Università di Pisa

Sant'Anna Scuola Universitaria Superiore Pisa

Consiglio Nazionale delle Ricerche

Book of Short Papers SIS 2020

SIS 2020 Pisa

SIS Società Italiana di Statistica

Editors: Alessio Pollice, Nicola Salvati and Francesco Schirripa Spagnolo

Published by Pearson

www.pearson.com

ISBN 9788891910776

Contents

Specialized sessions

Solicited Sessions

Contributed papers and Posters

New algorithms and goodness-of-fit diagnostics for ranked data modelling with the Extended Plackett-Luce distribution

Nuovi algoritmi e diagnostiche di bontà di adattamento per la modellizzazione dei dati di ranking con la distribuzione di Plackett-Luce estesa

Cristina Mollica and Luca Tardella

Abstract The *forward order* assumption postulates that the ranking process of the items is carried out by assigning the positions from the top (most-liked) to the bottom (least-liked) alternative. This assumption has been recently relaxed in the *Extended Plackett-Luce model* (EPL) through the introduction of the discrete *reference order* parameter, describing the rank attribution path. By starting from two formal properties of the EPL, we derive novel diagnostic tools for testing appropriateness of the EPL assumption. We also show how one of the two statistics can be exploited to construct a heuristic method, that surrogates the maximum likelihood approach for inferring the underlying reference order. The performance of the proposals was compared with more conventional approaches through an extensive simulation study.

Abstract *L'ipotesi di ordinamento in avanti postula che il processo di ranking venga eseguito assegnando le posizioni dalla prima all'ultima. Questa ipotesi è stata recentemente rilassata nel modello di Plackett-Luce esteso (EPL) attraverso l'introduzione del parametro discreto detto ordine di riferimento, che descrive il percorso di attribuzione del rango. Partendo da due proprietà formali dell'EPL, deriviamo nuovi strumenti diagnostici per testare l'adeguatezza del modello EPL. Mostriamo anche come una delle due statistiche possa essere impiegata per costruire un metodo euristico, che surroga l'approccio di massima verosimiglianza per l'inferenza sull'ordine di riferimento. La performance delle proposte è stata confrontata con approcci più convenzionali attraverso un ampio studio di simulazione.*

Key words: ranking data, Plackett-Luce model, model specification, goodness-of-fit assessment, bootstrap, heuristic methods

---

Cristina Mollica
Dipartimento MEMOTEF, Sapienza Università di Roma
e-mail: cristina.mollica@uniroma1.it

Luca Tardella
Dipartimento di Scienze Statistiche, Sapienza Università di Roma
e-mail: luca.tardella@uniroma1.it

1 Introduction

Let us consider an experiment in which a sample of N judges is asked to rank a set $I = \{1,\dots,K\}$ of K labeled alternatives, namely *items*, according to a certain criterion. A *ranking* is a vector $\pi = (\pi(1),\dots,\pi(K))$ collecting the ranks assigned to each item, specifically the entry $\pi(i)$ indicates the position attributed to the i-th alternative. Equivalently, data can be recorded in the *ordering* format $\pi^{-1} = (\pi^{-1}(1),\dots,\pi^{-1}(K))$, where the generic component $\pi^{-1}(j)$ indicates the item ranked in the j-th position. Therefore, ranking data are multivariate ordinal data taking values in the set of permutations $\mathscr{S}_K$ of the first K integers.

This work concentrates on the parametric family of stagewise ranking models, in particular on the *Extended Plackett-Luce model* (EPL) introduced by [1]. The EPL generalizes the *Plackett-Luce model* (PL) [2, 3] by relaxing the implicit forward order assumption with the introduction of the *reference order parameter* $\rho = (\rho(1),...,\rho(K))$. It indicates the rank assignment order, i.e., $\rho(t)$ is the position attributed at the stage t. The probability of an ordering π^{-1} under the EPL is

$$\mathbf{P}_{\text{EPL}}(\pi^{-1}|\rho,\underline{p}) = \mathbf{P}_{\text{PL}}(\pi^{-1}\rho|\underline{p}) = \prod_{t=1}^{K} \frac{p_{\pi^{-1}(\rho(t))}}{\sum_{v=t}^{K} p_{\pi^{-1}(\rho(v))}} \qquad \pi^{-1} \in \mathscr{S}_K, \qquad (1)$$

where ρ is the discrete parameter, specifically a permutation of the first K integers, and the positive quantities p_i's are referred to as *support parameters*, proportional to the probabilities for each item to be ranked in the position indicated by the first entry of ρ. Hereinafter, we will shortly refer to (1) as EPL$(\rho,\underline{p})$. Inference on the EPL and its generalization into a finite mixture framework was originally addressed from the Maximum Likelihood Estimation (MLE) perspective in [1], via the hybrid Expectation-Maximization-Minorization (EMM) algorithm. Recently, [4] introduced the Bayesian version of the EPL, where a tuned joint Metropolis-within-Gibbs sampling was developed to conduct approximate posterior inference. The MCMC algorithm was also adapted to infer on the EPL with order constrains on ρ [6] and on the Bayesian EPL mixture [7].

