

PhD THESIS DECLARATION

The undersigned

Marta Crispino

PhD Registration Number: 1502574

Thesis title:
Bayesian Learning of the Mallows rank model

PhD in Statistics

29th Cycle

External Advisor: Professor Arnaldo Frigessi

External Co-advisor: Professor Elja Arjas

Internal Advisor: Professor Sonia Petrone

Year of Discussion: 2018

DECLARES

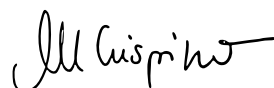
Under her responsibility:

- 1) that, according to Italian Republic Presidential Decree no. 445, 28th December 2000, mendacious declarations, falsifying records and the use of false records are punishable under the Italian penal code and related special laws. Should any of the above prove true, all benefits included in this declaration and those of the temporary embargo are automatically forfeited from the beginning;
- 2) that the University has the obligation, according to art. 6, par. 11, Ministerial Decree no. 224, 30th April 1999, to keep a copy of the thesis on deposit at the "Biblioteche Nazionali Centrali" (Italian National Libraries) in Rome and Florence, where consultation will be permitted, unless there is a temporary embargo protecting the rights of external bodies and the industrial/commercial exploitation of the thesis;

- 3) that the Bocconi Library will file the thesis in its “Archivio istituzionale ad accesso aperto” (institutional registry) which permits online consultation of the complete text (except in cases of a temporary embargo);
- 4) that, in order to file the thesis at the Bocconi Library, the University requires that the thesis be submitted online by the candidate in unalterable format to Società NORMADEC (acting on behalf of the University), and that NORMADEC will indicate in each footnote the following information:
 - thesis: Bayesian Learning of the Mallows rank model;
 - by Marta Crispino;
 - defended at Università Commerciale “Luigi Bocconi” – Milano in 2018;
 - the thesis is protected by the regulations governing copyright (Italian law no. 633, 22th April 1941 and subsequent modifications). The exception is the right of Università Commerciale “Luigi Bocconi” to reproduce the same for research and teaching purposes, quoting the source;
- 5) that the copy of the thesis submitted online to NORMADEC is identical to the copies handed in/sent to the members of the Thesis Board and to any other paper or digital copy deposited at the University offices, and, as a consequence, the University is absolved from any responsibility regarding errors, inaccuracy or omissions in the contents of the thesis;
- 6) that the contents and organization of the thesis is an original work carried out by the undersigned and does not in any way compromise the rights of third parties (Italian law, no. 633, 22nd April 1941 and subsequent integrations and modifications), including those regarding security of personal details; therefore the University is in any case absolved from any responsibility whatsoever, civil, administrative or penal, and shall be exempt from any requests or claims from third parties;
- 7) that the thesis meets one of the temporary embargo hypotheses included in the declaration “TEMPORARY EMBARGO REQUEST OF THE PhD THESIS” undersigned elsewhere.

November 22, 2017

Marta Crispino



Tesi di dottorato "Bayesian learning of the Mallows ranking model"

di CRISPINO MARTÀ

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2018

La tesi è tutelata dalla normativa sul diritto d'autore(Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.

Tesi di dottorato "Bayesian learning of the Mallows ranking model"

di CRISPINO MARTÀ

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2018

La tesi è tutelata dalla normativa sul diritto d'autore(Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.

I don't estimate probabilities, I provide them.

Elja Arjas

The only relevant thing is uncertainty - the extent of our knowledge and ignorance. The actual fact of whether or not the events considered are in some sense determined, or known by other people, and so on, is of no consequence.

Bruno de Finetti

There is no subject so old that something new cannot be said about it.

Fyodor Dostoevsky

Tesi di dottorato "Bayesian learning of the Mallows ranking model"

di CRISPINO MARTÀ

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2018

La tesi è tutelata dalla normativa sul diritto d'autore(Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.

Contents

Introduction	1
Motivating application	3
Pair comparison experiment and data	5
Outline of thesis	7
1 Background	9
1.1 Probabilistic models on ranking data	10
1.1.1 Thurstonian order statistics models	11
1.1.2 Ranking models induced by pairwise comparisons	13
1.1.3 Distance based ranking models: the Mallows model	14
1.1.4 Multistage ranking models: the Plackett-Luce model	20
1.2 Probabilistic models on pair comparison data	21
1.2.1 Non-transitivity	23
1.3 Machine learning approaches for ranking data	27
2 The Bayesian Mallows model for ranking data	29
2.1 The Bayesian Mallows model for full rankings	31
2.1.1 Prior distributions	31
2.1.2 Inference	33
2.1.3 Metropolis-Hastings algorithm for full rankings	33
2.1.4 Tuning the proposal distributions parameters	36
2.2 Approximating the partition function $Z_n(\alpha)$	37
2.2.1 Exact formula for footrule and Spearman distances	39
2.2.2 Off-line importance sampling, IS, for $Z_n(\alpha)$	39
2.2.3 Comparisons with other methods	51
2.3 Extensions to partial rankings and heterogeneous assessors	54
2.3.1 Ranking of the top ranked items	54
2.3.2 Pairwise comparisons	60
2.3.3 Clustering assessors giving full rankings	61
2.3.4 Clustering assessors giving pairwise comparisons	62

2.3.5	Example: preference prediction	63
2.4	Experiments	67
2.4.1	Meta-analysis of differential gene expression	67
2.4.2	Beach preference data	71
2.4.3	Sushi data	73
2.4.4	Movielens data	76
2.5	Discussion	79
Appendices		81
2.A	Pseudo-codes of the algorithms	81
2.B	Sampling from the Mallows model	81
3	The Bayesian Mallows model for non-transitive pair comparisons	85
3.1	The main model	87
3.1.1	Bernoulli model (BM) for mistakes	89
3.1.2	Logistic model (LM) for mistakes	91
3.1.3	Logistic-consensus model (LCM) for mistakes	92
3.1.4	Mixture model on θ	93
3.1.5	Mixture model on α and ρ	95
3.2	The MCMC algorithm for non-transitive pairwise preferences	97
3.3	Simulation study	101
3.3.1	Simulations with Bernoulli mistake model	102
3.3.2	Simulations with logistic mistake model	105
3.3.3	Ability to detect mistakes	108
3.4	Examples	112
3.4.1	Beach preference data revisited	112
3.4.2	Movie survey	115
3.5	Discussion	117
Appendices		119
3.A	Algorithms	119
3.B	Sample simulated data from the Mallows model with mistakes	123
3.C	A generalized version of the Bradley Terry model	127
4	An application to Electroacoustic music data	137
4.1	Introduction	138
4.2	Background studies	139
4.3	Aim and method	141
4.3.1	Creating the test stimuli	142
4.3.2	Test procedure	146

4.4	Results	148
4.5	Discussion	154
5	The Bayesian Mallows model with Cayley distance	157
5.1	The Cayley distance and its properties	157
5.2	The Mallows model with Cayley distance	159
5.2.1	Bayesian Learning of the MMC	160
5.2.2	Algorithm for the MMC	162
5.3	Ongoing work and discussion	164
6	Preliminary results on the conjugate prior elicitation problem	165
6.1	Sufficient statistic and MLE	165
6.2	The conjugate prior for ρ when θ is known	167
	Discussion	171
	Future work	174
	Bibliography	175

Tesi di dottorato "Bayesian learning of the Mallows ranking model"

di CRISPINO MARTÀ

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2018

La tesi è tutelata dalla normativa sul diritto d'autore(Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.

List of Figures

1	A scheme of the process generating sounds.	5
1.1	The Mallows density for the six right-invariant distances, for $n = 5$	16
1.2	Directed graph representing the set of preferences	26
2.1	Simulated data: Plot of the acceptance probability and marginal IAT. . . .	37
2.2	Simulated data: Ratio of the IS approximate partition function to the exact. . . .	41
2.3	Simulated data: A comparison among different approaches to compute the partition function.	44
2.4	Simulated data: Posterior density of α , for varying N , and for different choices of the prior. $n = 20$	45
2.5	Simulated data: Posterior CDF of $d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ for varying N , and for different choices of the prior. $n = 20$	46
2.6	Simulated data: Posterior density of α and posterior CDF of $d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$, for varying N , and for different choices of the prior. $n = 50$	47
2.7	Simulated data: Heatplot of the posterior marginal density of each item in the consensus. $n = 50$	48
2.8	Simulated data: Posterior density of α and posterior CDF of $d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$, for varying N , and for different choices of the prior. $n = 100$	49
2.9	Simulated data: Heatplot of the posterior marginal density of each item in the consensus. $n = 100$	49
2.10	Simulated data: Boxplots of the within-cluster sum of footrule distances, and of the within-cluster indicator of mis-fit to the data.	65
2.11	Simulated data: Barplots of the frequency of successes and failures for predictions; $\lambda_T = 20$	66
2.12	Simulated data: Barplots of the frequency of successes and failures for predictions; $\lambda_T = 10$	67
2.13	Meta Analysis: Heatplot of the marginal posterior probabilities, for the 89 genes, for being ranked as the k -th most preferred.	69
2.14	Beach preference data: The 15 images used for producing the Beach dataset.	72

2.15	Beach preference data: Posterior marginal probability, for each beach, of being ranked among the top-3 in the consensus, and in the individual rankings.	73
2.16	Sushi data: Boxplots of the within-cluster sum of footrule distances.	74
2.17	Sushi data: Heatplot of the posterior probabilities for all assessors of being assigned to each cluster.	75
2.18	Movielens data: Boxplots of the within-cluster indicator of mis-fit to the data.	77
2.19	Movielens data: Boxplots of the posterior probability for correctly predicting the discarded preference conditionally on the number of preferences stated by the user.	78
3.1	Graphical representation of the Bernoulli model for mistakes.	90
3.2	Graphical representation of the mixture model on θ	95
3.3	Graphical representation of the mixture model on $(\alpha, \boldsymbol{\rho})$	97
3.1	Simulated data: Boxplots of the average distance of the sampled rankings to the consensus.	101
3.2	BM simulated data: Posterior CDFs of $d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ for varying N , α , θ and λ_T	103
3.3	BM simulated data: Posterior CDFs of $d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ for varying N . $n = 15, 25$	104
3.4	BM simulated data: Boxplots of the posterior probabilities of correctly predicting the missing preferences.	105
3.5	Logistic theoretical probabilities of making a mistake as a function of the distance between the items compared.	106
3.6	LM simulated data: Boxplots of the posterior probability of correctly predicting the missing preferences stratified by the distance between the items compared	107
3.7	BM simulated data: Heatplot representing the posterior probabilities of estimating the preferences as in the ground truth.	109
3.8	BM Simulated data: ROC curves.	111
3.9	LM Simulated data: ROC curves	111
3.1	Beach preference data revisited: Posterior probability, for each beach, of being ranked among the top-3 in the LM consensus, and in the individual rankings.	114
3.2	Movie data: Posterior probability, for each movie, of being ranked among the top-3 in the LM consensus, and in the individual rankings.	116
3.3	Movie data: Preference matrices.	117
3.C.1	Trace plots of $a_{\boldsymbol{\mu}}$ (left) and $\boldsymbol{\mu}$ (right)	131
3.C.2	Trace plots of $\boldsymbol{\mu}_j$ for some users.	131

4.1	Details of the motion capture lab at the Department for Musicology, University of Oslo.	143
4.1	Acousmatic data. Boxplots of the within-cluster sum of footrule distances, and of the within-cluster indicator of mis-fit to the data.	149
4.2	Acousmatic data: Posterior consensus rankings of the three clusters.	150
4.3	Acousmatic data: Heatplot of the posterior probabilities for all listeners of being assigned to each cluster.	151
4.4	Acousmatic data: Heatplot of the marginal posterior probabilities for all the sounds of being ranked among the top-4.	152
4.5	Acousmatic data: relationship between the SAA and the posterior probabilities for some sounds of being ranked among the top-4 in the individual rankings.	153
4.6	Acousmatic data: relationship between the MSI and the posterior probabilities for some sounds of being ranked among the bottom-4 in the individual rankings.	153
4.7	Acousmatic data: Heatplot representing the cycle co-memberships of the sounds in the non-transitive patterns of the data.	154
5.1	The Cayley distance as a function of the dispersion parameter, for different values of n	160
5.2	Computational time of the Cayley distance.	164
6.1	The generic permutation polytope for $n = 3$ items.	169
6.2	Posterior distribution of ρ for varying θ_0 . $\theta = 0.5$	170
6.3	Posterior distribution of ρ for varying θ_0 . $\theta = 1.5$	170

Tesi di dottorato "Bayesian learning of the Mallows ranking model"

di CRISPINO MARTÀ

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2018

La tesi è tutelata dalla normativa sul diritto d'autore (Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.

List of Tables

1	Summary of the test sounds.	6
1.1	An example of right invariance.	18
2.1	Simulated data: Acceptance probability and IAT as a function of the standard deviation of the log-normal proposal.	37
2.2	Simulated data: Maximum relative error, of the approximation of the partition function via IS.	41
2.3	Simulated data: Maximum relative error, between the IS and limiting approximations of the partition function.	44
2.4	Simulated data: Comparisons with other methods in terms of the posterior consensus ranking and of the goodness of fit.	53
2.5	Meta-Analysis: Top-25 genes in the MAP consensus ranking from a total of 89 genes.	68
2.6	Meta-Analysis: Results given by other methods.	70
2.7	Meta-Analysis: Values of the average footrule distance for partial data, for each of the considered methods.	71
2.8	Beach preference data: Beaches arranged according to the CP consensus ordering.	72
2.9	Beach preference data: Consensus ordering given by other methods.	73
2.10	Sushi data: Sushi items arranged according to the MAP consensus.	74
2.11	Sushi data: lists of Sushi items produced by the <code>rankdist</code> package.	76
2.12	Movielens data: Movies arranged according to the CP consensus.	78
3.1	BM simulated data: Percentage of users for which the estimated top-3 items belong to the true top-5.	104
3.2	LM simulated data: Percentage of users for which the estimated top-3 items belong to the true top-5.	107
3.3	Simulated data. AUC statistics.	112
3.1	Beach preference data revisited: Beaches arranged according to the CP consensus ordering obtained with LM and BM.	113

3.2	Beach preference data revisited: Number of assessors for which we correctly identify the preferred beaches.	115
3.3	Movie data: Movies arranged by the CP consensus ordering (LM).	116
3.B.1	Logistic theoretical values of the probability of making a mistake depending on the value of the distance between the compared items.	126
3.C.1	MSE of $\boldsymbol{\mu}$	132
3.C.2	MSE of $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N$	132
3.C.3	Results from the comparison in terms of loss. Data generated from BM - $M = 0.5n(n - 1)/2$	133
3.C.4	Results from the comparison in terms of posterior distance. Data generated from BM - $M = 0.5n(n - 1)/2$	133
3.C.5	Results from the comparison in terms of loss. Data generated from BM - $M = 0.7n(n - 1)/2$	133
3.C.6	Results from the comparison in terms of posterior distance. Data generated from BM - $M = 0.7n(n - 1)/2$	134
3.C.7	Results from the comparison in terms of loss. Data generated from BTI - $M = 0.5n(n - 1)/2$	134
3.C.8	Results from the comparison in terms of posterior distance. Data generated from BTI - $M = 0.5n(n - 1)/2$	134
4.1	Acousmatic data: Sounds arranged according to the the CP ordering (BM).	149

Acknowledgements

I would like to thank all the people who supported me in these years, and made this thesis possible.

Arnoldo, for his guidance. In his role, not only he taught me to make scientific research, but also he showed me the way to work in a team and to feel part of it.

Elja, for his excellent advice. His ideas and rigor inspired me during these years.

Sonia, for giving me the opportunity to follow my research interests, and for carefully reading through the thesis and providing very helpful and detailed comments. During all these years as a PhD student, she accommodated all my needs and choices. Thank you!

My PhD mate Lorenzo, with whom I shared this experience and whose temper helped me in many difficult situations.

My coauthors Valeria, Natasha, and Øystein.

I am also grateful to all the people at OCBE for the warm welcome they reserved to me during the periods I spent in Oslo. I must in particular thank Kathrine Frey Frøslie for having introduced me to Natasha. Part of this thesis couldn't be there if it weren't for your insight.

Matteo, for being there always for me. You are my pillar.

I cannot forget my lifetime friends, Chiara, Giulia, Giulia 'Cuja', Wank, Claudia, Alice, Max, Gabri. Having you around makes me feel home.

My Oslo friends, Alice, Jacopo, Emily, Andrea, who enlightened the dark and cold Norwegian winters.

Giulia and Isa, I am indebted to you for your patience, and for the time you devoted to listening to me.

My dad and my aunt. Your love will guide me forever.

Abstract

This thesis studies the Mallows rank model in a Bayesian framework. This model, widely used for analyzing ranking data, assumes that the probability distribution of a ranking decays as the distance between the ranking and the modal ranking increases. However, inferential complexity has restricted its use to few distance functions between rankings. Our main contributions are the following: (a) to develop a framework to perform Bayesian inference in the Mallows model with most of the distances used in the literature; (b) to propose a strategy to approximate the intractable normalizing constant of the model, and even to derive it exactly in some special cases; and (c) to generalize the developed model in order to handle pairwise comparisons in the presence of non-transitive data. We extensively document the accuracy of the model, both theoretically and with simulated data, and report the analysis of many benchmark and toy datasets, to show the ability of the method to adapt to different data types. Our model is then applied to a study in the field of musicology, whose aim is to learn how people perceive electronically synthesized sounds as having human origin.

Introduction

This thesis revolves around the problem of rank aggregation and preference learning. Generally speaking, rank aggregation is about summarizing or aggregating the preferences of a population, while preference learning is about producing predictive preferences from data regarding user preferences. It is already clear from these two definitions that rank aggregation and preference learning are intimately related.

Different types of data have ranks as their natural scale, and can be classified into two main classes: explicit data, when people express a direct opinion over some items, or implicit, when people express an indirect preference by choosing some items and not others. Moreover, the ranking data can be either observed directly, or transformed from sets of assigned scores. For instance, given a set of scored items, one can order them according to the score value, which naturally gives rise to a ranking. In this thesis we deal only with explicit data, appearing in three main forms: full rankings, partial rankings, and pairwise comparisons.

Notable examples of fields where ranking data arise are: companies who recruit panels to rank novel products; market researches, which can be based on interviews where competing services, or items, are compared or ranked; political polls where is usually asked voters to rank electoral candidates; search engines, where retrieval results have to be ranked according to a user's preferences; recommender systems used by online stores to recommend products to their customers.

Some typical goals when dealing with rankings or preference data are: (i) aggregate the data coming from a group of homogeneous users, and summarize their preferences into a shared consensus ranking; (ii) estimate the individual rankings of the items, in case the users express only incomplete preferences, which amounts to predict the ranks of unranked items at the individual level; (iii) cluster the users into classes, each sharing a consensus ranking of the items, and classify new users.

In this thesis we handle all these tasks and their combinations in a unified Bayesian in-

ferential framework, which enables us to quantify posterior uncertainty of the estimates. Uncertainty evaluations of the estimated preferences and class memberships are a fundamental aspect of information in marketing and decision making. Actions based on unreliable predictions might better be postponed until more data are available and safer predictions can be made, in order not to unnecessarily annoy users or clients.

The proposed framework provides two main quantities of interest: the posterior distribution of the consensus ranking of a population, and the posterior distribution of the individual rankings for each user, when not readily available from the data. The consensus ranking can be seen as a model-based Bayesian aggregation of individual preferences of a group of users. It is analogous to the quantity which is of interest in the rank aggregation literature. The individual rankings are of great interest, for example, when performing personalized recommendations, or in studying how individual preferences change with user related covariates.

The Mallows rank model is one of the most widely used statistical models for analyzing ranking data. It assumes that the probability distribution of a ranking decays as the distance between the ranking and the modal - or consensus - ranking increases, and is thus known in the literature as a distance-based model. Inference in the Mallows model is difficult due to its intractable normalizing constant, which has closed form only for few distances, while is very time-consuming to numerically calculate in many instances.

The main contribution of this thesis is to develop a unifying framework to make inference in the Mallows model with most of the distances used in the literature. This is made possible thanks to a newly developed strategy able to handle the normalizing constant even in many of the cases when it was judged intractable so far. This strategy in itself is a separate methodological contribution of this thesis. The principal advantage of the Bayesian paradigm in this context comes from its ability to coherently quantify posterior uncertainties of estimates of any quantity of interest. Indeed, since our method provides the full posterior distribution of the parameters of interest, it makes possible to select any strategy to summarize it, driven by the application at hand. This is useful in applications, where the interest is often in computing posterior probabilities of more complex functions of the consensus ranking, for example the posterior probability that a certain item has rank lower than a given level (“among the top-3”, say), or that the rank of a certain item is higher than the rank of another one. These probabilities cannot be readily obtained within the maximum likelihood approach, while the Bayesian setting very

naturally allows to approximate any posterior summary of interest by means of a Markov Chain Monte Carlo algorithm. Moreover, the Bayesian framework allows to naturally combine different types of uncertainty in the reported data, coming from different sources, and to convert such data into the form of meaningful probabilistic inferences.

A second core contribution of the thesis is to generalize the Mallows model to handle pairwise comparison data in the presence of non-transitivity, that is, when one or more pairwise preferences contradict what is implied by other pairwise preferences given by the same individual. It is important to underline that in this thesis we consider non-transitivity at an individual level, thus not arising when aggregating preferences across users, as under majority rule. Non-transitivity of preferences can arise for many reasons, for example users' inattentiveness, uncertainty in their preferences, or actual confusion, even when one specific criterion for ranking is used. These situations are so common that most pairwise comparison data are in fact non-transitive, thus creating the need for methods able to predict individual preferences from data that lack logical transitivity. To our knowledge, most methods designed to estimate individual rankings from pairwise comparison data do not handle individual-level non-transitivity. Usually, one either drops such pairs, or only focuses on the estimation of the consensus ranking, without specifically modeling the non-transitivity characterizing the data. Instead, we handle individual non-transitive patterns with a latent layer of uncertainty which captures the generation of preference misreporting. We believe that the novelty of our strategy for individual-level non-transitivity is an important contribution to the literature on preference learning.

The motivation for the present work is not only methodological, but also driven by a specific application in the field of musicology, which is described in the next section.

Application

The practical problem in this application is to learn how listeners perceive sounds as having human origin. Specifically, we consider pairwise comparison data coming from an experiment where people were asked to hear a series of two different abstract sounds, and to tell which one was perceived as more human. Non-transitivity in the reported data arises in this experiment, and is mainly due to the intrinsic complexity of the task, and to the heterogeneity of the cohort of listeners, who had backgrounds varying from musicologists to university students.

The results of this test are of interest for musicologists, composers and sound designers, whose aim is to understand how computer generated sounds can appear more life-like.

The experiment relates to *acousmatic* music, which is a type of electronic music composed for presentation using loudspeakers, as opposed to live or video recorded performance. In *acousmatic* music the composer manipulates digitally recorded sounds, so that the cause of the sound, being a musical instrument or any other sound making system, remains hidden. Indeed, when sounds are played over loudspeakers there are no visual cues to help listeners understand how the sounds were made. On the other hand, when we hear the sound of musical instruments or sounds from our everyday environment, we are able to recognize their cause. This happens because in visual music we obtain the information that indicates the sounding object, that is, its causation. To give an example: if you listen to the crying of a baby, you guess that a baby produced this sound.

Since the advent of recording technology, abstract sounds - that is, sounds transformed with computer tools - have been used not only in *acousmatic* music, but also in much of the sound-world we experience over the Internet, TV and cinema.

The research question of this study is related to the capacity of listeners to identify the presence of human causation through the spatial behavior of sounds. *Spatial* here describes the fact that the causation of sound happens as an action in 3-D space.

The starting point for the experiment was a high-speed motion tracking recording of the physical movement used to produce one selected sound: a cellist bowing a down-bow chord. Features of this 3-D movement were successively subtracted, resulting in a series of 12 motion data-sets of varying proximity to the original. The motion data were then made audible by a process called parameter-mapping sonification (Grond and Berger 2011), where parameters in the data are mapped to parameters controlling computer generated sound (see Figure 1). The mapping rules are chosen to draw on our everyday perception of spatial motion, which involves not only absolute 3-D spatial location but, in addition, changes in volume, intensity and pitch correlated with changes in proximity and speed. In other words, listeners heard the physical spatial motion through sonification, rather than hearing the sound that the motion created, which, in this instance, was the sound of the cello.

Testing how listeners perceive a sound for which we lack a clear and commonly understood descriptive vocabulary is problematic. Moreover, listeners' varying familiarity with

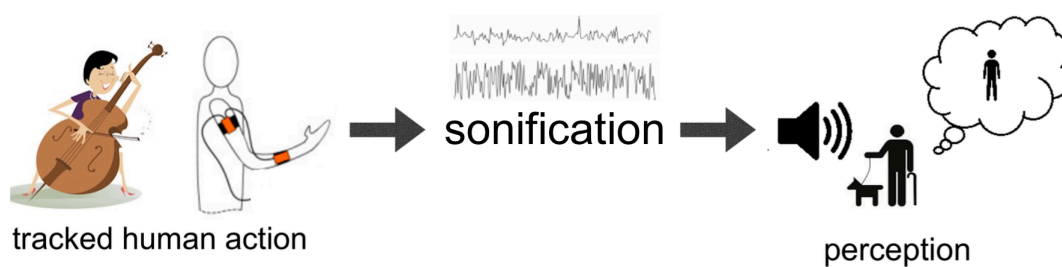


Figure 1: A scheme of the process generating sounds.

acousmatic music - or more generally with abstract sounds - affects the test results. For these reasons, a pair comparison test is the most appropriate design.

Pair comparison experiment and data

A group of 46 listeners were presented with a series of sounds, here sometimes called spatial audio stimuli because they originate from a spatial 3-D performance. The number of stimuli was 12. Test stimulus 1 (S1) was designed to more clearly sonify all features of the data. Each of the other 11 test stimuli were sonified by modifying one or more features of the data. This involved removing pitch and volume variation, flattening directional changes in the motion, or slowing the overall motion speed, as summarized in Table 1. Each listener was then exposed to 30 pairs of these sounds, which is ca. 45% of the total number of possible combinations out of 12 stimuli. The pairs were chosen randomly and independently for each user. Also the order in which the sounds were played was randomized.

Listeners were then asked to indicate, for each pair, which of the two stimuli most evoked a sensation of human physical movement of any kind. They were asked to follow their feelings, rather than imagining to watch a performance. The listeners were not told that the source motion stemmed from a cellist, nor were they asked to identify a specific human spatial movement. Before the test, participants were informed that the sounds were made by sonification and were warned that the sounds may be heard as strange. Each listener carried out the test sitting centrally to the loudspeaker array. Prior to the experiment, listeners were presented with a short training session of three sounds not used in the test sequence. When the experiment began, the test number was displayed on a computer screen, answers were written on paper, and listeners were requested to always

S1:	Pitch, volume-2, grain duration and spatial variations set to their most dynamic ranges.
S2:	The same as S1, but with spatial motion placed in front of the listening location.
S3:	Pitch, volume-2 and grain duration set to their most dynamic ranges, but all spatial motion removed and the sound rendered as mono. Played over one loudspeaker located in front of the listener.
S4:	3-D spatial variation in the original data partially reduced, leaving global direction changes. The same as S1, but reduced in pitch, volume-2 and grain duration variations.
S5:	3-D spatial variation in the original data flattened further, consisting of just three changes in direction between two points in space. The same as for S1, but with little variation in pitch, volume-2 and grain duration.
S6:	The same as S1, but with volume-2 variation removed.
S7:	The same as S1, but with pitch variation removed.
S8:	The same as S1, but with pitch and volume-2 variation removed.
S9:	The same as S4, but with pitch and volume-2 variation removed.
S10:	The same as S5, but with pitch and volume-2 variation removed.
S11:	The same as S1 played 30% slower in tempo.
S12:	The same as S1 played 50% slower in tempo (half speed).

Table 1: Summary of the test sounds.

make a choice even if they found it difficult to decide. If needed, they could ask to hear a test pair for a second time. At the end, they were asked to complete two questionnaires, the aim of which was to assign a Musical Sophistication Index score (MSI) and a rating of Spatial Audio awareness (SAA) to each listener. The MSI used was the Ollen musical sophistication index (Ollen 2006), which is an online survey that tests the validity of 29 indicators of musical sophistication used in published music research literature. The SAA index consisted of five questions as indicators of how aware listeners were of spatial audio regardless of musical background. Such a test did not exist in the literature, and needed to be custom designed for the experiment.

The choice to rely on a pairwise comparison experiment is crucially based on the listeners' lack of experience with abstract sounds. It is easier for the participants to compare two sounds, rather than to be exposed to several, which could create confusion. The experiment, indeed, was difficult as expected: 37 listeners (80%) reported non-transitivities in their pair comparisons, only 9 out of 46 listeners were able to stay consistent with themselves. The full description of the background studies, hypotheses, experimental setup and discussion of results is reported in Chapter 4.

Outline of thesis

The Bayesian Mallows model for ranking data is the focus of this thesis. We begin with a review of the main models for ranking data in Chapter 1, with a particular interest in the probabilistic approaches more closely related to the Mallows model.

In Chapter 2, we present the Bayesian Mallows model for rankings, along with a detailed explanation of the algorithm. We then study its properties, both theoretically and with simulations, and report the results of the analysis of some benchmark datasets. In Chapter 3, we present the Bayesian Mallows model for non-transitive pairwise comparisons. We explain the algorithm and report both simulation results and some examples on real data. In Chapter 4 we then show the results of the application to the sound data.

Chapter 5 is devoted to the specialization of the Bayesian Mallows model with Cayley distance. In Chapter 6, which is still preliminary, we discuss the elicitation of the conjugate prior for the Mallows model parameter.

Lastly, we conclude with a discussion of the contributions of this thesis.

Chapters 2-4 contain the main contributions of this thesis, that originate the works [Vitelli et al. \(2017\)](#), [Crispino et al. \(2017\)](#), and [Barrett and Crispino \(2017\)](#). Chapters 5 and 6 are ongoing work.

Chapter 2 is a development of Paper III of the PhD thesis of Øystein Sørensen from the University of Oslo (thesis defense: March 2015).

In particular, Chapter 2 improves and innovates in the following aspects:

1. We thoroughly revised the computational strategy by (i) studying the mixing properties of the convergence of the MCMC to the approximate target posterior, as well as the tuning of the proposal distributions and their effect on acceptance probabilities and convergence; (ii) theoretically studying how the approximated target distribution converges to the correct posterior as the number of IS samples grows; (iii) systematically exploring the effect of various approximations of the partition function on inference; (iv) presenting a simulation experiment with heterogeneous assessors and incomplete pairwise data; (v) deriving the exact calculation of the partition function for footrule up to 50 items; (vi) speeding-up the algorithm consistently (more efficient computation of distances between permutations, improved proposal

distributions both for the IS procedure, and in the MH steps of the MCMC);

2. We revised the section about prior densities;
3. We compared our method with other existing competitors, both on simulated data, and in the case studies;
4. We removed some of the case studies (the Potato, the Premier League and the Breast cancer datasets) and substituted them with simulated data, and a more suitable case study (the Beach dataset);
5. We entirely removed the study about time-dependent rankings;
6. We developed and used a new measure for selecting the number of clusters in order to be more robust to data sparsity. This criterion has been tested on simulated data.

Chapter 1

Background

In this chapter we briefly review the statistical literature on preference data, with a particular focus on the Mallows model, and on the other probabilistic approaches most closely related to our method. Attention will be given also to methods designed for pair comparison data, which are relevant for Chapters 3 and 4 of the thesis.

We here follow the lecture notes by [Diaconis \(1988\)](#), the monograph by [Marden \(1995\)](#), and the books by [Critchlow et al. \(1993\)](#) and [Alvo and Yu \(2014\)](#).

This chapter contains joint work with Valeria Vitelli, Øystein Sørensen, Natasha Barrett, Arnoldo Frigessi and Elja Arjas from [Crispino et al. \(2017\)](#) and [Vitelli et al. \(2017\)](#).

Outline

There are many ways of reviewing the literature on preference data. We here try both to follow the historical developments, and to separate the existing methods into sub-classes, for the sake of clarification.

In Section 1.1, we introduce four probabilistic models on ranking data, each of which corresponds to a specific generative process of the ranking: the Thurstonian order statistics models (Section 1.1.1), the pair comparison models, like the Babington Smith and the Mallows-Bradley Terry models (Section 1.1.2), the Mallows distance-based model (Section 1.1.3), and the multistage Plackett-Luce model (Section 1.1.4). In Section 1.2 we give an overview of the probability models for real pair comparison data, focusing mainly on the Bradley Terry model. We also outline the problem of non-transitivity present in pair comparison data, and briefly review the literature pertaining to this topic (Section 1.2.1). Finally, in Section 1.3, we present some ideas of the relevant machine learning literature,

and mention the papers that are mostly connected to the present work.

This chapter is not intended to broadly cover all the topics mentioned, but rather to equip the reader with the basic notions useful for contextualizing this thesis, and that will be considered familiar in the next chapters.

1.1 Probabilistic models on ranking data

Let us first clarify the terminology, notation and conventions, that we will use throughout the thesis.

A full ranking of n items is a mapping $\mathbf{R} : \mathcal{A} \rightarrow \mathcal{P}_n$ from a finite set, $\mathcal{A} = \{A_1, \dots, A_n\}$, which denotes the set of labeled items to be ranked, to the space of n -dimensional permutations \mathcal{P}_n , that results from the attribution of a rank $R_i \in \{1, \dots, n\}$ to each item, according to some criterion.

\mathcal{P}_n is a non-abelian group under composition, that is, in general, the composition is not commutative, and is also called the symmetric group of order n , denoted by \mathcal{S}_n .

We here denote a generic full ranking by $\mathbf{R} = (R_1, \dots, R_n)$, where R_i is the rank assigned to item A_i . By convention, $R_i < R_k$ means that item A_i is preferred to item A_k , since the rank assigned to A_i is *lower* to the one assigned to A_k (the most preferred item has rank $R_i = 1$), but is read A_i is ranked *higher* than A_k .

The full ordering is an alternative way of representing ranking data, and is here denoted by \mathbf{X} . The ordering and ranking vectors are in one-to-one correspondence: the components of $\mathbf{X} = (X_1, \dots, X_n)$ are items in \mathcal{A} , ordered from the most preferred to the worst, according to \mathbf{R} . In other words, it holds the following relationship: $X_i = A_k \iff R_k = i$, $\forall i, k = 1, \dots, n$, which will be shortcut as, $\mathbf{X} = \mathbf{R}^{-1}$. Then, $\mathbf{X} \in \mathcal{X}_n$, the set of permutations of the labels in \mathcal{A} .

For example, given the following full ranking of the items labelled $\mathcal{A} = \{A_1, \dots, A_{10}\}$

$$\mathbf{R} = \begin{matrix} & A_1 & A_2 & A_3 & A_4 & A_5 & A_6 & A_7 & A_8 & A_9 & A_{10} \\ \left(\right. & 1, & 7, & 8, & 2, & 10, & 4, & 6, & 9, & 3, & 5 \end{matrix} \left. \right),$$

the corresponding ordering vector is the following:

$$\mathbf{X} = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ \left(\right. & A_1, & A_4, & A_9, & A_6, & A_{10}, & A_7, & A_2, & A_3, & A_8, & A_5 \end{matrix} \left. \right).$$

Some models that we will introduce in the next sections are naturally defined on the rankings, others on the orderings, but they are equivalent, being induced one from the other because of the one-to-one correspondence between \mathbf{R} and \mathbf{X} .

It is important to clarify here the intimate relation that exists between a ranking and pairwise preferences. Given an unordered pair of items $\{A_i, A_k\}$, throughout the thesis we denote a pairwise preference between the two items as $(A_i \prec A_k)$, meaning that item A_i is preferred to item A_k . Given a full ranking $\mathbf{R} \in \mathcal{P}_n$, it is immediate to evince all the possible $n(n-1)/2$ pair orderings among the items, according to the following rule:

$$(A_{t_1} \prec A_{t_2}) \iff R_{jt_1} < R_{jt_2}, \quad t_1, t_2 = 1, \dots, n, \quad t_1 \neq t_2, \quad (1.1)$$

that is, when considering a pair of items, the item with lowest rank is the preferred one.

We will call pairwise preferences obtained as above *derived pairwise preferences* (DPP), to distinguish them to *real pairwise preferences* (RPP), which arise when people are asked to compare items in pairs, rather than to perform a full ranking. The main difference between these two types of pairwise data is that DPP are always complete and transitive, while RPP can be incomplete (if a user does not perform all the possible pairs), and non-transitive (if a user happens to contradict herself).

Therefore, when dealing with RPP data, we can face the following three cases:

1. the data are complete and transitive;
2. the data are partial and transitive;
3. the data are non-transitive (complete or partial).

In case 1, only one ranking is consistent with the data. In case 2 generally multiple rankings can be consistent with the data, while in case 3 no ranking is consistent with the data, because, by definition, a ranking is transitive (see also Section 1.2.1).

In Chapter 2, Section 2.3.2, we deal with transitive data, both DPP and RPP, with a particular focus on partial data. In Chapter 3 we will use RPP data, extending the model of Section 2.3.2 to non-transitivity.

1.1.1 Thurstonian order statistics models

To model data arising from a ranking or paired comparisons experiment, [Thurstone \(1927\)](#) introduced its *Law of Comparative Judgment*. The idea is that, in a ranking experiment,

each item has a utility (or score) and users' preferences depend on this utility, so that the item with the largest utility at the moment of comparison is preferred. In such models the utilities are latent variables, and each of these corresponds to a specific item. This law was initially proposed for real paired comparison data (see Section 1.2), later used on ranking data by Thurstone himself some years later, who converted ranking data into pair comparisons (according to eq. (1.1)), which were then analyzed using the Law of Comparative Judgment, and more recently Daniels (1950) extended it to data in the form of full orderings of the n items.

Formally, in the order statistics model is assumed that, in a ranking task involving n items, $\{A_1, \dots, A_n\}$, there exist n random utilities (or scores) Y_1, \dots, Y_n , one for each item, which are assumed independent and distributed according to F_i . The model then assigns to a ranking $\mathbf{R} \in \mathcal{P}_n$ the probability

$$P(\mathbf{R}) = P(Y_{i_1} < Y_{i_2} < \dots < Y_{i_n}) \quad (1.2)$$

where $i_r = k$ if and only if $R_k = r$. In a nutshell, under the order statistics model, the generative process of a ranking of n items is determined by the relative ordering of the n random utilities.

The most common simplification of (1.2) is to assume a linear model for the random utilities, that is, $Y_i = u_i + \epsilon_i$, $i = 1, \dots, n$, where the u_i is the mean score associated with item A_i and ϵ_i captures its variability. Such models are known as Thurstone order statistics models. Depending on the probabilistic statement on $F_i(y) = F(y - u_i)$ different models arise: the Thurstone model (Thurstone 1927) assumes that F is Gaussian, and the Bradley-Terry-Luce (BTL) model (Bradley and Terry 1952, Luce 1959) assumes that F is Gumbel (see also Section 1.1.4). Since under the BTL model the probability in (1.2) has close form, many works dealing with order statistics models for rankings are based on the the latter.

In addition to the order statistics models, there is a plethora of methods for constructing ranking models based on independent latent variables. For instance, one may employ paired comparison probabilities to construct a probability model on rankings. This idea is the topic of the next section.

1.1.2 Ranking models induced by pairwise comparisons

A popular approach to ranking data is the class of paired comparison models (David 1963, Alvo and Yu 2014). These models build on the connection between a ranking of n items in a set $\mathcal{A} = \{A_1, \dots, A_n\}$, and all $\binom{n}{2} = n(n-1)/2$ pairwise comparisons between the items themselves. The principal assumption of paired comparison models is that a ranking arises by making all the comparisons between paired items independently, and only accepting the result if it is consistent with a ranking (in the sense of equation (3.1)).

The saturated model in this class is the Babington Smith (BS) model (Smith 1950), which assumes exactly that a ranking is evinced by the $\binom{n}{2}$ paired comparison probabilities if they are consistent with each other. Formally, define the parameter $p_{ik} = P(A_i \prec A_k)$, to be the probability that item A_i is preferred to item A_k . Assuming mutual independence of the p_{ik} , the likelihood of a ranking \mathbf{R} is proportional to the product of the probabilities of all the pair comparisons that generated that ranking, and is then given by

$$P(\mathbf{R}|\mathbf{p}) = \frac{1}{c(\mathbf{p})} \prod_{(A_i, A_k): R_i < R_k} p_{ik}, \quad (1.3)$$

where the normalizing constant, $c(\mathbf{p}) = \sum_{\mathbf{r} \in \mathcal{P}_n} \prod_{(A_i, A_k): r_i < r_k} p_{ik}$, represents the probability that the set of comparisons is consistent with a ranking $\mathbf{r} \in \mathcal{P}_n$. The main drawback of the BS model, that in fact limited its use, is the inferential complexity growing with the number of items. It is indeed parametrized by the $\binom{n}{2}$ parameters, $\mathbf{p} = \{p_{ik}\}_{i=1, k>i}^n$, one for each pair of items. For this reason, in the past years some subclasses of the BS model arose, that added further constraints on the parameters: the two most important are the Mallows-Bradley-Terry and the Mallows models, both developed by Mallows (1957).

The Mallows-Bradley-Terry (MBT) model (Mallows 1957), is perhaps the most famous descendant of the BS model. It assumes that the probability that an item A_i is preferred to an item A_k , has the Bradley-Terry form (Bradley and Terry 1952),

$$p_{ik} = \frac{\mu_i}{\mu_i + \mu_k}, \quad (1.4)$$

where $\mu_i > 0$, $i = 1, \dots, n$, and $\sum_{i=1}^n \mu_i = 1$, that is, it only depends on item-specific parameters representing the score rating - or skill - of the two items under comparison (larger values of μ_i correspond to more preferred items).

Substituting (1.4) into (1.3) leads to the MBT ranking model,

$$P(\mathbf{R}|\boldsymbol{\mu}) = \frac{1}{c(\boldsymbol{\mu})} \prod_{i=1}^{n-1} \mu_{ik}^{n-i}. \quad (1.5)$$

The complexity of the original BS is then greatly reduced in the MBT, as the number of free parameters is set to $n-1$. The goal of the MBT model is then to make inference on $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)$, the vector of items' scores, that best represents the consensus preferences of all the users.

In the same paper, Mallows (1957) introduced a further simplification of the BS model, defining the ϕ -model and the ρ -model, both parametrized by only two parameters. The assumption of both is that p_{ik} depends only on the relative order of items A_i and A_k in the ranking \mathbf{R} . Then p_{ik} is the same for all pairs of items $\{A_i, A_k\}$ such that $R_i < R_k$. In the ϕ -model p_{ik} depends on whether $(R_i - R_k) > 0$ or not, and in the ρ -model p_{ik} depends also on the absolute difference of the ranks $|R_i - R_k|$.

The ϕ -model and the ρ -model belong to a more general distance-based family of distributions for rankings, $\mathbf{R} \in \mathcal{P}_n$, usually referred to as Mallows models, and formalized in its general form by Diaconis (1988), which is the topic of next section.

1.1.3 Distance based ranking models: the Mallows model

The Mallows model (MM) specifies the probability density, $\mathcal{M}(\boldsymbol{\rho}, \alpha)$, for a ranking $\mathbf{R} = (R_1, \dots, R_n) \in \mathcal{P}_n$, as follows

$$P(\mathbf{R} | \alpha, \boldsymbol{\rho}) := \frac{1}{Z_n(\alpha, \boldsymbol{\rho})} \exp \left[-\frac{\alpha}{n} d(\mathbf{R}, \boldsymbol{\rho}) \right], \quad (1.6)$$

where $\boldsymbol{\rho} \in \mathcal{P}_n$ is the location parameter (representing the shared consensus ranking), $\alpha > 0$ is the scale parameter (describing the concentration around the shared consensus), $d(\cdot, \cdot)$ is a distance function between two n -dimensional permutations, and

$$Z_n(\alpha, \boldsymbol{\rho}) = \sum_{\mathbf{r} \in \mathcal{P}_n} e^{-\frac{\alpha}{n} d(\mathbf{r}, \boldsymbol{\rho})}$$

is the normalizing constant (that we will often refer to as partition function). The MM is based on the assumption that there exists a modal ranking $\boldsymbol{\rho} \in \mathcal{P}_n$, and that the likelihood of a ranking \mathbf{R} decreases geometrically as some given distance between $\boldsymbol{\rho}$ and \mathbf{R} increases.

As a consequence, every permutation at the same distance to $\boldsymbol{\rho}$ has equal probability.

Depending on the choice of the distance, several models arise: the ϕ -model, mentioned earlier, is equivalent to the MM with Kendall distance,

$$d_K(\mathbf{R}, \boldsymbol{\rho}) = \sum_{i < k} \mathbb{1}[(\rho_i - \rho_k)(R_i - R_k) < 0], \quad 1 \leq i < k \leq n,$$

and the ρ -model is equivalent to the Mallows with Spearman's distance (that is l_2),

$$d_S(\mathbf{R}, \boldsymbol{\rho}) = \sum_{i=1}^n (\rho_i - R_i)^2,$$

both with $\boldsymbol{\rho} = (1, \dots, n)$. The Kendall distance, that measures the number of adjacent transpositions which convert \mathbf{R} into $\boldsymbol{\rho}$ or, equivalently, the number of discordant pairs in \mathbf{R} and $\boldsymbol{\rho}$, is by far the distance most frequently considered in the literature of MM, for reasons that will become clear soon. Other distance functions that appear frequently in the literature are the footrule distance (that is l_1),

$$d_F(\mathbf{R}, \boldsymbol{\rho}) = \sum_{i=1}^n |\rho_i - R_i|,$$

the Hamming distance,

$$d_H(\mathbf{R}, \boldsymbol{\rho}) = n - \sum_{i=1}^n \mathbb{1}_{\rho_i}(R_i),$$

the Cayley distance, $d_C(\mathbf{R}, \boldsymbol{\rho})$, which measures the minimum number of transpositions which convert \mathbf{R} into $\boldsymbol{\rho}$, and the Ulam distance, $d_U(\mathbf{R}, \boldsymbol{\rho})$, which is the number of deletion-insertion operations to convert \mathbf{R} into $\boldsymbol{\rho}$ (we refer to [Marden 1995](#), for detailed description of these distances).

All the distance functions mentioned so far satisfy the usual axioms, namely

$$d(\boldsymbol{\rho}, \boldsymbol{\rho}) = 0 \quad \forall \boldsymbol{\rho} \in \mathcal{P}_n \quad (\text{reflexivity}) \tag{1.7}$$

$$d(\boldsymbol{\rho}, \boldsymbol{\sigma}) > 0 \quad \forall \boldsymbol{\rho}, \boldsymbol{\sigma} \in \mathcal{P}_n, \text{ s.t. } \boldsymbol{\rho} \neq \boldsymbol{\sigma} \quad (\text{positivity}) \tag{1.8}$$

$$d(\boldsymbol{\rho}, \boldsymbol{\sigma}) = d(\boldsymbol{\sigma}, \boldsymbol{\rho}) \quad \forall \boldsymbol{\rho}, \boldsymbol{\sigma} \in \mathcal{P}_n \quad (\text{symmetry}). \tag{1.9}$$

Some of them, like Kendall and footrule, are also metrics on \mathcal{P}_n , in that they satisfy also

the triangle inequality

$$d(\boldsymbol{\rho}, \boldsymbol{\sigma}) \leq d(\boldsymbol{\rho}, \boldsymbol{\tau}) + d(\boldsymbol{\tau}, \boldsymbol{\sigma}) \quad \forall \boldsymbol{\rho}, \boldsymbol{\sigma}, \boldsymbol{\tau} \in \mathcal{P}_n. \quad (1.10)$$

The role of the distance plays a fundamental role in the MM. In fact, each distance induces a different partition of the space of permutations into level sets.

To clarify, given a permutation $\boldsymbol{\rho}$, for varying $\boldsymbol{R} \in \mathcal{P}_n$, the distance $d(\boldsymbol{\rho}, \boldsymbol{R})$ takes only a finite set of discrete values in $\mathcal{D} = \{d_1, d_2, \dots, d_a\}$, where a depends on n and on the chosen distance $d(\cdot, \cdot)$. By defining the level sets $L_i = \{\boldsymbol{R} \in \mathcal{P}_n : d(\boldsymbol{\rho}, \boldsymbol{R}) = d_i\} \subset \mathcal{P}_n$, $i = 1, \dots, a$, to be the set of permutations at the same given distance from $\boldsymbol{\rho}$, we can notice that $\mathcal{P}_n = \cup_{i=1}^a L_i$. The MM assigns the same probability density to all the permutations belonging to the same level set. In Figure 1.1, we provide a graphical representation of the MM for the six distances introduced earlier, in the case $n = 5$. On the x-axis it is represented the space of permutations, partitioned into the level sets defined by the corresponding distance. The length of each interval corresponds to the cardinality of the corresponding level set, L_i . On the y-axis is represented the Mallows density of eq. (1.6), for varying values of α , as stated in the legend on the right of the plots.

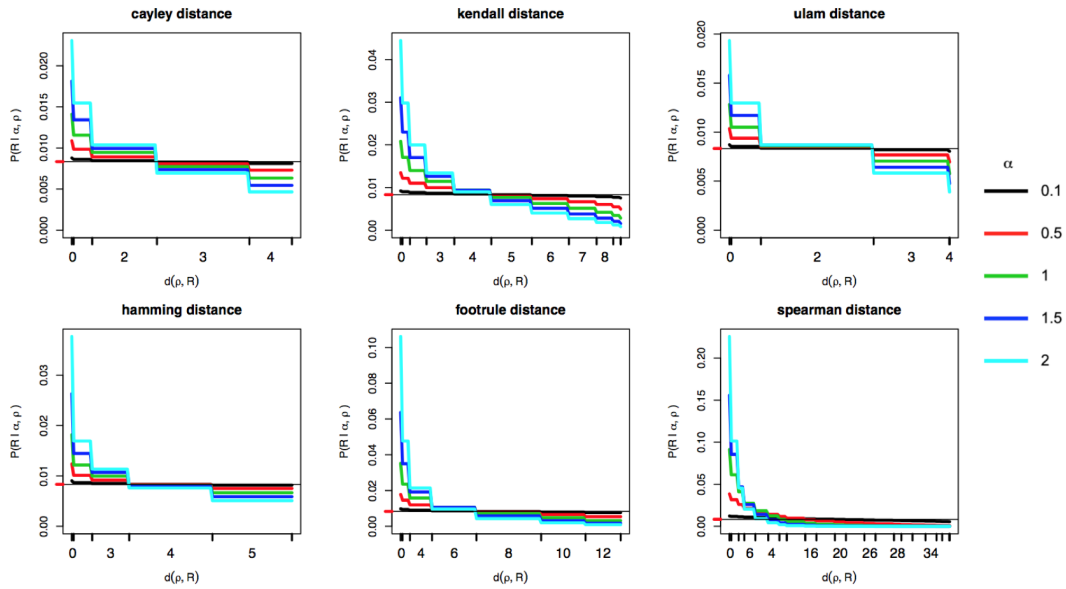


Figure 1.1: The Mallows density for the six right-invariant distances, for $n = 5$. On the x-axis is represented the space of permutations, partitioned into level sets. On the y-axis is represented the Mallows density of eq. (1.6), for varying values of α , as stated in the legend on the right of the plots.

This representation will be exploited in Section 2.2.1, where we will provide an exact

formula for the partition function of the MM.

Maximum Likelihood inference about the consensus ranking in the MM is generally very difficult, and in many cases NP-hard (Bartholdi et al. 1989a,b), which led to the development of heuristic algorithms (e.g. Busse et al. 2007). An additional inferential problem is that the MM, with α and ρ unknown, is not regular, since the parameter space is $\mathbb{R}^+ \times \mathcal{P}_n$. For these reasons, a Bayesian approach, based on sampling, rather than optimizing, could prove to be crucial.

In Chapter 2 we indeed propose a Bayesian approach for estimating the MM, that handles all the distances defined above. In practice though, we focus on the Kendall, the footrule and the Spearman distances in Chapters 2 and 3. In Chapter 5 we study the MM with Cayley distance, and in Chapter 6 we focus only on Spearman distance. Despite these specializations, the algorithms are available also for other distance functions mentioned earlier, and easily extendable to any right-invariant distance.

Distance functions and $Z_n(\alpha, \rho)$

The partition function $Z_n(\alpha, \rho)$ represents the main obstacle for performing inference in the MM. In principle, it can be solved numerically by summing $e^{-\frac{\alpha}{n}d(\mathbf{r}, \rho)}$ over the $n!$ rankings, $\mathbf{r} \in \mathcal{P}_n$. However, the computational time of this calculation increases more than exponentially with the number of items, and thus is not feasible for large n .

Yet, if the distance function, $d(\cdot, \cdot)$, is right-invariant (Diaconis 1988), $Z_n(\alpha, \rho)$ does not depend on the location parameter ρ .

Definition 1. (*Right-invariant distance*). A distance function is right-invariant, if $d(\rho, \sigma) = d(\rho\eta, \sigma\eta)$ for all $\eta, \rho, \sigma \in \mathcal{P}_n$. With $\rho\eta$ we denote the composition function of two permutations, $\rho, \eta \in \mathcal{P}_n$, which is defined as $\rho \circ \eta = \rho\eta = \rho_\eta = (\rho_{\eta_1}, \dots, \rho_{\eta_n})$.

A right-invariant distance is independent on any relabeling of the items, which is a natural assumption when dealing with rankings. Consider for example 4 students (i.e. the items), $\{A_1, A_2, A_3, A_4\}$, admitted in a PhD program with the ranking, $\rho_1 = (1, 3, 4, 2)$, and later in that same year ranked according to their performance in the PhD program, $\rho_2 = (3, 4, 1, 2)$. The distance between ρ_1 and ρ_2 can be thought of as a measure of the goodness of judgement of the PhD admission board. If the students are now relabelled in a different ordering, for example $\{A_4, A_2, A_1, A_3\}$, the two rankings are now permuted according to the different labeling as $\rho_1\eta = (2, 3, 1, 4)$ and $\rho_2\eta = (2, 4, 3, 1)$, where

$\eta = (4, 2, 1, 3)$ determines the relabelling of the students. It is however natural to assume that the distance between the initial ranking and the final one is the same under the two labelings, that is, $d(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) = d(\boldsymbol{\rho}_1\eta, \boldsymbol{\rho}_2\eta)$, because the situation depicted is the same (see also Table 1.1).

	A_1	A_2	A_3	A_4	→		A_4	A_2	A_1	A_3
$\boldsymbol{\rho}_1$	1	3	4	2		$\boldsymbol{\rho}_1\eta$	2	3	1	4
$\boldsymbol{\rho}_2$	3	4	1	2		$\boldsymbol{\rho}_2\eta$	2	4	3	1

Table 1.1: An example of right invariance.

Given $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2 \in \mathcal{P}_n$, for a right-invariant distance it holds $d(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) = d(\boldsymbol{\rho}_1\boldsymbol{\rho}_2^{-1}, \mathbf{1}_n)$, where $\mathbf{1}_n = (1, 2, \dots, n)$, from which it follows that $Z_n(\alpha, \boldsymbol{\rho})$ is independent on the latent consensus ranking $\boldsymbol{\rho}$, as

$$Z_n(\alpha, \boldsymbol{\rho}) = \sum_{\mathbf{r} \in \mathcal{P}_n} e^{-\frac{\alpha}{n}d(\mathbf{r}, \boldsymbol{\rho})} = \sum_{\mathbf{r} \in \mathcal{P}_n} e^{-\frac{\alpha}{n}d(\mathbf{r}\boldsymbol{\rho}^{-1}, \mathbf{1}_n)} = \sum_{\mathbf{r}' \in \mathcal{P}_n} e^{-\frac{\alpha}{n}d(\mathbf{r}', \mathbf{1}_n)}. \quad (1.11)$$

When $d(\cdot, \cdot)$ is right-invariant, we can thus write $Z_n(\alpha, \boldsymbol{\rho}) = Z_n(\alpha, \mathbf{1}_n) = Z_n(\alpha)$. All distances introduced in this section, and considered in this thesis, are right-invariant.

For some choices of right-invariant distances, the partition function has the additional advantage of being available in closed form. For this reason, most work has been limited to the MM with Kendall distance (Lu and Boutilier 2014, Meilă and Chen 2010), with Hamming distance (Irurozki et al. 2014), and with Cayley distance (Irurozki et al. 2016b), for which the closed form of $Z_n(\alpha)$ was given in Fligner and Verducci (1986).

Still, there are important and natural right-invariant distances for which the computation of the partition function is not feasible, in particular the footrule and the Spearman's distances. One of the contributions of this thesis, is to give a strategy to compute $Z_n(\alpha)$ exactly, in case of footrule and Spearman distances for moderate values of n (see Section 2.2.1). When the partition function is needed for larger values of n , we propose an importance sampling scheme which approximates $Z_n(\alpha)$ to an arbitrary precision (see Section 2.2.2). The approximation is performed off-line over a grid for α , given n , since $Z_n(\alpha)$ is free of $\boldsymbol{\rho}$.

An asymptotic approximation of $Z_n(\alpha)$, when $n \rightarrow \infty$, has been studied in Mukherjee (2016), where the author proposed an Iterative Proportional Fitting Procedure (IPFP) to numerically compute it. In Section 2.2.2, we report a comparison between our Importance Sampling procedure and the Mukherjee's proposal to approximate the partition function

$Z_n(\alpha)$ for the Mallows footrule model, and show that both methods work very well.

Some recent developments of the Mallows model

Since [Mallows \(1957\)](#) and [Diaconis \(1988\)](#), many generalizations of the Mallows models emerged, extending the main model of eq. (1.6) to handle different kind of data, like partial rankings ([Critchlow 2012](#), [Lebanon and Mao 2008](#), [Jacques and Biernacki 2014](#)), pair comparisons ([Lu and Boutilier 2014](#)), and heterogeneous data ([Murphy and Martin 2003](#), [Busse et al. 2007](#)). The majority of these approaches focuses on the Kendall distance only, and the few ones that handle also other distances, generally apply the methodology to datasets with very small number of items n . For example [Murphy and Martin \(2003\)](#) studied mixtures of Mallows with Kendall, footrule and Cayley distances, applying their method to the benchmark American psychological association election data set ([Diaconis 1988](#)), where only $n = 5$ candidates (items) are ranked. The difficulties in the computation of the partition function for the footrule distance, which arise for larger values of n , were therefore not discussed in the paper.

In the frequentist framework, the MM with other distances than Kendall was studied in [Iruozki et al. \(2014\)](#) and in [Iruozki et al. \(2016b\)](#), who also developed the `PerMallows` R package ([Iruozki et al. 2016a](#)). In all these works however neither the footrule, nor the Spearman distances were considered, which are the two distances for which the computation of the partition function is not straightforward.

[Lu and Boutilier \(2014\)](#) propose the Generalized Repeated Insertion Model (GRIM), based on the Mallows with Kendall distance only, that extends the Repeated Insertion Method (RIM) of [Doignon et al. \(2004\)](#), a technique for unconditional sampling of the MM. The authors perform maximum likelihood estimation of the consensus ranking from pairwise comparisons. They also allow for multi-modality in the data, and perform preference learning and prediction. Their approach is related to our extension to pairwise preference data (Section 2.3.2), but differs notably in the model and algorithm. In particular, [Lu and Boutilier \(2014\)](#) do not provide any strategy to deal with uncertainty quantification for their estimates; our target instead is the full posterior distribution of the unknown consensus ranking. Yet, the fact that for the uniform prior the maximum a posteriori (MAP) estimates and the ML estimates coincide, establishes a natural link between these inferential targets. Further, two of our illustrations, reported in Sections 2.4.3 and 2.4.4, use the same datasets as in [Lu and Boutilier \(2014\)](#).

One of the most significant extensions of the Mallows model is perhaps the Generalized Mallows Model (GMM) of [Fligner and Verducci \(1986\)](#), that allows for item-specific dispersion parameters. Instead of a single dispersion parameter α , it considers a vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{n-1})$ of $n - 1$ dispersion parameters, each acting on a particular position of the permutation. This work has become increasingly popular, because of its flexibility and computational tractability, especially for Kendall distance. In [Fligner and Verducci \(1986\)](#), the GMM was set for Kendall and Cayley distances, and [Meilă and Chen \(2010\)](#), [Meilă and Bao \(2010\)](#) studied the GMM with Kendall distance in the Bayesian framework.

1.1.4 Multistage ranking models: the Plackett-Luce model

The Plackett-Luce (PL) ([Luce 1959](#), [Plackett 1975](#)) model for rankings differs from the three classes of models in the previous sections in being a multistage model. The main idea of a such a model is that the ranking process can be decomposed into a sequence of independent stages, that serve to sequentially arrange the items from the top to the bottom.

The PL model assumes that, given a score μ_i , $i = 1, 2, \dots, n$, corresponding to each item A_i , an ordering \mathbf{X} arises through the following process: the top ranked item, X_1 , is chosen with probability

$$\frac{\mu_1}{\sum_{i=1}^n \mu_i};$$

then the second to the top item, X_2 , is chosen, among the remaining items, with probability

$$\frac{\mu_2}{\sum_{i=2}^n \mu_i};$$

the process continues until a full ordering is formed, so that the probability density of an ordering $\mathbf{X} = (X_1, \dots, X_n)$ is

$$P(\mathbf{X}|\boldsymbol{\mu}) = \prod_{i=1}^{n-1} \frac{\mu_i}{\sum_{j=i}^n \mu_j}.$$

Notice that, in view of the one-to-one correspondence between ordering and ranking vectors (see Section 1.1), the previous density holds also for a ranking, \mathbf{R} , after simple manipulations.

Inferring the parameters of the PL model is typically done by maximum likelihood estimation, using a Majorize-Minimization algorithm ([Hunter 2004](#)), and mixtures of PL

models in a maximum likelihood framework for clustering were used in [Gormley and Murphy \(2006\)](#). A Bayesian approach was considered by [Guiver and Snelson \(2009\)](#) and [Caron and Doucet \(2012\)](#). [Caron and Teh \(2012\)](#) developed a nonparametric extension of the PL model that is able to handle an infinite number of items, and generalizes to time-dependent preference probabilities. Their framework is formalized in [Caron et al. \(2014\)](#), where a Dirichlet process mixture is used to cluster assessors based on their preferences. Recently, [Mollica and Tardella \(2016a\)](#) propose a Bayesian finite mixture of PL models to account for unobserved sample heterogeneity of partially ranked data, and develop the efficient PLMIX R package ([Mollica and Tardella 2016b](#)), that focuses on Bayesian inference of the PL model and its extension within the finite mixture approach.

Some of the models mentioned in the previous sections can be seen as belonging to this class: for instance the ϕ -model and the GMM (see [Marden 1995](#), for details). The Bradley Terry model ([Bradley and Terry 1952](#)), which will be the topic of the next section, is also a special case of the PL, where the generative process implies that the above probabilities only depend on pairs of scores. As such, it can be classified as a multistage model induced by paired comparisons.

The parameters in the PL model are continuous, which gives it a lot of flexibility relative to the Mallows model, which instead has a location parameter that takes value in the discrete parameter space consisting of all $n!$ permutations of the integers $1, \dots, n$. However, compared to the PL model, the Mallows model has the advantage of being flexible in the choice of the distance function between permutations.

1.2 Probabilistic models on pair comparison data

Paired comparison data originate from the comparison of items in pairs. This type of data arises for instance when the perception of a user is involved: it is easier for people to compare items in pairs rather than ranking all of them. Another situation that gives rise to pair comparisons is tournament data, with a game between two players or teams interpreted as a pair comparison, and win as preference. These models are then designed for real pair preferences (RPP), but can be obviously used also with derived pair preferences (DPP), as explained in Section 1.1.

The two traditional probabilistic models for paired comparison data are the [Thurstone \(1927\)](#) and the [Bradley and Terry \(1952\)](#) models. Based on these, many extensions arose

in the past decades, mostly in the econometric and psychometric literatures.

Denote, as in Section 1.1.2, the probability that item A_i is preferred to item A_k as $p_{ik} = P(A_i \prec A_k)$, and suppose that it can be expressed in the parametric form $P(A_i \prec A_k | \mathbf{u})$, where $\mathbf{u} = (u_1, \dots, u_n)$ is a latent vector of item specific score parameters. Two classical models for pairwise comparisons arise if this probability has the form $F(u_i - u_k)$, where F is a CDF. When F is normal, we recover the [Thurstone \(1927\)](#) model, while if F is logistic CDF, then the [Bradley and Terry \(1952\)](#) arises. Notice the connection between these models and both the order statistic models introduced in Section 1.1.1, and the paired comparison models of Section 1.1.2, which use these pair comparison probabilities as building blocks for defining probabilistic models for ranking data. Specifically, notice that the Bradley-Terry (BT) pair probability can be equivalently expressed in the form,

$$\Pr(A_i \prec A_k | \boldsymbol{\mu}) = \frac{\mu_i}{\mu_i + \mu_k}, \quad (1.12)$$

where $\mu_i = e^{u_i}$, $i = 1, \dots, n$, which was already introduced in Section 1.1.2.

The key assumption of the models for pair comparison data is that all pairwise probabilities are conditionally independent given \mathbf{u} (or $\boldsymbol{\mu}$), and that they depend only on the relative sizes of the corresponding score parameters.

The BT model has been deeply studied and extended since its first appearance, and many scholars have generalized it in several directions (see e.g. [Davidson 1970](#), [Agresti 1996](#), [Wu et al. 2015](#)). Maximum likelihood estimation is typically performed through iterative algorithms ([Zermelo 1929](#)) and MM algorithm ([Hunter 2004](#)). In the Bayesian framework the more convincing approach was developed by [Caron and Doucet \(2012\)](#), who proposed an efficient Gibbs sampling based on a clever data augmentation scheme.

The BT model has a main drawback: it suffers when the data are very sparse, and in particular when the strong connection condition ([Ford 1957](#)) fails. This condition guarantees the existence and uniqueness of the MLE of the BT parameters (similarly for Thurstone). This condition is equivalent to the property that for any partition of the items into two sets, some items in the second set has been preferred to some items in the first set at least once by some user, see [Yan \(2016\)](#). As a consequence, the posterior inference based on the BT, will, in such case, be driven by the prior density, as shown by [Yan \(2016\)](#) with some examples. This effect, which we also see in our simulations (see Appendix 3.C), is one of the reasons why we chose to rely on the Mallows model to

analyze sparse preference data (see Chapter 3 and 4).

The BT model was also represented and fitted as a log-linear model [Dittrich et al. \(1998, 2002\)](#). In these works, the authors introduced user specific covariates into their framework, and extended it to the case of dependent pair comparisons. Building on [Dittrich et al. \(1998\)](#), [Francis et al. \(2010\)](#) further introduced random effects for each user in order to account for residual heterogeneity, that is not included in user-specific covariates. The authors proposed to treat the ranked data as a set of paired comparisons, and extend the model to allow for heterogeneity. As such, their data are in the form of derived pair comparison data (DPP), which, as already mentioned, are notably different from real paired comparison data (RPP). In particular, their data are in the form of full pairwise comparisons (i.e. they have all the possible $n(n - 1)/2$ pairs among n items) without non-transitivities. Therefore, their method cannot be used on our sparse data, that is, where each assessor provides a limited number of pairwise preferences, typically smaller than the maximum $n(n - 1)/2$, and is allowed to contradict herself, thus leading to non-transitive patterns in the data. The `prefmod` R package ([Hatzinger et al. 2012](#)) collects these results, and deals with maximum likelihood estimation of pair comparison data, also allowing for data with missing values, and the possibility to include user and item-specific covariates into the analysis.

An interesting literature that builds on the Thurstone's model is the psychometric one ([Bockenholt 1988](#), [Böckenholt 2001](#), [Böckenholt and Tsai 2001](#), [Böckenholt 2006](#)). In these works, the authors develop different generalizations of the Thurstone model, accounting for instance for multi-dimensional data, in case the items are evaluated with respect to multiple aspects, or introducing dependency among the observed pairs, by the inclusion of random effects in the model.

1.2.1 Non-transitivity

When dealing with real pairwise preference (RPP) data a major challenge arises: pairwise preferences are not always transitive ([Tversky 1969](#)). By transitivity, we mean that, for every triplet of items, $\{x, y, z\}$, $x \prec y$ and $y \prec z$ imply $x \prec z$. The data considered, indeed, may contain preferences of the form $x \prec y$, $y \prec z$ but $z \prec x$. Throughout the thesis, we refer to these pathological patterns as non-transitive, intransitive or inconsistent, interchangeably. Notice that the kind of non-transitivity that we consider in this thesis (Chapter 3) is only individual-level non-transitivity. A different type of non-transitivity

arises when aggregating preferences across assessors, as under Condorcet (Marquis of Condorcet 1785) or Borda (de Borda 1781) voting rules. Moreover, with individual-level non-transitivity, we don't mean situations where the same user repeatedly compares the same pair of items, sometimes giving different answers. We instead focus on sparse data, where each user compares at most once each pair of items, and where non transitivity arises because of later contradiction, exactly like in the circular triad $x \prec y$, $y \prec z$ but $z \prec x$.

