EXTENDING PARAMETRIC MODELS FOR RANKED DATA

CRISTINA MOLLICA
ADVISOR: PROF. LUCA TARDELLA

This thesis develops some original extensions of a widely used parametric distribution for random rankings.

Ranked data arise in several research fields, specifically in those experiments where a sample of N people is asked to rank a finite set of K items according to certain criteria, typically their personal preferences or attitudes. Interest in the analysis of ranking data is motivated, for example, by marketing and political surveys or psychological and behavioral studies. Another typical context is sports, where teams or players compete and the final outcome is a ranking among competitors. Formally, a *full* (or *complete*) *ranking* $\pi = (\pi(1), \ldots, \pi(K))$ is a sequence in which the generic component $\pi(i)$ must be read as the rank attributed to the i-th item. The set of all possible rankings is identified with the symmetric group $\mathcal{S}_K$ of permutations, whose fast-growing dimension $K!$ leads to the need of simplifying assumptions on the ranking process and to the wide assortment of restricted parametric models in the rank data theory (Marden, 1995). Following Critchlow et al. (1991), we review the basic approaches developed in the literature to construct non-uniform models, which can be classified in four main categories: (i) order statistics models, (ii) models based on paired comparisons, (iii) distance-based models and (iv) stagewise models. The main focus of the thesis is on a parametric distribution belonging to the last class, the *Plackett-Luce model* (PL), arisen from independent work by Luce (1959) and Plackett (1975). It is a very popular parametric model indexed by the vector $\underline{\mathrm{p}} = (p_1, \ldots, p_K)$ of *item support parameters*: a higher value of the support parameter p_i implies a greater probability for item i to be preferred at each selection level.

Probability distributions based on sequential construction of the ranking, such as the multistage ranking models introduced by Fligner and Verducci (1988), the PL and related extensions, implicitly suppose that preferences are expressed with the canonical *forward procedure*, meaning that the judge proceeds from the elicitation of their best choice (rank 1) up to the worst one (rank K). To our knowledge, the rank

assignment order has not received explicit consideration in a model setup aimed at improving the description of ranked data, although any other order for the rank assignment process is admissible and potentially leads to different results (Fligner and Verducci, 1988). This aspect has inspired us to relax the conventional forward assumption and to extend the PL in the following way: rather than fixing *a priori* the stepwise order leading the judge to their final ranked sequence, we represent it with a specific free parameter $\rho \in \mathcal{S}_K$ in the model and let data guide inference about the reference order followed in the rank assignment scheme. We named the novel generalization *Extended Plackett-Luce model* (EPL) and employed a representative property of the PL to formally prove the actual greater flexibility of the proposal in covering a proper wider class of probability ranking distributions. We solved the inferential issue of maximizing the likelihood function over a parameter space which turns out to be of mixed-type due to the discreteness of the reference order ρ, adapting and combining different estimation devices for ranking models. For the support parameters $\underline{\text{p}}$ we applied the *Minorization/Maximization* (MM) algorithm illustrated in Hunter (2004), which is an iterative optimization technique based on the replacement of the original objective function with a more tractable minorizing surrogate function. For the reference order, instead, we implemented a local search in the discrete space $\mathcal{S}_K$ similarly to the method suggested by Busse et al. (2007) and Lee and Yu (2010), constraining the optimization within a fixed distance from the current estimate of ρ. To evaluate the sensitivity of the algorithm w.r.t. the choice of a particular distance in the local search step, we focused on two frequently used metrics for rankings and compared the corresponding estimation performances. In order to address the common situation of unobserved sample heterogeneity and increase the applicability of the EPL, we also considered the natural extension of the novel model in the mixture modeling setting. In this framework the likelihood maximization required the derivation of a hybrid procedure, called EMM algorithm, which integrates the standard EM with the above MM algorithm in a similar manner as Gormley and Murphy (2006).