2 Novel EPL diagnostics and comparative evaluation

Specific diagnostic tools to evaluate model adequacy of multistage ranking models are very limited in the ranking literature and their effectiveness has not been deeply explored. The present work aims at providing some contributions in this direction.

Let us suppose that we have some data simulated from an EPL$(\rho,\underline{p})$. We expect the marginal frequencies of the items at the first stage to be ranked according to the order of the corresponding support parameter component. On the other hand, we expect the marginal frequencies of the items at the last stage to be ranked according to the reverse order of the corresponding support parameter component. One can then derive that the ranking of the marginal frequencies of the items corresponding

to the first and last stage should sum up to $(K+1)$, no matter what their support is. Of course, this is less likely to happen when the sample size is small or when the support parameters are not so different of each other. In any case, one can define a test statistic by considering, for each couple of integers (j, j') candidate to represent the first and the last stage ranks, namely $\rho(1)$ and $\rho(K)$, a discrepancy measure $T_{jj'}(\pi)$ between $K+1$ and the sum of the observed ranks of the frequencies corresponding to the same item extracted in the first and in the last stage. Formally, let $\underline{r}_j^{[1]} = (r_{j1}^{[1]}, \ldots, r_{jK}^{[1]})$ and $\underline{r}_{j'}^{[K]} = (r_{j'1}^{[K]}, \ldots, r_{j'K}^{[K]})$ be the marginal item frequency distributions for the j-th and j'-th positions, to be assigned respectively at the first [1] and last [K] stage. In other words, the generic entry $r_{ji}^{[s]}$ is the number of times that item i is ranked j-th at the s-th stage. The proposed EPL diagnostic relies on the following discrepancy

$$T_{jj'}(\pi) = \sum_{i=1}^{K} |\text{rank}(\underline{r}_j^{[1]})_i + \text{rank}(\underline{r}_{j'}^{[K]})_i - (K+1)|, \tag{2}$$

implying that the smaller the value $T_{jj'}(\pi)$, the larger the plausibility that the two integers (j, j') represent the first and the last components of the reference order. In this sense, $T_{jj'}(\pi)$ is a measure of the closeness of the positions j and j' in ρ. To globally assess the conformity of the sample with the EPL, we consider the statistic

$$T_m(\pi) = \min_{j<j'} T_{jj'}(\pi). \tag{3}$$

With the aim at further enlarging the collection of diagnostics for the EPL class, we focus our attention also on another specific property of the EPL, knows as *Luce's choice axiom* or *independence of irrelevant alternatives* (IIA). The IIA hypothesis implies that the probability ratio of selecting item i over item i' is constant over the stages of the ranking process (constant ratio rule), as long as the two items are both still available. Formally, it implies that the expected paired comparison frequency at stage t of choosing item i over item i' is $\tau^*_{ii't} = T_{ii't} p_i/(p_i + p_{i'})$, given by the product between the total number $T_{ii't}$ of pairwise comparisons between i and i' at stage t and the theoretical pairwise comparison probability under the EPL. Hence, a chi-squared statistic for the IIA assumption can be defined as follows

$$X^2_{\text{IIA}} = \sum_{t=1}^{K-1} \sum_{i<i'} \frac{(\tau_{ii't} - \tau^*_{ii't})^2}{\tau^*_{ii't}},$$

where $\tau_{ii't}$ is the observed paired comparison frequency at stage t.

After introducing novel test statistics, one should enquire into their power, for instance through a bootstrap approach, and preferably compare it with that of some standard goodness-of-fit tools for ranking models. To this aim, we conducted a simulation study under alternative model specifications, involving in the comparison the chi-squared discrepancies based on the top frequencies, the pairwise comparisons and the first-order marginals, given respectively by

