

A Model for Correlated Paired Comparison Data

Manuela Cattelan and Cristiano Varin

Abstract Paired comparison data arise when objects are compared in couples. This type of data occurs in many applications. Traditional models developed for the analysis of paired comparison data assume independence among all observations, but in some instances this seems unrealistic. Indeed, it is reasonable to believe that comparisons with a common object are correlated. A model that introduces correlation between comparisons with at least a common object is discussed. The likelihood function of the proposed model involves the approximation of a high dimensional integral. To overcome numerical difficulties a pairwise likelihood approach is adopted. The methodology is illustrated through the analysis of the 2008/2009 season of the Italian water polo league.

Key words: crossed random effects; latent variable models; paired comparison data; pairwise likelihood

1 Paired Comparison Data

Paired comparison data arise from the comparison of objects in couples. This type of data occurs in many applications such as consumer behaviour, preference testing, sensory testing, biology, acoustics, sports and many other areas.

The traditional models developed for the analysis of paired comparison data are the Thurstone-Mosteller model [7, 6] and the Bradley-Terry model [2]. In both models, the probability that an object is preferred over another is a function of the dif-

Manuela Cattelan

Department of Statistical Sciences, University of Padova, via C. Battisti 241, 35121 Padova, Italy
e-mail: manuela.cattelan@unipd.it

Cristiano Varin

Department of Statistics, University Ca' Foscari, San Giobbe, Cannaregio 873, 30121 Venice, Italy
e-mail: sammy@unive.it

ference of the true “worth” of the objects. The main difference between the two models lies in the link function: probit for the Thurstone-Mosteller model and logit for the Bradley-Terry model. Extensions of the original models have been developed to take into account specific features of paired comparison data as the existence of an order effect that advantages the object presented first, or situations in which there are three possible outcomes of the comparisons, i.e. preference for one of the two objects or impossibility to express a preference.

In general, traditional models for paired comparison data are applied assuming independence among all observations. This implies that the likelihood function is the product of the probabilities of the observed outcomes for each paired comparison. Often, the independence assumption seems unrealistic. For example, in sports data it is reasonable to suppose that the results of two matches involving a common player are correlated. In the following section a model that takes into account the correlation between observations with at least a common object is illustrated.

2 A Correlation Model for Paired Comparison Data

Let Y_{ij} be a binary random variable taking value 1 if object i is preferred to object j , and 0 otherwise. In traditional models for paired comparison data, the following generalised linear model is assumed. The density of Y_{ij} is distributed as a Bernoulli random variable whose mean is related to the worth of objects through

$$g\{E(Y_{ij})\} = \lambda_i - \lambda_j, \quad (1)$$

where g is a suitable link function and λ_i is the worth parameter for object i , $i = 1, \dots, n$. The worth parameter may depend on explanatory variables through the relation

$$\lambda_i = x_i' \beta,$$

where x_i is a p -dimensional vector of explanatory variables related to object i and β is a vector of p regression parameters.

Correlation between observations with a common object can be introduced by including an object-specific zero mean random effect u_i

$$\lambda_i = x_i' \beta + u_i.$$

Hence, eq. (1) becomes

$$g\{E(Y_{ij}|u_i, u_j)\} = (x_i - x_j)' \beta + u_i - u_j.$$

The addition of the random effects, besides introducing correlation among observations with common objects, allows also to account for the imperfect representation of the worth by a linear predictor.

The binary observations Y_{ij} can be represented as censored continuous random variables $Y_{ij} = I\{Z_{ij} > 0\}$, where $I\{A\}$ denotes the indicator function of the set A and

$$Z_{ij} = (x_i - x_j)' \beta + u_i - u_j + \varepsilon_{ij},$$

with the ε_{ij} 's being zero mean continuous random variables. Assume that the random effects u_i are independent, identically distributed normal random variables with zero mean and variance σ^2 , the ε_{ij} are independent, identically distributed standard normal variables and that errors and random effects are uncorrelated. The proposed model is therefore a Thurstone-Mosteller model with normal random effects. Hence, the correlation between a pair of hidden variables Z_{ij} and Z_{kl} is

$$\text{corr}(Z_{ij}, Z_{kl}) = \begin{cases} \sigma^2 / (1 + 2\sigma^2) & \text{if } i = k \text{ or } j = l, \\ 0 & \text{if } i \neq j \neq k \neq l, \\ -\sigma^2 / (1 + 2\sigma^2) & \text{if } i = l \text{ or } j = k, \end{cases}$$

