Methods for the Statistical Analysis of Preferences, with Applications to Social Science Data

Michael Pearce

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Elena A. Erosheva, Chair
Adrian E. Raftery
Marina Meilă

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Methods for the Statistical Analysis of Preferences,
with Applications to Social Science Data

Michael Pearce

Chair of the Supervisory Committee:
Elena A. Erosheva
Department of Statistics, School of Social Work, and the
Center for Statistics and the Social Sciences

Preference data, such as rankings and ratings, are prevalent in the social sciences for expressing and measuring attitudes or opinions. Oftentimes, deterministic algorithms or summary statistics are used to aggregate preferences, which lack the ability to measure uncertainty or identify preference heterogeneity in a population. This dissertation proposes new methodologies for statistical preference analysis that aid accurate estimation, inference, and decision-making with preference data in social science applications.

Motivated by previous attempts to integrate ordinal and cardinal data in psychometrics and computer science, we propose two of the first joint statistical models for rankings and ratings. Our models exploit the distinct and complementary properties of rankings and ratings to estimate fine-grained preferences in a population and identify potential heterogeneity. The proposed models impose few assumptions and permit many common preference data types, allowing their use in a variety of applications. We propose computationally efficient frequentist and Bayesian estimation frameworks, and apply the models to real peer review and preference survey data.

Additionally, we propose a Bayesian methodology for estimating rank-clusters from rankings. Rank-clusters denote cases when objects in a collection are equal in quality and thus should be clustered in their population-level rank. We extend previous frequentist work on

rank-clustering pairwise comparison data to permit analysis of more flexible ordinal data types. Furthermore, the model relies on a Bayesian framework that naturally allows for incorporating prior information and uncertainty quantification. We apply our model to real ranked-choice election data to analyze voters' perceptions of candidates.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AIBS: American Institute of Biological Sciences.

BTL: Bradley-Terry-Luce, a family of distributions for ordinal data that includes the Bradley-Terry and Plackett-Luce distributions (sometimes called the generalized Bradley-Terry).

COMPLETE RANKING: a ranking in which all objects in a collection are considered and ranked.

CONSENSUS: the quality of a population exhibiting similar preferences.

CONSENSUS RANKING: see *overall ranking*.

GROUPWISE COMPARISON: a type of ordinal preference data in which three or more objects are compared, but not the entire collection.

HETEROGENEOUS PREFERENCES: when groups within a population exhibit distinct preference ideologies, e.g., distinct political ideologies among members of different political parties.

IIA: Independence from Irrelevant Alternatives, a criterion from Social Choice Theory which is satisfied when the population-level preference between two objects depends only on the individual preferences between those two objects.

INCOMPLETE RANKING: a ranking formed by a judge who only considers a subset of the complete collection of objects; e.g., rankings formed under *separate ballots*.

INDUCED RANKING: a ranking formed by ordering objects on the basis of their ratings.

JUDGE: an individual or system expressing preferences on a collection of objects.

MFM: Mixture of Finite Mixtures, a Bayesian approach to cluster analysis in which both the number of clusters and cluster-specific parameters are estimated simultaneously.

NIH: National Institutes of Health.

OBJECT: an item which is assessed or judged based on its perceived worth, value, or quality.

OVERALL RANKING: the ordering of objects from best to worst with respect to the preferences of an entire group or population. This term may be used interchangeably with *consensus ranking*, *social order*, or *modal ranking*.

PAIRWISE COMPARISON: a type of ordinal preference data in which only two objects are compared.

PARTIAL RANKING: a ranking in which all objects are considered but only a subset are ranked, e.g., a top-3 ranking of 10 objects.

QUALITY: the overall value or worth of an object, which is assumed to exist and be measurable.

RANK: the level or place an object is assigned in a ranking, e.g., first place or rank 1.

RANK-CLUSTER: a group of objects that are assigned an identical rank in an overall ranking. Rank-clusters are formed among objects that are indistinguishable or identical in quality at the population level.

RANKING: a relative ordering of objects from best to worst.

RATING: an absolute assessment of an object on the basis of a pre-defined scale; sometimes called a score.

SEPARATE BALLOTS: the scenario in which each judge sees only a subset of objects from the complete collection when expressing preferences; see *incomplete ranking*.

ACKNOWLEDGMENTS

First and foremost, thank you to my advisor, Professor Elena A. Erosheva. I appreciate your encouragement and understanding as I encountered both personal and professional challenges during graduate school. You perfectly balanced structure with flexibility, allowing me to learn from your expertise while still developing as an independent researcher.

Thank you to my dissertation committee. I especially would like to thank Professor Adrian E. Raftery for sparking my interest in Bayesian statistics through his course on statistical demography, graciously welcoming me into his working group, and mentoring me on our joint research project that is not included in this dissertation. I would also like to thank Professor Marina Meilă for her constructive feedback on this dissertation, and Professor Conor Mayo-Wilson for his kindness and support during the final stages of this Ph.D.

Thank you to the many additional individuals who influenced this work. Specifically, thank you to Professors Abel Rodriguez, Yen-Chi Chen, and Carole Lee at the University of Washington, Professors T. Brendan Murphy and I. Claire Gormley at University College Dublin, and Dr. Stephen Gallo at the American Institute of Biological Sciences. Our discussions provided invaluable insights that elevated this dissertation.

Last but not least, thank you to my friends and family for your constant support throughout my time in Seattle. I could not have made it here without you.

DEDICATION

To my grandfather, Pete Phillips.

Rest in Peace

Chapter 1

INTRODUCTION

This dissertation proposes application-driven methodologies for the statistical analysis of preferences. Despite the ubiquity of preference data in many social science disciplines, methods to analyze preferences are underdeveloped. This chapter provides an overview of the primary contributions of the dissertation to statistical methodology and social science applications.

1.1 Contributions to Statistical Methodology

Rankings and ratings are two common types of preference data that are usually analyzed independently. However, rankings and ratings arise simultaneously in many settings and exhibit distinct and complementary properties that could be exploited (e.g., Biernat (1995); Ovadia (2004)). Chapter 2 describes these properties and summarizes existing techniques for analyzing preference data. Until recently, no statistical models existed for their joint analysis without reliance on data conversion.

In Chapter 3, we fill this gap by proposing the first joint statistical model for rankings and ratings, the Mallows-Binomial. Our novel methodology estimates preferences in a population via shared parameters between ranking and rating component distributions. Through theory and simulation studies, we demonstrate desirable statistical properties of the model, such as identifiability and consistency of maximum likelihood estimators. Furthermore, we propose computationally efficient algorithms for frequentist estimation to address known challenges with estimating related models, and prove the asymptotic validity of bootstrapped estimates of standard errors. Still, the Mallows-Binomial has certain limitations: The model cannot handle pairwise comparison data, does not satisfy the desirable Luce's Choice Axiom, and

lacks a framework for estimating preference heterogeneity.

These limitations lead us to propose a second joint statistical model for rankings and ratings in Chapter 4, the Bradley-Terry-Luce-Binomial (BTL-Binomial). BTL-Binomial satisfies Luce's Choice Axiom and allows for complex ordinal data, such as pairwise comparisons and rankings made under *separate ballots*. Furthermore, we develop the model in a Mixture of Finite Mixtures framework (Miller and Harrison 2018), which allows for Bayesian estimation of both the level and type of preference heterogeneity in a population. We compare assumptions and properties of the Mallows-Binomial and BTL-Binomial models in Chapter 5.

Chapter 6 shifts gears by addressing rank-clustering of objects based on ranking preference data. In this context, rank-clustering refers to the scenario in which multiple objects are indistinguishable or equal in quality at the population level, and thus are clustered in their overall rank. Existing work in the literature has applied frequentist techniques, such as the fused lasso (Tibshirani et al. 2005), to pairwise comparison data for the purpose of inducing rank-clusters (e.g., Masarotto and Varin (2012); Vana et al. (2016). However, these specific methods have not been developed for richer types of ordinal data, such as partial or complete rankings. Drawing on techniques from Bayesian variable selection, we propose a novel spike-and-slab prior for variable fusion that induces rank-clusters via partitions. The prior is applied to the flexible BTL family of ranking distributions in what we call the Rank-Clustered BTL model.

1.2 Contributions to Social Science Applications

Chapters 3 and 4 are primarily motivated by the setting of quality assessment in scientific peer review. A large literature documents myriad concerns regarding the accuracy and fairness of current peer review processes (see Chapter 2.5). Our proposed Mallows-Binomial and BTL-Binomial models provide formal mechanisms for sharing the distinct and complementary preference information provided by rankings and ratings into a unified approach to peer review quality assessment. The models are applicable under a variety of realistic peer

review settings, including preferences made under separate ballots and reviewer heterogeneity. Furthermore, we demonstrate our models' ability to accurately estimate preferences with uncertainty and aid informed decision-making.

Our joint models are useful to social science settings beyond peer review. In this work, we analyze survey responses that measure preferences using both ordinal and cardinal data. Even in the presence of respondent preference heterogeneity and substantial data missingness, we estimate heterogeneous preferences alongside their inherent uncertainty. The application suggests our models may be applicable to any social science field with interest in understanding the preferences or beliefs of a population. These fields may include political science, psychology, sociology, and education, among others.

Mallows-Binomial and BTL-Binomial address existing gaps in the literatures on measurement and psychometrics, which have long been documenting the differences and deficiencies of ratings and rankings as expressions of preferences (Alwin and Krosnick 1985; Russell and Gray 1994; Sung and Wu 2018). As a result, numerous elaborate variations of measurement approaches have been proposed that attempt to combine the good qualities of each data type (e.g., Smith and Kendall (1963); Goffin et al. (2009)). In contrast, our joint models provide simple and principled methods for joint analysis of ordinal and cardinal preference data without conversion or modification from their original form.