All above estimation procedures have been implemented developing original code in R language. We first tested the computational effectiveness and efficiency of the algorithm performing it in multiple simulation scenarios and subsequently verified the practical utility of the EPL with an application to the Large Fragment Phage Display (LFPD) real data set. This data set comes from a bioassay experiment

and collects the binding measurements of human blood exposed to $K = 11$ partially overlapping fragments of the HER2 oncoprotein. Raw quantitative outcomes have been obtained from $N = 67$ samples of human blood taken from three different disease groups: healthy patients, patients diagnosed with breast cancer at an early stage and patients diagnosed with metastatic breast cancer. For reasons due to the numerical instability of measurements and the absence of universally accepted methods of rescaling the original data, we were interested in verifying the possible usefulness of the underlying ordinal information as a more robust and unambiguously defined evidence of the sample heterogeneity. Specifically, we addressed the heterogeneous nature of experimental units via model-based clustering and compared the performance of the mixture model using the new distribution as mixture components with alternative mixture models for random rankings. BIC values for the mixture of EPL turned out to be significantly smaller than those of the competitor models. This indicated the EPL as the best model and proved the successful introduction of the discrete parameter ρ, which drastically improved the data fitting. Moreover, the mixture of EPL exhibited a very good accuracy in the discrimination of sample units w.r.t. the real disease status. It follows that, besides the methodological contribution to ranked data modeling, our work suggests the analysis of ranking data as a promising tool in an epitope-mapping experiment, allowing to partially overcome difficulties related to the preliminary choice of a normalization method.

The successful application of the EPL encouraged us to explore further PL generalizations in different directions and implement inference on them also within the Bayesian approach. These ideas include the possibility to combine the novel EPL with the popular *Benter model* (BM) described in Benter (1994). It generalizes the PL introducing additional *dampening parameters* to account for a variable selection accuracy over the stages of the ranking process. This means that the EPL and the BM move from substantially different but compatible attributes of the ranking procedure and their merging can add further flexibility to the PL. Another proposal contemplates the extension of the Bayesian device recently introduced by Caron and Doucet (2012) for the inference on the PL to the finite mixture context. We describe an efficient way to incorporate the latent group structure in their data augmentation approach and how to interpret previous maximum likelihood procedures as special instances of the proposed Bayesian analysis.

References

William Benter. Computer based horse race handicapping and wagering systems: A report. In Donald B. Hausch, Victor S.Y. Lo, and William T. Ziemba, editors, *Efficiency of Racetrack Betting markets*, pages 183–198. Academic Press, 1994.

Ludwig M. Busse, Peter Orbanz, and Joachim M. Buhmann. Cluster analysis of heterogeneous rank data. In Zoubin Ghahramani, editor, *Proceedings of the 24th International Conference on Machine Learning – ICML 2007*, pages 113–120. Omnipress, 2007.

François Caron and Arnaud Doucet. Efficient bayesian inference for generalized bradley-terry models. *J. Comput. Graph. Statist.*, 21(1):174–196, 2012. ISSN 1061-8600. doi: 10.1080/10618600.2012.638220.

Douglas E. Critchlow, Michael A. Fligner, and Joseph S. Verducci. Probability models on rankings. *J. Math. Psych.*, 35(3):294–318, 1991. ISSN 0022-2496. doi: 10.1016/0022-2496(91)90050-4.

Michael A. Fligner and Joseph Stephen Verducci. Multistage ranking models. *J. Amer. Statist. Assoc.*, 83(403):892–901, 1988. ISSN 0162-1459.

Isobel Claire Gormley and Thomas Brendan Murphy. Analysis of irish third-level college applications data. *J. Roy. Statist. Soc. Ser. A*, 169(2):361–379, 2006. ISSN 0964-1998. doi: 10.1111/j.1467-985X.2006.00412.x.

David R. Hunter. Mm algorithms for generalized bradley-terry models. *Ann. Statist.*, 32(1):384–406, 2004. ISSN 0090-5364. doi: 10.1214/aos/1079120141.

Paul H. Lee and Philip L. H. Yu. Distance-based tree models for ranking data. *Comput. Statist. Data Anal.*, 54(6):1672–1682, 2010. ISSN 0167-9473. doi: 10.1016/j.csda.2010.01.027.

R. D. Luce. *Individual choice behavior: A theoretical analysis*. John Wiley & Sons Inc., 1959.

John I. Marden. *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, 1995. ISBN 0-412-99521-2.

Robin L. Plackett. The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, 24(2):193–202, 1975. ISSN 0035-9254.