It should be clear that given a set of pairwise preferences containing a non-transitive pattern, it is not straightforward to infer an ordering of the items, as one cannot readily order the items involved in the non-transitive pattern. This challenge is ignored in the statistical methods of Section 1.1, which, being models for ranking data, exclude inconsistencies because of the transitivity property of a ranking. As a matter of fact, the literature accounting for non-transitive pair comparisons is limited to models dealing with pair comparison data, and in general not always interested in the individual-level preferences.

Here we discuss briefly some of the approaches that deal with this issues, and explain why our proposal of Chapter 3 is innovative and different from the present works.

A first examples of works that deal with non-transitive pairs are two generalizations of the BT model: Causeur and Husson (2005), who proposed a two-dimensional BT model (that is, a BT model parametrized by a two-dimensional worth parameter vector), and Usami (2010), who proposed a multidimensional generalization of the same model. These methods are based on the assumption that non-transitive patterns arise because assessors compare the items based on different scales of judgement, and as such do not lead to a final linear ordering of all items, but rather to multiple linear orderings).

In the psychometric literature, an interesting paper is Tsai and Böckenholt (2008), where the authors introduce a general class of Thurstonian-like models that can account simultaneously for transitive choice behavior and systematic deviations from it. However inference is performed when the data include repeated comparisons for each user, and all items are compared by each user. Our method of Chapter 3 is instead specifically developed for sparse data (see also Chapter 4), where the users perform at most once each comparison.

Another interesting approach to non-transitivity was presented by Volkovs and Zemel (2014), who developed a score-based model for pairwise preferences that generalizes the

PL model (see Section 1.1.4), known as the Multinomial Preference Model (MPM). The MPM deals with pairwise preferences, even non-transitive ones, and extends to supervised problems. Their method is connected to our logistic model for mistakes (see Section 3.1.2). The main difference between their model and ours is the data generating mechanism, which in Volkovs and Zemel (2014) is assumed to be a multinomial score based process, while our proposal assumes that the data are generated following a distance-based model between rankings. Moreover, one difficulty of the MPM is the use of gradient optimization in a non-convex problem (which could lead to local optima), and the somewhat arbitrary way of imputing missing ranks. In addition, their goal is to learn a single consensus ranking of the items, or multiple consensus rankings in case of clustering. Our method instead has the ability to further learn the individual latent rankings for each user.

A recent development to handle non-transitive pair data is Ding et al. (2015), who propose a novel mixed membership of Mallows models (M4) for dealing with noisy pair preference data, which generalizes the mixture model by Lu and Boutilier (2014). Their proposal is connected to our method of Chapter 3, in that they also postulate the existence of a latent variable, but differs most notably in three points: (a) they model the presence of non-transitive patterns in the data as arising because each user has multiple latent linear orderings; (b) they focus on the Kendall distance only, and (c) they assume a separability property, which means that each of the Mallows components must have at least one characterizing item pair, say (A_i, A_k) , such that, with very high probability, $(A_i \prec A_k)$ in that component of the mixture, whereas $(A_k \prec A_i)$ holds in the other Mallows components. Our idea instead is to rely on a mistake model for explaining the non-transitive patterns in the data, which can handle every right-invariant distance, and does not postulate any separability condition, which can be violated in practice. Our model for dealing with data showing non-transitive patterns, which is completely integrated in the Bayesian Mallows model of Chapter 2, is outlined in Chapter 3. We believe that our method is the first to exploit the Mallows model, in its general formulation - that is, based on a generic right-invariant distance - for non-transitive pair comparison data, and stands out as the only approach to non-transitive pair data, when the individual hidden rankings are of interest, the pairs are not repeatedly assessed by each user and are few, and a Bayesian model based approach is of interest.

In the machine learning community, the problem of finding a ranking based on possibly non-transitive pairwise comparisons is known as the minimum feedback arc set prob-

lem in a digraph (MFAS). Consider a digraph, or directed graph, $D = (V, E)$, where $V = \{v_1, \dots, v_n\}$ is the set of its vertices, and $E = \{e_1, \dots, e_M\}$ is the set of directed edges, that is, $e_m = (v_i, v_k)$, $v_i, v_k \in V$, $m = 1, \dots, M$, indicates that there is an edge from v_i to v_k . A feedback arc set of D is a (possibly empty) subset of arcs $E^* \subset E$ whose removal makes the graph acyclic, that is, $D^* = (V, E \setminus E^*)$ is a directed acyclic graph (DAG). The minimum feedback arc set problem consists in finding the smallest (in the sense of minimum cardinality) feedback arc set, and is a well-known NP-hard problem. The analogy between the MFAS problem and the problem of finding a ranking based on non-transitive pairwise comparisons is clear if one represents the pairwise preferences between items as directed arcs in a digraph whose vertices coincide with the items to be ranked. For example the set of preferences $\mathcal{B} = \{A_2 \prec A_1, A_5 \prec A_4, A_5 \prec A_3, A_5 \prec A_2, A_5 \prec A_1, A_3 \prec A_2, A_1 \prec A_3\}$ is represented as a directed graph in Figure 1.2, left.

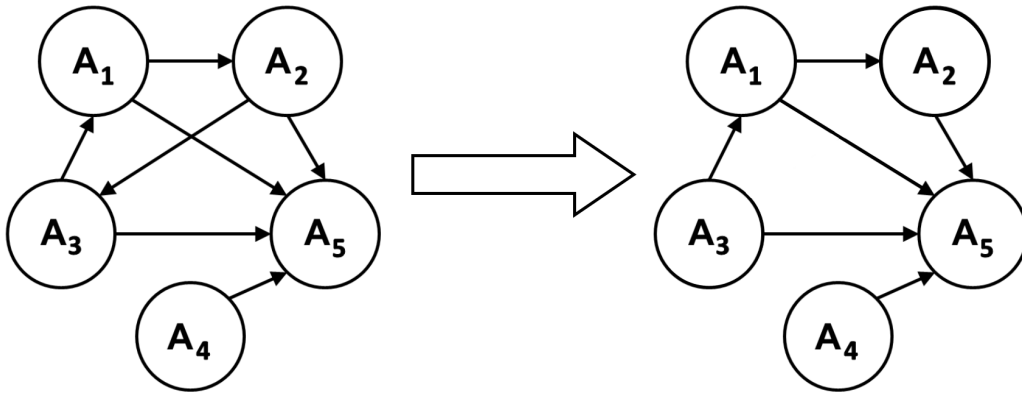


Figure 1.2: Left: Directed graph representing the set of preferences \mathcal{B} . Right: DAG resulting from the removal of the arc (A_2, A_3)

The non-transitive pattern $A_1 \prec A_2 \prec A_3 \prec A_1$, is then represented as a cycle in the directed graph. In this simple illustration, one possible solution of the MFAS is $E^* = \{(A_2, A_3)\}$. As a matter of fact, removing this edge results in the DAG of Figure 1.2, right. Equivalently, the reduced set of preferences $\mathcal{B} \setminus \{A_3 \prec A_2\}$ is transitive.

In this literature relevant papers are [Kenyon-Mathieu and Schudy \(2007\)](#), [Ailon \(2012\)](#), who both aim at finding the linear ordering of items with the smallest number of disagreements with the preferences of the data, by using methods from combinatorial optimization. As a consequence, these models are not probabilistic, and cannot be used for expressing uncertainty in the derived estimates.

1.3 A short detour into the machine learning approaches for ranking data

We can distinguish two main classes of models among the machine learning approaches that deal with ranking data. First, those pertaining to the area of learning to rank (LETOR) or rank aggregation, whose aim is to learn the best objective ranking (the analog of the consensus ranking of the Mallows model) from data regarding user preferences. The second class is the one of personalization and preference elicitation in recommender systems, where the interest is usually in learning individual preferences of users who have distinct preferences.

The first works dealing with the rank aggregation problem, which came out way before they were popularized by the machine learning community, regarded the political theory of elections. Indeed, in the late 18th century, Condorcet ([Marquis of Condorcet 1785](#)) and Borda ([de Borda 1781](#)) studied a way to aggregate political preferences in elections with more than two candidates, basically designing the first methods to aggregate rankings from noisy data. In the last decades, rank aggregation has been studied from a mathematical perspective, starting with the work of [Kemeny and Snell \(1962\)](#), who proposed a precise criterion for determining the *best* aggregate ranking, that is the one that minimizes the number of pairwise disagreements with the (aggregated) data. Recent applications of rank aggregation include sport tournaments ([Glickman 1999](#)), social choice theory ([Bartholdi et al. 1989a](#)), peer grading in Massive Open Online Courses (MOOC) ([Raman and Joachims 2015](#)) and, importantly, web ranking applications (like the Yahoo! Learning to Rank Challenge). The problem has then been studied from a computational perspective: [Bartholdi et al. \(1989b\)](#) showed that finding the Kemeny optimal ranking is NP-hard, which motivated the recent works aimed at finding approximations to the rank aggregation criteria ([Ali and Meilă 2012](#), [Dwork et al. 2001](#), [Kenyon-Mathieu and Schudy 2007](#)), also from pairwise comparisons ([Hüllermeier et al. 2008](#), [Liu et al. 2009](#), [Negahban et al. 2012](#), [Rajkumar et al. 2015](#), [Shah et al. 2015](#)).

A relevant research field linked to rank aggregation is the so-called label ranking ([Fürnkranz and Hüllermeier 2010](#)), which investigates the problem of learning a mapping from items to rankings over a finite number of predefined items' labels. The interest is then in assigning a complete preference order - that is, a ranking - of labels over the set of items, in order to perform prediction and classification. The label ranking literature is

huge and, as can be evinced by its definition, intimately linked to the Mallows model. As a matter of fact, one way of dealing with label ranking is to assume a probability model on rankings, such as the Mallows model, for performing learning and inference ([Cheng and Hüllermeier 2008](#)).

The second class of works dealing with ranking data, the area of personalized recommendation, aims at assessing individual or group rankings, by modeling the heterogeneity of user preferences. With the rise of e-commerce, many commercial websites are now using recommender systems to suggest their users products they may like. The most successful approach so far is the well-known collaborative filtering (CF), whose principal idea is to identify users with similar tastes and use them to generate the recommendations. CF is grounded on matrix factorization in reduced dimensional spaces and is thus related to singular value decomposition and principal component analysis. In this area lies the Netflix competition, which has started a massive research on predicting a user's movie ratings given the ratings for other movies, including both their own and those of other users. Recommendations in this area are based on the idea that a set of users which liked the same items in the past, will probably share the same preferences in the future.

The collaborative ranking subclass of CF seeks to predict the ranking of the items, and to perform recommendations based on this. In these works, it is usually postulated a global preference structure (a consensus ranking) then used to link the users' preferences ([Rendle et al. 2009](#), [Lu and Negahban 2015](#), [Park et al. 2015](#)). The main problem that this literature faces is to estimate a binary variable (a yes/no answer, for example whether a user will like an item or not), in order to make personalized recommendations. As such, their aim is often not to obtain a final full ranking over the items, neither global nor individual, which is the main task we consider.

Chapter 2

The Bayesian Mallows model for ranking data

In this chapter we develop a Bayesian framework for inference in the Mallows model. The method is potentially able to handle any right-invariant distance. This has the theoretical interest to exploit the main advantage of the Mallows model, namely its flexibility in the choice of the distance. Inference is based on a Metropolis-Hastings algorithm, which converges to the posterior distribution of the parameters of the Mallows model, if the exact partition function is available. In case the exact partition function is not available, we propose to approximate it using an off-line importance sampling scheme, and we document the quality and efficiency of this approximation. Using data augmentation techniques, our method extends to data in the form of incomplete rankings, like top-k rankings, and pairwise comparisons, and can be easily adapted to the case of ranks missing at random. In case of heterogeneous assessors, we develop a mixture model which embeds the Bayesian Mallows model. Our approach unifies clustering, classification and preference prediction in a single inferential procedure, thus leading to coherent posterior credibility levels of the learned parameters. The Bayesian setting indeed allows to naturally compute complex probabilities of interest, like the probability that an item has consensus rank higher than a given level, or the probability that the consensus rank of an item is higher than that of another item of interest. For incomplete rankings this can be performed also at the individual assessor level, allowing for individual recommendations.

This chapter contains joint work with Valeria Vitelli, Øystein Sørensen, Arnoldo Frigessi and Elja Arjas and is based on [Vitelli et al. \(2017\)](#).

Outline

In Section 2.1, we introduce the Bayesian Mallows model for complete rankings. In Section 2.1.1 we discuss the choice of the prior distributions, Sections 2.1.2 and 2.1.3, are devoted to show how efficient Bayesian computation can be performed for this model, and tuning of the hyperparameters is discussed in Section 2.1.4. In Section 2.2 we develop and test an importance sampling scheme for computing the partition function, based on a pseudo-likelihood approximation of the Mallows model. After a short section regarding the exact computation of the partition function (Section 2.2.1), we carefully test and study the importance sampling approximation of the partition function, and its effect on inference, both theoretically and by simulations (Section 2.2.2). In Section 2.2.3, we then present an extensive comparison with heuristic and ML approaches, performed on simulated complete data. Section 2.3 is dedicated to the model extensions. In Section 2.3.1 we extend the Bayesian Mallows approach to partial rankings, and we prove some results on the effects of unranked items on the consensus ranking. Section 2.3.2 considers data in the form of pairwise comparisons. In Sections 2.3.3 and 2.3.4 we describe mixture models dealing with heterogeneous assessors expressing full rankings or pairwise comparisons, able to find cluster-specific consensus rankings. Section 2.3.5 presents an example of preference prediction in our framework, based on data simulated from a realistic setup, which calls both for cluster assignment and individual preference learning. We show that our approach works well in a simulation context. In Section 2.4, we then show some illustrations of the performance of our method on real data: the selected case studies exemplify the different incomplete data situations considered. The Meta-Analysis (Section 2.4.1) is a case of very sparse top-k rankings, the Beaches data (Section 2.4.2) consist of pairwise comparisons of an homogeneous data sample, the Sushi (Section 2.4.3) is an example of complete rankings with clusters, and the Movielens benchmark data (Section 2.4.4) is an illustration of pairwise comparisons with clusters. Section 2.5 provides a discussion and directions for future research.

2.1 The Bayesian Mallows model for full rankings

We here follow the notation introduced earlier in Section 1.1.

Let $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$, be a set of n items, and assume that N assessors rank all items individually with respect to a considered feature. The ordering provided by assessor j is here denoted by \mathbf{X}_j , and the observations $\mathbf{X}_{1:N} = \mathbf{X}_1, \dots, \mathbf{X}_N$ are N permutations of the labels in \mathcal{A} . The observations usually come directly in the form of rankings, and denoted by $\mathbf{R}_{1:N} = \mathbf{R}_1, \dots, \mathbf{R}_N$. Then $\mathbf{R}_j = (R_{1j}, R_{2j}, \dots, R_{nj}) \in \mathcal{P}_n$, $j = 1, \dots, N$, denotes the full ranking of assessor j , and R_{ij} , $i = 1, \dots, n$, denotes the rank given to item A_i by assessor j . We assume that the N observed rankings $\mathbf{R}_1, \dots, \mathbf{R}_N \in \mathcal{P}_n$ are conditionally independent given the parameters α and $\boldsymbol{\rho}$, and that each of them is distributed according to the Mallows model, of equation (1.6), with these parameters. The likelihood of the data takes then the form

$$P(\mathbf{R}_1, \dots, \mathbf{R}_N | \alpha, \boldsymbol{\rho}) = \frac{1}{Z_n(\alpha)^N} \exp \left[-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right]. \quad (2.1)$$

where $d(\cdot, \cdot)$ is assumed right-invariant (see Section 1.1.3).

2.1.1 Prior distributions

To complete the specification of the Bayesian model for the rankings $\mathbf{R}_1, \dots, \mathbf{R}_N$, a prior for its parameters is needed. Commonly, the parameter of direct inferential interest is the consensus ranking $\boldsymbol{\rho}$. The scale parameter α has a more indirect role, in controlling variation between the individual rankings $\mathbf{R}_{1:N}$. We here assume a priori that α and $\boldsymbol{\rho}$ are independent.

An obvious choice for the prior for $\boldsymbol{\rho}$ in the context of the Mallows likelihood is to use the Mallows model family also in setting up a prior for $\boldsymbol{\rho}$, that is, let $\pi(\boldsymbol{\rho}) = \pi(\boldsymbol{\rho} | \alpha_0, \boldsymbol{\rho}_0) \propto \exp \left\{ -\frac{\alpha_0}{n} d(\boldsymbol{\rho}, \boldsymbol{\rho}_0) \right\}$. Here α_0 and $\boldsymbol{\rho}_0$ are fixed hyperparameters, with $\boldsymbol{\rho}_0$ specifying the ranking that is a priori thought most likely, and α_0 controlling the tightness of the prior around $\boldsymbol{\rho}_0$. Since α_0 is fixed, $Z_n(\alpha_0)$ is a constant. Note that combining the likelihood with the prior $\pi(\boldsymbol{\rho} | \alpha_0, \boldsymbol{\rho}_0)$ above has the same effect on inference as involving an additional hypothetical assessor $j = 0$, say, who then provides the ranking $\mathbf{R}_0 = \boldsymbol{\rho}_0$ as data, with α_0 fixed. If we were to elicit a value for α_0 , we could reason as follows. Consider, for $\boldsymbol{\rho}_0$ fixed, the prior expectation $g_n(\alpha_0) := E_{\pi(\boldsymbol{\rho})}[d(\boldsymbol{\rho}, \boldsymbol{\rho}_0) | \alpha_0, \boldsymbol{\rho}_0]$. Because of the assumed right

invariance of the distance $d(\cdot, \cdot)$, this expectation is independent of $\boldsymbol{\rho}_0$, which is why $g_n(\cdot)$ depends only on α_0 . Moreover, $g_n(\alpha_0)$ is obviously decreasing in α_0 . For the footrule and Spearman distances, which are defined as sums of item specific deviations $|\rho_{0i} - \rho_i|$ or $|\rho_{0i} - \rho_i|^2$, $g_n(\alpha_0)$ can be interpreted as the expected (average, per item) error in the prior ranking $\pi(\boldsymbol{\rho}|\alpha_0, \boldsymbol{\rho}_0)$ of the consensus. A value for α_0 can then be elicited by first choosing a target level τ_0 , say, which would realistically correspond to such an a priori expected error size, and then finding the value α_0 such that $g_n(\alpha_0) = \tau_0$. This procedure requires numerical evaluation of the function $g_n(\alpha_0)$ over a range of suitable α_0 values.

In this chapter and in Chapter 3, we employ only the uniform prior on \mathcal{P}_n , $\boldsymbol{\rho} \sim \mathcal{U}(\mathcal{P}_n)$, that is $\pi(\boldsymbol{\rho}) = (n!)^{-1} \mathbb{1}_{\mathcal{P}_n}(\boldsymbol{\rho})$, corresponding to $\alpha_0 = 0$, while in Chapter 5, we will go deeper in the previous proposal, and provide some alternatives for the elicitation of the prior on $\boldsymbol{\rho}$, in the particular case of the Mallows model with Cayley distance. In addition, in Chapter 6 we discuss the elicitation of a conjugate prior when the distance is set to Spearman.

For the scale parameter α , we use a truncated exponential prior, with density $\pi(\alpha|\lambda) = \lambda e^{-\lambda\alpha} \mathbb{1}_{[0, \alpha_{\max})}(\alpha) / (1 - e^{-\lambda\alpha_{\max}})$, where the cut-off point $\alpha_{\max} < \infty$ is large compared to the values supported by the data. In practice, in the computations involving sampling of values for α , truncation was never applied. We show in Section 2.2.2, using simulated data, that inference on $\boldsymbol{\rho}$ is almost completely insensitive on the choice of λ . A theoretical argument for this fact is provided in that same section, although it is tailored more specifically to the numerical approximations of $Z_n(\alpha)$. For these reasons, in all our data analyses, we assign λ a fixed value. We chose small values for λ , typically of the order of 10^{-2} , thus implying a prior density for α which is quite flat in the region supported in practice by the likelihood. If a more elaborate elicitation of the prior for α was preferred, this could be achieved by computing, by numerical integration, values of the function $E_{\pi(\alpha)}[g_n(\alpha)|\lambda]$, selecting a realistic target τ , and solving $E_{\pi(\alpha)}[g_n(\alpha)|\lambda] = \tau$ for λ . In a similar fashion as earlier, also $E_{\pi(\alpha)}[g_n(\alpha)|\lambda]$ can be interpreted as an expected (average, per item) error in the ranking, but now *errors* are meant as those made by the assessors, relative to the consensus, and expectation is with respect to the exponential prior $\pi(\alpha|\lambda)$.

2.1.2 Inference

Having assumed prior independence between $\boldsymbol{\rho}$ and α , and given the prior densities $\pi(\boldsymbol{\rho})$ and $\pi(\alpha)$ as in Section 2.1.1, the posterior distribution for $\boldsymbol{\rho}$ and α is given by

$$P(\boldsymbol{\rho}, \alpha | \mathbf{R}_1, \dots, \mathbf{R}_N) \propto \frac{\pi(\boldsymbol{\rho}) \pi(\alpha)}{Z_n(\alpha)^N} \exp \left[-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right]. \quad (2.2)$$

In applications, often the interest is in computing posterior summaries of (2.2). One such summary is the marginal posterior mode of $\boldsymbol{\rho}$, (that is, the maximum a posteriori, MAP) of (2.2), which does not depend on α and, in case of uniform prior on $\boldsymbol{\rho}$, coincides with the maximum likelihood estimator. The marginal posterior distribution of $\boldsymbol{\rho}$ is in fact given by

$$P(\boldsymbol{\rho} | \mathbf{R}_1, \dots, \mathbf{R}_N) \propto \pi(\boldsymbol{\rho}) \int \frac{\pi(\alpha)}{Z_n(\alpha)^N} \exp \left[-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right] d\alpha. \quad (2.3)$$

Given the data $\mathbf{R}_{1:N}$, for varying consensus ranking $\boldsymbol{\rho}$, the sum of distances, $T(\boldsymbol{\rho}, \mathbf{R}_{1:N}) = \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho})$, takes only a finite set of discrete values $\{t_1, t_2, \dots, t_m\}$, where m depends on the distance chosen, $d(\cdot, \cdot)$, on the sample size N , and on n . Therefore, the set of all permutations \mathcal{P}_n can be partitioned into the sets $H_i = \{\mathbf{r} \in \mathcal{P}_n : T(\mathbf{r}, \mathbf{R}_{1:N}) = t_i\}$ for each distance t_i . These sets are level sets of the posterior marginal distribution in (2.3), as all $\mathbf{r} \in H_i$ have the same posterior marginal probability. The level sets do not depend on α but the posterior distribution shared by the permutations in each set does.

Sometimes the interest is in computing posterior probabilities of more complex functions of $\boldsymbol{\rho}$, for example the posterior probability that a certain item i has consensus rank lower than a given level k , $P(\rho_i < k | \text{data})$, or that a certain item i_1 is ranked higher than another one, i_2 , in the consensus, $P(\rho_{i_1} < \rho_{i_2} | \text{data})$. These probabilities cannot be readily obtained within the maximum likelihood approach, while the Bayesian setting very naturally allows to approximate any posterior summary of interest, by means of a Markov Chain Monte Carlo algorithm, which at convergence samples from (2.2).

2.1.3 Metropolis-Hastings algorithm for full rankings

In order to obtain samples from the posterior density of equation (2.2), we iterate between two steps. First we update the consensus ranking, $\boldsymbol{\rho}$, by proposing $\boldsymbol{\rho}'$ according to a

distribution which is centered around the current rank ρ .

Definition 2. (*Leap-and-Shift Proposal, L-S*). Fix an integer $L \in \{1, \dots, \lfloor (n-1)/2 \rfloor\}$ and draw a random number $u \sim \mathcal{U}\{1, \dots, n\}$. Define, for a given ρ , the set of integers $\mathcal{S} = \{\max(1, \rho_u - L), \min(n, \rho_u + L)\} \setminus \{\rho_u\}$, $\mathcal{S} \subseteq \{1, \dots, n\}$, and draw a random number r uniformly in \mathcal{S} . Let $\rho^* \in \{1, 2, \dots, n\}^n$ have elements $\rho_u^* = r$ and $\rho_i^* = \rho_i$ for $i \in \{1, \dots, n\} \setminus \{u\}$, constituting the leap step.

Now, defining $\Delta = \rho_u^* - \rho_u$, the elements of the proposed $\rho' \in \mathcal{P}_n$ are defined, for all $i = 1, \dots, n$, as

$$\rho'_i = \begin{cases} \rho_u^* & \text{if } \rho_i = \rho_u \\ \rho_i - 1 & \text{if } \rho_u < \rho_i \leq \rho_u^* \text{ and } \Delta > 0 \\ \rho_i + 1 & \text{if } \rho_u > \rho_i \geq \rho_u^* \text{ and } \Delta < 0 \\ \rho_i & \text{otherwise .} \end{cases}$$

This constitutes the shift step.

Proposition 1. The L-S proposal $\rho' \in \mathcal{P}_n$ is a local perturbation of ρ , separated from ρ by a Ulam distance 1 .

Proof. From the definition and by construction, $\rho^* \notin \mathcal{P}_n$, since there exist two indices $i \neq j$ such that $\rho_i^* = \rho_j^*$. The shift of the ranks by Δ brings ρ^* to ρ' back into \mathcal{P}_n . The Ulam distance $d(\rho, \rho')$ counts the number of edit operations needed to convert ρ into ρ' , where each edit operation involves deleting a character and inserting it in a new place (see also Section 1.1.3). This is equal to 1, following [Gopalan et al. \(2006\)](#). \square

The L-S proposal is not symmetric, and the probability mass function associated to the transition is given by

$$\begin{aligned} P_L(\rho'|\rho) &= \sum_{u=1}^n P_L(\rho'|U=u, \rho) P(U=u) = \\ &= \frac{1}{n} \sum_{u=1}^n \mathbb{1}_{\{\rho_{-u}\}}(\rho_{-u}^*) \mathbb{1}_{\{0 < |\rho_u - \rho_u^*| \leq L\}}(\rho_u^*) \left[\frac{\mathbb{1}_{\{L+1, \dots, n-L\}}(\rho_u)}{2L} + \sum_{l=1}^L \frac{\mathbb{1}_{\{l\}}(\rho_u) + \mathbb{1}_{\{n-l+1\}}(\rho_u)}{L+l-1} \right] \\ &+ \frac{1}{n} \sum_{u=1}^n \mathbb{1}_{\{\rho_{-u}\}}(\rho_{-u}^*) \mathbb{1}_{\{|\rho_u - \rho_u^*| = 1\}}(\rho_u^*) \left[\frac{\mathbb{1}_{\{L+1, \dots, n-L\}}(\rho_u^*)}{2L} + \sum_{l=1}^L \frac{\mathbb{1}_{\{l\}}(\rho_u^*) + \mathbb{1}_{\{n-l+1\}}(\rho_u^*)}{L+l-1} \right], \end{aligned}$$

where $\rho_{-u} = \{\rho_i; i \neq u\}$. The acceptance probability of ρ' in the M-H algorithm is then

$\min\{1, \eta_\rho\}$, where

$$\eta_\rho = \frac{P_L(\boldsymbol{\rho}|\boldsymbol{\rho}')\pi(\boldsymbol{\rho}')}{P_L(\boldsymbol{\rho}'|\boldsymbol{\rho})\pi(\boldsymbol{\rho})} \exp \left\{ -\frac{\alpha}{n} \sum_{j=1}^N [d(\mathbf{R}_j, \boldsymbol{\rho}') - d(\mathbf{R}_j, \boldsymbol{\rho})] \right\}. \quad (2.4)$$

The term $\sum_{j=1}^N [d(\mathbf{R}_j, \boldsymbol{\rho}') - d(\mathbf{R}_j, \boldsymbol{\rho})]$ in (2.4) can be computed efficiently, since most elements of $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ are equal. Let $\rho_i = \rho'_i$ for $i \in E \subset \{1, \dots, n\}$, and $\rho_i \neq \rho'_i$ for $i \in E^c$. When $d(\cdot, \cdot)$ is the footrule or the Spearman distance, we have, for $p \in \{1, 2\}$ respectively,

$$\sum_{j=1}^N [d(\mathbf{R}_j, \boldsymbol{\rho}') - d(\mathbf{R}_j, \boldsymbol{\rho})] = \sum_{j=1}^N \left\{ \sum_{i \in E^c} |R_{ij} - \rho'_i|^p - \sum_{i \in E^c} |R_{ij} - \rho_i|^p \right\}. \quad (2.5)$$

For the Kendall distance, instead, we get

$$\begin{aligned} & \sum_{j=1}^N [d(\mathbf{R}_j, \boldsymbol{\rho}') - d(\mathbf{R}_j, \boldsymbol{\rho})] = \\ &= \sum_{j=1}^N \sum_{1 \leq k < l \leq n} \{ \mathbb{1} [(R_{kj} - R_{lj})(\rho'_k - \rho'_l) > 0] - \mathbb{1} [(R_{kj} - R_{lj})(\rho_k - \rho_l) > 0] \} = \\ &= \sum_{j=1}^N \sum_{k \in E^c \setminus \{n\}} \sum_{l \in \{E^c \cap \{l > k\}\}} \mathbb{1} \{ [(R_{kj} - R_{lj})(\rho'_k - \rho'_l) > 0] - \mathbb{1} [(R_{kj} - R_{lj})(\rho_k - \rho_l) > 0] \}. \end{aligned}$$

Hence, by storing the set E^c at each MCMC iteration, the computation of (2.4) involves a sum over fewer terms, speeding up the algorithm consistently. The parameter L is used for tuning the acceptance probability (2.4).

In the second step we sample a proposal α' from a lognormal distribution $\ln \mathcal{N}(\ln \alpha, \sigma_\alpha^2)$, centered at the current value of α , and accept it with probability $\min\{1, \eta_\alpha\}$, where

$$\eta_\alpha = \frac{Z_n(\alpha)^N \pi(\alpha') \alpha'}{Z_n(\alpha')^N \pi(\alpha) \alpha} \exp \left[-\frac{(\alpha' - \alpha)}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right]. \quad (2.6)$$

The variance σ_α^2 can be tuned to obtain a desired acceptance probability.

A further parameter, named α_{jump} , can be used to update α only every α_{jump} updates of $\boldsymbol{\rho}$: the possibility to tune it ensures a better mixing of the MCMC in the different sparse data applications. The above described MCMC algorithm is summarized as Algorithm 1 of Appendix 2.A.

Proposition 2. *(Convergence of the MCMC algorithm). The MCMC Algorithm 1 using*

the exact partition function $Z_n(\alpha)$ samples from the Mallows posterior in equation (2.2), as the number of MCMC iterations tends to infinity.

Proof. Because of reversibility of the proposals, detailed balance holds for the Markov chain. Ergodicity follows by aperiodicity and positive recurrence. \square

2.1.4 Tuning the proposal distributions parameters

We here study the effect of the L-S proposal on η_ρ , and the tuning of its parameter L in relation to MCMC convergence, on simulated data. Also the role of parameter σ_α in the log-normal proposal for α , is briefly explored.

Data were generated from the Mallows model with footrule distance, $\alpha_{\text{true}} = 2$ and $\rho_{\text{true}} = (1, \dots, n)$. Two scenarios were used, with $n = 20$ and $n = 50$, because the choice of L would likely depend on the number of items. For generating the data, we run our MCMC sampler (see Appendix 2.B) for 10^5 burn-in iterations, and collected one sample every 100 iterations after that. We collected samples from $N = 500$ assessors. The data analyses were carried out by using the same distance as in the data generation (footrule), and the MCMC was run for 10^6 iterations after 10^5 iterations of burn-in, with a 1 to 100 thinning for α . 10 different chains were started from random points of the parameter space, and posterior inference was based on merging the results from these chains, as the MCMC converged to the same limit. The same analyses were also performed for Kendall distance, on data generated by using the `PerMallows` R package (Irurozki et al. 2016a). Equivalent results (not shown) as for the footrule were obtained.

In the MCMC, we controlled for (i) mixing, aiming at an acceptance rate of approximately 1/3 for each parameter (Gelman et al. 1996, Roberts et al. 1997), and (ii) autocorrelation, monitoring the Integrated Autocorrelation Time (IAT) τ (Green and Han 1992). Since ρ is multivariate, we monitored the IAT for each component of ρ . As expected, the acceptance rate η_ρ decreases with increasing L , and depends also on the value of n (Figure 2.1, top panels). Based on the results shown in Figure 2.1 (bottom panels), we propose as a rule of thumb that L should be set equal to $n/5$. This choice seems reasonable also from the perspective of η_ρ (Figure 2.1, upper panels). Not surprisingly, the acceptance rate η_α decreases with increasing σ_α (Table 2.1). Aiming at a value of η_α close to 1/3 sets us also close to the minimal value of τ_α . In the case of $n = 20$ values close to 0.2 appear to be good choices for σ_α , while for $n = 50$ values near 0.1 might be slightly preferred.

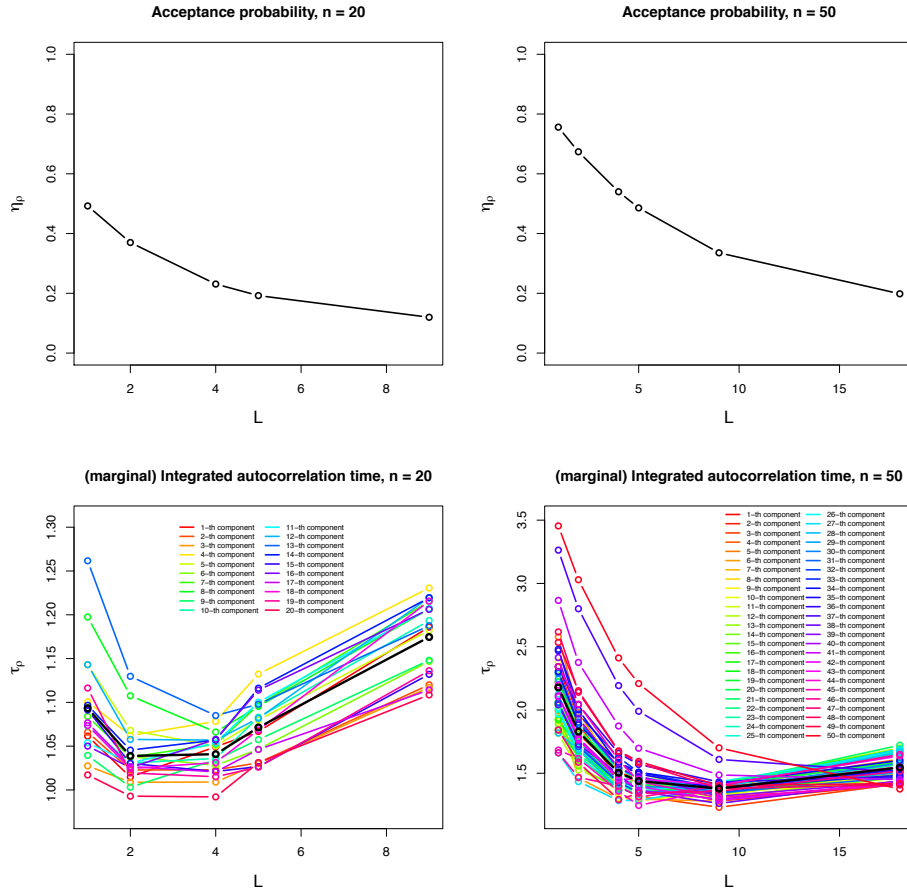


Figure 2.1: Results of the simulations described in Section 2.1.4. Top panels: acceptance probability η_ρ along MCMC iterations; bottom panels: marginal IAT τ_ρ of ρ . Left and right panels show the results when $n = 20$ and 50 , respectively.

	$n = 20$		$n = 50$	
σ_α	η_α	τ_α	η_α	τ_α
0.01	0.93	4.32	0.88	3.83
0.02	0.86	3.6	0.76	3.4
0.05	0.67	2.67	0.51	2.64
0.1	0.47	2.65	0.3	2.41
0.2	0.27	2.24	0.16	2.2
0.5	0.11	2.53	0.07	2.74

Table 2.1: Results of the simulations described in Section 2.1.4. Acceptance probability η_α and IAT τ_α of α along MCMC iterations, for two simulations with $n = 20$ and 50 . In each row, the value of σ_α (standard deviation of the log-normal proposal for α) used in the MCMC.

2.2 Approximating the partition function $Z_n(\alpha)$

For Kendall's, Hamming and Cayley distances, the partition function $Z_n(\alpha)$ is available in close form (Fligner and Verducci 1986), but this is not the case for footrule and Spearman

distances. In Section 2.2.1, we propose a strategy to compute $Z_n(\alpha)$ exactly, in case of footrule and Spearman distances for moderate values of n . For larger values of n , we propose an approximation of the partition function $Z_n(\alpha)$ based on importance sampling (Section 2.2.2). The main idea, motivated by the fact that $Z_n(\alpha)$ does not depend on ρ , is that we can approximate it off-line over a grid of α values, and then interpolate them, in order to yield an estimate over a continuous range. With this approximation, we can then read off the needed values to compute the acceptance probability η_α very rapidly. We study the convergence of the importance sampler theoretically and numerically, with a series of experiments aimed at demonstrating the quality of the approximation, and the impact of it in inference. In Section 2.2.3, we also provide a comparison among the inferential results obtained with our method and with existing competitors, based on simulated data.

We here describe two alternative methods for dealing with intractable normalizing constants in MCMC algorithms: the exchange algorithm, and the pseudo-marginal approaches. Contrarily to our approach, both methods would target the exact posterior distribution. We here explain why these methods can't be applied in our case in practice. The exchange algorithm (Møller et al. 2006, Murray et al. 2012) is based on the idea to include into the MCMC an auxiliary variable, which, if chosen appropriately, eliminates the intractable term from the M-H ratio. The assumption of this method is that we can draw independent, and exact samples from the proposal distribution. This is not the case in our model, as there are no algorithms available to exactly sample from the Mallows model with many of the distances considered (e.g. footrule and Spearman). The pseudo-marginal class of methods (Beaumont 2003, Andrieu and Roberts 2009), sometimes referred to as Exact-approximate methods, are based on the property that the invariant distribution of the Markov chain produced is the exact target distribution despite the use of an approximation in the Metropolis-Hastings acceptance probability. The idea is to replace the posterior density $P(\rho, \alpha | \mathbf{R}_1, \dots, \mathbf{R}_N)$, of eq. (2.2) with a non-negative unbiased estimator P^* , such that for some $C > 0$ it holds that $E[P^*] = CP$. The approximate acceptance ratio then uses P^* , but this results in an algorithm still targeting the exact posterior. An unbiased estimate of the posterior P can be obtained via importance sampling if it is possible to simulate directly from the likelihood. Again, this is not the case in our model, neither is use of exact simulation possible for our model. Therefore our conclusion is to resort to the pseudo-likelihood based Importance Sampling (IS) approach (Section 2.2.2),

and to the asymptotic results (Mukherjee 2016). Both methods work very well.

2.2.1 Exact formula for footrule and Spearman distances

We here follow the reasoning in Irurozki et al. (2016a), by noting that the partition function of equation (1.11), $Z_n(\alpha) = \sum_{\mathbf{r} \in \mathcal{P}_n} e^{-\frac{\alpha}{n}d(\mathbf{r}, \mathbf{1}_n)}$, can be written in a more convenient way. We notice that $d(\mathbf{r}, \mathbf{1}_n)$ takes only the finite number of discrete values $\mathcal{D} = \{d_1, \dots, d_a\}$, where a depends on n and on the chosen distance $d(\cdot, \cdot)$. We define $L_i = \{\mathbf{r} \in \mathcal{P}_n : d(\mathbf{r}, \mathbf{1}_n) = d_i\} \subset \mathcal{P}_n$, $i = 1, \dots, a$, to be the set of permutations at the same given distance from $\mathbf{1}_n$, and denote by $|L_i|$ its cardinality. Then

$$Z_n(\alpha) = \sum_{d_i \in \mathcal{D}} |L_i| e^{-\frac{\alpha}{n}d_i}. \quad (2.7)$$

In order to compute $Z_n(\alpha)$ one thus needs $|L_i|$, for all values $d_i \in \mathcal{D}$. In the case of the footrule distance, the set \mathcal{D} is made of all even numbers, from 0 to $\lfloor n^2/2 \rfloor$, and $|L_i|$, corresponds to the sequence A062869 available for $n \leq 50$ on the On-Line Encyclopedia of Integer Sequences (OEIS) (Sloane 2017). In the case of Spearman's distance, the set \mathcal{D} is made of all even numbers, from 0 to $2\binom{n+1}{3}$, and $|L_i|$ corresponds to the sequence A175929 available only until $n \leq 14$ in the OEIS. Having discovered these tabulated sequences, we can exploit them to compute $Z_n(\alpha)$ exactly for many different values of n . This will be crucial, both for speeding-up the algorithm, and for evaluating the performance of the IS approximation (discussed in Section 2.2.2), when compared to alternative methods.

2.2.2 Off-line importance sampling, IS, for $Z_n(\alpha)$

For K rank vectors $\mathbf{R}^1, \dots, \mathbf{R}^K$ sampled from an IS auxiliary distribution $q(\mathbf{R})$, the unbiased IS estimate of $Z_n(\alpha)$ is given by

$$\hat{Z}_n(\alpha) = K^{-1} \sum_{k=1}^K \exp[-(\alpha/n)d(\mathbf{R}^k, \mathbf{1}_n)] q(\mathbf{R}^k)^{-1}. \quad (2.8)$$

The more $q(\mathbf{R})$ resembles the Mallows likelihood (2.1), the smaller is the variance of $\hat{Z}_n(\alpha)$. On the other hand, it must be computationally feasible to sample from $q(\mathbf{R})$. We use the following pseudo-likelihood approximation of the target (2.1). Let $\{i_1, \dots, i_n\}$ be

a uniform sample from \mathcal{P}_n , giving the order of the pseudo-likelihood factorization. Then

$$\begin{aligned}
P(R_{i_n} | \mathbf{1}_n) &= \frac{\exp[-(\alpha/n)d(R_{i_n}, i_n)] \cdot \mathbb{1}_{[1, \dots, n]}(R_{i_n})}{\sum_{r_n \in \{1, \dots, n\}} \exp[-(\alpha/n)d(r_n, i_n)]}, \\
P(R_{i_{n-1}} | R_{i_n}, \mathbf{1}_n) &= \frac{\exp[-(\alpha/n)d(R_{i_{n-1}}, i_{n-1})] \cdot \mathbb{1}_{[\{1, \dots, n\} \setminus \{R_{i_n}\}]}(R_{i_{n-1}})}{\sum_{r_{n-1} \in \{1, \dots, n\} \setminus \{R_{i_n}\}} \exp[-(\alpha/n)d(r_{n-1}, i_{n-1})]}, \\
&\vdots \\
P(R_{i_2} | R_{i_3}, \dots, R_{i_n}, \mathbf{1}_n) &= \frac{\exp[-(\alpha/n)d(R_{i_2}, i_2)] \cdot \mathbb{1}_{[\{1, \dots, n\} \setminus \{R_{i_3}, \dots, R_{i_n}\}]}(R_{i_2})}{\sum_{r_2 \in \{1, \dots, n\} \setminus \{R_{i_3}, \dots, R_{i_n}\}} \exp[-(\alpha/n)d(r_2, i_2)]}, \\
P(R_{i_1} | R_{i_2}, \dots, R_{i_n}, \mathbf{1}_n) &= \mathbb{1}_{[\{1, \dots, n\} \setminus \{R_{i_2}, \dots, R_{i_n}\}]}(R_{i_1}).
\end{aligned}$$

Each factor is a simple univariate distribution. We sample R_{i_n} first, and then conditionally on that, $R_{i_{n-1}}$ and so on. The k -th full sample \mathbf{R}^k has probability $q(\mathbf{R}^k) = P(R_{i_n}^k | \mathbf{1}_n) P(R_{i_{n-1}}^k | R_{i_n}^k, \mathbf{1}_n) \cdots P(R_{i_2}^k | R_{i_3}^k, \dots, R_{i_n}^k, \mathbf{1}_n)$.

We observe that this pseudo-likelihood construction is similar to the sequential representation of the Plackett-Luce model with a Mallows parametrization of probabilities.

Note that, in principle, we could sample rankings \mathbf{R}^k from the Mallows model with a different distance than the one of the target model (for example Kendall), or use the pseudo-likelihood approach with a different ‘‘proposal distance’’ other than the target distance. We experimented with these alternatives, but keeping the pseudo-likelihood with the same distance as the one in the target was most accurate and efficient (results not shown). In what follows the distance in (2.8) is the same as the distance in (2.2).

Testing the Importance Sampler

We experimented by increasing the number of importance samples in powers of ten, over a discrete grid of 100 equally spaced α values between 0.01 and 10 (this is the range of α which turned out to be relevant in all our applications when footrule distance is used, typically $\alpha < 5$). We produced a smooth partition function simply using a polynomial of degree 10. The ratio $\hat{Z}_n^K(\alpha)/Z_n(\alpha)$ as a function of α is shown in Figure 2.2 for $n = 10, 20, 50$ and when using different values of K : the ratio quickly approaches 1 when increasing K ; for larger n , a larger K is needed to ensure precision, but $K = 10^6$ seems enough to give very precise estimates.

When n is larger than 50, no exact expression for $Z_n(\alpha)$ is available. Then, we directly compare the estimated $\hat{Z}_n^K(\alpha)$ for increasing K , to check whether the estimates stabilize.

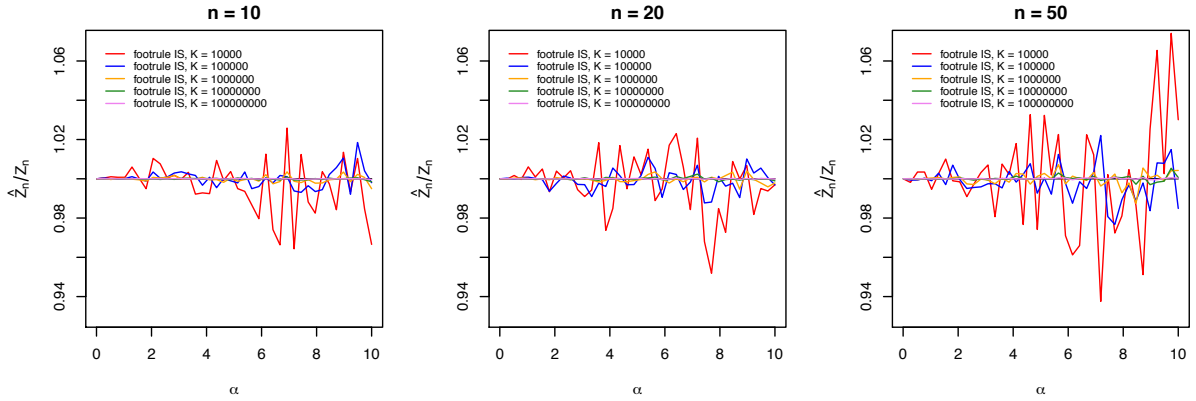


Figure 2.2: Ratio of the approximate partition function computed via IS to the exact, $\hat{Z}_n(\alpha)/Z_n(\alpha)$, as a function of α , when using the footrule distance. From left to right, $n = 10, 20, 50$; different colors refer to different values of K , as stated in the legend.

We thus inspect the maximum relative error

$$\epsilon_K = \max_{\alpha} \left[\frac{\left| \hat{Z}_n^K(\alpha) - \hat{Z}_n^{K/10}(\alpha) \right|}{\left| \hat{Z}_n^{K/10}(\alpha) \right|} \right] \quad (2.9)$$

for $K = 10^2, \dots, 10^8$. Results are shown in Table 2.2 for $n = 75$ and 100 . For both values of n we see that the estimates quickly stabilize, and $K = 10^6$ appears to give good approximations. The computations shown here were performed on a desktop computer, and the off-line computation with $K = 10^6$ samples for $n = 10$ took less than 15 minutes, with no efforts for parallelizing the algorithm, which would be easy and beneficial. $K = 10^6$ samples for $n = 100$ were obtained on a 64-cores computing cluster in 12 minutes.

K	10^2	10^3	10^4	10^5	10^6	10^7	10^8
$n = 75$	152.036	0.921	0.373	0.084	0.056	0.005	0.004
$n = 100$	67.487	1.709	0.355	0.187	0.045	0.018	0.004

Table 2.2: Approximation of the partition function via the importance sampling for the footrule model: maximum relative error ϵ_K , eq. (2.9), between the current and the previous K , for $n = 75$ and 100 .

Effect of $\hat{Z}_n(\alpha)$ on the MCMC

In this section, we report theoretical results regarding the convergence of the MCMC, when using the pseudo-likelihood approximation of the partition function.

Proposition 3. *Algorithm 1 of Appendix 2.A using $\hat{Z}_n(\alpha)$ of eq. (2.8) instead of $Z_n(\alpha)$*

converges to

$$\frac{1}{\hat{C}(\mathbf{R}_{1:N})} \pi(\alpha) \pi(\boldsymbol{\rho}) \hat{Z}_n(\alpha)^{-N} \exp \left[-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right], \quad (2.10)$$

with the normalizing factor

$$\hat{C}(\mathbf{R}_{1:N}) = \int \pi(\alpha) \hat{Z}_n(\alpha)^{-N} \sum_{\boldsymbol{\rho} \in \mathcal{P}_n} \pi(\boldsymbol{\rho}) \exp \left[-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right] d\alpha.$$

Proof. The acceptance probability of the MCMC in Algorithm 1 with the approximate partition function is given by (2.6) using $\hat{Z}_n(\alpha)$ of (2.8) instead of $Z_n(\alpha)$, which is exactly the acceptance probability needed for (2.10). \square

The fact that $\hat{C}(\mathbf{R}_{1:N}) < \infty$ is an obvious consequence of our assumption, in Section 2.1.1, that the prior $\pi(\alpha)$ is defined on the support $[0, \alpha_{\max})$. The approximation $\hat{Z}_n(\alpha)$ converges to $Z_n(\alpha)$ as the number K of IS samples converges to infinity. We here change the notation, in order to explicitly show this dependence, and write $\hat{Z}_n^K(\alpha)$. Clearly, the approximate posterior (2.10) converges to the correct posterior (2.2) if K increases with N , $K = K(N)$, and

$$\lim_{N \rightarrow \infty} \left[\frac{\hat{Z}_n^{K(N)}(\alpha)}{Z_n(\alpha)} \right]^N = 1, \quad \text{for all } \alpha. \quad (2.11)$$

Proposition 4. *There exists a factor $c(\alpha, n, d(\cdot, \cdot))$ not depending on N , such that, if $K = K(N)$ tends to infinity as $N \rightarrow \infty$ faster than $c(\alpha, n, d(\cdot, \cdot)) \cdot N^2$, then (2.11) holds.*

Proof. We see that the ratio

$$\left[\frac{\hat{Z}_n^{K(N)}(\alpha)}{Z_n(\alpha)} \right]^N = \exp \left[N \ln \left(1 + \frac{\hat{Z}_n^{K(N)}(\alpha) - Z_n(\alpha)}{Z_n(\alpha)} \right) \right]$$

tends to 1 in probability as $K(N) \rightarrow \infty$ when $N \rightarrow \infty$ if

$$\frac{\hat{Z}_n^{K(N)}(\alpha) - Z_n(\alpha)}{Z_n(\alpha)} \quad (2.12)$$

tends to 0 in probability faster than $1/N$. Since (2.8) is a sum of i.i.d. variables, there exists a constant $c = c(\alpha, n, d(\cdot, \cdot))$ depending on α , n and the distance d chosen (but not

on N) such that as $K(N) \rightarrow \infty$,

$$\sqrt{K(N)}(\hat{Z}_n^{K(N)}(\alpha) - Z_n(\alpha)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, c^2).$$

Therefore, for (2.12) tending to 0 faster than $1/N$, it is sufficient that $K(N)$ grows faster than N^2 . The speed of convergence to 1 of (2.11) depends on c . \square

Testing approximations of the MCMC in inference

We report results from extensive simulation experiments carried out in several different parameter settings, to investigate if our algorithm provides correct posterior inferences. In addition, we study the sensitivity of the posterior distributions to differences in the prior specifications, and demonstrate their increased precision when the sample size N grows. We explore the robustness of inference when using approximations of the partition function $Z_n(\alpha)$, both when obtained by applying our IS approach described in the previous section, and when using, for large n , the asymptotic approximation $Z_{\text{lim}}(\alpha)$ proposed in Mukherjee (2016). In order to implement Mukherjee's strategy for the computation of the limiting partition function, we fixed his parameter k to 10^3 , and used the his Iterative Proportional Fitting Procedure (IPFP) (Mukherjee 2016, Theorem 1.9) with $m = 10^4$ (after verifying in different situations that the IPFP had typically already converged after 10^3 iterations; not shown).

A comparison between the limiting partition function obtained by running the Mukherjee's procedure, and our IS approximation for $n = 50, 75, 100$ is shown in Figure 2.3, where the plot is in log scale because of better visualization.

We see from the left panel, where the exact normalizing constant is also plotted (red points), that our IS approximation overlaps with the exact, while the Mukherjee limit is slightly biased. In particular, for $n = 50$ we obtain the maximum relative errors $\max_{\alpha} \left[\frac{|\log Z_{50}(\alpha) - \log \hat{Z}_{50}^{IS}(\alpha)|}{|\log Z_{50}(\alpha)|} \right] = 6.2 \cdot 10^{-6}$, and $\max_{\alpha} \left[\frac{|\log Z_{50}(\alpha) - \log \hat{Z}_{\text{lim},50}(\alpha)|}{|\log Z_{50}(\alpha)|} \right] = 4.68$. We then check if the asymptotic and the IS approximations move closer as n grows. Therefore we compute the maximum relative error $\epsilon_n = \max_{\alpha} \left[\frac{|\log \hat{Z}_n^{IS}(\alpha) - \log \hat{Z}_{\text{lim},n}(\alpha)|}{|\log \hat{Z}_n^{IS}(\alpha)|} \right]$, for $n = 50, 75, 100$ and obtain the results of Table 2.3.

We notice that the two approximations become closer as n grows, and thus we can reasonably assume that the approximation will be good also for larger n . This result is crucial in applications where n is so large that the importance sampling approximation is

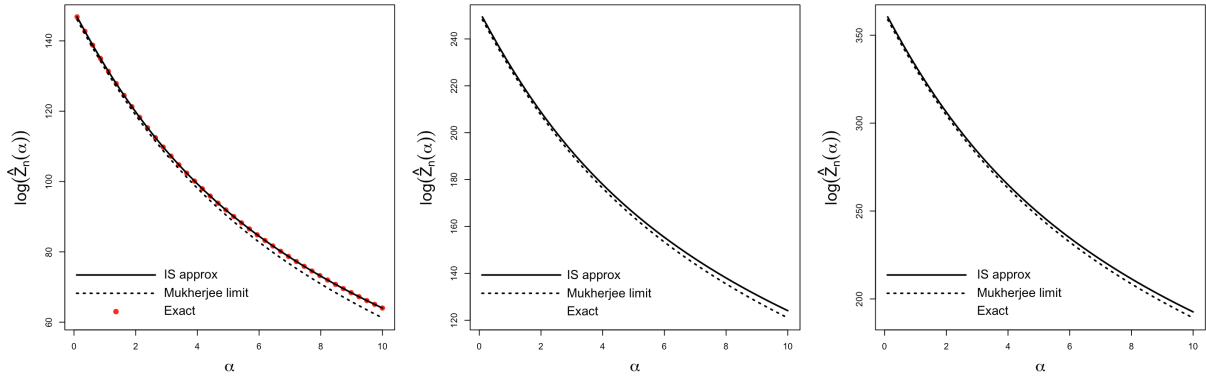


Figure 2.3: A comparison among different approaches to compute the partition function $Z_n(\alpha)$ for the Mallows footrule model. The left panel refers to $n = 50$, the middle panel to $n = 75$, and the right panel to $n = 100$. The solid line refers to the IS approximation, the dashed line to the Mukherjee limit. Note that the exact $Z_n(\alpha)$ is only available for $n = 50$, and depicted with red dots.

	$n = 50$	$n = 75$	$n = 100$
ϵ_n	4.68	4.08	3.79

Table 2.3: Maximum relative error between the IS and the limiting approximations of $Z_n(\alpha)$, for $n = 50, 75$ and 100 .

computationally not feasible, and thus using the limiting partition function $Z_{\text{lim}}(f, \alpha)$ for approximating $Z_n(f, \alpha)$ proves to be an excellent alternative (see Section 2.4.4).

In the remaining of this section we investigate if our algorithm provides correct posterior inferences, by experiments carried out in different parameter settings, while focusing on the footrule distance, since it enables to explore all the different settings, and being the preferred distance in the experiments reported in Section 2.4. Some model parameters are kept fixed in the various cases: $\alpha_{\text{jump}} = 10$, $\sigma_\alpha = 0.15$, and $L = n/5$ (for the tuning of the two latter parameters, see the simulation study in Section 2.1.4). Computing times for the simulations, performed on a laptop computer, varied depending on the values of n and N , from a minimum of 24'' in the smaller case with $n = 20$ and $N = 20$, to a maximum of 3'22'' for $n = 100$ and $N = 1000$.

First, we generated data from a Mallows model with $n = 20$ items, using samples from $N = 20, 50$, and 100 assessors, a setting of moderate complexity. The value of α_{true} was chosen to be either 1 or 3, and ρ_{true} was fixed at $(1, \dots, n)$. To generate the data, we run the MCMC sampler (see Appendix 2.B) for 10^5 burn-in iterations, and collected one sample every 100 iterations after that (these settings were kept in all data generations).

In the analysis, we considered the performance of the method when using the IS

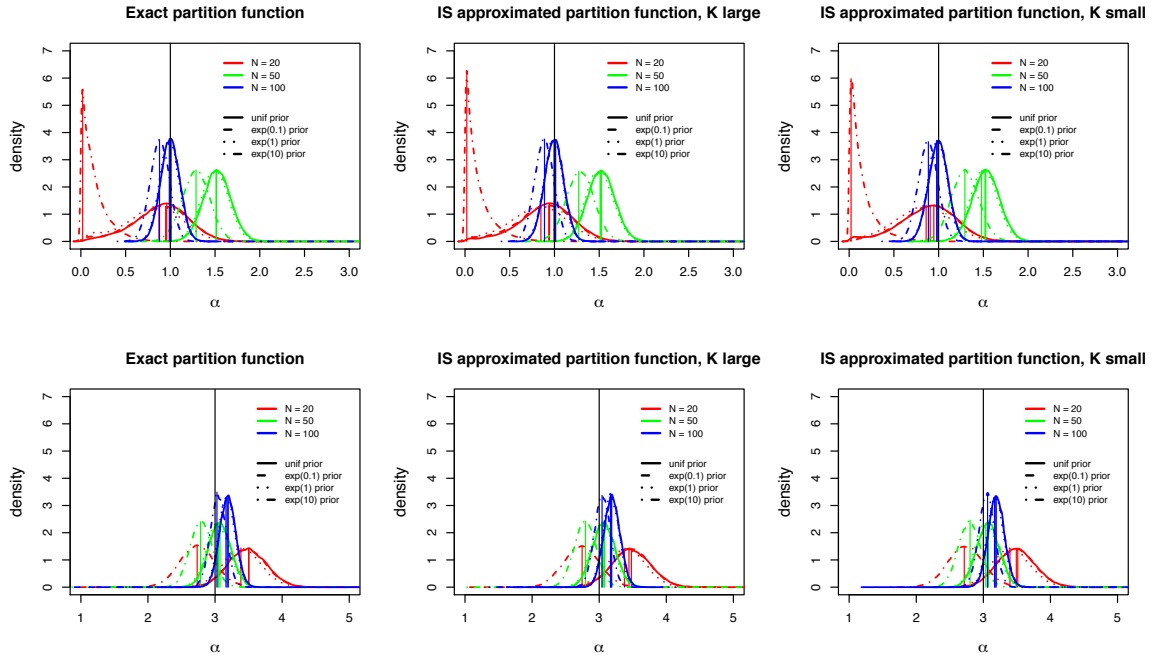


Figure 2.4: Results of the simulations described in Section 2.2.2, when $n = 20$. In each plot is represented the posterior density of α (the black vertical line indicates α_{true}) obtained for various choices of N (different colors), and for different choices of the prior for α (different line types), as stated in the legend. From left to right, MCMC run with the exact $Z_n(\alpha)$, with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^8$, and with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^4$. First row: $\alpha_{\text{true}} = 1$; Second row: $\alpha_{\text{true}} = 3$.

approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^4$ and 10^8 , then comparing the results with those based on the exact $Z_n(\alpha)$. In each case, we run the MCMC for 10^6 iterations, with 10^5 iterations for burn-in, and updated α at every 10th update of $\boldsymbol{\rho}$. Finally, we varied the prior for α to be either the nonintegrable uniform or the exponential using hyperparameter values $\lambda = 0.1, 1$ and 10 . The results are shown in Figure 2.4 for α and Figure 2.5 for $\boldsymbol{\rho}$. As expected, we can see the precision and the accuracy of the marginal posterior distributions increasing, both for α and $\boldsymbol{\rho}$, with N becoming larger. For smaller values of α_{true} , the marginal posterior for α is more dispersed, and $\boldsymbol{\rho}$ is stochastically farther from $\boldsymbol{\rho}_{\text{true}}$. These results are remarkably stable against varying choices of the prior for α , even when the quite strong exponential prior with $\lambda = 10$ was used (with one exception: in the case of $N = 20$ the rather dispersed data generated by $\alpha_{\text{true}} = 1$ were not sufficient to overcome the control of the exponential prior with $\lambda = 10$, which favored even smaller values of α ; see Figure 2.4, top panels). Finally, and most importantly, we see that inference on both α and $\boldsymbol{\rho}$ is completely unaffected by the approximation of $Z_n(\alpha)$ already when $K = 10^4$.

In a second experiment we generated data using $n = 50$ items, $N = 50$ or 500 assessors,

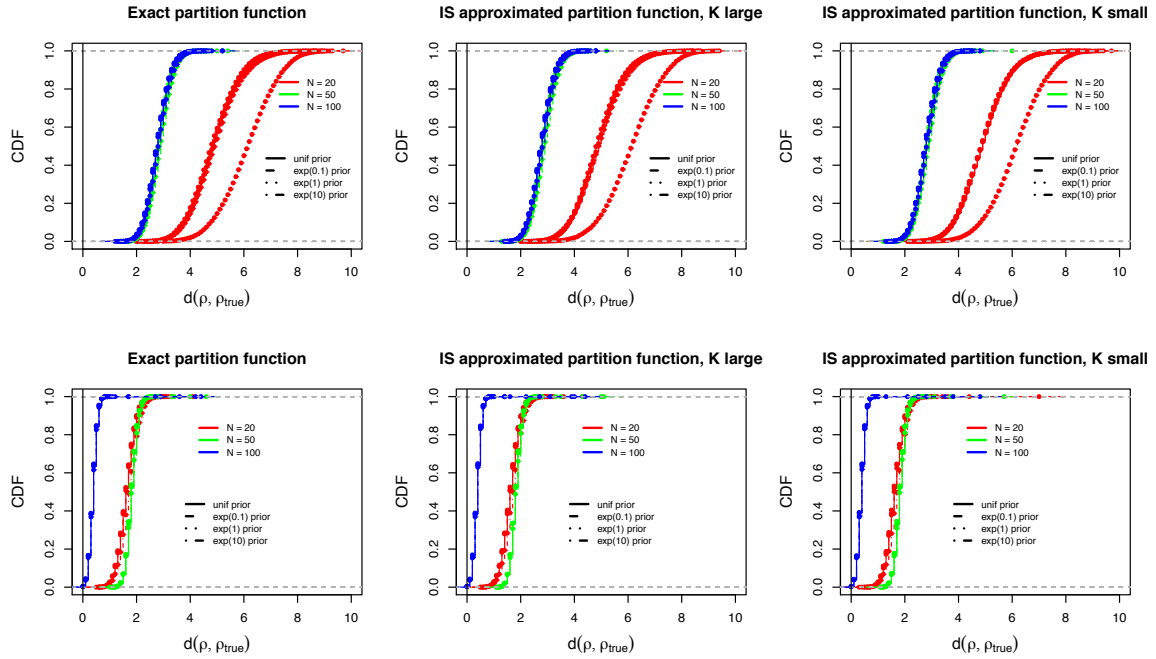


Figure 2.5: Results of the simulations described in Section 2.2.2, when $n = 20$. In each plot is represented the posterior CDF of $d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ obtained for various choices of N (different colors), and for different choices of the prior for α (different line types), as stated in the legend. From left to right, MCMC run with the exact $Z_n(\alpha)$, with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^8$, and with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^4$. First row: $\alpha_{\text{true}} = 1$; Second row: $\alpha_{\text{true}} = 3$.

and scale parameter $\alpha_{\text{true}} = 1$ or 5 . This increase in the value of n gave us some basis for comparing the results obtained by using the IS approximation of $Z_n(\alpha)$ with those from the asymptotic approximation $Z_{\text{lim}}(\alpha)$ of Mukherjee (2016), while still retaining also the possibility of using the exact $Z_n(\alpha)$. For the analysis, all the previous MCMC settings were kept, except for the prior for α : since results from $n = 20$ turned out to be independent of the choice of the prior, here we used the same exponential prior with $\lambda = 0.1$ in all comparisons, as suggested in Section 2.1.1. The results are shown in Figures 2.6 and 2.7. Again, we observe substantially more accurate results for larger values of N and α_{true} . Concerning the impact of approximations to $Z_n(\alpha)$, we notice that, even in this case, the marginal posterior of $\boldsymbol{\rho}$ appears completely unaffected by the partition function not being exact (see Figure 2.6, right panels, and Figure 2.7). In the marginal posterior for α (Figure 2.6, left panels), there are no differences between using the IS approximations and the exact, but there is a difference between Z_{lim} and the other approximations: Z_{lim} appears to be systematically slightly worse.

Finally, we generated data from the Mallows model with $n = 100$ items, $N = 100$

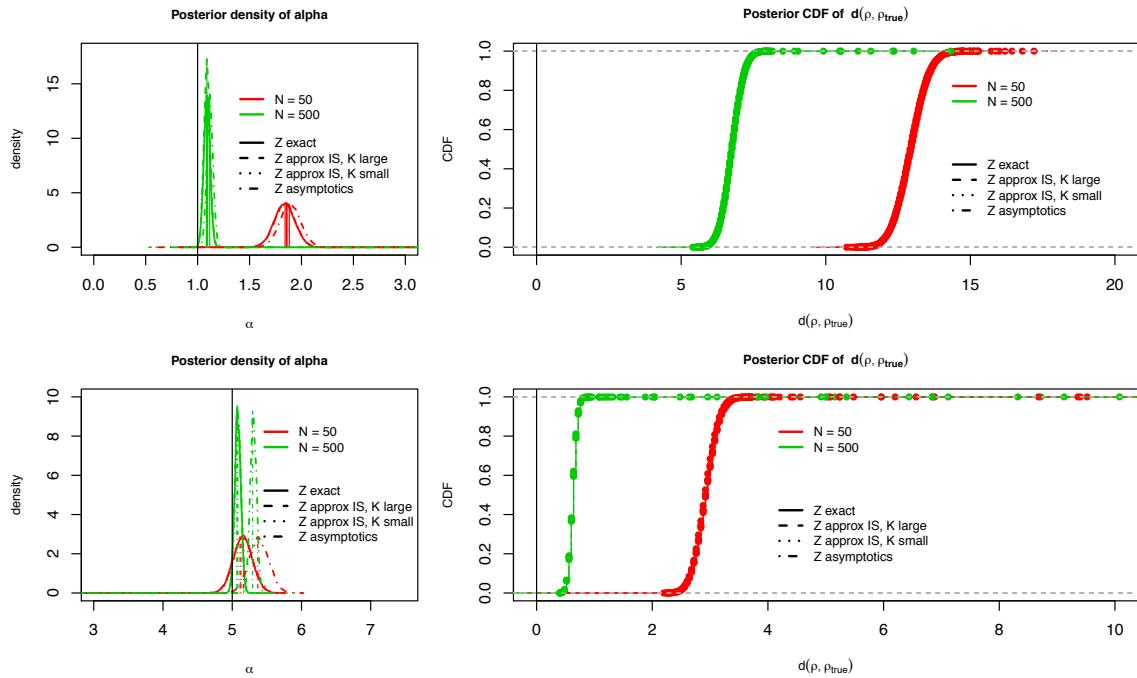


Figure 2.6: Results of the simulations described in Section 2.2.2, when $n = 50$. Left, posterior density of α (the black vertical line indicates α_{true}) obtained for various choices of N (different colors), and when using the exact, or different approximations to the partition function (different line types), as stated in the legend. Right, posterior CDF of $d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ in the same settings. First row: $\alpha_{\text{true}} = 1$; Second row: $\alpha_{\text{true}} = 5$.

or 1000 assessors, and using $\alpha_{\text{true}} = 5$ or 10. Because of this large value of n we were no longer able to compute the exact $Z_n(\alpha)$, hence we only compared results from the different approximations. We kept the same MCMC settings as for $n = 50$, both in data generation and analysis. The results are shown in Figures 2.8 and 2.9. Also in this case, we observe substantially more accurate estimates with larger values of N and α_{true} , establishing an overall stable performance of the method. Here, using the small number $K = 10^4$ of samples in the IS approximation has virtually no effect on accuracy of the marginal posterior for α , while a small effect can be detected from using the asymptotic approximation (Figure 2.8, left panels). However, again, the marginal posterior for $\boldsymbol{\rho}$ appears completely unaffected by the considered approximations in the partition function (Figure 2.8, right panels, and Figure 2.9).