In Chapter 6, we apply our Rank-Clustered BTL model to ranked choice voting data from the 2021 Minneapolis mayoral election to simultaneously estimate the overall ordering of candidates and rank-clusters of candidates. Related work on variable fusion with pairwise comparison data suggests future applications in diverse fields such as sports modeling and academic rankings (e.g., Tutz and Schauberger (2015); Varin et al. (2016)).

1.3 Organization of Dissertation

To summarize, Chapter 2 provides important background information on preference data and analysis, Social Choice Theory, and peer review. Chapters 3 and 4 present statistical models for joint analysis of rankings and ratings. Specifically, Chapter 3 proposes the Mallows-

Binomial model in a frequentist framework, while Chapter 4 proposes the BTL-Binomial model in a Bayesian framework that allows for the estimation of preference heterogeneity. Theoretical and practical qualities of the two joint models are compared in Chapter 5. Chapter 6 proposes a Bayesian methodology for estimating rank-clusters among objects in ordinal preference data. Last, Chapter 7 discusses the complete work and proposes areas for future research.

Chapter 2
BACKGROUND

In this chapter, we provide key background information that will be useful for understanding the subsequent chapters of this dissertation. Specifically, we define key concepts from the fields of preference learning and Social Choice Theory, review existing methods for analyzing preference data, and describe our primary application area, peer review.

2.1 Preference Data

Preference data is common to our world: Citizens express preferences through voting in elections, critics rank movies when creating annual top-10 lists, judges score figure skaters in the Olympics using numerical scales, wine critics use Likert scales with words such as "mediocre" to rate wines, the Google PageRank algorithm sorts webpages based on relevance to a query, and so on. Although preference data exists in many forms, in all cases it expresses the preferences of a *judge* on one or more *objects*. Oftentimes, preference data is provided in response to a question or query, a process known as *preference elicitation*.

As can be seen from the previous examples, two common types of preference data are *rankings* and *ratings*. We describe each type in turn before comparing their properties.

2.1.1 Rankings

Rankings are a type of ordinal preference data that denote a relative ordering of objects from best to worst (potentially allowing ties). In a ranking, an object's *rank* is the place it receives in the ranking. Although some authors have drawn a distinction between the terms "ranking" and "ordering," in this dissertation we choose to use solely the former in accordance with its popular usage. For example, a voter may provide the ranking {Candidate A $\prec$

Candidate B $\prec$ Candidate C} to suggest that they prefer Candidate A first, B second, and C third. Rankings provide relative judgements by utilizing other objects as reference points. Thus, rankings are thought to provide objective comparisons between ranked objects because they require the judge to make explicit comparisons that are scale-free (Biernat 1995). However, rankings force demarcation even when it may not exist and may lack granularity in comparisons. For example, given the ranking $\{A \prec B \prec C\}$, there is no way to determine if A and B are nearly tied, or if C is far less-preferred than B.

Rankings arise in different forms. Given a collection of objects, a ranking is called *complete* when all objects are ranked. In contrast, a ranking is called *partial* when all objects were considered, but only a subset of the most-preferred are ranked (e.g., a top-5 ranking). In a partial ranking, we assume that unranked objects are less-preferred than those ranked, but also that the preference order among the unranked objects is unknown. Next, we call a ranking *incomplete* when a judge is asked only to rank a subset of the complete collection of objects. In incomplete rankings, no information can be gleaned regarding objects not considered. For example, if a judge is asked to rank the music genres "classical" and "jazz", the ranking should provide no information on their preferences regarding any other music genre. We call incomplete rankings involving two objects a *pairwise comparison*, and incomplete rankings involving more than two objects a *groupwise comparison*. Rankings may be both partial and incomplete (e.g., a top-3 ranking of mayoral candidates, but only among the 5 candidates from a specific political party).

2.1.2 Ratings

Ratings are a form of cardinal preference data, which are sometimes called *scores*. Ratings are absolute judgements in the sense that they do not directly use other objects as reference points. Instead, ratings reflect preferences in relation to some standard or target level of performance that is indicated with verbal descriptions of a scale: low to high, poor to excellent, etc. When the scale has many values, ratings provide granular assessments. Furthermore, when ratings are calibrated across judges, they allow for global comparisons. However, there

are issues which limit the use of ratings for making such global comparisons. For example, some judges may be naturally more lenient or harsh (e.g., one judge's 5/10 may be another's 7/10), or may become cognitively burdened by the number of scores they need to provide and stop expressing internally consistent scores (Johnson 2008; Wang and Shah 2018; Poston 2008; Griffin and Brenner 2004). For these reasons, scores have sometimes been described as highly subjective and inconsistent in different bodies of literature (Biernat 1995; Biernat and Kobrynowicz 1997; Biernat et al. 2009; Mallard et al. 2009). When a single judge provides more than one rating, they may be interpreted as allowing the judge to make relative comparisons implicitly. However, it has been observed that ties are commonly present when rating two or more objects, thus limiting the use of ratings for demarcating (Feather 1973; Shah et al. 2018).

Like rankings, ratings also arise in different forms. Ratings are generally provided on a pre-defined numerical scale. The specific numerical scale used depends on context and the range of reasonable options. Thus, ratings may be treated as continuous or discrete measures. Ratings may also be obtained via conversion from Likert-type scales (Likert 1932). Such ratings have been criticized on the grounds that they only reflect qualitative ordinal judgements and do not possess interval properties, in that numeric differences between any two values may not be meaningful. We note that these criticisms, while valid, do not prevent the widespread use of numerical summaries for high-stakes decisions, as in federal research grant funding (National Institutes of Health 2021).

2.1.3 Comparison Between Rankings and Ratings

Rankings and ratings exhibit many distinct but complementary properties. Many of these contrasting properties have already been suggested by our descriptions: Rankings provide ordinal, scale-free, coarse, and objective comparisons, while ratings provide cardinal, scaled, granular, and subjective comparisons. We may consider ratings as providing more information than rankings, as they may be ordered into an *induced ranking* that also contains information on the relative distances between the induced rank places.

Psychological and psychometric literatures have long been documenting the comparative properties and deficiencies of both ratings and rankings as expressions of preferences. Rankings have been criticized for imposing high cognitive load, especially when the number of options is large (Alwin and Krosnick 1985); for potentially forcing judges to make invalid distinctions in cases that have low discriminability (Russell and Gray 1994); and for being difficult to analyze or summarize by means of common statistical techniques (Sung and Wu 2018). On the other hand, ratings have been criticized for providing measurements that are only coarsely granular as well as for allowing judges to use the same numeric value for more than one case (Russell and Gray 1994). Comparative judgements are less susceptible to noise: research on job performance measurement (industrial/organizational psychology), measurement of attitudes (social psychology), and person perception (personality psychology) on comparative and absolute judgements has found that rankings have better validity and may have more accuracy than absolute ratings (Goffin and Olson 2011). To reconcile and draw on both approaches, psychological research has moved to suggest variations of measurement approaches that try to combine good qualities of both expressions of preferences. Examples include—but are not limited to—the Behaviorally Anchored Rating Scale (Smith and Kendall 1963); the Visual Analogue Scale for Rating, Ranking, and Paired Comparison (Sung and Wu 2018); and the Relative Percentile Method (Goffin et al. 2009). As the respective names suggest, these measurement approaches offer much more elaborate and time-consuming data collection mechanisms as compared to either rankings or ratings.

2.2 Preference Learning

Under a general definition, preference learning is "the problem of learning from observations which reveal, either explicitly or implicitly, information about the preferences of an individual (e.g., a user of a computer system) or a class of individuals" (Fürnkranz and Huellermeier 2010, page v).

The output of a preference learning problem varies. In general, models output summary statistics of the overall group preferences. In the machine learning or AI literature, the

problem is often formulated in terms of a utility function which aggregates preference data in terms of preference relations; the utility function is often learned using training data and later used for prediction (Fürnkranz and Huellermeier 2010). The output may be an estimated *consensus ranking*, which expresses the overall preferences of a population. In other contexts, the consensus ranking is called a *social order*, *modal ranking*, or simply the *overall ranking*. For simplicity, we use only the term consensus ranking in this subsection.

Still, the term consensus ranking is somewhat of a misnomer. While consensus rankings are meant to reflect the overall preferences of a population, they may not reflect actual consensus. Even when the consensus ranking reflects true consensus, that consensus may be weak (e.g., a moderate proportion of judges exhibit preferences that align only partially with the identified consensus ranking). Additionally, a single consensus ranking may be an inappropriate method to summarize preference data when judges exhibit *heterogeneous preferences*, which occur when judges use distinct ideologies when deciding and expressing preferences. In such cases, we may consider whether consensus exists locally within subgroups of the judges and form consensus rankings among them. A real-world example of this phenomenon is in political preferences, where voters of different political parties exhibit substantially different preferences when expressing their preferred candidates for office. Furthermore, some objects may be of equal quality or indistinguishable in order by the populations of judges. In such cases, an accurate consensus ranking could potentially include rank-clusters of objects.

Consensus rankings are sometimes formed deterministically under the rules of a system. For example, in instant-runoff voting systems, well-defined procedures dictate how the ranked votes provided by constituents (judges) of the candidates (objects) are tabulated to determine the winner of the election. Here, the winning candidate is an explicit winner of the election, but the remaining candidates are implicitly ranked according to when in the tallying process they are removed from consideration. In an example with ratings, movies may be ordered into a consensus ranking according to their average rating (e.g., 3.56 stars out of 5) assigned by critics or general audiences. In this context, movies are ranked higher or lower than the others based solely upon their average rating. Therefore, the consensus ranking does

not depend upon the number of ratings available for each movie nor how close the average ratings are between movies.