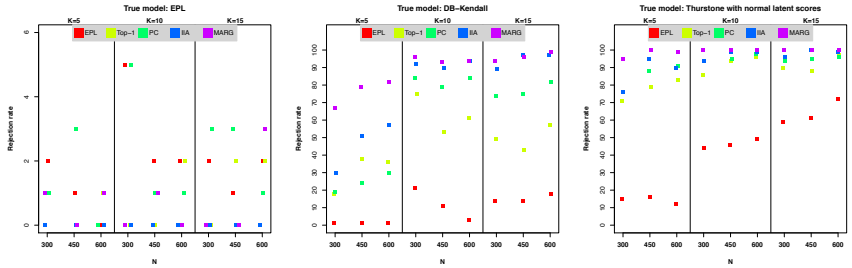

Fig. 1: Rejection rates of the EPL assumption for alternative goodness-of-fit diagnostics computed on simulated data from different model scenarios.

$$X^2_{\text{TOP}} = \sum_{i=1}^{K} \frac{(m_{1i} - m^*_{1i})^2}{m^*_{1i}} \qquad X^2_{\text{PC}} = \sum_{i<i'} \frac{(\tau_{ii'} - \tau^*_{ii'})^2}{\tau^*_{ii'}} \qquad X^2_{\text{M}} = \sum_{j=1}^{K} \sum_{i=1}^{K} \frac{(m_{ji} - m^*_{ji})^2}{m^*_{ji}}$$

where the expected frequencies are $m^*_{1i} = Np_i$ and $\tau^*_{ii'} = N\frac{p_i}{p_i + p_{i'}}$, whereas m^*_{ji} were estimated with a Monte Carlo simulation.

A comparative evaluation of the competing model specification tools was carried out by means of an extensive simulation study. For each possible combination (K,N), with values varying respectively in the grids $K \in \{5,10,15\}$ and $N \in \{300,450,600\}$, we drew 100 datasets with N orderings of K items from the following ranking distributions: i) EPL, ii) distance-based model (DB) with the Kendall metric (DB-Kend), iii) DB with the Hamming distance (DB-Ham) and iv) TH with normal latent scores (TH-norm). where the true parameter values were uniformly generated. To approximate the reference distribution of the test statistics under the EPL assumption, the bootstrap method was applied. Finally, for each model adequacy criterion, we estimated the mis- and correct rejection rates with the relative frequency of the times that the p-value was smaller than or equal to 0.05.

The simulation study revealed a satisfactory performance of all the considered diagnostics regarding the rates of mis-rejections, testified by estimated Type I error probabilities below 0.05 (Figure 1, left). On the other hand, noteworthy differences emerged in terms of the power. Firstly, measure (3) exhibited a consistent poor behavior of the estimated power under each considered model scenario, see for instance the plots in the middle and on the right of Figure 1. At least two motivations can be put forward to argue this evidence. The former is related to the formal definition of $T_m(\pi)$; in fact, (3) is a parameter-free measure based on the ranks of the expected marginal frequencies, rather than on the computation of the parameter-dependent first- and last-stage theoretical probabilities. This makes $T_m(\pi)$ by construction a rougher diagnostic in the comparison with the other statistics. Secondly, the remarkably low power of (3) under the DB with the Kendall metric (Figure 1, center) suggested that the monotonicity property of the first- and last-stage item probabilities is not specific of the EPL, but it is shared by other rankings models too. Another stable evidence highlighted by the comparative analysis concerns the

best-performing diagnostic, which turned to be the one relying on the marginal item distributions. However, it is no less apparent that, for higher values of K and N, the performance of the new IIA statistic is pretty much equivalent to that of the chi-squared based on the marginal distributions and, in general, always better than the remaining competing statistics typically used in the real-data applications.

3 Likelihood-free estimation of the reference order

Rather than from a model specification perspective, in this section we explored the utility of the statistic (3) from the inferential point of view.

Let $\mathbf{T}(\pi) = (T_{jj'}(\pi))$ be the $K \times K$ matrix with entries defined in (2). For each component $T_{jj'}(\pi)$, the inequality $T_{jj'}(\pi) \leq u_K$ holds, where the upper bound u_K corresponds to the constant value in the main diagonal of the matrix, i.e., $u_K = T_{jj}(\pi) = \sum_{l=1}^{K} |2l-(K+1)|$ depending on data only through K. Our heuristic method to estimate the unknown parameter ρ is composed of the following steps:

1. compute $\mathbf{D}(\pi) = |\mathbf{T}(\pi) - u_k \mathbf{J}_K|$, where $\mathbf{J}_K$ is $K \times K$ all-ones matrix, so that each component $D_{jj'}(\pi)$ can be interpreted as a measure of the distance between positions j and j' in the sequential rank assignment process;
2. use the matrix $\mathbf{D}(\pi)$ as the input of a Principal Component Analysis (PCA);
3. estimate ρ by taking the non-decreasing ordering of the scores $(s_1, \ldots, s_K)$ of the K positions on the first PC.