Unfortunately, such a model has an intractable likelihood function. In fact, the likelihood function is computed integrating out all the random effects

$$L(\theta; y) = \int_{R^n} \left\{ \prod_{i=1}^{n-1} \prod_{j=i+1}^n P(Y_{ij} = y_{ij} | u_i, u_j; \theta) \right\} \left\{ \prod_{i=1}^n \phi(u_i; \theta) du_i \right\}, \quad (2)$$

where $\phi(u_i; \sigma^2)$ denotes the normal density function of the random effects u_i depending on the parameter σ^2 and $\theta = (\beta', \sigma^2)'$. The computation of the full likelihood requires the approximation of the integral (2), which has dimension equal to the number of objects being compared. To overcome numerical difficulties, inference for the correlation model is based on a special type of pseudo-likelihood called pairwise likelihood.

3 Pairwise Likelihood

Pairwise likelihood [4] belongs to the larger family of composite likelihoods [5]. A composite likelihood is a type of pseudo likelihood constructed by compounding marginal or conditional probabilities for subsets of events. In particular, pairwise likelihood consists of the product of bivariate marginal probabilities. The general formula of pairwise likelihood is

$$PL(\theta; y) = \prod_{i=1}^{N-1} \prod_{j=i+1}^N f(y_i, y_j; \theta)^{w_{ij}},$$

where N is the total number of observations and w_{ij} is a weight associated with the bivariate density of y_i and y_j . The logarithm of the pairwise likelihood, $pl(\theta; y) = \log PL(\theta; y)$, is the associated pairwise log-likelihood and its maximum, $\hat{\theta}$, is the

maximum pairwise likelihood estimator. Under usual regularity conditions, the maximum pairwise likelihood estimator is consistent and asymptotically normally distributed with mean θ and covariance matrix $G(\theta) = H(\theta)^{-1}J(\theta)H(\theta)^{-1}$, where $J(\theta) = \text{var} \{ \nabla pl(\theta; Y) \}$ and $H(\theta) = E \{ -\nabla^2 pl(\theta; Y) \}$, see [3]. For the purpose of hypothesis testing and interval estimation, it is possible to define a pairwise likelihood version of the log-likelihood ratio statistics. Suppose that δ is a q -dimensional subvector of the p -dimensional vector $\theta = (\delta', \lambda')'$ and that it is of interest to test $H_0 : \delta = \delta_0$, then the pairwise log-likelihood ratio statistic is

$$CLR(Y) = 2 \left[pl(\hat{\theta}; Y) - pl(\delta_0, \hat{\lambda}(\delta_0); Y) \right],$$

where $\hat{\lambda}(\delta_0)$ denotes the maximum pairwise likelihood estimator in the subspace where $\delta = \delta_0$. The composite log-likelihood ratio statistics has asymptotic distribution given by the weighted sum $\sum_{i=1}^q \xi_i \chi_{i(1)}^2$, where $\chi_{i(1)}^2$ are independent chi-square random variables with 1 degree of freedom and the ξ 's are the eigenvalues of $(H^{\delta\delta})^{-1}G_{\delta\delta}$, where $H^{\delta\delta}$ denotes the block of the inverse of $H(\theta)$ pertaining to δ and $G_{\delta\delta}$ is the block of the matrix $G(\theta)$ pertaining to δ .

Pairwise likelihood can be usefully employed for making inference in the correlation model for paired comparison data. In this case, the pairwise likelihood is the product of the bivariate probabilities of all distinct pairs of comparisons

$$PL(\theta; y) = \prod_{(ij) \neq (kl)} P(Y_{ij} = y_{ij}, Y_{kl} = y_{kl}; \theta)^{w_{ij,kl}},$$

where $w_{ij,kl}$ is the weight associated with the pair of observations (y_{ij}, y_{kl}) . Under the assumptions of normality of the random effects and of probit link function, the corresponding censored random variables Z_{ij} have bivariate normal distributions. Then, the probability that i loses against both j and k is

$$\begin{aligned} P(Y_{ij} = 0, Y_{ik} = 0; \theta) &= P(Z_{ij} < 0, Z_{ik} < 0; \theta) \\ &= \Phi_2 \left(-\frac{(x_i - x_j)'\beta}{\sqrt{1 + 2\sigma^2}}, -\frac{(x_i - x_k)'\beta}{\sqrt{1 + 2\sigma^2}}, \frac{\sigma^2}{1 + 2\sigma^2} \right), \end{aligned}$$