2.3 Social Choice Theory

Social Choice Theory refers to the analysis and aggregation of individual preferences to understand the overall preferences of a group or population. Here, preferences are aggregated using a *social welfare function*, which takes as input the preferences and returns as output the ordered preferences of the group, called the *social order*, which is akin to the consensus ranking discussed previously (Sen 1986).

A primary result from Social Choice Theory is *Arrow's Impossibility Theorem* (Arrow 1950). The theorem states that in a system with 3 or more objects, there is no social welfare function that simultaneously satisfies all of the following desirable criteria: *unrestricted domain*, *non-dictatorship*, *Pareto efficiency*, and *independence from irrelevant alternatives* (IIA). Unrestricted domain states that any coherent voter preference is allowed. Non-dictatorship requires that there is no single, specific voter whose preferences always prevail in the system; anonymous voting methods satisfy this condition. Pareto efficiency states that if every judge prefers one object to another, the social order will as well. Finally, IIA states that the social order between two objects only depends on the individual preferences between those two objects. IIA is related to Luce's Choice Axiom, which states that the probability of selecting one object over another is unaffected by the presence or absence of other objects (Luce 1959). Luce's Choice Axiom is a stronger condition that implies IIA.

Arrow's Theorem may be interpreted as suggesting that there is no "perfect" system of aggregating preferences[1]. Nonetheless, the Social Choice literature is full of proposed social

[1]There are some cases in which it may be beneficial to *not* satisfy one or more of the criteria specified by Arrow's Theorem. As examples, medal orders are determined in Olympic sport climbing using a rank-multiplication system across multiple events (Nguyen et al. 2021), and in US secondary and collegiate cross-country running competitions using a rank-sum system across teams (Hammond 2007), which are known to violate Social Choice Theory axioms such as transitivity and IIA (Nguyen et al. 2021; Hammond 2007; Mixon Jr and King 2012; Boudreau et al. 2018). However, Boudreau and Sanders (2015) argue that these systems actually increase competitive balance (i.e., "level the playing field") and thus increase fan engagement and enthusiasm.

welfare functions. For example, the Kemeny-Young method identifies the social order with the smallest total distance (under a certain metric) to the observed preferences (Kemeny 1959; Young 1988). Kemeny-Young is an example of a Condorcet method, which identifies a Condorcet winner if one exists. A Condorcet winner is one that wins a majority of all pairwise comparisons against each of the other objects. Another example is the Borda rule, which assigns scores to an object based on the number of objects ranked lower and tallies scores to determine the social order (de Borda 1781). Interestingly, many of these ideas rely on preferences expressed as rankings, yet convert ranking information into numerical ratings. The methods also provide deterministic outcomes; no statistical uncertainty is introduced or expressed.

2.4 Statistical Models for Preference Learning

In this section, we review statistical models for preference learning based on (1) rankings, (2) ratings, and (3) rankings and ratings jointly. In contrast to deterministic preference aggregators or utility functions, statistical models for preference data can explore the uncertainty inherent to estimated preferences. Specifically, they may be used to identify overall or local consensus, measure its strength, and ultimately provide information on the uncertainty of the estimated consensus ranking.

2.4.1 Ranking Models

Statistical ranking models have been proposed since at least the early 20th century (Marden 1996). Thurstone (1927) modeled rankings using order statistics of Normal distributions. Another method to study rankings is through pairwise comparisons. The Bradley-Terry model (proposed by Zermelo (1929) and discovered independently by Bradley and Terry (1952)) specifies the probability that object i will be ranked above object j using the likelihood function,

$$P[i \succ j | p_i, p_j] = \frac{p_i}{p_i + p_j} \tag{2.1}$$

where $p_i, p_j \geq 0$. The model was extended to allow for multiple comparisons in the Plackett-Luce model (sometimes called the generalized Bradley-Terry model), which was proposed by Plackett (1975) and justified under Luce's Choice Axiom. In this model, a ranking $\pi = \{1 \prec 2 \prec \cdots \prec J\}$ of J objects is assigned probability

$$P[\Pi = \pi | p_1, \ldots, p_J] = \prod_{j=1}^{J} \frac{p_j}{\sum_{j'=j}^{J} p_{j'}} \tag{2.2}$$

where often we set $\sum_j p_j = 1$ for identifiability. Rankings drawn from the Plackett-Luce model may be interpreted as being created sequentially, where in the first stage an object is selected among all the options, in the second stage an object is selected among all the remaining, and so on. Together, we refer to the Bradley-Terry and Plackett-Luce distributions as the Bradley-Terry-Luce (BTL) family.

Many extensions of BTL models have been proposed, such as those that incorporate judge-level covariates (Gormley and Murphy 2010) or object-level covariates (Chapman and Staelin 1982; Tutz and Schauberger 2015). Furthermore, mixtures of BTL distributions have been studied by numerous authors to account for heterogeneous populations of judges (Gormley and Murphy 2006, 2008; Gormley et al. 2009; Mollica and Tardella 2017; Zhao et al. 2016; Chierichetti et al. 2018; Liu et al. 2019; Zhao and Xia 2019; Zhang et al. 2022). Estimation of BTL models is often challenging due to the data-dependent normalizing constant. However, various works have proposed speedy and accurate estimation methods in both frequentist and Bayesian settings (Hunter et al. 2004; Guiver and Snelson 2009; Caron et al. 2014; Maystre and Grossglauser 2015; Mollica and Tardella 2017; Turner et al. 2020; Nguyen and Zhang 2023).

Another common model for rankings is the Mallows model (Mallows 1957). The Mallows model is sometimes called the normal distribution for rankings, as it is a location-scale family in which the centrality parameter, π_0, is the consensus ranking itself and the scale parameter $\theta \geq 0$ dictates how likely rankings of a given distance to π_0 are to be drawn, where the probability decreases exponentially with the distance. Specifically, the probability

of drawing a ranking π from a Mallows(π_0, θ) distribution is

$$P[\Pi = \pi | \pi_0, \theta] = \frac{e^{-\theta d(\pi, \pi_0)}}{\psi(\theta)} \tag{2.3}$$

where $d(\cdot, \cdot)$ is a distance metric and $\psi(\theta)$ is a function which provides an appropriate normalizing constant. Two common choices for distance metrics are based on Kendall's τ (Kendall 1938), leading to the Mallows ϕ model, and Spearman's ρ (Spearman 1904), leading to the Mallows θ model. The Kendall distance is the minimum number of adjacent object swaps needed to convert one ranking into another. The Spearman distance is the squared Euclidean distance between two rankings. Henceforth referred to as simply the Mallows model, the Mallows ϕ model has received particular attention as a natural fit in many ranking applications. In seminal works, Fligner and Verducci (1986) and Fligner and Verducci (1988) extend the model to allow for partial, top-R rankings and explicitly define the normalizing constant in closed form,

$$\psi_{R,J}(\theta) = \prod_{j=1}^{R} \frac{1 - e^{-\theta(J-j+1)}}{1 - e^{-\theta}}. \tag{2.4}$$

Furthermore, Fligner and Verducci (1986) propose the Generalized Mallows model that introduces rank level-specific scale parameters. More recently, Meila and Bao (2010) proposed the Infinite Generalized Mallows model to aggregate rankings over infinite collections of objects. Mixtures of Mallows models have been studied by Marden (1996), Murphy and Martin (2003), Busse et al. (2007), Lu and Boutilier (2011), and Collas and Irurozki (2021).

Due to its location-scale form, the Mallows model may be considered as a holistic selection model (in contrast to the sequential selection form of the Bradley-Terry-Luce family). Under the Mallows, rankings are drawn with probability that depends on their overall distance to the central permutation, and not on the specific ordering of objects in that permutation. As a result, the Mallows model does not generally satisfy the IIA criterion or Luce's Choice Axiom (Marden 1996). For completeness, we provide a counterexample: Suppose $\pi_0 = 1 \prec 2 \prec 3$

and $\theta > 0$ is fixed. Across all rankings of three objects, the overall probability that $1 \prec 3$ is,

$$\begin{aligned}\Pr[1 \prec 3] &= \Pr[\{1 \prec 2 \prec 3\}] + \Pr[\{1 \prec 3 \prec 2\}] + \Pr[\{2 \prec 1 \prec 3\}] \\ &= \frac{1 + 2e^{-\theta}}{\psi(\theta)}. \end{aligned} \tag{2.5}$$

However, suppose object 2 is ranked first. The conditional probability that $1 \prec 3$ is then,

$$\begin{aligned}\Pr[1 \prec 3 | 2 \text{ is ranked first}] &= \frac{\Pr[\{2 \prec 1 \prec 3\}]}{\Pr[\{2 \prec 1 \prec 3\}] + \Pr[\{2 \prec 3 \prec 1\}]} \\ &= \frac{e^{-\theta}}{e^{-\theta} + e^{-2\theta}}. \end{aligned} \tag{2.6}$$

Since the probabilities in Equations 2.5 and 2.6 are not equal regardless of θ, the Mallows model does not satisfy Luce's Choice Axiom or IIA.