In conclusion, the main results, from the perspective of practical applications, are (1) the relative lack of sensitivity of the posterior inferences to the specification of the prior for the scale parameter α , and (2) the robustness of the marginal posterior inference on $\boldsymbol{\rho}$ on the choice of the approximation of the partition function $Z_n(\alpha)$. Point (1) was

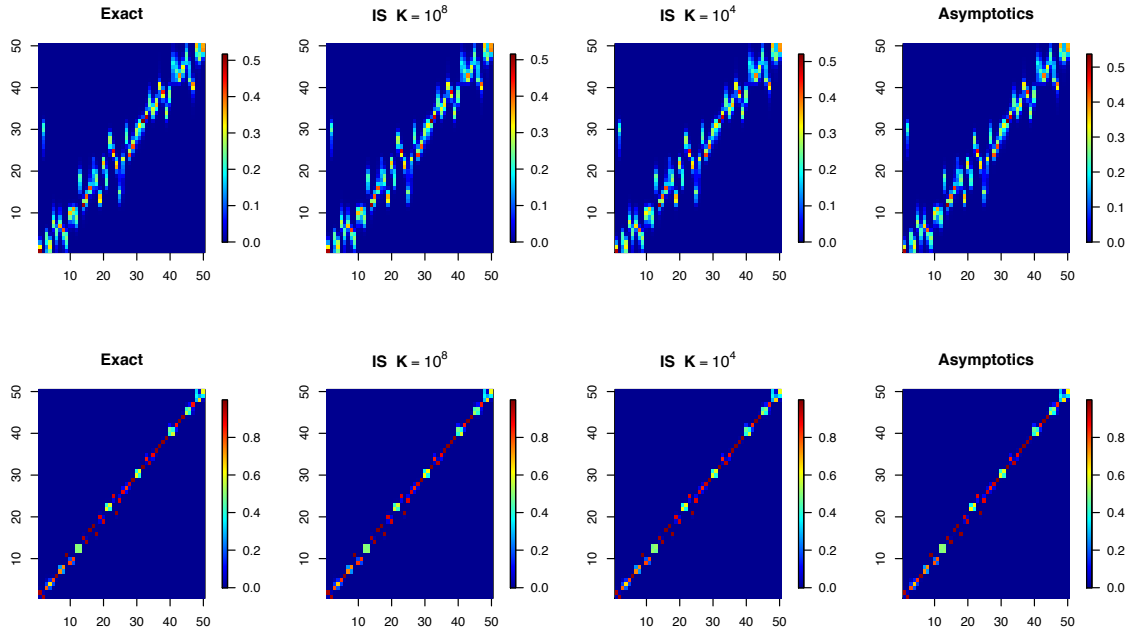


Figure 2.7: Results of the simulations described in Section 2.2.2, when $n = 50$ and $\alpha_{\text{true}} = 5$. In the x-axis items are ordered according to the true consensus $\boldsymbol{\rho}_{\text{true}}$. Each column j represents the posterior marginal density of item j in the consensus $\boldsymbol{\rho}$. Concentration along the diagonal is a sign of success of inference. From left to right, results obtained with the exact $Z_n(\alpha)$, with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^8$, with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^4$, and with $Z_{\text{lim}}(\alpha)$. First row: $N = 50$; Second row: $N = 500$.

not an actual surprise, as it can be understood to be a consequence of the well-known Bernstein-von Mises principle.

Observation (2) deserves a somewhat closer inspection.

The marginal posterior $P(\alpha|\mathbf{R}_{1:N})$, considered in Figures 2.4, 2.6 and 2.8, is obtained from the joint posterior (2.2) by simple summation over $\boldsymbol{\rho}$, then getting the expression

$$P(\alpha|\mathbf{R}_{1:N}) = \frac{\pi(\alpha)}{[Z_n(\alpha)]^N} C(\alpha; \mathbf{R}_{1:N}), \quad (2.13)$$

where $C(\alpha; \mathbf{R}_{1:N}) = \sum_{\boldsymbol{\rho} \in \mathcal{P}_n} \pi(\boldsymbol{\rho}) \exp\left[-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho})\right]$ is the required normalization. For a proper understanding of the structure of the joint posterior and its modification (2.10), it is helpful to first factorize (2.2) into the product

$$P(\alpha, \boldsymbol{\rho}|\mathbf{R}_{1:N}) = P(\alpha|\mathbf{R}_{1:N})P(\boldsymbol{\rho}|\alpha, \mathbf{R}_{1:N}), \quad (2.14)$$

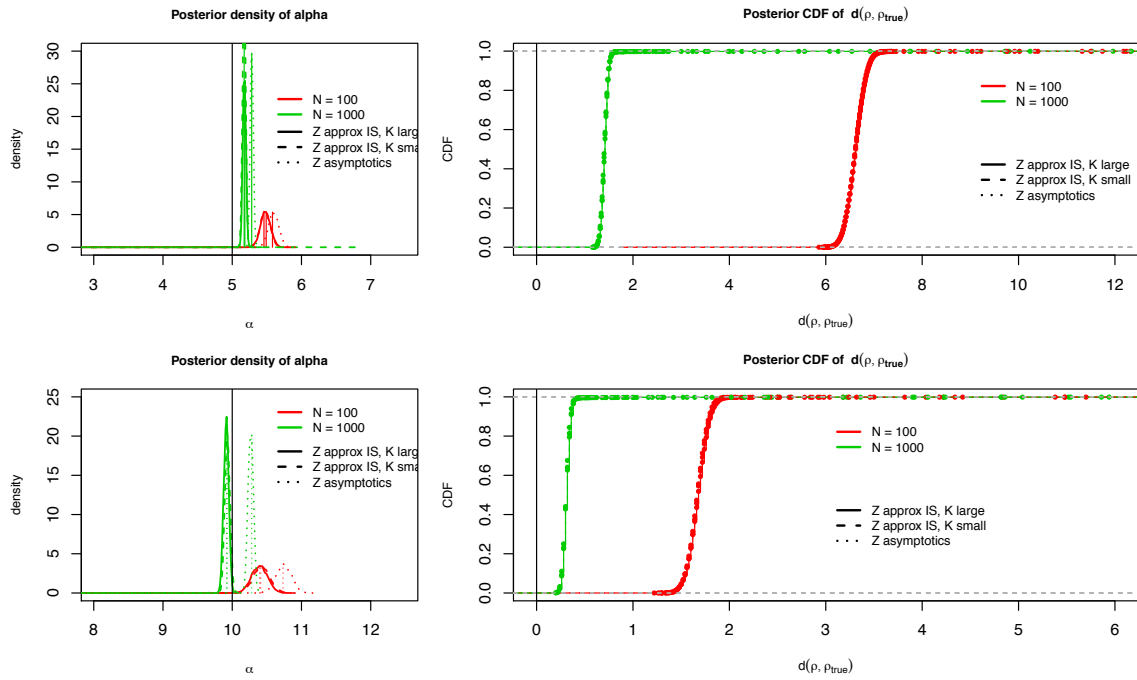


Figure 2.8: Results of the simulations described in Section 2.2.2, when $n = 100$. Left, posterior density of α (the black vertical line indicates α_{true}) obtained for various choices of N (different colors), and when using different approximations to the partition function (different line types), as stated in the legend. Right, posterior CDF of $d(\rho, \rho_{\text{true}})$ in the same settings. First row: $\alpha_{\text{true}} = 5$; Second row: $\alpha_{\text{true}} = 10$.

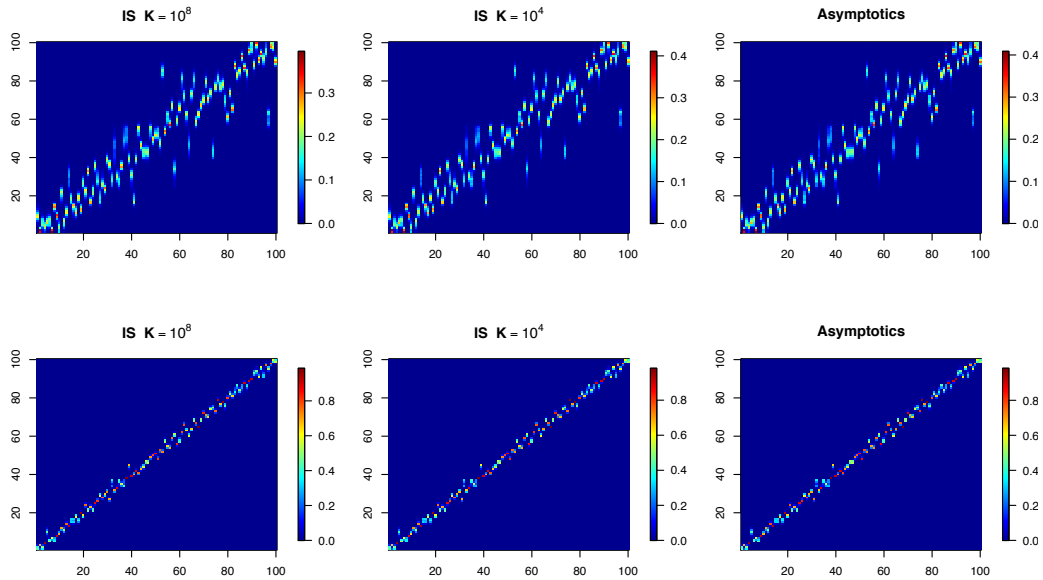


Figure 2.9: Results of the simulations described in Section 2.2.2, when $n = 100$ and $\alpha_{\text{true}} = 5$. In each heatplot, posterior marginal distribution of ρ . From left to right, results obtained with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^8$, with the IS approximation $\hat{Z}_n^K(\alpha)$ with $K = 10^4$, and with $Z_{\text{lim}}(\alpha)$. First row: $N = 100$; Second row: $N = 1000$.

where then

$$P(\boldsymbol{\rho}|\alpha, \mathbf{R}_{1:N}) = [C(\alpha; \mathbf{R}_{1:N})]^{-1} \pi(\boldsymbol{\rho}) \exp \left[-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right]. \quad (2.15)$$

The joint posterior (2.10), which arises from replacing the partition function $Z_n(\alpha)$ by its approximation $\hat{Z}_n(\alpha)$, can be similarly expressed as the product

$$\hat{P}(\alpha, \boldsymbol{\rho}|\mathbf{R}_{1:N}) = \hat{P}(\alpha|\mathbf{R}_{1:N}) P(\boldsymbol{\rho}|\alpha, \mathbf{R}_{1:N}), \quad (2.16)$$

where

$$\hat{P}(\alpha|\mathbf{R}_{1:N}) = [\hat{C}(\mathbf{R}_{1:N})]^{-1} \left[Z_n(\alpha)/\hat{Z}_n(\alpha) \right]^N P(\alpha|\mathbf{R}_{1:N}). \quad (2.17)$$

This requires that the normalizing factor $\hat{C}(\mathbf{R}_{1:N})$ already introduced in (2.10), and here expressed as

$$\hat{C}(\mathbf{R}_{1:N}) \equiv \int \left[Z_n(\alpha)/\hat{Z}_n(\alpha) \right]^N P(\alpha|\mathbf{R}_{1:N}) d\alpha, \quad (2.18)$$

is finite. By comparing (2.14) and (2.16) we see that, under this condition, the posterior $\hat{P}(\alpha, \boldsymbol{\rho}|\mathbf{R}_{1:N})$ arises from $P(\alpha, \boldsymbol{\rho}|\mathbf{R}_{1:N})$ by changing the expression (2.13) of the marginal posterior for α into (2.17), while the conditional posterior $P(\boldsymbol{\rho}|\alpha, \mathbf{R}_{1:N})$ for $\boldsymbol{\rho}$, given α , remains the same in both cases. Thus, the marginal posteriors $P(\boldsymbol{\rho}|\mathbf{R}_{1:N})$ and $\hat{P}(\boldsymbol{\rho}|\mathbf{R}_{1:N})$ for $\boldsymbol{\rho}$ arise as mixtures of the same conditional posterior $P(\boldsymbol{\rho}|\alpha, \mathbf{R}_{1:N})$ with respect to two different mixing distributions, $P(\alpha|\mathbf{R}_{1:N})$ and $\hat{P}(\alpha|\mathbf{R}_{1:N})$.

It is obvious from (2.17) and (2.18) that $\hat{P}(\alpha|\mathbf{R}_{1:N}) = P(\alpha|\mathbf{R}_{1:N})$ would hold if the ratio $Z_n(\alpha)/\hat{Z}_n(\alpha)$ would be constant in α , and this would also entail the exact equality $\hat{P}(\boldsymbol{\rho}|\mathbf{R}_{1:N}) = P(\boldsymbol{\rho}|\mathbf{R}_{1:N})$. It was established in (2.11) that, in the IS scheme, $Z_n(\alpha)/\hat{Z}_n(\alpha) \rightarrow 1$ as $K \rightarrow \infty$. Thus, for large enough K , $\left[Z_n(\alpha)/\hat{Z}_n(\alpha) \right]^N \approx 1$ holds as an approximation (see Proposition 4). Importantly, however, (2.17) shows that the approximation is only required to hold well on the effective support of $P(\alpha|\mathbf{R}_{1:N})$, and this support is narrow when N is large. This is evident from Figures 2.4, 2.6 (left) and 2.8 (left). On this support, because of uniform continuity in α , also the integrand $P(\boldsymbol{\rho}|\alpha, \mathbf{R}_{1:N})$ in (2.15) remains nearly a constant. In fact, experiments (results not shown) performed by varying α over a much wider range of fixed values, while keeping the same $\mathbf{R}_{1:N}$, gave remarkably stable results for the conditional posterior $P(\boldsymbol{\rho}|\alpha, \mathbf{R}_{1:N})$. This contributes to the high degree of robustness in the posterior inference on $\boldsymbol{\rho}$, making requirements of

using large values of K much less stringent.

In Figures 2.5 - 2.9 we consider and compare the marginal posterior CDF's of the distance $d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ under the schemes $P(\cdot|\mathbf{R}_{1:N})$ and $\hat{P}(\cdot|\mathbf{R}_{1:N})$.

Using the shorthand $d^* = d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$, let

$$\begin{aligned} F_{d^*}(x|\alpha, \mathbf{R}_{1:N}) &\equiv P(d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}}) \leq x|\alpha, \mathbf{R}_{1:N}) = \sum_{\{\boldsymbol{\rho}: d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}}) \leq x\}} P(\boldsymbol{\rho}|\alpha, \mathbf{R}_{1:N}), \\ F_{d^*}(x|\mathbf{R}_{1:N}) &\equiv \sum_{\{\boldsymbol{\rho}: d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}}) \leq x\}} P(\boldsymbol{\rho}|\mathbf{R}_{1:N}) = \int F_{d^*}(x|\alpha, \mathbf{R}_{1:N}) P(\alpha|\mathbf{R}_{1:N}) d\alpha, \\ \hat{F}_{d^*}(x|\mathbf{R}_{1:N}) &\equiv \sum_{\{\boldsymbol{\rho}: d(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}}) \leq x\}} \hat{P}(\boldsymbol{\rho}|\mathbf{R}_{1:N}) = \int F_{d^*}(x|\alpha, \mathbf{R}_{1:N}) \hat{P}(\alpha|\mathbf{R}_{1:N}) d\alpha. \end{aligned} \quad (2.19)$$

For example, in Figure 2.5 we display, for different priors, the CDF's $F_{d^*}(x|\mathbf{R}_{1:N})$ on the left, and $\hat{F}_{d^*}(x|\mathbf{R}_{1:N})$ in the middle and on the right, corresponding to two different IS approximations of the partition function. Like the marginal posteriors $P(\boldsymbol{\rho}|\mathbf{R}_{1:N})$ and $\hat{P}(\boldsymbol{\rho}|\mathbf{R}_{1:N})$ above, $F_{d^*}(x|\mathbf{R}_{1:N})$ and $\hat{F}_{d^*}(x|\mathbf{R}_{1:N})$ can be thought of as mixtures of the same function, here $F_{d^*}(x|\alpha, \mathbf{R}_{1:N})$, but with respect to two different mixing distributions, $P(\alpha|\mathbf{R}_{1:N})$ and $\hat{P}(\alpha|\mathbf{R}_{1:N})$. The same arguments, which were used above in support of the robustness of the posterior inference on $\boldsymbol{\rho}$, apply here as well. Extensive empirical evidence for their justification is provided in Figures 2.5-2.9.

Finally note that these arguments also strengthen considerably our earlier conclusion of the lack of sensitivity of the posterior inference on $\boldsymbol{\rho}$ to the specification of the prior for α . For this, we only need to consider alternative priors, say, $\pi(\alpha)$ and $\hat{\pi}(\alpha)$, *mutatis mutandis* in place of the mixing distributions $P(\alpha|\mathbf{R}_{1:N})$ and $\hat{P}(\alpha|\mathbf{R}_{1:N})$.

2.2.3 Comparisons with other methods

The procedure we propose is Bayesian, and one of its strengths is its ability to quantify the uncertainty related the parameter estimates and predictions. In order to compare our results with the ones obtained by other methods which provide only point estimates, we need to summarize the posterior density of the model parameters into a single point estimate, for example MAP, mode, mean, cumulative probability consensus. The cumulative probability (CP) consensus ranking is the ranking arising from the following sequential scheme: first select the item which has the maximum a posteriori marginal probability of being ranked 1st; then the item which has the maximum a posteriori marginal posterior

probability of being ranked 1st or 2nd among the remaining ones, etc. The CP consensus can then be seen as a sequential MAP. We generated the data from the Mallows model (for details refer to Appendix 2.B) with Kendall distance, since this is the unique distance handled by all the considered competitors based on the Mallows model. We compare our procedure (here denoted by `BayesMallows`) with the following methods:

- `PerMallows` (Irurozki et al. 2016a): MLE of the Mallows and the Generalized Mallows models, with some right-invariant distance functions, but not footrule nor Spearman.
- `rankcluster` (Jacques et al. 2014): Inference for the Insertion Sorting Rank (ISR) model.
- `RankAggreg` (Pihur et al. 2009): Rank aggregation via several different algorithms. Here we use the Cross-Entropy Monte Carlo algorithm.
- Borda count (de Borda 1781): Easy and classic way to aggregate ranks. Basically equivalent to the average rank method, thus not a probabilistic approach.

The results of the comparisons are shown in Table 2.4. The `BayesMallows` estimates are obtained through Algorithm 1 of Appendix 2.A, with the available exact partition function corresponding to Kendall distance, and for 10^5 iterations (after a burn-in of 10^4 iterations). All quantities shown are averages over 50 independent repetitions of the whole simulation experiment. $\hat{\alpha}$ is the posterior mean (for `BayesMallows`) or the MLE (for `PerMallows`), while $\hat{\pi}$ is the MLE estimate of the dispersion parameter of ISR (for `rankcluster`). $\hat{\rho}$ is the consensus ranking estimated by the different procedures. For `BayesMallows` (CP) it is given by the CP consensus, while for `BayesMallows` (MAP) it is given by the MAP. We compare the goodness of fit of the methods by evaluating two quantities: first, the normalized Kendall distance between the estimated consensus ranking and the true one, used to generate the data, $d_K^n(\hat{\rho}, \rho_{\text{true}})$. Second, the average Kendall distance between the data points and the estimated consensus ranking, $T(\hat{\rho}, \mathbf{R}_{1:N}) = \frac{1}{N} \sum_{j=1}^N d_K(\hat{\rho}, \mathbf{R}_j)$. This quantity makes sense here, being independent on the likelihood assumed by the different models.

The first remark about the results in Table 2.4 is the clear improvement of the performance in terms of $d_K^n(\hat{\rho}, \rho_{\text{true}})$, of all the methods, for increasing α . This obvious result is a consequence of the easier task of rank aggregation when the assessors are more concentrated around the consensus. Because the data were generated with the same model which

α_{true}	method	$\hat{\alpha}$ or $\hat{\pi}$	$d_K^n(\hat{\boldsymbol{\rho}}, \boldsymbol{\rho}_{\text{true}})$	$T(\hat{\boldsymbol{\rho}}, \mathbf{R}_{1:N})$
1	BayesMallows (CP)	1.01 (0.22)	0.53 (0.26)	19.07 (0.54)
	BayesMallows (MAP)		0.57 (0.31)	19.07 (0.56)
	PerMallows	1.10 (0.19)	0.54 (0.26)	19.12 (0.56)
	rankcluster	0.60 (0.02)	0.86 (0.34)	19.4 (0.58)
	RankAggreg	n.a.	0.66 (0.27)	19.25 (0.58)
	Borda	n.a.	0.54 (0.27)	19.12 (0.56)
2	BayesMallows (CP)	2.05 (0.18)	0.17 (0.12)	16.29 (0.47)
	BayesMallows (MAP)		0.18 (0.13)	16.28 (0.47)
	PerMallows	2.07 (0.17)	0.23 (0.13)	16.33 (0.46)
	rankcluster	0.66 (0.02)	0.37 (0.22)	16.52 (0.54)
	RankAggreg	n.a.	0.29 (0.14)	16.41 (0.49)
	Borda	n.a.	0.23 (0.14)	16.33 (0.46)
3	BayesMallows (CP)	3.02 (0.07)	0.06 (0.08)	13.88 (0.5)
	BayesMallows (MAP)		0.07 (0.09)	13.87 (0.5)
	PerMallows	3.02 (0.21)	0.09 (0.08)	13.9 (0.51)
	rankcluster	0.72 (0.01)	0.15 (0.11)	13.96 (0.49)
	RankAggreg	n.a.	0.14 (0.11)	13.94 (0.52)
	Borda	n.a.	0.09 (0.08)	13.91 (0.51)
4	BayesMallows (CP)	3.96 (0.20)	0.02 (0.05)	11.83 (0.41)
	BayesMallows (MAP)		0.02 (0.04)	11.83 (0.41)
	PerMallows	3.95 (0.20)	0.03 (0.05)	11.85 (0.4)
	rankcluster	0.76 (0.01)	0.08 (0.08)	11.9 (0.44)
	RankAggreg	n.a.	0.06 (0.05)	11.87 (0.42)
	Borda	n.a.	0.03 (0.05)	11.85 (0.4)

Table 2.4: Results of the simulations. $\hat{\alpha}$, refers to the posterior mean (row: BayesMallows) or to MLE (row: PerMallows). $\hat{\pi}$ is the dispersion parameter of ISR. $\hat{\boldsymbol{\rho}}$ is the consensus ranking estimated by the different procedures: MAP (row: BayesMallows (MAP)), CP (row: BayesMallows (CP)), MLE (row: PerMallows and rankcluster), point estimate (row: RankAggreg and Borda). In parenthesis is reported the standard deviation. Parameters setting: $N = 100$, $n = 10$.

BayesMallows and PerMallows used for inference, we expected that the Mallows-based methods would perform better than the rank aggregation methods we considered. The results of Table 2.4 confirm this claim: BayesMallows and PerMallows outperform the other rank aggregation methods, with the exception of Borda count, which gives the same results as PerMallows. This is not surprising, since the PerMallows MLE of the consensus is approximated through the Borda algorithm. Moreover, when the summary of the Bayesian posterior is the CP consensus, the performance of BayesMallows, both in terms of $d_K^n(\hat{\boldsymbol{\rho}}, \boldsymbol{\rho}_{\text{true}})$ and $T(\hat{\boldsymbol{\rho}}, \mathbf{R}_{1:N})$, was better than the others. This is another advantage of our approach on the competitors: being the output a full posterior distribution of the consensus, we can select any strategy to summarize it, possibly driven by the application

at hand. To conclude, our approach gives slightly better results than the other existing methods, and in the worst cases the performance is still equivalent. In Section 2.4 we will compare inferential results on real data, not necessarily generated from the Mallows model.

2.3 Extensions to partial rankings and heterogeneous assessors

We now relax two assumptions of the previous section, namely that each assessor ranks all n items and that the assessors are homogeneous, all sharing a common consensus ranking. This allows us to treat the important situation of pairwise comparisons, and of multiple classes of assessors, as incomplete data cases, within the same Bayesian Mallows framework.

2.3.1 Ranking of the top ranked items

Often only a subset of the items is ranked: ranks can be missing at random, the assessors may only have ranked the, in-their-opinion, top- k items, or can be presented with a subset of items that they have to rank. These situations can be handled conveniently in our Bayesian framework, by applying data augmentation techniques (Tanner and Wong 1987). In this section we discuss the top- k ranking, but the algorithm can easily be generalized to the other cases mentioned.

Suppose that each assessor j has ranked the subset of items $\mathcal{A}_j \subseteq \{A_1, A_2, \dots, A_n\}$, giving them top ranks from 1 to $n_j = |\mathcal{A}_j|$. Let $R_{ij} = \mathbf{X}_j^{-1}(A_i)$ if $A_i \in \mathcal{A}_j$, while for $A_i \in \mathcal{A}_j^c$, R_{ij} is unknown, except for the constraint $R_{ij} > n_j$, $j = 1, \dots, N$. We define augmented data vectors $\tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N$ by assigning ranks to these non-ranked items randomly, using an MCMC algorithm, and do this in a way which is compatible with the rest of the data. Let $\mathcal{S}_j = \{\tilde{\mathbf{R}}_j \in \mathcal{P}_n : \tilde{R}_{ij} = \mathbf{X}_j^{-1}(A_i) \text{ if } A_i \in \mathcal{A}_j\}$, $j = 1, \dots, N$, be the set of possible augmented random vectors, that is the original partially ranked items together with the allowable “fill-ins” of the missing ranks. Our goal is to sample from the posterior distribution

$$P(\alpha, \boldsymbol{\rho} | \mathbf{R}_1, \dots, \mathbf{R}_N) = \sum_{\tilde{\mathbf{R}}_1 \in \mathcal{S}_1} \cdots \sum_{\tilde{\mathbf{R}}_N \in \mathcal{S}_N} P(\alpha, \boldsymbol{\rho}, \tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N | \mathbf{R}_1, \dots, \mathbf{R}_N).$$

Our MCMC algorithm alternates between sampling the augmented ranks given the current values of α and $\boldsymbol{\rho}$, and sampling α and $\boldsymbol{\rho}$ given the current values of the augmented ranks. For the latter, we sample from the posterior $P(\alpha, \boldsymbol{\rho} | \tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N)$ as in Section 2.1.3. For the former, fixing α and $\boldsymbol{\rho}$ and the observed ranks $\mathbf{R}_1, \dots, \mathbf{R}_N$, we see that $\tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N$ are conditionally independent, and moreover, that each $\tilde{\mathbf{R}}_j$ only depends on the corresponding \mathbf{R}_j . This enables us to consider the sampling of new augmented vectors $\tilde{\mathbf{R}}'_j$ separately for each $j, j = 1, \dots, N$. Specifically, given the current $\tilde{\mathbf{R}}_j$ (which embeds information contained in \mathbf{R}_j) and the current values for α and $\boldsymbol{\rho}$, $\tilde{\mathbf{R}}'_j$ is sampled in \mathcal{S}_j from a uniform proposal distribution, meaning that the highest ranks from 1 to n_j have been reserved for the items in \mathcal{A}_j , while compatible ranks are randomly drawn for items in \mathcal{A}_j^c . The proposed $\tilde{\mathbf{R}}'_j$ is then accepted with probability $\min\{1, \eta_{\mathbf{R}_j}\}$, where

$$\eta_{\mathbf{R}_j} = \exp \left[-\frac{\alpha}{n} \left(d(\tilde{\mathbf{R}}'_j, \boldsymbol{\rho}) - d(\tilde{\mathbf{R}}_j, \boldsymbol{\rho}) \right) \right]. \quad (2.20)$$

The MCMC scheme described above and used in the case of partial rankings is sketched as Algorithm 3 of Appendix 2.A.

Effects of unranked items on the consensus ranking

In applications in which the number of items is large there are often items which none of the assessors included in their top-list. What is the exact role of such “left-over” items in the top- k consensus ranking of all items? Can we ignore such “left-over” items and consider only the items explicitly ranked by at least one assessor? In the following we first show that only items explicitly ranked by the assessors appear in top positions of the consensus ranking. We then show that, when considering the MAP consensus ranking, excluding the left-over items from the ranking procedure already at the start has no effect on how the remaining ones will appear in such consensus ranking.

For a precise statement of these results, we need some new notation. Suppose that assessor j has ranked a subset \mathcal{A}_j of n_j items. Let $\mathcal{A} = \bigcup_{j=1, \dots, N} \mathcal{A}_j$, and denote $n = |\mathcal{A}|$. Let n^* be the total number of items, including left-over items which have not been explicitly ranked by any assessor. Denote by $\mathcal{A}^* = \{A_i; i = 1, \dots, n^*\}$ the collection of all items, and by $\mathcal{A}^c = \mathcal{A}^* \setminus \mathcal{A}$ the left-over items. Each rank vector \mathbf{R}_j for assessor j contains, in some order, the ranks from 1 to n_j given to items in \mathcal{A}_j . In the original data the ranks of all remaining items are left unspecified, apart from the fact that implicitly,

for assessor j , they would have values which are at least as large as $n_j + 1$.

The results below are formulated in terms of the two different modes of analysis, which we need to compare and which correspond to different numbers of items being included. The first alternative is to include in the analysis the complete set \mathcal{A}^* of n^* items, and to complement each data vector \mathbf{R}_j by assigning (originally missing) ranks to all items which are not included in \mathcal{A}_j ; their ranks will then form some permutation of the sequence $(n_j + 1, \dots, n^*)$. We call this mode of analysis *full analysis*, and denote the corresponding probability measure by P_{n^*} . The second alternative is to include in the analysis only the items which have been explicitly ranked by at least one assessor, that is, items belonging to the set \mathcal{A} . We call this second mode *restricted analysis*, and denote the corresponding probability measure by P_n . The probability measure P_n is specified as before, including the uniform prior on the consensus ranking $\boldsymbol{\rho}$ across all $n!$ permutations of $(1, 2, \dots, n)$, and the uniform prior of the unspecified ranks R_{ij} of items $A_i \in \mathcal{A}_j^c$ across the permutations of $(n_j + 1, \dots, n)$. The definition of P_{n^*} is similar, except that then the uniform prior distributions are assumed to hold in the complete set \mathcal{A}^* of items, that is, over permutations of $(1, 2, \dots, n^*)$ and $(n_j + 1, \dots, n^*)$, respectively. In the posterior inference carried out in both modes of analysis, the augmented ranks, which were not recorded in the original data, are treated as random variables, with values being updated as part of the MCMC sampling.

Proposition 5. *Consider two latent consensus rank vectors $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ such that*

- (i) *in the ranking $\boldsymbol{\rho}$ all items in \mathcal{A} have been included among the top- n -ranked, while those in \mathcal{A}^c have been assigned ranks between $n + 1$ and n^* ,*
- (ii) *$\boldsymbol{\rho}'$ is obtained from $\boldsymbol{\rho}$ by a permutation, where the rank in $\boldsymbol{\rho}$ of at least one item belonging to \mathcal{A} has been transposed with the rank of an item in \mathcal{A}^c .*

Then, $P_{n^}(\boldsymbol{\rho}|\text{data}) \geq P_{n^*}(\boldsymbol{\rho}'|\text{data})$, for the footrule, Kendall and Spearman distances in the full analysis mode.*

Proof. Having assumed the uniform prior across all permutations of latent consensus ranks, the desired result will hold if and only if $\sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \leq \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}')$. This is true if $d(\mathbf{R}_j, \boldsymbol{\rho}) \leq d(\mathbf{R}_j, \boldsymbol{\rho}')$ holds separately for each assessor j , for $j = 1, \dots, N$. We consider first the footrule distance d , and then show that the result holds also for the Kendall and Spearman distances. This proof follows Proposition 4 in [Meilă and Bao](#)

(2010).

Suppose first, for simplicity, that all assessors have ranked the same n items, that is, $\mathcal{A}_1 = \mathcal{A}_2 = \dots = \mathcal{A}_N = \mathcal{A}$. Later we allow the sets \mathcal{A}_j of ranked items to be different for different assessors. Thus there are $n^* - n$ items, which nobody ranked in the original data.

We now introduce synthetic rankings for all these items as well, that is, we augment each \mathbf{R}_j as recorded in the data by replacing the missing ranks of the items $A_i \in \mathcal{A}^c$ by some permutation of their possible ranks from $n + 1$ to n^* . We then show that the desired inequality holds regardless of how these ranks $\{R_{ij}, A_i \in \mathcal{A}^c\}$ were assigned. The proof is by induction, and it is carried out in several steps.

For the first step, let $\boldsymbol{\rho}$ be a rank vector where the ranks from 1 to n , in any order, have been assigned to the items in \mathcal{A} , and the ranks R_{ij} between $n + 1$ and n^* are given to items in \mathcal{A}^c . Let $\boldsymbol{\rho}'$ be a rank vector obtained from $\boldsymbol{\rho}$ by a transposition of the ranks of two items, say, of $A_{i_0} \in \mathcal{A}^c$ and $A_{i_1} \in \mathcal{A}$, with $\rho_{i_0} = \rho'_{i_1} \geq n + 1$ and $\rho_{i_1} = \rho'_{i_0} \leq n$. Fixing these two items, we want to show that $d(\mathbf{R}_j, \boldsymbol{\rho}) \leq d(\mathbf{R}_j, \boldsymbol{\rho}')$. For the footrule distance we have to show that $\sum_{i=1}^n |R_{ij} - \rho_i| \leq \sum_{i=1}^n |R_{ij} - \rho'_i|$. Since $\boldsymbol{\rho}$ and $\boldsymbol{\rho}'$ coincide for all their coordinates $i \neq i_0, i_1$, it is enough to compare here the terms $|R_{i_0j} - \rho_{i_0}|$ and $|R_{i_1j} - \rho_{i_1}|$ on the left to the corresponding terms $|R_{i_0j} - \rho'_{i_0}|$ and $|R_{i_1j} - \rho'_{i_1}|$ on the right. We need to distinguish between two situations:

- (i) Suppose $R_{i_1j} \leq \rho_{i_1}$. Then, $\rho'_{i_1} - R_{i_1j} > \rho_{i_1} - R_{i_1j}$. On the other hand, $\rho_{i_0} \geq n + 1$ implies that $A_{i_0} \in \mathcal{A}^c$, and it is therefore ranked by assessor j with $R_{i_0j} \geq n + 1$. Therefore, $|R_{i_0j} - \rho'_{i_0}| \geq |R_{i_0j} - \rho_{i_0}|$. By combining these two results we get that $|R_{i_0j} - \rho_{i_0}| + |R_{i_1j} - \rho_{i_1}| \leq |R_{i_0j} - \rho'_{i_0}| + |R_{i_1j} - \rho'_{i_1}|$.
- (ii) Now, suppose that $R_{i_1j} > \rho_{i_1}$. Then, $R_{i_1j} - \rho_{i_1} \leq n - \rho_{i_1} \leq R_{i_0j} - \rho'_{i_0}$. Moreover, since $|R_{i_0j} - \rho_{i_0}| \leq |R_{i_1j} - \rho_{i_1}| = |R_{i_1j} - \rho'_{i_1}|$, we have that again $|R_{i_0j} - \rho_{i_0}| + |R_{i_1j} - \rho_{i_1}| \leq |R_{i_0j} - \rho'_{i_0}| + |R_{i_1j} - \rho'_{i_1}|$ holds.

The same reasoning holds also for the Kendall distance, since the Kendall distance between the two rank vectors, which are obtained from each other by a transposition of a pair of items, is the same as the footrule distance. For the Spearman distance, we only need to form squares of the distance between pairs of items, and the inequality remains valid.

For the general step of the induction, suppose that $\boldsymbol{\rho}$ has been obtained from its original version with all items in \mathcal{A} ranked to the first n positions, via a sequence of transpositions

between items originally in \mathcal{A} and items originally in \mathcal{A}^c . Let $\boldsymbol{\rho}'$ be a rank vector where one more transposition of this type from $\boldsymbol{\rho}$ to $\boldsymbol{\rho}'$ has been carried out. Then the argument of the proof can still be carried through, and the conclusion $d(\mathbf{R}_j, \boldsymbol{\rho}) \leq d(\mathbf{R}_j, \boldsymbol{\rho}')$ holds. This argument needs to be complemented by considering the uniform random permutations, corresponding to the assumed prior of the ranks originally missing in the data, across their possible values from $n+1$ to n^* . But this is automatic, because the conclusion holds separately for all permutations of such ranks.

Finally, the argument needs to be extended to the situation in which the sets \mathcal{A}_j of ranked items can be different for different assessors. In this case we are led to consider, as a by-product of the data augmentation scheme, a joint distribution of the rank vectors $\{\tilde{\mathbf{R}}_j; j = 1, \dots, N\}$. Here, for each j , the n_j items which were ranked first have been fixed by the data. The remaining $n - n_j$ items are assigned augmented random ranks with values between $n_j + 1$ and n , where the probabilities, corresponding to the model P_{n^*} , are determined by the inference from the assumed Mallows model and the data. The conclusion remains valid regardless of the particular way in which the augmentation was done, and so it holds also when taking an expectation with respect to P_{n^*} . \square

Remark. The above proposition says, in essence, that any consensus lists of top- n ranked items, which contains one or more items with their ranks completely missing in the data (that is, the item was not explicitly ranked by any of the assessors), can be improved *locally*, in the sense of increasing the associated posterior probability with respect to P_{n^*} . This happens by trading such an item in the top- n list against another, which had been ranked but which had not yet been selected to the list. In particular, the MAP estimate(s) for consensus ranking assign n highest ranks to explicitly ranked items in the data (which corresponds to the result in [Meilă and Bao 2010](#), for Kendall distance). The following statement is an immediate implication of Proposition 5, following from a marginalization with respect to P_{n^*} .

Corollary 1. *Consider, for $k \leq n$, collections $\{A_{i_1}, A_{i_2}, \dots, A_{i_k}\}$ of k items and the corresponding ranks $\{\rho_{i_1}, \rho_{i_2}, \dots, \rho_{i_k}\}$. In full analysis mode, the maximal posterior probability $P_{n^*}(\{\rho_{i_1}, \rho_{i_2}, \dots, \rho_{i_k}\} = \{1, 2, \dots, k\} | \text{data})$, is attained when $\{A_{i_1}, A_{i_2}, \dots, A_{i_k}\} \subset \mathcal{A}$.*

Another consequence of Proposition 5 is the coincidence of the MAP estimates under the two probability measures P_n and P_{n^*} . For proving this result a further argument,

which goes beyond Proposition 5, has to be made explicit.

Corollary 2. Denote by $\boldsymbol{\rho}^{MAP*}$ the MAP estimate for the consensus ranking obtained in a full analysis, $\boldsymbol{\rho}^{MAP*} := \operatorname{argmax}_{\boldsymbol{\rho} \in \mathcal{P}_{n^*}} P_{n^*}(\boldsymbol{\rho} | \text{data})$, and by $\boldsymbol{\rho}^{MAP}$ the MAP estimate for the consensus ranking obtained in a restricted analysis, $\boldsymbol{\rho}^{MAP} := \operatorname{argmax}_{\boldsymbol{\rho} \in \mathcal{P}_n} P_n(\boldsymbol{\rho} | \text{data})$. Then, $\boldsymbol{\rho}^{MAP*}|_{i:A_i \in \mathcal{A}} \equiv \boldsymbol{\rho}^{MAP}$.

Proof. It follows from Proposition 5 that the n top ranks in $\boldsymbol{\rho}^{MAP*}$ are all assigned to items $A_i \in \mathcal{A}$. Therefore, using shorthand $\boldsymbol{\rho}_{\mathcal{A}} = (\boldsymbol{\rho}_i; A_i \in \mathcal{A})$ and $\boldsymbol{\rho}_{\mathcal{A}^c} = (\boldsymbol{\rho}_i; A_i \in \mathcal{A}^c)$ we see that $\boldsymbol{\rho}^{MAP*}$ must be of the form $\boldsymbol{\rho}^{MAP*} = (\boldsymbol{\rho}_{\mathcal{A}}^{MAP*}, \boldsymbol{\rho}_{\mathcal{A}^c}^{MAP*}) = (\boldsymbol{\pi}, \boldsymbol{\pi}')$, where $\boldsymbol{\pi}$ is a permutation of the set $(1, 2, \dots, n)$, and similarly $\boldsymbol{\pi}'$ is some permutation of $(n+1, \dots, n^*)$. To prove the statement, we show the following: (i) the posterior probabilities $P_{n^*}(\boldsymbol{\rho}_{\mathcal{A}} = \boldsymbol{\pi}, \boldsymbol{\rho}_{\mathcal{A}^c} = \boldsymbol{\pi}' | \text{data})$ and $P_{n^*}(\boldsymbol{\rho}_{\mathcal{A}} = \boldsymbol{\pi} | \boldsymbol{\rho}_{\mathcal{A}^c} = \boldsymbol{\pi}', \text{data})$ are invariant under permutations of $\boldsymbol{\pi}'$, and (ii) the latter conditional probabilities $P_{n^*}(\boldsymbol{\rho}_{\mathcal{A}} = \boldsymbol{\pi} | \boldsymbol{\rho}_{\mathcal{A}^c} = \boldsymbol{\pi}', \text{data})$ coincide with $P_n(\boldsymbol{\rho}_{\mathcal{A}} = \boldsymbol{\pi} | \text{data})$. As a consequence, a list of top- n items obtained from the *full analysis* estimate $\boldsymbol{\rho}^{MAP*}$ qualifies also as the *restricted analysis* estimate $\boldsymbol{\rho}^{MAP}$, and conversely, $\boldsymbol{\rho}^{MAP}$ can be augmented with any permutation $\boldsymbol{\pi}'$ of $(n+1, \dots, n^*)$ to jointly form $\boldsymbol{\rho}^{MAP*}$.

The first part of (i) follows by noticing that the likelihood in the *full analysis*, when considering consensus rankings of the form $\boldsymbol{\rho} = (\boldsymbol{\rho}_{\mathcal{A}}, \boldsymbol{\rho}_{\mathcal{A}^c}) = (\boldsymbol{\pi}, \boldsymbol{\pi}')$, only depends on the observed data via $\boldsymbol{\pi}$. Since the assessors act independently, each imposing a uniform prior on their unranked items, also the posterior $P_{n^*}(\boldsymbol{\rho}_{\mathcal{A}} = \boldsymbol{\pi}, \boldsymbol{\rho}_{\mathcal{A}^c} = \boldsymbol{\pi}' | \text{data})$ will depend only on $\boldsymbol{\pi}$. The second part follows from the first, either by direct conditioning in the joint distribution, or by first computing the marginal $P_{n^*}(\boldsymbol{\rho}_{\mathcal{A}^c} = \boldsymbol{\pi}' | \text{data})$ by summation, and then dividing. (ii) follows then because, for both posterior probabilities, the sample space, the prior, and the likelihood are the same. \square

Remark. The above result is very useful in the context of applications, since it guarantees that the top- n items in the MAP consensus ranking do not depend on which version of the analysis is performed. Recall that a full analysis cannot always be carried out in practice, due to the fact that left-over items might be unknown, or their number might be too large for any realistic computation.

2.3.2 Pairwise comparisons

In many situations, assessors compare pairs of items rather than ranking all or a subset of items. We extend our Bayesian data augmentation scheme to handle such data. Our approach is an alternative to [Lu and Boutilier \(2014\)](#), who treated preferences by applying their Repeated Insertions Model (RIM). Our approach is simpler, it is fully integrated into our Bayesian inferential framework, and it works with any right-invariant distance.

As an example of paired comparisons, assume assessor j stated the preferences $\mathcal{B}_j = \{A_1 \prec A_2, A_2 \prec A_5, A_4 \prec A_5\}$. Here $A_r \prec A_s$ means that A_r is preferred to A_s , so that A_r has a lower rank than A_s . Let \mathcal{A}_j be the set of items constrained by assessor j , in this case $\mathcal{A}_j = \{A_1, A_2, A_4, A_5\}$. Differently from [Section 2.3.1](#), the items which have been considered by each assessor are now not necessarily fixed to a given rank. Hence, in the MCMC algorithm, we need to propose augmented ranks which obey the partial ordering constraints given by each assessor, with the difficulty that none of the items is now fixed to a given rank. Note that we can also handle the case when assessors give ties as a result of some pairwise comparisons: in such a situation, each pair of items resulting in a tie is randomized to a preference at each data augmentation step inside the MCMC, thus correctly representing the uncertainty of the preference between the two items. None of the experiments included in this chapter involve ties, thus this randomization is not needed.

In this chapter we assume that the pairwise orderings in \mathcal{B}_j are mutually compatible, and define by $\text{tc}(\mathcal{B}_j)$ the transitive closure of \mathcal{B}_j , that is, the smallest set that consistently extends the original preference set. It is defined as the set union of \mathcal{B}_j and all pairwise preferences that are not explicitly given but are induced by \mathcal{B}_j by transitivity. In the example above, $\text{tc}(\mathcal{B}_j) = \mathcal{B}_j \cup \{A_1 \prec A_5\}$. For the case of ordered subsets of items, the transitive closure is simply the single set of pairwise preferences compatible with the ordering, for example $\{A_1 \prec A_2 \prec A_5\}$ yields $\text{tc}(\mathcal{B}_j) = \{A_1 \prec A_2, A_2 \prec A_5, A_1 \prec A_5\}$. The R packages `sets` ([Meyer and Hornik 2009](#)) and `relations` ([Hornik and Meyer 2014](#)) efficiently compute the transitive closure.

The main idea of our method for handling such data remains the same as in [Section 2.3.1](#), and the algorithm is the same as [Algorithm 3](#). However, here a “modified” L-S proposal distribution, rather than a uniform one, is used to sample augmented ranks which are compatible with the partial ordering constraint. Suppose that, from the latest step of the MCMC, we have a full augmented rank vector $\tilde{\mathbf{R}}_j$ for assessor j , which is

compatible with $\text{tc}(\mathcal{B}_j)$. Draw a random number u uniformly from $\{1, \dots, n\}$. If $A_u \in \mathcal{A}_j$, let $l_j = \max\{\tilde{R}_{kj} : A_k \in \mathcal{A}_j, k \neq u, (A_k \succ A_u) \in \text{tc}(\mathcal{B}_j)\}$, with the convention that $l_j = 0$ if the set is empty, and $r_j = \min\{\tilde{R}_{kj} : A_k \in \mathcal{A}_j, k \neq u, (A_k \prec A_u) \in \text{tc}(\mathcal{B}_j)\}$, with the convention that $r_j = n + 1$ if the set is empty. Now complete the leap step by drawing a new proposal \tilde{R}'_{uj} uniformly from the set $\{l_j + 1, \dots, r_j - 1\}$. Otherwise, if $A_u \in \mathcal{A}_j^c$, we complete the leap step by drawing \tilde{R}'_{uj} uniformly from $\{1, \dots, n\}$. The shift step remains unchanged. Note that this modified L-S is symmetric.

In Chapter 3 we will come back to the subject of this section, and provide more mathematical details. We will then move to explaining a strategy to relax the assumption of mutually compatible pair comparisons, that will be the main topic of that chapter.

2.3.3 Clustering assessors giving full rankings

So far we have assumed that there exists a unique consensus ranking shared by all assessors. In many cases the assumption of homogeneity is unrealistic: the possibility of clustering assessors into more homogeneous subsets, each sharing a consensus ranking of the items, brings the model closer to reality. We then introduce a mixture of Mallows models, able to handle heterogeneity. We here assume that the data consist of complete rankings

Let $z_1, \dots, z_N \in \{1, \dots, C\}$ be the class labels indicating how individual users are assigned to one of the C clusters. The assessments within each cluster $c \in \{1, \dots, C\}$ are described by a Mallows model with parameters α_c and $\boldsymbol{\rho}_c$. Assuming conditional independence given the Mallows parameters and the class labels, the augmented data formulation of the likelihood for the observed rankings $\mathbf{R}_{1:N}$ is given by

$$P(\mathbf{R}_{1:N} | \boldsymbol{\rho}_{1:C}, \alpha_{1:C}, z_{1:N}) = \prod_{j=1}^N \frac{1}{Z_n(\alpha_{z_j})} \exp \left[-\frac{\alpha_{z_j}}{n} d(\mathbf{R}_j, \boldsymbol{\rho}_{z_j}) \right].$$

For the scale parameters, following Section 2.1.1, we assume the prior

$$\pi(\alpha_1, \dots, \alpha_C) = \left[\frac{\lambda}{1 - e^{\lambda \alpha_{\max}}} \right]^C e^{-\lambda \sum_{c=1}^C \alpha_c} \prod_{c=1}^C \mathbb{1}_{[0, \alpha_{\max})}(\alpha_c).$$

We further assume that the cluster labels are a priori conditionally independent given the

mixing parameters of the clusters, τ_1, \dots, τ_C , and distributed according to

$$P(z_1, \dots, z_N | \tau_1, \dots, \tau_C) = \prod_{j=1}^N \tau_{z_j},$$

where $\tau_c \geq 0$, $c = 1, \dots, C$ and $\sum_{c=1}^C \tau_c = 1$. Finally τ_1, \dots, τ_C are assigned the standard Dirichlet prior with parameter ψ , $\pi(\tau_1, \dots, \tau_C) = \Gamma(\psi C) \Gamma(\psi)^{-C} \prod_{c=1}^C \tau_c^{\psi-1}$, where $\Gamma(\cdot)$ denotes the gamma function.

The number of clusters C is often not known, and the selection of C can be based on different criteria. Here we inspect the posterior distribution of the within-cluster sum of distances of the observed ranks from the corresponding cluster consensus, $\boldsymbol{\rho}_c$, $T(\boldsymbol{\rho}_{1:C}, \mathbf{R}_{1:N}) = \sum_{c=1}^C \sum_{j:z_j=c}^N d(\mathbf{R}_j, \boldsymbol{\rho}_c)$ (see also Section 2.4.3). This approach is a Bayesian version of the more classical within-cluster sum-of-squares criterion for model selection, and we expect to observe an elbow when plotting $T(\boldsymbol{\rho}_{1:C}, \mathbf{R}_{1:N})$ as a function of C , driving the choice of the number of clusters.

Label switching is not explicitly handled inside our MCMC, to ensure full convergence of the chain (Jasra et al. 2005, Celeux et al. 2000). MCMC iterations are re-ordered after convergence is achieved, as in Papastamoulis (2015).

The MCMC algorithm alternates between sampling $\boldsymbol{\rho}_1, \dots, \boldsymbol{\rho}_C$ and $\alpha_1, \dots, \alpha_C$ in a M-H step, and τ_1, \dots, τ_C and z_1, \dots, z_N in a Gibbs sampler step. The former step is straightforward, since $(\boldsymbol{\rho}_c, \alpha_c)_{c=1, \dots, C}$ are conditionally independent given z_1, \dots, z_N . In the latter, we exploit the fact that the Dirichlet prior for τ_1, \dots, τ_C is conjugate to the multinomial conditional prior for z_1, \dots, z_N given τ_1, \dots, τ_C . Therefore in the Gibbs step for τ_1, \dots, τ_C , we sample from $\mathcal{D}(\psi + n_1, \dots, \psi + n_C)$, where $\mathcal{D}(\cdot)$ denotes the Dirichlet distribution and $n_c = \sum_{j=1}^N \mathbb{1}_c(z_j)$, $c = 1, \dots, C$. Finally, in the Gibbs step for z_j , $j = 1, \dots, N$, we sample from $P(z_j = c | \tau_c, \boldsymbol{\rho}_c, \alpha_c, R_j) \propto \tau_c P(\mathbf{R}_j | \boldsymbol{\rho}_c, \alpha_c) = \tau_c Z_n(\alpha_c)^{-1} \exp[-(\alpha_c/n) d(\mathbf{R}_j, \boldsymbol{\rho}_c)]$. The pseudo-code of the clustering algorithm is sketched in Algorithm 2 of Appendix 2.A.

2.3.4 Clustering assessors giving pairwise comparisons

It happens frequently in applications that the model extensions described in Sections 2.3.1, 2.3.2, and 2.3.3 occur jointly: the assessors only provide partial rankings or pairwise comparisons, and they cannot be assumed to form a sample from a homogeneous population. For example, we can think of situations where internet users provide their

preferences on some selected movies (see the data experiment described in Section 2.4.4).

It is straightforward to treat such complexities in our Bayesian Mallows model: Algorithms 3 and 2 of Appendix 2.A can be simply merged by iterating between augmentation, clustering, and α and ρ updates.

We cluster the data using the same mixture approach described in Section 2.3.3. However, assessors do not provide complete rankings $\mathbf{R}_1, \dots, \mathbf{R}_N$, but they rather give partial rankings described by the sets $\mathcal{S}_1, \dots, \mathcal{S}_N$, or some preferences contained in the sets $\mathcal{B}_1, \dots, \mathcal{B}_N$. Hence, differently from Algorithm 2, we here need to provide these sets as input to the MCMC algorithm, as already sketched in Algorithm 3. Embedded in the new algorithm, after the Gibbs step devoted to the updating of cluster assignments z_1, \dots, z_N , we need to perform the updating of the augmented rankings $\tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N$. For each assessor j , $j = 1, \dots, N$, a new augmented rank vector $\tilde{\mathbf{R}}'_j$ is proposed with the same strategy used in Algorithm 3: we use the L-S distribution centered at $\tilde{\mathbf{R}}_j$, and subject to the constraints given by either \mathcal{S}_j or \mathcal{B}_j .

The MCMC algorithm for clustering based on partial rankings or pairwise preferences is sketched in Algorithm 4 of Appendix 2.A.

2.3.5 Example: preference prediction

Consider a situation in which the assessors have expressed their preferences on a collection of items, by performing only partial rankings. Or, suppose that they have been asked to respond to some queries containing different sets of pairwise comparisons. One may then ask how the assessors would have ranked a subset of the items when such ranking could not be concluded directly from the data they provided. Sometimes the interest is to predict the assessors' top preferences, accounting for the possibility that such top lists could contain items which some assessors had not seen. Problems of this type are commonly referred to as personalized ranking, or preference learning (Fürnkranz and Hüllermeier 2010), being a step towards personalized recommendation (see also Section 1.3). There is a large and rapidly expanding literature describing a diversity of methods in this area.

Our framework, based on the Mallows model, and its estimation algorithms as described in the previous sections, form a principled approach for handling such problems. Assuming a certain degree of similarity in the individual preferences, and with different assessors providing partly complementary information, it is natural to try to borrow strength from such partial preference information from different assessors for forming a

consensus. Expanding the model to include clusters allows handling heterogeneity that may be present in the assessment data. The Bayesian estimation procedure provides then the joint posterior distribution, expressed numerically in terms of the MCMC output consisting of sampled values of all cluster membership indicators, z_j , and of complete individual rankings, $\tilde{\mathbf{R}}_j$. For example, if assessor j did not compare A_1 to A_2 , we might be interested in computing $P(A_1 \prec_j A_2 | \text{data})$, the predictive probability that this assessor would have preferred item A_1 to item A_2 . This probability is then readily obtained from the MCMC output, as a marginal of the posterior $P(\tilde{\mathbf{R}}_j | \text{data})$.

To illustrate how this is possible with our approach, we present a small simulated experiment, where we considered heterogeneous assessors expressing some of their pairwise preferences, and then wanted to predict the full individual latent ranking $\tilde{\mathbf{R}}_j$ of all items, for all j . For this, we generated pairwise preference data from a mixture of Mallows models with footrule distance, using the procedure explained in Appendix 2.B. We generated the data with $N = 200$, $n = 15$, $C = 3$, $\alpha_1, \dots, \alpha_C = 4$, $\psi_1, \dots, \psi_C = 50$, obtaining the true $\tilde{\mathbf{R}}_{j,\text{true}}$ for every assessor. Then, we assigned to each assessor j a different number, $T_j \sim \text{TruncPoiss}(\lambda_T, T_{\max})$, of pair comparisons, sampled from a truncated Poisson distribution with $\lambda_T = 20$, denoting by $T_{\max} = n(n-1)/2$ the total number of possible pairs from n items. Each pair comparison was then ordered according to the true $\tilde{\mathbf{R}}_{j,\text{true}}$. The average number of pairs per assessor was around 20, less than 20% of T_{\max} .

In the analysis, we ran Algorithm 4 of Appendix 2.A on these data, using the exact partition function, for 10^5 iterations (of which 10^4 were for burn-in). Separate analyses were performed for $C \in \{1, \dots, 6\}$. Then, in order to inspect if our method correctly identified the true number of clusters, we computed two quantities: a version of the within-cluster sum of footrule distances introduced earlier, $T(\boldsymbol{\rho}_{1:C}, \tilde{\mathbf{R}}_{1:N}) = \sum_{c=1}^C \sum_{j:z_j=c} d(\tilde{\mathbf{R}}_j, \boldsymbol{\rho}_c)$, here computed with respect to the the estimated full rankings $\tilde{\mathbf{R}}_{1:N}$, and a within-cluster indicator of mis-fit to the data, $\sum_{c=1}^C \sum_{j:z_j=c} |\{B \in \text{tc}(\mathcal{B}_j) : B \text{ not consistent with } \boldsymbol{\rho}_c\}|$, where a pair comparison $B \in \text{tc}(\mathcal{B}_j)$, $B = (A_r \prec A_s)$ is not consistent with $\boldsymbol{\rho}_c$ if $\rho_{c,r} > \rho_{c,s}$. The number of such non-consistent pairs in \mathcal{B}_j gives an indication of the mis-fit of the j -th assessor to its cluster. Notice that, while the latter measure takes into account the data directly, the former is based on the augmented rankings $\tilde{\mathbf{R}}_j$ only. Hence, the within-cluster sum of footrule distances could be more sensitive to possible mis-specifications in $\tilde{\mathbf{R}}_j$ when the data are very sparse. Notice also that the second measure is a ‘modified’ version of the Kendall distance between the data and the cluster centers. The boxplots of

the posterior distributions of these two quantities are shown in Figure 2.10: the two measures are very consistent in indicating a clear elbow at $C = 3$, thus correctly identifying the value we used to generate the data.

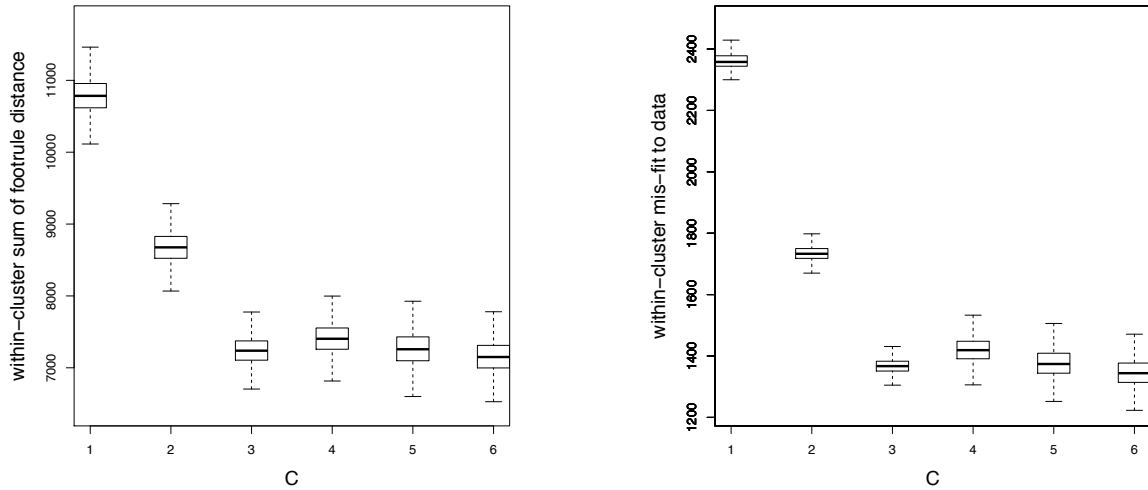


Figure 2.10: Results of the simulation in Section 2.3.5. Boxplots of the posterior distribution of the within-cluster sum of footrule distances (left), and of the within-cluster indicator of mis-fit to the data (right), for different choices of C .

We then studied the success rates of correctly predicting missing individual pairwise preferences. A pairwise preference between items A_{i_1} and A_{i_2} was considered missing for assessor j if it was not among the sampled pairwise comparisons included in the data as either $A_{i_1} \prec_{j,\text{true}} A_{i_2}$ or $A_{i_2} \prec_{j,\text{true}} A_{i_1}$, nor could such ordering be concluded from the data indirectly by transitivity. Thus we computed, for all assessors j , the predictive probabilities $P(A_{i_1} \prec_j A_{i_2} | \text{data})$ for all pairs of items $\{A_{i_1}, A_{i_2}\}$ not ordered in $\text{tc}(\mathcal{B}_j)$. The rule for practical prediction was to always bet on the ordering with the larger predictive probability of these two probabilities, then at least 0.5. Each resulting predictive probability is a direct quantification of the uncertainty in making the bet: a value close to 0.5 expresses a high degree of uncertainty, while a value close to 1 would signal greater confidence in that the bet would turn out right. In the experiment, these bets were finally compared to the orderings of the same pairs in the simulated true rankings $\tilde{\mathbf{R}}_{j,\text{true}}$. If they matched, this was registered as a success, and if not, as a failure. In Figure 2.11 are shown the barplots of the results from this experiment, expressed in terms of the frequency of successes (red columns) and failures (blue columns), obtained by combining the outcomes from all individual assessors. For this presentation, the predictive probabilities used for betting were grouped into the respective intervals $[0.50, 0.55]$, $(0.55, 0.60]$, \dots , $(0.95, 1.00]$

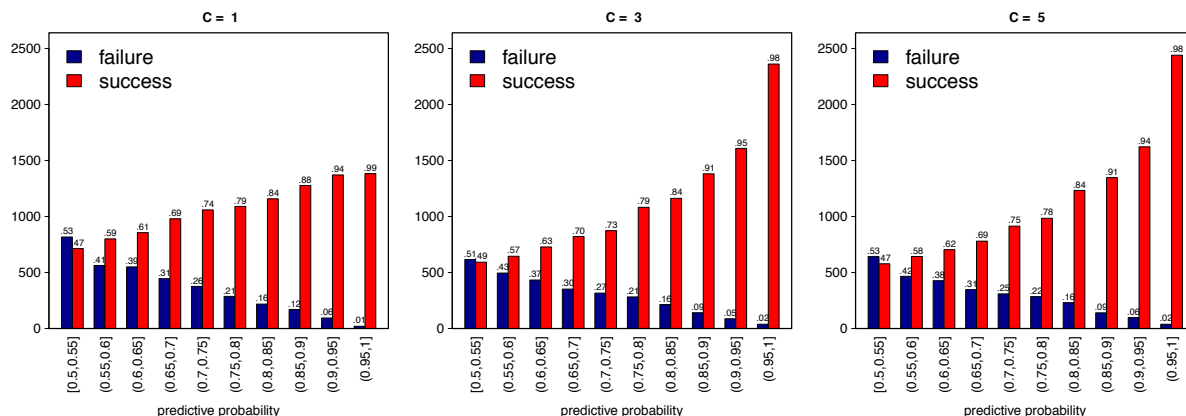


Figure 2.11: Results of the simulation in Section 2.3.5. Barplots of the frequency of successes (red columns) and failures (blue columns) obtained fixing $C = 1$ (left), 3 (middle), and 5 (right), for the data generated with $\lambda_T = 20$. For $C = 1$, 75% of all predictions were correct, for $C = 3$, 79.1%, and for $C = 5$, 79%.

on the horizontal axis, so that pair preferences become more difficult to predict the more one moves to the left, along the x-axis. On top of each column the percentage of successes, or failures, of the corresponding bets is shown. For the results considered on the left, the predictions were made without assuming a cluster structure ($C = 1$) in the analysis, in the middle graph the same number ($C = 3$) of clusters was assumed in the analysis as in the data generation, and on the right, we wanted to study whether assuming an even larger number ($C = 5$) of clusters in the analysis might influence the performance of our method for predicting missing preferences.

Two important conclusions can be made from the results of this experiment. First, from comparing the three graphs, we can see that not assuming a cluster structure ($C = 1$) in the data analysis led to an overall increased proportion of uncertain bets, in the sense of being based on predictive probabilities closer to the 0.5 end of the horizontal axis, than if either $C = 3$ or $C = 5$ was assumed. On the other hand, there is almost no difference between the graphs corresponding to $C = 3$ and $C = 5$. Thus, moderate overfitting of clusters neither improved nor deteriorated the quality of the predictions (this seems consistent with the very similar within-cluster distances in these two cases, shown in Figure 2.10). A second, and more interesting, observation is that, in all three cases considered, the predictive probabilities used for betting turned out to be empirically very well calibrated (see, for example, Dawid 1982, Little 2011). For example, of the bets based on predictive probabilities in the interval $(0.70, 0.75]$, 74% were successful for $C = 1$, 73% when $C = 3$, and 75% when $C = 5$. By inspection, such correspondence can be seen

to hold quite well on all intervals in all three graphs. That the same degree of empirical calibration holds also when an incorrect number of clusters was fitted to the data as with the correct one, signals a certain amount of robustness of this aspect towards variations in the modeling.

We repeated the same experiment with $\lambda_T = 10$. This gives an average number of pairs per assessor around 10% of T_{\max} . Results are displayed in Figure 2.12. Predictive probabilities are still very well calibrated, but of course the quality of prediction is worse. Nonetheless, for $C = 3$, 76.8% of all predictions were correct.

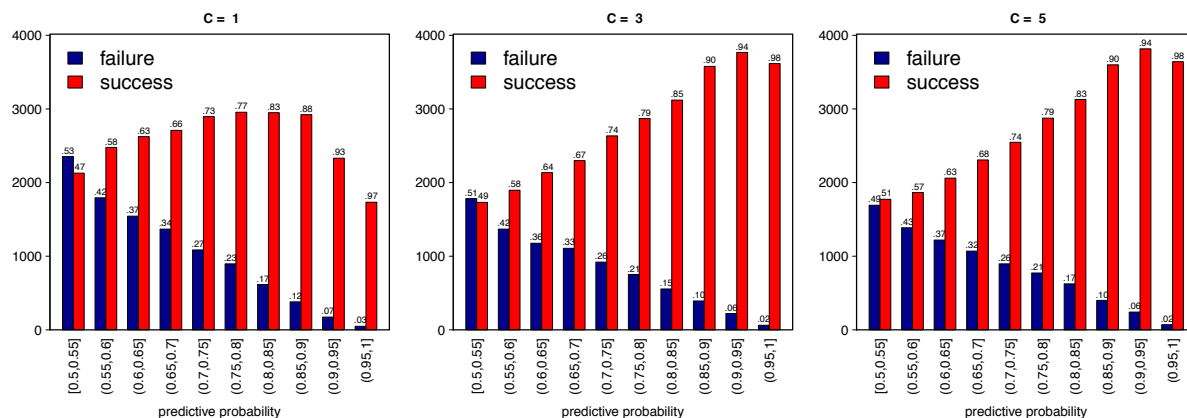


Figure 2.12: Results of the simulation in Section 2.3.5. Barplots of the total numbers of successes (red columns) and failures (blue columns) obtained fixing $C = 1$ (left), 3 (middle), and 5 (right), for the data generated with $\lambda_T = 10$. For $C = 1$, 71% of all predictions was correct, for $C = 3$, 76.8%, and for $C = 5$, 76.7%.

2.4 Experiments

In this section we illustrate the use of our Bayesian Mallows model in various situations corresponding to different data structures. In Section 2.4.1 we consider a very sparse dataset consisting of top- k rankings, and in Section 2.4.2 we illustrate the method on a pairwise comparisons dataset, both without clusters. We then illustrate the mixture model extension, both on complete rankings in Section 2.4.3, and on pairwise comparisons, in Section 2.4.4.

2.4.1 Meta-analysis of differential gene expression

Studies of differential gene expression between two conditions produce lists of genes, ranked according to their level of differential expression as measured by, for example,

p -values. There is often little overlap between gene lists found by independent studies comparing the same condition. This situation raises the question of whether a consensus top list over all available studies can be found.

We handle this situation in our Bayesian Mallows model for top- k rankings (Section 2.3.1). We consider each study $j \in \{1, \dots, N\}$ to be an assessor, providing a top- n_j list of differentially expressed genes, which constitute the ranked items. This problem was studied by DeConde et al. (2006), Deng et al. (2014), and Lin and Ding (2009), who all used the same 5 studies comparing prostate cancer patients with healthy controls (Dhanasekaran et al. 2001, Luo et al. 2001, Singh et al. 2002, True et al. 2006, Welsh et al. 2001). We consider the same 5 studies, and we aim at estimating a consensus with uncertainty. Data consist of the top-25 lists of genes from each study, in total 89 genes. Here we perform a restricted analysis with $n_j = 25$ for all $j = 1, \dots, 5$, and $n = 89$.