where $\Phi_2(\cdot, \cdot; \rho)$ denotes the cumulative distribution function of a normal random variable with standardised marginals and correlation ρ . The probabilities of the other possible outcomes (win-loss, loss-win and win-win) are

$$\begin{aligned} P(Y_{ij} = 1, Y_{ik} = 0) &= \Phi \left(-\frac{(x_i - x_k)'\beta}{\sqrt{1 + 2\sigma^2}} \right) - \Phi_2 \left(-\frac{(x_i - x_j)'\beta}{\sqrt{1 + 2\sigma^2}}, -\frac{(x_i - x_k)'\beta}{\sqrt{1 + 2\sigma^2}}, \frac{\sigma^2}{1 + 2\sigma^2} \right), \\ P(Y_{ij} = 0, Y_{ik} = 1) &= \Phi \left(-\frac{(x_i - x_j)'\beta}{\sqrt{1 + 2\sigma^2}} \right) - \Phi_2 \left(-\frac{(x_i - x_j)'\beta}{\sqrt{1 + 2\sigma^2}}, -\frac{(x_i - x_k)'\beta}{\sqrt{1 + 2\sigma^2}}, \frac{\sigma^2}{1 + 2\sigma^2} \right), \end{aligned}$$

$$P(Y_{ij} = 1, Y_{ik} = 1) = 1 - \Phi\left(-\frac{(x_i - x_j)'\beta}{\sqrt{1 + 2\sigma^2}}\right) - \Phi\left(-\frac{(x_i - x_k)'\beta}{\sqrt{1 + 2\sigma^2}}\right) \\ + \Phi_2\left(-\frac{(x_i - x_j)'\beta}{\sqrt{1 + 2\sigma^2}}, -\frac{(x_i - x_k)'\beta}{\sqrt{1 + 2\sigma^2}}; \frac{\sigma^2}{1 + 2\sigma^2}\right),$$

where Φ denotes the cumulative distribution function of a standard normal random variable.

4 An Application to Water Polo Data

Sports data are a natural field of application of models for paired comparison data. The data set considered here consists of the results of the water polo matches played by teams in the male A1 league during the 2008/2009 regular season. The water polo tournament has a double round robin structure, so in each half of the season every team competes once against all the other teams in the league. The A1 league includes 12 teams playing altogether 132 matches. At the end of the regular season, the best eight teams access to the playoffs to compete for the title of Italian Champion. The available covariate is the number of accesses to the playoffs in the previous six years. The analysis is focused on determining whether there is a sort of “tradition effect” in water polo, that is whether teams strong in past seasons tend to be strong also in the present one. Water polo matches can also end in ties, hence the model presented so far needs a further extension in order to account for the three possible outcomes of the matches. This extension can be accomplished through the introduction of a threshold parameter τ . Thus, the probability that i loses against j is equal to the probability that the corresponding latent random variable Z_{ij} is smaller than $-\tau$. The probability of a tie between i and j is equal to the probability that the corresponding latent variable Z_{ij} is between $-\tau$ and τ . Finally, the probability that i wins against j is equal to the probability that Z_{ij} is larger than τ . Then, the probability that i loses both the matches against j and k can be computed as

$$P(Z_{ij} < -\tau, Z_{ik} < -\tau) = \Phi_2\left(-\frac{\tau + (x_i - x_j)'\beta}{\sqrt{1 + 2\sigma^2}}, -\frac{\tau + (x_i - x_k)'\beta}{\sqrt{1 + 2\sigma^2}}; \frac{\sigma^2}{1 + 2\sigma^2}\right),$$

while the probability that i and j draws and i loses against k is equal to

$$P(-\tau < Z_{ij} < \tau, Z_{ik} < -\tau) = P(Z_{ij} < \tau, Z_{ik} < -\tau) - P(Z_{ij} < -\tau, Z_{ik} < -\tau) \\ = \Phi_2\left(\frac{\tau - (x_i - x_j)'\beta}{\sqrt{1 + 2\sigma^2}}, -\frac{\tau + (x_i - x_k)'\beta}{\sqrt{1 + 2\sigma^2}}; \frac{\sigma^2}{1 + 2\sigma^2}\right) + \\ - \Phi_2\left(-\frac{\tau + (x_i - x_j)'\beta}{\sqrt{1 + 2\sigma^2}}, -\frac{\tau + (x_i - x_k)'\beta}{\sqrt{1 + 2\sigma^2}}; \frac{\sigma^2}{1 + 2\sigma^2}\right).$$

The probabilities of the other possible outcomes can be similarly computed.

