Efficient estimation of finite mixtures of Mallows models with the Spearman distance

Marta Crispino [1] Cristina Mollica [2] Valerio Astuti [1] Luca Tardella [2]

[1]Dipartimento di Economia e Statistica, Banca d'Italia

[2]Dipartimento di Scienze Statistiche, Sapienza Università di Roma
cristina.mollica@uniroma1.it

Dodicesima Giornata della Ricerca MEMOTEF

Piazza dei Cavalieri di Malta 2, Rome, Italy

31 May – 1 June 2022

Ranking data

Ranking data are common in contexts where

- the phenomenon cannot be measured in objective and precise manner
- experiment: N subjects rank n items according to a certain criterion

Examples of research fields requiring rank data analysis:

- social and behavioral sciences
  - preference studies (items = degree courses or jobs)
  - marketing surveys (items = consumer goods)
  - political/election studies (items = political candidates or goals)
  - psychological studies (items = words or topics)
- sport/racing contexts
  - national soccer championships (items = soccer teams)
  - horse or car races (items = horses or cars).

What is a ranking?

A **complete** (or **full**) **ranking** is a bijective mapping

$$\pi : I \to R$$

- $I = \{1, \ldots, n\}$ is the set of labeled **items**
- $R = \{1, \ldots, n\}$ is the set of **ranks**
- $n =$ number of items to be ranked.

$$\pi = (\pi(1), \ldots, \pi(n))$$
$$\downarrow$$
$$\pi(i) = \text{rank attributed to the } i\text{-th item}$$

Example: $\pi = (3, 5, 2, 1, 4) \Leftrightarrow$ *Item 1 ranked 3rd, Item 2 ranked 5th...*

The **ranking space**

$\mathcal{P}_n$ = set of all $n!$ *permutations* + *composition operation* $\circ$

$$\pi\sigma^{-1} = \pi \circ \sigma^{-1} = (\pi(\sigma^{-1}(1)), \ldots, \pi(\sigma^{-1}(n)))$$

Marden. 1995. *Analyzing and modeling rank data*, Monographs on Statistics and Applied Probability, vol. 64, Chapman & Hall, London.

Distance-based models

Features of the **Mallows models** (MMs):

- their paternity is attributed to Mallows (1957)
- they represent exponential families for random permutations
- based on the notion of metric between rankings

$$\mathbb{P}(\boldsymbol{r}\,|\boldsymbol{\rho},\theta) = \frac{e^{-\theta\, d(\boldsymbol{r},\boldsymbol{\rho})}}{Z(\theta,\boldsymbol{\rho})} \qquad \boldsymbol{r} \in \mathcal{P}_n$$

- $\boldsymbol{\rho} \in \mathcal{P}_n$ is the consensus ranking
- $\theta \in \mathbb{R}_0^+$ is the concentration parameter
- $d(\cdot,\cdot)$ is a distance over $\mathcal{P}_n$
- $Z(\theta,\boldsymbol{\rho}) = \sum_{\boldsymbol{r}\in\mathcal{P}_n} e^{-\theta\, d(\boldsymbol{r},\boldsymbol{\rho})}$ is the normalizing constant

---

Mallows. 1957. *Non-Null Ranking Models. I*, Biometrika **44**, no. 1/2, 114–130.

Metrics for rankings

Some of the most popular metrics for rankings are

- the Kendall distance $d_K(\boldsymbol{r}, \boldsymbol{\rho}) = \sum\sum_{1 \leq i < i' \leq n} I_{[(r(i)-r(i'))(\rho(i)-\rho(i'))<0]}$
- the Cayley distance $d_C(\boldsymbol{r}, \boldsymbol{\rho})$ corresponding to the minimum number of transpositions needed to transform $\boldsymbol{r}^{-1}$ into $\boldsymbol{\rho}^{-1}$
- the Hamming distance $d_H(\boldsymbol{r}, \boldsymbol{\rho}) = \#\{i = 1, \ldots, n : r(i) \neq \rho(i)\}$
- the Spearman distance

$$d_S(\boldsymbol{r}, \boldsymbol{\rho}) = \sum_{i=1}^{n} (r_i - \rho_i)^2$$

Properties:

1. all metrics are right-invariant $\implies$ $Z(\theta, \boldsymbol{\rho}) = Z(\theta)$
2. only some distances are decomposable $\implies$ closed-form for $Z(\theta)$

☹ Property 2 does not hold for the Spearman distance ☹

Diaconis. 1988. *Group representations in probability and statistics*, Institute of Mathematical Statistics, vol. 11, Chapman & Hall, Hayward, CA, USA.