ρ	MAP	$P(\rho_i \leq i)$	$P(\rho_i \leq 10)$	$P(\rho_i \leq 25)$
1	HPN	0.58	0.72	0.84
2	AMACR	0.59	0.69	0.8
3	NME2	0.26	0.56	0.64
4	GDF15	0.32	0.67	0.79
5	FASN	0.61	0.65	0.76
6	SLC25A6	0.19	0.63	0.71
7	OACT2	0.61	0.63	0.71
8	UAP1	0.62	0.64	0.74
9	KRT18	0.6	0.61	0.72
10	EEF2	0.64	0.64	0.75
11	GRP58	0.13	0.07	0.61
12	NME1	0.68	0.15	0.79
13	STRA13	0.49	0.06	0.56
14	ALCAM	0.33	0.05	0.65
15	SND1	0.51	0.07	0.71
16	CANX	0.59	0.07	0.64
17	TMEM4	0.34	0.05	0.58
18	DAPK1	0.15	0.04	0.21
19	CCT2	0.59	0.05	0.62
20	MRPL3	0.36	0.06	0.6
21	MTHFD2	0.43	0.06	0.58
22	PPIB	0.51	0.06	0.57
23	SLC19A1	0.42	0.06	0.53
24	FMO5	0.58	0.05	0.59
25	TRAM1	0.14	0.04	0.14

Table 2.5: Top-25 genes in the MAP consensus ranking from a total of 89 genes. The cumulative probability of each gene in the top-25 positions in the MAP of being in that position, or higher, is shown in the third column of the table, $P(\rho_i \leq i)$. The probabilities of being among the top-10 and top-25 are also shown for each gene.

We analyze the five gene lists with the Bayesian Mallows model for partial data (Section 2.3.1), with footrule distance. We run 20 different chains, for a total of 10^7 iterations (computing time was 16'4''), and discarded the first $5 \cdot 10^4$ iterations of each as burn-in. For the partition function, we used the IS approximation $Z_n^K(\alpha)$ with $K = 10^7$, computed off-line on a grid of α 's in $(0, 40]$. After some tuning, we set $L = 40$, $\sigma_\alpha = 0.95$, $\lambda = 0.05$

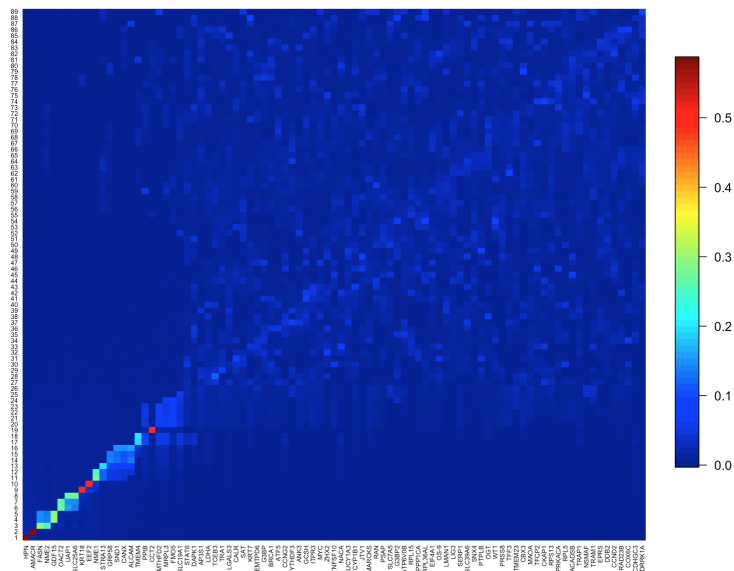


Figure 2.13: Heatplot of the posterior probabilities, for 89 genes, for being ranked as the k -th most preferred, for $k = 1, \dots, 89$. On the x-axis the genes are ordered according to the estimated CP consensus.

and $\alpha_{\text{jump}} = 1$.

In Figure 2.13 we report the heatmap of the marginal posterior probabilities, for the 89 genes (on the x-axis), for being ranked as the k -th most preferred, for $k = 1, \dots, 89$ (on the y-axis). The genes are ordered according to the CP consensus.

Like DeConde et al. (2006), Deng et al. (2014), and Lin and Ding (2009), our method ranked the genes HPN and AMACR first and second in the MAP consensus ranking. The low value of the posterior mean of α , being 0.56 (mode 0.43, high posterior density interval HPDI=(0.04, 1.29)), is an indicator of a generally low level of agreement between the studies. In addition, the fact that $n > N$, and having partial data, both contribute to keeping α small. However, the posterior probability for each gene to be among the top-10 or top-25 is not so low (see columns 4 and 5 of Table 2.5), thus demonstrating that our approach can provide a valid criterion for consensus. In the hypothetical situation in which we had included in our analysis all n^* genes following a *full analysis* mode, with n^* being at least 7567, the largest number of genes included in in any of the five original studies (DeConde et al. 2006), this would have had the effect of making the posterior probabilities in Table 2.5 smaller. On the other hand, because of Corollary 2 of Section 2.3.1, the ranking order obtained from such a hypothetical analysis based on all n^* genes would remain the same as in Table 2.5.

ρ	CE algorithm	GA algorithm	ρ	mean	median	geo.mean	l2norm
1	HPN	HPN	1	HPN	HPN	HPN	HPN
2	AMACR	AMACR	2	AMACR	AMACR	AMACR	AMACR
3	FASN	NME2	3	GDF15	FASN	FASN	GDF15
4	GDF15	0ACT2	4	FASN	KRT18	GDF15	NME1
5	NME2	GDF15	5	NME1	GDF15	NME2	FASN
6	0ACT2	FASN	6	KRT18	NME1	SLC25A6	KRT18
7	KRT18	KRT18	7	EEF2	EEF2	EEF2	EEF2
8	UAP1	SLC25A6	8	NME2	UAP1	0ACT2	NME2
9	NME1	UAP1	9	0ACT2	CYP1B1	OGT	UAP1
10	EEF2	SND1	10	SLC25A6	ATF5	KRT18	0ACT2
11	STRA13	EEF2	11	UAP1	BRCA1	NME1	SLC25A6
12	ALCAM	NME1	12	CANX	LGALS3	UAP1	STRA13
13	GRP58	STRA13	13	GRP58	MYC	CYP1B1	CANX
14	CANX	ALCAM	14	STRA13	PCDHGC3	ATF5	GRP58
15	SND1	GRP58	15	SND1	WT1	CBX3	SND1
16	SLC25A6	TMEM4	16	OGT	TFF3	SAT	ALCAM
17	TMEM4	CCT2	17	ALCAM	MARCKS	CANX	TMEM4
18	PPIB	FM05	18	CYP1B1	OS-9	BRCA1	MTHFD2
19	CCT2	CANX	19	MTHFD2	CCND2	GRP58	MRPL3
20	MRPL3	DYRK1A	20	ATF5	DYRK1A	MTHFD2	PPIB
21	MTHFD2	MTHFD2	21	CBX3	TRAP1	STRA13	OGT
22	SLC19A1	CALR	22	SAT	FM05	LGALS3	CYP1B1
23	FM05	MRPL3	23	BRCA1	ZHX2	ANK3	SLC19A1
24	PRSS8	TRA1	24	MRPL3	RPL36AL	GUCY1A3	ATF5
25	NACA	NACA	25	LGALS3	ITPR3	LDHA	CBX3

Table 2.6: Results given by the `RankAggreg` R package (left) and by the `TopKLists` R package (right).

Next we compared the result shown in Table 2.5 with other approaches: Table 2.6 (left) reports results obtained with `RankAggreg` (Pihur et al. 2009), which targets meta-analysis problems, while in Table 2.6 (right) different aggregation methods implemented in `TopKLists` (Schimek et al. 2015) are considered. The results obtained via `RankAggreg` turned out unstable, with the final output changing in every run, and the list shown in Table 2.6 differs from that in Pihur et al. (2009). Overall, apart from the genes ranked to the top-2 places, there is still considerable variation in the exact rankings of the genes. Rather than considering such exact rankings, however, it may in practice be of more interest to see to what extent the same genes are shared between different top- k lists. Here the results are more positive. For example, of the 10 genes on top of the MAP consensus list of Table 2.5, always 9 genes turned out to be in common with each of the lists of Table 2.6, with the exception of the median (column 3 of Table 2.6, right), where only 7 genes are shared. Column 4 of Table 2.5 provides additional support to the MAP selection of the top-10: all genes included in that list have posterior probability at least 0.56 for being among the top-10, while for those outside the list it is maximally 0.15.

In order to have a quantification of the quality of the different estimates, we compute the footrule distance for partial data (Critchlow 2012, p. 30) between ρ and \mathbf{R}_j , averaged

over the assessors, defined as follows

$$T_{\text{partial}}(\boldsymbol{\rho}, \mathbf{R}) = \frac{1}{N} \sum_{j=1}^N \sum_{i=1}^n |\nu_{R_{ij}} - \nu_{\rho_i}|,$$

where $\nu_{\boldsymbol{\rho}}, \nu_{\mathbf{R}_j} \in \mathcal{P}_n$ are equal to $\boldsymbol{\rho}$ and \mathbf{R}_j in their top- n_j ranks (top-25 in the case of gene lists), while the rank $\frac{n+n_j+1}{2}$ is assigned to the items whose rank in $\boldsymbol{\rho}$ and \mathbf{R}_j is not in their top- n_j . Note that $\frac{n+n_j+1}{2}$ (equal to 57.5 in this case) is the average of the ranks of the excluded items. Table 2.7 reports the values of T_{partial} for the various methods. We notice that the minimum value is achieved by the Mallows MAP consensus list.

	MAP	CE	GA	mean	median	geo.mean	l2norm
$T_{\text{partial}}(\boldsymbol{\rho}, \mathbf{R})$	12.56	12.67	12.98	13.52	15.26	14.05	13.04

Table 2.7: Values of the average footrule distance for partial data T_{partial} between the partial gene lists and the different estimated consensus rankings.

2.4.2 Beach preference data

Here we consider pair comparison data generated as follows: first we chose $n = 15$ images of tropical beaches, shown in Figure 2.14, such that they differ in terms of presence of building and people. For example, beach B9 depicts a very isolated scenery, while beach B2 presents a large hotel seafront.

The pairwise preference data were collected as follows. Each assessor was shown a sequence of 25 pairs of images, and asked on every pair the question: *Which of the two beaches would you prefer to go to in your next vacation?*. Each assessor was presented with a random set of pairs, arranged in random order. As there are 105 possible pairs, 25 pairs is less than 25% of the total. We collected $N = 60$ answers. Seven assessors did not answer to all questions, but we kept these responses as our method is able to analyze also incomplete data. Nine assessors returned orderings which contained at least one non-transitive pattern of comparisons. In this analysis we dropped the non-transitive patterns from the data. Systematic methods for dealing with non-transitive ranking data will be considered in Chapter 3, and in Section 3.4.1, we will come back to this dataset, considering also the non-transitive patterns here dropped.

We analyze these data with the Mallows model for pairwise comparisons (Section 2.3.2), with footrule distance. Since the number of items is $n = 15$, we here make use the

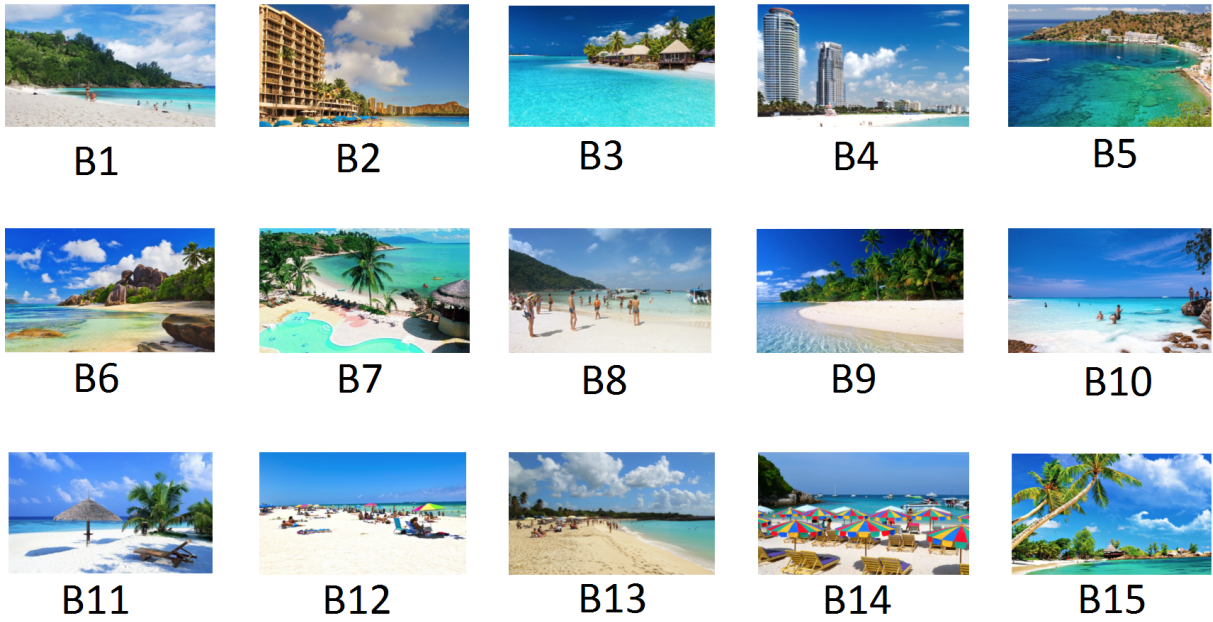


Figure 2.14: The 15 images used for producing the Beach dataset.

exact partition function (see Section 2.2.1). We run the MCMC (Algorithm 3 of Appendix 2.A) for 10^6 iterations, and discarded the first 10^5 iterations as burn-in. We set $L = 2$, $\sigma_\alpha = 0.1$, $\lambda = 0.1$ and $\alpha_{\text{jump}} = 100$. Computing time was less than $2'$.

The posterior mean of α was $\mathbb{E}(\alpha|\text{data}) = 3.38$ (2.94, 3.82). In Table 2.8 we report the consensus ranking of the beaches arranged according to the CP procedure. The 95% HPDI for each item represents the posterior uncertainty. In column 3 is also reported the cumulative probability of each image to be ranked in that position, or higher, $P(\rho_i \leq i)$.

ρ	CP	$P(\rho_i \leq i)$	95% HPDI
1	B9	0.81	(1,2)
2	B6	1	(1,2)
3	B3	0.83	(3,4)
4	B11	0.75	(3,5)
5	B15	0.68	(4,7)
6	B10	0.94	(4,7)
7	B1	1	(6,7)
8	B13	0.69	(8,10)
9	B5	0.55	(8,10)
10	B7	1	(8,10)
11	B8	0.41	(11,14)
12	B4	0.62	(11,14)
13	B14	0.81	(11,14)
14	B12	0.94	(12,15)
15	B2	1	(14,15)

Table 2.8: Results of the pair comparisons. Beaches arranged according to the CP consensus ordering together with the corresponding 95% highest posterior density intervals.

In Table 2.9 we also report the consensus ranking obtained by two other methods, for comparison. **BT** denotes the Bradley Terry ordering, obtained by ordering the score

vectors μ_1, \dots, μ_n (see Section 1.1.2), as returned by the BradleyTerry2 R package (Firth and Turner 2012); **PR** denotes the popular Google PageRank output (Brin and Page 1998) given by the igraph R package (Csardi and Nepusz 2006).

ρ	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
BT	B6	B9	B3	B11	B10	B15	B1	B5	B7	B13	B4	B8	B14	B12	B2
PR	B6	B9	B10	B15	B3	B1	B11	B13	B7	B5	B8	B12	B4	B14	B2

Table 2.9: Consensus ordering given by other methods: **BT** denotes the Bradley Terry ordering as returned by the BradleyTerry2 R package; **PR** denotes the popular Google PageRank output given by the igraph R package.

One advantage of our method is that it is designed to also estimate the latent full rankings of each assessor. Figure 2.15 was obtained as follows: in the separate column on the left, we display the posterior probability $P(\rho_{B_i} \leq 3 | \text{data})$ that a given image B_i , $i = 1, \dots, 15$, was among the top-3 in the consensus ρ . In the other columns we show, for each beach B_i , the individual posterior probabilities $P(\tilde{R}_{j,B_i} \leq 3 | \text{data})$, of being among the top-3 for each assessor j , $j = 1, \dots, 60$. We see for example that beach B5, which was ranked only 9th in the consensus, had, for 4 assessors, posterior probability very close to 1 (as indicated by purple cells) of being included among their top-3 beaches.

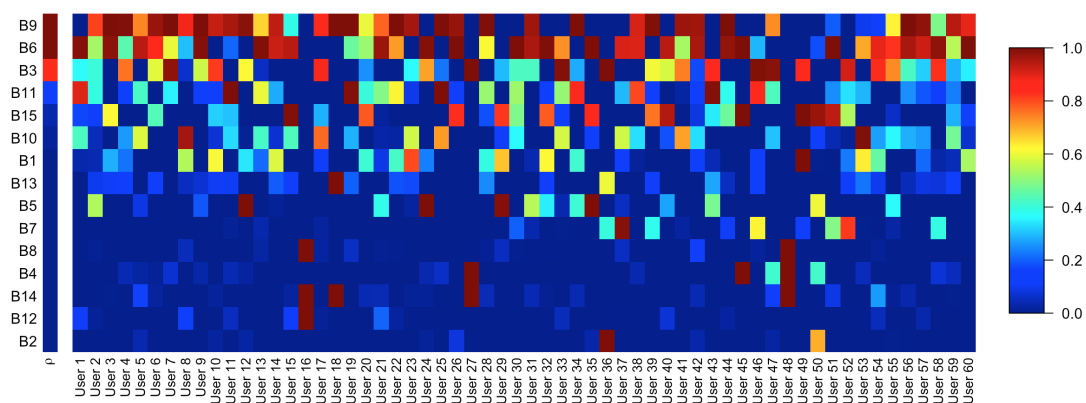


Figure 2.15: Posterior probability, for each beach, of being ranked among the top-3 in ρ (column 1), and in R_j , $j = 1, \dots, 60$ (next columns).

2.4.3 Sushi data

We illustrate clustering based on full rankings using the benchmark dataset of sushi preferences collected across Japan (Kamishima 2003), see also Lu and Boutilier (2014). $N = 5000$ people were interviewed, each giving a complete ranking of $n = 10$ sushi variants. Cultural differences among Japanese regions influence food preferences, so we expect

the assessors to be clustered according to different shared consensus rankings. We analyzed the sushi data using mixtures of Mallows models (Section 2.3.3) with the footrule distance (with the exact partition function, see Section 2.2.1). We run the MCMC (Algorithm 2 of Appendix 2.A) for 10^6 iterations, and discarded the first 10^5 iterations as burn-in. After some tuning, we set L to its minimum value 1, $\sigma_\alpha = 0.1$, $\lambda = 0.1$ and $\alpha_{\text{jump}} = 100$. In the Dirichlet prior for τ , we set the hyper-parameter $\psi = N/C$, thus favoring high-entropy distributions. For each possible number of clusters $C \in \{1, \dots, 10\}$, we used a thinned subset of MCMC samples to compute the posterior footrule distance between ρ_c and the ranking of each assessor assigned to that cluster, $T(\rho_{1:C}, \mathbf{R}_{1:N})$, introduced in Section 2.3.3. The posterior of this quantity, over all assessors and cluster centers, was then used for choosing the appropriate value for C , see Figure 2.16. We found an elbow at $C = 6$, which was then used to further inspect results.

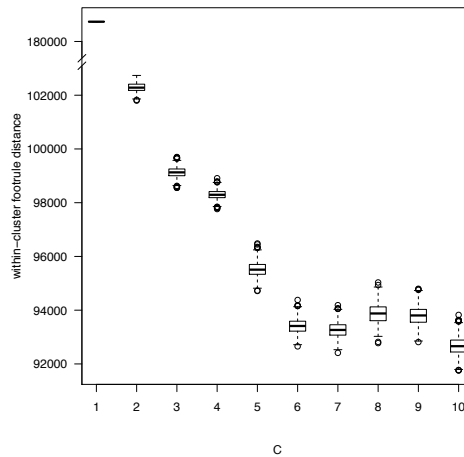


Figure 2.16: Results of the Sushi experiment. Boxplots of the posterior distributions of the within-cluster sum of footrule distances of assessors' ranks from the corresponding cluster consensus for different choices of C (note the y-axis break, for better visualization).

	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
τ_c	0.243 (0.23,0.26)	0.131 (0.12,0.14)	0.107 (0.1,0.11)	0.117 (0.11,0.12)	0.121 (0.11,0.13)	0.278 (0.27,0.29)
α_c	3.62 (3.52,3.75)	2.55 (2.35,2.71)	3.8 (3.42,4.06)	4.02 (3.78,4.26)	4.46 (4.25,4.68)	1.86 (1.77,1.94)
1	fatty tuna	shrimp	sea urchin	fatty tuna	fatty tuna	fatty tuna
2	sea urchin	sea eel	fatty tuna	salmon roe	tuna	tuna
3	salmon roe	egg	shrimp	tuna	tuna roll	sea eel
4	sea eel	squid	tuna	tuna roll	shrimp	shrimp
5	tuna	cucumber roll	squid	shrimp	squid	salmon roe
6	shrimp	tuna	tuna roll	egg	sea eel	tuna roll
7	squid	tuna roll	salmon roe	squid	egg	squid
8	tuna roll	fatty tuna	cucumber roll	cucumber roll	cucumber roll	sea urchin
9	egg	salmon roe	egg	sea eel	salmon roe	egg
10	cucumber roll	sea urchin	sea eel	sea urchin	sea urchin	cucumber roll

Table 2.10: Results of the Sushi experiment when setting $C = 6$. Sushi items arranged according to the MAP consensus ranking found from the posterior distribution of ρ_c , $c = 1, \dots, 6$. At the top of the table, corresponding MAP estimates for τ and α , with 95% HPDIs (in parenthesis). Results are based on 10^6 MCMC iterations.

Table 2.10 shows the results when the number of clusters is set to $C = 6$: for each cluster, the MAP estimates for $\boldsymbol{\tau}$ and $\boldsymbol{\alpha}$, together with their 95% HPDIs, are shown on the top of the Table. Table 2.10 also shows the sushi items, arranged in cluster-specific lists according to the MAP consensus ordering (in this case equal to the CP consensus). Our results can be compared with the ones in Lu and Boutilier (2014) (reported in Table 1 of their section 5.3.2): the correspondence of the clusters could be 1-4, 2-1,3-2,4-5,5-4,6-0. Note that the dispersion parameter α in our Bayesian Mallows model is connected to the dispersion parameter ϕ in Lu and Boutilier (2014) by the link $\alpha = -n \ln(\phi)$. Hence, we can also observe that the cluster-specific α values reported in Table 2.10 are quite comparable to the dispersion parameters of Lu and Boutilier (2014).

We investigate the stability of the clustering in Figure 2.17, which shows the heatmap of the posterior probabilities, for all 5000 assessors (on the x-axis), of being assigned to each of the 6 clusters in Table 2.10 (clusters $c = 1, \dots, 6$ from bottom to top in Figure 2.17): most of these individual probabilities were concentrated on some particular preferred value of c among the six possibilities, indicating a reasonably stable behavior in the cluster assignments.

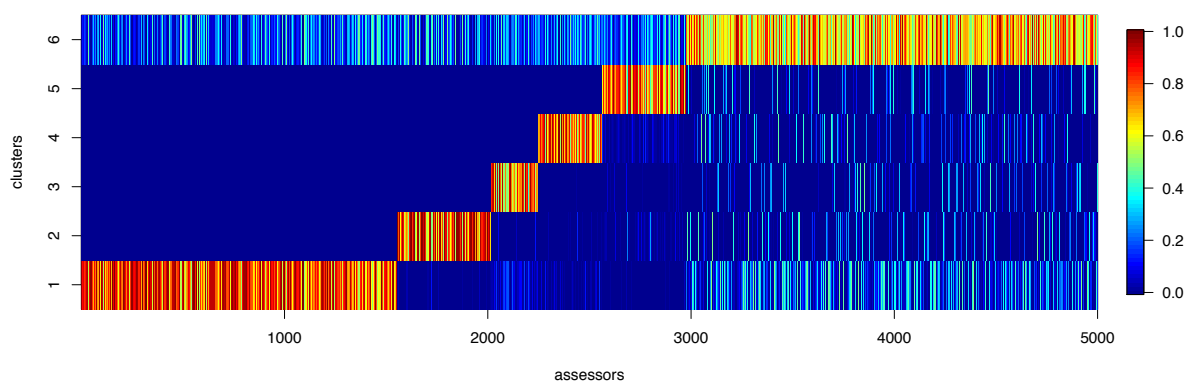


Figure 2.17: Heatplot of posterior probabilities for all 5000 assessors (on the x-axis) of being assigned to each cluster ($c = 1, \dots, 6$ from bottom to top).

In a second moment, we tried to compare the performance of our model with the one given by the `rankcluster` R package, in terms of $T(\boldsymbol{\rho}_{1:C}, \mathbf{R}_{1:N})$. However the method had difficulties in converging, probably because of the high dimensionality of the data. We then compared the results with the `rankdist` R package that implements the Mallows model with Kendall distance. Since the main function does not return the cluster assignments, we could not compute $T(\boldsymbol{\rho}_{1:C}, \mathbf{R}_{1:N})$. We then only compare the results in terms of the

estimated cluster consensus, reported in Table 2.11: note that there is good agreement with our MAP consensus lists in Table 2.10, and the correspondence `BayesMallows-rankdist` could be 1-3, 2-5, 3-6, 4-4, 5-2, 6-1.

	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
1	fatty tuna	fatty tuna	fatty tuna	fatty tuna	shrimp	fatty tuna
2	salmon roe	tuna	sea urchin	tuna	sea eel	sea eel
3	tuna	tuna roll	salmon roe	salmon roe	squid	sea urchin
4	shrimp	shrimp	tuna	shrimp	egg	tuna
5	tuna roll	squid	shrimp	sea urchin	fatty tuna	salmon roe
6	squid	sea eel	sea eel	tuna roll	tuna	shrimp
7	sea eel	egg	squid	squid	tuna roll	tuna roll
8	egg	cucumber roll	tuna roll	sea eel	cucumber roll	squid
9	cucumber roll	salmon roe	egg	egg	salmon roe	egg
10	sea urchin	sea urchin	cucumber roll	cucumber roll	sea urchin	cucumber roll

Table 2.11: Results of the Sushi experiment when setting $C = 6$ and using the `rankdist` package. Sushi items arranged according to the estimated consensus ranking.

2.4.4 Movielens data

The Movielens dataset¹ contains movie ratings from 6040 users. In this example, we focused on the $n = 200$ most rated movies, and on the $N = 6004$ users who rated (not equally) at least 3 movies. Each user had considered only a subset of the n movies (30.2 on average). We converted the ratings given by each user from a 1-5 scale to pairwise preferences as described in Lu and Boutilier (2014): each movie was preferred to all movies which the user had rated strictly lower. We selected users whose rating included at least 3 movies, because two of them were needed to create at least a pairwise comparison, and the third one was needed for prediction, as explained in the following.

Since we expected heterogeneity among users, due for example to age, gender, social factors or education, we applied the clustering scheme for pairwise preferences (sketched in Section 2.3.3), with the footrule distance. Since $n = 200$, we used the asymptotic approximation for $Z_n(\alpha)$ described in Mukherjee (2016) (see Section 2.2.2 for details). We then run the MCMC (Algorithm 4 of Appendix 2.A) for 10^5 iterations, after a burn-in of $5 \cdot 10^4$ iterations. We set: $L = 20$, $\sigma_\alpha = 0.05$, $\alpha_{\text{jump}} = 10$ and $\lambda = 0.1$, after some tuning. Note that the label switching problem only affects inference on cluster-specific parameters, but it does not affect predictive distributions (Celeux et al. 2006). We varied the number C of clusters in the set $\{1, \dots, 15\}$, and inspected the within-cluster indicator of mis-fit to the data introduced in Section 2.3.5, see Figure 2.18: the posterior within-cluster indicator shows two possible elbows: $C = 5$, and $C = 11$. Hence, according to

¹www.grouplens.org/datasets/.

these criteria, both choices seemed initially conceivable. However, it is beyond the scope of this chapter to discuss ways to decide the number of clusters.

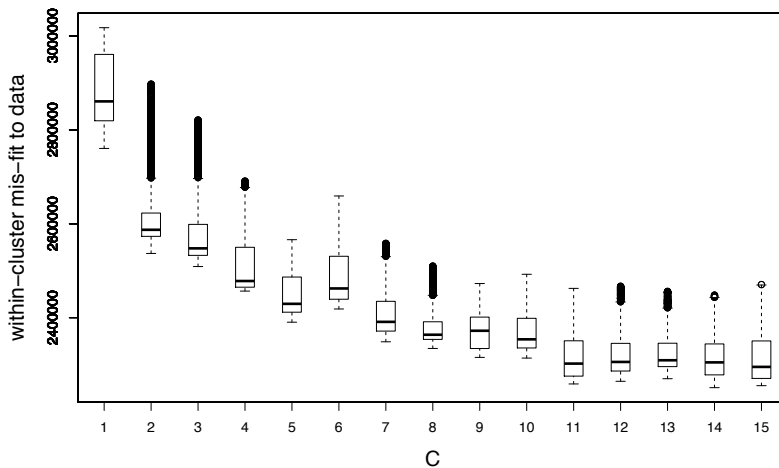


Figure 2.18: Results of the Movielens experiment. Boxplots of the posterior distributions of the within-cluster indicator of mis-fit to the data, as introduced in Section 2.3.5, for different choices of C .

In order to select one of these two models, we examined their predictive performance. Before converting ratings to preferences, we discarded for each user j one of the rated movies at random. Then, we randomly selected one of the other movies rated by the same user, and used it to create a pairwise preference involving the discarded movie. This preference was then not used for inference. After running the Bayesian Mallows model, we computed for each user the predictive probabilities $P(\tilde{\mathbf{R}}_j | \text{all data})$, and thereby the probabilities for correctly predicting the discarded preference. The median, across all users, of these probabilities was 0.8225 for the model with $C = 5$ clusters, and 0.796 for $C = 11$ clusters. Moreover, for $C = 5$, 88 % of these probabilities were higher than 0.5. These are very positive results, and they suggest that the predictive performance of the model with 5 clusters is slightly better than the one with 11 clusters. It appears that the larger number of clusters in the latter model leads to a slight overfitting, and this is likely to be the main cause of the loss in the predictive success. Figure 2.19 shows the boxplots of the posterior distribution of the probability for correct preference prediction of the left-out comparison, stratified with respect to the number of preferences given by each user, for the model with $C = 5$. The histogram on the right shows the same posterior probability for correctly predicting the discarded preference for all users, for the same model, regardless of how many preferences each user had expressed. Interestingly, in this

data, the predictive power is rather stable and high, irrespectively from how many movies the users rated. In other applications, we would expect the predictions to become better the more preferences are expressed by a user. In this case, a figure similar to Figure 2.19 could guide personal recommendation algorithms, which should not rely on estimated point preferences, if these are too uncertain, as happens for users who have given just a few ratings.

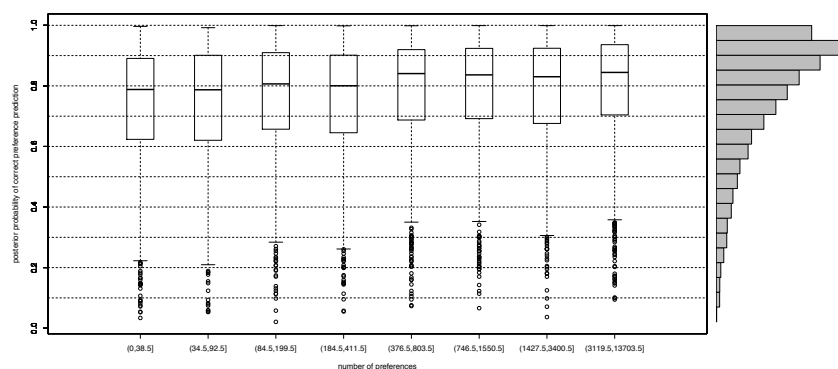


Figure 2.19: Results of the Movielens experiment. Boxplots of the posterior probability for correctly predicting the discarded preference conditionally on the number of preferences stated by the user, for the model with $C = 5$. The histogram on the right shows the marginal posterior probability for correct preference prediction.

	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$
τ_c	0.325 (0.32,0.33)	0.219 (0.21,0.23)	0.156 (0.15,0.17)	0.145 (0.14,0.15)	0.155 (0.15,0.16)
α_c	2.53 (2.36,2.7)	3.33 (3.2,3.48)	2.58 (2.27,2.81)	1.87 (1.67,2.02)	2.68 (2.47,2.89)
1	A Christmas Story	Citizen Kane	The Sting	Indiana Jones (I)	Shawshank Redempt.
2	Schindler's List	The Godfather	Dr. Strangelove	A Christmas Story	Indiana Jones (I)
3	The Godfather	Pulp Fiction	2001: Space Odyssey	Star Wars (IV)	Braveheart
4	Casablanca	Dr. Strangelove	The Maltese Falcon	The Princess Bride	Star Wars (IV)
5	Star Wars (IV)	A Clockwork Orange	Casablanca	Schindler's List	Saving Private Ryan
6	Shawshank Redempt.	Casablanca	Taxi Driver	The Matrix	The Green Mile
7	Saving Private Ryan	The Usual Suspects	Citizen Kane	Shawshank Redempt.	Schindler's List
8	The Sting	2001: Space Odyssey	Schindler's List	Indiana Jones (III)	The Sixth Sense
9	The Sixth Sense	American Beauty	Chinatown	The Sting	The Matrix
10	American Beauty	Star Wars (IV)	The Godfather	The Sixth Sense	Star Wars (V)

Table 2.12: Results of the Movielens experiment. Movies arranged according to the CP consensus ranking, from the posterior distribution of ρ_c , $c = 1, \dots, 5$.

In Table 2.12 the MAP estimates for τ and α , together with their 95% HPDIs, are shown at the top. The table also shows a subset of the movies, arranged in cluster-specific top-10 lists according to the CP consensus ranking, from the posterior distribution of ρ_c , $c = 1, \dots, 5$. We note that all α values correspond to a reasonable within-cluster variability. Moreover, the lists reported in Table 2.12 characterize the users in the same cluster as individuals sharing a reasonably well interpretable preference profile. Since in the Movielens dataset additional information on the users is available, we compared the estimated cluster assignments with the age, gender, and the occupation of the users.

While occupation showed no interesting patterns, the second and fifth clusters had more males than expected, in contrast to the first and fourth clusters which included more females than average, the former above 45 and the latter below 35 of age.

2.5 Discussion

In this chapter, we developed a fully Bayesian hierarchical framework for the analysis of ranking data. An important advantage of the Bayesian approach is that it offers coherently propagated and directly interpretable ways to quantify posterior uncertainties of estimates of any quantity of interest. Earlier Bayesian treatments of the Mallows ranking model are extended in many ways: we develop an importance sampling scheme for $Z_n(\alpha)$ allowing to use other distances than Kendall's, and our MCMC algorithm efficiently samples from the posterior distribution of the consensus ranking and of the latent assessor-specific full rankings. We also develop various extensions of the model, motivated by applications in which data take particular forms.

All methods presented have been implemented in C++, and run efficiently on a desktop computer, with the exception of the Movielens experiment, which needed to be run on a cluster. Obtaining a sufficiently large sample from the posterior distribution takes from a few seconds, for small problems, to several minutes, in the examples involving massive data augmentation. We are currently working on distributed versions of the MCMC.

The proposed models perform very well with a large number of assessors N , but may not be computationally feasible when the number of items is extremely large, for example $n \geq 10^4$, which is not uncommon in certain applications (Volkovs and Zemel 2014). Already in the considered case of $n = 200$, the MCMC converges slowly, a problem also shared by maximum likelihood estimation of ρ (Aledo et al. 2013, Ali and Meilă 2012). For footrule and Spearman distances, there exist an asymptotic approximation for $Z_n(\alpha)$ as $n \rightarrow \infty$ (Mukherjee 2016), which we compared to our IS procedure in Section 2.2.2, and used in a real data application (Section 2.4.4). Many of the extensions we propose for solving specific problems (for example, clustering, preference prediction, pairwise comparisons) are needed jointly in real applications, as we illustrate for example in the Movielens data. Our general framework is flexible enough to handle such extensions.

There are many situations in which rankings vary over time, as in political surveys (Regenwetter et al. 1999) or book bestsellers (Caron and Teh 2012). This case is dealt

with in [Asfaw et al. \(2017\)](#), where the approach described in this chapter is extended to dynamic rankings.

A natural generalization of the Mallows model is to allow for a multivariate dispersion parameter, α . This is known as generalized Mallows's model (GMM), first implemented in [Fligner and Verducci \(1986\)](#), for Kendall and Cayley distances, and further extended in [Meilă and Bao \(2010\)](#), for Kendall distance only, to the Bayesian framework (see also . The generalized Mallows model with footrule and Spearman cannot be enforced, because these two distances do not factorize into $(n - 1)$ terms. An alternative generalization was proposed by [Lee and Yu \(2010\)](#) (further developed in [Lee and Yu 2012](#)), where the authors propose their weighted distance-based ranking models by using weighted distances between rankings. Starting from their model, we believe that, also within our framework, it is feasible to generalize the Bayesian Mallows model with a multidimensional dispersion parameter of length n .

Appendix

2.A Pseudo-codes of the algorithms

We here report the pseudo-codes of the algorithms. Our codes can handle any right invariant distance. The distances currently implemented are Kendall, footrule, Spearman and Cayley. For Kendall and Cayley there is no need to run the IS to approximate $Z_n(\alpha)$, as it is implemented the available closed form of [Fligner and Verducci \(1986\)](#). The same is true for footrule ($n \leq 50$) and Spearman ($n \leq 14$), thanks to the results of Section 2.2.1. For footrule ($n > 50$) and Spearman ($n > 14$) the IS procedure have to be un off-line, before the MCMC procedure.

Algorithm 1: Basic MCMC Algorithm for Complete Rankings

```

input :  $\mathbf{R}_1, \dots, \mathbf{R}_N; \lambda, \sigma_\alpha, \alpha_{\text{jump}}, L, d(\cdot, \cdot), Z_n(\alpha), M.$ 
output: Posterior distributions of  $\rho$  and  $\alpha$ .
Initialization of the MCMC: randomly generate  $\rho_0$  and  $\alpha_0$ .

for  $m \leftarrow 1$  to  $M$  do
  M-H step: update  $\rho$ :
  sample:  $\rho' \sim \text{L-S}(\rho_{m-1}, L)$  and  $u \sim \mathcal{U}(0, 1)$ 
  compute:  $\text{ratio} \leftarrow$  equation (2.4) with  $\rho \leftarrow \rho_{m-1}$  and  $\alpha \leftarrow \alpha_{m-1}$ 
  if  $u < \text{ratio}$  then  $\rho_m \leftarrow \rho'$ 
  else  $\rho_m \leftarrow \rho_{m-1}$ 

  if  $m \bmod \alpha_{\text{jump}} = 0$  then M-H step: update  $\alpha$ :
  sample:  $\alpha' \sim \text{LnN}(\alpha_{m-1}, \sigma_\alpha^2)$  and  $u \sim \mathcal{U}(0, 1)$ 
  compute:  $\text{ratio} \leftarrow$  equation (2.6) with  $\rho \leftarrow \rho_m$  and  $\alpha \leftarrow \alpha_{m-1}$ 
  if  $u < \text{ratio}$  then  $\alpha_m \leftarrow \alpha'$ 
  else  $\alpha_m \leftarrow \alpha_{m-1}$ 
end

```

2.B Sampling from the Mallows model

We here explain the procedure we used to sample data from the Mallows model. To sample N full rankings $\mathbf{R}_1, \dots, \mathbf{R}_N \sim \mathcal{M}(\rho, \alpha)$, we use the following scheme (sketched in Algorithm 5). We run a basic Metropolis-Hastings algorithm with fixed consensus $\rho \in \mathcal{P}_n$, $\alpha > 0$ and with a given distance measure, $d(\cdot, \cdot)$, until convergence. Then we

Algorithm 2: MCMC Algorithm for Clustering Complete Rankings

input : $\mathbf{R}_1, \dots, \mathbf{R}_N; C, \psi, \lambda, \sigma_\alpha, \alpha_{\text{jump}}, L, d(\cdot, \cdot), Z_n(\alpha), M$.
output: Posterior distributions of $\rho_1, \dots, \rho_C, \alpha_1, \dots, \alpha_C, \tau_1, \dots, \tau_C, z_1, \dots, z_N$.
Initialization of the MCMC: randomly generate $\rho_{1,0}, \dots, \rho_{C,0}, \alpha_{1,0}, \dots, \alpha_{C,0}, \tau_{1,0}, \dots, \tau_{C,0}$, and $z_{1,0}, \dots, z_{N,0}$.

```

for  $m \leftarrow 1$  to  $M$  do
  Gibbs step: update  $\tau_1, \dots, \tau_C$ 
  compute:  $n_c = \sum_{j=1}^N \mathbb{1}_c(z_{j,m-1})$ , for  $c = 1, \dots, C$ 
  sample:  $\tau_1, \dots, \tau_C \sim \mathcal{D}(\psi + n_1, \dots, \psi + n_C)$ 

  for  $c \leftarrow 1$  to  $C$  do
    M-H step: update  $\rho_c$ 
    sample:  $\rho'_c \sim \text{L-S}(\rho_{c,m-1}, L)$  and  $u \sim \mathcal{U}(0, 1)$ 
    compute:  $\text{ratio} \leftarrow$  equation (2.4) with  $\rho \leftarrow \rho_{c,m-1}$  and  $\alpha \leftarrow \alpha_{c,m-1}$ , and where the sum is over  $\{j : z_{j,m-1} = c\}$ 
    if  $u < \text{ratio}$  then  $\rho_{c,m} \leftarrow \rho'_c$ 
    else  $\rho_{c,m} \leftarrow \rho_{c,m-1}$ 

    if  $m \bmod \alpha_{\text{jump}} = 0$  then M-H step: update  $\alpha_c$  sample:  $\alpha'_c \sim \mathcal{N}(\alpha_{c,m-1}, \sigma_\alpha^2)$  and  $u \sim \mathcal{U}(0, 1)$ 
    compute:  $\text{ratio} \leftarrow$  equation (2.6) with  $\rho \leftarrow \rho_{c,m}$  and  $\alpha \leftarrow \alpha_{c,m-1}$ , and where the sum is over  $\{j : z_{j,m-1} = c\}$ 
    if  $u < \text{ratio}$  then  $\alpha_{c,m} \leftarrow \alpha'_c$ 
    else  $\alpha_{c,m} \leftarrow \alpha_{c,m-1}$ 

  end

  Gibbs step: update  $z_1, \dots, z_N$ 
  for  $j \leftarrow 1$  to  $N$  do
    foreach  $c \leftarrow 1$  to  $C$  do compute cluster assignment probabilities:  $p_{cj} = \frac{\tau_{c,m}}{Z_n(\alpha_{c,m})} \exp\left[-\frac{\alpha_{c,m}}{n} d(\mathbf{R}_j, \rho_{c,m})\right]$ 
    sample:  $z_{j,m} \sim \mathcal{Mn}(p_{1j}, \dots, p_{Cj})$ 
  end
end

```

Algorithm 3: MCMC Algorithm for Partial Rankings or Pairwise Preferences

input : $\{S_1, \dots, S_N\}$ or $\{\text{tc}(\mathcal{B}_1), \dots, \text{tc}(\mathcal{B}_N)\}; \lambda, \sigma_\alpha, \alpha_{\text{jump}}, L, d(\cdot, \cdot), Z_n(\alpha), M$.
output: Posterior distributions of ρ, α and $\tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N$.
Initialization of the MCMC: randomly generate ρ_0 and α_0 .

```

if  $\{S_1, \dots, S_N\}$  among inputs then
  foreach  $j \leftarrow 1$  to  $N$  do randomly generate  $\tilde{\mathbf{R}}_j^0$  in  $S_j$ 
else
  foreach  $j \leftarrow 1$  to  $N$  do randomly generate  $\tilde{\mathbf{R}}_j^0$  compatible with  $\text{tc}(\mathcal{B}_j)$ 
end

for  $m \leftarrow 1$  to  $M$  do
  M-H step: update  $\rho$ :
  sample:  $\rho' \sim \text{L-S}(\rho_{m-1}, L)$  and  $u \sim \mathcal{U}(0, 1)$ 
  compute:  $\text{ratio} \leftarrow$  equation (2.4) with  $\rho \leftarrow \rho_{m-1}$  and  $\alpha \leftarrow \alpha_{m-1}$ 
  if  $u < \text{ratio}$  then  $\rho_m \leftarrow \rho'$ 
  else  $\rho_m \leftarrow \rho_{m-1}$ 

  if  $m \bmod \alpha_{\text{jump}} = 0$  then M-H step: update  $\alpha$ :
  sample:  $\alpha' \sim \mathcal{N}(\alpha_{m-1}, \sigma_\alpha^2)$  and  $u \sim \mathcal{U}(0, 1)$ 
  compute:  $\text{ratio} \leftarrow$  equation (2.6) with  $\rho \leftarrow \rho_m$  and  $\alpha \leftarrow \alpha_{m-1}$ 
  if  $u < \text{ratio}$  then  $\alpha_m \leftarrow \alpha'$ 
  else  $\alpha_m \leftarrow \alpha_{m-1}$ 

  M-H step: update  $\tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N$ :
  for  $j \leftarrow 1$  to  $N$  do
    if  $\{S_1, \dots, S_N\}$  among inputs then sample:  $\tilde{\mathbf{R}}'_j$  in  $S_j$  from the L-S distribution centered at  $\tilde{\mathbf{R}}_j^{m-1}$ 
    else sample:  $\tilde{\mathbf{R}}'_j$  from the L-S distribution centered at  $\tilde{\mathbf{R}}_j^{m-1}$  and compatible with  $\text{tc}(\mathcal{B}_j)$ 
    compute:  $\text{ratio} \leftarrow$  equation (2.20) with  $\rho \leftarrow \rho_m, \alpha \leftarrow \alpha_m$  and  $\tilde{\mathbf{R}}_j \leftarrow \tilde{\mathbf{R}}_j^{m-1}$ 
    sample:  $u \sim \mathcal{U}(0, 1)$ 
    if  $u < \text{ratio}$  then  $\tilde{\mathbf{R}}_j^m \leftarrow \tilde{\mathbf{R}}'_j$ 
    else  $\tilde{\mathbf{R}}_j^m \leftarrow \tilde{\mathbf{R}}_j^{m-1}$ 
  end
end

```

took N sampled rankings with a large enough interval between each of them to achieve independence.

In case of heterogeneous rankings, we sampled from Algorithm 6. As inputs, we give the number of clusters C , the fixed consensuses ρ_1, \dots, ρ_C , the fixed $\alpha_1, \dots, \alpha_C$, the hyper-

Algorithm 4: MCMC Algorithm for Clustering Partial Rankings or Pairs

input : $\{S_1, \dots, S_N\}$ or $\{tc(\mathcal{B}_1), \dots, tc(\mathcal{B}_N)\}$; $C, \psi, \lambda, \sigma_\alpha, \alpha_{\text{jump}}, L, d(\cdot, \cdot), Z_n(\alpha), M$.
output: Posterior distributions of $\rho_1, \dots, \rho_C, \alpha_1, \dots, \alpha_C, \tau_1, \dots, \tau_C, z_1, \dots, z_N$, and $\tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N$.
Initialization of the MCMC: randomly generate $\rho_{1,0}, \dots, \rho_{C,0}, \alpha_{1,0}, \dots, \alpha_{C,0}, \tau_{1,0}, \dots, \tau_{C,0}$, and $z_{1,0}, \dots, z_{N,0}$.

if $\{S_1, \dots, S_N\}$ among inputs then
 | **foreach** $j \leftarrow 1$ to N **do** randomly generate $\tilde{\mathbf{R}}_j^0$ in S_j
else
 | **foreach** $j \leftarrow 1$ to N **do** randomly generate $\tilde{\mathbf{R}}_j^0$ compatible with $tc(\mathcal{B}_j)$
end

for $m \leftarrow 1$ to M **do**
 Gibbs step: update τ_1, \dots, τ_C
 compute: $n_c = \sum_{j=1}^N \mathbb{1}_c(z_{j,m-1})$, for $c = 1, \dots, C$
 sample: $\tau_1, \dots, \tau_C \sim \mathcal{D}(\psi + n_1, \dots, \psi + n_C)$

for $c \leftarrow 1$ to C **do**
 M-H step: update ρ_c
 sample: $\rho'_c \sim \text{L-S}(\rho_{c,m-1}, L)$ and $u \sim \mathcal{U}(0, 1)$
 compute: *ratio* \leftarrow equation (2.4) with $\rho \leftarrow \rho_{c,m-1}$ and $\alpha \leftarrow \alpha_{c,m-1}$, and where the sum is over $\{j : z_{j,m-1} = c\}$
 if $u < \text{ratio}$ **then** $\rho_{c,m} \leftarrow \rho'_c$
 else $\rho_{c,m} \leftarrow \rho_{c,m-1}$

 if $m \bmod \alpha_{\text{jump}} = 0$ **then M-H step: update** α_c
 sample: $\alpha'_c \sim \mathcal{N}(\alpha_{c,m-1}, \sigma_\alpha^2)$ and $u \sim \mathcal{U}(0, 1)$
 compute: *ratio* \leftarrow equation (2.6) with $\rho \leftarrow \rho_{c,m}$ and $\alpha \leftarrow \alpha_{c,m-1}$, and where the sum is over $\{j : z_{j,m-1} = c\}$
 if $u < \text{ratio}$ **then** $\alpha_{c,m} \leftarrow \alpha'_c$
 else $\alpha_{c,m} \leftarrow \alpha_{c,m-1}$

end

Gibbs step: update z_1, \dots, z_N
 for $j \leftarrow 1$ to N **do**
 foreach $c \leftarrow 1$ to C **do** compute cluster assignment probabilities: $p_{cj} = \frac{\tau_{c,m}}{Z_n(\alpha_{c,m})} \exp\left[\frac{-\alpha_{c,m}}{n} d(\tilde{\mathbf{R}}_j^{m-1}, \rho_{c,m})\right]$
 sample: $z_{j,m} \sim \mathcal{Mn}(p_{1j}, \dots, p_{Cj})$
 end

M-H step: update $\tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N$:
 for $j \leftarrow 1$ to N **do**
 if $\{S_1, \dots, S_N\}$ among inputs **then** sample: $\tilde{\mathbf{R}}'_j$ in S_j from the L-S distribution centered at $\tilde{\mathbf{R}}_j^{m-1}$
 else sample: $\tilde{\mathbf{R}}'_j$ from the L-S distribution centered at $\tilde{\mathbf{R}}_j^{m-1}$ and compatible with $tc(\mathcal{B}_j)$
 compute: *ratio* \leftarrow equation (2.20) with $\rho \leftarrow \rho_{z_{j,m},m}$, $\alpha \leftarrow \alpha_{z_{j,m},m}$ and $\tilde{\mathbf{R}}_j \leftarrow \tilde{\mathbf{R}}_j^{m-1}$
 sample: $u \sim \mathcal{U}(0, 1)$
 if $u < \text{ratio}$ **then** $\tilde{\mathbf{R}}_j^m \leftarrow \tilde{\mathbf{R}}'_j$
 else $\tilde{\mathbf{R}}_j^m \leftarrow \tilde{\mathbf{R}}_j^{m-1}$
 end

end

Algorithm 5: MCMC Sampler for full rankings

input : ρ, α, d, N, L
output: $\mathbf{R}_1, \dots, \mathbf{R}_N$
Initialization of the MCMC: randomly generate $\mathbf{R}_{1,0}, \dots, \mathbf{R}_{N,0}$

for $m \leftarrow 1$ to M **do**
 for $j \leftarrow 1$ to N **do**
 sample $\mathbf{R}'_j \sim \text{L-S}(\mathbf{R}_{j,m-1}, L)$ and $u \sim \mathcal{U}(0, 1)$
 compute: *ratio* $= \frac{P_L(\mathbf{R}_j | \mathbf{R}'_j)}{P_L(\mathbf{R}'_j | \mathbf{R}_j)} \exp\left\{-\frac{\alpha}{n} \sum_{j=1}^N [d(\mathbf{R}'_j, \rho) - d(\mathbf{R}_j, \rho)]\right\}$ with $\mathbf{R}_j \leftarrow \mathbf{R}_{j,m-1}$
 if $u < \text{ratio}$ **then** $\tilde{\mathbf{R}}_j^m \leftarrow \mathbf{R}'_j$
 else $\tilde{\mathbf{R}}_j^m \leftarrow \tilde{\mathbf{R}}_j^{m-1}$
 end

end

parameter $\psi = (\psi_1, \dots, \psi_C)$ of the Dirichlet density over the proportion of assessors in the clusters, and $d(\cdot, \cdot)$. The algorithm then returns the rankings $\mathbf{R}_1, \dots, \mathbf{R}_N$, sampled from a Mixture of Mallows models, as well as the the cluster assignments z_1, \dots, z_N .

For generating top- k rankings, we simply generate $\mathbf{R}_1, \dots, \mathbf{R}_N$ with Algorithm 5, and

Algorithm 6: MCMC Sampler for full rankings with clusters

input : $C, \rho_{1:C}, \alpha_{1:C}, \psi, d, N, L$
output: $\mathbf{R}_1, \dots, \mathbf{R}_N$ and z_1, \dots, z_N
Initialization of the MCMC: randomly generate $\mathbf{R}_{1,0}, \dots, \mathbf{R}_{N,0}, \tau_1, \dots, \tau_C \sim \text{Dir}(\psi)$, and $z_1, \dots, z_N \sim \mathcal{M}n(1, \tau_1, \dots, \tau_C)$
for $m \leftarrow 1$ **to** M **do**
 for $c \leftarrow 1$ **to** C **do**
 | | compute: $N_c = \sum_{j=1}^N \mathbb{1}_c(z_j)$, and sample N_c ranks with Algoritihm 5
 | | **end**
 end
end

then keep only the top- k items. In case of clusters, we do the same as above, but starting with Algorithm 6.

Finally, to sample data made of pairwise comparisons, we first generate $\mathbf{R}_1, \dots, \mathbf{R}_N$ with Algoritihm 5. Then, we select the number of pairwise comparisons, $T_1, \dots, T_N, T_j < n(n-1)/2$, for all $j = 1, \dots, N$, that each assessor will evaluate². Finally, given $\mathbf{R}_1, \dots, \mathbf{R}_N$ and T_1, \dots, T_N , for each assessor j , we randomly sample without replacement T_j pairs in the collection of all possible $n(n-1)/2$ pairs, and order each of them according to \mathbf{R}_j . For generating pairwise comparisons with clusters, we follow the previous procedure, but starting with Algorithm 6.

²It is possible to choose a different number of pairs per assessor.

Chapter 3

The Bayesian Mallows model for non-transitive pair comparisons

In this chapter, we propose a flexible extension of the model of Chapter 2, able to learn the individual rankings of a set of items from non-transitive pairwise comparison data. Lack of transitivity may naturally arise when the items compared are perceived as rather similar, if the pairwise comparisons are presented sequentially without allowing for consistency checks, or simply because of users' inattentiveness. Situations of this kind are very common when the set of items is large, and when the users are unlikely to be able, or willing, to compare all of them in order to perform a ranking. In such cases, a pairwise comparison experiment is then often preferred, and sometimes it is the only possible experimental procedure (Agresti 1996, David 1963). As already discussed in Section 1.2.1, to our knowledge most of the methods to estimate individual rankings from sparse (not repeated) pairwise comparison data are not able to handle individual-level non-transitivity. They either drop such pairs, or they only focus on the estimation of the consensus ranking, which can indeed contradict individual preferences, and in such way they do not specifically model the non-transitivity characterizing the data. For a complete description of these methods we refer to Section 1.2.1. Instead, we incorporate the non-transitive patterns of the data directly into the Bayesian Mallows model for pairwise comparisons of Section 2.3.2, and provide a strategy to estimate, with uncertainty, the individual preferences of the users. We accomplish this by postulating the existence of a true latent individual ranking of the items, and assuming that non-transitive patterns arise because users make mistakes by switching the order between some pairs under comparison.

This chapter contains joint work with Valeria Vitelli, Elja Arjas, Natasha Barrett and Arnaldo Frigessi, and is based on Crispino et al. (2017).

Outline

In Section 3.1, we present the basic Mallows model for non-transitive pair comparisons, and we propose some model specifications, including two mixture models. In particular, in Section 3.1.1 we model the probability of making a mistake as a constant, independent of the pairs being assessed, and also independent of all other comparisons made by the same user. This models, for example, a mouse click mistake or a random preference between a pair of items. In Sections 3.1.2 and 3.1.3, instead, we assume that the probability of making a mistake depends on the items compared: the stronger is the preference between a pair of items, the smaller is the probability of making a mistake. This models a situation when items are more easily mis-compared by the user when they are rather similar in the personal ranking (Section 3.1.2), or in the consensus ranking (Section 3.1.3). We then develop two mixture model extensions: in 3.1.4 we consider users who differ in their ability to stay consistent with logical transitivity when announcing their pairwise comparisons, which results in a mixture on the probability of making a mistake; in 3.1.5, we deal with the case when users are suspected to be heterogeneous in their preferences, that calls for a model able to learn individual preferences associated to multiple consensus rankings. The Markov Chain Monte Carlo algorithm, is outlined in Section 3.2 (and further specified in 3.A), while Section 3.3 is devoted to simulations.

In Section 3.4 we then report the analysis of two toy datasets, also including the Beach preference data of Section 2.4.2. Finally, in Section 3.5 we discuss the contributions of this chapter.

3.1 The main model

We consider the situation where N users independently express their preferences between pairs of items in a set $\mathcal{A} = \{A_1, \dots, A_n\}$, as in Section 2.3.2. In many situations of practical interest the users do not decide on the set of pairs to be considered, which are instead assigned to the users by an external authority. We here don't model the way in which the pairs are chosen, and simply assume a general framework, where each user j receives a different subset $\mathcal{C}_j = \{\mathcal{C}_{j1}, \dots, \mathcal{C}_{jT_j}\}$ of $T_j \leq n(n-1)/2$ random pairs. Let $\mathcal{B}_j = \{\mathcal{B}_{j1}, \dots, \mathcal{B}_{jT_j}\}$ be the set of pairwise preferences given by user j , where \mathcal{B}_{jt} is the order that user j assigned to the pair \mathcal{C}_{jt} . For example, if $\mathcal{C}_{jt} = \{A_{t_1}, A_{t_2}\}$, it could be that $\mathcal{B}_{jt} = (A_{t_1} \prec A_{t_2})$, $t_1, t_2 \in \{1, \dots, n\}$, meaning that item A_{t_1} is preferred to item A_{t_2} . Such data are incomplete since not all items, nor pairs, are handled by each user. We here assume no ties in the data, that is, users are forced to express their preference for all pairs in the list \mathcal{C}_j assigned to them, and neither indifference nor abstention are permitted.

Like in Section 2.3.2, we assume that each user j has a personal latent ranking, $\tilde{\mathbf{R}}_j \in \mathcal{P}_n$, distributed according the Mallows density of eq. (1.6), $\tilde{\mathbf{R}}_1, \dots, \tilde{\mathbf{R}}_N | \boldsymbol{\rho}, \alpha \stackrel{i.i.d.}{\sim} \mathcal{M}(\boldsymbol{\rho}, \alpha)$. In this chapter we will denote the individual latent rankings $\tilde{\mathbf{R}}_j$ by \mathbf{R}_j , for simplifying the notation. This should not create confusion, since the data considered in this chapter are only in the form of pairwise preferences, $\mathcal{B}_{1:N}$, while by $\mathbf{R}_{1:N}$ we denote the latent individual rankings. As a consequence, since \mathbf{R}_j is here a random variable, it is object of inference itself.

We model the situation where each user j , when announcing her preferences, mentally matches the items under comparison with her latent ranking \mathbf{R}_j . The situation considered in Section 2.3.2, corresponded to users who were consistent with \mathbf{R}_j . Then the pairwise orderings in \mathcal{B}_j were induced by \mathbf{R}_j following the rule:

$$(A_{t_1} \prec A_{t_2}) \iff R_{jt_1} < R_{jt_2}, \quad (3.1)$$

where R_{jt_i} denotes the rank of item A_{t_i} in \mathbf{R}_j . Being the preferences in \mathcal{B}_j induced by a full ranking, the set contained only transitive preferences. Under these assumptions, inference for the Mallows parameters α and $\boldsymbol{\rho}$, and the individual rankings $\mathbf{R}_{1:N}$, was performed by first computing the transitive closure of each preference set, $\text{tc}(\mathcal{B}_j)$, and second, by integrating out all the rankings $\mathbf{R} \in \mathcal{P}_n$ that were compatible with the transitive closure of the preference sets, here denoted by $\mathbf{R}_j \leftarrow \text{tc}(\mathcal{B}_j)$. This corresponded to the following

posterior distribution,

$$\begin{aligned}
P(\alpha, \boldsymbol{\rho} | \mathcal{B}_{1:N}) &= \sum_{\mathbf{R}_1 \leftarrow \text{tc}(\mathcal{B}_1)} \dots \sum_{\mathbf{R}_N \leftarrow \text{tc}(\mathcal{B}_N)} P(\alpha, \boldsymbol{\rho}, \mathbf{R}_{1:N} | \mathcal{B}_{1:N}) = \sum_{\mathbf{R}_1 \leftarrow \text{tc}(\mathcal{B}_1)} \dots \sum_{\mathbf{R}_N \leftarrow \text{tc}(\mathcal{B}_N)} P(\alpha, \boldsymbol{\rho} | \mathbf{R}_{1:N}) \propto \\
&\propto \pi(\alpha) \pi(\boldsymbol{\rho}) \prod_{j=1}^N \left[\sum_{\mathbf{R}_j \leftarrow \text{tc}(\mathcal{B}_j)} P(\mathbf{R}_j | \alpha, \boldsymbol{\rho}) \right]. \tag{3.2}
\end{aligned}$$

In equation (3.2) is implicitly assumed: (i) $(\mathcal{B}_{1:N} \perp\!\!\!\perp \alpha, \boldsymbol{\rho}) | \mathbf{R}_{1:N}$; (ii) $\forall j, k \in \{1, \dots, N\}$, $(\mathcal{B}_j \perp\!\!\!\perp \mathcal{B}_k) | \mathbf{R}_{1:N}$; (iii) $\forall j \in \{1, \dots, N\}$, $P(\mathcal{B}_j | \mathbf{R}_{1:N}) = P(\mathcal{B}_j | \mathbf{R}_j) = 1$, for all $\mathbf{R}_j \leftarrow \text{tc}(\mathcal{B}_j)$.

In this chapter, instead, we consider the case of users who are not fully consistent with their latent rankings: the pairwise orderings in \mathcal{B}_j may not be mutually compatible. Since the transitive closure of a non-transitive set does not exist, the previous procedure cannot be followed in such a case. We propose a probabilistic strategy to deal with this issue, based on the assumption that non-transitivities are due to mistakes in deriving the pair order from the latent ranking \mathbf{R}_j . The likelihood of set of preferences \mathcal{B}_j , analogous to the summation of eq. (3.2), second line, is

$$P(\mathcal{B}_j | \alpha, \boldsymbol{\rho}) = \sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathcal{B}_j, \mathbf{R}_j | \alpha, \boldsymbol{\rho}) = \sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathbf{R}_j | \alpha, \boldsymbol{\rho}) P(\mathcal{B}_j | \mathbf{R}_j), \tag{3.3}$$

where $P(\mathcal{B}_j | \mathbf{R}_j)$ is the probability of ordering the pairs in \mathcal{C}_j as in \mathcal{B}_j , possibly generating non-transitivities, when the latent ranking for user j is \mathbf{R}_j . It can therefore be seen as forming the error model in this context, which will be specified in complete detail in the next sections. The joint posterior of the model parameters is then:

$$P(\alpha, \boldsymbol{\rho} | \mathcal{B}_1, \dots, \mathcal{B}_N) \propto \pi(\alpha) \pi(\boldsymbol{\rho}) \prod_{j=1}^N \left[\sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathbf{R}_j | \alpha, \boldsymbol{\rho}) P(\mathcal{B}_j | \mathbf{R}_j) \right],$$

where we assumed a truncated gamma prior, $\pi(\alpha) \propto \alpha^{\gamma-1} e^{-\lambda\alpha} \mathbb{1}_{[0, \alpha_{\max})}(\alpha)$, for α , and the uniform prior on \mathcal{P}_n , $\pi(\boldsymbol{\rho}) = \frac{\mathbb{1}_{\mathcal{P}_n}(\boldsymbol{\rho})}{n!}$, for $\boldsymbol{\rho}$ (see Section 2.1.1, for justifications, and Chapters 5 and 6 for alternatives). This strategy is able to recover possible linear orderings close, in terms of some given distance, to the non-transitive sets of preferences. We developed two basic models for $P(\mathcal{B}_j | \mathbf{R}_j)$, that is the probability of making a mistake: the Bernoulli model (BM) and the Logistic model (LM). In BM, outlined in Section 3.1.1, we assume that non-transitivities arise from random mistakes while LM, presented

in Section 3.1.2, assumes that non-transitivities arise from mistakes due to difficulty in ordering items that are perceived individually rather similarly. In Sections 3.1.3, 3.1.4 and 3.1.5, we then present some extensions of BM and LM.

3.1.1 Bernoulli model (BM) for mistakes

We first assume that the pairwise comparisons given by a user are conditionally independent given her latent ranking \mathbf{R}_j ,

$$P(\mathcal{B}_j | \mathbf{R}_j) = \prod_{t=1}^{T_j} P(\mathcal{B}_{jt} | \mathbf{R}_j). \quad (3.4)$$

We here define $g(\mathcal{B}_{jt}, \mathbf{R}_j)$, an indicator function of a given comparison $\mathcal{B}_{jt} = (A_{t_1} \prec A_{t_2})$ and of a given ranking $\mathbf{R}_j = \{R_{j1}, \dots, R_{jn}\} \in \mathcal{P}_n$,

$$g(\mathcal{B}_{jt}, \mathbf{R}_j) = \begin{cases} 0 & \text{if } R_{jt_1} < R_{jt_2} \\ 1 & \text{otherwise,} \end{cases}$$

where t_1 is the index of the preferred item A_{t_1} in \mathcal{B}_{jt} , the t -th comparison of user j , and t_2 is the index of the least preferred item. Thus $g(\mathcal{B}_{jt}, \mathbf{R}_j) = 1$ if the preference order of \mathcal{B}_{jt} is not implied by the ranking \mathbf{R}_j , in the sense of eq. (3.1), i.e. \mathcal{B}_{jt} and \mathbf{R}_j disagree in their preference ordering between items A_{t_1} and A_{t_2} .

We then assume the following Bernoulli type model for modeling the probability that a user j makes a mistake in a given pairwise comparison \mathcal{B}_{jt} , i.e. the probability that she reverses the true latent preference implied by her latent ranking \mathbf{R}_j :

$$P(\mathcal{B}_{jt} \text{ mistake} | \theta, \mathbf{R}_j) = P[g(\mathcal{B}_{jt}, \mathbf{R}_j) = 1 | \theta, \mathbf{R}_j] = \theta, \quad \theta \in [0, 0.5).$$

The probability of eq. (3.4) is then given by

$$P(\mathcal{B}_j | \theta, \mathbf{R}_j) = \left(\frac{\theta}{1 - \theta} \right)^{\sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)} (1 - \theta)^{T_j}. \quad (3.5)$$

As prior density for θ , we choose the Beta distribution truncated on the interval $[0, 0.5)$, with given hyperparameters κ_1 and κ_2 : $\pi(\theta) \propto \theta^{\kappa_1 - 1} (1 - \theta)^{\kappa_2 - 1} \mathbb{1}_{[0, 0.5)}(\theta)$. We truncate it on $[0, 0.5)$ mainly for identification purposes, but also because we want to

force the probability of making a mistake to be less than 0.5. The reason for this, lies in the fact that we do not admit the possibility that a users makes more than 50% of mistakes in data she provides. The posterior density of the model parameters, defined on the support, $S = \mathbb{1}(\{0 \leq \alpha < \alpha_{\max}\} \cap \{\boldsymbol{\rho} \in \mathcal{P}_n\} \cap \{\mathbf{R}_j \in \mathcal{P}_n\}_{j=1}^N \cap \{0 \leq \theta < 0.5\})$, has then the following form,

$$P(\alpha, \boldsymbol{\rho}, \theta | \mathcal{B}_{1:N}) \propto \pi(\alpha) \pi(\boldsymbol{\rho}) \pi(\theta) \prod_{j=1}^N \left[\sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathbf{R}_j | \alpha, \boldsymbol{\rho}) P(\mathcal{B}_j | \theta, \mathbf{R}_j) \right]. \quad (3.6)$$

We sample from the density of equation (3.6), through an augmented sampling scheme by first updating $\alpha, \boldsymbol{\rho}$ and θ given $\mathcal{B}_{1:N}$ and $\mathbf{R}_{1:N}$, and then, updating $\mathbf{R}_{1:N}$ given $\alpha, \boldsymbol{\rho}, \theta$ and $\mathcal{B}_{1:N}$. The former step is done by using the conditional density

$$P(\alpha, \boldsymbol{\rho}, \theta | \mathcal{B}_{1:N}, \mathbf{R}_{1:N}) = \alpha^{\gamma-1} e^{-\alpha [\lambda + \frac{1}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho})] - N \ln[Z_n(\alpha)]} \cdot \left(\frac{\theta}{1-\theta} \right)^{\kappa_1 - 1 + \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)} (1-\theta)^{\kappa_2 + \kappa_1 - 2 + \sum_{j=1}^N T_j}. \quad (3.7)$$

The second step, is performed by using the density

$$P(\mathbf{R}_{1:N} | \alpha, \boldsymbol{\rho}, \theta, \mathcal{B}_{1:N}) \propto P(\mathbf{R}_{1:N} | \alpha, \boldsymbol{\rho}) P(\mathcal{B}_{1:N} | \theta, \mathbf{R}_{1:N}) = \frac{e^{-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho})}}{[Z_n(\alpha)]^N} \left(\frac{\theta}{1-\theta} \right)^{\sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)} (1-\theta)^{\sum_{j=1}^N T_j}. \quad (3.8)$$

Figure 3.1 shows the graphical representation of the Bernoulli model for mistakes. The MCMC algorithm is presented in details in Section 3.2.