Estimation of the Mallows model is often challenging due to the discrete centrality parameter π_0. In a frequentist setting, Meila et al. (2012) proved that the MLE of π_0 is precisely the solution in the Kemeny-Young model. Later research proposed computationally efficient tree-search algorithms for exact computation of the MLE and approximate estimation methods to use when the number of objects is large or when consensus is weak (Mandhani and Meila 2009; Meila and Bao 2010). Tang (2019), He (2022), and Busa-Fekete et al. (2021) study other asymptotic and finite-sample properties of the Mallows model in frequentist settings such as bias, consistency, and hypothesis testing. Vitelli et al. (2018), Liu et al. (2019), and Crispino and Antoniano-Villalobos (2022) study Bayesian estimation of the Mallows model.

2.4.2 Rating Models

Ratings are rarely modeled statistically. Instead, simple summary statistics such as the mean or median, or variations thereof, are commonly used (Lee et al. 2013; Tay et al. 2020; National Institutes of Health 2021). For example, the *trimmed mean*, which is defined as the mean rating after removal of the highest and lowest values, is used in Olympic figure skating in an effort to reduce the effects of rating anomalies and bias (Emerson and Arnold 2011).

Although uncommon, ratings can be modeled using a variety of standard probability distributions. As continuous measures, the Normal, Truncated Normal, Beta, or Exponential distributions, among others, may be appropriate depending on the range of allowable ratings and the patterns observed. Discrete ratings often arise from an ordinal, well-defined, and equally-spaced set. In such cases, we may linearly transform the set of allowable ratings into a set of integers that match the support of an appropriate probability distribution. If the original set is theoretically infinite (e.g., the number of goals scored by a soccer team; see Egidi and Torelli (2021)), the Poisson or Negative Binomial distributions may be appropriate. If the original set is discrete (e.g., Likert scale, ratings between 1 and 5 in single decimal increments), the Binomial or Beta-Binomial may be appropriate. In these parametric models, a central tendency parameter is useful for interpretability. Nonparametric rating models may also be reasonable depending upon the data context (Munzel and Bandelow 1998).

2.4.3 Joint Models for Rankings and Ratings

In recent years, a growing body of literature has suggested that collecting and analyzing both rankings and ratings may be beneficial for understanding preferences. That is, using rankings and ratings in tandem may retain the benefits and minimize the downsides of each data type, and thus remove an existing false dichotomy (Belkin et al. 1995; Lee 1997; Ovadia 2004; van Herk and van de Velden 2007; Macdonald and Ounis 2009; de Chiusole and Stefanutti 2011; Balog et al. 2012; Shah et al. 2018; Liu et al. 2022; Su 2022). However, few such methods exist, and none of the statistical models described in Sections 2.4.1 or 2.4.2 can be used directly to jointly model rankings and ratings. Literatures in a variety of disciplines have navigated this issue in different ways.

In the social and health sciences, the literature on *mixed-outcomes* includes proposed methods for combining preference data of different types via *conversion*, such as converting rankings into ratings or ratings into rankings prior to performing a statistical analysis (Salomon 2003; Kim et al. 2015; Venkatraghavan et al. 2019). Converting ratings into rankings is generally straightforward, as ratings from a single judge may be converted into a ranking

by simple ordering. The reverse is not true, and often requires a model or other assumptions to convert rankings into ratings. Li et al. (2009) provide one such method, in which ranking data is converted into ratings for each judge such that the first-ranked object receives that judge's best rating, the second-ranked object receives that judge's second-best rating, etc. However, in general we find that conversion is suboptimal because it distorts or discards the observed data.

Motivated by problems in metasearch, information retrieval, and peer review, authors in computer science have proposed algorithmic approaches for merging ordinal and cardinal preferences that are known as *data fusion* (Fagin 2002; Hsu and Taksa 2005; Bhamidipati and Pal 2008; Li et al. 2009; Ailon 2010; Somers et al. 2017; Zeng and Shen 2020; Xu et al. 2020; Hochbaum and Moreno-Centeno 2021; Liu et al. 2022). In general, these algorithmic methods do not allow for the quantification of uncertainty. Furthermore, many assume that a judge's rankings and ratings must be internally consistent, an assumption that is often unrealistic in practice (Biernat et al. 1998; Biernat and Vescio 2002; Biernat 2003; Kamishima 2003; Biernat et al. 2009; Gallo 2020) and has to be enforced during data collection.

2.5 Peer Review

We conclude this chapter with an introduction to scientific peer review. Peer review may refer to a broad range of activities, such as the review of funding applications or the assessment of research for admittance into journals or conferences (Lee et al. 2013). A fair and unbiased peer review process is important to ensure the high quality of disseminated research and to create fairness in the scientific community.

Many potential issues and criticisms with peer review have been identified (see Lee et al. (2013) for a thorough review). These criticisms include, but are not limited to, purportedly low inter-rater reliability of assessments (McGraw and Wong 1996; Brezis and Birukou 2020; Pier et al. 2018; Resnik and Elmore 2016; Erosheva et al. 2021); discrepancies with respect to author characteristics such as prestige, affiliation, nationality, language, race, or gender (Wenneras and Wold 2010; Grant et al. 1997; Bornmann et al. 2007; Erosheva et al. 2020; Lee

et al. 2020; Ginther et al. 2018); discrepancies with respect to reviewer characteristics such as differences in leniency or harshness among reviewers from certain fields or social groups (Hug and Ochsner 2022); and content-based discrepancies such as preferences for basic or translational research, confirmation bias, or conservatism (Armstrong 1997; Mallard et al. 2009; Hug and Ochsner 2022).

The subjectivity of peer review assessments is a common concern. For NIH funding applications, Johnson (2008) demonstrated a "discussion effect" between pre- and post-discussion ratings, in which systematic shifts were observed in ratings after group discussion that may or may not be related to funding applications' scientific merits. Furthermore, Johnson (2008) found a general tendency of some reviewers to be more or less stringent when assigning ratings. In a series of works, Biernat (1995); Biernat and Kobrynowicz (1997); Biernat et al. (2009) described the subjectivity of ratings based on a "shifting standards model", or the idea that the scale of ratings is context- or individual-dependent and may be impacted by reviewer bias. Instead, Biernat (1995) argued that rankings constitute a more "objective" form of preference data since they rely on absolute comparisons. Relatedly, Wang and Shah (2018, 2020) discussed how ratings often suffer from "arbitrary miscalibrations." That said, they also demonstrated that conversion of ratings into rankings is worse from the perspective of statistical risk, even with adversarially chosen shifting standards.

These works demonstrate the potential challenges and complexities of quality assessments in peer review. We propose alternative methods for conducting peer review in Chapters 3 and 4 that allow for collecting and analyzing assessments in the form of rankings and ratings simultaneously. As we will demonstrate, our models allow for better separation of proposals based on their perceived quality, estimate the relative and absolute quality of proposals with uncertainty, and aid accurate and transparent decision-making.

Chapter 3

THE FIRST JOINT STATISTICAL MODEL FOR RANKINGS AND RATINGS

This chapter is based on (Pearce and Erosheva 2022a) and (Pearce and Erosheva 2022b). For consistency with the rest of this dissertation, the word "scores" in the original published works has been replaced by "ratings" throughout. We would like to thank Dr. Stephen Gallo (formerly of the American Institute of Biological Sciences) for providing the data used in this chapter, and Dr. Yen-Chi Chen for his helpful insights regarding the use of the nonparametric bootstrap for estimating standard errors in our model. This work was supported by the National Science Foundation under Grant No. 2019901.

3.1 Introduction

Rankings and ratings arise simultaneously in a number of contexts. In peer review, judges may rate proposals numerically and subsequently rank their top few favorites (Gallo 2023). In information retrieval, distinct algorithms may assess the relevance of documents to a query using either ratings or rankings (Hsu and Taksa 2005). In film studies, critics may rate movies as they are released and later provide year-end rankings of their favorites. In these examples, the same judges may provide both kinds of information, or distinct sets of judges may provide solely rankings or ratings.

However, there are few principled methods for analyzing rankings and ratings jointly (see Chapter 2.4.3). Existing methods are either non-statistical (and thus cannot be used to understand uncertainty in estimated preferences) or require the practitioner to convert preference data of one type into another. Data conversion is suboptimal as may distort or discard the original information, and results usually depend on the chosen data conversion

procedure.

In this chapter, we propose a unified statistical model to capture information from both rankings and ratings. Conditional on the latent qualities of objects being assessed, Binomial ratings and Mallows rankings are assumed to have independent error distributions that reflect the distinct tasks of formulating ratings and rankings and realistically allow judges to be internally inconsistent when expressing preferences using distinct data types. Model parameters quantify both the absolute and relative qualities of the objects, identify a consensus ranking, and measure the strength of consensus using an existing metric in the literature. To estimate the model, we formulate exact and approximate algorithms to find maximum likelihood estimators and demonstrate regimes in which each may be useful. In addition to simulation studies, we apply the model to real data from grant panel review in which ratings and rankings were collected from the same judges. We show how the estimated parameters can be used to learn the rank ordering of grant proposals and the associated statistical uncertainty to make funding decisions.

The rest of this paper is organized as follows. In Section 3.2, we propose the Mallows-Binomial model for rankings and ratings, and explore its assumptions and statistical properties. We develop exact and approximate frequentist estimation algorithms and compare their performance in a simulation study in Section 3.3. We illustrate the model on real ranking and rating data collected during a Fall 2020 grant panel review cycle at the American Institute of Biological Sciences in Section 3.4. The chapter concludes in Section 3.5 with a brief discussion.

3.2 Mallows-Binomial Model

Suppose a judge assesses J objects using both rankings and ratings. We assume that each object $j \in \{1, \ldots, J\}$ has a true underlying *quality*, $p_j \in [0, 1]$. We use the convention that lower values of p_j denote better quality. Let $X = [X_1 \ X_2 \ \ldots X_J]^T$ be a vector of integer ratings, where each $X_j \in \{0, 1, \ldots, M\}$ is the rating assigned to object j. Let Π be the top-R ranking of the objects, $R \leq J$, such that no ties are allowed. Π is called a *partial ranking*

when $R < J$ and a *complete ranking* when $R = J$. Rankings need not align with the order of the ratings.

