 & 2.5 & 3 \end{pmatrix}.$$

Differently from the previous setting, this specification of the model allows to estimate also the variance of the differences $T_i - T_n$ and to check whether they are different for the various objects. This aspect is of particular interest in psychometrics.

Tsai and Böckenholt (2008) propose a specification of the matrix $\boldsymbol{B}$ which depends only on one parameter b whose value is set equal to 0.5. Table 7 presents the results of the simulations. Maximum likelihood based on numerical integration is the method that performs best, but it can be exploited only because n is small. However, pairwise likelihood estimation seems to perform quite well, especially if compared to limited information estimation which seems not satisfactory in this case with $S = 100$, as already noticed in Tsai and Böckenholt (2008). It is possible to consider an increase in the number of objects n or an increase in the number of subjects S that perform the comparisons. Tsai and Böckenholt (2008) conduct also a larger simulation with $n = 4$ and $S = 300$. The Authors conclude that in this case limited information estimation produces estimates which are accurate enough. However, pairwise likelihood yields estimates which are acceptable even with $S = 100$ and larger S may reduce the standard errors of the estimates.

Finally, notice that pairwise likelihood can be employed also in situations with large n and $S = 1$ or S very small as happens in sports data or animal behaviour applications (Cattelan, 2009).

Example We fit model (5) to universities data by means of pairwise likelihood. A full likelihood approach based on numeric approximation is not used since it is necessary to approximate 303 integrals of dimension 15 and this would take very long. It is assumed that $\boldsymbol{\Omega} = \omega^2 \mathbf{I}_{15}$. The constraints employed for estimation are those proposed in Maydeu-Olivares and Hernández (2007). In this case the diagonal elements of $\boldsymbol{\Sigma}_T$ are set equal to 1 and the additional constraint $\sum_{k=2}^{n} \sigma_{k1} = 1$, where σ_{kl} denotes the element in row k and column l of the matrix $\boldsymbol{\Sigma}_T$, is used. The constraints proposed by Maydeu-Olivares and Hernández (2007) are employed because they facilitate the interpretation of the results. Indeed, with these constraints a positive covariance means that strong preference for a stimulus is associated with strong preference for the other stimulus, while a negative covariance means that strong preference for one stimulus is associated with weak preference for the other stimulus. The estimate of the threshold parameter (with standard error in brackets) is $\hat{\tau}_2 = 0.205\,(0.018)$ while the variance parameter is $\hat{\omega}^2 = 0.180\,(0.026)$. Table 8 shows the estimates of the mean and correlation matrix for the six universities. A high correlation is estimated between Barcelona and Milan, so strong preference for Barcelona is associated with strong preference for Milan. Even though some correlations do not seem significant, it appears that a strong preference for St. Gallen is

Table 7: Empirical means and standard errors of 1,000 simulations estimated by maximum likelihood (`ML`), limited information (`LI`) and pairwise likelihood (`PL`) with $S = 100$.

		ML		LI		PL	
	True value	Est.	S.E.	Est.	S.E.	Est.	S.E.
μ_1	-0.2	-0.207	0.191	-0.226	0.205	-0.215	0.192
μ_2	1	1.003	0.307	1.068	0.417	1.025	0.331
μ_3	-1.5	-1.511	0.317	-1.594	0.489	-1.542	0.357
σ_1^2	1.5	1.535	0.785	2.058	1.967	1.698	1.054
σ_2^2	4	4.015	1.678	5.342	4.424	4.446	2.420
σ_3^2	3	2.996	1.277	3.913	3.170	3.322	1.929
σ_{12}	1	0.942	0.615	1.340	1.443	1.116	0.874
σ_{13}	1.3	1.273	0.649	1.716	1.476	1.433	0.954
σ_{23}	2.5	2.467	1.007	3.351	2.673	2.767	1.534
b	0.5	0.530	0.407	0.720	0.820	0.581	0.501

associated with a weak preference for all the other universities but Stockholm. The worth parameters denote the same ranking of all universities as the one arising from Table 3.