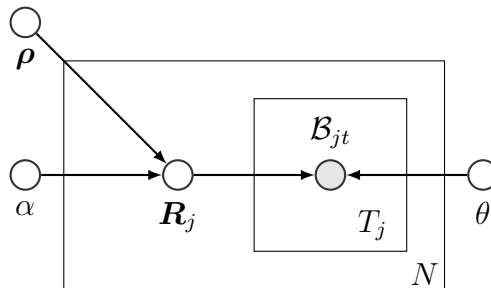


Figure 3.1: Graphical representation of the Bernoulli model for mistakes.

3.1.2 Logistic model (LM) for mistakes

The idea of the logistic model for mistakes is that a user j is more likely to be confused, and consequently to make a mistake, if two items in pair are more similar according to her latent ranking vector \mathbf{R}_j . We assume the following logistic type model for the probability of making a mistake in a given pairwise comparison

$$\text{logit } P[g(\mathcal{B}_{jt}, \mathbf{R}_j) = 1 \mid \mathbf{R}_j, \beta_0, \beta_1] = -\beta_0 - \beta_1 \frac{d_{\mathbf{R}_j, t} - 1}{n - 2},$$

where $d_{\mathbf{R}_j, t}$ is the l_1 distance of the ranks of the two items under comparison in \mathcal{B}_{jt} , according to \mathbf{R}_j : if $\mathcal{B}_{jt} = (A_{t_1} \prec A_{t_2})$, then $d_{\mathbf{R}_j, t} = |R_{jt_1} - R_{jt_2}|$. Note that $d_{\mathbf{R}_j, t} \in \{1, \dots, n - 1\}$, and thus its minimum value $d_{\mathbf{R}_j, t} = 1$ serves as a reference value, and β_0 is the corresponding parameter for a mistake in case of paired items which are ranked as neighbors. This corresponds to the usual practice in logistic regressions, where one of the covariate levels is chosen as a reference, with its own intercept parameter.

We assume that β_1 and β_0 are a priori independent and distributed according to a gamma prior, $\beta_1 \sim \Gamma(\lambda_{11}, \lambda_{12})$, and $\beta_0 \sim \Gamma(\lambda_{01}, \lambda_{02})$. These choices are motivated by the fact that we want to model a negative dependence between the distance of the items and the probability of making a mistake ($\beta_1 > 0$), and second, we want to force the probability of making a mistake when the items have ranks differing by 1 to be less than 0.5 ($\beta_0 > 0$), for the same reasons that drove this choice in the case of the BM.

The rationale behind this model is that if two items have ranks close to each other in the latent ranking \mathbf{R}_j , their relative preference is presumably rather vague, and this could lead to inverting their order in \mathcal{B}_j . If this happens, it should have only a relatively small influence on the likelihood. In contrast, if two items have far away ranks in \mathbf{R}_j , then their mutual preference should be clearer, and incorrectly reversing their ordering should have a large influence on the likelihood. Therefore, the posterior density of the model, defined on the support, $S = \mathbb{1}(\{0 \leq \alpha < \alpha_{\max}\} \cap \{\boldsymbol{\rho} \in \mathcal{P}_n\} \cap \{\mathbf{R}_{1:N} \in \mathcal{P}_n\} \cap \{\beta_1 > 0\} \cap \{\beta_0 > 0\})$, is given by

$$P(\alpha, \boldsymbol{\rho}, \beta_0, \beta_1 \mid \mathcal{B}_{1:N}) \propto \pi(\beta_0) \pi(\beta_1) \pi(\boldsymbol{\rho}) \pi(\alpha) \prod_{j=1}^N \left[\sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathbf{R}_j \mid \alpha, \boldsymbol{\rho}) P(\mathcal{B}_j \mid \beta_0, \beta_1, \mathbf{R}_j) \right]. \quad (3.9)$$

Analogously to equation (3.6), we sample from the posterior of equation (3.9) by first updating $\alpha, \boldsymbol{\rho}, \beta_0$ and β_1 , given $\mathcal{B}_{1:N}$ and $\mathbf{R}_{1:N}$, that is from the conditional

$$\begin{aligned}
P(\alpha, \boldsymbol{\rho}, \beta_0, \beta_1 | \mathcal{B}_{1:N}, \mathbf{R}_{1:N}) &\propto \alpha^{\gamma-1} \beta_0^{\lambda_{01}-1} \beta_1^{\lambda_{11}-1} \left[\prod_{j=1}^N \prod_{t=1}^{T_j} \left(1 + e^{-\beta_0 - \beta_1 \frac{d_{\mathbf{R}_j, t} - 1}{n-2}} \right) \right]^{-1} \\
&\cdot e^{-\alpha \left[\lambda + \frac{1}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right] - \beta_0 \left[\lambda_{02} + \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j) \right]} \\
&\cdot e^{-\beta_1 \left[\lambda_{12} + \frac{1}{n-2} \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)(d_{\mathbf{R}_j, t} - 1) \right] - N \ln[Z_n(\alpha)]}.
\end{aligned} \tag{3.10}$$

Secondly, we update $\mathbf{R}_{1:N}$, given $\alpha, \boldsymbol{\rho}, \beta_0, \beta_1$, and $\mathcal{B}_{1:N}$, from

$$\begin{aligned}
P(\mathbf{R}_{1:N} | \alpha, \boldsymbol{\rho}, \beta_0, \beta_1, \mathcal{B}_{1:N}) &\propto P(\mathbf{R}_{1:N} | \alpha, \boldsymbol{\rho}) P(\mathcal{B}_{1:N} | \beta_0, \beta_1, \mathbf{R}_{1:N}) \propto \\
&\propto e^{-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) - N \ln[Z_n(\alpha)] - \beta_0 \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)} \\
&\cdot e^{-\frac{\beta_1}{n-2} \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)(d_{\mathbf{R}_j, t} - 1)} \left[\prod_{j=1}^N \prod_{t=1}^{T_j} \left(1 + e^{-\beta_0 - \beta_1 \frac{d_{\mathbf{R}_j, t} - 1}{n-2}} \right) \right]^{-1}.
\end{aligned} \tag{3.11}$$

The algorithm for the Logistic mistake model is sketched in Appendix 3.A.

3.1.3 Logistic-consensus model (LCM) for mistakes

It is natural to think of a modification of the LM, where the probability of making a mistake depends on some dissimilarity between the items, shared by all the users, rather than on the individual ranking. For example, each user could be more likely to make a mistake, when assessing a pairwise comparison, if the two items in pair are more similar according to some given item-dependent covariate, $\mathbf{Y} = (Y_1, \dots, Y_n)$. We here develop a model where \mathbf{Y} coincides with the unknown consensus ranking of the items $\boldsymbol{\rho} = (\rho_1, \dots, \rho_n)$.

We then assume the following specialization of the logistic model of Section 3.1.2, for the probability of making a mistake

$$\text{logit } P[g(\mathcal{B}_{jt}, \mathbf{R}_j) = 1 | \boldsymbol{\rho}, \beta_0, \beta_1] = -\beta_0 - \beta_1 \frac{d_{\boldsymbol{\rho}, t} - 1}{n-2},$$

where $d_{\boldsymbol{\rho}, t}$ is defined, similarly to Section 3.1.2, as the l_1 distance of the ranks of the two items under comparison in \mathcal{B}_{jt} , according to $\boldsymbol{\rho}$: if $\mathcal{B}_{jt} = (A_{t_1} \prec A_{t_2})$, then $d_{\boldsymbol{\rho}, t} = |\rho_{t_1} - \rho_{t_2}|$. Note, as before, that $d_{\boldsymbol{\rho}, t} \in \{1, \dots, n-1\}$, thus $d_{\boldsymbol{\rho}, t} = 1$ serves as a reference value, and β_0 is the corresponding parameter for a mistake in case of paired items which are ranked as

neighbors. The prior densities of β_1 and β_0 are the same as in Section 3.1.2. The posterior density of this model is therefore given by

$$P(\alpha, \boldsymbol{\rho}, \beta_0, \beta_1 | \mathcal{B}_{1:N}) \propto \pi(\beta_0)\pi(\beta_1)\pi(\boldsymbol{\rho})\pi(\alpha) \prod_{j=1}^N \sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathbf{R}_j | \alpha, \boldsymbol{\rho}) P(\mathcal{B}_j | \beta_0, \beta_1, \boldsymbol{\rho}, \mathbf{R}_j), \quad (3.12)$$

where

$$P(\mathcal{B}_j | \beta_0, \beta_1, \boldsymbol{\rho}, \mathbf{R}_j) = \prod_{t=1}^{T_j} P(\mathcal{B}_{jt} | \beta_0, \beta_1, \boldsymbol{\rho}, \mathbf{R}_j) = \prod_{t=1}^{T_j} \frac{\left(e^{-\beta_0 - \beta_1 \frac{d_{\boldsymbol{\rho}, t-1}}{n-2}} \right)^{g(\mathcal{B}_{jt}, \mathbf{R}_j)}}{1 + e^{-\beta_0 - \beta_1 \frac{d_{\boldsymbol{\rho}, t-1}}{n-2}}}.$$

The sampling scheme is similar to the one outlined in Section 3.1.2: we sample from the posterior of equation (3.12) by first updating $\alpha, \boldsymbol{\rho}, \beta_0$ and β_1 , given $\mathcal{B}_{1:N}$ and $\mathbf{R}_{1:N}$ from the conditional distribution

$$\begin{aligned} P(\alpha, \boldsymbol{\rho}, \beta_0, \beta_1 | \mathcal{B}_{1:N}, \mathbf{R}_{1:N}) &\propto \alpha^{\gamma-1} \beta_0^{\lambda_{01}-1} \beta_1^{\lambda_{11}-1} \left[\prod_{j=1}^N \prod_{t=1}^{T_j} \left(1 + e^{-\beta_0 - \beta_1 \frac{d_{\boldsymbol{\rho}, t-1}}{n-2}} \right) \right]^{-1} \\ &\cdot e^{-\alpha \left[\lambda + \frac{1}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right] - \beta_0 \left[\lambda_{02} + \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j) \right]} \\ &\cdot e^{-\beta_1 \left[\lambda_{12} + \frac{1}{n-2} \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)(d_{\boldsymbol{\rho}, t-1}) \right] - N \ln[Z_n(\alpha)]}. \end{aligned} \quad (3.13)$$

Secondly, we update $\mathbf{R}_{1:N}$, given $\alpha, \boldsymbol{\rho}, \beta_0, \beta_1$, and $\mathcal{B}_{1:N}$, from

$$\begin{aligned} P(\mathbf{R}_{1:N} | \alpha, \boldsymbol{\rho}, \beta_0, \beta_1, \mathcal{B}_{1:N}) &\propto P(\mathbf{R}_{1:N} | \alpha, \boldsymbol{\rho}) P(\mathcal{B}_{1:N} | \beta_0, \beta_1, \boldsymbol{\rho}, \mathbf{R}_{1:N}) \propto \\ &\propto e^{-\frac{\alpha}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) - N \ln[Z_n(\alpha)] - \beta_0 \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)} \\ &\cdot e^{-\frac{\beta_1}{n-2} \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)(d_{\boldsymbol{\rho}, t-1})} \left[\prod_{j=1}^N \prod_{t=1}^{T_j} \left(1 + e^{-\beta_0 - \beta_1 \frac{d_{\boldsymbol{\rho}, t-1}}{n-2}} \right) \right]^{-1}. \end{aligned} \quad (3.14)$$

The algorithm for the Logistic-consensus mistake model is in a straightforward specialization of Algorithm 8, presented in Appendix 3.A.

3.1.4 Mixture model on θ

Suppose that the users are thought to differ in their ability to stay consistent with logical transitivity when announcing their pairwise comparisons. Accounting for such heterogeneity leads to a generalization of the model of Section 3.1.1, where a mixture of Bernoulli

distributions is proposed for modeling the generation of mistakes, and a consequent grouping of the users into clusters is performed. In this section, we assume that all the N users share the same consensus ranking $\boldsymbol{\rho}$ and dispersion parameter, α , and only differ in the θ parameter, that is the probability of making a mistake when announcing a pairwise preference between a pair of items. Letting $\xi_1, \dots, \xi_N \in \{1, \dots, K\}$ assign each user to one of K clusters, each described by a different θ_k , $k = 1, \dots, K$, the likelihood is

$$P(\mathcal{B}_{1:N} | \alpha, \boldsymbol{\rho}, \theta_{1:K}, \xi_{1:N}) = \prod_{j=1}^N \left[\sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathbf{R}_j | \alpha, \boldsymbol{\rho}) P(\mathcal{B}_j | \theta_{\xi_j}, \mathbf{R}_j) \right],$$

where

$$P(\mathcal{B}_j | \theta_{\xi_j}, \mathbf{R}_j) = \left(\frac{\theta_{\xi_j}}{1 - \theta_{\xi_j}} \right)^{\sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)} (1 - \theta_{\xi_j})^{T_j}. \quad (3.15)$$

We assume that the θ_k are i.i.d. according to a Beta distribution truncated on the interval $[0, 0.5)$, with hyperparameters κ_1 and κ_2 : $\pi(\theta_k) \propto \theta_k^{\kappa_1-1} (1 - \theta_k)^{\kappa_2-1} \mathbb{1}_{[0,0.5)}(\theta_k)$, $k = 1, \dots, K$. We further assume that the cluster labels are a priori conditionally independent given the mixing parameters of the clusters, τ_1, \dots, τ_K , and distributed according to a categorical distribution

$$P(\xi_1, \dots, \xi_N | \tau_1, \dots, \tau_K) \propto \prod_{j=1}^N \tau_{\xi_j} = \prod_{j=1}^N \prod_{k=1}^K \tau_k^{\mathbb{1}_k(\xi_j)},$$

where $\tau_k \geq 0$, $\forall k = 1, \dots, K$ and $\sum_k \tau_k = 1$. Finally, we assign to τ_1, \dots, τ_K the Dirichlet density with parameter ψ . These choices lead to the following posterior density

$$P(\alpha, \boldsymbol{\rho}, \theta_{1:K}, \xi_{1:N}, \tau_{1:K} | \mathcal{B}_{1:N}) \propto \pi(\alpha) \pi(\boldsymbol{\rho}) \prod_{k=1}^K [\pi(\theta_k) \pi(\tau_k)] \cdot \prod_{j=1}^N \left\{ P(\xi_j | \tau_{1:K}) \sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathbf{R}_j | \alpha, \boldsymbol{\rho}) P(\mathcal{B}_j | \theta_{\xi_j}, \mathbf{R}_j) \right\}. \quad (3.16)$$

Similarly to the homogeneous case, we sample from the posterior of equation (3.16) by first updating $\alpha, \boldsymbol{\rho}, \tau_{1:K}, \xi_{1:N}$ and $\theta_{1:K}$ given $\mathcal{B}_{1:N}$ and $\mathbf{R}_{1:N}$, and then updating $\mathbf{R}_{1:N}$ given $\alpha, \boldsymbol{\rho}, \tau_{1:K}, \xi_{1:N}, \theta_{1:K}$ and $\mathcal{B}_{1:N}$. The former step is done by using the conditional density,

$$\begin{aligned}
P(\alpha, \boldsymbol{\rho}, \tau_{1:K}, \theta_{1:K}, \xi_{1:N} | \mathcal{B}_{1:N}, \mathbf{R}_{1:N}) &\propto \alpha^{\gamma-1} \prod_{k=1}^K \left[\tau_k^{\psi-1 + \sum_{j=1}^N \mathbb{1}_k(\xi_j)} \theta_k^{\kappa_1-1} (1-\theta_k)^{\kappa_2-1} \right] \\
&\cdot \prod_{j=1}^N \left[\frac{e^{-\alpha[\lambda + \frac{1}{n}d(\mathbf{R}_j, \boldsymbol{\rho})]}}{Z_n(\alpha)} \left(\frac{\theta_{\xi_j}}{1-\theta_{\xi_j}} \right)^{\sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)} (1-\theta_{\xi_j})^{T_j} \right], \tag{3.17}
\end{aligned}$$

The second step, is performed by using the density,

$$P(\mathbf{R}_{1:N} | \alpha, \boldsymbol{\rho}, \tau_{1:K}, \theta_{1:K}, \xi_{1:N}, \mathcal{B}_{1:N}) \propto \prod_{j=1}^N \frac{e^{-\frac{\alpha}{n}d(\mathbf{R}_j, \boldsymbol{\rho})}}{Z_n(\alpha)} P(\mathcal{B}_j | \theta_{\xi_j}, \mathbf{R}_j), \tag{3.18}$$

where $P(\mathcal{B}_j | \theta_{\xi_j}, \mathbf{R}_j)$ is defined in eq. (3.15). Figure 3.2 shows the graphical representation of the hierarchical construction of the mixture model for mistakes. The algorithm for this mixture extension is in Appendix 3.A.

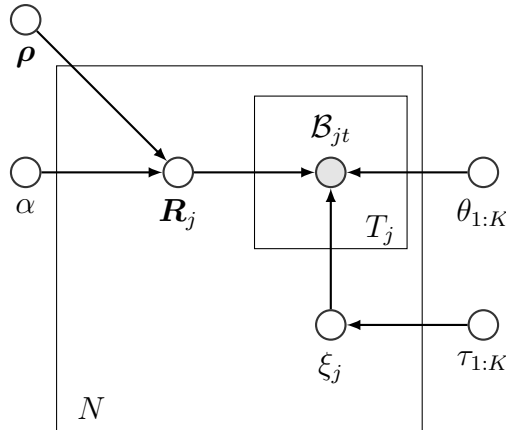


Figure 3.2: Graphical representation of the mixture model on θ .

3.1.5 Mixture model on α and ρ

So far we assumed that a unique consensus ranking was shared by all users. Since in many situations the assumption of homogeneity with respect to an underlying common consensus ranking is unrealistic, we here allow for clustering the users into separate subsets, each sharing a consensus ranking of the items, similarly to Section 2.3.3. In this section we then propose a mixture model generalization of the Bernoulli model of Section 3.1.1 to deal with heterogeneous users expressing pairwise preferences with mistakes.

Let $z_1, \dots, z_N \in \{1, \dots, C\}$ be the class labels indicating how individual users are assigned to one of the C clusters. Each cluster is described by a different pair of Mallows parameters $(\alpha_c, \boldsymbol{\rho}_c)$, $c = 1, \dots, C$, so that the likelihood has the form,

$$P(\mathcal{B}_{1:N} | \alpha_{1:C}, \boldsymbol{\rho}_{1:C}, \theta, \eta_{1:C}, z_{1:N}) = \prod_{j=1}^N \left[\sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathbf{R}_j | \alpha_{z_j}, \boldsymbol{\rho}_{z_j}) P(\mathcal{B}_j | \theta, \mathbf{R}_j) \right],$$

where

$$P(\mathbf{R}_j | \alpha_{z_j}, \boldsymbol{\rho}_{z_j}) = \frac{1}{Z_n(\alpha_{z_j})} \exp \left\{ -\frac{\alpha_{z_j}}{n} d(\mathbf{R}_j, \boldsymbol{\rho}_{z_j}) \right\}.$$

Again, we assume that the cluster labels are a priori conditionally independent given the mixing parameters of the clusters, η_1, \dots, η_C , and distributed according to a categorical distribution

$$P(z_1, \dots, z_N | \eta_1, \dots, \eta_C) \propto \prod_{j=1}^N \eta_{z_j} = \prod_{j=1}^N \prod_{c=1}^C \eta_c^{\mathbb{1}_c(z_j)},$$

where $\eta_c \geq 0$, $\forall c = 1, \dots, C$, $\sum_c \eta_c = 1$; further η_1, \dots, η_C are assumed to have Dirichlet density with parameter χ . These choices lead to the following posterior density,

$$P(\alpha_{1:C}, \boldsymbol{\rho}_{1:C}, \eta_{1:C}, z_{1:N}, \theta | \mathcal{B}_{1:N}) \propto \pi(\theta) \prod_{c=1}^C [\pi(\alpha_c) \pi(\boldsymbol{\rho}_c) \pi(\eta_c)] \cdot \prod_{j=1}^N \left[P(z_j | \eta_{1:C}) \sum_{\mathbf{R}_j \in \mathcal{P}_n} P(\mathbf{R}_j | \alpha_{z_j}, \boldsymbol{\rho}_{z_j}) P(\mathcal{B}_j | \theta, \mathbf{R}_j) \right]. \quad (3.19)$$

We sample from the posterior of eq. (3.19) by first updating $\alpha_{1:C}, \boldsymbol{\rho}_{1:C}, \eta_{1:C}, z_{1:N}$ and θ given $\mathcal{B}_{1:N}$ and $\mathbf{R}_{1:N}$, and second updating $\mathbf{R}_{1:N}$ given $\alpha_{1:C}, \boldsymbol{\rho}_{1:C}, \eta_{1:C}, z_{1:N}, \theta$ and $\mathcal{B}_{1:N}$. The former step is done by using the conditional density,

$$P(\alpha_{1:C}, \boldsymbol{\rho}_{1:C}, \eta_{1:C}, z_{1:N}, \theta | \mathcal{B}_{1:N}, \mathbf{R}_{1:N}) \propto \prod_{c=1}^C \left[\alpha_c^{\gamma-1} e^{-\lambda \alpha_c} \eta_c^{\chi-1 + \sum_{j=1}^N \mathbb{1}_c(z_j)} \right] \cdot \left(\frac{\theta}{1-\theta} \right)^{\kappa_1-1 + \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j)} (1-\theta)^{\kappa_2 + \kappa_1 - 2 + \sum_{j=1}^N T_j} \prod_{j=1}^N \left[\frac{e^{-\frac{\alpha_{z_j}}{n} d(\mathbf{R}_j, \boldsymbol{\rho}_{z_j})}}{Z_n(\alpha_{z_j})} \right]. \quad (3.20)$$

The second step, is performed by using the density,

$$P(\mathbf{R}_{1:N} | \alpha_{1:C}, \boldsymbol{\rho}_{1:C}, \eta_{1:C}, \theta, z_{1:N}, \mathcal{B}_{1:N}) \propto \prod_{j=1}^N \frac{e^{-\frac{\alpha_{z_j}}{n} d(\mathbf{R}_j, \boldsymbol{\rho}_{z_j})}}{Z_n(\alpha_{z_j})} P(\mathcal{B}_j | \theta, \mathbf{R}_j), \quad (3.21)$$

where $P(\mathcal{B}_j|\theta, \mathbf{R}_j)$ is defined in eq. (3.5). Figure 3.3 shows the graphical representation of the hierarchical construction of the Mixture model for the Mallows parameters. The algorithm for this second mixture extension is Appendix 3.A.

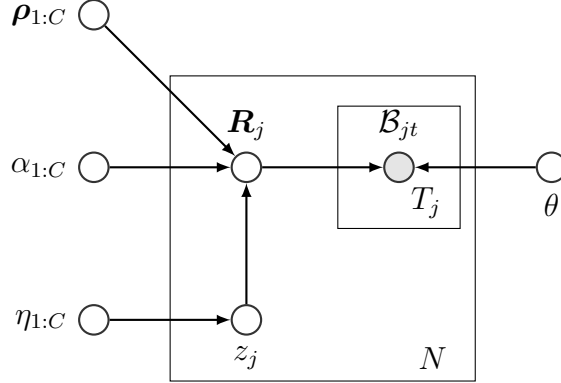


Figure 3.3: Graphical representation of the mixture model on (α, ρ) .

3.2 The MCMC algorithm for non-transitive pairwise preferences

We here outline the developed MCMC algorithm to sample from the posterior density of eq. (3.6). Details on the MCMC adaptations to the models on Sections 3.1.2-3.1.5, are provided in Appendix 3.A.

As mentioned in Section 3.1.1, the MCMC iterates between two main steps:

1. Update α, ρ and θ given $\mathcal{B}_{1:N}$ and $\mathbf{R}_{1:N}$, using eq. (3.7):
 - (a) Metropolis update of ρ
 - (b) Metropolis update of α
 - (c) Gibbs update of θ
2. Update $\mathbf{R}_{1:N}$ given α, ρ, θ and $\mathcal{B}_{1:N}$, using eq. (3.8).

In step 1(a), we propose a new consensus ranking ρ' according to a symmetric proposal which is centered around the current consensus ranking ρ .

Definition 3. *Swap proposal.* Denote the current version of the consensus ordering vector by $\mathbf{x} = (\rho)^{-1}$, which is the vector whose n components are the items in \mathcal{A} ordered from best to worst according to ρ , i.e., $x_i^t = A_k \iff \rho_k^t = i$ (see Section 1.1). Let $L^* \in \{1, \dots, n\}$.

Sample uniformly an integer l from $U\{1, 2, \dots, L^*\}$ and draw a random number u uniformly in $\{1, 2, \dots, n - l\}$. The proposal \mathbf{x}' has components

$$x'_i = \begin{cases} x_i & \text{if } i \neq \{u, u + l\} \\ x_{u+l} & \text{if } i = u \\ x_u & \text{if } i = u + l, \end{cases}$$

and the proposed ranking is $\boldsymbol{\rho}' = (\mathbf{x}')^{-1}$.

The parameter L^* plays the role of the maximum allowed distance between the ranks of the swapped items, and is used for tuning the acceptance probability in the Metropolis-Hastings step. The transition probability of the **Swap** proposal is symmetric, as

$$\begin{aligned} q(\boldsymbol{\rho}' \rightarrow \boldsymbol{\rho}) &= q(\boldsymbol{\rho} \rightarrow \boldsymbol{\rho}') = \sum_{l=1}^{L^*} P(L=l)P(\boldsymbol{\rho}' \rightarrow \boldsymbol{\rho}|L=l)\mathbb{1}(|\boldsymbol{\rho}' - \boldsymbol{\rho}| = 2l) = \\ &= \frac{1}{L^*} \sum_{l=1}^{L^*} \frac{1}{n-l} \mathbb{1}(|\boldsymbol{\rho}' - \boldsymbol{\rho}| = 2l). \end{aligned}$$

This is a very intuitive and simple way of exploring \mathcal{P}_n , but has appealing properties for us, that will become clear in step 2. In this step, alternative proposals could be considered, for example the L-S introduced in Chapter 2, Section 2.1.3.

Remark 1. The **Swap** proposal $\boldsymbol{\rho}'$ is a local perturbation of $\boldsymbol{\rho}$, separated from $\boldsymbol{\rho}$ by Cayley distance $d_C = 1$, by Hamming distance $d_H = 2$, expected Kendall distance $\mathbb{E}(d_K) = L^*$, expected footrule distance $\mathbb{E}(d_F) = L^* + 1$, and expected Spearman' distance $\mathbb{E}(d_S) = \frac{(L^*+1)(2L^*+1)}{3}$. This follows by the definitions of the various distances and by simple calculations.

The proposed ranking is then accepted with probability $\min\{1, a_\rho\}$, where

$$\ln a_\rho = -\frac{\alpha}{n} \sum_{j=1}^N [d(\mathbf{R}_j, \boldsymbol{\rho}') - d(\mathbf{R}_j, \boldsymbol{\rho})]. \quad (3.22)$$

Notice that (3.22) is very similar to (2.4), but does not include the correction accounting for the L-S asymmetric proposal.

In step 1(b) we propose α' from a log-normal density $\ln \mathcal{N}(\ln \alpha, \sigma_\alpha^2)$ and accept it with

probability equal to $\min\{1, a_\alpha\}$, where

$$\ln a_\alpha = \gamma \ln(\alpha'/\alpha) - \left[\lambda + \frac{1}{n} \sum_{j=1}^N d(\mathbf{R}_j, \boldsymbol{\rho}) \right] (\alpha' - \alpha) - N \ln[Z_n(\alpha')/Z_n(\alpha)]. \quad (3.23)$$

The previous acceptance probability takes into account the asymmetric transition probability of the chain, that results from the log-normal proposal, like in (2.6).

The partition function $Z_n(\alpha)$ can be either computed exactly or approximated by the Importance Sampling introduced in Chapter 2, depending on the distance function chosen and on the number n of items considered. In this chapter we always use footrule distance with $n \leq 50$, so that $Z_n(\alpha)$ is always to be intended as exact.

In step 1(c) we sample θ from the beta distribution, truncated to the interval $[0, 0.5)$, with updated hyper-parameters,

$$\kappa'_1 = \kappa_1 + \sum_{j=1}^N \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j), \quad \kappa'_2 = \kappa_2 + \sum_{j=1}^N \sum_{t=1}^{T_j} [1 - g(\mathcal{B}_{jt}, \mathbf{R}_j)]. \quad (3.24)$$

Step 2 is a Metropolis-Hastings for the individual rankings. We exploit, like in Chapter 2, the fact that fixing all other parameters and the data, $\mathbf{R}_1, \dots, \mathbf{R}_N$ are conditionally independent, and that each \mathbf{R}_j only depends on the corresponding set of pairwise comparisons, \mathcal{B}_j . We thus sample \mathbf{R}'_j from the Swap proposal, separately for each $j = 1, \dots, N$. The Swap proposal is here advantageous because it perturbs locally not only the current individual ranking \mathbf{R}_j , but also the function $g(\mathcal{B}_{jt}, \mathbf{R}_j)$.

Remark 2. *The Swap proposal always gives a proposed individual ranking $\mathbf{R}'_j \neq \mathbf{R}_j$. However, it may happen that $g(\mathcal{B}_{jt}, \mathbf{R}'_j) = g(\mathcal{B}_{jt}, \mathbf{R}_j)$, $\forall t = 1, \dots, T_j$.*

This is important for what concerns the acceptance probability of \mathbf{R}'_j . If $g(\mathcal{B}_{jt}, \mathbf{R}'_j) = g(\mathcal{B}_{jt}, \mathbf{R}_j)$, $\forall t = 1, \dots, T_j$, the acceptance probability depends only on the ratio of the Mallows kernels of \mathbf{R}'_j and \mathbf{R}_j , and is equal to $\min\{1, a_1\}$, where

$$\ln a_1 = -\frac{\alpha}{n} [d(\mathbf{R}'_j, \boldsymbol{\rho}) - d(\mathbf{R}_j, \boldsymbol{\rho})]. \quad (3.25)$$

If $g(\mathcal{B}_{jt}, \mathbf{R}'_j) \neq g(\mathcal{B}_{jt}, \mathbf{R}_j)$, for some $t = 1, \dots, T_j$, the acceptance probability depends

also on the mistakes model, and is equal to $\min\{1, a_2\}$ where

$$\ln a_2 = \ln a_1 + \sum_{t=1}^{T_j} [g(\mathcal{B}_{jt}, \mathbf{R}'_j) - g(\mathcal{B}_{jt}, \mathbf{R}_j)] \ln [\theta/(1 - \theta)]. \quad (3.26)$$

Example To illustrate this step of the algorithm, suppose that a user expresses the following set of preferences,

$$\mathcal{B}_j = \{A_2 \prec A_1, A_5 \prec A_4, A_5 \prec A_3, A_5 \prec A_2, A_5 \prec A_1, A_3 \prec A_2, A_1 \prec A_3\}.$$

This set contains the non-transitive pattern $A_2 \prec A_1 \prec A_3 \prec A_2$. For the illustration, suppose that the current value of the individual ranking vector is $\mathbf{R}_j = (5, 4, 3, 2, 1)$, which corresponds to the individual ordering vector $\mathbf{X}_j = (A_5, A_4, A_3, A_2, A_1)$, and for which $\sum_{t=1}^7 g(\mathcal{B}_{jt}, \mathbf{R}_j) = 1$, because the only pairwise preference in \mathcal{B}_j that contradicts \mathbf{R}_j is $A_1 \prec A_3$. If we sample the proposal $\mathbf{X}'_j = (A_5, A_3, A_4, A_2, A_1)$, this gives $g(\mathcal{B}_{jt}, \mathbf{R}'_j) = g(\mathcal{B}_{jt}, \mathbf{R}_j)$, $\forall t = 1, \dots, 7$, and $\mathbf{R}'_j = (5, 4, 2, 3, 1) \neq \mathbf{R}_j$. However, if we sample $\mathbf{X}'_j = (A_4, A_5, A_3, A_2, A_1)$, then $\mathbf{R}'_j = (5, 4, 3, 1, 2) \neq \mathbf{R}_j$ and also $\sum_{t=1}^7 g(\mathcal{B}_{jt}, \mathbf{R}'_j) = 2 \neq \sum_{t=1}^7 g(\mathcal{B}_{jt}, \mathbf{R}_j)$ since, also the preference $A_5 \prec A_4$ contradicts the sampled \mathbf{R}' .

Algorithm 7: MCMC Algorithm for the Bernoulli model for mistakes.

input : $\mathcal{B}_1, \dots, \mathcal{B}_N$; $\lambda, \gamma, \sigma_\alpha, L^*, \kappa_1, \kappa_2, d(\cdot, \cdot), Z_n(\alpha), M$.
output: Posterior distributions of $\rho, \alpha, \theta, \mathbf{R}_1, \dots, \mathbf{R}_N$.
Initialization of the MCMC: randomly generate $\rho_0, \alpha_0, \theta_0$ and $\mathbf{R}_1^0, \dots, \mathbf{R}_N^0$.

for $m \leftarrow 1$ **to** M **do**

M-H step: update ρ
sample: $\rho' \sim \text{Swap}(\rho_{m-1}, L^*)$ and $u \sim \mathcal{U}(0, 1)$
compute: $ratio \leftarrow$ eq. (3.22) with $\rho \leftarrow \rho_{m-1}, \alpha \leftarrow \alpha_{m-1}$, and $\mathbf{R}_{1:N} \leftarrow \mathbf{R}_{1:N}^{m-1}$
if $u < ratio$ **then** $\rho_m \leftarrow \rho'$
else $\rho_m \leftarrow \rho_{m-1}$

M-H step: update α
sample: $\alpha' \sim \ln \mathcal{N}(\ln \alpha_{m-1}, \sigma_\alpha^2)$ and $u \sim \mathcal{U}(0, 1)$
compute: $ratio \leftarrow$ eq. (3.23) with $\rho \leftarrow \rho_m, \alpha \leftarrow \alpha_{m-1}$, and $\mathbf{R}_{1:N} \leftarrow \mathbf{R}_{1:N}^{m-1}$
if $u < ratio$ **then** $\alpha_m \leftarrow \alpha'$
else $\alpha_m \leftarrow \alpha_{m-1}$

Gibbs step: update θ
compute: κ'_1 and κ'_2 from eq. (3.24) with $\mathbf{R}_{1:N} \leftarrow \mathbf{R}_{1:N}^{m-1}$, and sample: $\theta' \sim \text{Be}(\kappa'_1, \kappa'_2)$ truncated to the interval $[0, 0.5]$

M-H step: update $\mathbf{R}_1, \dots, \mathbf{R}_N$
for $j \leftarrow 1$ **to** N **do**
sample: $\mathbf{R}'_j \sim \text{Swap}(\mathbf{R}_j^{m-1}, L^*)$ and $u \sim \mathcal{U}(0, 1)$
if $\sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}'_j) = \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j^{m-1})$ **then** compute: $ratio \leftarrow$ eq. (3.25) with $\rho \leftarrow \rho_m$ and $\alpha \leftarrow \alpha_m$
else compute: $ratio \leftarrow$ eq. (3.26) with $\rho \leftarrow \rho_m, \alpha \leftarrow \alpha_m$ and $\theta \leftarrow \theta_m$
if $u < ratio$ **then** $\mathbf{R}_j^m \leftarrow \mathbf{R}'_j$
else $\mathbf{R}_j^m \leftarrow \mathbf{R}_j^{m-1}$
end

end

The pseudo-code of the MCMC for Bernoulli mistakes is here reported as Algorithm 7.

Appropriate convergence of the MCMC must in practice be checked by inspecting the trace

plots of the parameters, and by monitoring, for example, the integrated autocorrelation, like we did in Chapter 2. In Appendix 3.A we explain in detail how the algorithm is adapted to the logistic mistake model and to the mixture extensions, and report the pseudo-codes of the Algorithms.

3.3 Simulation study

Several simulation experiments were carried out to assess the performance of the methodology introduced in this chapter. An important aspect in the design of these experiments was finding appropriate values for the model parameters. The parameter α_{true} controls the concentration of the individual latent rankings $\mathbf{R}_{1,\text{true}}, \dots, \mathbf{R}_{N,\text{true}}$ around the true consensus $\boldsymbol{\rho}_{\text{true}}$: the larger α_{true} is, the more concentrated the individual rankings are. To give an idea of this effect, we plot in Figure 3.1, for a range of different α_{true} values, average distances $\frac{1}{N} \sum_{j=1}^N d(\mathbf{R}_{j,\text{true}}, \boldsymbol{\rho}_{\text{true}})$ obtained when $\mathbf{R}_{j,\text{true}} \sim \mathcal{M}(\alpha_{\text{true}}, \boldsymbol{\rho}_{\text{true}})$, with $N = 100$ and $n = 10$ (left), $n = 30$ (right). For each considered α_{true} , each boxplot in the Figures is computed from a set of 100 simulated samples.

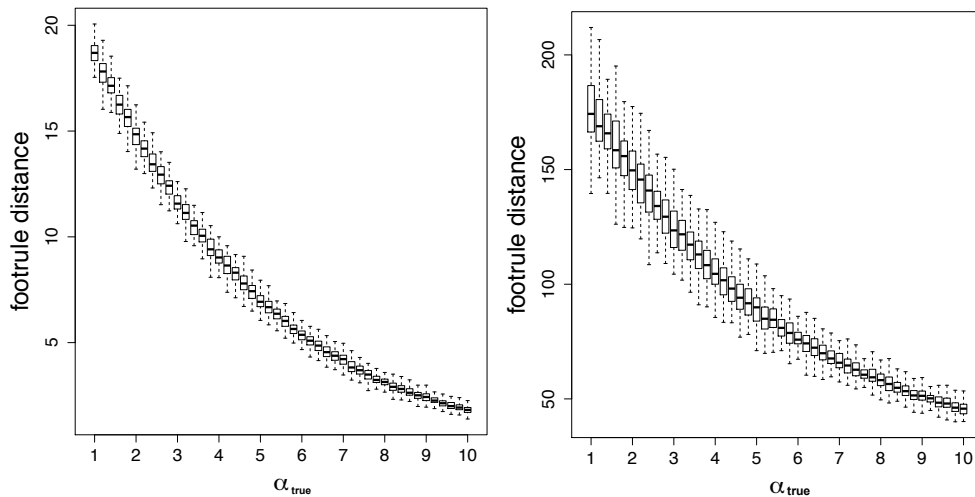


Figure 3.1: Boxplots of the average footrule distance, $\frac{1}{N} \sum_{j=1}^N d(\mathbf{R}_{j,\text{true}}, \boldsymbol{\rho}_{\text{true}})$, for $N = 100$, and for individual rankings, $\mathbf{R}_{1:N,\text{true}}$, generated from the Mallows model, $\mathcal{M}(\alpha_{\text{true}}, \boldsymbol{\rho}_{\text{true}})$, for increasing values of α_{true} . Left: $n = 10$; Right: $n = 30$.

With the number of items growing, identifying a consensus ranking becomes increasingly hard due to the $n!$ possible permutations. To balance this, N , the number of users, and λ_T , the average number of pairs compared by each user, must be chosen accordingly. For instance, if $n = 10$, the maximal number of pairs that can be formed is

$T_{\max} = \binom{n}{2} = 45$, while for $n = 20$, $T_{\max} = 190$. Choosing $\lambda_T = 30$ would in the case of $n = 10$ correspond to providing the individual users, on average, with the proportion $\frac{\lambda_T}{T_{\max}} = \frac{30}{45} \simeq 0.67$ of all possible pair comparisons, while for $n = 20$ the corresponding proportion is only $\frac{\lambda_T}{T_{\max}} = \frac{30}{190} \simeq 0.16$.

One should also account for the effect of the parameter θ (or the logistic parameters β_0 and β_1) controlling the level of noise in the data, in the form of mistakes in reporting individual pairwise comparisons. While larger values of N will generally facilitate the estimation of the consensus ranking, larger θ will render individual ranking estimates much less reliable.

3.3.1 Simulations with Bernoulli mistake model

The aim of the experiments was to validate the method and to evaluate its performance in some test situations. The data were simulated from the Mallows model with Bernoulli mistakes, varying parameters θ , α , n , N , and T_j , $j = 1, \dots, N$, while always using the footrule distance. The number of items n was always kept below 50, thus enabling us to use exact values for the partition function, see Section 2.2.1. For a more detailed description of the data generation used in this section, see Appendix 3.B.

Various point estimates can be deduced from the posterior distribution of $\boldsymbol{\rho}$, one being the maximum a posteriori (MAP). In this section we always choose the cumulative probability (CP) consensus ordering, defined in Chapter 2, Section 2.2.3.

In order to assess the performance of our methods, in Figure 3.2 we plot the posterior distribution of the normalized footrule distance (denoted by the apex ‘ n ’) of the estimated consensus $\boldsymbol{\rho}$ and the true consensus, $d_F^n(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}}) = \frac{1}{n} \sum_{i=1}^n |\rho_i - \rho_{i,\text{true}}|$, for varying parameters α , θ , λ_T and N , while keeping fixed $n = 10$.

As expected, the performance of the method improves as the number of users N increases (Figure 3.2, top-left), as the probability of making mistakes θ decreases (Figure 3.2, top-right), as the dispersion of the individual latent rankings $\boldsymbol{R}_{j,\text{true}}$ around $\boldsymbol{\rho}_{\text{true}}$ decreases, that is when α increases, (Figure 3.2, bottom-left), and when the average number of pairwise comparisons λ_T becomes larger, (Figure 3.2, bottom-right). Interestingly, in the last case, the method performs generally well also if the average number of pairs is $\lambda_T = 15$, being only 1/3 of the maximal number of pairs possible.

In Figure 3.3 we plot the posterior distribution of $d_F^n(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$, corresponding to simulation experiments with a larger number of items, $n \in \{15, 25\}$, for increasing N . Note that

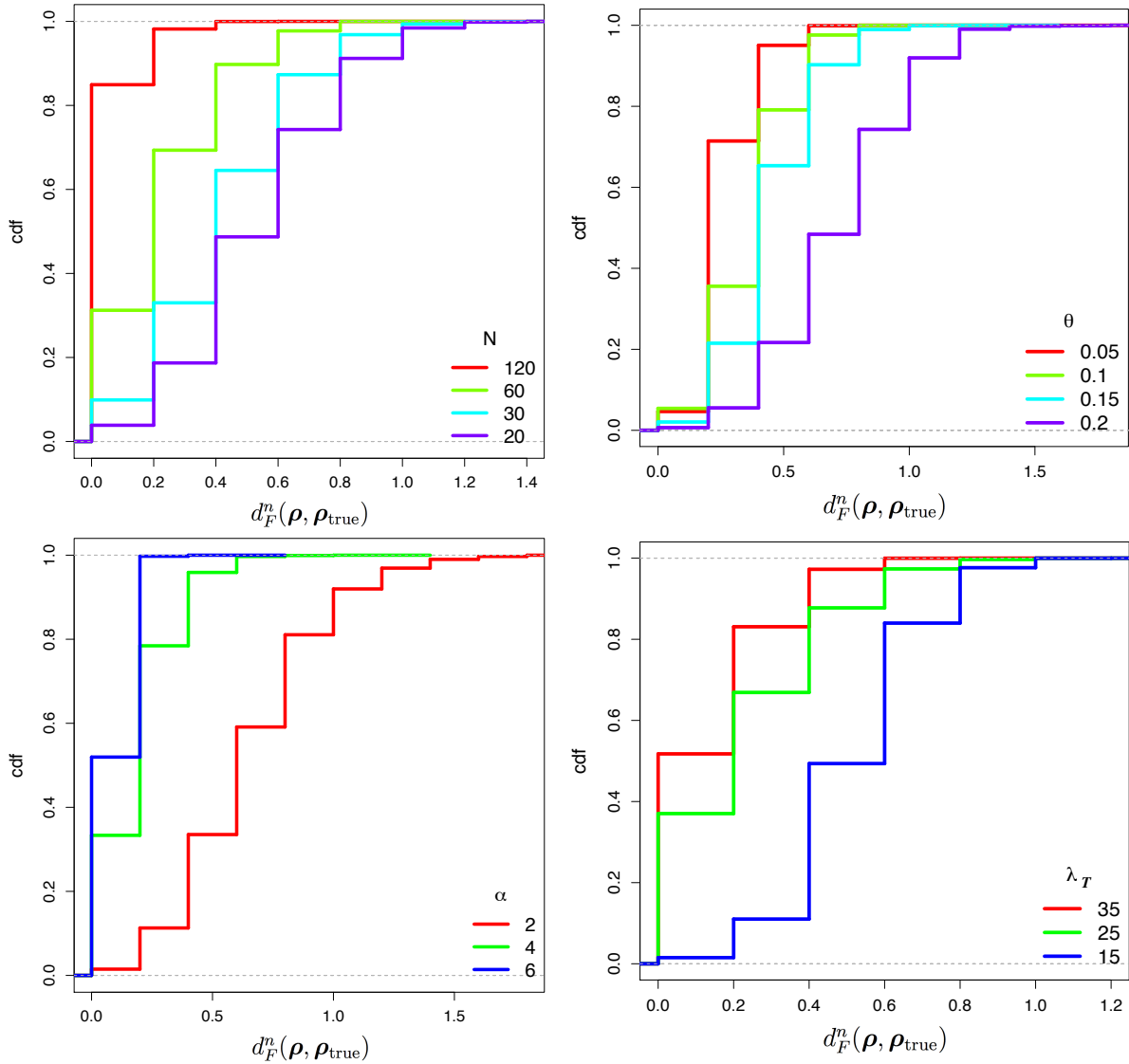


Figure 3.2: Results of the simulated data. Posterior CDFs of $d_F^n(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ as a function of N for $\alpha = 3$, $\lambda_T = 25$, $\theta = 0.1$ (top-left); as a function of θ for $\alpha = 3$, $N = 40$, $\lambda_T = 25$ (top-right); as a function of α , for $\theta = 0.1$, $N = 40$, $\lambda_T = 25$ (bottom-left); as a function of λ_T for $\alpha = 3$, $N = 40$, $\theta = 0.1$ (bottom-right).

the number of pairs assessed by each user in the case $n = 25$ is around 50, which is only 1/6 of all the possible pairs.

Next, we studied the performance of the method in terms of the precision of the individual ranking estimation. We quantify the results by the probability of getting at least 3 items right, among the top-5, defined as follows. For each user $j = 1, \dots, N$, we found the triplet of items $D_3^j = \{A_{i_1}, A_{i_2}, A_{i_3}\}$ that had maximum posterior probability of being ranked jointly among the top-3 items, that is the triplet that maximized

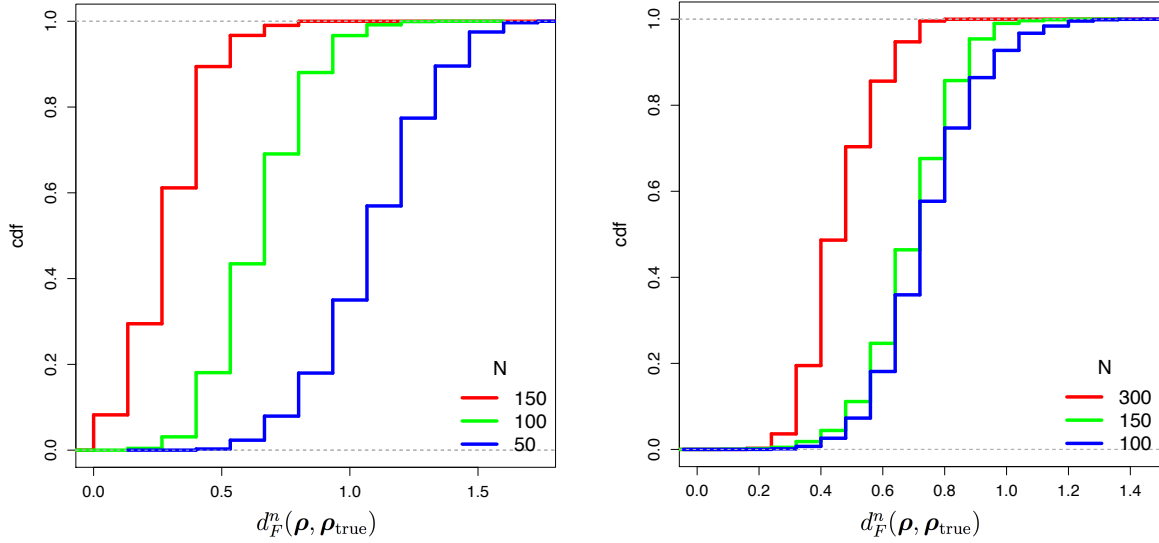


Figure 3.3: Results of the simulated data. Posterior CDFs of $d_F^n(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ as a function of N , for $\theta = 0.1$, $\alpha = 3.5$, $\lambda_T = 25$, $n = 15$ (left), and for $\theta = 0.1$, $\alpha = 4.5$, $\lambda_T = 50$, $n = 25$ (right).

$\sum_{\sigma \in \mathcal{P}_3} P(\{R_{ji_1}, R_{ji_2}, R_{ji_3}\} = \sigma \mid \text{data})$, where σ denotes a permutation of the set $\{1, 2, 3\}$. This posterior quantity is estimated along the MCMC trajectory. We defined H_5^j to be the set of 5 highest ranked items in $\mathbf{R}_{j, \text{true}}$, for each user j . We then checked whether $D_3^j \subset H_5^j$ (that is the top-3 estimated items are all among the top-5 of each user). The percentages of users for which this is true is reported in Table 3.1.

We notice that the results are overall very good: when n is set to 10 (first 4 sub-tables from the left in Table 3.1), we consistently learn 3 out of the top-5 items in more than 70% of the users (with a peak of 100%). Also in the more difficult cases of $n = 15$ and $n = 25$ (first 2 sub-tables from the right in Table 3.1) the results are very good, especially considering that this percentage does not include the cases where only 2 (or 1) items where correctly estimated in the top positions.

N	%	θ	%	α	%	λ_M	%	N	%	N	%
20	88	0.05	92.5	2	82.5	15	85	50	65	100	44
30	83	0.1	87.5	4	95	25	97.5	100	58	150	46
60	83	0.15	75	6	92.5	35	100	150	60	300	45
120	75	0.2	72.5								

Table 3.1: Results of the simulated data. Percentage of users for which the estimated top-3 items belong to the true top-5. Data corresponding to simulations with parameter settings of Figures 3.2 and 3.3: from left to right, same parameters as in Figure 3.2 top-left, top-right, bottom-left, bottom-right; Figure 3.3, left, right.

We then chose one of the simulated data cases and computed the posterior probabili-

ties of correctly predicting the preference order of all pairs not assessed by the users, i.e. $P[g(\mathcal{B}_{j,\text{new}}, \mathbf{R}_j) = g(\mathcal{B}_{j,\text{new}}, \mathbf{R}_{j,\text{true}}) \mid \text{data}]$. Figure 3.4 shows the boxplots for these predictive probabilities, (left) stratified according to the number of pairs each user assessed in the data, and (right) stratified according to the footrule distance between the true individual ranking $\mathbf{R}_{j,\text{true}}$ and the true consensus $\boldsymbol{\rho}_{\text{true}}$, $d_F(\boldsymbol{\rho}_{\text{true}}, \mathbf{R}_{j,\text{true}}) = \sum_{i=1}^n |\rho_{i,\text{true}} - R_{ji,\text{true}}|$.

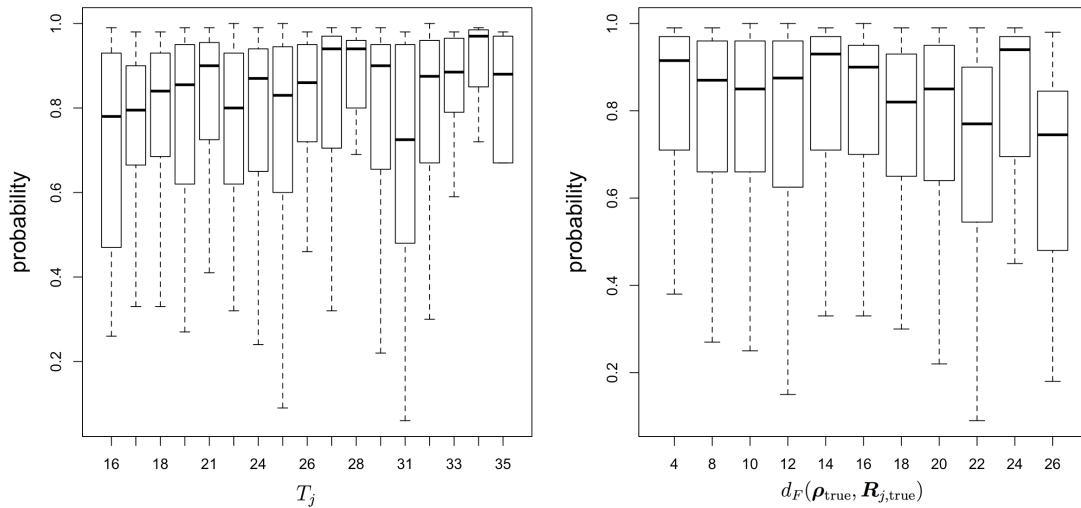


Figure 3.4: Results of the simulated data. Posterior probabilities of correctly predicting the preference order of all pairs not assessed by the users, (left) stratified according to the number of pairs each user assessed in the data, and (right) stratified according to $d_F(\boldsymbol{\rho}_{\text{true}}, \mathbf{R}_{j,\text{true}})$.

In the case considered, the model had a very good predictive power, especially considering that the simulated data had many mistakes (around 10%). We also notice a slight increase of the predictive probabilities as T_j increases (left panel) and as $d(\boldsymbol{\rho}_{\text{true}}, \mathbf{R}_{j,\text{true}})$ decreases (right panel). These results are not surprising: it is easier to predict correct orderings of new pairs when (i) the user assesses more pairs, and (ii) the user's own ranking resembles more the shared consensus.

As a final check, we applied the logistic MCMC on these BM generated data. The results were very similar to those obtained above and the posterior distribution of β_1 was highly concentrated around 0, which is consistent with the fact that, at $\beta_1 = 0$, LM collapses to BM.

3.3.2 Simulations with logistic mistake model

In this section we show results obtained from experiments on simulated data generated from the logistic model for mistakes of Section 3.1.2. The procedure is similar to the one described in the previous section, as well as the data generation procedure (see Appendix

3.B). We varied the parameters N , β_0 , β_1 , α , and λ_T , while, as in Section 3.3.1, we fixed the true consensus ranking $\boldsymbol{\rho}_{\text{true}}$ and the footrule distance. We then analyzed the generated datasets by applying both the logistic (LM) and the Bernoulli (BM) models.

No systematic differences in the results could be detected in the accuracy of the consensus ranking estimate, evaluated in terms of the posterior CDF of $d_F^n(\boldsymbol{\rho}, \boldsymbol{\rho}_{\text{true}})$ (not shown). We then studied whether clear differences could be found when comparing the estimates of the individual ranking vectors to the corresponding true values. For this, we inspected the performance of the LM model by using two sets of simulations where $n = 10$, $N = 100$, $T_j = 25 \forall j$, $\alpha = 2.5$, but with different settings of logistic parameters:

S1: We varied β_0 while keeping constant $\beta_1 = 5$, with the nested procedure explained in Appendix 3.B;

S2: We varied the parameters β_0 and β_1 together, keeping the probability of making a mistake (averaged across the simulated distances) constant to ca. 0.1.

In Figure 3.5 we plot the theoretical values of the logistic mistake probability, $P[g(\mathcal{B}_{jt}, \mathbf{R}_j) = 1 \mid \mathbf{R}_j, \beta_0, \beta_1]$, as a function of the distance $d_{\mathbf{R}_j, t}$ between the items compared, when varying β_0 and β_1 according to the schemes S1 (left) and S2 (right).

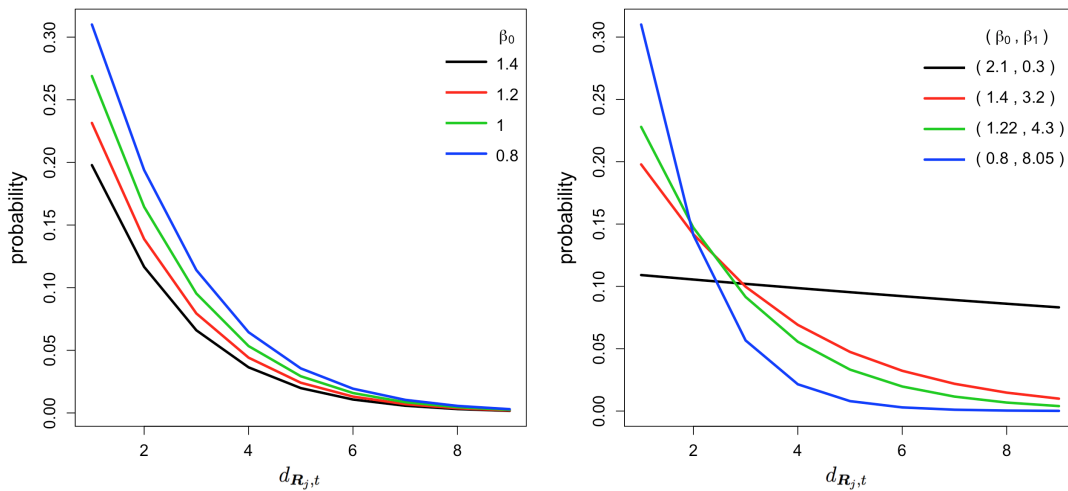


Figure 3.5: Logistic theoretical probabilities of making a mistake as a function of the distance between the items compared, for the two sets of simulations, S1 left, S2 right.

In S1, where β_1 is fixed, the smaller the value of β_0 , the more likely are mistakes in the data. With scheme S2 instead, we want to inspect whether the performance of LM changes as the dependence on the distance $d_{\mathbf{R}_j, t}$ becomes stronger. For this, we used the same measure as in Table 3.1 of Section 3.3.1, namely the percentage of users for which

the estimated top-3 items belong to the true top-5. In Table 3.2, we report the results for all the simulated datasets, when estimated with the BM and the LM algorithms.

Simulation	Model	%	Simulation	Model	%
S1: $\beta_0 = 1.4$	LM	94%	S2: black line	LM	91%
	BM	90%		BM	92%
S1: $\beta_0 = 1.2$	LM	92%	S2: red line	LM	89%
	BM	92%		BM	87%
S1: $\beta_0 = 1$	LM	93%	S2: green line	LM	92%
	BM	91%		BM	91%
S1: $\beta_0 = 0.8$	LM	88%	S2: blue line	LM	91%
	BM	85%		BM	91%

Table 3.2: Percentage of users for which the estimated top-3 items belong to the true top-5. Simulations with parameter settings of Figure 3.5.

As expected, the performance of both models deteriorates as β_0 decreases when β_1 is fixed (left panel in Table 3.2, corresponding to simulations in Figure 3.5 left). Somewhat surprisingly, BM and LM perform in very similar ways (right panel in Table 3.2, corresponding to simulations parameters in Figure 3.5 right), and this remains true when the dependence in LM on the distance becomes stronger.

We then computed the posterior probabilities of correctly predicting the preference order of all pairs not assessed by the users. Figure 3.6 shows the boxplots for these predictive probabilities, stratified according to the true distance between the items.

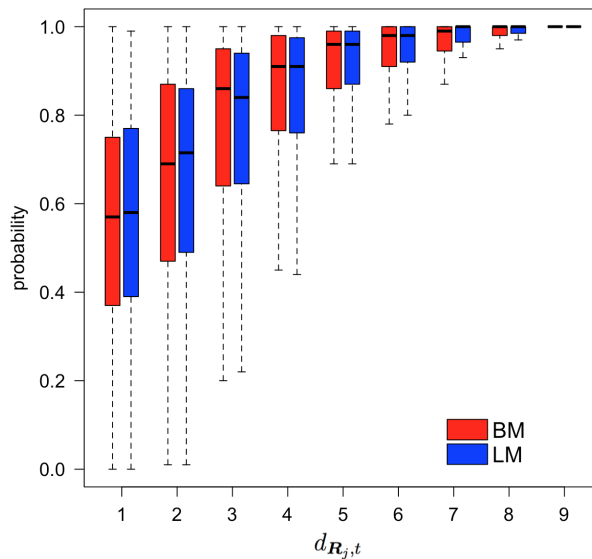


Figure 3.6: Box-plots of the posterior probability for correctly predicting the missing preferences stratified by $d_{R_{j,t}}$, the distance between the items in $R_{j,true}$. Simulation S2, red line in Figure 3.5 right.

In this comparison, both models had a very good predictive power, especially considering that the simulated data had many mistakes (around 10%) and the assessments provided by different users were quite variable ($\alpha = 2.5$). In many instances, the more general LM model appears to have had a slight edge over the simpler BM, but this was not true always, and overall, both methods produced very similar estimates of the individual rank vectors. The similarity of their performances may be because the transitivity property required in constructing versions of complete rankings \mathbf{R}_j is so strong that the precise form of the error model no longer has a major impact on the results.

3.3.3 Ability to detect mistakes

One way to measure the performance of our procedure is to study its ability in terms of detection of the mistakes made by the users.

When we simulate the data, we know which preferences were mistakes, and which ones were not. We can then look at the sensitivity and specificity of the results. The sensitivity is the proportion of positives (i.e. mistakes) that are correctly identified as positives (also known as true positive rate, TPR). The specificity is the proportion of negatives that are correctly identified as negatives (also equal to 1-FPR, where FPR denotes the false positive rate). In particular, we can compute, for each pairwise comparison assessed in the data \mathcal{B}_{jt} , the posterior probability that it is identified as a mistake, given that it was a mistake (i.e. the probability that \mathcal{B}_{jt} is a true positive)

$$P [g(\mathcal{B}_{jt}, \mathbf{R}_j) = 1 \mid \text{data}, g(\mathcal{B}_{jt}, \mathbf{R}_{j,\text{true}}) = 1] . \quad (3.27)$$

In Figure 3.7 we show the results from data simulated from the BM with $\theta = 0.2$, $\alpha = 3$ and $n = 10$. It is reported the posterior probability, for each comparison (columns) and for each user (rows), to estimate the preference as in the ground truth. In each cell, it is therefore reported the posterior of equation (3.27) if the comparison was a mistake in the data (represented as a white 1 on the cell) and

$$P [g(\mathcal{B}_{jt}, \mathbf{R}_j) = 0 \mid \text{data}, g(\mathcal{B}_{jt}, \mathbf{R}_{j,\text{true}}) = 0] \quad (3.28)$$

if was not a mistake in the data (represented as a white 0 on the cell). Eq. (3.28) represents the probability that \mathcal{B}_{jt} is a true negative.

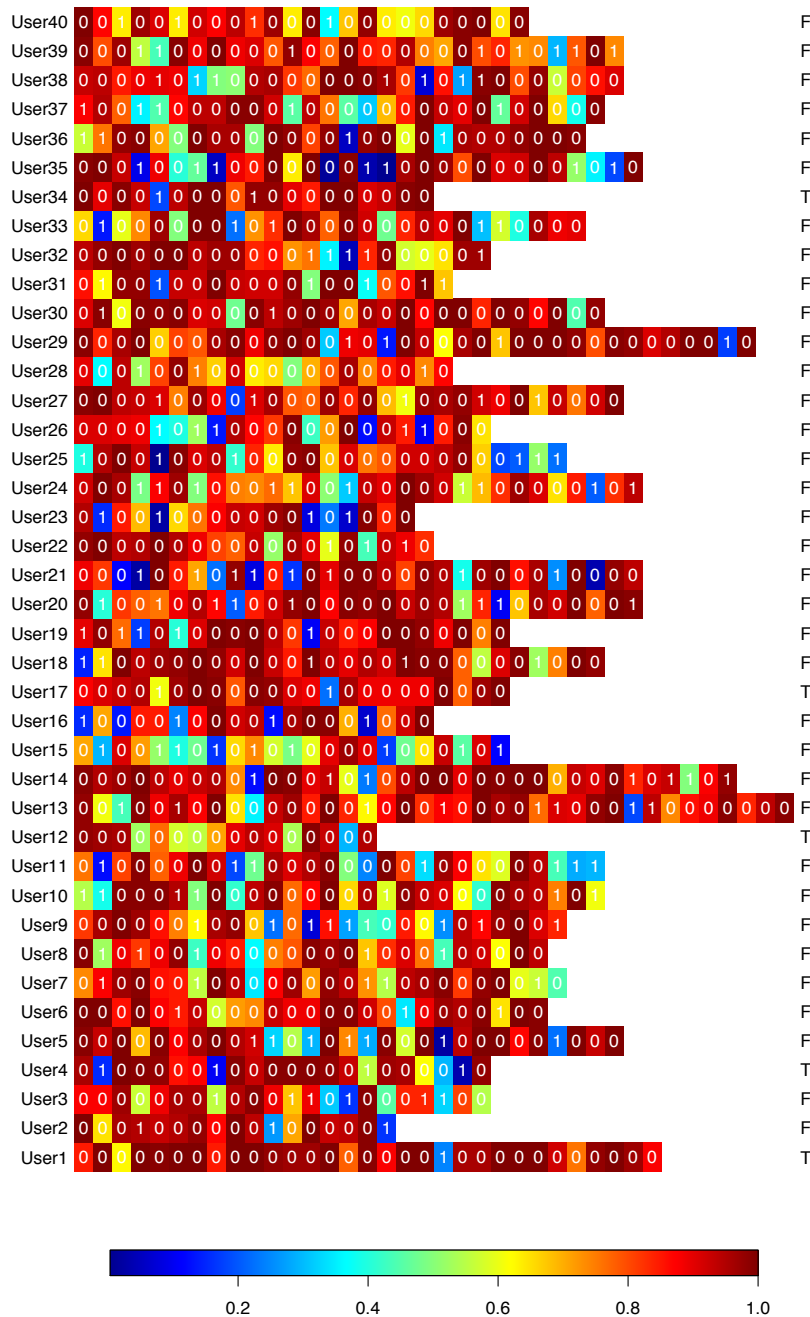


Figure 3.7: Posterior probability of estimating the preferences as in the ground truth. Cells labeled with a white ‘1’, show the posterior probability of equation (3.27), cells labeled with a white ‘0’, show the posterior probability of equation (3.28). The column on the right side of the table indicates if the original set of data given by user j was transitive (T) or non-transitive (F). Data simulated with $\theta = 0.2$, $\alpha = 3$ and $n = 10$.

In case of perfect classification, that is when the procedure correctly identifies all the pair preferences expressed by the users, all the cells would be red. A blue cell represents a mis-classification. On the one hand, a blue cell labelled with a ‘1’, means that the user’s

choice was a mistake, but the algorithm fails to identify it as a mistake (a false negative). On the other hand, a blue cell labelled with a ‘0’ means that the user’s choice was correct, but the algorithm erroneously classify it as a mistake (a false positive).

The column on the right side of the Figure indicates if the original set of data given by user j was transitive (T) or non-transitive (F). This latter information is useful to understand the different behavior of the model when the set of preferences is not transitive. Indeed, in such a case the algorithm is forced to detect at least one mistake in the individual data; this is because of the augmentation scheme, where is proposed an individual ranking that embeds only transitive preferences. If the set of preferences is transitive, instead, there is no need to change any pair preference, but, at the same time, it is allowed. Interestingly, we notice that the algorithm barely corrects preferences from transitive users: looking at the rows relative to users 1, 4, 12, 17, 34, we see that most of the mistakes (labelled with a white 1 on the cell), are blue, thus not identified as mistakes. This can be due to the fact that the algorithm seeks to minimize the number of mistakes in the data, and mostly corrects those necessary to make the sets of preferences transitive. From Figure 3.7 it is indeed clear that the majority of mis-classified cells are labeled ‘1’, which means that the majority of mis-classified cells are false negatives. This result is encouraging, because it is more important not to change the ordering of a preference that was correct, rather than failing to identifying a mistake in the data. The reason for this will be clear in Section 3.4.1, where, in analyzing a real dataset, our main concern regards identifiability.