We propose a joint probability model for the judge's ranking Π and ratings X,

$$\begin{aligned} P\big[\Pi = \pi, X = x | p, \theta\big] =& \frac{e^{-\theta d_{R,J}(\pi,\pi_0)}}{\psi_{R,J}(\theta)} \times \prod_{j=1}^{J} \binom{M}{x_j} p_j^{x_j} (1-p_j)^{M-x_j} \\ & p = [p_1 \ \ldots \ p_J]^T \in [0,1]^J, \ \pi_0 = \text{Order}(p), \theta > 0, \\ & X_1, \ldots, X_J, \Pi \text{ are all mutually independent,} \end{aligned} \tag{3.1}$$

where $d_{R,J}(\cdot,\cdot)$ is the Kendall τ distance between two rankings and $\psi_{R,J}(\theta)$ is the normalizing constant of a (partial) Mallows model, as seen in Equation 2.4. We refer to this model as the *Mallows-Binomial*(p, θ) distribution.

A key aspect of this model is the incorporation of two distinct types of preference data. It can be seen directly from Equation 3.1 that our model corresponds to $J + 1$ joint observations per judge, with J ratings and one (partial) ranking. The Mallows-Binomial model incorporates information from both data types without conversion to learn object quality parameters, p_j, $j = 1, \ldots, J$. The joint likelihood ties together the ratings and ranking by assuming that the modal consensus ranking of the Mallows component is the same as the ranking induced by the Binomial rating parameters, p_j, $j = 1, \ldots, J$. This formulation naturally reflects the relationship between ratings and rankings given each object's true underlying quality and the order of all objects induced by their true underlying qualities. The parameter θ is the *consensus scale parameter*, which can be interpreted exactly as in the Mallows model with respect to the rankings: Large values of θ suggest strong ranking consensus among judges. As θ decreases to 0, the model approaches a uniform distribution over the possible rankings. The Mallows-Binomial model constitutes a proper probability distribution as the product of $J + 1$ independent component distributions given the parameters (p, θ).

3.2.1 Assumptions

The Mallows-Binomial distribution makes few assumptions. Primarily, the model assumes that both rankings and ratings reflect the true underlying qualities of the objects. As a consequence, deviations from the true underlying qualities are the result of the independent error distributions corresponding to Mallows and Binomial models for rankings and ratings, respectively.[1]

On the other hand, the Mallows-Binomial does not assume that each ranking is of the same length, that the ratings and ranking of each judge align, or even that the same judges provide both rankings and ratings. Inconsistent preferences arise in the peer review context considered in Section 3.4. In our grant peer review data, judges first rate objects (grant proposals) and openly share their ratings during a panel discussion, and then provide a separate partial ranking after the discussion of all objects is completed. The partial ranking is made in private, potentially leading to changes in perception of quality. Inconsistent preferences may also arise when ratings and rankings are provided by different sets of judges. For example, in database search or information retrieval, relevancy criteria used by algorithms may arise from completely separate systems, such as when one system (e.g., a machine learning algorithm) provides numerical ratings and another (e.g., a human judge) ranks the most relevant objects. Such situations do not affect estimation or interpretation of estimated parameters; our model can still capture distinct preferences.

3.2.2 Statistical Properties

For the rest of this subsection, we explore frequentist properties of the Mallows-Binomial distribution. Specifically, we study identifiability, bias and consistency of maximum likelihood estimators (MLE), and the estimation of standard errors via the nonparametric bootstrap.

---

[1] See Chapter 5 for a detailed exploration of assumptions imposed by the Mallows-Binomial model.

Identifiability

We prove that the Mallows-Binomial(p, θ) model is identifiable via Proposition 1.

Proposition 1 *Let M, J, and R be fixed and positive integers such that $R \leq J$. Then the Mallows-Binomial(p, θ) model is identifiable.*

Proof Let $P_{p,\theta}$ denote the probability distribution of ratings X and rankings Π under a Mallows-Binomial(p, θ) model. Let $\theta_1, \theta_2 > 0$ and $p_1, p_2 \in [0,1]^J$ such that $P_{p_1,\theta_1} = P_{p_2,\theta_2}$. Then,

$$
\begin{aligned}
&P_{p_1,\theta_1} = P_{p_2,\theta_2} \\
&\iff \frac{e^{-\theta_1 d_{R,J}(\Pi, \text{Order}(p_1))}}{\psi_{R,J}(\theta_1)} \prod_{j=1}^{J} p_{1j}^{X_j}(1-p_{1j})^{M-X_j} = \frac{e^{-\theta_2 d_{R,J}(\Pi, \text{Order}(p_2))}}{\psi_{R,J}(\theta_2)} \prod_{j=1}^{J} p_{2j}^{X_j}(1-p_{2j})^{M-X_j} \\
&\iff 0 = \big(\theta_2 d_{R,J}(\Pi, \text{Order}(p_2)) - \theta_1 d_{R,J}(\Pi, \text{Order}(p_1))\big) + \log \frac{\psi_{R,J}(\theta_2)}{\psi_{R,J}(\theta_1)} + \\
&\qquad \sum_{j=1}^{J} \big[X_j \log \frac{p_{1j}}{p_{2j}} + (M - X_j) \log \frac{1-p_{j1}}{1-p_{j2}}\big].
\end{aligned}
$$

For each $j = 1, \ldots, J$, and for any arbitrary X_j, the expression $X_j \log \frac{p_{1j}}{p_{2j}} + (M - X_j) \log \frac{1-p_{j1}}{1-p_{j2}} = 0$ if and only if $p_{1j} = p_{2j}$ by the identifiability of the Binomial distribution. Thus, for any arbitrary collection $X_1, \ldots, X_J$, the final sum is 0 if and only if $p_1 = p_2$. Continuing under the assumption that $p_1 = p_2$, we have Order(p_1)=Order(p_2) and thus,

$$
P_{p_1,\theta_1} = P_{p_2,\theta_2} \iff 0 = d_{R,J}(\Pi, \text{Order}(p_1))(\theta_2 - \theta_1) + \log \frac{\psi_{R,J}(\theta_2)}{\psi_{R,J}(\theta_1)}
$$

which for any arbitrary Π is 0 if and only if $\theta_1 = \theta_2$. Therefore, the Mallows-Binomial model is identifiable. ■

Bias and Consistency

Bias and consistency of the MLE in the Mallows and Binomial distributions is a natural starting point to examine bias and consistency of the MLE in the joint Mallows-Binomial.

Tang (2019) demonstrated that in the Mallows model, the MLE $\hat{\pi}_0$ of the consensus ranking π_0 is consistent whereas its bias is difficult to quantify due to the discrete nature of the parameter, and $\hat{\theta}$ is biased upward for a finite number of judges, I, but consistent as I increases to infinity. As a univariate exponential family, $\hat{p}$ in a Binomial(M, p) distribution with M known is unbiased and consistent. Therefore, we expect Mallows-Binomial(p, θ) MLEs $\hat{p}$ and $\hat{\theta}$ to be consistent but potentially biased.

It is straightforward to prove that $\hat{\theta}$ is biased upward. This is because $\hat{\theta} = \infty$ whenever all rankings are identical to $\hat{\pi}_0 = \text{Order}(\hat{p})$, which occurs with positive probability for any $\theta > 0$. However, excluding such situations, bias is difficult to demonstrate. An illustration of small but non-zero bias can be found in Appendix A.1. On the other hand, we prove consistency of the MLE $(\hat{p}, \hat{\theta})$ via Proposition 2.

Proposition 2 *Let M, J, and R be fixed and positive integers such that $R \leq J$. Let $\theta_0 \in (0, \infty)$ and $p_0 \in (0, 1)^J$. Let $(X, \Pi)_I$ denote a sample of I independent and identically distributed samples from a Mallows-Binomial(p_0, θ_0) distribution, and $(\hat{p}, \hat{\theta})_I$ be the maximum likelihood estimators based on that sample. Then, $(\hat{p}, \hat{\theta})_I \xrightarrow{p} (p_0, \theta_0)$.*

Proof of Proposition 2 is sufficiently technical and thus relegated to Appendix A.2.

Since the magnitude of bias and rate of convergence are challenging to derive analytically, we explore these concepts through simulation. We run simulations for different values of the following parameters: the number of judges $I \in \{5, 20, 80\}$, maximum integer rating $M \in \{10, 20, 40\}$, number of objects $J \in \{6, 12, 18\}$, size of each partial ranking $R \in \{6, 12, 18 | R \leq J\}$, and consensus scale parameter $\theta \in \{1, 2, 3\}$. For each unique combination of I, M, J, R, and θ, we performed 20 simulations, where in each simulation we sampled a new object quality vector p from a Uniform$[0, 1]^J$. After examining results separately for different values of I, M, J, and R, we noticed minimal differences based on M or R. Therefore, we present aggregated results for given I and J in Figure 3.1.