6 Software

Critchlow and Fligner (1991) show that through the loglinear representation of the Bradley-Terry model standard programs can be used to estimate it. In the literature some estimation algorithms for the model assuming independent observations have been proposed. For example, Hunter (2004) develops a maximization-minimization algorithm for estimating Bradley-Terry models both with order effects and ties. However, the implementation of the estimation algorithms proposed in the literature can be difficult for the final user. Fortunately, fitting models to paired comparison data is facilitated by some `R` packages which allow fitting of the classical models and in some cases also of more complicated models.

The `eba` package (Wickelmaier and Schmid, 2004) fits elimination by aspects models (Tversky, 1972) to paired comparison data. The elimination by aspects model assumes that different objects present various aspects. The worth of each object is the sum of the worth associated with each aspect possessed by the object. When all objects possess only one relevant aspect, then the elimination by aspects model reduces to the Bradley-Terry model. Therefore, in case only one aspect per object is specified, the function `eba` fits an unstructured Bradley-Terry model. Covariates cannot be included and ties are not allowed, but `eba` can estimate an order effect in case one of the objects enjoys some benefits from the order of the presentation. The `eba` function requires only that all worth parameters are positive, hence any multiple of the worth parameters produces the same value of the likelihood. The function `strans` checks how many violations of weak, moderate and strong stochas-

Table 8: Estimates and standard errors (in brackets) of mean and correlation parameters of model (5) for universities data.

	Barcelona	London	Milan	Paris	St. Gallen	Stockholm	μ
Barcelona	1						0.405
	(fixed)						(0.073)
London	0.058	1					1.346
	(0.084)	(fixed)					(0.087)
Milan	0.724	0.185	1				0.308
	(0.062)	(0.097)	(fixed)				(0.074)
Paris	0.171	0.054	0.331	1			0.748
	(0.094)	(0.117)	(0.113)	(fixed)			(0.086)
St. Gallen	-0.303	-0.139	-0.298	-0.496	1		0.371
	(0.113)	(0.139)	(0.144)	(0.157)	(fixed)		(0.081)
Stockholm	0.350	0.316	0.339	0.144	0.287	1	0
	(0.079)	(0.091)	(0.097)	(0.113)	(0.130)	(fixed)	(fixed)

tic transitivity are present in the data and the function `thurstone` fits a Thurstone model to the data assuming independent observations.

The `prefmod` package (Hatzinger, 2010) fits Bradley-Terry models exploiting their loglinear representation. Ordinal paired comparisons are allowed, but the model reduces the categories to three or two depending on whether there is a no preference category or not. There are three different functions for estimating models for paired comparison data: the `llbt.fit` function which estimates the loglinear version of the Bradley-Terry model through the estimation algorithm described in Hatzinger and Francis (2004), the `llbtPC.fit` function that estimates the loglinear model exploiting the `gnm` (Turner and Firth, 2010b) function for fitting generalised nonlinear models and the `pattPC.fit` function which fits paired comparison data using a pattern design, that is all possible patterns of paired comparisons. Missing covariates are not allowed, in such a case the corresponding data are removed. The function `pattPC.fit` allows to include a covariate for the interaction between comparisons that have one item in common. However, the response table grows dramatically with the number of objects since in case of only two possible outcomes the number of patterns is 2^N, so no more than 6 objects can be included with 2 response categories and not more than 5 with three response categories. The function `pattPC.fit` handles also some cases in which the responses are missing not at random, while the function `pattnpml.fit` fits a mixture model to overdispersed paired comparison data using nonparametric maximum likelihood in which the number of latent classes is specified by the user.

The `BradleyTerry2` package (Turner and Firth, 2010a) expands the previous `BradleyTerry` (Firth, 2008) package and allows to fit both unstructured and structured paired comparison models with logit, probit and cauchit link functions. Model fitting is either by maximum likelihood, penalised quasi-likelihood (when there are random effects as specified in Section 5.2) or bias-reduced maximum likelihood (Firth, 1993). In case of unstructured models it is possible to choose the reference category whose worth is set to zero. The structured model can include also object-specific random effects that are assumed normally distributed, in this case the model is estimated by means of penalised quasi-likelihood (Breslow and Clayton, 1993). If there are missing explanatory variables, an additional worth parameter for the object with missing covariates is estimated. Order effects and more general comparison-specific covariates can be included, but only win-loss responses are allowed. It is possible to use bias-reduced maximum likelihood to produce finite estimates and standard errors even in case of complete separation of the data, situation which occurs for example when an object is preferred in all the paired comparisons in which it is involved.