MM with the Spearman distance (MMS)

The MMS can be written as

$$\mathbb{P}(\boldsymbol{r} \,|\, \boldsymbol{\rho}, \theta) = \frac{e^{-2\theta\left(c_n - \boldsymbol{\rho}'\boldsymbol{r}\right)}}{Z(\theta)} \qquad \boldsymbol{r} \in \mathcal{P}_n$$

where $\boldsymbol{e} = (1, 2, ..., n)$ and $c_n = n(n+1)(2n+1)/6$.

Remarks:

- it is also known as θ-model
- it is the analogue of the Gaussian distribution over $\mathcal{P}_n$
- importantly, Feigin and Cohen (1978) pointed out that

😄 The MMS admits a closed-form for the MLE of ρ 😄

$$\Downarrow$$

$$\hat{\boldsymbol{\rho}} = (\hat{\rho}_1, \ldots, \hat{\rho}_i, \ldots, \hat{\rho}_n) \quad \text{with} \quad \hat{\rho}_i = \text{rank}(\bar{r}_i) \text{ in } \{\bar{r}_1, \ldots, \bar{r}_n\},$$

Feigin and Cohen. 1978. *On a Model for Concordance Between Judges*, Journal of the Royal Statistical Society. Series B (Methodological) **40**, no. 2, 203–213.

MLE of MMS mixtures via EM

To account for unobserved sample heterogeneity, we assume

$$\mathbb{P}(\boldsymbol{r}|\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\omega}) = \sum_{g=1}^{G} \omega_g \mathbb{P}(\boldsymbol{r}\,|\boldsymbol{\rho}_g,\theta_g) = \sum_{g=1}^{G} \omega_g \frac{e^{-2\theta_g\left(c_n-\boldsymbol{\rho}_g'\boldsymbol{r}\right)}}{Z(\theta_g)}$$

We conducted MLE with the EM algorithm by extending the approach by Beckett (1993) for partial rankings.

- N_l is the frequency of the observed partial sequence $\boldsymbol{r}_l$ where only a subset $\mathcal{I}_l \subseteq \{1,2,\ldots,n\}$ of $n_l = |\mathcal{I}_l|$ items are actually ranked
- $\mathcal{C}(\boldsymbol{r}_l) \subset \mathcal{P}_n$ is the set of full rankings which are compatible with $\boldsymbol{r}_l$
- $\boldsymbol{r}_m^* \in \mathcal{C}(\boldsymbol{r}_l)$ is a generic full ranking compatible with $\boldsymbol{r}_l$

The complete-data log-likelihood of the G-component MMS mixture is

$$\ell_c(\boldsymbol{\rho},\boldsymbol{\theta},\boldsymbol{\omega},\boldsymbol{z},\underline{\boldsymbol{r}}^*) = \sum_{m=1}^{M}\sum_{g=1}^{G} N_m z_{mg}\left(\log\omega_g - 2\theta_g\left(c_n - \boldsymbol{\rho}_g'\boldsymbol{r}_m^*\right) - \log Z(\theta_g)\right)$$

Beckett. 1993. *Maximum Likelihood Estimation in Mallows's Model Using Partially Ranked Data*, Probability Models and Statistical Analyses for Ranking Data, 92–107.

E-step

For $m = 1, \dots, M$ and $g = 1, \dots, G$, at iteration (t+1) compute

$$\hat{N}_m^{(t+1)} = \sum_{l:\, \boldsymbol{r}_m^* \in \mathcal{C}(\boldsymbol{r}_l)} N_l \hat{p}_{lm}^{(t)}$$

$$\hat{z}_{mg}^{(t+1)} = \frac{\omega_g^{(t)} \mathbb{P}\left(\boldsymbol{r}_m^* | \boldsymbol{\rho}_g^{(t)}, \theta_g^{(t)}\right)}{\sum_{g'=1}^G \omega_{g'}^{(t)} \mathbb{P}\left(\boldsymbol{r}_m^* | \boldsymbol{\rho}_{g'}^{(t)}, \theta_{g'}^{(t)}\right)}$$