The simulation indicates that the MLE $\hat{p}$ is unbiased and consistent in I, and that $\hat{\theta}$ is minimally biased and consistent in I. Estimation error and bias for p and θ appear

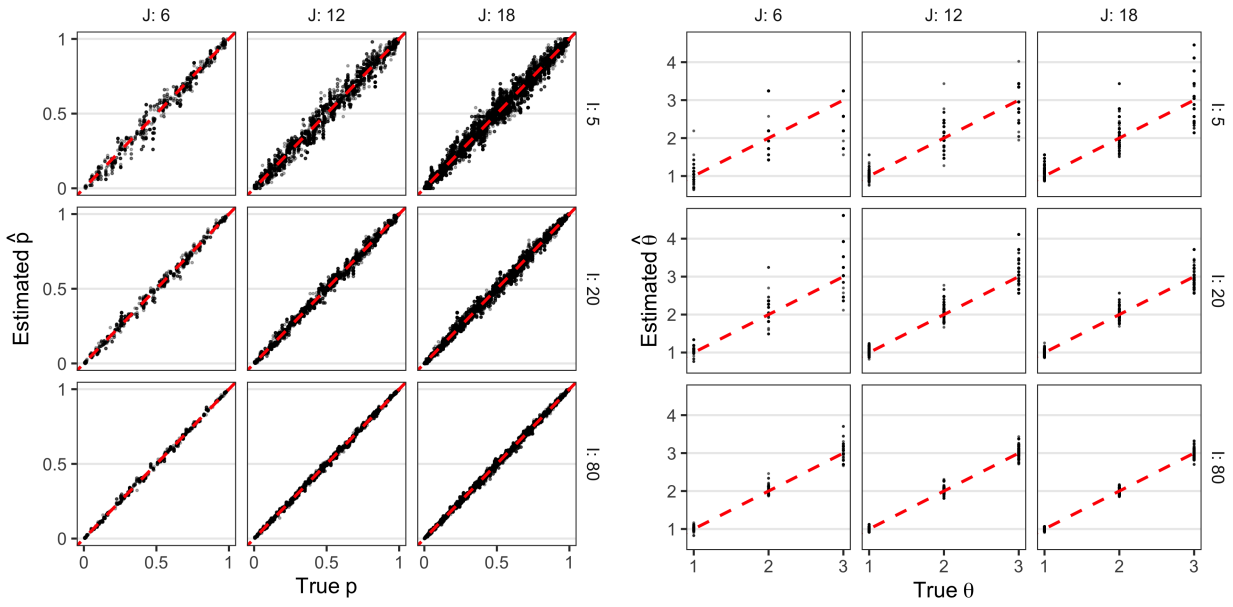

Figure 3.1: True vs maximum likelihood estimates of p (left) and θ (right) in simulated data from the Mallows-Binomial model across different values of I and J. Results are aggregated over M and R. The right panel excludes cases where all sampled rankings were equivalent ($\hat{\theta} = \infty$). Red dotted lines represent perfect estimation accuracy.

similar in scale to that when estimating Binomial probabilities or Mallows scale parameters in independent models, respectively, even for modest numbers of judges, I.

Estimation of Standard Errors

We propose estimating standard errors via the nonparametric bootstrap (Efron and Tibshirani 1994). The asymptotic validity of bootstrap estimates of standard errors is proved via Proposition 3.

Proposition 3 *Let M, J, and R be fixed and positive integers such that $R \leq J$. Let $(X, \Pi)_I$ denote a sample of I independent and identically distributed samples from a Mallows-*

Binomial(p, θ) *distribution. Then, nonparametric bootstrap estimates of standard errors for* $(\hat{p}, \hat{\theta})$ *based on* $(X, \Pi)_I$ *are parameter-wise asymptotically valid as* $I \to \infty$.

Proof of Proposition 3, as well as a more thorough description of the nonparametric bootstrap, is relegated to Appendix A.3.

Given the presence of $J+1$ parameters, we recommend a relatively large number of bootstrap samples in order to obtain a proper empirical distribution of the estimators. We also note that bootstrapped confidence intervals for $\hat{p}$ and $\hat{\theta}$ do not directly provide confidence intervals for the estimated consensus ranking of objects, $\hat{\pi}_0$. To create confidence intervals for consensus rankings, we again propose using the nonparametric bootstrap. Specifically, for each bootstrap sample and the associated MLE, the order of the estimated object quality parameters can be treated as one observation in the empirical distribution of the estimated consensus ranking. We can subsequently form confidence intervals from the empirical distribution in a straightforward manner. Conveniently, the same bootstrap samples used when creating confidence intervals for $\hat{p}$ and $\hat{\theta}$ may be used again here for computational efficiency.

3.3 Frequentist Estimation

Analytic solutions for the maximum likelihood estimator (MLE) of a Mallows distribution do not exist. Even more, finding the MLE is an NP-hard problem (Meila et al. 2012). Difficulty arises from the discrete consensus ranking, which may be one of $J!$ unique possibilities. Although the Mallows-Binomial model contains $J+1$ *continuous* parameters, $(p, \theta) \in [0,1]^J \times \mathbb{R}_{>0}$, the *discrete* order of p affects the likelihood. Thus, frequentist estimation of the Mallows-Binomial model is both a continuous and discrete problem.

The discrete aspect of estimation in the Mallows-Binomial model allows us to leverage existing Mallows model estimation algorithms. As we will demonstrate, the inclusion of ratings in the proposed model generally speeds up estimation as ratings provide information on the strength of differences in object qualities, beyond their induced ranking. Still, exact computation of the MLE is difficult, or even intractable, as the number of objects increases.

In this section, after some preliminaries, we propose exact and approximate algorithms to estimate the Mallows-Binomial MLEs and compare them in a simulation study.

3.3.1 Preliminaries

Suppose I judges assess a collection of J objects using integer ratings in the range $\{0, 1, \ldots, M\}$ and rankings of length R, where M, J, and R are all known and fixed integers and $R \leq J$,. We assume that each judge's ranking and ratings are drawn independently from the same Mallows-Binomial(p, θ) distribution, where p and θ are unknown and will be estimated via the method of maximum likelihood. Let $\pi_0 = \text{Order}(p)$, $\Pi = \{\Pi_i\}_{i=1,\ldots,I}$ denote the judges' rankings and $X = \{X_{ij}\}_{i=1,\ldots,I}^{j=1,\ldots,J}$ denote the judges' ratings.

We begin by stating a useful property of the Kendall distance: For any two specific rankings π_1, π_2 of length R and J, respectively, the Kendall distance can be written as,

$$d_{R,J}(\pi_1, \pi_2) = \sum_{j=1}^{R} V_j(\pi_1, \pi_2), \tag{3.2}$$

where $V_1(\pi_1, \pi_2)$ is the number of adjacency swaps needed to place the first object of π_1 in the first position of π_2, $V_2(\pi_1, \pi_2)$ is the number of additional adjacency swaps needed to place the second object of π_1 in the second position of π_2, and so on (Fligner and Verducci 1986). Note that each $V_j \in \{0, \ldots, J - j\}$.

Then, the joint log likelihood of the ratings X and rankings Π is,

$$\begin{aligned}
\ell(p, \theta | X = x, \Pi = \pi) &= \log \prod_{i=1}^{I} \left[\frac{e^{-\theta \sum_{j=1}^{R} V_j(\pi_i, \pi_0)}}{\psi_{R,J}(\theta)} \prod_{j=1}^{J} \binom{M}{x_{ij}} p_j^{x_{ij}} (1 - p_j)^{M - x_{ij}} \right] \\
&= \sum_{i=1}^{I} \Bigg[- \theta \sum_{j=1}^{R} V_j(\pi_i, \pi_0) - \log \psi_{R,J}(\theta) \\
&\qquad + \sum_{j=1}^{J} \left[\log \binom{M}{x_{ij}} + x_{ij} \log p_j + (M - x_{ij}) \log(1 - p_j) \right] \Bigg].
\end{aligned}$$

The maximum likelihood estimators, $(\hat{p}, \hat{\theta})$, are therefore,

$$
\begin{aligned}
(\hat{p}, \hat{\theta}) &= \arg\max_{p,\theta} \sum_{i=1}^{I} \Bigg[-\theta \sum_{j=1}^{R} V_j(\pi_i, \pi_0) - \log \psi_{R,J}(\theta) + \\
&\qquad \sum_{j=1}^{J} \Big[x_{ij} \log p_j + (M - x_{ij}) \log(1 - p_j) \Big] \Bigg] \\
&= \arg\min_{p,\theta} \left\{ \theta \sum_{j=1}^{R} \overline{V}_j \right\} + \left\{ \log \psi_{R,J}(\theta) \right\} + \left\{ \sum_{j=1}^{J} \overline{x}_j \log \frac{1}{p_j} + (M - \overline{x}_j) \log \frac{1}{1 - p_j} \right\} \\
&\equiv \arg\min_{p,\theta} \; f(p, \theta), \qquad (3.3)
\end{aligned}
$$

where $\overline{V}_j = I^{-1} \sum_{i=1}^{I} V_j(\pi_i, \pi_0)$ and $\overline{x}_j = I^{-1} \sum_{i=1}^{I} x_{ij}$. As no analytic solution exists, the function f within Equation 3.3 will be referred to interchangeably as a "cost" or "objective" function to be minimized via numerical optimization.

3.3.2 Exact Estimation Algorithms

The MLE $(\hat{p}, \hat{\theta})$ induces an ordering of the true underlying object qualities, $\hat{\pi}_0 = \text{Order}(\hat{p})$. To find the MLE, we flip the problem around. Instead of optimizing over p and θ directly, we first obtain $\hat{\pi}_0$ and then optimize for $\hat{p}$ and $\hat{\theta}$ under the constraints implied by $\hat{\pi}_0$ on $\hat{p}$.