The package `psychotree` (Strobl *et al.*, 2011) implements the method for recursive partitioning of the subjects on the basis of their explanatory variables and estimates an unstructured Bradley-Terry model for each of the final subsamples of subjects, see Section 4.2.

There are various packages for estimating models for paired comparison data, nevertheless even some of the most simple specifications cannot be straightforwardly fitted. For example, there is no function that estimates a Thurstone model when ties are allowed.

7 Conclusions

Paired comparison data are often criticised because they are relative and not absolute measures. However, this is the reason why they are so widely employed in applications, especially when the judgement of a person is required. Thurstonian and Bradley-Terry models are usually applied for the analysis of this type of data and the paper reviews some of the many extensions of those models that have been proposed in the literature. However, there are many other aspects which have not been considered here, for example the problem of optimal design (Graßhoff and Schwabe, 2008; Goos and Großmann, 2011), the development of models for multi-dimensional data when objects are evaluated with respect to multiple aspects (Böckenholt, 1988; Dittrich *et al.*, 2006), the temporal extension for comparisons repeated in time (Fahrmeir and Tutz, 1994; Glickman, 2001; Böckenholt, 2002; Dittrich *et al.*, 2008), the estimation of abilities of individuals belonging to a team that performs the paired comparisons (Huang *et al.*, 2006; Menke and Martinez, 2008) and many more.

Here, particular attention has been focused on models for dependent data. Dependencies arise when different paired comparisons are performed by the same judge. Hence, the issue of the dependence structure of the data has been investigate both in the statistic and psychometric literature. Thurstonian models seem particularly suitable to account for dependence between observations. Unfortunately, inference in those models requires the approximation of integrals whose dimension is equal to the number of paired comparisons, which increases rapidly with n. The proposed pairwise likelihood method seems to perform well in different scenarios.

Psychometric models are as unstructured and unrestricted as possible, but this produces identifiability issues and the necessary constraints cause difficulties in the interpretation of the estimates. Moreover, since different covariance matrices yield the same value of the likelihood, the best solution has to be chosen on the basis of goodness of fit statistics. These statistics may be problematic since the corresponding contingency table may be very sparse. In some instances, the mean of Pearson chi-square statistics associated with all the 2 × 2 tables that can be formed with the results of the paired comparisons is considered, in other instances the fit to the triples of results is employed. Maydeu-Olivares (2001) proposes a goodness of fit statistic which overcomes the problem of sparseness of the contingency table, but it actually regards the fitting of thresholds and tetrachoric correlations and not of the observed data.

All the models for dependent data are presented assuming win-loss responses, however it is easy to extend them to ordinal paired comparison data exploiting the latent variable $\boldsymbol{Z}_s$ and introducing threshold parameters as described in Section 3.

In some instances, for example when n is large, not all paired comparisons are made by all judges, but each person performs only a part of all the comparisons. These data can be analysed by means of the presented models with small modifications.

References

Agresti, A. (1992). Analysis of ordinal paired comparison data. *Applied Statistics* **41**, 287-297.

Agresti, A. (2002). *Categorical Data Analysis.* New York: Wiley.

Barry, D. and Hartigan, J. A. (1993) Choice models for predicting divisional winners in major league baseball. *Journal of the American Statistical Association* **88**, 766-774.

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

Böckenholt, U. (1988). A logistic representation of multivariate paired-comparison model. *Journal of Mathematical Psychology* **32**, 44-63.

Böckenholt, U. (2001). Hierarchical modeling of paired comparison data. *Psychological Methods* **6**, 49-66.

Böckenholt, U. (2002). A Thurstonian analysis of preference change. *Journal of Mathematical Psychology* **46**, 300-314.

Böckenholt, U. and Dillon, W. R. (1997a). Modeling within-subject dependencies in ordinal paired comparison data. *Psychometrika* **62**, 411-434.

Böckenholt, U. and Dillon, W. R. (1997b). Some new methods for an old problem: modeling preference changes and competitive market structures in pretest market data. *Journal of Marketing Research* **34**, 130-142.

Böckenholt, U. and Tsai, R.-C. (2001). Individual differences in paired comparison data. *British Journal of Mathematical and Statistical Psychology* **54**, 265-277.