We then investigate more deeply the intuition explained above, with a first series of simulated data, where we vary the θ parameter controlling the average number of mistakes in the BM generated data. We here quantify the performance of the method by considering ROC-curves. The ROC curve is drawn by considering different thresholds u , $0 < u < 1$, and then classifying for each u a pair as positive if the posterior probability $P[g(\mathcal{B}_{jt}, \mathbf{R}_j) = 1 \mid \text{data}] > u$. The curve is then formed by plotting the TPR versus FPR, for varying u .

In Figure 3.8 (left), we draw the ROC curves corresponding to data simulated with BM, for increasing (and nested) θ , and fixed $n = 10$, $N = 50$, $\alpha = 3.5$, $\lambda_T = 25$, while in Figure 3.8 (right), the ROC curves correspond to the data simulated with BM for increasing α , already considered in Section 3.3.1.

In Figure 3.9, we draw the ROC curves corresponding to the simulated data of Section

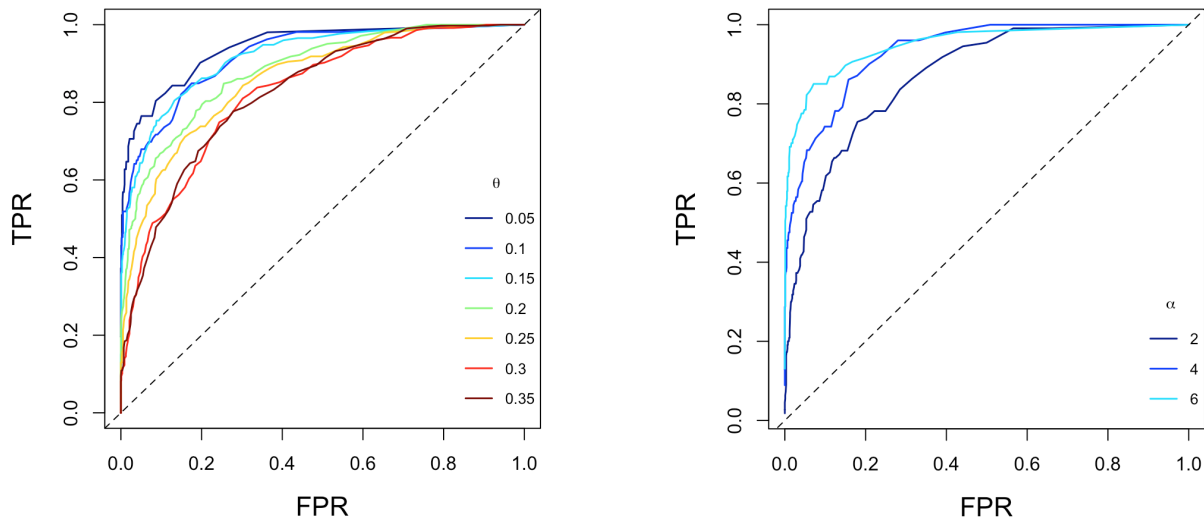


Figure 3.8: BM Simulated data. Left: ROC curve for the binary classification for increasing θ . $n = 10$, $N = 50$, $\alpha = 3.5$, $\lambda_T = 25$; Right: ROC curve for the binary classification for increasing α . Simulated data of Section 3.3.1, Figure 3.2 bottom-left.

3.3.2, Figure 3.5 left, for increasing β_0 . On the left are shown the results obtained by LM estimation, on the right by BM estimation.

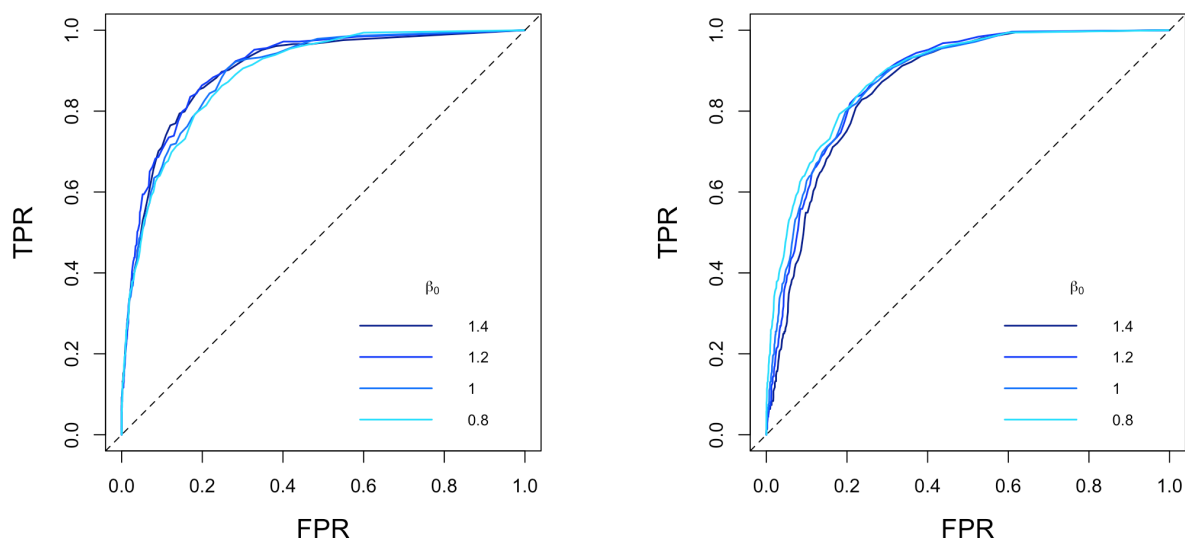


Figure 3.9: LM Simulated data. Left: ROC curve for simulated data with increasing β_0 estimated with LM. Right: ROC curve for the same data estimated with BM.

Since the differences are tiny as expected, in order to inspect whether there is any difference we also report the results of the AUC statistics, in Table 3.3.

We see that, in terms of the AUC statistics, the results obtained with the LM are slightly better than those obtained by BM.

β_0	LM	BM
1.4	0.90	0.88
1.2	0.91	0.89
1	0.90	0.87
0.8	0.89	0.88

Table 3.3: Simulated data. AUC statistics.

3.4 Examples

3.4.1 Beach preference data revisited

We here analyze the same dataset introduced in Section 2.4.2, but considering the non-transitive patterns of the data.

As already mentioned earlier, in our data 9 out of 60 (the 15%) users¹ returned queries which contained at least one non-transitive pattern of comparisons. In Section 2.4.2 we dropped such patterns, and analyzed the remaining transitive data with the model of Section 2.3.2. Here, instead, we account for the non-transitivities and analyze the data with the Bernoulli and logistic models for mistakes for homogeneous users. We run both algorithms with 10^5 iterations (computing time was 10' each), and discarded the first $2 \cdot 10^4$ iterations of each as burn-in. For the partition function, we used the exact $Z_n(\alpha)$, since $n = 15$. After some tuning, we set $L^* = 1$, $\sigma_\alpha = 0.3$, $\lambda = 0.1$, $\gamma = 2$, in both algorithms and $\sigma_{\beta_0} = \sigma_{\beta_1} = 0.4$, in the logistic. The posterior means of the parameters of interest, when using the Bernoulli (BM) and the Logistic (LM) models were $\mathbb{E}_{\text{BM}}(\alpha | \text{data}) = 5.15$ (4.53, 5.78), $\mathbb{E}_{\text{BM}}(\theta | \text{data}) = 0.023$ (0.013, 0.035), $\mathbb{E}_{\text{LM}}(\alpha | \text{data}) = 5.25$ (4.62, 5.9), $\mathbb{E}_{\text{LM}}(\beta_0 | \text{data}) = 2.29$ (1.29, 3.26), $\mathbb{E}_{\text{LM}}(\beta_1 | \text{data}) = 4.43$ (1, 8.47).

In Table 3.1 we report the CP consensus lists of the beaches obtained with the two procedures (columns 2 and 5), along with the cumulative probability for each beach of being ranked in that position or higher, $P(\rho_i < i)$ (columns 3 and 6), and the corresponding 95% HPD interval for each (columns 4 and 7). We see that the two procedures converge to two slightly different consensus orderings: the pairs B4 and B8, and B12-B14 are inverted. The uncertainty on the exact ordering of these beaches pairs is indeed indicated by both $P(\rho_i < i)$ and the corresponding HPDIs. We also notice that the resulting list is very similar to the one obtained without the model for mistakes in Table 2.8. The reason for this is that these data have very few non-transitive patterns, which is also indicated

¹The users labelled 2, 5, 6, 10, 17, 20, 28, 42, 59.

by the posterior means of the parameters governing the number mistakes in the data in BM and LM (θ for BM and β_0, β_1 for LM).

ρ	CP_{LM}	$P(\rho_i < i)$	95% HPDI	CP_{BM}	$P(\rho_i < i)$	95% HPDI
1	B9	0.87	(1,2)	B9	0.91	(1,2)
2	B6	1	(1,2)	B6	1	(1,2)
3	B3	0.85	(3,4)	B3	0.85	(3,4)
4	B11	0.98	(3,4)	B11	0.98	(3,4)
5	B15	0.96	(4,5)	B15	0.98	(5,5)
6	B10	0.93	(6,7)	B10	0.85	(6,7)
7	B1	1	(6,7)	B1	1	(6,7)
8	B7	0.55	(8,10)	B7	0.58	(8,10)
9	B5	0.87	(8,10)	B5	0.8	(8,10)
10	B13	1	(8,10)	B13	1	(8,10)
11	B4	0.6	(11,13)	B8	0.49	(11,13)
12	B8	0.55	(11,14)	B4	0.78	(11,14)
13	B14	0.8	(11,14)	B12	0.58	(12,14)
14	B12	0.98	(12,14)	B14	0.99	(12,14)
15	B2	1	(15,15)	B2	1	(15,15)

Table 3.1: Beaches arranged by the estimated CP consensus ranking. Left columns: logistic procedure; Right columns: Bernoulli procedure.

Therefore, with few non-transitivities in the data, it appears to be non crucial, at least for the estimation of the consensus ordering, the use of the more complicated model for mistakes introduced in this chapter. On the other hand this model allows not to drop the non-transitive pairs of the data, which is, in our opinion, the best procedure when approaching a dataset. Indeed, it may happen that the information contained is of interest, which could change markedly the results of the analysis.

We then repeated the same analysis on the latent full rankings of each assessor of Section 2.4.2, and report the results obtained with LM (similar results, not shown, were obtained with BM). Figure 3.1 is the corresponding of Figure 2.15. The main difference is that in 3.1 the estimated users' individual rankings resemble more the consensus. In particular, in 2.15 some of the users' estimated top-3 beaches appear in the bottom positions of the consensus ranking (the users labelled 16, 18, 27, 36, 45, 48, 50). In 3.1 instead, there is not such effect (with the exception of user 50).

This result indicates that the model for mistakes has indeed an effect on the analysis, and that, as we expected, this effect is mainly related with the estimation of the individual latent rankings. It is however important to comment on a possible identification problem. The homogeneity of the users, clear from Figure 3.1, follows from the fact that LM has

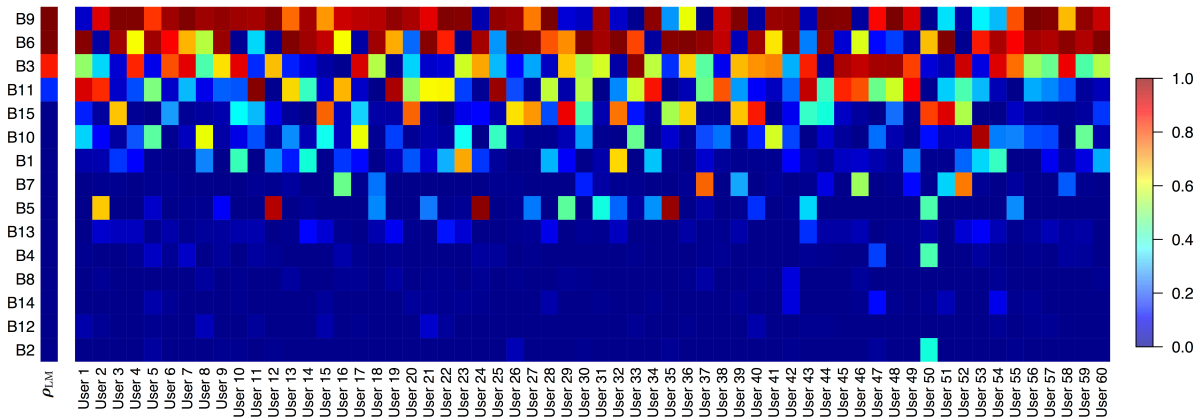


Figure 3.1: Posterior probability, for each beach, of being ranked among the top-3 in ρ_{LM} (column 1), and in \mathbf{R}_j , $j = 1, \dots, 60$ (next columns). Results from the LM.

the effect of correcting the mistakes by the users. However, since the model for mistakes corrects the individual preferences also when they are transitive (indeed none of the above mentioned users produced non-transitivities in the reported data), this effect could be due to two mechanisms:

- (i) The users were indeed more homogeneous than what seemed by looking at Figure 2.15, but made mistakes that we could not notice by looking at the data (because their sets of preferences were transitive, that is they managed to stay consistent with themselves);
- (ii) The LM model corrects pairs which were not mistakes, because of the relevance of the borrowing strength effect.

We studied this possible mis-identification in Section 3.3.3, where we inspected the ability to detect mistakes on simulated data. In particular, the false positive rate was always very low, thus reassuring about this matter.

Since we are here working with real data, that are not necessarily generated by our model, we perform an additional check, to test for the mis-identification concern. We studied how well the different procedures (namely the model without accounting for mistakes, here denoted by NoM, the BM and the LM) are able to recover the true top-3 beaches for each respondent. Indeed, in the questionnaire we also asked, as a final question after all the pairwise comparisons, to choose the top-3 beaches among all the images, and to order them from the most preferred to the least preferred. Like in the previous sections, we denote these sets by H_3^j , $j = 1, \dots, 53$, which are available for only 53 (out of 60) users, who answered the final question.

We then found the triplet of beaches that had maximum posterior probability of being ranked jointly among the top-3, and denote it with $D_3^j = \{Bi_1, Bi_2, Bi_3\}$ (for details on the calculations see Section 3.1.1). Finally, we checked how many (0-3 out of 3) of the beaches in D_3^j , $j = 1, \dots, 53$, were correctly identified in H_3^j by the 3 procedures, and give the results in Table 3.2, where we report the number of assessors for which we correctly identify the number of beaches indicated in the corresponding column by the three procedures (rows).

	0	1	2	3
LM	-	10	31	12
BM	-	11	33	9
NoM	6	10	31	6

Table 3.2: Number of assessors for which we correctly identify the number of beaches indicated in the corresponding column in their top-3, by the three procedures (rows).

It is clear that allowing for a model for mistakes (BM and LM rows) outperforms not considering them (NoM row). We also notice that, for these data, LM performs better than BM. This suggests that, in this survey, the more realistic logistic model for mistakes is the most appropriate.

3.4.2 Movie survey

In this section we consider data collected through an *ad hoc* survey created as follows. We selected the 15 highest grossing movies worldwide ($n = 15$), adjusted for ticket inflation price², and then assigned to each assessor j , randomly, $T_j = 30$ pairs of movies to be compared. The question put to the assessors, for each pair, was: “Which of the following two movies has so far brought in more box-office revenues (when adjusted for inflation)?”. We randomized the pairs of movies both within the same survey and across the surveys: every user was asked to answer to a different randomized collection of pairwise comparisons. Notice that the maximum number of pairwise comparisons out of $n = 15$ movies is $n(n - 1)/2 = 105$, thus we asked less than 30% of the total number of pairwise comparisons. We then sent the survey to 34 colleagues ($N = 34$).

This survey was much more difficult than the one of Section 3.4.1, thus we expected that the data showed many non-transitive patters. Indeed 10 out of 34 (the 30%) users³

²<http://www.imdb.com/list/ls077140585/>.

³The users labelled 2, 6, 7, 9, 13, 14, 18, 23, 30, 32.

returned pairwise sets which contained at least one non-transitive pattern of comparisons.

In Table 3.3, we report the CP consensus of the movies obtained when applying the LM Algorithm with the same settings as in Section 3.4.1. The estimated consensus ranking is quite a lot different from the truth. This is not surprising since the respondents, who were our department colleagues, had no expertise in cinema, and also because some of the true cash magnets were quite old. We notice though a consensus of the users as regards the top movies and the last movies (as indicated by the narrower HPDIs). On the other hand, there is much uncertainty in the central rankings.

ρ_{true}	ρ_{LM}	CP_{LM}	$P(\rho_i < i)$	95% HPDI
7	1	Star Wars	0.7	(1,2)
3	2	Titanic	0.94	(1,3)
10	3	Jurassic Park	0.57	(2,5)
4	4	Avatar	0.51	(2,6)
14	5	The Godfather	0.81	(3,6)
12	6	The Lion King	0.87	(3,7)
9	7	E.T.	0.72	(6,9)
1	8	Gone with the wind	0.67	(6,10)
8	9	The sound of music	0.43	(7,12)
11	10	Jaws	0.68	(8,12)
15	11	Jurassic World	0.86	(8,12)
5	12	Ben Hur	0.85	(10,13)
2	13	Snow White	0.68	(12,15)
13	14	The Exorcist	0.56	(13,15)
6	15	The Ten Commandments	1	(13,15)

Table 3.3: Movies arranged by the estimated CP ordering with the logistic model for mistakes.

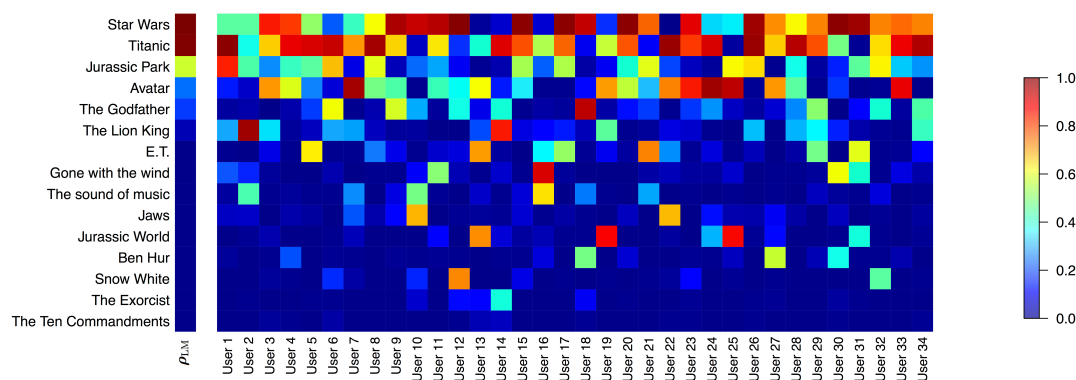


Figure 3.2: Posterior probability, for each movie, of being ranked among the top 3 in ρ_{LM} (column 1), and in \mathbf{R}_j , $j = 1, \dots, 34$ (next columns).

In Figure 3.2 is represented the posterior probability $P(\rho_{A_i} \leq 3 | \text{data})$ that a given movie A_i , $i = 1, \dots, 15$, was among the top-3 in the consensus ranking ρ_{LM} (first column), and in the individual rankings of all users, $j = 1, \dots, 34$ (remaining columns). As is clear

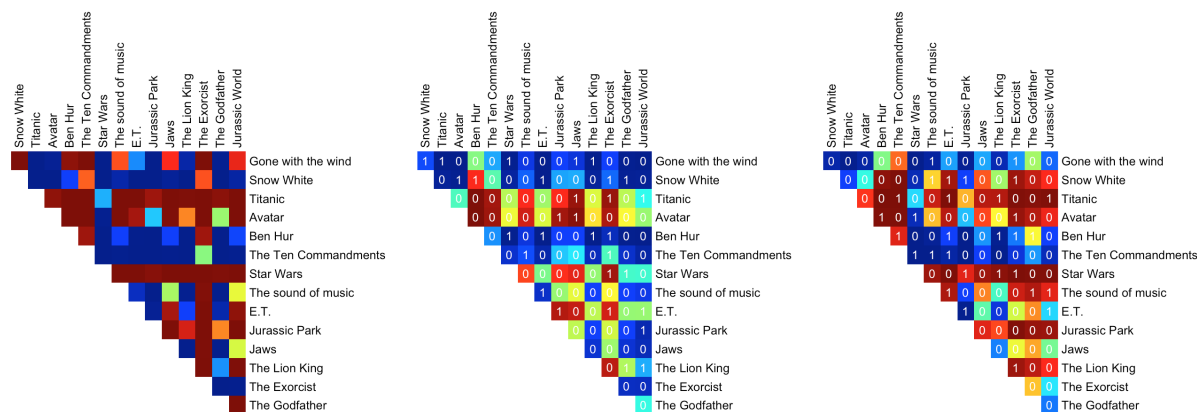


Figure 3.3: Preference matrices. Each entry (i, k) of the matrix represents the posterior probability that movie of row i is preferred to movie of column k in the CP consensus (left), in user 13 (middle), and in user 15 (right). The white numbers on each cell indicate whether the preference was assessed (1) or not (0) by the user in the data.

from the Figure, the top-3 movies are not shared by all users. For example Stars Wars, which is the top ranked movie in the consensus, was not estimated among the top-3 movies in 7 out of 34 users. We then computed the probability that movie A_i is preferred to movie A_k , for all pairs of movies, based on ρ_{LM} , and obtained the left half matrix of Figure 3.3. Each entry (i, k) of the matrix represents the posterior probability that movie i is preferred to movie k in the consensus ordering, that is $P(\rho_{A_i} < \rho_{A_k} | \text{data})$. We then produced the same half matrix on the basis of the individual rankings of the two users that were farthest from the consensus (in terms of top-3 detection). The results are showed in the middle and right matrices of Figure 3.3, where is clear that both the users are very different from the consensus.

3.5 Discussion

The main contribution of this chapter is to introduce a new Bayesian method for considering and correcting non-transitive pairwise preference data. The principal advantage of the Bayesian approach in this context comes from its ability to combine different types of uncertainty in the reported data, coming from different sources, and from being able to convert such data into the form of meaningful probabilistic inferences. Our method provides the posterior distribution of the consensus ranking, based on pairwise assessment data from a pool of users, who may have individually violated logical transitivity in their reporting. The method is also able to produce the posterior distributions of the

latent individual rankings of the users. Importantly such rankings can be used in the construction of personalized recommendations, or in studying how individual preferences change with user related covariates. We also developed mixture models generalization of the main model, able to handle heterogeneity in pairwise and non-transitive preference data, that will prove to be crucial in applications (see Chapter 4).

Appendix

3.A Algorithms

Algorithm for the logistic mistake model

The structure of the MCMC algorithm for the logistic model is the same as the one of Section 3.2. The differences in the LM version of the algorithm are (1) the acceptance ratios for $\mathbf{R}_1, \dots, \mathbf{R}_N$ (step 2 of Section 3.2), and (2) the way in which the parameters of the error model are updated (step 1(c) of Section 3.2).

The acceptance probability of an individual ranking, \mathbf{R}'_j , is the following.

If $g(\mathcal{B}_{jt}, \mathbf{R}'_j) = g(\mathcal{B}_{jt}, \mathbf{R}_j)$, $\forall t = 1, \dots, T_j$, the acceptance probability, conditioned on the current values of α , $\boldsymbol{\rho}$, β_0 , and β_1 , is $\min\{1, a_3\}$ where

$$\begin{aligned} \ln a_3 = & \ln a_1 - \frac{\beta_1}{n-2} \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j) \left[d_{\mathbf{R}'_j, t} - d_{\mathbf{R}_j, t} \right] + \\ & + \sum_{t=1}^{T_j} \ln \frac{1 + \exp \left[-\beta_0 - \beta_1 \frac{d_{\mathbf{R}_j, t-1}}{n-2} \right]}{1 + \exp \left[-\beta_0 - \beta_1 \frac{d_{\mathbf{R}'_j, t-1}}{n-2} \right]}. \end{aligned} \quad (3.29)$$

and where $\ln a_1$ is given by eq. (3.25). If $g(\mathcal{B}_{jt}, \mathbf{R}'_j) \neq g(\mathcal{B}_{jt}, \mathbf{R}_j)$, for some $t = 1, \dots, T_j$, the acceptance probability is $\min\{1, a_4\}$ where

$$\begin{aligned} \ln a_4 = & \ln a_1 - \beta_0 \sum_{t=1}^{T_j} \left[g(\mathcal{B}_{jt}, \mathbf{R}'_j) - g(\mathcal{B}_{jt}, \mathbf{R}_j) \right] + \sum_{t=1}^{T_j} \ln \frac{1 + \exp \left[-\beta_0 - \beta_1 \frac{d_{\mathbf{R}_j, t-1}}{n-2} \right]}{1 + \exp \left[-\beta_0 - \beta_1 \frac{d_{\mathbf{R}'_j, t-1}}{n-2} \right]} + \\ & - \frac{\beta_1}{n-2} \sum_{t=1}^{T_j} \left[g(\mathcal{B}_{jt}, \mathbf{R}'_j) (d_{\mathbf{R}'_j, t} - 1) - g(\mathcal{B}_{jt}, \mathbf{R}_j) (d_{\mathbf{R}_j, t} - 1) \right]. \end{aligned} \quad (3.30)$$

In place of the Gibbs step for θ , there are two Metropolis steps for updating β_0 and β_1 .

The β_1 step, conditioning on the current values of α , $\boldsymbol{\rho}$, β_0 and $\mathbf{R}_1, \dots, \mathbf{R}_N$, is performed as follows. We sample the proposal β'_1 from $\ln \mathcal{N}(\ln \beta_1, \sigma_{\beta_1})$, and accept it with probability $\min\{1, a_{\beta_1}\}$, where:

$$\begin{aligned} \ln a_{\beta_1} = & (\beta_1 - \beta'_1) \left[\lambda_{12} + \frac{1}{n-2} \sum_{j=1}^N \sum_{t=1}^{T_j} g(B_{jt}, \mathbf{R}_j)(d_{\mathbf{R}_j,t} - 1) \right] + \\ & + \sum_{j=1}^N \sum_{t=1}^{T_j} \ln \frac{1 + \exp \left[-\beta_0 - \beta_1 \frac{d_{\mathbf{R}_j,t-1}}{n-2} \right]}{1 + \exp \left[-\beta_0 - \beta'_1 \frac{d_{\mathbf{R}_j,t-1}}{n-2} \right]} + \lambda_{11} [\ln(\beta'_1/\beta_1)]. \end{aligned} \quad (3.31)$$

The β_0 step is performed by conditioning on the current values of α , $\boldsymbol{\rho}$, β_1 and $\mathbf{R}_1, \dots, \mathbf{R}_N$. We sample the proposal β'_0 from $\ln \mathcal{N}(\ln(\beta_0), \sigma_{\beta_0})$, and accept with probability $\min\{1, a_{\beta_0}\}$, where:

$$\begin{aligned} \ln a_{\beta_0} = & (\beta_0 - \beta'_0) \left[\lambda_{01} + \sum_{j=1}^N \sum_{t=1}^{T_j} g(B_{jt}, \mathbf{R}_j) \right] + \lambda_{00} [\ln(\beta'_0/\beta_0)] \\ & + \sum_{j=1}^N \sum_{t=1}^{T_j} \ln \frac{1 + \exp \left[-\beta_0 - \beta_1 \frac{d_{\mathbf{R}_j,t-1}}{n-2} \right]}{1 + \exp \left[-\beta'_0 - \beta_1 \frac{d_{\mathbf{R}_j,t-1}}{n-2} \right]}. \end{aligned} \quad (3.32)$$

The pseudo-code of the MCMC for logistic mistakes is reported as Algorithm 8.

Algorithm 8: MCMC Algorithm for logistic model for mistakes.

input : $\mathcal{B}_1, \dots, \mathcal{B}_N$; $\lambda, \gamma, \sigma_\alpha, L^*, \lambda_{01}, \lambda_{02}, \lambda_{11}, \lambda_{12}, \sigma_{\beta_0}, \sigma_{\beta_1}, d(\cdot, \cdot), Z_n(\alpha), M$.

output: Posterior distributions of $\boldsymbol{\rho}, \alpha, \beta_0, \beta_1$ and $\mathbf{R}_1, \dots, \mathbf{R}_N$.

Initialization of the MCMC: randomly generate $\boldsymbol{\rho}_0, \alpha_0, \beta_{0,0}, \beta_{1,0}$ and $\mathbf{R}_1^0, \dots, \mathbf{R}_N^0$.

for $m \leftarrow 1$ **to** M **do**

M-H step: update $\boldsymbol{\rho}$: Same as Algorithm 7

M-H step: update α : Same as Algorithm 7

M-H step: update β_1 :

 sample: $\beta'_1 \sim \ln \mathcal{N}(\beta_{1,m-1}, \sigma_{\beta_1}^2)$ and $u \sim \mathcal{U}(0, 1)$

 compute: $ratio \leftarrow$ eq. (3.31) with $\beta_0 = \beta_{0,m-1}$ and $\mathbf{R}_{1:N} \leftarrow \mathbf{R}_{1:N}^{m-1}$

if $u < ratio$ **then** $\beta_{1,m} \leftarrow \beta'_1$

else $\beta_{1,m} \leftarrow \beta_{1,m-1}$

M-H step: update β_0 :

 sample: $\beta'_0 \sim \ln \mathcal{N}(\beta_{0,m-1}, \sigma_{\beta_0}^2)$ and $u \sim \mathcal{U}(0, 1)$

 compute: $ratio \leftarrow$ eq. (3.32) with $\beta_1 = \beta_{1,m}$ and $\mathbf{R}_{1:N} \leftarrow \mathbf{R}_{1:N}^{m-1}$

if $u < ratio$ **then** $\beta_{0,m} \leftarrow \beta'_0$

else $\beta_{0,m} \leftarrow \beta_{0,m-1}$

M-H step: update $\mathbf{R}_1, \dots, \mathbf{R}_N$:

for $j \leftarrow 1$ **to** N **do**

 sample: $\mathbf{R}'_j \sim \text{Swap}(\mathbf{R}_j^{m-1}, L^*)$ and $u \sim \mathcal{U}(0, 1)$

if $\sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}'_j) = \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j^{m-1})$ **then** compute: $ratio \leftarrow$ eq. (3.29) with $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho}_m, \alpha \leftarrow \alpha_m$,

$\beta_1 \leftarrow \beta_{1,m}$ and $\beta_0 \leftarrow \beta_{0,m}$

else compute: $ratio \leftarrow$ eq. (3.30) with $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho}_m, \alpha \leftarrow \alpha_m, \beta_1 \leftarrow \beta_{1,m}$ and $\beta_0 \leftarrow \beta_{0,m}$

if $u < ratio$ **then** $\mathbf{R}_j^m \leftarrow \mathbf{R}'_j$

else $\mathbf{R}_j^m \leftarrow \mathbf{R}_j^{m-1}$

end

end

Algorithm for the mixture on θ

As mentioned in Section 3.1.4, the structure of the MCMC is:

1. Update $\alpha, \boldsymbol{\rho}, \tau_{1:K}, \theta_{1:K}$ and $\xi_{1:N}$ given $\mathcal{B}_{1:N}$ and $\mathbf{R}_{1:N}$, using eq. (3.17):
 - (a) Metropolis update of $\boldsymbol{\rho}$ (same as Algorithm 7)
 - (b) Metropolis update of α (same as Algorithm 7)
 - (c) Gibbs update of $\tau_{1:K}$
 - (d) Gibbs update of $\xi_{1:N}$
 - (e) Gibbs update of $\theta_{1:K}$
2. Update $\mathbf{R}_{1:N}$ given $\alpha, \boldsymbol{\rho}, \tau_{1:K}, \theta_{1:K}, \xi_{1:N}$ and $\mathcal{B}_{1:N}$, using eq. (3.18).

In step 1(c) we update τ_1, \dots, τ_K , by sampling from a Dirichlet density with updated hyperparameters, $\tau_1, \dots, \tau_K \sim \mathcal{D}(\psi + n_1, \dots, \psi + n_K)$, where $n_k = \sum_{j=1}^N \mathbb{1}_k(\xi_j)$, $k = 1, \dots, K$. In step 1(d) we then sample each ξ_j , independently, from the categorical distribution with probabilities,

$$P(\xi_j = k | \tau_k, \theta_k, \mathbf{R}_j, \mathcal{B}_j) \propto \tau_k \left(\frac{\theta_k}{1 - \theta_k} \right)^{\sum_{t=1}^{T_j} g(\mathcal{B}_{j,t}, \mathbf{R}_j)} (1 - \theta_k)^{T_j}, \forall k. \quad (3.33)$$

Step 1(e) consists of K independent Gibbs steps for updating $\theta_1, \dots, \theta_K$: for each $k = 1, \dots, K$ we sample θ_k from the beta distribution, truncated to the interval $[0, 0.5)$, with updated hyper-parameters,

$$\kappa'_1 = \kappa_1 + \sum_{j:\xi_j=k} \sum_{t=1}^{T_j} g(\mathcal{B}_{j,t}, \mathbf{R}_j), \quad \kappa'_2 = \kappa_2 + \sum_{j:\xi_j=k} \sum_{t=1}^{T_j} [1 - g(\mathcal{B}_{j,t}, \mathbf{R}_j)]. \quad (3.34)$$

Step 2 goes as follows. For each user $j = 1, \dots, N$, a new rank vector \mathbf{R}'_j is proposed with the Swap proposal centered at \mathbf{R}_j , and accepted with probability $\min\{1, a_1\}$, like in (3.25), if $g(\mathcal{B}_{j,t}, \mathbf{R}'_j) = g(\mathcal{B}_{j,t}, \mathbf{R}_j)$, $\forall t = 1, \dots, T_j$. If $g(\mathcal{B}_{j,t}, \mathbf{R}'_j) \neq g(\mathcal{B}_{j,t}, \mathbf{R}_j)$, for some $t = 1, \dots, T_j$, \mathbf{R}'_j is accepted with probability $\min\{1, a_5\}$, where

$$\ln a_5 = \ln a_1 + \sum_{t=1}^{T_j} [g(\mathcal{B}_{j,t}, \mathbf{R}'_j) - g(\mathcal{B}_{j,t}, \mathbf{R}_j)] \ln [\theta_{\xi_j} / (1 - \theta_{\xi_j})]. \quad (3.35)$$

The pseudo-code of the MCMC for the mixture on θ is reported as Algorithm 9.

Algorithm 9: MCMC Algorithm for mixture on θ .

input : $\mathcal{B}_1, \dots, \mathcal{B}_N; \lambda, \gamma, \sigma_\alpha, L^*, \kappa_1, \kappa_2, d(\cdot, \cdot), Z_n(\alpha), n, M, K, \psi$
output: Posterior distributions of $\rho, \alpha, \theta_1, \dots, \theta_K, \tau_1, \dots, \tau_K, \xi_1, \dots, \xi_N$ and $\mathbf{R}_1, \dots, \mathbf{R}_N$
Initialization of the MCMC: randomly generate $\rho_0, \alpha_0, \tau_{1,0}, \dots, \tau_{K,0}, \xi_{1,0}, \dots, \xi_{N,0}, \theta_{1,0}, \dots, \theta_{K,0}$ and $\mathbf{R}_1^0, \dots, \mathbf{R}_N^0$

for $m \leftarrow 1$ to M do

M-H step: update ρ : Same as Algorithm 7
M-H step: update α : Same as Algorithm 7

Gibbs step: update τ_1, \dots, τ_K
compute: $n_k = \sum_{j=1}^N \mathbb{1}_k(\xi_{j,m-1})$, for $k = 1, \dots, K$
sample: $\tau_{1,m}, \dots, \tau_{K,m} \sim \mathcal{D}(\psi + n_1, \dots, \psi + n_K)$

Gibbs step: update ξ_1, \dots, ξ_N
for $j \leftarrow 1$ to N do
 foreach $k \leftarrow 1$ to K **do** compute cluster assignment probabilities $p_{j,k}$ from equation (3.33), with $\tau_{1:K} \leftarrow \tau_{1:K,m}$,
 $\theta_{1:K} \leftarrow \theta_{1:K,m-1}$ and $\mathbf{R}_{1:N} \leftarrow \mathbf{R}_{1:N}^{m-1}$
 sample: $\xi_{j,m} \sim \text{Cat}(p_{j,1}, \dots, p_{j,K})$
end

Gibbs steps: update $\theta_1, \dots, \theta_K$
for $k \leftarrow 1$ to K do
 compute: κ'_1 and κ'_2 from eq. (3.34), with $\mathbf{R}_{1:N} \leftarrow \mathbf{R}_{1:N}^{m-1}$ and $\xi_{1:N} \leftarrow \xi_{1:N,m}$
 sample: $\theta_{k,m} \sim \text{Be}(\kappa'_1, \kappa'_2)$ truncated to the interval $[0, 0.5]$
end

Update $\mathbf{R}_1, \dots, \mathbf{R}_N$
for $j \leftarrow 1$ to N do
 sample: $\mathbf{R}'_j \sim \text{Swap}(\mathbf{R}_j^{m-1}, L^*)$ and $u \sim \mathcal{U}(0, 1)$
 if $\sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}'_j) = \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j^{m-1})$ **then** compute: $\text{ratio} \leftarrow$ eq. (3.25) with $\rho \leftarrow \rho_m$ and $\alpha \leftarrow \alpha_m$
 else compute: $\text{ratio} \leftarrow$ eq. (3.35) with $\rho \leftarrow \rho_m, \alpha \leftarrow \alpha_m, \theta_{\xi_j} \leftarrow \theta_{\xi_{j,m,m}}$
 if $u < \text{ratio}$ **then** $\mathbf{R}_j^m \leftarrow \mathbf{R}'_j$
 else $\mathbf{R}_j^m \leftarrow \mathbf{R}_j^{m-1}$
end

end

Algorithm for the mixture on ρ and α

As mentioned in Section 3.1.5, the structure of the MCMC is:

1. Update $\alpha_{1:C}, \rho_{1:C}, \eta_{1:C}, \theta$ and $z_{1:N}$ given $\mathcal{B}_{1:N}$ and $\mathbf{R}_{1:N}$, using eq. (3.20):
 - (a) Metropolis update of $\rho_{1:C}$
 - (b) Metropolis update of $\alpha_{1:C}$
 - (c) Gibbs update of $\eta_{1:C}$
 - (d) Gibbs update of $z_{1:N}$
 - (e) Gibbs update of θ (same as Algorithm 7)
2. Update $\mathbf{R}_{1:N}$ given $\alpha_{1:C}, \rho_{1:C}, \eta_{1:C}, \theta, z_{1:N}$ and $\mathcal{B}_{1:N}$, using eq. (3.21).

Steps 1(a) and 1(b) are straightforward, since $(\rho_c, \alpha_c)_{c=1, \dots, C}$ are conditionally independent given z_1, \dots, z_N . The proposal ρ'_c for each cluster is sampled from the **Swap** distribution centered at ρ_c , and accepted with probability $\min\{1, a_{\rho_c}\}$, where:

$$\ln a_{\rho_c} = -\frac{\alpha_c}{n} \sum_{j: z_j=c} [d(\mathbf{R}_j, \rho'_c) - d(\mathbf{R}_j, \rho_c)]. \quad (3.36)$$

Next, $\alpha'_c \sim \log\mathcal{N}(\ln \alpha_c, \sigma_\alpha^2)$ is accepted with probability $\min\{1, a_{\alpha_c}\}$, where

$$\ln a_{\alpha_c} = \gamma \ln(\alpha'_c/\alpha_c) - \left[\lambda + \frac{1}{n} \sum_{j: z_j=c} d(\mathbf{R}_j, \boldsymbol{\rho}_c) \right] (\alpha' - \alpha) - N_c \ln [Z_n(\alpha'_c)/Z_n(\alpha_c)], \quad (3.37)$$

and $N_c = \sum_{j=1}^N \mathbb{1}_c(z_j)$, $\forall c = 1, \dots, C$.

In 1(c), η_1, \dots, η_C are sampled from a Dirichlet density with updated hyperparameters, $\eta_1, \dots, \eta_C \sim \mathcal{D}(\chi + N_1, \dots, \chi + N_C)$, and in 1(d) we sample each z_j , independently, from the categorical distribution with probabilities,

$$P(z_j = c | \eta_c, \boldsymbol{\rho}_c, \alpha_c, \mathbf{R}_j) \propto \eta_c P(\mathbf{R}_j | \boldsymbol{\rho}_c, \alpha_c) = \eta_c \frac{e^{-\frac{\alpha_c}{n} d(\mathbf{R}_j, \boldsymbol{\rho}_c)}}{Z_n(\alpha_c)} \quad c = 1, \dots, C. \quad (3.38)$$

The Gibbs Sampler step for θ , is the same as step 1(c) of Section 3.2.

In step 2, for each $j = 1, \dots, N$, we sample a new rank vector \mathbf{R}'_j from the the Swap proposal centered at \mathbf{R}_j , and accept it with probability $\min\{1, a_6\}$, where

$$\ln a_6 = -\frac{\alpha_{z_j}}{n} [d(\mathbf{R}'_j, \boldsymbol{\rho}_{z_j}) - d(\mathbf{R}_j, \boldsymbol{\rho}_{z_j})], \quad (3.39)$$

if $g(\mathcal{B}_{jt}, \mathbf{R}'_j) = g(\mathcal{B}_{jt}, \mathbf{R}_j)$, $\forall t = 1, \dots, T_j$, and with probability $\min\{1, a_7\}$, where

$$\ln a_7 = \ln a_6 + \sum_{t=1}^{T_j} [g(\mathcal{B}_{jt}, \mathbf{R}'_j) - g(\mathcal{B}_{jt}, \mathbf{R}_j)] \ln [\theta/(1 - \theta)], \quad (3.40)$$

if $g(\mathcal{B}_{jt}, \mathbf{R}'_j) \neq g(\mathcal{B}_{jt}, \mathbf{R}_j)$, for some $t = 1, \dots, T_j$.

The pseudo-code of the MCMC for for the mixture on the Mallows parameters is reported as Algorithm 10.

3.B Sample simulated data from the Mallows model with mistakes

In this section we explain the procedure we used to sample a set of pairwise preferences, $\mathcal{B} = \mathcal{B}_1, \dots, \mathcal{B}_N$, where $\mathcal{B}_j = \{\mathcal{B}_{j1}, \dots, \mathcal{B}_{jT_j}\}$, from the Mallows model in the presence of mistakes, thereby extending the procedure of Appendix 2.B. The scheme is the following:

1. Sample $\mathbf{R}_{1,\text{true}}, \dots, \mathbf{R}_{N,\text{true}}$, from the Mallows density, $\mathcal{M}(\alpha_{\text{true}}, \boldsymbol{\rho}_{\text{true}})$;

Algorithm 10: MCMC Algorithm for mixture on ρ and α .

input : $\mathcal{B}_1, \dots, \mathcal{B}_N; \lambda, \gamma, \sigma_\alpha, L^*, \kappa_1, \kappa_2, d(\cdot, \cdot), Z_n(\alpha), n, M, K, \chi, C$
output: Posterior distributions of $\rho_1, \dots, \rho_C, \alpha_1, \dots, \alpha_C, \theta, \eta_1, \dots, \eta_C, z_1, \dots, z_N$ and $\mathbf{R}_1, \dots, \mathbf{R}_N$
Initialization of the MCMC: randomly generate $\rho_{1,0}, \dots, \rho_{C,0}, \alpha_{1,0}, \dots, \alpha_{C,0}, \eta_{1,0}, \dots, \eta_{C,0}, z_{1,0}, \dots, z_{N,0}$, and $\mathbf{R}_1^0, \dots, \mathbf{R}_N^0$

```

for  $m \leftarrow 1$  to  $M$  do
  for  $c \leftarrow 1$  to  $C$  do
    M-H step: update  $\rho_c$ 
    sample:  $\rho'_c \sim \text{Swap}(\rho_{c,m-1}, L^*)$  and  $u \sim \mathcal{U}(0, 1)$ 
    compute:  $ratio \leftarrow$  eq. (3.36) with  $\rho_c \leftarrow \rho_{c,m-1}, \alpha_c \leftarrow \alpha_{c,m-1}$ , and  $z_{1:N} \leftarrow z_{1:N,m-1}$ 
    if  $u < ratio$  then  $\rho_{c,m} \leftarrow \rho'_c$ 
    else  $\rho_{c,m} \leftarrow \rho_{c,m-1}$ 

    M-H step: update  $\alpha_c$ 
    sample:  $\alpha'_c \sim \ln \mathcal{N}(\ln \alpha_{c,m-1}, \sigma_\alpha^2)$  and  $u \sim \mathcal{U}(0, 1)$ 
    compute:  $ratio \leftarrow$  eq. (3.37) with  $\rho_c \leftarrow \rho_{c,m}, \alpha_c \leftarrow \alpha_{c,m-1}$ , and  $z_{1:N} \leftarrow z_{1:N,m-1}$ 
    if  $u < ratio$  then  $\alpha_{c,m} \leftarrow \alpha'_c$ 
    else  $\alpha_{c,m} \leftarrow \alpha_{c,m-1}$ 
  end

  Gibbs step: update  $\eta_1, \dots, \eta_C$ 
  compute:  $N_c = \sum_{j=1}^N 1_c(z_{j,m-1})$ , for  $c = 1, \dots, C$ 
  sample:  $\eta_{1,m}, \dots, \eta_{C,m} \sim \mathcal{D}(\chi + N_1, \dots, \chi + N_C)$ 

  Gibbs step: update  $z_1, \dots, z_N$ 
  for  $j \leftarrow 1$  to  $N$  do
    foreach  $c \leftarrow 1$  to  $C$  do compute cluster assignment probabilities  $p_{j,c}$  from equation (3.38)
    with  $\eta_{1:C} \leftarrow \eta_{1:C,m}, \alpha_{1:C} \leftarrow \alpha_{1:C,m}, \mathbf{R}_{1:N} \leftarrow \mathbf{R}_{1:N}^{m-1}$  and  $\rho_{1:C} \leftarrow \rho_{1:C,m}$ 
    sample:  $z_{j,m} \sim \text{Cat}(p_{j,1}, \dots, p_{j,C})$ 
  end

  Gibbs step: update  $\theta$ : Same as Algorithm 7
  Update  $\mathbf{R}_1, \dots, \mathbf{R}_N$ 
  for  $j \leftarrow 1$  to  $N$  do
    sample:  $\mathbf{R}'_j \sim \text{Swap}(\mathbf{R}_j^{m-1}, L^*)$  and  $u \sim \mathcal{U}(0, 1)$ 
    if  $\sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}'_j) = \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \mathbf{R}_j^{m-1})$  then compute:  $ratio \leftarrow$  eq. (3.39) with  $\rho_c \leftarrow \rho_{c,m}$  and  $\alpha_c \leftarrow \alpha_{c,m}$ 
    else compute:  $ratio \leftarrow$  eq. (3.40) with  $\rho_c \leftarrow \rho_{c,m}, \alpha_c \leftarrow \alpha_{c,m}, \theta_{z_j} \leftarrow \theta_{z_j,m}$ 
    if  $u < ratio$  then  $\mathbf{R}_j^m \leftarrow \mathbf{R}'_j$ 
    else  $\mathbf{R}_j^m \leftarrow \mathbf{R}_j^{m-1}$ 
  end
end
end

```

2. For each $j = 1, \dots, N$, select the numbers of pairwise comparisons, $T_j < T_{\text{Max}}$ and, sample, without replacement, T_j unordered pairs from the collection of T_{Max} possible pairs $\mathcal{C}_j = \{\mathcal{C}_{j1}, \dots, \mathcal{C}_{jT_j}\}$;
3. For each pair \mathcal{C}_{jt} , generate the ordered comparison \mathcal{B}_{jt} , either correctly (w.r.t. $\mathbf{R}_{j,\text{true}}$) or reversed, with probability depending on the model used, BM or LM.

Steps 1-3 are almost the same as in Appendix 2.B, the only difference being in that the data were here produced in a nested fashion. This was done to facilitate the comparability of the simulation results generated under different parameter settings.

When increasing N , the number of users

For fixed $n, \theta, \lambda_T, \alpha$ and ρ , we first generated $\mathcal{B} = \mathcal{B}_1, \dots, \mathcal{B}_N$ with the largest N , through steps 1-4 of Section 3.B. Then we created the nested datasets by subsampling from \mathcal{B} the intended smaller number of users.

When increasing λ_T , the average number of pairs per user

Fix n , θ , N , α and ρ , and let $1 \leq \lambda_T^1 < \dots < \lambda_T^T \leq M_{\max}$. The goal here was to create, for all users j , individual nested pairwise datasets, $\mathcal{B}_{\lambda_T^t}$, $1 \leq t \leq T$, such that (i) all pairs present in $\mathcal{B}_{\lambda_T^t}$ had to be present also in $\mathcal{B}_{\lambda_T^{t+1}}$, and (ii) the cardinalities $T_j^t = |\mathcal{B}_{j, \lambda_T^t}|$ would satisfy the approximation $\mathbb{E}(T_j^t) \approx \lambda_T^t$. This was achieved by first sampling $T_j^T = |\mathcal{B}_{j, \lambda_T^T}|$ from the truncated Poisson distribution with parameter λ_T^T , truncated at T_{\max} , and then performing sequential thinning of the pairs, moving first from $\mathcal{B}_{\lambda_T^T}$ to $\mathcal{B}_{\lambda_T^{T-1}}$, then from $\mathcal{B}_{\lambda_T^{T-1}}$ to $\mathcal{B}_{\lambda_T^{T-2}}$, etc., until finally reaching $\mathcal{B}_{\lambda_T^1}$. Thinning was done independently for different users j , and so that approximately the proportion $\lambda_T^{t-1}/\lambda_T^t$ of the T_j^t pairs in $\mathcal{B}_{\lambda_T^t}$ were kept also in $\mathcal{B}_{\lambda_T^{t-1}}$, $1 \leq t \leq T$, with the other pairs being removed.

When increasing the number of mistakes in the BM model

The goal here was to create nested datasets, $\mathcal{B}_{\theta_1}, \dots, \mathcal{B}_{\theta_T}$, corresponding to increasing average numbers of mistakes, $0 < \theta_1 < \dots < \theta_T < 0.5$. Each $\mathcal{B}_{\theta_t} = \{\mathcal{B}_{1, \theta_t}, \dots, \mathcal{B}_{N, \theta_t}\}$, $t \in \{1, \dots, T\}$, is then the collection of the pairwise comparisons of the N assessors, corresponding to θ_t number of mistakes. For generating the nested sequence of datasets, the rule was that the mistakes of $\mathcal{B}_{j, \theta_t}$ had to remain in $\mathcal{B}_{j, \theta_{t+1}}$, and the probability of a mistake in each pair in $\mathcal{B}_{j, \theta_t}$ was θ_t .

We implemented points 1-3 of section B.1, while step 4 was done as follows. For each \mathcal{C}_{jt} , $j = 1, \dots, N$, $t = 1, \dots, T_j$:

- Divide the interval $[0, 1)$ into subintervals: $[0, \theta_1), \dots, [\theta_{T-1}, \theta_T), [\theta_T, 1)$;
- Sample $u \sim U(0, 1)$;
- If $u \in [\theta_t, \theta_{t+1})$, $t \in \{0, \dots, T\}$, then, in generating the datasets $\mathcal{B}_{j, \theta_1}, \dots, \mathcal{B}_{j, \theta_{t-1}}$, keep the order of the pair comparison $\mathcal{B}_{j, t}$ the same as it is in $\mathbf{R}_{j, \text{true}}$, and reverse it in the data sets $\mathcal{B}_{j, \theta_t}, \dots, \mathcal{B}_{j, \theta_T}$.

When increasing the number of mistakes in the LM model

The goal here was to create nested data with an increasing number of mistakes generated from the logistic model:

$$\text{logit } P(\mathcal{B}_{jt} \text{ mistake} | \mathbf{R}_j, \beta_0, \beta_1) = -\beta_0 - \beta_1 \frac{d_{\mathbf{R}_j, t} - 1}{n - 2}.$$

To do so, we either fixed β_1 but varied β_0 , or fixed β_0 but varied β_1 . Here we explain the case of β_1 fixed at β_1^* and β_0 varying. Let us denote the nested datasets as $\mathcal{B}_{\beta_{0,1}}, \dots, \mathcal{B}_{\beta_{0,T}}$, corresponding to the decreasing sequence $\beta_{0,1} > \dots > \beta_{0,T} > 0$, and assume, as in the previous section, that the mistakes in $\mathcal{B}_{j,\beta_{0,t}}$ had to remain in $\mathcal{B}_{j,\beta_{0,t-1}}$ (notice that here the larger β_0 is, the less mistakes are in the data).

For an illustration, consider datasets of $n = 10$ items.

In Table 3.B.1, we show $P(\mathcal{B}_{jt} \text{ mistake} | \mathbf{R}_j, \beta_0, \beta_1 = \beta_1^*)$, for $\beta_1^* = 5$, for some chosen values of β_0 (columns), depending on the value of the distance between the compared items (rows).

$d_{\mathbf{R}_{j,m}}$	$\beta_{0,1} = 1.6$	$\beta_{0,2} = 1.1$	$\beta_{0,3} = 0.6$	$\beta_{0,4} = 0.1$
1	0.17	0.25	0.35	0.48
2	0.11	0.17	0.25	0.35
3	0.07	0.11	0.17	0.25
4	0.04	0.07	0.11	0.17
5	0.03	0.04	0.07	0.11
6	0.02	0.03	0.04	0.07
7	0.01	0.02	0.03	0.04
8	0.01	0.01	0.02	0.03
9	0.00	0.01	0.01	0.02

Table 3.B.1: Logistic theoretical values of the probability of making a mistake depending on the value of the distance between the compared items.

We denote the matrix of values of Table 3.B.1 by Λ , and its d -th row by Λ_d . For each user j , we first sampled a set of pair comparisons \mathcal{C}_j , as in steps 1-3 of 3.B, and then generated four (corresponding to the 4 values of β_0) nested datasets with increasing number of mistakes as follows. For each \mathcal{C}_{jt} , $j = 1, \dots, N$, $t = 1, \dots, T_j$:

- Compute $d_{\mathbf{R}_{j,t}}$ and select the corresponding row $\Lambda_{d_{\mathbf{R}_{j,t}}}$;
- Divide the interval $[0, 1)$ into 4 subintervals with, as extremes, the values of $\Lambda_{d_{\mathbf{R}_{j,t}}}$ $[0, \beta_{0,4}(d_{\mathbf{R}_{j,t}})), \dots, [\beta_{0,1}(d_{\mathbf{R}_{j,t}}), 1)$;
- Sample $u \sim U(0, 1)$;
- If $u \in [\beta_{0,t+1}(d_{\mathbf{R}_{j,t}}), \beta_{0,t}(d_{\mathbf{R}_{j,t}}))$, $t \in \{0, \dots, T\}$, with $\beta_{0,0} = 1$ and $\beta_{0,T+1} = 0$, then, in generating the datasets $\mathcal{B}_{\beta_{0,T}}, \dots, \mathcal{B}_{\beta_{0,t}}$, keep the order of the pair comparison \mathcal{B}_{jt} the same as it is in $\mathbf{R}_{j,\text{true}}$, and reverse it in the data sets $\mathcal{B}_{\beta_{0,t+1}}, \dots, \mathcal{B}_{\beta_{0,1}}$.

3.C A generalized version of the Bradley Terry model

In this appendix we implement a model extension proposed by an anonymous referee, which is an extension of the Bradley Terry model (Bradley and Terry 1952), with an additional layer for individual variability. Here, we show how we implemented this model (which we henceforth call BTI), and make a fair comparison between our proposed Bayesian Mallows with Bernoulli mistakes (of Section 3.1.1) and the BTI, based on data that resemble the ones for which our method is designed for. All in all we find that the BTI is better suited to data of different type than the ones motivating our approach.

Throughout we only consider homogeneous data, even if both models can be extended to include clustering of users, as we explicitly do for the BM.

The BTI model

As in the Bradley Terry model (Section 1.2), suppose that the preferences expressed by a user in pair comparisons have the form (1.12), but now allowing for the score parameters to vary from one user to another. Denoting the parameter vector of user j by $\boldsymbol{\mu}_j = (\mu_{j1}, \dots, \mu_{jn})$, we assume that

$$\Pr(A_i \prec_j A_k | \mu_{ji}, \mu_{jk}) = \frac{\mu_{ji}}{\mu_{ji} + \mu_{jk}}, \quad (3.41)$$

where with \prec_j is meant the preference according to user j .

Thus, given $\boldsymbol{\mu}_j$, the outcomes of the pair comparisons for user j depend only on the relative sizes of the individual score parameters of the items being compared. Assuming in addition that such outcomes, given $\boldsymbol{\mu}_j$, are conditionally independent leads to the corresponding product form likelihood of expressions of the form (3.41).

Suppose that N users express a number of pairwise comparisons among n items. We denote by $\mathbf{D} = \{\mathbf{D}_1, \dots, \mathbf{D}_N\}$ the associated data, where \mathbf{D}_j denotes the data coming from user j . Let w_{jik} denote the number of comparisons of user j where A_i is preferred to A_k , $w_{ji} := \sum_{k \neq i}^n w_{jik}$ the number of times item A_i is preferred to any other item by user j , and $n_{jik} = w_{jik} + w_{jki}$ the number of comparisons that user j assesses between A_i and A_k . Assume that the data $\mathbf{D}_1, \dots, \mathbf{D}_N$, are conditionally independent given $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N$, and the distribution of the pairs in each \mathbf{D}_j depends only on the corresponding parameters $\boldsymbol{\mu}_j$.

The log-likelihood function of $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N$ for the joint data is then

$$\begin{aligned} l(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N; \mathbf{D}) &= \sum_{j=1}^N l(\boldsymbol{\mu}_j | \mathbf{D}_j) = \sum_{j=1}^N \sum_{1 \leq i \neq k \leq n} w_{jik} [\ln \mu_{ji} - \ln(\mu_{ji} + \mu_{jk})] = \\ &= \sum_{j=1}^N \sum_{i=1}^n w_{ji} \ln \mu_{ji} - \sum_{j=1}^N \sum_{1 \leq i < k \leq n} n_{jik} \ln(\mu_{ji} + \mu_{jk}). \end{aligned} \quad (3.42)$$

Next, we specify a prior for all the user specific parameters $\boldsymbol{\mu}_j$, centered around the shared consensus parameter vector $\boldsymbol{\mu}$, again postulating conditional independence. This is done in terms of a lower level gamma model, specified as

$$\pi(\boldsymbol{\mu}_{1:N} | \boldsymbol{\mu}) = \prod_{j=1}^N \prod_{i=1}^n \mathcal{G}a(\mu_{ji}; a_{\boldsymbol{\mu}}, a_{\boldsymbol{\mu}}/\mu_i), \quad (3.43)$$

where with $\mathcal{G}a(x; a, b)$ is denoted the gamma density with shape a and rate b , so that in this case the expected mean of μ_{ij} is equal to μ_i , and the variance to $\mu_i^2/a_{\boldsymbol{\mu}}$.

Finally, we propose a prior for $\boldsymbol{\mu}$, and assume an inverse gamma, conjugate to the model

$$\pi(\boldsymbol{\mu}) = \prod_{i=1}^n \mathcal{IG}(\mu_i; a, b) \propto \prod_{i=1}^n \frac{e^{-\frac{b}{\mu_i}}}{\mu_i^{a+1}}. \quad (3.44)$$

Following the augmentation scheme of [Caron and Doucet \(2012\)](#), we define, for each j , and for each pair of items $\{A_i, A_k\}$, the following augmented variable

$$Z_{jik} = \sum_{m=1}^{n_{jik}} \min(Y_{mji}, Y_{mjk})$$

which, by definition, is gamma distributed, with parameters n_{jik} and $(\mu_{ji} + \mu_{jk})$.

The probability density of $\mathbf{Z}_j = \{Z_{j12}, Z_{j13}, \dots, Z_{j(n-1)n}\}$ is therefore given by

$$p(\mathbf{z}_j | \mathbf{D}_j, \boldsymbol{\mu}_j) = \prod_{1 \leq i < k \leq n: n_{jik} \geq 1} \mathcal{G}a(z_{jik}; n_{jik}, \mu_{ji} + \mu_{jk}),$$

and the resulting augmented data log-likelihood is given by

$$l^c(\mathbf{D}, \mathbf{Z}; \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N) = \sum_{j=1}^N l^c(\mathbf{D}_j, \mathbf{Z}_j; \boldsymbol{\mu}_j) = \sum_{j=1}^N \left[\sum_{i=1}^n w_{ji} \ln \mu_{ji} - \sum_{1 \leq i < k \leq n : n_{jik} \geq 1} [(\mu_{ji} + \mu_{jk}) z_{jik} - (n_{jik} - 1) \ln z_{jik} + \ln \Gamma(n_{jik})] \right]. \quad (3.45)$$

The posterior distribution based on the data \mathbf{D} is therefore given by

$$\begin{aligned} \pi(\boldsymbol{\mu}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N, \mathbf{Z} | \mathbf{D}, a_\mu) &\propto \pi(\boldsymbol{\mu}) \pi(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N | \boldsymbol{\mu}, a_\mu) l^c(\mathbf{D}, \mathbf{Z}; \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N) \propto \\ &\propto \left[\prod_{i=1}^n \frac{e^{-\frac{b}{\mu_i}}}{\mu_i^{a+1}} \right] \left[\prod_{j=1}^N \prod_{i=1}^n \frac{1}{\mu_i^{a_\mu}} \mu_{ji}^{a_\mu - 1} e^{-a_\mu \frac{\mu_{ji}}{\mu_i}} \right] \cdot \\ &\cdot \left[\prod_{j=1}^N \left(\prod_{i=1}^n \mu_{ji}^{w_{ji}} \right) \left(\prod_{1 \leq i < k \leq n : n_{jik} \geq 1} \frac{z_{jik}^{n_{jik} - 1} e^{-(\mu_{ji} + \mu_{jk}) z_{jik}}}{\Gamma(n_{jik})} \right) \right]. \end{aligned} \quad (3.46)$$

We can then sample from the posterior distribution of eq. (3.46), using the above data augmentation scheme, with Algorithm 11, when a_μ is given.

Algorithm 11: Gibbs Sampler for the BTL model

Data: $\mathbf{D}_1, \dots, \mathbf{D}_N, T, a_\mu, a, b$
Output: $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N$

```

for  $t = 1, \dots, T$  do
  Update  $\boldsymbol{\mu}$ 
  for  $i = 1, \dots, n$  do
    Sample:  $\mu_i^{(t)} | \mathbf{D}, \mathbf{Z}^{(t-1)}, \boldsymbol{\mu}_{1:N}^{(t-1)} \sim \mathcal{IG}(a + N a_\mu, b + a_\mu \sum_{j=1}^N \mu_{ji}^{(t-1)})$ 
  end
  Update  $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N$ 
  for  $1 \leq i < k \leq n$  do
    for  $i = 1, \dots, n$  such that  $n_{jik} \geq 1$  do
      Sample:  $z_{jik}^{(t)} | \mathbf{D}, \boldsymbol{\mu}_j^{(t-1)}, \boldsymbol{\mu}^{(t)} \sim \mathcal{Ga}(n_{jik}, \mu_{ji}^{(t-1)} + \mu_{jk}^{(t-1)})$ 
    end
    for  $i = 1, \dots, n$  do
      Sample  $\mu_{ji}^{(t)} | \mathbf{D}, \mathbf{Z}_j^{(t)}, \boldsymbol{\mu}^{(t)} \sim \mathcal{Ga}(a_\mu + w_{ji}, a_\mu / \mu_i^{(t)} + \sum_{i < k : n_{jik} \geq 1} z_{jik}^{(t)} + \sum_{i > k : n_{jik} \geq 1} z_{jki}^{(t)})$ 
    end
  end
end
end

```

Hyperparameter estimation

We implemented a version of Algorithm 11, where we estimate a_μ from the data, and where we choose a gamma prior $\pi(a_\mu) \propto a_\mu^{\gamma-1} e^{-\lambda a_\mu}$ on a_μ (see Algorithm 12).

Then, we added a M-H random walk step to update a_μ : We propose $a_\mu^P = \ln \mathcal{N}(\ln(a_\mu^C), \sigma_a^2)$, where a_μ^C is the current value of a_μ , and accept it with probability $\min\{1, r_{a_\mu}\}$, where

$$\begin{aligned} \ln(r_{a_\mu}) = & (nNa_\mu^P + \gamma) \ln(a_\mu^P) - (nNa_\mu^C + \gamma) \ln(a_\mu^C) - nN[\ln \Gamma(a_\mu^P) - \ln \Gamma(a_\mu^C)] + \\ & -(a_\mu^P - a_\mu^C) \left[\lambda + \sum_{i=1}^n \left[\frac{1}{\mu_i} \left(\sum_{j=1}^N \mu_{ji} \right) + N \ln(\mu_i) - \sum_{j=1}^N \ln(\mu_{ji}) \right] \right]. \end{aligned} \quad (3.47)$$

The hyperparameters were chosen so that the prior is rather vague: $\lambda = 1/5$ and $\gamma = 2$, which correspond to prior mean $\mathbb{E}(a_\mu) = 10$, and variance $\mathbb{V}(a_\mu) = 50$.

Algorithm 12: Gibbs Sampler for the BTL model with M-H step for a_μ .

```

Data:  $D_1, \dots, D_N, T, a, b, \gamma, \lambda$ 
Output:  $\mu, a_\mu$  and  $\mu_1, \dots, \mu_N$ 
for  $t = 1, \dots, T$  do
    Update  $\mu$ 
    for  $i = 1, \dots, n$  do
        | Sample:  $\mu_i^{(t)} | D, Z^{(t-1)}, \mu_{1:N}^{(t-1)} \sim \mathcal{IG}(a + Na_\mu^{(t-1)}, b + a_\mu^{(t-1)} \sum_{j=1}^N \mu_{ji}^{(t-1)})$ 
    end

    Update  $a_\mu$ 
    Sample:  $a_\mu^P \sim \ln \mathcal{N}(\ln(a_\mu^{(t-1)}), \sigma_a^2)$  and  $u \sim \mathcal{U}(0, 1)$ 
    Compute:  $\ln(r_{a_\mu})$  (eq. (3.47)), with  $a_\mu = a_\mu^{(t-1)}$ ,  $\mu = \mu^{(t)}$ , and  $\mu_{1:N} = \mu_{1:N}^{(t-1)}$ 
    if  $u < r_{a_\mu}$  then
        |  $a_\mu^{(t)} = a_\mu^P$ 
    else
        |  $a_\mu^{(t)} = a_\mu^{(t-1)}$ 
    end

    Update  $\mu_1, \dots, \mu_N$ 
    for  $1 \leq i < k \leq n$  do
        for  $i = 1, \dots, n$  such that  $n_{jik} \geq 1$  do
            | Sample:  $Z_{jik}^{(t)} | D, \mu_j^{(t-1)}, \mu^{(t)} \sim \mathcal{Ga}(n_{jik}, \mu_{ji}^{(t-1)} + \mu_{jk}^{(t-1)})$ 
        end
        for  $i = 1, \dots, n$  do
            | Sample:  $\mu_j^{(t)} | D, Z_j^{(t)}, \mu^{(t)} \sim \mathcal{Ga}(a_\mu^{(t)} + w_{ji}, a_\mu^{(t)} / \mu_i^{(t-1)} + \sum_{i < k: n_{jik} \geq 1} Z_{jik}^{(t)} + \sum_{i > k: n_{jik} \geq 1} Z_{jki}^{(t)})$ 
        end
    end
end

```

Experimental results

Estimating parameters with the BTI

We choose $n = 5$ items and $N = 50$ users. For fixed $\mu = (\mu_1, \dots, \mu_5)$ and a_μ , we generated a dataset of $N = 50$ individual rankings μ_1, \dots, μ_N from a gamma distribution as in eq. (3.43). For each user $j = 1, \dots, 50$, we then simulated on average $M = 20n(n-1)/2$ pair comparisons, from the BTI likelihood, eq. (3.41), expecting roughly 20 repeated assessments for each pair. We then run 5000 iterations of Algorithm 12.

We report here the trace plots of a_μ, μ and of some of the μ_j , to show convergence, which should be checked more formally.