where

$$\hat{p}_{lm}^{(t)} = \mathbb{P}(\boldsymbol{r}_m^* | \boldsymbol{r}_l, \boldsymbol{\rho}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\omega}^{(t)}) = \frac{\sum_{g=1}^G \omega_g^{(t)} e^{-2\theta_g^{(t)}\left(c_n - \boldsymbol{\rho}_g^{\prime(t)} \boldsymbol{r}_m^*\right) - \log Z\left(\theta_g^{(t)}\right)}}{\sum_{\boldsymbol{s}^* \in \mathcal{C}(\boldsymbol{r}_l)} \sum_{g=1}^G \omega_g^{(t)} e^{-2\theta_g^{(t)}\left(c_n - \boldsymbol{\rho}_g^{\prime(t)} \boldsymbol{s}^*\right) - \log Z\left(\theta_g^{(t)}\right)}}$$

M-step

By setting $\hat{N}_g^{(t+1)} = \sum_{m=1}^{M} \hat{N}_m^{(t+1)} \hat{z}_{mg}^{(t+1)}$, for $g = 1, \ldots, G$ compute

$$\omega_g^{(t+1)} = \frac{\hat{N}_g^{(t+1)}}{N}$$

$$\boldsymbol{\rho}_g^{(t+1)} : \quad \rho_{gi}^{(t+1)} = \text{rank}\left(\bar{r}_{gi}^{*(t+1)}\right)$$

$$\theta_g^{(t+1)} : \quad \text{E}_{\theta_g}[D_S] = 2\left(c_n - \boldsymbol{\rho}_g'^{(t+1)} \bar{\boldsymbol{r}}_g^{*(t+1)}\right)$$

where $\bar{r}_{gi}^{*(t+1)} = \frac{\sum_{m=1}^{M} \hat{N}_m^{(t+1)} \hat{z}_{mg}^{(t+1)} r_{mi}^*}{\hat{N}_g^{(t+1)}}$ and $\text{E}_{\theta_g}[D_S] = \frac{\sum_{d \in \mathcal{D}_n} dN_d \, e^{-d\theta_g}}{\sum_{d \in \mathcal{D}_n} N_d \, e^{-d\theta_g}}$ with

$$\mathcal{D}_n = \left\{2n \,:\, n \in \mathbb{N}_0 \text{ and } 0 \leq d \leq 2\binom{n+1}{3}\right\}$$

$$N_d = |\{\boldsymbol{r}^* \in \mathcal{P}_n \,:\, d(\boldsymbol{r}^*, \boldsymbol{e}) = d\}|$$

☺ Novel approximation of N_d for $n \geq 15$ ☺

Application to the Reading Genres dataset

2. Reading Genres data (top-5 rankings): $N = 507$ people ranked $K = 11$ reading genres in order of preference

1. Classic 2. Novel 3. Thrillers 4. Fantasy 5. Biography

6. Teenage 7. Horror 8. Comics 9. Poetry 10. Essay 11. Humor

- brand new data from a survey conducted in Italy in 2019
- estimation of G-component MMS-mixture with $G = 1, \ldots, 5$

$G = 1$	$G = 2$	$G = 3$	$G = 4$	$G = 5$
16431.64	16001.46	**15930.49**	15904.41	15930.21

Table. BIC values of the MMS-mix fitted to the Reading Genres data.

Application to the Reading Genres dataset

	Group 1	Group 2	Group 3
ω	0.42	0.07	0.51
θ	0.048	0.036	0.038
Rank 1	Novel	Fantasy	Novel
Rank 2	Classic	Comics	Thrillers
Rank 3	Thrillers	Teenage	Fantasy
Rank 4	Essay	Humor	Classic
Rank 5	Biography	Classic	Teenage
Rank 6	Poetry	Horror	Horror
Rank 7	Fantasy	Novel	Biography
Rank 8	Comics	Thrillers	Comics
Rank 9	Humor	Essay	Poetry
Rank 10	Horror	Biography	Essay
Rank 11	Teenage	Poetry	Humor

Conclusions and future work

In conclusion...

- existence of a closed-form for the MLE of ρ for the MMS
- MLE of MMS mixtures via an efficient EM algorithm
- extension via data augmentation for various forms of partial rankings
- novel approximation of the Spearman distance distribution for large n
- successful application to real datasets

For the future...

- construction of a novel R package for mixtures of MMSs
- inclusion of individual and/or item-specific covariates