Bradley, R. A. (1976). Science, statistics and paired comparisons. *Biometrics* **32**, 213-239.

Bradley, R. A. and Terry, M. E. (1952). The rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39**, 324-345.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Carroll, J. D. and De Soete, G. (1991). Toward a new paradigm for the study of multiattribute choice behavior. Spatial and discrete modeling of pairwise preferences. *American Psychologist* **46**, 342-351.

Cattelan, M. (2009). Correlation models for paired comparisons data. *Ph.D. Thesis*, Department of Statistical Sciences, University of Padova.

Cattelan, M., Varin, C. and Firth, D. (2010). Stochastic dynamic Thurstone-Mosteller models for sports tournaments. CRiSM working paper N. 10-19.

Causeur, D. and Husson, F. (2005). A 2-dimensional extension of the Bradley-Terry model for paired comparisons. *Journal of Statistical Planning and Inference* **135**, 254-259.

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

Cox, D.R. and Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737.

Craig, P. (2008). A new reconstruction of multivariate normal orthant probabilities. *Journal of the Royal Statistical Society Series B* **70**, 227-243.

Critchlow, D. E. and Fligner, M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika* **56**, 517-533.

David, H. (1988). *The Method of Paired Comparisons.* Griffin, London.

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* **65**,317-328.

Davidson, R. R. and Farquhar, P. H. (1976). A bibliography on the method of paired comparisons. *Biometrics* **32**, 241-252.

De Soete, G. and Carroll, J. D. (1983). A maximum likelihood method for fitting the wandering vector model. *Psychometrika* **48**, 553-566.

De Soete, G., Carroll, J. D. and DeSarbo, W. S. (1989). The wandering ideal point model for analyzing paired comparisons data. In G. De Soete, H. Feger and K. C. Klauer (Eds.), *New Developments in Psychological Choice Modeling*, 123-137. Amsterdam: North-Holland.

De Soete, G. and Winsberg, S. (1993). A Thurstonian pairwise choice model with univariate and multivariate spline transformations. *Psychometrika* **58**, 233-256.

Dillon, W. R., Kumar, A. and de Borrero, M. S. (1993). Capturing individual differences in paired comparisons: an extended BTL model incorporating descriptor variables. *Journal of Marketing Research* **30**, 42-51.

Dittrich, R., Francis, B., Hatzinger, R. and Katzenbeisser, W. (2006). Modelling dependency in multivariate paired comparisons: a log-linear approach. *Mathematical Social Sciences* **52**, 197-209.

Dittrich, R., Francis, B. and Katzenbeisser, W. (2008). Temporal dependence in longitudinal paired comparisons. Research Report, Department of Statistics and Mathematics, WU Vienna University of Economics and Business.

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. *Applied Statistics* **47**, 511-525.

Dittrich, R., Hatzinger, R. and Katzenbeisser, W. (2001). Corrigendum: Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Applied Statistics* **50**, 247-249.

Dittrich, R., Hatzinger, R. and Katzenbeisser, W. (2002). Modelling dependencies in paired comparison data. A log-linear approach. *Computational Statistics and Data Analysis* **40**, 39-57.

Dittrich, R., Hatzinger, R. and Katzenbeisser, W. (2004). A log-linear approach for modelling ordinal paired comparison data on motives to start a PhD program. *Statistical Modelling* **4**, 1-13.

Duineveld, C. A. A., Arents, P. and King. B. M. (2000). Log-linear modelling of paired comparison data from consumer tests. *Food Quality and Preference* **11**, 63-70.

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

Fahrmeir, L. and Tutz, G. (1994) Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association* **89**, 1438-1449.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27-38.

Firth, D. (2005). Bradley-Terry models in R. *Journal of Statistical Software* **12**, 1-12.

Firth, D. (2008) BradleyTerry: Bradley-Terry models. (http://CRAN.R-project.org/package=BradleyTerry).

Firth, D. and Menezes, R. X. de (2004). Quasi-variances. *Biometrika* **91**, 65-80.

Ford, L. R. Jr. (1957). Solution to a ranking problem from binary comparisons. *The American Mathematical Monthly* **64**, 28-33.

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. *Applied Statistics* **51**, 319-336.