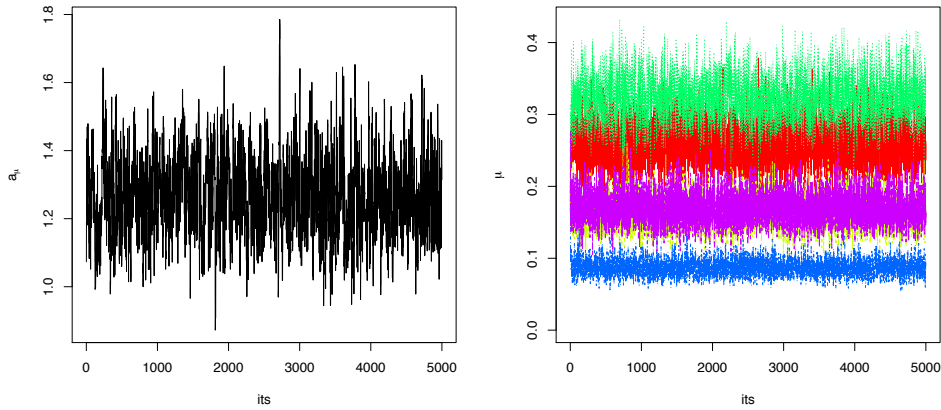


Figure 3.C.1: Trace plots of a_μ (left) and μ (right)

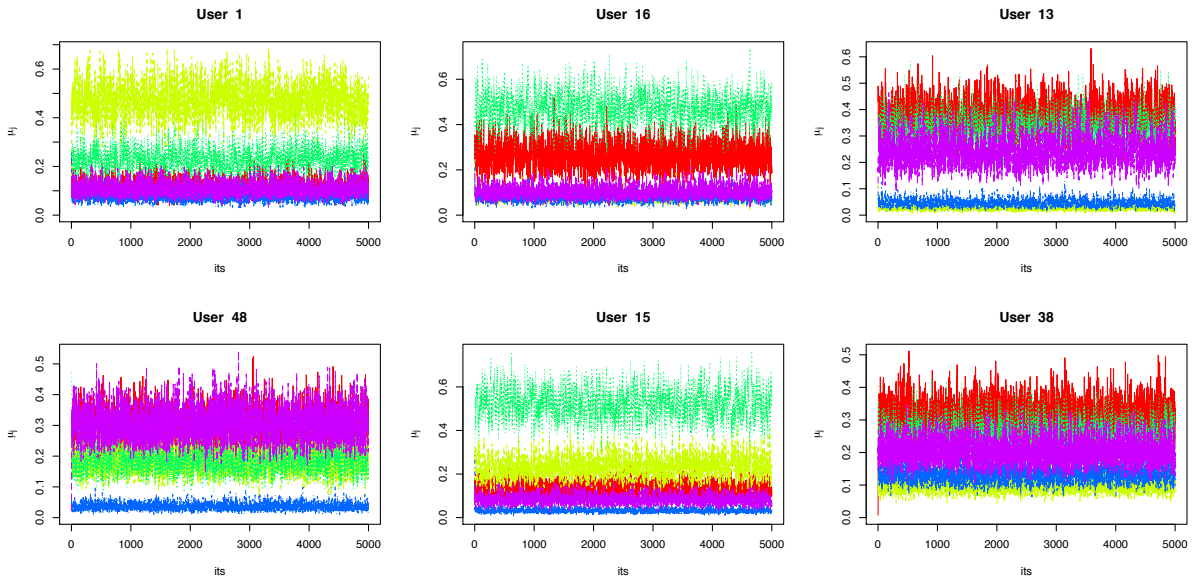


Figure 3.C.2: Trace plots of μ_j for some users.

Then, we reduce the average number of pair comparisons assessed by each user, considering the following cases: $M = 10n(n-1)/2, 5n(n-1)/2, n(n-1)/2, 0.5n(n-1)/2$. We are interested in studying the performance of the BTI in situations where there are few replicated assessments, as in our musicology case, where no replications at all were present. In the next tables we report the MSE of μ , $\text{MSE}(\mu) = \frac{1}{n} \sum_{i=1}^n (\mu_i - \hat{\mu}_i)^2$, and the average MSE of μ_1, \dots, μ_N , $\text{MSE}(\mu_1, \dots, \mu_N) = \frac{1}{N} \sum_{j=1}^N \frac{1}{n} \sum_{i=1}^n (\mu_{ij} - \hat{\mu}_{ij})^2$, where the estimated quantities $\hat{\mu}_i, \hat{\mu}_{ij}$ are posterior means.

The experiment should be in principle repeated several times, independently. We notice that, as expected, the performance deteriorates as the number of pairs decreases,

M	$20n(n-1)/2$	$10n(n-1)/2$	$5n(n-1)/2$	$n(n-1)/2$	$0.5n(n-1)/2$
$\text{MSE}(\boldsymbol{\mu})$	0.00038	0.00034	0.00025	0.00086	0.0025

Table 3.C.1: MSE of $\boldsymbol{\mu}$.

M	$20n(n-1)/2$	$10n(n-1)/2$	$5n(n-1)/2$	$n(n-1)/2$	$0.5n(n-1)/2$
$\text{MSE}(\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N)$	0.012	0.014	0.016	0.033	0.035

Table 3.C.2: MSE of $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N$.

and that estimates of the individual $\boldsymbol{\mu}_j$ are more difficult than the estimate of $\boldsymbol{\mu}$.

Comparison: data generated from BM

We first performed the following experiment.

- Simulate 6 datasets with the Mallows model with Bernoulli mistakes, BM.
 - Fix $n = 10$, $N = 50$, $\alpha^* = 4$, and simulate randomly $\boldsymbol{\rho}^* \in \mathcal{P}_n$;
 - Sample $\mathbf{R}_1^*, \dots, \mathbf{R}_N^* \sim \text{Mallows}(\alpha^*, \boldsymbol{\rho}^*)$;
 - Set the average number of pairwise comparisons per user to either $0.7n(n-1)/2$ or $0.5n(n-1)/2$;
 - Set $\theta^* = 0.05, 0.1, 0.15$;
 - Simulate the pair comparison data from the Bernoulli model from mistakes with the different values of θ^* , without repetitions.
- We then have the following cases, corresponding to different setting of parameters: (e1) $M = 0.5n(n-1)/2$, $\theta = 0.05$, (e2) $M = 0.5n(n-1)/2$, $\theta = 0.1$, (e3) $M = 0.5n(n-1)/2$, $\theta = 0.15$, (e4) $M = 0.7n(n-1)/2$, $\theta = 0.05$, (e5) $M = 0.7n(n-1)/2$, $\theta = 0.1$, (e6) $M = 0.7n(n-1)/2$, $\theta = 0.15$.
- Compare BM and BTI in terms of:

1. The posterior expected mean squared error

$$A_s(\mathbf{x}, \mathbf{x}^*) = \sum_{i=1}^n \sum_{k=1}^n P(x_i = k | x_i^* = i, \text{data}) (k - i)^2, \quad \mathbf{x}, \mathbf{x}^* \in \mathcal{P}_n$$

In order to compare the two procedures we translate the score vectors estimated through BTI $(\boldsymbol{\mu}, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N)$ into rankings. We do so by simply ordering the scores, leading to the rank vectors, $r(\boldsymbol{\mu}), r(\boldsymbol{\mu}_1), \dots, r(\boldsymbol{\mu}_N)$. Then,

for BTI we compute $A_s(r(\boldsymbol{\mu}), \boldsymbol{\rho}^*)$, $A_s(r(\boldsymbol{\mu}_1), \mathbf{R}_1^*), \dots, A_s(r(\boldsymbol{\mu}_N), \mathbf{R}_N^*)$, while for BM $A_s(\boldsymbol{\rho}, \boldsymbol{\rho}^*)$, $A_s(\mathbf{R}_1, \mathbf{R}_1^*), \dots, A_s(\mathbf{R}_N, \mathbf{R}_N^*)$.

2. The posterior l_1 distance between the estimated $\boldsymbol{\rho}$ and the true value of the consensus $\boldsymbol{\rho}^*$, $D(\boldsymbol{\rho}, \boldsymbol{\rho}^*) = \sum_{i=1}^n |\rho_i - \rho_i^*|$, and the average posterior distance between the estimated $\mathbf{R}_1, \dots, \mathbf{R}_N$ and true value of the individual rankings $\mathbf{R}_1^*, \dots, \mathbf{R}_N^*$, $\frac{1}{N} \sum_{j=1}^N D(\mathbf{R}_j, \mathbf{R}_j^*) = \sum_{i=1}^n |R_{ji} - R_{ji}^*|$

The results are summarized in the four tables below, where we compare BM and BTI in the different settings as above.

	θ	BM $A(\boldsymbol{\rho}, \boldsymbol{\rho}^*)$	BTI $A(r(\boldsymbol{\mu}), \boldsymbol{\rho}^*)$	BM $\overline{A_s(\mathbf{R}_j, \mathbf{R}_j^*)}$	BTI $\overline{A_s(r(\boldsymbol{\mu}_j), \mathbf{R}_j^*)}$
$M = 0.5n(n-1)/2$	0.05	1.82	4.24	20.6	37.45
	0.1	2.86	4.75	31.36	49.08
	0.15	3.33	7.79	42.69	51.53

Table 3.C.3: Results from the comparison in terms of loss. $\overline{A_s(\mathbf{R}_j, \mathbf{R}_j^*)}$ and $\overline{A_s(r(\boldsymbol{\mu}_j), \mathbf{R}_j^*)}$ are averages over the $N = 50$ users.

	θ	BM $D(\boldsymbol{\rho}, \boldsymbol{\rho}^*)$	BTI $D(r(\boldsymbol{\mu}), \boldsymbol{\rho}^*)$	BM $\overline{D(\mathbf{R}_j, \mathbf{R}_j^*)}$	BTI $\overline{D(r(\boldsymbol{\mu}_j), \mathbf{R}_j^*)}$
$M = 0.5n(n-1)/2$	0.05	1.81	3.76	9.29	14.16
	0.1	2.69	3.82	12.24	15.79
	0.15	3.12	5.74	14.87	16.43

Table 3.C.4: Results from the comparison in terms of posterior distance. $\overline{D(\mathbf{R}_j, \mathbf{R}_j^*)}$ and $\overline{D(r(\boldsymbol{\mu}_j), \mathbf{R}_j^*)}$ are sample averages.

	θ	BM $A(\boldsymbol{\rho}, \boldsymbol{\rho}^*)$	BTI $A(r(\boldsymbol{\mu}), \boldsymbol{\rho}^*)$	BM $\overline{A_s(\mathbf{R}_j, \mathbf{R}_j^*)}$	BTI $\overline{A_s(r(\boldsymbol{\mu}_j), \mathbf{R}_j^*)}$
$M = 0.7n(n-1)/2$	0.05	0.65	5.72	15.34	29.17
	0.1	0.58	5.34	24.18	39.81
	0.15	1.75	8.59	35.98	52.28

Table 3.C.5: Results from the comparison in terms of loss. $\overline{A_s(\mathbf{R}_j, \mathbf{R}_j^*)}$ and $\overline{A_s(r(\boldsymbol{\mu}_j), \mathbf{R}_j^*)}$ are sample averages.

In all the cases considered, BM outperforms the BTI model, and often with a large margin. In itself, the order of the outcomes is not surprising since the data were simulated by the BM model. For both methods, larger M and smaller θ increase the precision of the estimates. Next, we perform a comparison based on data simulated with the BTI.

	θ	BM $D(\boldsymbol{\rho}, \boldsymbol{\rho}^*)$	BTI $D(r(\boldsymbol{\mu}), \boldsymbol{\rho}^*)$	BM $\overline{D(\mathbf{R}_j, \mathbf{R}_j^*)}$	BTI $\overline{D(r(\boldsymbol{\mu}_j), \mathbf{R}_j^*)}$
$M = 0.7n(n-1)/2$	0.05	0.58	4.29	7.29	12.13
	0.1	0.55	3.75	10.06	14.65
	0.15	1.29	5.15	13	16.53

Table 3.C.6: Results from the comparison in terms of posterior distance. $\overline{D(\mathbf{R}_j, \mathbf{R}_j^*)}$ and $\overline{D(r(\boldsymbol{\mu}_j), \mathbf{R}_j^*)}$ are sample averages.

Comparison: data generated from BTI

We then perform a second experiment:

- Simulate 3 datasets with BTI.
 - Fix $n = 10$, $N = 50$.
 - Fix $\boldsymbol{\mu}^* = \mathbb{E}(\boldsymbol{\mu}|\text{data})$, i.e. equal to the posterior mean of $\boldsymbol{\mu}$ from experiment (e1) of the previous section;
 - Fix $a_{\boldsymbol{\mu}}^* = 0.8, 1, 1.2$;
 - Sample $\boldsymbol{\mu}_1^*, \dots, \boldsymbol{\mu}_N^* \sim \text{Gamma}(a_{\boldsymbol{\mu}}^*, a_{\boldsymbol{\mu}}^*/\boldsymbol{\mu}^*)$;
 - Set number of pairwise comparisons per user equal to the ones in experiment (e1);
 - Simulate the pair comparison data from the BTI.
- Compare BM and BTI with the same measures as before.

	$a_{\boldsymbol{\mu}}$	BM $A(\boldsymbol{\rho}, r(\boldsymbol{\mu}^*))$	BTI $A(r(\boldsymbol{\mu}), r(\boldsymbol{\mu}^*))$	BM $\overline{A_s(\mathbf{R}_j, r(\boldsymbol{\mu}_j^*))}$	BTI $\overline{A_s(r(\boldsymbol{\mu}_j), r(\boldsymbol{\mu}_j^*))}$
$M = 0.5n(n-1)/2$	0.8	17.52	15.9	54.02	54.74
	1	7.1	6.23	47.98	56.68
	1.2	8.63	10.52	49.76	53.46

Table 3.C.7: Results from the comparison in terms of loss. $\overline{A_s(\mathbf{R}_j, r(\boldsymbol{\mu}_j^*))}$ and $\overline{A_s(r(\boldsymbol{\mu}_j), r(\boldsymbol{\mu}_j^*))}$ are sample averages.

	$a_{\boldsymbol{\mu}}$	BM $D(\boldsymbol{\rho}, r(\boldsymbol{\mu}^*))$	BTI $D(r(\boldsymbol{\mu}), r(\boldsymbol{\mu}^*))$	BM $\overline{D(\mathbf{R}_j, r(\boldsymbol{\mu}_j^*))}$	BTI $\overline{D(r(\boldsymbol{\mu}_j), r(\boldsymbol{\mu}_j^*))}$
$M = 0.5n(n-1)/2$	0.8	9.44	8.9	17.12	18.01
	1	6.11	5.3	16.2	17.2
	1.2	5.79	7.11	16.62	17.47

Table 3.C.8: Results from the comparison in terms of posterior distance. $\overline{D(\mathbf{R}_j, r(\boldsymbol{\mu}_j^*))}$ and $\overline{D(r(\boldsymbol{\mu}_j), r(\boldsymbol{\mu}_j^*))}$ are sample averages.

We see from the tables that BTI performs almost always better in terms of the estimation of the global consensus parameter, while BM does always better in terms of the individual rankings.

We want to point out that the data simulated in this manner have much more mistakes (as measured with the estimated θ parameter of the BM model) with respect to the data we consider in our paper. This explains the much worse results of the two methods in this section.

It is worth pointing out that data simulated from the BTI model usually have many more orderings in the pairwise preferences ‘switched around’ than would be the case in a BM model.

Possible future applications of the BTI

A typical application of the BT model is the analysis of sport data, where teams/players repeatedly compete with each other. BT is used by the World Chess Federation to rank players (see the ELO system, Elo, 1978). A benchmark dataset used to illustrate the potentiality of the BT model is typically the NASCAR (Hunter 2004, Caron and Doucet 2012), which collects the results of the stock car racing in the US. BT is usually applied to one season’s results, thus leading to a final ranking of the drivers (or cars). We could apply the BTI model to this dataset, where each year of the NASCAR data is considered as a user, with an individual ranking μ_j (the ranking of that year), and a global ranking μ (across years) that can be inferred.

Comments

We developed the model proposed by the anonymous referee, and showed its suitability for data in the form of repeated pairwise comparisons performed by each user. We also compared the performance of the BTI with our proposed method on data where each user only performs a limited number of comparisons without repetitions, so that not all pairs of items are compared by each assessor. With this kind of data, it is not satisfied the strong connection condition (Ford 1957), which guarantees the existence and uniqueness of the MLE of the BTI parameters (see also Section 1.2).

In contrast, the model introduced in this Chapter is specifically designed for the description and analysis of heterogeneous data arising from different users performing a limited number of pairwise comparisons. The results from such comparisons are assumed

to be mostly consistent with their individual preference profiles, each such profile forming a linear ordering. On the other hand, the model allows for the possibility of occasional mistakes, in which the preference ordering in a pairwise comparison is reversed, and this can ruin their logical transitivity. In our musicology application (Chapter 4), considering the users individually is important, as different users may well disagree (on what their latent complete rankings of the sounds would have been, had they been reported, in a differently designed experiment, in full), but each of them can be expected to be internally consistent with the requirement of transitivity. The small/moderate amount of inconsistencies in such individual assessments is then in our case covered by either the binomial or the logistic model.

Chapter 4

An application to Electroacoustic music data

This chapter is devoted to the motivating application of this thesis. We consider data coming from an experiment where people were asked to hear a series of two different abstract sounds, and to tell which one was perceived as more human. The data consist of pairwise preferences, and show many non-transitive patterns. The cohort of listeners who took part in the experiment had varying backgrounds, ranging from musicologists to university students. Therefore, we expected listeners to cluster into groups, sharing different opinions about the degree of human causation behind the sounds. It appeared then natural to apply to these data the Bayesian Mallows model for non-transitive pairwise comparisons of Chapter 3. In particular, we apply the mixture model of Section 3.1.5, in order to account for the heterogeneity of the cohort. In addition to the grouping of the listeners around the shared consensus rankings, our method enables to study the association between the individual listeners' rankings and their own musical experience and background. The results are interesting for composers and sound designers, whose aim is to understand how human performance expression can be communicated through audio, leading to computer generated sounds appearing more life-like.

This chapter contains joint work with Natasha Barrett, Valeria Vitelli, Elja Arjas, and Arnaldo Frigessi and is based on [Crispino et al. \(2017\)](#) and [Barrett and Crispino \(2017\)](#).

Outline

Some sections of this chapter are mainly written by Natasha Barrett, and are thus very specialized in their language. Indeed, they will appear in a musicology journal. We decided to report here the whole work for completeness and for the interested reader. The introduction of this thesis reports a self-contained summary of the motivations behind this study, along with a short explanation of the method and the test procedure. The reader not interested in the details can skip Sections 4.1-4.3, and jump directly to Section 4.4, that is the main statistical contribution of this chapter.

We start, in Sections 4.1 and 4.2, by introducing the music problem and the related work. In Section 4.3 we explain the technical details regarding the way sounds were generated, as well as the test procedure of the experiment, which was completely designed by the authors. In Section 4.4 we present the results we obtained by applying the mixture model of Section 3.1.5 on the data at hand. We conclude in Section 4.5 with a short discussion of the relevance of this study and future developments.

4.1 Introduction

Music and motion are undoubtedly connected. Not only does music makes us move, but musical parameters have been shown to simulate listeners' imagined images of motion. Investigations into how listeners associate changes in sound with physical space and their bodies suggest a variety of connections. Simple connections include how temporal features (e.g. tempo or attack rate) are associated with speed or velocity, and how changes in pitch are associated with spatial ascent and descent. Studies also reveal more complex connections, where changes in one domain may stimulate changes in one or more different domains, for example that a crescendo, rather than a pitch rise, may stimulate upwards gestures (Eitan and Granot 2006). Such studies imply that the theory of embodied cognition is a central consideration. Embodied cognition is the theory of mentally re-coding sound into multi-modal gestural images involving a re-enactment of whatever we perceive (Godøy 2006). Developing the discussion, Leman (2012), suggests that the body is a mediator between our environment and our personal experience, through which we accumulate a repertoire of gestures and gesture/action consequences. Directly relevant to our acousmatic musical discourse, Godøy (2010) clarifies embodied cognition, proposing that it involves 'our capacity for having internal images of the world, as somehow originating

in, but not necessarily truthfully reflecting external experience, because bits and pieces from lived experience may be recombined in novel and/or fictional ways'. This process described by Godøy is aligned with the approach that some composers take when creating sounds and musical structures.

It is with this background that we can discuss the impact of 3-D sound spatialization on listeners' understanding of human agency, and further, how a sounding spatial representation of the non-sounding physical movement that lies behind the causation of the sound, can carry this information. Although composers and musicologists have at length discussed spatial information in a musical context, as far as we are aware, this study is the first to test for the influence of sound spatialization on how listeners may hear human agency.

4.2 Background studies

An understanding of the cross-model interactions in the perception of spatial sound can inform our own study. The ways in which human movement may relate to the organization of sound can be informed by 'sound tracing' experiments (Godøy et al. 2006) where subjects use their bodies to spatially describe what they hear. When considering the whole body, a study by Pedersen and Alsop (2012) showed that many sound stimuli were associated with an expected bodily response and a consistent interpretation of agency in the sound. For example, 'floating' sounds would stimulate 'floating' bodily movements. However, for some stimuli, agency in the participant was reversed. For example a 'punching' sound, rather than simulating a punch action, instead stimulated the response of being punched.

In a study by Marentakis and McAdams (2013) the authors conducted a number of tests to ascertain whether a performer's own gestures assisted in the identification of motion trajectories, and whether some trajectories are easier to identify than others. They also investigated the effect of congruent and incongruent audio-visual information. Although a detailed and relevant study in relation to our own work, their choice of spatialization method casts a question of the results. The authors chose to use Vector Base Amplitude Panning, VBAP, (Pulkki et al. 2001) to spatialize a variety of trajectories over eight loudspeakers. VBAP is a panning method, which is suitable for simple trajectories, but without the addition of advanced processing, it is not possible to render perceptually

clear trajectories other than panning at the perimeter of the loudspeaker array. Further, use of just eight loudspeakers offers a low angular resolution, far lower than that of our auditory perception¹ as well as being insufficient to capture the changes in spatial motion tested for. Although not possible to draw conclusions from their study concerning listening alone, in their experiments involving bimodal feedback (an interaction between sound and sight), congruent information was seen to improve performance. If we assume that the listeners were not able to clearly hear the differences between auditory spatial features, the added visual stimuli appear to have convinced them to hear information that was absent.

In another study concerning how auditory-visual cues effect spatial sound streaming (or the segregation of sound in space), [Shestopalova et al. \(2015\)](#) showed that movement-congruent visual cues did not necessarily strengthen the effects of spatial separation. They also conclude that the congruency between auditory and visual stimuli may use mental resources that could have been utilized for more accurate auditory processing, supporting models of modality-specific competition for perceptual awareness.

More generally, we find studies that assess the effect of bi-modality on the subjective experience of the music. Although many studies demonstrate that we primarily use visual information when making judgements about music performance, for example [Tsay \(2013\)](#), cross-modal interactions may involve a more complex network of connections that we may have initially assumed. [Vines et al. \(2006\)](#) demonstrated three contrasting scenarios: an independence of information transmitted through the visual and auditory domains; that an experience of tension (emotion) and phrasing (structure) could be enhanced by bimodal cues; and that the addition of visual information can dampen the intensity of emotional response.

In our own work, we can summarize that, (a) visual information should be removed if listeners are to successfully engage in the challenging task of spatial listening, (b) that an accurate and robust spatialization method is required when creating the test stimuli, and (c) we can anticipate a relationship between agency in the sound and listeners' understanding, but that the mapping between the two may not appear immediately straight forward.

¹In the horizontal plane, our spatial discrimination has been tested to occupy a range of between 0.75-10 degrees depending on the source angle in relation to the listening direction ([Blauert 1997](#)).

4.3 Aim and method

The aim of this experiment was to investigate the role of sound spatialization in listeners' mental representation of human agency when they hear non-visual sound. Listeners' linguistic descriptions of what they hear are notoriously inconsistent, despite often meaning the same. The test we designed was therefore intended to avoid the need for a descriptive language.

We considered allowing listeners to allocate each sound a score, indicating how strongly each evoked human agency in relation to the other sounds. However, we already knew (i) that listeners would span a large range of spatial audio skills, (ii) that the tests would be challenging, and (iii) that it would be ideal to ascertain the degree of certainty in the results and to detect any self-contradictions. For these reasons, we decided to custom design a pairwise test, which is often the preferred experiment when differences between items are small (David 1963).

In particular we chose the following setting,

- $n = 12$ test stimuli were paired into all possible $n(n - 1)/2 = 66$ combinations.
- $N = 46$ listeners spanning a broad range of ages (21-65 years) and musical abilities, were presented with $T_j = 30$ pairs of sounds (which is 45% of the total number of possible pairs out of 12 stimuli). This choice was motivated by the fact that evaluating all possible 66 pairs of sounds could have exceeded the listeners' attention span.
- The pairs were chosen randomly and independently for each user.
- The order in which the sounds were played was randomized.

When designing the experiment, it was necessary to find a balance between n the number of test stimuli, T_j the number of pairs of stimuli that each listener j needed to evaluate, and N the number of participants that could be obtained from the local environment. Ideally, the list of test stimuli would have consisted of the three spatial variations, each sonified with all permutations of pitch and volume-2 variation. This would however have resulted in four more stimuli (for a total of $n = 16$), with a knock-on effect of almost doubling the number of pairs, which would have been $n(n - 1)/2 = 120$. Moreover, in such a case also the difficulty of the test would have increased, since the differences between the stimuli would have been smaller and more difficult to hear. We

then opted for 12 stimuli, and fixed the number of pairs that each listener evaluated at 30, mainly because we believed that 30 is a reasonable trade-off between the amount of information we needed to obtain meaningful results, and the length of the test which had to be kept short enough so as not to generate confusion in the listeners. With different settings, the number of participants would have to be increased significantly.

4.3.1 Creating the test stimuli

The test stimuli were designed considering spatialization method, sound source and motion source for the spatial trajectory, each of which are connected in some way. For example, sine-tones are spatially vague regardless of precision in spatialization method, while a clearly recognizable source, such as that of running footsteps on gravel, will carry extra information biasing listeners' spatial interpretation. The following sections explain these three sides of the test stimuli.

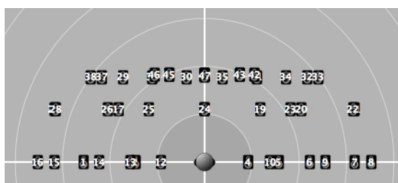
Spatialization method

To synthesize the spatial movement, we use a method called higher-order ambisonics, HOA, (Daniel and Moreau 2004), which is a way of synthesizing 3-D sound over a loudspeaker array. HOA involves a two-step process of spatially encoding the information irrespective of the loudspeaker array, and then applying the appropriate decoder for the array to be used. This approach creates an accurate projection of spatial information, especially for a centrally located listener. We applied a 6th order 3-D spatial encoding, decoded over the 47-loudspeaker 3-D array at the motion capture lab at the Department for Musicology, University of Oslo (Figure 4.1).

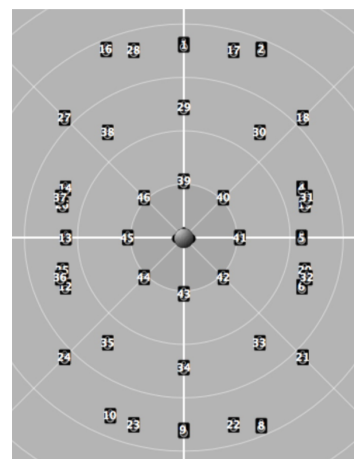
The decoder used was the Max-rE dual-band energy preserving decoder with a cross over frequency set to 400 Hz (Zotter et al. 2012). This method had been tested in a previous project and was found to be the best option for the available loudspeaker system (Barrett 2016). Although near-field coded higher-order ambisonics (Favrot and Buchholz 2012) can in theory recreate the sensation of changes in proximity resulting from variations in the curvature of the approaching wave-front, based on the author's previous experience this method is unsatisfactory in a practical application, which requires the use of regularization functions, found to have an impact on the reproduced sound field, explained briefly in Carpentier et al. (2017). Instead, variations in proximity were projected via three sets of perceptual cues: (a) by changing the relative weight of the



(a) 47-loudspeaker 3-D array at the motion capture lab at the Department for Musicology, University of Oslo.



(b) 47-loudspeaker 3-D array modelled from the side, rotated 45 degrees, visualised in IR-CAM's spat.viewer (part of the Spat package, IRCAM 2017).



(c) Speaker layout from above.

Figure 4.1: Details of the motion capture lab at the Department for Musicology, University of Oslo.

spherical harmonic components: when close-up, the sound is heard as larger and more enveloping, when further away it is heard more as a point source, (b) by changing the gain of the source in relation to distance, and (c) by changing the high frequency content of the source in relation to distance. All three modulations correlate with our perceptual understanding of proximity variations in the real-world. The Doppler effect (which is a pitch shift resulting from a sound moving at speed in relation to the listening position) was not used. In our everyday experience, motorized vehicles are the most common source of Doppler shifts, which would add an inappropriate layer of connotation to the test sounds. Reverberation was also avoided so as to not risk the addition of inappropriate source implications.

Motion sources

To ‘hear’ human agency, we need to somehow ‘hear’ human movement. Normally we hear the result of human movement, and not the movement itself. Therefore, to create the test sounds, it is first necessary to capture movement, or describe it, as spatial data. Human motion archetypes are those with which we are most familiar in terms of our own bodies, and serve as a natural starting point. These archetypes are actions produced by, to name few examples, a swing of the arm, swaying the head and torso, turning, throwing, catching, hitting, the drumming of fingers, punching, stroking, jumping or running. We

can refine these movements and consider motion archetypes that commonly stimulate objects of sound production, including classical musical instruments. These archetypes will include hit, scrape, push, pull, blow, and a variety of smaller and larger motoric variations on these themes. For the motion source of our experiment, the articulation of a music instrument was appropriate: the sound-causation is underpinned by logical Newtonian laws combined with biomechanics, while the spatial domain is constrained within a volume suitable for study.

Motion capture

In a recording session, different performers executed various motion archetypes on their instruments: a cellist played a variety of bowed articulations; a percussionist articulated a cymbal with his hands, and for a non-musical contrast, one person threw a tennis ball. 3-D motion data from these performances was captured using the Qualisys optical motion-capture system and eight Oqus 300 cameras. Passive markers were placed at critical locations over the performers bodies, and motion data was recorded at a rate of 250 Hz with a spatial resolution of less than a mm. The camera system tracks the location of each marker, recording a dataset of 3-D coordinates at intervals of 4 ms. The temporal and spatial resolution was important so as to capture micro activity. Although we may not see micro movements when watching another person, these movements are understood in our own bodies, and can further be made audible when sonifying data that has captured these movements. After assessing all recordings, a cellist bowing a single down-bow action over a ‘double-stop’ was chosen as the motion archetype, using only the one marker located on the lower side of the right hand. Listeners would not be expected to identify the origins of this action in terms of a cellist or a cello. Rather, the action embodied ‘push’ (with some friction resistance), changes of direction, changes of speed, a ‘throw’ as the bow leaves the strings, as well as creating a human motion archetype embodying small micro-movements.

Sonification: listening to the data

The data was sonified so that the spatial trajectory could be heard. Sonification is a process where data is mapped to sound, and for our data, parameter mapping sonification is the appropriate method, described in chapter 15 of *The Sonification Handbook* (Hermann et al. 2011), using the software Cheddar (Barrett 2016).

It was necessary to make appropriate decisions regarding how the parameters of the data should have mapped to sound parameters, as well as exploring scaling ranges that most clearly revealed the qualities in the data. Cheddar provides a framework designed to draw on the perceptual aspects of our spatial hearing, as well as allow scope for interesting sounding results.

The sonification was designed to create sounds that were unrecognizable as to their acoustic source, so as to avoid listeners attaching human causation by way of identifying non-spatial information. For example, when hearing a recognizable musical instrument, most listeners intuitively connect this with a human performance. With these considerations, the following mapping of data to sound was used:

- Each data point triggered a sound grain. The sound grains were initially identical, and chosen from a spectrally rich source.
- 3-D spatial-data points were mapped to 3-D grain location in a 6th order 3-D ambisonics synthesis. The dimensions of the motion source, which traversed a volume of 0.5m x 0.4m x 0.3m, were scaled up to occupy the width of the listening space, resulting in a sonification occupying a volume of 5 x 4 x 3 meters projected over the loudspeaker of size 8 x 5 x 3 meters (see Figure 4.1). By scaling in this way, smaller spatial motions in the source were more likely to be audible in the sonification. Further, it imposed a more dynamic result, intending to enhance listeners' sense of embodiment (discussed in Section 4.1), and in keeping with the way composers perform spatial ideas in their music.
- Velocity of motion was also mapped to grain duration, where higher velocity data values resulted in longer grains. As the data rate was constant, higher velocities would therefore result in a denser grain overlap and subsequent timbral changes, as well as a volume increase.
- Velocity of motion was mapped to the volume of each grain using the decibel scale, so that a doubling in data values resulted in a doubling of perceptual volume. This added an extra volume variation to the result of increased grain overlap, and will hereby be termed volume-2.
- Vertical movement was mapped to pitch, which has been shown to enhance our intuitive awareness of height, as well as relevant to the way in which the shape of our ears and head filters sound from different directions.

- A fixed attack and decay envelop of 5 ms was added to each grain. Maintaining this short attack and decay envelop regardless of grain size ensured a textural quality conducive to spatial identification.
- The timeline of the data was mapped to the timeline of the sonification.

A number of studies (e.g. [Bigand and Parncutt 1999](#), [Krumhansl 1996](#)) show that many structural features of music contribute to the experience of tension, such as loudness dynamics, note density and harmonic relations. The mapping of motion velocity to grain volume and grain duration results in changes of timbre and loudness, leading to changes in perceived tension which may enhance associations with human agency. In the first sonification, the ranges of each of these parameters were scaled to most clearly enhance features in the data. Likewise, larger scaling ranges were more likely to make audible the micro variations in spatial data based on aural evaluation. 11 more sonifications were made from the one dataset. Each version either suppressed features of the original movement captured in the data, reduced the scaling ranges of the sonification, or both (see [Table 1](#) for a description).

In ambisonics, it is necessary to specify the view point from which the spatial synthesis is calculated. We can think of this as the location of a ‘virtual’ listener inside the data, which will then also be the perspective of a real listener. For all but two test stimuli, the sonifications were made for the real listener located in the centre of the motion mean. One of these two stimuli (S3) was reduced to mono, while the other (S2) was spatialised with the virtual listener on the edge of the spatial domain, as if the motion occurred in front. In all but two cases (S11 and S12), time was treated as the original tempo (one data point triggering a sound grain every 4 ms), where the duration of each sonification was 5 seconds.

Based on the assumption that larger sonification ranges enhance spatial information in the data, S1 should be ranked at the top, while S10 should be ranked at the bottom. We can also speculate that as the added pitch variation serves to enhance vertical motion, yet is not a true part of the 3-D sonification, S7 may be evaluated similarly to S1.

4.3.2 Test procedure

A pilot listening session was carried out on two listeners who were not participating in the final experiment: one experienced in electroacoustic music and one inexperienced.

Both listeners were aware of the aims of the project, were asked to assess whether the sonifications made audible the intended information, whether the proposed questions were clear, and whether the total duration of the test and number of test stimuli was realistic. Based on listener feedback a number of changes were made:

- Sonification mapping ranges were optimized in pitch and volume-2 scaling so as to be sure that the qualities of test stimuli S1 were clearly audible.
- The original text asked the listener to identify ‘human agency’. This term was discussed as too specialized, and instead replaced with the phrase ‘human physical action’.

The 46 listeners completed the test, but one participant was turned down after reporting known hearing loss. The tests were carried out in a darkened black box room. Each listener was located at the centre of the space, which is the most accurate 3-D spatial listening point.

Listeners were presented with the following statement: “*This test investigates the role of sound spatialization in how we may associate sound that we hear, with human physical action.*” They were then told that they would be presented with 30 pairs of short sounds, and for each pair, they should choose the one that, “*most evokes a feeling of human causation or human physical origins*”. They were asked to judge the sound as they experience it in relation to their own body, this being to avoid the listener trying to ascertain a possible real source, which in the pilot was shown to be a rationalization that halted the intuitive process. Listeners were also informed that the sounds were made by sonification (with a brief explanation), that this process results in the sounds appearing abstract and could be experienced as somewhat strange. The test began with a training session, during which the listeners were asked to familiarize themselves with these strange qualities.

When the test began, the test number was displayed on a computer screen, listeners noted their answers on a chart, and were requested to always make a choice even if they found it difficult to decide. They were also allowed to repeat a test pair, but only in sequence and not at the end of the experiment. At the end, they were asked to complete two questionnaires that probed their background musical and spatial-audio experience. One questionnaire resulted in a musical sophistication index score (MSI) and the other rated spatial-audio awareness (SAA). The MSI used was the Ollen Musical Sophistication

Index (OMSI), which is an online survey that tests the validity of 29 indicators of musical sophistication used in published music research literature (Ollen 2006). The SAA, or spatial audio awareness index consisted of five questions as indicators of how aware listeners were of spatial audio regardless of musical background. Such a test does not already exist in the literature and was custom designed for the experiment.

The spoken introduction and training session lasted 4 minutes, the test lasted 16 minutes, and the questionnaires took on average a total of 5 minutes.

4.4 Results

We analyzed the data with the mixture model explained in Chapter 3, Section 3.1.5, with footrule distance. With $n = 12$ sounds we could use the exact the partition function. In the Dirichlet prior for η , we set the hyperparameter $\chi = 20$, which favors high-entropy distributions, thus reflecting our inability to express precise prior knowledge. In the Beta prior for θ , we set the hyperparameters at $\kappa_1 = \kappa_2 = 1$, i.e. the uniform distribution on the interval $[0, 0.5)$, and the hyperparameters of the prior for α at $\gamma = 1$ and $\lambda = 1/10$, as discussed in Chapter 2, Section 2.1.4. We run the MCMC sampler for 10^6 iterations, after a burn-in of $2 \cdot 10^5$.

Separate analyses were performed for $C \in \{1, \dots, 7\}$. In order to choose an appropriate number of clusters, we plot in Figure 4.1 two quantities: on the left, the within-cluster sum of footrule distances between the individual rankings and the consensus ranking of that cluster, $\sum_{c=1}^C \sum_{j:z_j=c} d_F(\mathbf{R}_j, \boldsymbol{\rho}_c)$; on the right, the within-cluster indicator of misfit to the data, $\sum_{c=1}^C \sum_{j:z_j=c} \sum_{t=1}^{T_j} g(\mathcal{B}_{jt}, \boldsymbol{\rho}_c)$. Both these measures were already used in Chapter 2. There appears to be an elbow at $C = 3$, to guide us in the choice of the number of clusters. We decided on $C = 3$, also motivated by the relatively small sample size of the experiment ($N = 46$).

Table 4.1 shows the results for $C = 3$: the maximum a posteriori (MAP) estimates for η and α , together with their 95% HPD intervals, are shown at the top of the table. The table also shows the estimated cluster-specific consensus lists of sounds, estimated by the CP procedure. We observe the differences in the three consensus lists. S1, the stimulus with the most dynamic spatial motion, is on top in cluster 3, but at the bottom in cluster 1; S8, the test stimulus that has maximum spatial details but no volume nor pitch change, is on top in cluster 1, but second to the last in clusters 2 and 3. Finally, S5, the

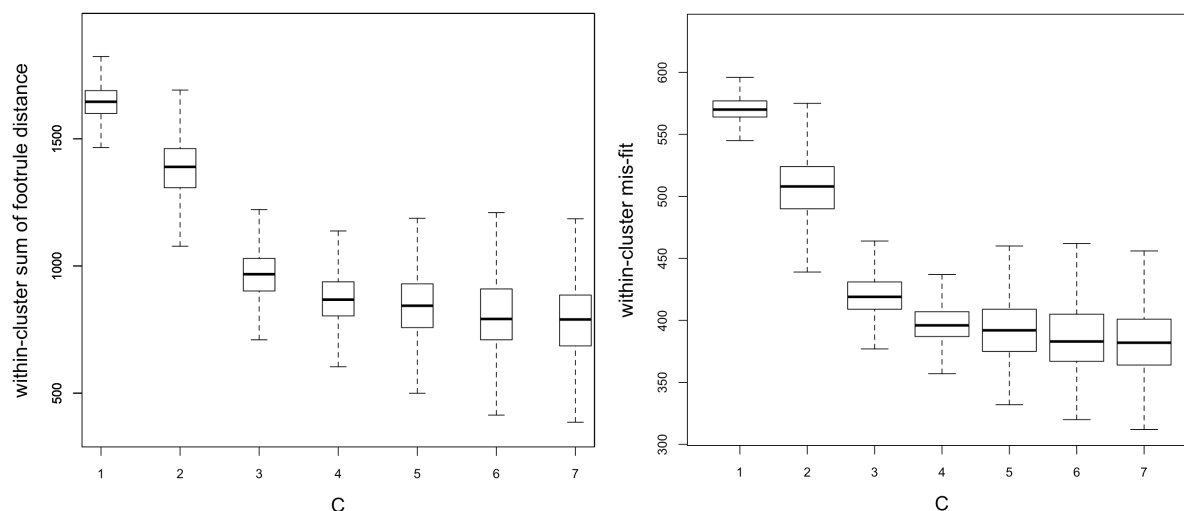


Figure 4.1: *Acousmatic* data. Boxplots of the within-cluster sum of footrule distances between the individual rankings and the consensus ranking of that cluster (left), and of the within-cluster indicator of mis-fit to the data (right), for different choices of C .

stimulus that contains the least movement variation where pitch and volume are naturally suppressed, is ranked third and first in clusters 1 and 2, but towards the bottom of the list in cluster 3. Figure 4.2 shows the heatmap of the posterior marginal probabilities, for each sound, of being ranked as the k -th highest, $k = 1, \dots, 12$. On the x-axis the sounds are ordered according to the CP consensus orderings of Table 4.1. Each cell represents, through colors, the probability that the corresponding sound (on the x-axis) has the rank reported on the y-axis.

Cluster 1	Cluster 2	Cluster 3
$\alpha_1 = 2.66$ (1.14,4.96)	$\alpha_2 = 5.16$ (3.15,9.29)	$\alpha_3 = 5.32$ (3.61,7.66)
$\eta_1 = 0.31$ (0.21,0.41)	$\eta_2 = 0.33$ (0.22,0.43)	$\eta_3 = 0.37$ (0.27,0.48)
S8	S5	S1
S10	S4	S7
S5	S12	S11
S9	S2	S2
S6	S11	S4
S4	S3	S12
S7	S6	S6
S11	S1	S3
S12	S7	S5
S2	S9	S9
S3	S8	S8
S1	S10	S10

Table 4.1: *Acousmatic* data. Sounds are ordered according to the CP consensus ordering, obtained from the posterior distribution of ρ_c , $c = 1, 2, 3$.

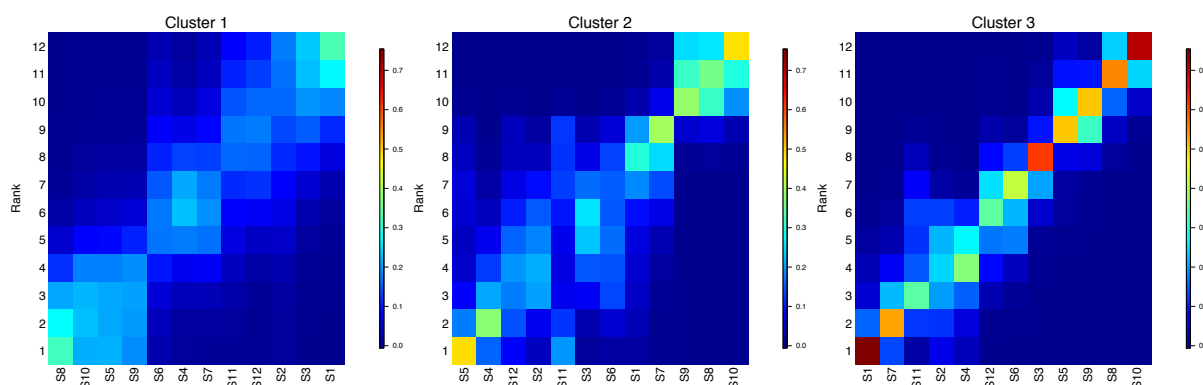


Figure 4.2: *Acousmatic* data. Posterior consensus ranking of the three clusters.

The explanation of the clusters' assignments is as follows:

Cluster 1 (C1)

Top-3 stimuli: S8, S10 and S5.

Bottom-3 stimuli: S2, S3 and S1.

Listeners in C1 found variation in volume or pitch as a negative or distracting feature, while space as an important feature. They rated S8 at the top, a test stimulus that contains all spatial movements but with pitch and volume variations removed. Also, S10, S5 and S9, which were ranked next, lack volume and pitch details. The bottom-4 stimuli, on the other hand, contain maximum pitch and volume variation. S1 is the same as S8 but with pitch and volume variations present. Also S3, being mono-sound, forms a strong contrast to the top ranked S8 that has maximum spatial movement.

Cluster 2 (C2)

Top-3 stimuli: S5, S4, and S12.

Bottom-3 stimuli: S9, S8, S10.

In C2 listeners prioritized pitch and volume variations above spatial variation, and preferred low spatial variation (slower, or more relaxed movements). This is sustained by the top-3 sounds, which feature a low amount of spatial variation, but also correlated pitch and volume, and the bottom-3 sounds which are the same as the top-3 but lack correlated pitch and volume variation. In particular, S5 contains the least movement variation where pitch and volume are suppressed. S4 and S12, although both similar to S1 (which is ranked lower), are each less dynamic in their own way: S12 is played half speed and S4 reflects the global but not smaller movement details. S9 and S10 are the same as S5 and S4, but lack pitch and volume variations, and S8 also lacks pitch and volume variation.

Cluster 3 (C3)

Top-3 stimuli: S1, S7 and S11.

Bottom-3 stimuli: S9, S8 and S10.

C3 consists of listeners who, in their evaluation of the test stimuli, appear to include all spatial cues that adhere to our everyday perception of spatial motion: they prioritize high levels of spatial detail above all other features, and their perception of these details is enhanced by correlated pitch and volume variations. The stimuli with most dynamic spatial motion, enhanced by spatially correlated pitch and volume variations, are in the top-3, while stimuli with the least of these features are in the bottom-3. In particular, S1 is the optimized full spatial representation of the source data, S7 is the same as S1 but with pitch variation removed, while S11 is the same as S1 played 30% slower. S9 and S8 contain significant spatial variation but lack both pitch and volume variations, while S10 contains the least of all information.

We investigate the stability of the clustering in Figure 4.3, that shows the heatmap of the posterior probabilities, for all the listeners (shown on the x-axis), for being assigned to each of the $C = 3$ clusters identified in Table 4.1. Most of the probabilities are concentrated on some particular value of c among the three possibilities, indicating a reasonably precise behavior in the cluster assignments.

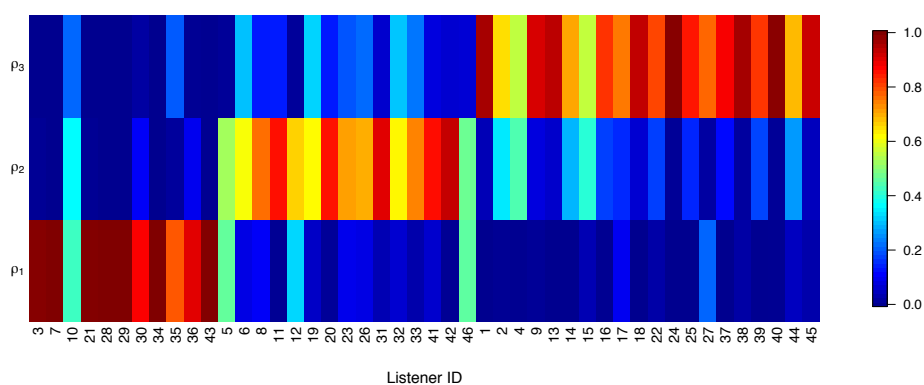


Figure 4.3: *Acousmatic* data. Heatplot, for all the listeners (on the x-axis), of the posterior probabilities of being assigned to each of the three clusters (on the y-axis).

We then computed, fixing these cluster assignments, the marginal posterior probability that each sound is among the top-4 in $\rho_{1:3}$ and in \mathbf{R}_j , $j = 1, \dots, 46$, respectively. The results are shown in Figure 4.4, which is the analog of Figures 2.15, 3.1 and 3.2.

Each heatmap refers to a cluster, C1 (left), C2 (center) and C3 (right), and represents

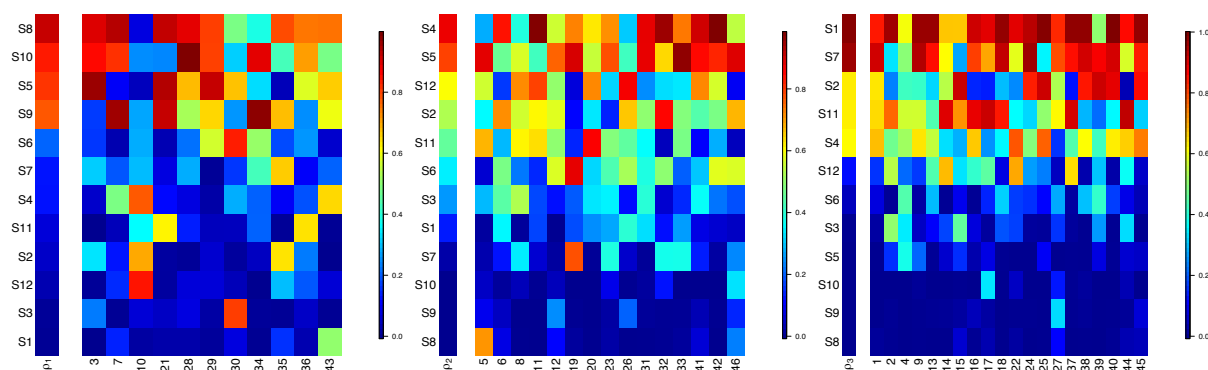


Figure 4.4: *Acousmatic* data. Heatplot of the marginal posterior probabilities for all the stimuli (y-axis) of being ranked among the top-4 for cluster 1 (left), 2 (center) and 3 (right).

the marginal posterior probabilities for each sound (y-axis) being ranked among the top-4 in the consensus of that cluster (first column), and in the individual rankings of listeners in that cluster (remaining columns, users on the x-axis). As Figure 4.4 shows, there is considerable variation in the estimated rankings of the sounds between individual listeners even when they are included in the same cluster. For example, looking at Figure 4.4 left, we see that S8, S10, and S5 have high (> 0.8) posterior probability of being ranked among the top-4 stimuli in the consensus ranking (column 1). However, looking at the estimates for the listeners in C1, we see that the variation is very high: for example, listener 30 (column labelled 30) has a very high posterior probability of ranking S3 and S6 among the top-4 stimuli. This aspect is important for what concerns individual estimates.

Here we investigate whether there is any relationship between listeners' musical background and the test results. In particular, we inspect the relationship between the posterior probability of placing some given stimuli in the top (bottom) ranks and the musical sophistication index (MSI), or the spatial audio awareness index (SAA). Figure 4.5 shows the relationship between listeners' SAA and the probability of S1 (left), S7 (middle), and S1 and S7 jointly (right) being ranked in the top-4 in the individual ranking of the listeners. Recall that S1 was the original sound, while S7 was identical to S1, but without pitch variation. The plot suggests that spatial listening is a skill that is enhanced through training.

In Figure 4.6 we display the relationship between listeners' MSI and the probability of S8 (left), S10 (middle), and S8 and S10 jointly (right), being ranked among the bottom-4 stimuli.

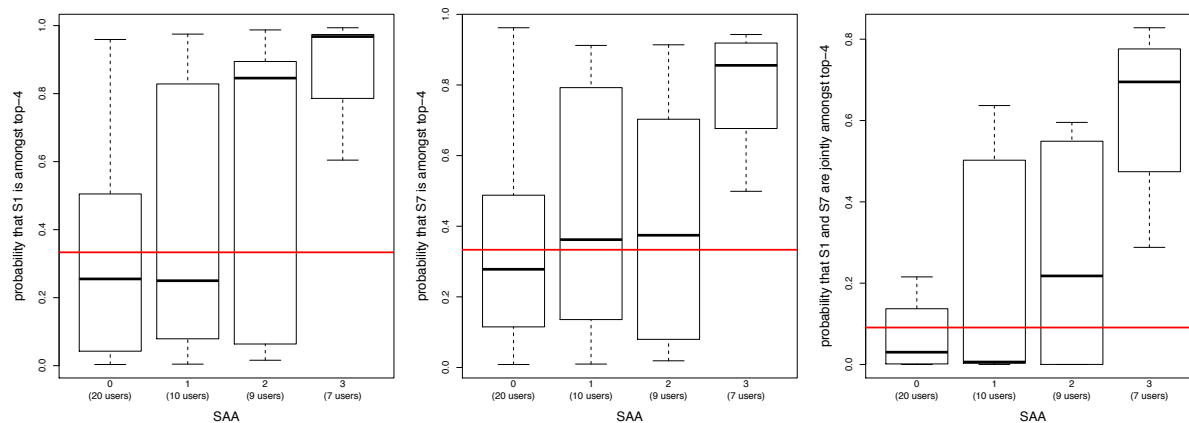


Figure 4.5: *Acousmatic* data. Boxplot of the posterior probabilities for sounds S1 (left), S7 (middle), S1 and S7 jointly (right), of being ranked among the top-4 in the individual ranking \mathbf{R}_j , stratified by the SAA index. The horizontal red line is the threshold in the case of random assignment. The scale of SAA goes from 0 to 3, where 3 is an indicator of awareness of spatial dimension of sounds.

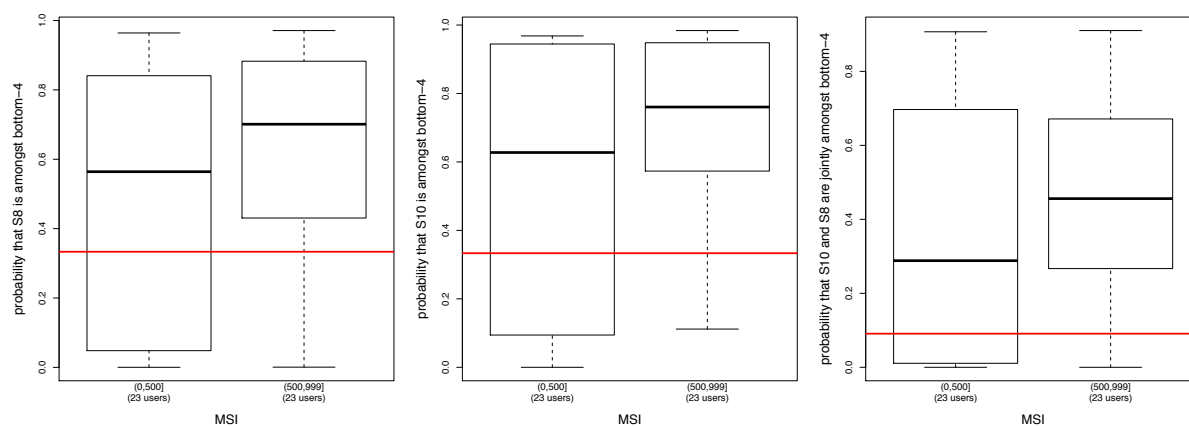


Figure 4.6: *Acousmatic* data. Boxplot of the posterior probabilities for sounds S8 (left), S10 (middle), S8 and S10 jointly (right), of being ranked among the bottom-4 in the individual ranking \mathbf{R}_j , stratified by the MSI index. The horizontal red line is the threshold in the case of random assignment.

Listeners with a score greater than 500 were classified as musically more sophisticated, and those with a score less than 500 as less sophisticated². Both S8 and S10 suppress pitch and volume variations, which are expected to enhance the implication of human causation. These two stimuli are more likely to be ranked among the bottom-4 by listeners with high MSI. This indicates that musically sophisticated listeners, find pitch and volume variations as important qualities for a stimulus to sound human.

The experiment was difficult as expected: 80% of the listeners reported non-transitivities

²As suggested in <http://marcs-survey.uws.edu.au/OMSI/omsi.php>.

in their pair comparisons and only 9 out of 46 listeners were able to stay consistent with themselves. The remaining 37 listeners produced many non-transitivities. Figure 4.7 is the heatplot representing the aggregated matrix of cycle co-memberships: each cell represents the probability that the corresponding sounds on the x-axis and on the y-axis are in a same non-transitive pattern (here called cycle), and thus the probability that they are confused by the listeners. This plot helps in understanding the extent to which a sound is more easily confused with another. For example, S10 and S12, which are very different (see Table 1), appear in the same cycle a small number of times (blue cell), while S8 and S9 (which both lack pitch and volume variation, and have very similar spatial information) appear in the same cycle a large number of times (burgundy cell).

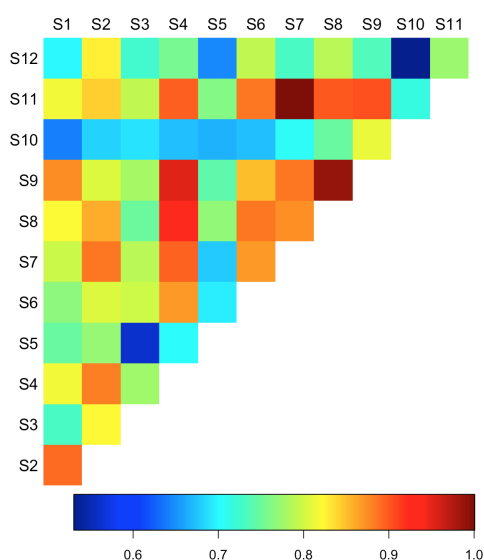


Figure 4.7: Heatplot representing the cycle co-memberships of the sounds in the non-transitive patterns of the data.

4.5 Discussion

The results of the analysis clearly reveal that spatial listening is a skill that is enhanced through experience and personal interest. Also, different groups of listeners latch on to different types of information in their experience of human agency. Yet, the identification of three clusters, as well as the uncertainty for each ranking, indicates that answering the question as to whether sound spatialization can suggest human agency is far from straight

forward. If we were to assume that human agency is projected by sound spatialization mimicking physical actions, and enhanced by volume and pitch modulation, S1 would optimally capture human agency through sound spatialization and S10 would fail in this respect. This trend is strongly identified for cluster 1 only (although cluster 2 also ranks S10 last). However, common to cluster 2 and 3 - which together account for 35 out of the 46 listeners - is a preference for the original, complete spatial movement. From the rank of S2 we can make assumptions as to the effect of spatial scaling on listeners' choices: in S2 indeed the spatial activity occupies a small spatial zone in front of the listener, more akin to the original source space. For all listeners, S2 occupies an uncertain middle ranking, where the reduction in distance traversed may serve to blur a listener's judgement.

The preference of cluster 2 for slower movements may point out that these listeners find the large and fast variations in pitch and volume-2 distracting, which can indicate that a future study would benefit from smaller variations in the sonification scaling range. Also, the results from clusters 1 and 3 suggest that pitch variation as an enhancement of verticality is an unnecessary addition, which may have served to make the tests trickier or the results less consistent. A further study may therefore choose to remove this aspect of the sonification.

Tesi di dottorato "Bayesian learning of the Mallows ranking model"

di CRISPINO MARTÀ

discussa presso Università Commerciale Luigi Bocconi-Milano nell'anno 2018

La tesi è tutelata dalla normativa sul diritto d'autore (Legge 22 aprile 1941, n.633 e successive integrazioni e modifiche).

Sono comunque fatti salvi i diritti dell'università Commerciale Luigi Bocconi di riproduzione per scopi di ricerca e didattici, con citazione della fonte.

Chapter 5

The Bayesian Mallows model with Cayley distance

In this chapter, we provide the specialization of the Bayesian Mallows model of Chapter 2 when the metric on the space of permutations is the Cayley distance.

We start in Section 5.1 by presenting the definition and properties of the Cayley distance. In Section 5.2, we give the main equations and properties of the Mallows model with Cayley distance, and outline a strategy to make Bayesian inference on this model (Section 5.2.1), also investigating some choices for the prior density over the consensus ranking. We then outline the adaptation of the MCMC of Chapter 2 for this case (in Section 5.2.2), where we introduce a new symmetric proposal for the consensus parameter of the MMC, which is particularly suited for this model.

This chapter is ongoing work.

5.1 The Cayley distance and its properties

As already mentioned in Section 1.1.3, the Cayley distance $d_C(\boldsymbol{\rho}, \boldsymbol{\sigma})$ between two permutations $\boldsymbol{\rho}, \boldsymbol{\sigma} \in \mathcal{P}_n$ counts the minimum number of swaps required to convert $\boldsymbol{\rho}$ into $\boldsymbol{\sigma}$. It is right-invariant (see definition 1 in Section 1.1.3), from which it follows that

$$d_C(\boldsymbol{\rho}\boldsymbol{\rho}^{-1}, \boldsymbol{\sigma}\boldsymbol{\rho}^{-1}) = d_C(\mathbf{1}_n, \boldsymbol{\sigma}\boldsymbol{\rho}^{-1}) := d_C(\boldsymbol{\sigma}\boldsymbol{\rho}^{-1}), \quad (5.1)$$

where $\mathbf{1}_n = (1, 2, \dots, n)$ is the identity permutation.

The Cayley distance, contrarily to most of the other distances considered in this thesis, is also left-invariant (and thus bi-invariant), meaning that it does not change

if renumbering the ranks. This property is somewhat counterintuitive when dealing with human preferences: the preferred item (ranked 1st), is different from the least preferred one (ranked n^{th}), say, which implies that the distance between them is large. The previous reasoning does not hold for Cayley distance, which indeed is a measure of pure disorder, rather than being a spatial distance like footrule or Spearman (Marden 1995). For this reason the Cayley distance is not used with preference data, but rather in applications concerning cryptography, genomics, or random number generators.

The bi-invariance property of the Cayley distance is intimately connected with the notion of cycles in a permutation.

Definition 4. Cycle of a permutation ρ . A cycle in a permutation ρ is an ordered set $\{i_1, \dots, i_k\} \subseteq \{1, \dots, n\}$ such that $\rho(i_1) = i_2, \rho(i_2) = i_3, \dots, \rho(i_k) = i_1$.

A way of representing a permutation is indeed through the cyclic notation, in which it is factorized into the set of disjoint cycles. For example, the permutation $\rho = (6, 3, 2, 1, 4, 5)$ has cyclic decomposition $(1\ 4\ 5\ 6)(2\ 3)$. Notice here the difference in notation: the elements of a permutation are separated by commas, while the elements of cycles are not.

It can be proven the following identity, that holds for each permutation $\rho \in \mathcal{P}_n$

$$d_C(\mathbf{1}_n, \rho) := d_C(\rho) := n - C(\rho), \quad (5.2)$$

where $C(\rho)$ is the number of cycles of ρ . In other words, the number of swaps to convert a given ranking, ρ , into the identity permutation $\mathbf{1}_n$ - which is exactly the definition of Cayley distance between ρ and $\mathbf{1}_n$ - equals the length of the ranking minus its number of cycles. Every permutation uniquely decomposes into the product of disjoint cycles, but a cyclic decomposition may correspond to many permutations. All permutations that have the same disjoint cycle decomposition form a conjugacy class.

Given $\rho, \sigma \in \mathcal{P}_n$, it holds

$$d_C(\rho, \sigma) = d_C(\mathbf{1}_n, \sigma\rho^{-1}) = n - C(\rho\sigma^{-1}). \quad (5.3)$$

An important property of the Cayley distance $d_C(\rho)$ is that it can be decomposed into a sum of $n - 1$ independent terms X_i , $d_C(\rho) = \sum_{i=1}^{n-1} X_i(\rho)$, where $X_i(\rho) = 0$ if i is the largest item in its cycle in ρ , $X_i(\rho) = 1$ otherwise (Feller 1968).

5.2 The Mallows model with Cayley distance

In this chapter we express the scale parameter of eq. (1.6) as $\theta = \alpha/n$, for simplifying the notation, thus leading to the following form of the Mallows density,

$$P(\mathbf{R} | \theta, \boldsymbol{\rho}) := \frac{1}{Z(\theta)} \exp[-\theta d(\mathbf{R}, \boldsymbol{\rho})]. \quad (5.4)$$

The Mallows model with Cayley distance, henceforth referred to as MMC, has the appealing property that the partition function has a closed form, due to [Fligner and Verducci \(1986\)](#):

$$Z(\theta) = \sum_{\mathbf{R} \in \mathcal{P}_n} e^{-\theta d_C(\mathbf{R}, \boldsymbol{\rho})} = \prod_{i=1}^{n-1} (1 + ie^{-\theta}) \quad (5.5)$$

Remark 3. Notice that 5.5, can equivalently be written as $Z(\theta) = e^{-\theta(n-1)}(1 + e^\theta)_{n-1}$, where $(x)_n = \prod_{i=1}^n (x + i - 1)$ denotes the Pochhammer symbol.

Proof. $\prod_{i=1}^{n-1} (1 + ie^{-\theta}) = e^{-\theta(n-1)} \prod_{i=1}^{n-1} (e^\theta + i) := e^{-\theta(n-1)}(1 + e^\theta)_{n-1}$. \square

The density of a ranking $\mathbf{R} \in \mathcal{P}_n$ in the MMC can then be written as follows,

$$P(\mathbf{R} | \theta, \boldsymbol{\rho}) = \frac{e^{-\theta[n-C(\mathbf{R}\boldsymbol{\rho}^{-1})]}}{e^{-\theta(n-1)}(1 + e^\theta)_{n-1}} = \frac{e^{-\theta[1-C(\mathbf{R}\boldsymbol{\rho}^{-1})]}}{(1 + e^\theta)_{n-1}}. \quad (5.6)$$

Given a sample $\mathbf{R}_1, \dots, \mathbf{R}_N | \theta, \boldsymbol{\rho} \stackrel{i.i.d}{\sim} \mathcal{M}_C(\theta, \boldsymbol{\rho})$, where \mathcal{M}_C is the density in eq. (5.6), the likelihood is simply

$$P(\mathbf{R}_1, \dots, \mathbf{R}_N | \theta, \boldsymbol{\rho}) = \frac{e^{-\theta \sum_{j=1}^N d_C(\mathbf{R}_j, \boldsymbol{\rho})}}{Z(\theta)^N} = \frac{e^{-\theta[N - \sum_{j=1}^N C(\mathbf{R}_j \boldsymbol{\rho}^{-1})]}}{[(1 + e^\theta)_{n-1}]^N}. \quad (5.7)$$

In Figure 5.1 we boxplot the Cayley distance samples $\mathbf{R}_1, \dots, \mathbf{R}_N | \theta, \boldsymbol{\rho} \stackrel{i.i.d}{\sim} \mathcal{M}_C(\boldsymbol{\rho}, \theta)$ as a function of θ , for different values of n , as stated in the titles.

Maximum likelihood estimation of the MMC is studied in [Irurozki et al. \(2016b\)](#), where the authors propose an exact algorithm based on a branch and bound search in order to estimate the MLE, which is the solution to the following combinatorial optimization problem,

$$\boldsymbol{\rho}_{MLE} = \operatorname{argmax}_{\boldsymbol{\rho} \in \mathcal{P}_n} \sum_{j=1}^N C(\mathbf{R}_j \boldsymbol{\rho}^{-1}).$$

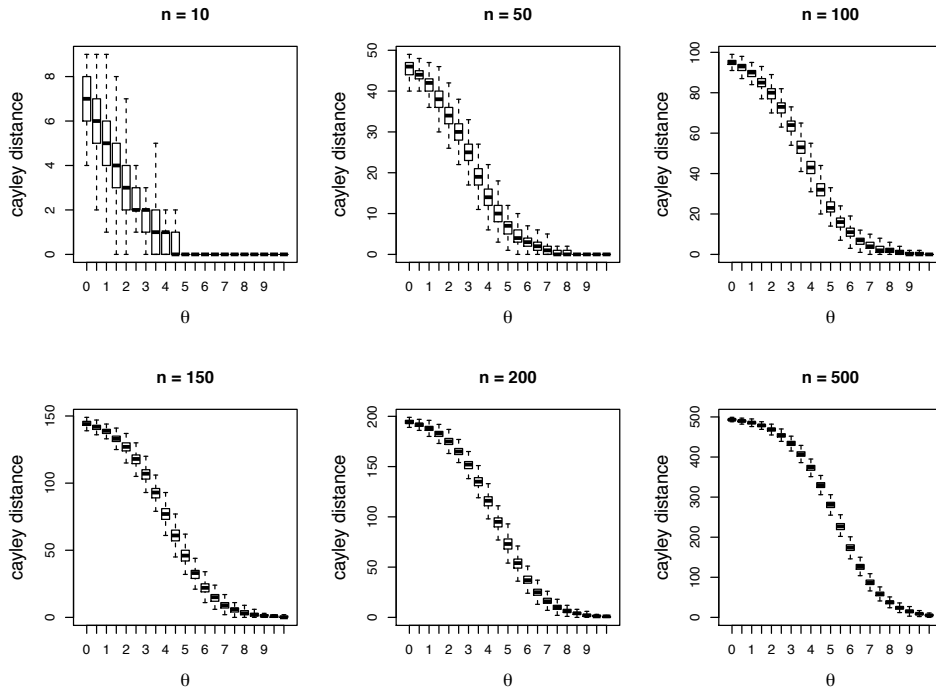


Figure 5.1: The Cayley distance of a sample $\mathbf{R}_1, \dots, \mathbf{R}_N \sim \mathcal{M}_C(\boldsymbol{\rho}, \theta)$ as a function of θ , for different values of n .