Mandhani and Meila (2009) and Meila et al. (2012) observed for the Mallows model that $\hat{\pi}_0$ could be estimated exactly using an A* algorithm. A* is a standard graph traversal algorithm developed by Hart et al. (1968). Given a graph, A* finds the shortest path between a starting node and any terminal node. The algorithm requires a *cost function* that measures the exact cost to get from the starting node to any other node, and a *heuristic function* that estimates the remaining cost from any node to the nearest terminal node. The heuristic function is called *admissible* when it guarantees a lower bound on the remaining cost. A* provably yields the shortest path when the heuristic is admissible. A trivial, admissible heuristic always returns 0, but results in an inefficient graph search. Oppositely, a maximal or near-maximal ("tight") admissible heuristic may reduce the number of nodes traversed during the search but be burdensome to compute and slow the overall algorithm.

A* algorithms traditionally define separate cost and heuristic functions but these functions are always used together (Hart et al. 1968). Thus, at each node the algorithm sums the cost and heuristic functions to lower bound the total cost possible given the current node. Due to the interdependent nature of the model parameters, we use an equivalent method of defining a single, admissible *total cost heuristic* function which outputs a guaranteed lower bound on the total cost possible at any node in the graph. In other words, this single function is the sum of the usual cost and heuristic functions.

We propose two A* algorithms to calculate the exact MLE of the Mallows-Binomial model. Both algorithms use the same graph as in Mandhani and Meila (2009) and Meila et al. (2012) but differ based on their admissible total cost heuristic functions; the first is crude but fast to compute, the second is tight but slow. We compare their overall speed in Section 3.3.4.

Graph

We define the graph G as a tree that progressively adds one object to the ranking as you move down its branches. To specify a single starting node, we let the zero$^{\text{th}}$ layer of G be empty. In the first layer, there is a node for each object in the collection. Traversing to any specific node in the first layer constrains the corresponding object to have the lowest-valued quality parameter (but does not specify any relationships among the remaining objects). For example, at node $n = (1)$ when $J = 3$, the quality parameters are required to satisfy $p_1 \leq p_2$ and $p_1 \leq p_3$, but no relationship is specified between p_2 and p_3. Subsequent layers are successively formed from each node by adding a unique branch for each object not yet in the path to the node. Nodes in the $(J-1)^{\text{th}}$ layer are terminal as the last object is implied. For example, when $J = 3$ the node $n = (3, 2)$ is terminal as it implies the complete ordering of objects $(3, 2, 1)$. An example search graph when $J = 3$ is shown in Figure 3.2 (adapted from Mandhani and Meila (2009)).

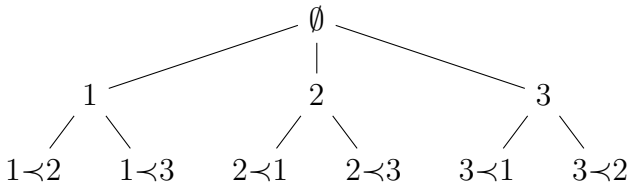

Figure 3.2: Graph for an A* search algorithm with $J = 3$ objects.

Crude Total Cost Heuristic

Before stating our first total cost heuristic, we define a useful quantity based on rankings only: Let Q be a $J \times J$ matrix such that each entry Q_{uv}, $u, v \in \{1, \ldots, J\}$, is

$$Q_{uv} = \frac{\sum_{i=1}^{I} I\{\text{object } u \text{ is ranked strictly higher than object } v \text{ in } \pi_i\}}{I} \tag{3.4}$$

When $u = v$, it follows that $Q_{uv} = 0$. If a comparison between objects cannot be deduced from any given ranking (due to partial rankings), we define the corresponding term in the numerator to be zero but do not change the denominator. Thus, $Q_{uv} + Q_{vu} = 1$ whenever a strict ordering can be deduced between objects u, v for all judges and is less than one otherwise. We are now ready to define the crude total cost heuristic.

Definition 4 (Crude Total Cost Heuristic) *Let $n \in G$ such that $n = (n_1, \ldots, n_k)$, $1 \leq k \leq J - 1$, where $n_1, \ldots, n_k$ indicate unique objects in the collection $\{1, \ldots, J\}$. Then, the crude total cost heuristic, $g_c(n) : G \to \mathbb{R}$, is*

$$g_c(n) = \Big\{\hat{\theta}^n L\Big\} + \Big\{\log \psi_{R,J}(\hat{\theta}^n)\Big\} + \Big\{\sum_{j=1}^{J} \overline{x}_j \log \frac{1}{\hat{p}_j^n} + (M - \overline{x}_j) \log \frac{1}{1 - \hat{p}_j^n}\Big\}$$

$$L = \Big(\sum_{\substack{v \in \{1:k\} \\ u \in \{(v+1):J\}}} Q_{n_u n_v}\Big) + \Big(\sum_{u,v \in \{(k+1):J\}} \min(Q_{n_u n_v}, Q_{n_v n_u})\Big)$$

$$\hat{\theta}^n = \arg\min_{\theta} \Big[\theta L + \log \psi_{R,J}(\theta)\Big]$$

$$\hat{p}^n = \arg\min_{p} \Big[\sum_{j=1}^{J} \overline{x}_j \log \frac{1}{p_j} + (M - \overline{x}_j) \log \frac{1}{1 - p_j}\Big] s.t.\ p_{n_1} \leq \cdots \leq p_{n_k},\ p_{n_k} \leq p_{n_l}, l > k.$$

The crude total cost heuristic may be seen as an extension of the quantity L from Meila et al. (2012). We prove that g_c is admissible via Proposition 5.

Proposition 5 *Under the conditions of Definition 4,*

$$g_c(n) \leq \arg\min_{p,\theta} f(p,\theta) \quad \textit{such that} \quad p_{n_1} \leq \cdots \leq p_{n_k},\ p_{n_k} \leq p_{n_l}, l > k$$

and therefore $g_c(n)$ is admissible.

Proof $g_c(n)$ consists of three terms which can each be mapped to a unique term in f. We prove the lower bound by proving (a) the first and second terms of g are a lower bound on the corresponding terms in f, and (b) the third term of g is a lower bound on the corresponding term in f.

(a) We first prove that $L \leq \sum_{j=1}^{R} \overline{V}_j$. Following closely the logic of Mandhani and Meila (2009),

$$\begin{aligned}
L &= \sum_{\substack{v\in\{1:k\}\\ u\in\{(v+1):k\}}} Q_{n_u n_v} + \sum_{u,v\in\{(k+1):J\}} \min(Q_{n_u n_v}, Q_{n_v n_u}) \\
&= \sum_{j\in\{1:k\}} \overline{V}_j + \sum_{u,v\in\{(k+1):J\}} \min(Q_{n_u n_v}, Q_{n_v n_u}) \\
&\leq \sum_{j\in\{1:k\}} \overline{V}_j + \sum_{j\in\{(k+1):J\}} \overline{V}_j \\
&= \sum_{j=1}^{R} \overline{V}_j
\end{aligned}$$

The second line above holds by definition of $\overline{V}_j$ and the third line holds since one of $Q_{n_u n_v}, Q_{n_v n_u}$ must appear in the expression $\sum_{j\in\{(k+1):J\}} \overline{V}_j$. The fourth and final line holds since each $\overline{V}_j = 0$ when $j > R$ definitionally. We complete (a) by again referencing Mandhani and Meila (2009), who proved that given L, $\hat{\theta}^n$ lower bounds the first two terms of f.

(b) Since $\hat{p}^n$ is defined as the arg min over p for the third term of f subject to the bare minimum constraints imposed by n, the third term of g must lower bound the total cost. This is because as we traverse down the graph from n, only additional constraints may be imposed. Each additional constraint cannot lower the objective function, leading to a lower bound.

Therefore, $g_c(n)$ is an admissible total cost heuristic. ■

Note that g_c is suitably called crude because it is not necessarily a tight lower bound. Instead, the function independently lower bounds components of the likelihood corresponding to the Mallows and Binomial models. However, it is easy and quick to compute L using matrix algebra, $\hat{\theta}^n$ via univariate optimization, and $\hat{p}^n$ via strictly convex optimization in a highly-constrained subspace of the J-dimensional unit hypercube.

LP Total Cost Heuristic

In the crude total cost heuristic, it can be seen that the lower bound on the cost corresponding to the ratings cannot be improved independently of the rankings, given n. A comparable statement is not true for the cost corresponding to rankings. The LP total cost heuristic makes the latter component tighter.

As a brief aside, the MLE of π_0 in the Mallows model is also the solution to the Kemeny ranking problem (Meila et al. 2012). Conitzer et al. (2006) proposed an algorithm to solve the Kemeny ranking problem based on an LP relaxation of the linear integer program that returns the minimum weight feedback edge set. Intuitively, the result can be understood as follows: In the crude lower bound, each pair of objects u, v must be ranked such that u is before v or v is before u. It does not take into account more complex relationships. For example, if u is before v and v is before an object w, the lower bound would still illogically allow w to be before u. The algorithm of Conitzer et al. (2006) removes this possibility. Mandhani and Meila (2009) applied their result to an A* search algorithm for the Mallows

model. In this chapter, we extend this result to the Mallows-Binomial case.