The inferential effectiveness of the proposal to recover the true discrete parameter was explored by means of a simulation study. For each possible combination (K,N), where $K \in \{5,10,15\}$ and $N \in \{50,200,1000,10000\}$, we drew 100 datasets $\underline{\pi}_{(R)}^{-1}$ with $R = 1, \ldots, 100$ from the EPL with uniformly generated parameters. For comparison purposes, we inferred the reference order of each simulated sample $\underline{\pi}_{(R)}^{-1}$ with the heuristic strategy described above, with the one replacing the PCA with the Multidimensional Scaling (MDS) and with the MLE approach [1], which is considered as the reference method for the present estimation task. Finally, the estimation performance of the competing strategies were compared in terms of: i) percentage of matching between $\dot{\rho}^{(R)}$ and $\hat{\rho}^{(R)}$ (% recoveries) and ii) average cograduation $\bar{r}_{\text{Spear}}(\dot{\rho}, \hat{\rho})$ between the estimated and the actual reference order computed with Spearman's rank correlation.

As apparent in Table 1, PCA and MDS exhibited essentially the same ability. Compared with the MLE, one can appreciate very good results for the heuristic methods. The percentage of matching consistently grows with N and, by checking also the cases where there is not an exact correspondence, on average an analogous trend is highlighted for the relative Spearman correlation. Additionally, if we look at a fixed N, the percentage of recoveries shows a worse tendency for larger values of K. In this regard, the cases $K \in \{10,15\}$ combined with a relatively very low ($N = 50$) and very high ($N = 10000$) sample size deserve some considerations to stress typical issues which can be encountered in a ranking data analysis. First, in a

Table 1: Inferential performance of the heuristic methods via PCA and MDS to estimate the reference order on simulated data compared to the MLE.

	% recoveries			$\bar{r}_{\text{Spear}}(\hat{\rho},\hat{\rho})$		
(K,N)	PCA	MDS	MLE	PCA	MDS	MLE
$(5,50)$	58	60	45	0.86	0.86	0.69
$(5,200)$	79	80	77	0.89	0.89	0.95
$(5,1000)$	91	90	100	0.92	0.93	1.00
$(5,10000)$	97	97		0.94	0.94	
$(10,50)$	3	5	3	0.90	0.89	0.87
$(10,200)$	14	16	23	0.93	0.93	0.98
$(10,1000)$	55	54	68	0.96	0.95	0.99
$(10,10000)$	79	78		0.96	0.96	
$(15,50)$	0	0	0	0.92	0.92	0.91
$(15,200)$	0	1	3	0.94	0.94	0.97
$(15,1000)$	13	15	26	0.97	0.97	0.99
$(15,10000)$	48	51		0.97	0.97	

sparse data situation, all of the estimation techniques exhibit a great uncertainty in recovering the actual ρ, testified by the negligible values of the recovery percentage. On the other hand, although a better behavior of the MLE is expected for $N = 10000$, this has not been implemented since, without a specialized program, fitting the EPL to a large sample can be deeply computational demanding if not actually unfeasible. Moreover, the computational burden is further aggravated by the multiple initialization needed to address the issue of local maxima. In the light of these remarks, the likelihood-free approach can be motivated as a straightforward method that can be combined with the MLE or with an MCMC method in the Bayesian framework. In fact, without computational costs, it can implemented as a preliminary step to obtain a promising initialization, that can guide the parameter space exploration towards the global optimum and substantially reduce the elaboration time.

References

1. Mollica, C., Tardella, L.: Epitope profiling via mixture modeling of ranked data. Stat Med. **33**(21), 3738–3758 (2014)
2. Luce, R.D.: Individual choice behavior: A theoretical analysis. John Wiley & Sons Inc (1959)
3. Plackett, R.L.: The analysis of permutations. J Royal Stat Soc C. **24**(2), 193–202 (1975)
4. Mollica, C., Tardella L.: Bayesian analysis of ranking data with the Extended Plackett-Luce model. Statistical Methods and Applications. doi: 10.1007/s10260-020-00519-5 (2020)
5. Mollica, C., Tardella L.: Bayesian mixture of Plackett-Luce models for partially ranked data. Psychometrika. **82**(2), 442–458 (2017)
6. Mollica, C., Tardella, L.: Constrained Extended Plackett-Luce model for the analysis of preference rankings. In: Book of Short Papers SIS 2018, (eds.) Springer Italia, pp 480-486, ISBN: 9788891910233 (2018)
7. Mollica, C., Tardella, L. Modelling unobserved heterogeneity of ranking data with the Bayesian mixture of Extended Plackett-Luce models. In: Book of Short Papers CLADAG 2019, pp. 346-349, ISBN: 978-88-8317-108-6 (2019)