As the authors claim, this optimization problem is commonly believed to be NP-complete, also if its computational complexity classification is still an open issue.

5.2.1 Bayesian Learning of the MMC

With Cayley distance, and a prior density $\pi(\cdot, \cdot)$ on the MMC parameters $\boldsymbol{\rho}$ and θ , the posterior distribution is given by

$$\pi(\boldsymbol{\rho}, \theta | \mathbf{R}_1, \dots, \mathbf{R}_N) \propto \frac{\pi(\boldsymbol{\rho}, \theta) e^{-\theta[N - \sum_{j=1}^N C(\mathbf{R}_j \boldsymbol{\rho}^{-1})]}}{[(1 + e^\theta)_{n-1}]^N}. \quad (5.8)$$

We briefly discuss two natural strategies to elicit the prior distribution.

The first road is to assume prior independence among the two parameters, as we did in Chapters 2 and 3. For θ , one possibility is to choose the exponential density, as suggested in Section 2.1.1. An informative prior for $\boldsymbol{\rho}$, as mentioned in Section 2.1.1, could be the Mallows family, that is $\pi(\boldsymbol{\rho}) = \pi(\boldsymbol{\rho} | \theta_0, \boldsymbol{\rho}_0) \propto \exp[-\theta_0 d(\boldsymbol{\rho}, \boldsymbol{\rho}_0)]$, where θ_0 and $\boldsymbol{\rho}_0$ are fixed hyperparameters. If one is able to provide a prior central permutation $\boldsymbol{\rho}_0$, with uncertainty around it, controlled by the hyperparameter θ_0 , then a natural prior for $\boldsymbol{\rho}$ would be a Mallows density with Cayley distance. Such prior distribution assigns

equal probability to all permutations with the same Cayley distance to ρ_0 , that is, to all permutations ρ such that $\rho\rho_0^{-1}$ has the same number of cycles. Given a permutation $\rho \in \mathcal{P}_n$, the possible number of cycles is $C(\rho) \in \{0, \dots, n-1\}$. So, this strategy amounts to set a n possible values to the prior. Linked to this idea, is the interesting proposal of [Gupta and Damien \(2002\)](#), who suggest to elicit a prior which is constant on conjugacy classes, this being is equivalent to assign a priori equal probability to all permutations with the same cyclic structure. Our proposal reduces to the one of [Gupta and Damien \(2002\)](#), if the density assumed on each conjugacy class is the MMC.

A second possibility is to model, through the joint prior, the possible dependency of the two parameters. A proposal is the following hierarchical scheme:

$$\mathbf{R}_1, \dots, \mathbf{R}_N | \theta, \rho \stackrel{i.i.d}{\sim} \mathcal{M}(\theta, \rho)$$

$$\rho | \theta, \rho_0 \sim \mathcal{M}(\theta, \rho_0)$$

$$\theta | N'_0, D'_0 \sim \exp\{-\theta N'_0 D'_0 - N'_0 \ln Z(\theta)\}$$

The posterior of eq. (5.8), would in this special case be given by,

$$\begin{aligned} \pi(\rho, \theta | \mathbf{R}_1, \dots, \mathbf{R}_N) &\propto \pi(\theta) \pi(\rho | \theta) P(\mathbf{R}_1, \dots, \mathbf{R}_N | \theta, \rho) = \\ &= \frac{\exp\left\{-\theta \left[N'_0 D'_0 + d_C(\rho, \rho_0) + \sum_{j=1}^N d_C(\mathbf{R}_j, \rho) \right]\right\}}{[Z(\theta)]^{N'_0 + N + 1}} = \\ &= \frac{\exp\left\{-\theta \left[N'_0 (D'_0 - n - 1) - N - 1 - C(\rho_0 \rho^{-1}) - \sum_{j=1}^N C(\mathbf{R}_j \rho^{-1}) \right]\right\}}{(1 + e^\theta)_{n-1}^{N'_0 + N + 1}}. \end{aligned} \quad (5.9)$$

This means that the joint posterior for the pair (ρ, θ) is as if one had $(N'_0 + N + 1)$ observations, of which:

- N'_0 have average Cayley distance from ρ given by D'_0 ;
- N are the sampled observations, $\mathbf{R}_1, \dots, \mathbf{R}_N$;
- 1 is ρ_0 , which is the hyperparameter of the prior density over ρ .

Thanks to this prior it is possible to elicit two quantities at the same time: by means of the hyperparameters N'_0 and D'_0 , one can provide information about the conjugacy class which is a priori thought more likely; through the hyperparameter ρ_0 , one can give weight to a specific central permutation, that may or may not belong to the same conjugacy class of the previous point.

5.2.2 Algorithm for the MMC

A simple adaptation of the random walk Metropolis-Hastings algorithm for full rankings of Chapter 2 enables us to sample from the posterior density (5.8). We here treat the case of prior independence among θ and $\boldsymbol{\rho}$. The calculations can be easily extended to the case of dependency (that is, to the posterior of eq. (5.9)).

As in Algorithm 1, we iterate between two steps. We first update $\boldsymbol{\rho}$, by sampling $\boldsymbol{\rho}'$ from the following customized proposal.

Definition 5. *Cayley proposal.* Denote the current version of the consensus ranking by $\boldsymbol{\rho}^m$. Let $L^* \in \{1, \dots, n\}$. Sample uniformly an integer l from $U\{1, 2, \dots, L^*\}$. The proposal $\boldsymbol{\rho}'$ is sampled uniformly between the permutations at Cayley distance l from $\boldsymbol{\rho}^m$: $\boldsymbol{\rho}' \sim \mathcal{U}(\{\boldsymbol{\sigma} \in \mathcal{P}_n : d_C(\boldsymbol{\sigma}, \boldsymbol{\rho}^m) = l\})$.

The parameter L^* is the maximum allowed Cayley distance of the proposal from the current value of $\boldsymbol{\rho}$, and is used for tuning the acceptance probability in the M-H step.

The process of simulating $\boldsymbol{\rho}'$ uniformly between the permutations at Cayley distance l from $\boldsymbol{\rho}^m$ is performed in two stages:

- i. Pick, uniformly at random, a permutation $\boldsymbol{\sigma}$ at Cayley distance l from the identity permutation $\mathbf{1}_n$ (function available in the the `PerMallows` R package);
- ii. Set $\boldsymbol{\rho}' = \boldsymbol{\sigma}\boldsymbol{\rho}^m$, since $l = d_C(\boldsymbol{\sigma}, \mathbf{1}_n) = d_C(\boldsymbol{\sigma}\boldsymbol{\rho}^m, \boldsymbol{\rho}^m)$, by bi-invariance of Cayley distance.

The number of permutations $\boldsymbol{\sigma}$ at Cayley distance l from $\boldsymbol{\rho}^m$ equals the number of permutations $\boldsymbol{\sigma}$, s.t. $\boldsymbol{\sigma}(\boldsymbol{\rho}^m)^{-1}$ has $(n-l)$ cycles (see eq. (5.3)), which is given by the unsigned Stirling numbers of the first kind (OEIS sequence A094638 Sloane 2017). As a consequence, the transition probability of the Cayley proposal is symmetric, and given by

$$\begin{aligned} q(\boldsymbol{\rho}' \rightarrow \boldsymbol{\rho}^m) &= q(\boldsymbol{\rho}^m \rightarrow \boldsymbol{\rho}') = \sum_{l=1}^{L^*} P(L=l)P(\boldsymbol{\rho}' \rightarrow \boldsymbol{\rho}^m | L=l)\mathbb{1}(d_C(\boldsymbol{\rho}', \boldsymbol{\rho}^m) = l) = \\ &= \frac{1}{L^*} \sum_{l=1}^{L^*} \frac{1}{S_n^{(n-l)}} \mathbb{1}(d_C(\boldsymbol{\rho}', \boldsymbol{\rho}^m) = l). \end{aligned}$$

The proposed value is then accepted with probability $\eta = \min\{1, a_\rho\}$, where:

$$\begin{aligned} \ln a_\rho &= -\theta \sum_{j=1}^N [d_C(\mathbf{R}_j, \boldsymbol{\rho}') - d_C(\mathbf{R}_j, \boldsymbol{\rho}^m)] - \theta_0 [d_C(\boldsymbol{\rho}_0, \boldsymbol{\rho}') - d_C(\boldsymbol{\rho}_0, \boldsymbol{\rho}^m)] = \\ &= -\theta \sum_{j=1}^N [C(\boldsymbol{\rho}^m \mathbf{R}_j^{-1}) - C(\boldsymbol{\rho}' \mathbf{R}_j^{-1})] - \theta_0 [C(\boldsymbol{\rho}^m (\boldsymbol{\rho}_0)^{-1}) - C(\boldsymbol{\rho}' (\boldsymbol{\rho}_0)^{-1})] . \end{aligned} \quad (5.10)$$

We then update θ by sampling θ' from a log-normal density, $\ln \mathcal{N}(\ln \theta^m, \sigma_\theta^2)$, and accepting it with probability $\eta = \min\{1, a_\theta\}$, where

$$\ln a_\theta = \ln(\theta'/\theta^m) - (\theta' - \theta^m) \left[\lambda + N - \sum_{j=1}^N C(\boldsymbol{\rho} \mathbf{R}_j^{-1}) \right] - N \ln \frac{(e^{\theta'} + 1)_{n-1}}{(e^{\theta^m} + 1)_{n-1}} . \quad (5.11)$$

Algorithm 13: Random walk MH for the MMC

input : $\mathbf{R}_1, \dots, \mathbf{R}_N, \lambda, \boldsymbol{\rho}_0, \theta_0, \sigma_\theta, L, Z(\theta), M, L^*$

output: Posterior distributions of $\boldsymbol{\rho}$ and θ

Initialization: randomly generate $\boldsymbol{\rho}_0$ and θ_0

for $m \leftarrow 1$ to M **do**

Update $\boldsymbol{\rho}$:

 sample $\boldsymbol{\rho}'$ form the Cayley proposal with parameter L^* and centered in $\boldsymbol{\rho}^m$

 compute: $ratio \leftarrow$ equation (5.10) with $\theta \leftarrow \theta^m$

 sample: $u \sim \mathcal{U}(0, 1)$

if $u < ratio$ **then**

 | $\boldsymbol{\rho}^{m+1} \leftarrow \boldsymbol{\rho}'$

else

 | $\boldsymbol{\rho}^{m+1} \leftarrow \boldsymbol{\rho}^m$

end

Update θ :

 sample: $\theta' \sim \ln \mathcal{N}(\ln \theta^m, \sigma_\theta^2)$

 compute: $ratio \leftarrow$ equation (5.11) with $\boldsymbol{\rho} \leftarrow \boldsymbol{\rho}^m$

 sample: $u \sim \mathcal{U}(0, 1)$

if $u < ratio$ **then**

 | $\theta^{m+1} \leftarrow \theta'$

else

 | $\theta^{m+1} \leftarrow \theta^m$

end

end

To save time, after the acceptance step for $\boldsymbol{\rho}$, we store the value of $\sum_{j=1}^N C(\boldsymbol{\rho} \mathbf{R}_j^{-1})$, corresponding to the accepted $\boldsymbol{\rho}$, which is then used in the step for θ . The algorithm computational time depends strongly on n and N : the computation of Cayley distance between n -dimensional permutations, requires increasing time in n (see Figure 5.2), and the larger N is, the larger is the number of times such distance must be calculated.

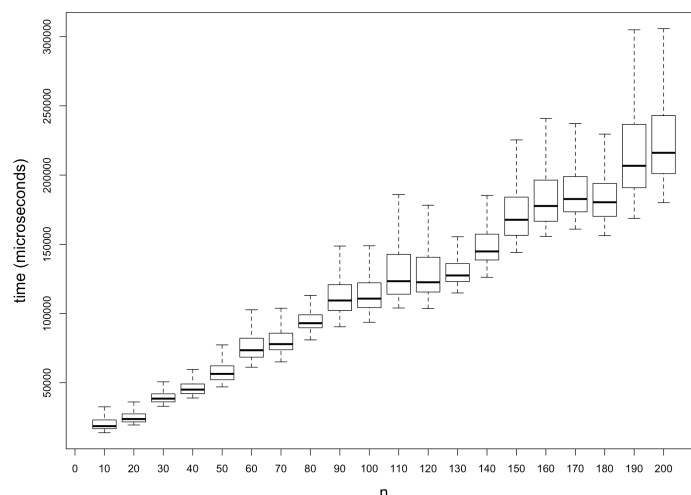


Figure 5.2: Boxplot of the time to compute the number of cycles in a permutation (i.e. the Cayley distance), stratified by the length of the permutation.

5.3 Ongoing work and discussion

We are currently working on applications of the model proposed in this chapter, and exploring its connections with some areas of Bayesian non-parametrics.

For example, a construction that gives rise to the MMC is the Chinese restaurant process, explained in a nutshell below. Imagine a restaurant containing n tables. People arrive at the restaurant and choose a table according to the following scheme: The first person seats at table 1. The second one seats to the right of the first person or at table 2 with probabilities $e^\theta/(1+e^\theta)$ and $1/(1+e^\theta)$ respectively. The $(i+1)$ -st person chooses to seat at an empty table with probability $1/(ie^\theta+1)$, and to the right of one of the previous people, i , randomly and with equal probability $ie^\theta/(ie^\theta+1)$. The final arrangement of the tables is a permutation expressed in cyclic notation.

A second interesting application of the MMC is that the partition given by the cycles in the Mallows model with Cayley distance arises in mathematical population genetics as the Ewens sampling formula (Aldous 1985, Gnedin and Gorin 2016).

Exploring inferential aspects of uses of the MMC in these areas is an envisaged development of the work outlined in this chapter.

Chapter 6

Preliminary results on the conjugate prior elicitation problem

In this chapter we discuss a model extension that we are working on, and give some preliminary results. The extension regards the prior elicitation problem. In particular, we are working on defining the conjugate prior for the parameters of the Mallows model in case of Spearman distance. In Section 6.1, we show that the Mallows model with Spearman distance has sufficient statistics, and in Section 6.2 we outline the first results regarding a conjugate prior for its consensus parameter.

6.1 Sufficient statistic and MLE

As in Chapter 5, let us parametrize the Mallows model (1.6) in the scale parameter $\theta = \alpha/n$, so that, in the case of Spearman distance, it specifies the probability density of a ranking $\mathbf{R} \in \mathcal{P}_n$ as

$$\begin{aligned} P(\mathbf{R}|\theta, \boldsymbol{\rho}) &= \frac{1}{Z(\theta)} \exp \left[-\theta \sum_{i=1}^n (R_i - \rho_i)^2 \right] \propto \exp \left[-\theta \sum_{i=1}^n [R_i^2 + \rho_i^2 - 2\rho_i R_{j_i}] \right] = \\ &= \exp \left[-\theta n(n+1)(2n+1)/3 + 2\theta \sum_{i=1}^n R_i \rho_i \right] \propto \exp \left[2\theta \sum_{i=1}^n R_i \rho_i \right]. \end{aligned} \quad (6.1)$$

Let us assume that the parameter θ is known. In this chapter we only focus on the inference about $\boldsymbol{\rho}$.

Given a sample $\mathbf{R}_1, \dots, \mathbf{R}_N | \boldsymbol{\rho} \stackrel{i.i.d}{\sim} \mathcal{M}_S(\theta, \boldsymbol{\rho})$, where $\mathcal{M}_S(\cdot, \cdot)$ is the density defined in (6.1),

the likelihood is then

$$P(\mathbf{R}_1, \dots, \mathbf{R}_N | \boldsymbol{\rho}) \propto \exp \left[-\theta \sum_{j=1}^N \sum_{i=1}^n (\rho_i - R_{ji})^2 \right] \propto \exp \left(2\theta N \sum_{i=1}^n \rho_i \bar{R}_i \right), \quad (6.2)$$

where $\bar{R}_i = \frac{1}{N} \sum_{j=1}^N R_{ji}$, $i = 1, \dots, n$, is the sample average of the i -th rank.

The previous calculation shows that the sufficient statistic for $\boldsymbol{\rho} = (\rho_1, \dots, \rho_n)$, is given by $\bar{\mathbf{R}} = (\bar{R}_1, \dots, \bar{R}_n)$, and the MLE is the solution to the following maximization

$$\operatorname{argmax}_{\boldsymbol{\rho} \in \mathcal{P}_n} \sum_{i=1}^n \rho_i \bar{R}_i \quad .$$

Denote by $\mathbf{Y}(\bar{\mathbf{R}}) = (Y_1(\bar{\mathbf{R}}), \dots, Y_n(\bar{\mathbf{R}})) \in \mathcal{P}_n$ the rank vector of $\bar{\mathbf{R}}$, that is, $Y_i(\bar{\mathbf{R}}) = \sum_{h=1}^n \mathbb{1}(\bar{R}_h \leq \bar{R}_i)$, $i = 1, \dots, n$. The following proposition shows that $\boldsymbol{\rho}_{MLE} = \operatorname{argmax}_{\boldsymbol{\rho} \in \mathcal{P}_n} \sum_{i=1}^n \rho_i \bar{R}_i = \mathbf{Y}(\bar{\mathbf{R}})$.

Proposition 6. *Let $\mathbf{R}_1, \dots, \mathbf{R}_N | \boldsymbol{\rho} \stackrel{i.i.d.}{\sim} \mathcal{M}_S(\theta, \boldsymbol{\rho})$, and define the vector of sample ranks as $\bar{\mathbf{R}} = (\bar{R}_1, \dots, \bar{R}_n)$, where $\bar{R}_i = \frac{1}{N} \sum_{j=1}^N R_{ji}$. Assume $\bar{R}_i \neq \bar{R}_j$, for each $i \neq j$, and denote by $\mathbf{Y}(\bar{\mathbf{R}}) = (Y_1(\bar{\mathbf{R}}), \dots, Y_n(\bar{\mathbf{R}})) \in \mathcal{P}_n$ the rank vector of $\bar{\mathbf{R}}$, defined as above. Then $\boldsymbol{\rho}_{MLE} = \mathbf{Y}(\bar{\mathbf{R}})$.*

Proof. The following two identities hold by right-invariance (Section 1.1.3, Definition 1):

$$\sum_{i=1}^n \rho_i \bar{R}_i = \sum_{i=1}^n i(\bar{\mathbf{R}} \circ \boldsymbol{\rho}^{-1})_i \quad (6.3)$$

$$\sum_{i=1}^n \rho_i Y_i(\bar{\mathbf{R}}) = \sum_{i=1}^n i(Y(\bar{\mathbf{R}}) \circ \boldsymbol{\rho}^{-1})_i = \sum_{i=1}^n i Y_i(\bar{\mathbf{R}} \circ \boldsymbol{\rho}^{-1}) \quad (6.4)$$

Eq. (6.3) implies that $\hat{\boldsymbol{\rho}}_1 = \operatorname{argmax}_{\boldsymbol{\rho} \in \mathcal{P}_n} \sum_{i=1}^n i(\bar{\mathbf{R}} \circ \boldsymbol{\rho}^{-1})_i$ is such that $(\bar{\mathbf{R}} \circ \hat{\boldsymbol{\rho}}_1^{-1})_1 \leq (\bar{\mathbf{R}} \circ \hat{\boldsymbol{\rho}}_1^{-1})_2 \leq \dots \leq (\bar{\mathbf{R}} \circ \hat{\boldsymbol{\rho}}_1^{-1})_n$ (by Lemma 2 in Hüllermeier et al. (2008)).

By (6.4), it follows that $\hat{\boldsymbol{\rho}}_2 = \operatorname{argmax}_{\boldsymbol{\rho} \in \mathcal{P}_n} \sum_{i=1}^n i Y_i(\bar{\mathbf{R}} \circ \boldsymbol{\rho}^{-1})$, is such that $Y_i(\bar{\mathbf{R}} \circ \hat{\boldsymbol{\rho}}_2^{-1}) = i$, for each $i = 1, \dots, n$.

Now, notice that $(\bar{\mathbf{R}} \circ \hat{\boldsymbol{\rho}}_1^{-1})_1 \leq (\bar{\mathbf{R}} \circ \hat{\boldsymbol{\rho}}_1^{-1})_2 \leq \dots \leq (\bar{\mathbf{R}} \circ \hat{\boldsymbol{\rho}}_1^{-1})_n$ if and only if $Y_i(\bar{\mathbf{R}} \circ \hat{\boldsymbol{\rho}}_1^{-1}) = i$, for each $i = 1, \dots, n$. This proves that $\hat{\boldsymbol{\rho}}_1 = \hat{\boldsymbol{\rho}}_2$. \square

6.2 The conjugate prior for ρ when θ is known

We here give a strategy to elicit a conjugate prior on the consensus parameter of the Mallows model with Spearman distance. Notice that the likelihood of eq. (6.2) can be written as

$$P(\mathbf{R}_1, \dots, \mathbf{R}_N | \rho) \propto \exp \left[-\theta \sum_{j=1}^N \sum_{i=1}^n (\rho_i - R_{ij})^2 \right] \propto \exp \left[-\theta N \sum_{i=1}^n (\rho_i - \bar{R}_i)^2 \right].$$

It is worth pointing out here the notion of permutation polytope, which is intimately connected to the vector of sample ranks $\bar{\mathbf{R}}$.

Definition 6. The *permutation polytope* for given n , pp_n , is the convex hull of the points $\rho \in \mathcal{P}_n \subset \mathbb{R}^n$,

Then, the permutations $\rho \in \mathcal{P}_n \subset \mathbb{R}^n$ form the vertices of pp_n , and lie on the same $(n-1)$ -dimensional hyperplane $\mathcal{H}_n = \left\{ \mathbf{x} \in \mathbb{R}^n \mid \sum_{i=1}^n x_i = \frac{n(n+1)}{2} \right\}$.

Proposition 7. Let $\mathbf{R}_1, \dots, \mathbf{R}_N \in \mathcal{P}_n$, and define the vector of sample ranks as $\bar{\mathbf{R}} = (\bar{R}_1, \dots, \bar{R}_n)$, where $\bar{R}_i = \frac{1}{N} \sum_{j=1}^N R_{ji}$, for all $i = 1, \dots, n$. Then $\bar{\mathbf{R}} \in \text{pp}_n$.

Proof. The result follows directly from the following proposition by [Rado \(1952\)](#):

Proposition 8. ([Rado 1952](#)) Let us assume that $x_1 \geq x_2 \geq \dots \geq x_n$. Then a point $(t_1, \dots, t_n) \in \mathbb{R}^n$ belongs to the permutohedron $\text{pp}(x_1, \dots, x_n)$ if and only if $\sum_{i=1}^n t_i = \sum_{i=1}^n x_i$ and, for any nonempty subset $\{i_1, \dots, i_k\} \subset \{1, \dots, n\}$, $\sum_{i=1}^k t_{i_k} = \sum_{i=1}^k x_i$.

If we take $(x_1, x_2, \dots, x_n) = (n, n-1, \dots, 1)$, and $(t_1, \dots, t_n) = (\bar{R}_1, \dots, \bar{R}_n)$, it is easy to show that $(\bar{R}_1, \dots, \bar{R}_n) \in \text{pp}(n, n-1, \dots, 1) \equiv \text{pp}_n$. \square

Keeping θ fixed, the conjugate prior for $\rho \in \mathcal{P}_n$ is

$$\pi(\rho | \rho_0, \theta_0) = \frac{1}{Z_n^*(\theta_0, \rho_0)} \exp \left[-\theta_0 \sum_{i=1}^n (\rho_{0i} - \rho_i)^2 \right] \propto \exp \left[2\theta_0 \sum_{i=1}^n \rho_i \rho_{0i} \right], \quad (6.5)$$

where ρ_0 , the central parameter, belongs to the permutation polytope of order n , pp_n . What is appealing in this density, if compared with the Mallows model, is that its parameter space is regular, being given by $\text{pp}_n \times \mathbb{R}^+$.

The posterior density for $\boldsymbol{\rho}$ is

$$\pi(\boldsymbol{\rho}|\mathbf{R}_1, \dots, \mathbf{R}_N) \propto \exp \left\{ 2(\theta_0 + \theta N) \sum_{i=1}^n \rho_i \left[\frac{\theta N}{\theta_0 + \theta N} \bar{R}_i + \frac{\theta_0}{\theta_0 + \theta N} \rho_{0,i} \right] \right\}, \quad (6.6)$$

that is, the posterior has the same parametric form as the prior distribution. In particular, $\boldsymbol{\rho}|\mathbf{R}_1, \dots, \mathbf{R}_N \sim \pi \left(\frac{\theta N}{\theta_0 + \theta N} \bar{\mathbf{R}} + \frac{\theta_0}{\theta_0 + \theta N} \boldsymbol{\rho}_0, \theta_0 + \theta N \right)$, is the same parametric density of the prior (6.5), with updated parameters:

$$\boldsymbol{\rho}_N = \frac{\theta N}{\theta_0 + \theta N} \bar{\mathbf{R}} + \frac{\theta_0}{\theta_0 + \theta N} \boldsymbol{\rho}_0 \in \mathbb{P}\mathbb{P}_n \quad (6.7)$$

$$\theta_N = \theta_0 + \theta N \quad . \quad (6.8)$$

The posterior consensus parameter is expressed as a weighted average of the prior hyperparameter $\boldsymbol{\rho}_0$ and the observed mean value, $\bar{\mathbf{R}}$, with weights proportional to the spread parameters.

In the limiting cases, the posterior mean equals the prior mean or the observed value:

$$\boldsymbol{\rho}_N = \begin{cases} \boldsymbol{\rho}_0, & \text{if } \bar{\mathbf{R}} = \boldsymbol{\rho}_0 \text{ or } \theta_0 \rightarrow \infty \\ \bar{\mathbf{R}}, & \text{if } \bar{\mathbf{R}} = \boldsymbol{\rho}_0 \text{ or } \theta \rightarrow \infty \end{cases} .$$

If $\theta_0 \rightarrow \infty$ the prior density is infinitely more precise than the data, and so the posterior and prior distributions are identical and concentrated at the value $\boldsymbol{\rho}_0$. The other way around is true if $\theta \rightarrow \infty$, that is, the data are infinitely more precise than the prior, and the posterior density is concentrated at the observed value $\bar{\mathbf{R}}$. Notice that, by means of Proposition 6, the MAP of $\boldsymbol{\rho}$ is $\boldsymbol{\rho}_{MAP} = \mathbf{Y}(\boldsymbol{\rho}_N)$.

Figure 6.1 shows the $n = 3$ polytope in two dimensions, obtained with the R package **ConsRank** (D'Ambrosio et al. 2017). Rankings are represented by each vertex. The Euclidean distance between any two vertices is proportional to the Spearman distance between the two rankings corresponding to the two vertices (equality holds if the edges have length $\sqrt{2}$). Permutation polytopes are a common way of displaying the frequencies of a set of rankings, as they correspond to histograms for continuous data. In particular, the idea (due to Thompson 1993) is to visualize ranking data by placing at each vertex of the polytope a ball with radius proportional to the frequency of the ranking corresponding to that vertex.

We use this tool to illustrate the effects of the prior on $\boldsymbol{\rho}$ on inference. Since the

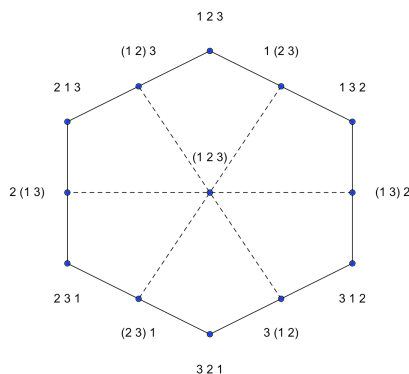


Figure 6.1: The generic permutation polytope for $n = 3$ items.

permutation polytope is difficult to visualize for high dimensions, we make an example with $n = 3$.

We simulate $N = 40$ rankings from the Mallows with Spearman distance, with $\theta = 0.5$ and $\boldsymbol{\rho} = (3, 2, 1)$, obtaining the vector of sample ranks $\bar{\mathbf{R}} = (2.25, 2.125, 1.625)$. We then sample 1000 rankings from the posterior density of $\boldsymbol{\rho}$, having assumed the prior of equation (6.5), with $\boldsymbol{\rho}_0 = (1, 2, 3)$, and varying $\theta_0 = 0, 10, 20, 30, 40, 50$.

In Figure 6.2 we represent through permutation polytopes the posterior samples corresponding to the different values of the prior hyperparameter θ_0 . The blue balls centered at each vertex of the polytopes, have radius proportional to the frequency of rankings corresponding to that vertex. As expected, the larger θ_0 is, the more concentrated the posterior ranks are at $\boldsymbol{\rho}_0$. When $\theta_0 = 0$ (Figure 6.2 top left), the posterior is only driven by $\bar{\mathbf{R}}$. Since the data were sampled from a Mallows density with very small spread parameter ($\theta = 0.5$) the frequencies of the rankings in this plot are not clearly concentrated at the true generating consensus $\boldsymbol{\rho} = (3, 2, 1)$. Repeating the same analysis with $\theta = 1.5$, we obtain the vector of sample ranks $\bar{\mathbf{R}} = (2.675, 2.025, 1.3)$, and the posterior samples represented in Figure 6.3, where is clear the impact of the sample concentration of the rankings around the consensus $\boldsymbol{\rho} = (3, 2, 1)$, due to the larger value of α .

Notice that, when choosing the prior hyperparameter $\boldsymbol{\rho}_0$ we are not forced to choose an element of the space of permutations. This flexibility enables to elicit a central parameter with uncertainty in some positions and not in others. For instance, we may be confident that the consensus rank ρ_1 is equal to 1, but are uncertain about ρ_2 and ρ_3 . This can be expressed through the prior hyperparameter $\boldsymbol{\rho}_0 = (1, 2.5, 2.5) \in \mathbb{P}\mathbb{P}_3$.

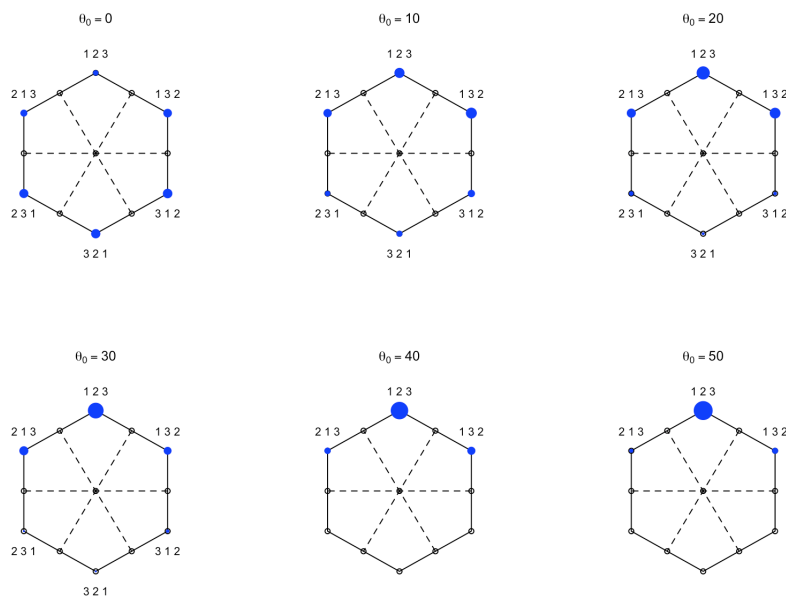


Figure 6.2: The generic permutation polytope for $n = 3$ items. $\theta = 0.5$.

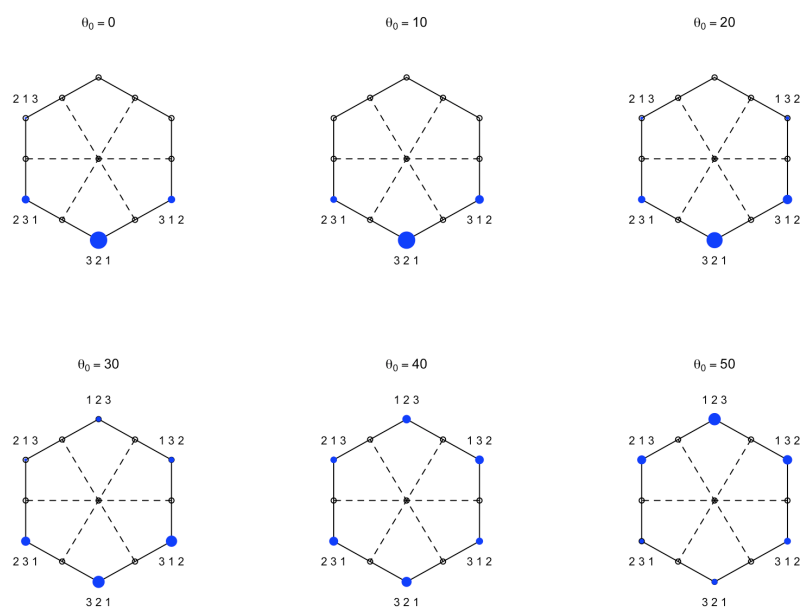


Figure 6.3: The generic permutation polytope for $n = 3$ items. $\theta = 1.5$.

The results outlined in this chapter are preliminary and have been derived in the last weeks of work on this thesis. More work is necessary in order to apply the findings to the general Bayesian Mallows model. However, we are confident that the idea of defining a conjugate prior for the consensus parameter is an interesting line of research. We will pursue the study of this topic, also exploring the possibility to elicit a joint conjugate prior for the two parameters of the Mallows model.

Discussion

After its first appearance in [Mallows \(1957\)](#), the Mallows model received much attention in the literature on ranking data. It was originally developed for dealing with pairwise comparisons, and it only considered two distances: Kendall and Spearman. [Diaconis \(1988\)](#) developed the theory for a general class of distance-based models, thus extending the Mallows model to any arbitrary right-invariant distance.

The principal advantage of the Mallows model, when compared with other statistical approaches to ranking data, is its ability to capture the main features of the data with only two parameters. However, many scholars believe that this feature does not allow the model to be as flexible as a stage-wise model, for example the Plackett-Luce, whose parameters are real-valued. We argue that the Mallows model is versatile in that it can adapt to many different distance measures; this results into richer expressiveness than many other probabilistic models on rankings. The choice of the appropriate distance depends on the specific application. Some problems require a distance able to measure only the disorder in the given domain, like Cayley, while others require a distance more suited to learn preferences of a population, like Kendall or footrule. In other words, the application field influences the most accurate distance to use: specific problems are better modeled under a particular distance. However, the Mallows model has one main shortcoming, namely that the computation of the normalizing constant is intractable. This problem limited its use to some specific distances, such as the Kendall distance, for which a closed form of the normalizing constant exists and is tractable.

One of the main contributions of this thesis is to show how to handle all the other well-known right-invariant distances, so that the versatility of the Mallows model is fully exploited. In particular, the implemented algorithm allows to use Kendall, Cayley, footrule and Spearman distances, and it can be immediately generalized to any other right-invariant distance (e.g. Hamming or Ulam). For Kendall and Cayley distances, we exploit the well-known closed form of the normalization constant due to [Fligner and](#)

[Verducci \(1986\)](#). For footrule ($n < 51$) and Spearman ($n < 15$), we provide the exact form of the normalizing constant, while for larger values of n we develop a strategy to approximate it. In particular, we propose an off-line importance sampling scheme, and we largely document the quality and efficiency of this approximation. The above strategy is fully integrated into the developed Bayesian framework for the analysis of the Mallows model (Chapter 2). Therefore, our framework makes possible the Bayesian learning of the Mallows model for any right-invariant distance. As an illustrative example, we specialize our framework to the Cayley distance, which has many appealing properties and connections with other statistical areas (Chapter 5). The possibility to perform Bayesian analysis of the Mallows model with any right-invariant distance is a novel contribution to the literature on ranking data, where the Mallows model was studied in the Bayesian paradigm only in few specific cases (see for instance [Meilă and Bao 2010](#)).

One of the well known advantages of the Bayesian approach is its ability to provide posterior uncertainty related to any quantity of interest. Indeed, the availability of the full posterior distribution of the parameters allows to obtain any summary of interest, driven by the application at hand. This feature proved to be crucial when comparing our model with existing competitors. In particular, we compared our results with some of the works in the learning to rank (LETOR) field, whose aim is to find the ranking that best describes the preferences expressed in the data. Our method is capable not only to perform rank aggregation, but it also provides uncertainty quantification around the estimated rankings.

Another advantage of Bayesian statistics is the possibility to include prior information into the analysis. However, in this thesis we always used the uniform prior over the space of permutations, because the main interest was in developing a general framework in the simplest possible case. Nevertheless, we believe that studying the elicitation problem in this context is very interesting and a promising avenue for future research. In Chapter 6 we move the first steps in this direction by providing some preliminary results on the conjugate prior in the special case of the Mallows model with Spearman distance.

A second core contribution of this work is the development of a model to deal with non-transitive patterns in individual pairwise preference data (Chapter 3). The literature on inferential models for non-transitive pair comparisons is limited, and it mostly appears in the machine learning community, under the name of feedback arc set problem (see Section 1.2.1), or of linear ordering problem. These works generally do not provide uncertainty

quantifications to the derived point estimates. On the contrary, our method leads to a probabilistic model of rankings. In this case it provides not only the posterior distribution of the consensus ranking, but also of the individual rankings for each user, which can be used for performing personalized recommendation, or to study the association between individual preferences and user-related covariates.

It is easy to understand the relevance of this second contribution when dealing with data showing many non-transitive patterns. Indeed, only in few cases it is possible for a person to compare many items at the same time, to assign ranks to all of them, and thus to produce a unique ranking. Often instead, when differences between items are small, or the number of items under comparison is large, it is difficult for a person to provide a full ranking. In such cases, the preferred method is to let the person repeatedly compare the items in pairs, that is, to design a paired comparisons experiment (David 1963). However, this method admits the possibility that the person contradicts herself. This is very common, especially because this kind of experiment is chosen in situations where the items under evaluation are rather similar, as already mentioned. Clearly, the presence of non-transitive patterns makes impossible to readily identify the unique ranking describing a person's preferences. To our knowledge most of the existing statistical methods to estimate individual rankings from pairwise comparison data do not specifically model the non-transitivity characterizing the data. Instead we incorporate the non-transitive patterns of the data directly into the developed Bayesian framework of Chapter 2, thus enabling the statistician not to lose possibly relevant information.

The advantages discussed so far proved to be very important in the application to sound data discussed in Chapter 4. The aim of the experiment was to investigate the impact of 3-D sound spatialization on listeners' understanding of human agency, when they hear abstract sounds. Listeners' linguistic descriptions of what they hear are notoriously inconsistent, despite often meaning the same. The experiment therefore had to be designed to avoid the need for a descriptive language. We first considered allowing listeners to allocate each sound a score, indicating how strongly each evoked human agency in relation to the other sounds. However, since we already knew that listeners would span a large range of spatial audio skills and that the tests would be challenging, the obvious choice was to let them evaluate the sounds in pairs. We then opted for analyzing the data at hand with the Bayesian Mallows model for non-transitive pair comparisons. As expected, the collected data were very noisy and ambiguous, showing many non-transitive

patterns. In this case, a model able to detect and correct these non-transitivities is more appropriate. The results of the experiment showed three clusters of listeners, each sharing different opinions about the degree of human causation behind sounds. This grouping indicates that answering the question as to whether sound spatialization can suggest human agency is far from straightforward. In addition to the grouping of the listeners around the shared consensus rankings, we also studied the association between individual listeners' rankings and their own musical experience or musical background. This enabled to evince that spatial listening is a skill that is enhanced through experience and personal interest.

All methods presented have been implemented in R and C++. We are working on the development of an R package, in order to make available the theory and algorithms to the interested reader.

Future work

In addition to the line of research outlined in Chapter 6, we intend to extend the framework of this thesis in the following directions.

First, we aim at studying the Generalized Mallows model (Fligner and Verducci 1986) with Cayley distance in the Bayesian framework, and to investigate whether within our framework it is feasible to handle the weighted Mallows model of Lee and Yu (2010).

Second, we plan to allow for the possibility to incorporate covariates into the model. One direction in this regard is to let the probability of making mistakes depend on item-specific covariates, as discussed in Section 3.1.3. Another possibility is to exploit user-specific covariates to model the dependency between the users, direction that could be crucial in the classification and prediction of new instances.

Third, we aim at generalizing the procedure to the non-parametric case, thus enabling to automatically select the number of clusters. A first insight in this direction is to simply put a prior on the number of clusters, and let the algorithm estimate its posterior probability. The computational aspect of this method can be tackled by using the well-known reversible jump approach (Green 1995, Richardson and Green 1997).

Finally, we aim at explicitly model the presence of ties and/or indifference in the data.

Bibliography

- Agresti, A. (1996), *Categorical data analysis*, New York: John Wiley and Sons. [22](#), [85](#)
- Ailon, N. (2012), ‘An active learning algorithm for ranking from pairwise preferences with an almost optimal query complexity’, *Journal of Machine Learning Research* **13**, 137–164. [26](#)
- Aldous, D. J. (1985), Exchangeability and related topics, in ‘École d’Été de Probabilités de Saint-Flour XIII’, Springer, pp. 1–198. [164](#)
- Aledo, J. A., Gámez, J. A. and Molina, D. (2013), ‘Tackling the rank aggregation problem with evolutionary algorithms’, *Applied Mathematics and Computation* **222**, 632–644. [79](#)
- Ali, A. and Meilă, M. (2012), ‘Experiments with Kemeny ranking: What works when?’, *Mathematical Social Sciences* **64**(1), 28–40. [27](#), [79](#)
- Alvo, M. and Yu, P. L. H. (2014), *Statistical Methods for Ranking Data*, Frontiers in Probability and the Statistical Sciences, Springer, New York, NY, USA. [9](#), [13](#)
- Andrieu, C. and Roberts, G. O. (2009), ‘The pseudo-marginal approach for efficient Monte Carlo computations’, *The Annals of Statistics* **37**(2), 697–725. [38](#)
- Asfaw, D., Vitelli, V., Sørensen, Ø., Arjas, E. and Frigessi, A. (2017), ‘Time-varying rankings with the Bayesian Mallows model’, *Stat* **6**(1), 14–30. [80](#)
- Barrett, N. (2016), ‘Interactive spatial sonification of multidimensional data for composition and auditory display’, *Computer Music Journal* . [142](#), [144](#)
- Barrett, N. and Crispino, M. (2017), ‘The Impact of 3-D Sound Spatialisation on Listeners’ Understanding of Human Agency in Acousmatic Music’, *Submitted* . [7](#), [137](#)
- Bartholdi, J. J., Tovey, C. A. and Trick, M. A. (1989a), ‘The computational difficulty of manipulating an election’, *Social Choice and Welfare* **6**(3), 227–241. [17](#), [27](#)
- Bartholdi, J., Tovey, C. and Trick, M. A. (1989b), ‘Voting schemes for which it can be difficult to tell who won the election’, *Social Choice and Welfare* **6**(2), 157–165. [17](#), [27](#)
- Beaumont, M. A. (2003), ‘Estimation of population growth or decline in genetically monitored populations’, *Genetics* **164**(3), 1139–1160. [38](#)

- Bigand, E. and Parncutt, R. (1999), ‘Perceiving musical tension in long chord sequences’, *Psychological Research* **62**(4), 237–254. [146](#)
- Blauert, J. (1997), *Spatial hearing: the psychophysics of human sound localization*, MIT press. [140](#)
- Böckenholt, U. (1988), ‘A logistic representation of multivariate paired-comparison models’, *Journal of mathematical psychology* **32**(1), 44–63. [23](#)
- Böckenholt, U. (2001), ‘Hierarchical modeling of paired comparison data.’, *Psychological Methods* **6**(1), 49. [23](#)
- Böckenholt, U. (2006), ‘Thurstonian-based analyses: Past, present, and future utilities’, *Psychometrika* **71**(4), 615–629. [23](#)
- Böckenholt, U. and Tsai, R.-C. (2001), ‘Individual differences in paired comparison data’, *British Journal of Mathematical and Statistical Psychology* **54**(2), 265–277. [23](#)
- Bradley, R. A. and Terry, M. E. (1952), ‘Rank analysis of incomplete block designs: I. The method of paired comparisons’, *Biometrika* **39**(3/4), 324–345. [12](#), [13](#), [21](#), [22](#), [127](#)
- Brin, S. and Page, L. (1998), ‘The anatomy of a large-scale hypertextual web search engine’, *Computer networks and ISDN systems* **30**(1), 107–117. [73](#)
- Busse, L. M., Orbanz, P. and Buhmann, J. M. (2007), Cluster Analysis of Heterogeneous Rank Data, in ‘Proceedings of the 24th International Conference on Machine Learning’, ICML ’07, ACM, New York, NY, USA, pp. 113–120. [17](#), [19](#)
- Caron, F. and Doucet, A. (2012), ‘Efficient Bayesian inference for generalized Bradley–Terry models’, *Journal of Computational and Graphical Statistics* **21**(1), 174–196. [21](#), [22](#), [128](#), [135](#)
- Caron, F. and Teh, Y. W. (2012), Bayesian nonparametric models for ranked data, in ‘Advances in Neural Information Processing Systems’, pp. 1520–1528. [21](#), [79](#)
- Caron, F., Teh, Y. W. and Murphy, T. B. (2014), ‘Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes’, *The Annals of Applied Statistics* **8**(2), 1145–1181. [21](#)
- Carpentier, T., Barrett, N., Gottfried, R. and Noisternig, M. (2017), ‘Holophonic sound in IRCAM’s concert hall: Technological and aesthetic practices’, *Computer Music Journal* . [142](#)
- 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**(2), 245–259. [24](#)

- Celeux, G., Forbes, F., Robert, C. P. and Titterton, D. M. (2006), ‘Deviance information criteria for missing data models’, *Bayesian Analysis* **1**(4), 651–674. [76](#)
- Celeux, G., Hurn, M. and Robert, C. (2000), ‘Computational and Inferential Difficulties with Mixture Posterior Distribution’, *Journal of the American Statistical Association* **95**(451), 957–970. [62](#)
- Cheng, W. and Hüllermeier, E. (2008), Instance-Based Label Ranking using the Mallows Model, in ‘ECCBR Workshops’, pp. 143–157. [28](#)
- Crispino, M., Vitelli, V., Barrett, N., Arjas, E. and Frigessi, A. (2017), ‘A Bayesian Mallows approach to non-transitive pair comparison data: how human are sounds?’, *Submitted, available at <https://arxiv.org/abs/1705.08805>*. [7](#), [9](#), [85](#), [137](#)
- Critchlow, D. E. (2012), *Metric methods for analyzing partially ranked data*, Vol. 34, Springer Science and Business Media. [19](#), [70](#)
- Critchlow, D., Fligner, M. and Verducci, J. (1993), ‘Probability Models and Statistical Analyses for Ranking Data’, *Springer, New York* **220**, 90–98. [9](#)
- Csardi, G. and Nepusz, T. (2006), ‘The igraph software package for complex network research’, *InterJournal, Complex Systems* p. 1695.
URL: <http://igraph.org> [73](#)
- D’Ambrosio, A., Amodio, S. and Mazzeo, G. (2017), ‘ConsRank: Compute the Median Ranking(s) According to the Kemeny’s Axiomatic Approach’. R package version 2.0.1.
URL: <https://CRAN.R-project.org/package=ConsRank> [168](#)
- Daniel, J. and Moreau, S. (2004), Further study of sound field coding with higher order ambisonics, in ‘Audio Engineering Society Convention 116’, Audio Engineering Society. [142](#)
- Daniels, H. E. (1950), ‘Rank Correlation and Population Models’, *Journal of the Royal Statistical Society. Series B (Methodological)* **12**(2), 171–191. [12](#)
- David, H. A. (1963), *The method of paired comparisons*, Vol. 12, DTIC Document. [13](#), [85](#), [141](#), [173](#)
- Davidson, R. R. (1970), ‘On extending the Bradley-Terry model to accommodate ties in paired comparison experiments’, *Journal of the American Statistical Association* **65**(329), 317–328. [22](#)
- Dawid, A. P. (1982), ‘The well-calibrated Bayesian’, *Journal of the American Statistical Association* **77**(379), 605–610. [66](#)
- de Borda, J. C. (1781), ‘Mémoire sur les élections au scrutin, histoire de l’académie royale des sciences’, *Paris, France*. [24](#), [27](#), [52](#)

- DeConde, R. P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B. and Etzioni, R. (2006), ‘Combining Results of Microarray Experiments: A Rank Aggregation Approach’, *Statistical Applications in Genetics and Molecular Biology* **5**(1), Article 15. [68](#), [69](#)
- Deng, K., Han, S., Li, K. J. and Liu, J. S. (2014), ‘Bayesian Aggregation of Order-Based Rank Data’, *Journal of the American Statistical Association* **109**(507), 1023–1039. [68](#), [69](#)
- Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. A. and Chinnaiyan, A. M. (2001), ‘Delineation of prognostic biomarkers in prostate cancer’, *Nature* **412**, 822–826. [68](#)
- Diaconis, P. (1988), *Group representations in probability and statistics*, Vol. 11 of *Lecture Notes - Monograph Series*, Institute of Mathematical Statistics, Hayward, CA, USA. [9](#), [14](#), [17](#), [19](#), [171](#)
- Ding, W., Ishwar, P. and Saligrama, V. (2015), ‘Learning mixed membership mallows models from pairwise comparisons’, *arXiv preprint arXiv:1504.00757*. [25](#)
- 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’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **47**(4), 511–525. [23](#)
- Dittrich, R., Hatzinger, R. and Katzenbeisser, W. (2002), ‘Modelling dependencies in paired comparison data: A log-linear approach’, *Computational statistics and data analysis* **40**(1), 39–57. [23](#)
- Doignon, J. P., Pekeč, A. and Regenwetter, M. (2004), ‘The repeated insertion model for rankings: Missing link between two subset choice models’, *Psychometrika* **69**(1), 33–54. [19](#)
- Dwork, C., Kumar, R., Naor, M. and Sivakumar, D. (2001), Rank aggregation methods for the web, in ‘Proceedings of the 10th international conference on World Wide Web’, ACM, pp. 613–622. [27](#)
- Eitan, Z. and Granot, R. Y. (2006), ‘How music moves’, *Music Perception: An Interdisciplinary Journal* **23**(3), 221–248. [138](#)
- Favrot, S. and Buchholz, J. (2012), ‘Reproduction of nearby sound sources using higher-order ambisonics with practical loudspeaker arrays’, *Acta Acustica United with Acustica* **98**(1), 48–60. [142](#)
- Feller, W. (1968), *An introduction to probability theory and its applications: volume I*, Vol. 3, John Wiley and Sons New York. [158](#)

- Firth, D. and Turner, H. L. (2012), ‘Bradley-Terry models in R: the BradleyTerry2 package’, *Journal of Statistical Software* **48**(9). [73](#)
- Fligner, M. A. and Verducci, J. S. (1986), ‘Distance based ranking models’, *Journal of the Royal Statistical Society B* **48**(3), 359–369. [18](#), [20](#), [37](#), [80](#), [81](#), [159](#), [171](#), [174](#)
- Ford, L. R. (1957), ‘Solution of a ranking problem from binary comparisons’, *The American Mathematical Monthly* **64**(8), 28–33. [22](#), [135](#)
- Francis, B., Dittrich, R. and Hatzinger, R. (2010), ‘Modeling heterogeneity in ranked responses by nonparametric maximum likelihood: how do Europeans get their scientific knowledge?’, *The Annals of Applied Statistics* **4**(4), 2181–2202. [23](#)
- Fürnkranz, J. and Hüllermeier, E. (2010), *Preference learning: An introduction*, Springer. [27](#), [63](#)
- Gelman, A., Roberts, G. O. and Gilks, W. R. (1996), ‘Efficient Metropolis jumping rules’, *Bayesian statistics* **5**(599-608), 42. [36](#)
- Glickman, M. E. (1999), ‘Parameter estimation in large dynamic paired comparison experiments’, *Applied Statistics* pp. 377–394. [27](#)
- Gnedin, A. and Gorin, V. (2016), ‘Spherically Symmetric Random Permutations’, *arXiv preprint arXiv:1611.01860*. [164](#)
- Godøy, R. I. (2006), ‘Gestural-Sonorous Objects: embodied extensions of Schaeffer’s conceptual apparatus’, *Organised Sound* **11**(02), 149–157. [138](#)
- Godøy, R. I. (2010), ‘Images of sonic objects’, *Organised Sound* **15**(01), 54–62. [138](#)
- Godøy, R. I., Haga, E. and Jensenius, A. R. (2006), ‘Exploring music-related gestures by sound-tracing: A preliminary study’. [139](#)
- Gopalan, P., Jayram, T., Krauthgamer, R. and Kumar, R. (2006), Approximating the Longest Increasing Sequence and Distance from Sortedness in a Data Stream. Research Microsoft Publications. [34](#)
- Gormley, I. C. and Murphy, T. B. (2006), ‘Analysis of Irish third-level college applications data’, *Journal of the Royal Statistical Society A* **169**(2), 361–379. [21](#)
- Green, P. J. (1995), ‘Reversible jump Markov chain Monte Carlo computation and Bayesian model determination’, *Biometrika* **82**(4), 711–732.
URL: <http://biomet.oxfordjournals.org/content/82/4/711.abstract> [174](#)
- Green, P. J. and Han, X. I. (1992), Metropolis methods, Gaussian proposals and anti-thetic variables, in ‘Stochastic Models, Statistical methods, and Algorithms in Image Analysis’, Springer, pp. 142–164. [36](#)

- Grond, F. and Berger, J. (2011), ‘Parameter Mapping Sonification’, In *Thomas Hermann, Andrew D. Hunt, and John Neuhoff (eds.), The Sonification Handbook* pp. 363–398. [4](#)
- Guiver, J. and Snelson, E. (2009), Bayesian inference for Plackett-Luce ranking models, in ‘proceedings of the 26th annual international conference on machine learning’, ACM, pp. 377–384. [21](#)
- Gupta, J. and Damien, P. (2002), ‘Conjugacy class prior distributions on metric-based ranking models’, *Journal of the Royal Statistical Society B* **64**(3), 433–445. [161](#)
- Hatzinger, R., Dittrich, R. et al. (2012), ‘prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings’, *Journal of Statistical Software* **48**(10), 1–31. [23](#)
- Hermann, T., Hunt, A. and Neuhoff, J. G. (2011), *The sonification handbook*, Logos Verlag Berlin. [144](#)
- Hornik, K. and Meyer, D. (2014), ‘relations: Data Structures and Algorithms for Relations’, *R package version 0.5* . [60](#)
- Hüllermeier, E., Fürnkranz, J., Cheng, W. and Brinker, K. (2008), ‘Label ranking by learning pairwise preferences’, *Artificial Intelligence* **172**(16), 1897–1916. [27](#), [166](#)
- Hunter, D. R. (2004), ‘MM algorithms for generalized Bradley-Terry models’, *The Annals of Statistics* **32**(1), 384–406. [20](#), [22](#), [135](#)
- Irurozki, E., Calvo, B. and Lozano, A. (2014), ‘Sampling and learning the Mallows and generalized Mallows models under the Hamming distance’, *Bernoulli (submitted)* . [18](#), [19](#)
- Irurozki, E., Calvo, B. and Lozano, A. (2016a), ‘PerMallows: An R package for Mallows and generalized Mallows models’, *Journal of Statistical Software* **71**. [19](#), [36](#), [39](#), [52](#)
- Irurozki, E., Calvo, B. and Lozano, A. (2016b), ‘Sampling and learning the Mallows and Generalized Mallows models under the Cayley distance’, *Methodology and Computing in Applied Probability* . [18](#), [19](#), [159](#)
- Jacques, J. and Biernacki, C. (2014), ‘Model-based clustering for multivariate partial ranking data’, *Journal of Statistical Planning and Inference* **149**, 201–217. [19](#)
- Jacques, J., Grimonprez, Q. and Biernacki, C. (2014), ‘Rankcluster: An R package for clustering multivariate partial rankings’, *The R Journal* **6**(1), 10. [52](#)
- Jasra, A., Holmes, C. and Stephens, D. (2005), ‘Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling’, *Statistical Science* **20**(1), 50–67. [62](#)

- Kamishima, T. (2003), Nantonac Collaborative Filtering: Recommendation Based on Order Responses, *in* ‘Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining’, ACM, New York, NY, USA, pp. 583–588. [73](#)
- Kemeny, J. G. and Snell, J. L. (1962), *Mathematical models in the social sciences*, Blaisdell Publishing Company. [27](#)
- Kenyon-Mathieu, C. and Schudy, W. (2007), How to rank with few errors, *in* ‘Proceedings of the thirty-ninth annual ACM symposium on Theory of computing’, pp. 95–103. [26](#), [27](#)
- Krumhansl, C. L. (1996), ‘A perceptual analysis of Mozart’s Piano Sonata K. 282: Segmentation, tension, and musical ideas’, *Music Perception: An Interdisciplinary Journal* **13**(3), 401–432. [146](#)
- Lebanon, G. and Mao, Y. (2008), ‘Non-parametric modeling of partially ranked data’, *Journal of Machine Learning Research* **9**, 2401–2429. [19](#)
- Lee, P. H. and Yu, P. L. H. (2010), ‘Distance-based tree models for ranking data’, *Computational Statistics and Data Analysis* **54**(6), 1672–1682. [80](#), [174](#)
- Lee, P. H. and Yu, P. L. H. (2012), ‘Mixtures of weighted distance-based models for ranking data with applications in political studies’, *Computational Statistics and Data Analysis* **56**(8), 2486–2500. [80](#)
- Leman, M. (2012), Musical gestures and embodied cognition, *in* ‘Journées d’informatique musicale (JIM-2012)’, Université de Mons, pp. 5–7. [138](#)
- Lin, S. and Ding, J. (2009), ‘Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA studies’, *Biometrics* **65**(1), 9–18. [68](#), [69](#)
- Little, R. (2011), ‘Calibrated Bayes, for statistics in general, and missing data in particular’, *Statistical Science* **26**(2), 162–174. [66](#)
- Liu, N. N., Zhao, M. and Yang, Q. (2009), Probabilistic latent preference analysis for collaborative filtering, *in* ‘Proceedings of the 18th ACM conference on Information and knowledge management’, ACM, pp. 759–766. [27](#)
- Lu, T. and Boutilier, C. (2014), ‘Effective sampling and learning for Mallows models with pairwise-preference data’, *Journal of Machine Learning Research* **15**, 3783–3829. [18](#), [19](#), [25](#), [60](#), [73](#), [75](#), [76](#)
- Lu, Y. and Negahban, S. N. (2015), Individualized rank aggregation using nuclear norm regularization, *in* ‘Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on’, IEEE, pp. 1473–1479. [28](#)

- Luce, R. D. (1959), *Individual choice behavior: A theoretical analysis*, Wiley, New York, NY, USA. [12](#), [20](#)
- Luo, J., Duggan, D. J., Chen, Y., Sauvageot, J., Ewing, C. M., Bittner, M. L., Trent, J. M. and Isaacs, W. B. (2001), ‘Human Prostate Cancer and Benign Prostatic Hyperplasia: Molecular Dissection by Gene Expression Profiling’, *Cancer Research* **61**(12), 4683–4688. [68](#)
- Mallows, C. L. (1957), ‘Non-null ranking models. I’, *Biometrika* **44**(1/2), 114–130. [13](#), [14](#), [19](#), [171](#)
- Marden, J. I. (1995), *Analyzing and Modeling Rank Data*, Vol. 64 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, Cambridge, MA, USA. [9](#), [15](#), [21](#), [158](#)
- Marentakis, G. and McAdams, S. (2013), ‘Perceptual impact of gesture control of spatialization’, *ACM Transactions on Applied Perception (TAP)* **10**(4), 22. [139](#)
- Marquis of Condorcet, M. J. A. N. d. C. (1785), ‘Essai sur l’application de l’analyse à la probabilité des décisions rendues à la pluralité des voix’, *Paris: De l’imprimerie royale* . [24](#), [27](#)
- Meilă, M. and Bao, L. (2010), ‘An Exponential Model for Infinite Rankings’, *Journal of Machine Learning Research* **11**, 3481–3518. [20](#), [56](#), [58](#), [80](#), [172](#)
- Meilă, M. and Chen, H. (2010), Dirichlet Process Mixtures of Generalized Mallows Models, *in* ‘Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)’, AUAI Press, Corvallis, OR, USA, pp. 358–367. [18](#), [20](#)
- Meyer, D. and Hornik, K. (2009), ‘Generalized and Customizable Sets in R’, *Journal of Statistical Software* **31**(2), 1–27. [60](#)
- Møller, J., Pettitt, A. N., Reeves, R. and Berthelsen, K. K. (2006), ‘An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants’, *Biometrika* **93**(2), 451–458. [38](#)
- Mollica, C. and Tardella, L. (2016a), ‘Bayesian Plackett-Luce mixture models for partially ranked data’, *Psychometrika* . (published on line). [21](#)
- Mollica, C. and Tardella, L. (2016b), ‘PLMIX: An R package for modeling and clustering partially ranked data’, *arXiv preprint arXiv:1612.08141* . [21](#)
- Mukherjee, S. (2016), ‘Estimation in exponential families on permutations’, *The Annals of Statistics* **44**(2), 853–875. [18](#), [39](#), [43](#), [46](#), [76](#), [79](#)
- Murphy, T. B. and Martin, D. (2003), ‘Mixtures of distance-based models for ranking data’, *Computational Statistics and Data Analysis* **41**(3–4), 645 – 655. [19](#)

- Murray, I., Ghahramani, Z. and MacKay, D. (2012), ‘MCMC for doubly-intractable distributions’, *arXiv preprint arXiv:1206.6848* . 38
- Negahban, S., Oh, S. and Shah, D. (2012), Iterative ranking from pair-wise comparisons, *in* ‘Advances in Neural Information Processing Systems’, pp. 2474–2482. 27
- Ollen, J. E. (2006), A criterion-related validity test of selected indicators of musical sophistication using expert ratings, PhD thesis, The Ohio State University. 6, 148
- Papastamoulis, P. (2015), ‘label. switching: An R package for dealing with the label switching problem in MCMC outputs’, *arXiv preprint:1503.02271* . 62
- Park, D., Neeman, J., Zhang, J., Sanghavi, S. and Dhillon, I. (2015), Preference completion: Large-scale collaborative ranking from pairwise comparisons, *in* ‘International Conference on Machine Learning’, pp. 1907–1916. 28
- Pedersen, M. and Alsop, R. (2012), An approach to feature extraction of human movement qualities and its application to sound synthesis, *in* ‘Interactive: Proceedings of the 2012 Australasian Computer Music Conference’. 139
- Pihur, V., Datta, S. and Datta, S. (2009), ‘RankAggreg, an R package for weighted rank aggregation’, *BMC bioinformatics* **10**(1), 62. 52, 70
- Plackett, R. L. (1975), ‘The Analysis of Permutations’, *Journal of the Royal Statistical Society C* **24**(2), 193–202. 20
- Pulkki, V. et al. (2001), *Spatial sound generation and perception by amplitude panning techniques*, Helsinki University of Technology. 139
- Rado, R. (1952), ‘An inequality’, *Journal of the London Mathematical Society* **1**(1), 1–6. 167
- Rajkumar, A., Ghoshal, S., Lim, L.-H. and Agarwal, S. (2015), Ranking from Stochastic Pairwise Preferences: Recovering Condorcet Winners and Tournament Solution Sets at the Top, *in* ‘ICML’, pp. 665–673. 27
- Raman, K. and Joachims, T. (2015), Bayesian ordinal peer grading, *in* ‘Proceedings of the Second (2015) ACM Conference on Learning@ Scale’, ACM, pp. 149–156. 27
- Regenwetter, M., Falmagne, J. C. and Grofman, B. (1999), ‘A stochastic model of preference change and its application to 1992 presidential election panel data’, *Psychological Review* **106**(2), 362–384. 79
- Rendle, S., Freudenthaler, C., Gantner, Z. and Schmidt-Thieme, L. (2009), BPR: Bayesian personalized ranking from implicit feedback, *in* ‘Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence’, AUAI Press, pp. 452–461. 28

- Richardson, S. and Green, P. J. (1997), ‘On Bayesian analysis of mixtures with an unknown number of components (with discussion)’, *Journal of the Royal Statistical Society: series B (statistical methodology)* **59**(4), 731–792. [174](#)
- Roberts, G. O., Gelman, A. and Gilks, W. R. (1997), ‘Weak convergence and optimal scaling of random walk Metropolis algorithms’, *The annals of applied probability* **7**(1), 110–120. [36](#)
- Schimek, M. G., Budinská, E., Kugler, K. G., Švendová, V., Ding, J. and Lin, S. (2015), ‘TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists’, *Statistical Applications in Genetics and Molecular Biology* **14**(3), 311–316. [70](#)
- Shah, N., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K. and Wainwright, M. (2015), Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence, in ‘Artificial Intelligence and Statistics’, pp. 856–865. [27](#)
- Shestopalova, L., Bohm, T. M., Bendixen, A., Andreou, A. G., Georgiou, J., Garreau, G., Hajdu, B., Denham, S. L. and Winkler, I. (2015), ‘Do audio-visual motion cues promote segregation of auditory streams?’, *Probing auditory scene analysis* p. 50. [140](#)
- Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D’Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R. and Sellers, W. R. (2002), ‘Gene expression correlates of clinical prostate cancer behavior’, *Cancer Cell* **1**(2), 203 – 209. [68](#)
- Sloane, N. J. A. (2017), ‘The Encyclopedia of Integer Sequences’.
URL: <http://oeis.org> [39](#), [162](#)
- Smith, B. B. (1950), ‘Discussion of professor Ross’s paper’, *Journal of the Royal Statistical Society B* **12**(1), 41–59. [13](#)
- Tanner, M. and Wong, W. (1987), ‘The calculation of posterior distributions by data augmentation (with discussion)’, *Journal of the American Statistical Association* **82**, 528–550. [54](#)
- Thompson, G. (1993), ‘Generalized permutation polytopes and exploratory graphical methods for ranked data’, *The Annals of Statistics* pp. 1401–1430. [168](#)
- Thurstone, L. L. (1927), ‘A law of comparative judgment.’, *Psychological review* **34**(4), 273. [11](#), [12](#), [21](#), [22](#)
- True, L., Coleman, I., Hawley, S., Huang, C., Gifford, D., Coleman, R., Beer, T. M., Gelmann, E., Datta, M., Mostaghel, E., Knudsen, B., Lange, P., Vessella, R., Lin, D., Hood, L. and Nelson, P. S. (2006), ‘A molecular correlate to the Gleason grading system for prostate adenocarcinoma’, *Proceedings of the National Academy of Sciences* **103**(29), 10991–10996. [68](#)

- 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**(1), 10–20. [24](#)
- Tsay, C.-J. (2013), ‘Sight over sound in the judgment of music performance’, *Proceedings of the National Academy of Sciences* **110**(36), 14580–14585. [140](#)
- Tversky, A. (1969), ‘Intransitivity of Preferences’, *Preference, Belief, and Similarity* p. 433. [23](#)
- Usami, S. (2010), ‘Individual differences multidimensional Bradley-Terry model using reversible jump Markov Chain Monte Carlo algorithm’, *Behaviormetrika* **37**(2), 135–155. [24](#)
- Vines, B. W., Krumhansl, C. L., Wanderley, M. M. and Levitin, D. J. (2006), ‘Cross-modal interactions in the perception of musical performance’, *Cognition* **101**(1), 80–113. [140](#)
- Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A. and Arjas, E. (2017), ‘Probabilistic preference learning with the Mallows rank model’, *Accepted for publication on Journal of Machine Learning Research, available at <https://arxiv.org/abs/1405.7945v4>*. [7](#), [9](#), [29](#)
- Volkovs, M. N. and Zemel, R. S. (2014), ‘New Learning Methods for Supervised and Un-supervised Preference Aggregation’, *Journal of Machine Learning Research* **15**, 1135–1176. [24](#), [25](#), [79](#)
- Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F. and Hampton, G. M. (2001), ‘Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer’, *Cancer Research* **61**(16), 5974–5978. [68](#)
- Wu, R., Xu, J., Srikant, R., Massoulié, L., Lelarge, M. and Hajek, B. (2015), Clustering and inference from pairwise comparisons, *in* ‘ACM SIGMETRICS Performance Evaluation Review’, Vol. 43/1, ACM, pp. 449–450. [22](#)
- Yan, T. (2016), ‘Ranking in the generalized Bradley–Terry models when the strong connection condition fails’, *Communications in Statistics-Theory and Methods* **45**(2), 340–353. [22](#)
- Zermelo, E. (1929), ‘Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung’, *Mathematische Zeitschrift* **29**(1), 436–460. [22](#)
- Zotter, F., Pomberger, H. and Noisternig, M. (2012), ‘Energy-preserving ambisonic decoding’, *Acta Acustica united with Acustica* **98**(1), 37–47. [142](#)