Table 1 Estimates and standard errors of independence (first two columns) and correlation models (last two columns) for the water polo data

	INDEPENDENCE		CORRELATION	
	Est.	SE	Est.	SE
Playoffs	0.238	0.034	0.315	0.090
Home effect	0.223	0.116	0.294	0.129
Threshold	0.218	0.051	0.290	0.066
σ^2	-	-	0.380	0.165

Besides the accesses to the playoffs in the previous six years, the effect of playing at home is taken into account, too. In fact, the team playing at home usually enjoys some advantages over the away team, including the acquaintance with the playing field and usually a larger number of supporters. Table 1 shows the results of the estimates of an independence model, corresponding to the restriction $\sigma^2 = 0$, (first two columns) and the proposed correlation model (last two columns). The standard errors for the correlation model are obtained through parametric bootstrap. Both models show that the team playing at home has actually an advantage over the away team. The estimate of the accesses to the playoffs in the previous six years is significant, denoting that teams which were strong in the recent past tend to be strong also in the present season.

The variance of the random effects in the correlation model is 0.380. It is interesting to test whether the variance component can be null, thus cancelling the correlation between matches with a common player. The test of the hypothesis $H_0 : \sigma^2 = 0$ against $H_1 : \sigma^2 > 0$ is not straightforward since the parameter value under the null hypothesis lies in the boundary of the parameter space. In this case it is convenient to apply parametric bootstrap as in [1]. First, compute the test statistic

$$CLR(y) = 2 [pl(\hat{\theta}(y); y) - pl(\hat{\theta}_0(y); y)],$$

where $\hat{\theta}_0(y)$ is the maximum pairwise likelihood estimator under the null hypothesis. Then, setting $\theta = \hat{\theta}_0(y)$, generate M data sets $y^{(1)}, \dots, y^{(M)}$. For each simulated data set, compute the maximum pairwise likelihood estimator $\hat{\theta}(y^{(m)})$, the maximum pairwise likelihood estimator under the null hypothesis $\hat{\theta}_0(y^{(m)})$ and the relative pairwise log-likelihood ratio statistic $CLR(y^{(m)})$. The p -value of the test is

$$p = \frac{1 + \sum_{m=1}^M I \{ CLR(y^{(m)}) \geq CLR(y) \}}{M + 1}.$$

For the water polo data, the p -value computed from 1000 simulations is smaller than 0.001, therefore the hypothesis of null variance may not be accepted.

5 Discussion

In this paper traditional models for paired comparison data are extended in order to introduce correlation among observations with common objects. In many instances, as for example in sports data, it is apparent that a model which includes correlation is more realistic. In the water polo data analysed here, the presence of correlation between matches with common teams is borne out by the significance of the variance component. Including correlation in the model is important because it has an impact on the standard errors of the regression parameters, and in some cases this may change the significance of a parameter.

The correlation model can be usefully applied also in other areas. For example, in biological studies scientists are interested in determining whether some specific covariates affect the outcomes of contests between animals. In this instance it seems more realistic to include correlation between fights involving the same animal. This consideration may be relevant in order to check which covariates really affect the outcomes of fights.

The model for sports data can be further extended allowing for a temporal evolution of abilities of teams. For example, it is possible to include time-varying covariates which yield different abilities of teams in different matches. An alternative currently under study is the specification of a temporal evolution of the random effects which induce a temporal variation of abilities.

Acknowledgements The authors would like to thank professor David Firth for useful suggestions and discussion.

References

1. Bellio, R., Varin, C.: A pairwise likelihood approach to generalized linear models with crossed random effects. *Stat. Model.* **5**, 217-227 (2005)
2. Bradley, R.A., Terry, M.E.: The rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika* **39**, 324-345 (1952)
3. Cox, D.R., Reid, N.: A note on pseudolikelihood constructed from marginal densities. *Biometrika* **91**, 729-737 (2004)
4. Le Cessie, S., Van Houwelingen, J.C.: Logistic regression for correlated binary data. *Appl. Stat.* **43**, 95-108 (1994)
5. Lindsay, B.G.: Composite likelihood methods. *Contemp. Math.* **80**, 221-239 (1988)
6. Mosteller, F.: Remarks on the method of paired comparisons. I. The least square solution assuming equal standard deviation and equal correlations. *Psychometrika* **16**, 203-206 (1951)
7. Thurstone, L.L.: A law of comparative judgement. *Psychol. Rev.* **34**, 273-286 (1927)