Genz, A. and Bretz, F. (2002). Methods for the computation of multivariate t-probabilities. *Journal of Computational and Graphical Statistics* **11**, 950-971.

Glenn, W. A. and David, H. A. (1960). Ties in paired-comparison experiments using a modified Thurstone-Mosteller model. *Biometrics* **16**, 86-109.

Glickman, M.E. (2001). Dynamic paired comparison models with stochastic variances. *Journal of Applied Statistics* **28**, 673-689.

Goos, P. and Großmann, H. (2011). Optimal design of factorial paired comparison experiments in the presence of within-pair order effects. *Food Quality and Preference* **22**, 198-204.

Graßhoff, U. and Schwabe, R. (2008). Optimal design for the Bradley-Terry paired comparison model. *Statistical Methods and Applications* **17**, 275-289.

Hatzinger, R. (2010). prefmod: Utilities to fit paired comparison models for preferences. (http://CRAN.R-project.org/package=prefmod).

Hatzinger, R. and Francis, B. J. (2004). Fitting paired comparison models in R. Research report, University of Wien (http://epub.wu-wien.ac.at/dyn/openURL?id=0ai:epub.wu-wien.ac.at:epub-wu-01_709).

Head, M. L., Doughty, P., Blomberg, S. P. and Keogh, S. (2008). Chemical mediation of reciprocal mother-offspring recognition in the Southern Water Skink (*Eulamprus heatwolei*). *Australian Ecology* **33**, 20-28.

Huang, T.-Z., Weng, R. C. and Lin, C.-J. (2006). Generalized Bradley-Terry models and multi-class probability estimates. *Journal of Machine Learning Research* **7**, 85-115.

Henery, R. J. (1992). An extension to the Thurstone-Mosteller model for chess. *The Statistician* **41**, 559-567.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Annals of Statistics* **32**, 384-406.

Jerome, S., Hicks, C. and Herron-Marx, S. (2009). Designing sexual health services for young people: a methodology for capturing the user voice. *Health and Social Care in the Community* **17**, 350-357.

Joe, H. (1990). Extended use of paired comparison models, with application to chess rankings. *Applied Statistics* **39**, 85-93.

Kissler, J. and Bäuml, K.-H. (2000). Effects of the beholder's age on the perception of facial attractiveness. *Acta Psychologica* **104**, 145-166.

Knorr-Held, L. (2000) Dynamic rating of sports teams. *The Statistician* **49**, 261-276.

Lancaster, J. F. and Quade, D. (1983). Random effects in paired-comparison experiments using the Bradley-Terry model. *Biometrics* **39**, 245-249.

Le Cessie, S. and Van Houwelingen, J. C. (1994). Logistic regression for correlated binary data. *Applied Statistics* **43**, 95-108.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics* **80**, 221-239.

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

Maydeu-Olivares, A. (2001). Limited information estimation and testing of Thurstonian models for paired comparison data under multiple judgement sampling. *Psychometrika* **66**, 209-228.

Maydeu-Olivares, A. (2002). Limited information estimation and testing of Thurstonian models for preference data. *Mathematical Social Sciences* **43**, 467-483.

Maydeu-Olivares, A. (2003). Thurstonian covariance and correlation structures for multiple judgement paired comparison data. Working Papers Economia, Instituto de Empresa, Area of Economic Environment (http://econpapers.repec.org/RePEc:emp:wpaper:wp03-04).

Maydeu-Olivares, A. and Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychometrika* **10**, 285-304.

Maydeu-Olivares, A. and Böckenholt, U. (2008). Modeling subject health outcomes. Top 10 reasons to use Thurstone's method. *Medical Care* **46**, 346-348.

Maydeu-Olivares, A. and Hernández, A. (2007). Identification and small sample estimation of Thurstone's unrestricted model for paired comparisons data. *Multivariate Behavioral Research* **42**, 323-347.

Mazzucchi, T. A., Linzey, W. G. and Bruning, A. (2008). A paired comparison experiment for gathering expert judgement for an aircraft wiring risk assessment. *Reliability Engineering and System Safety* **93**, 722-731.

McHale, I. and Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting* **27**, 619-630.

Menke, J. E. and Martinez, T. R. (2008). A Bradley-Terry artificial neural network model for individual ratings in group competitions. *Neural Computing & Applications* **17**, 175-186.

Miwa, A., Hayter J. and Kuriki, S. (2003). The evaluation of general non-centred orthant probabilities. *Journal of the Royal Statistical Society Series B* **65**, 223-234.

Mosteller, F. (1951). Remarks on the method of paired comparisons. I. The least squares solution assuming equal standard deviations and equal correlations. II. The effect of an aberrant standard deviation when equal standard deviations and equal correlations are assumed. III. A test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika* **16**, 3-9 203-218.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika* **43**, 551-560.

Muthén, B. (1993). Goodness of fit with categorical and other non normal variables. In: Bollen, K. A., Long, J., S. (Eds.), *Structural Equation Models*. Sage, Newbury Park, CA, 205-234.

Muthén, B., du Toit, S. H. C. and Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Technical Report.

R Development Core Team (2011). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. ISBN 3-900051-07-0 (http://www.R-project.org).

Rao, P. V. and Kupper, L. L. (1967). Ties in paired-comparison experiments: a generalization of the Bradley-Terry model. *Journal of the American Statistical Association* **62**, 194-204.

Sham, P. C. and Curtis, D. (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Annals of Human Genetics* **59**, 323-336.

Sinclair, C. D. (1982). GLIM for preference. *Lecture Notes in Statistics* **14**, 164-178.

Springall, A. (1973). Response surface fitting using a generalization of the Bradley-Terry paired comparison model. *Applied Statistics* **22**, 59-68.

Stern, H. (1990). A continuum of paired comparisons models. *Biometrika* **77**, 265-273.

Stern, S. E. (2011). Moderated paired comparisons: a generalized Bradley-Terry model for continuous data using a discontinuous penalized likelihood function. *Applied Statistics*. To appear.

Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science* **9**, 94-108.

Strobl, C., Wickelmaier, F. and Zeileis, A. (2011). Accounting for individual differences in Bradley-Terry models by means of recursive partitioning. *Journal of Educational and Behavioral Statistics* **36**, 135-153.

Stuart-Fox, D. M., Firth, D., Moussalli, A. and Whiting, M. J. (2006). Multiple signals in chameleon contests: designing and analysing animal contests as a tournament. *Animal Behaviour* **71**, 1263-1271.

Takane, Y. (1989). Analysis of covariance structures and probabilistic binary choice data. In *new Developments in Psychological Choice Modeling*, eds. De Soete, G., Feger, H. and klauser, K. C. Elsevier, North-Holland.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review* **34**, 368-389.

Tsai, R.-C. (2000). Remarks on the identifiability of Thurstonian ranking models: Case V, Case III, or neither? *Psychometrika* **65**, 233-240.

Tsai, R.-C. (2003). Remarks on the identifiability of Thurstonian paired comparison models under multiple judgement. *Psychometrika* **68**, 361-372.

Tsai, R.-C. and Böckenholt, U. (2002). Two-level linear paired comparison models: estimation and identifiability issues. *Mathematical Social Sciences* **43**, 429-449.

Tsai, R.-C. and Böckenholt, U. (2006). Modelling intransitive preferences: A random effect approach. *Journal of Mathematical Psychology* **50**, 1-14.

Tsai, R.-C. and Böckenholt, U. (2008). On the importance of distinguishing between within- and between-subject effects in intransitive intertemporal choice. *Journal of Mathematical Psychology* **52**, 10-20.

Turner, H. and Firth, D. (2010a). Bradley-Terry models in R: The `BradleyTerry2` package. (http://CRAN.R-project.org/package=BradleyTerry2).

Turner, H. and Firth, D. (2010b). Generalized nonlinear models in R: An overview of the `gnm` package. (http://CRAN.R-project.org/package=gnm).

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

Usami, S. (2010). Individual differences multidimensional Bradley-Terry model using reversible jump Markov chain Monte Carlo algorithm. *Behaviormetrika* **37**, 135-155.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5-42.

Whiting, M. J., Stuart-Fox, D. M., O'Connor, D., Firth, D., Bennett, N. C. and Blomberg, S. P. (2006). Ultraviolet signals ultra-aggression in a lizard. *Animal Behaviour* **72**, 353-363.

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

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **29**, 436-460.

Acknowledgements

The Author would like to thank Cristiano Varin and Alessandra Salvan for useful suggestions.