Definition 6 (LP Total Cost Heuristic) *Let $n \in G$ such that $n = (n_1, \ldots, n_k)$, $1 \leq k \leq J-1$, where $n_1, \ldots, n_k$ indicate unique objects in the collection $\{1, \ldots, J\}$. Then, the LP Total Cost Heuristic, $g_{lp}(n) : G \to \mathbb{R}$, is*

$$g_{lp}(n) = \left\{\hat{\theta}^n L_{LP}\right\} + \left\{\log \psi_{R,J}(\hat{\theta}^n)\right\} + \left\{\sum_{j=1}^{J} \overline{x}_j \log \frac{1}{\hat{p}_j^n} + (M - \overline{x}_j) \log \frac{1}{1 - \hat{p}_j^n}\right\}$$

L_{LP} as defined in Conitzer et al. (2006)

$$\hat{\theta}^n = \arg\min_{\theta} \left[\theta L_{LP} + \log \psi_{R,J}(\theta)\right]$$

$$\hat{p}^n = \arg\min_{p} \left[\sum_{j=1}^{J} \overline{x}_j \log \frac{1}{p_j} + (M - \overline{x}_j) \log \frac{1}{1 - p_j}\right] \; s.t. \; p_{n_1} \leq \cdots \leq p_{n_k}, p_{n_k} \leq p_{n_l}, l > k$$

Note that g_{lp} is identical to g_c except for the replacement of L with L_{LP}. We prove that g_{lp} is a tighter lower bound than g_c and admissible via Proposition 7.

Proposition 7 *Under the conditions of Definition 6,*

$$g_c(n) \leq g_{lp}(n)$$

for all nodes $n \in G$. Furthermore,

$$g_{lp}(n) \leq \arg\min_{p,\theta} f(p, \theta) \quad \textit{such that} \quad p_{n_1} \leq \cdots \leq p_{n_k}, \; p_{n_k} \leq p_{n_l}, l > k$$

and therefore $g_{lp}(n)$ is admissible.

Proof Conitzer et al. (2006) prove that $L \leq L_{LP}$. Note that g_c and g_{lp} are identical besides the replacement of L with L_{LP}. Thus $g_c(x) \leq g_{lp}(x)$.

It was shown in Mandhani and Meila (2009) that $L_{LP} \leq \sum_j \overline{V}_j$. In tandem with the proof of Proposition 5, g_{lp} is admissible. ■

3.3.3 Approximate Estimation Algorithms

Exact MLE search algorithms in a Mallows model may be intractably slow when J is large or consensus among judges is weak (Mandhani and Meila 2009). To deal with such cases, approximate search algorithms have been proposed (Ali and Meilă 2012). Here, we extend two fast and accurate algorithms proposed by Fligner and Verducci (1988) and Cohen et al. (1999), respectively. We also state a third approximate algorithm which improves the accuracy of the latter algorithm at a computational cost. Each algorithm is described in turn.

FV Algorithm

Under certain weak conditions, Fligner and Verducci (1988) found that the average ranking is an unbiased estimator of the true consensus ranking in a Mallows model. The same paper proposed an approximate search algorithm for the MLE by averaging each object's rank place across judges and ordering the averages from best to worst into an "average ranking". Then, one calculates the joint density of the data given the average ranking, as well as given each ranking one Kendall distance unit away from the average ranking. The ranking with the highest density in this small collection becomes the approximate MLE.

We propose a simple extension to the Mallows-Binomial model which we call "FV". First, the algorithm calculates average rankings based on ratings alone and rankings alone. If a distinct ordering of objects cannot be determined due to ties or partial rankings, all possible ways to break those ties are included in the set. Second, we calculate the joint density of the data given each of the average rankings and all rankings within one Kendall distance unit from each of the average rankings. The ranking with the highest density becomes the approximate MLE, $\hat{\pi}_0$. Then, $\hat{p}$ and $\hat{\theta}$ are calculated conditional on $\hat{\pi}_0$.

Greedy Algorithm

Cohen et al. (1999) proposed a greedy algorithm to approximate $\hat{\pi}_0$. Specifically, their algorithm iteratively estimates $\hat{\pi}_0$ by choosing the best available object at each ranking level

from first to last. Here, "best" means the object which least lowers the joint likelihood of the data given the current partial ordering. The algorithm is similar to the A* algorithms from Section 3.3.2, except there is no side-to-side traversal in the tree, i.e., once an object is selected for first place, that choice is never reconsidered. $\hat{\theta}$ is calculated conditional on $\hat{\pi}_0$. We apply Cohen et al.'s algorithm to the Mallows-Binomial model using the density function of the Mallows-Binomial instead of the Mallows, in what we call the "Greedy" algorithm.

Greedy Local Algorithm

The "Greedy Local" algorithm extends the Greedy algorithm with a local search. Specifically, it runs the Greedy algorithm as stated and subsequently calculates the joint likelihood of the data given $\hat{\pi}_0$ and all rankings within one Kendall distance unit away from $\hat{\pi}_0$. If no ranking yields a higher likelihood than $\hat{\pi}_0$, the search stops. Else, $\hat{\pi}_0$ is updated to be the ranking with the current highest likelihood and the local search repeats until no better ranking is found. Then, $\hat{p}$ and $\hat{\theta}$ are calculated conditional on $\hat{\pi}_0$.

The Greedy Local algorithm is slower than the Greedy algorithm, but guaranteed to estimate a $\hat{\pi}_0$ which yields a likelihood at least as great as that from the Greedy algorithm. When the Greedy algorithm identifies the exact MLE, the computational expense of performing the Greedy Local algorithm will be minimal as only one round of local search is performed.

3.3.4 Comparison of Algorithms

We now compare the speed and accuracy of estimation algorithms through a simulation study. We ran 20 unique simulations for each combination of model constants $I \in \{5, 20, 80\}$, $M \in \{10, 20, 40\}$, $J \in \{6, 12, 18\}$, and $R \in \{6, 12, 18 | R \leq J\}$ and parameter $\theta \in \{1, 2, 3\}$. In each, we sampled p randomly from a Uniform$[0, 1]^J$. Then, estimation was performed on each data set using each of the 5 algorithms described in Section 3.3: Crude, LP, FV, Greedy, and Greedy Local. The first two are exact algorithms while the latter three are approximate. We now demonstrate results separately based on speed and accuracy of the algorithms.

Speed

There are two useful metrics to consider when evaluating speed in graph search algorithms. The first is overall time, which we measure in seconds. The second is the number of nodes traversed, which signifies an efficient algorithm with respect to memory. While time may be a more practically important metric, if the number of nodes traversed is substantially smaller for a slower algorithm then potential improvements to memory time or code efficiency may ultimately result in a faster algorithm. We compare algorithm speed in Figure 3.3 on the basis of these two metrics.

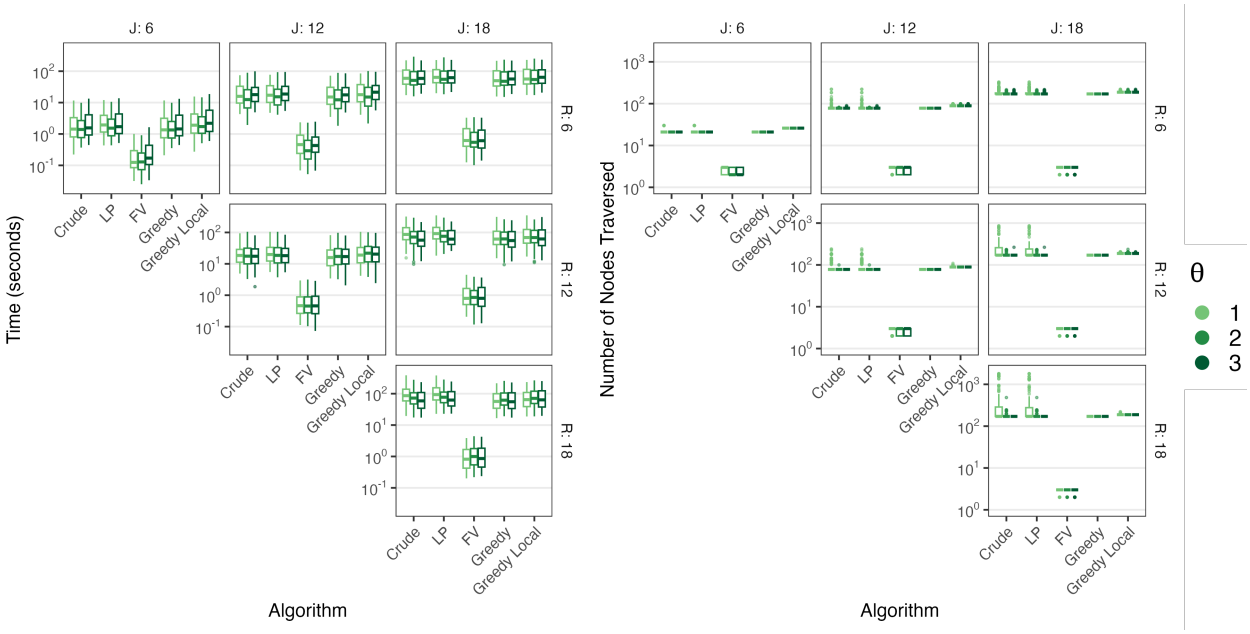

Figure 3.3: Comparison of Mallows-Binomial estimation algorithm speed based on time (left) and number of nodes traversed (right) across different values of J, R, and θ. Results are aggregated over M and I.

Among the exact algorithms, the number of nodes traversed is comparable between both yet computation time is somewhat higher for LP under most regimes. When exact search

is desired, we recommend the Crude algorithm on the basis of these results. Regarding the approximate algorithms, FV is substantially faster than the rest. This difference is by approximately an order of magnitude in all regimes. Greedy and Greedy Local are generally similar in speed to the Crude exact algorithm, a potentially disappointing result given that Greedy and Greedy Local operate under no guarantee of providing an exact solution. However, they exhibit consistent speed results, unlike the exact algorithms which have some extreme slow-speed outliers.

Overall, we observe that estimation time generally increases as J increases and decreases as θ increases. These results should not be surprising: For large J, the algorithms can be slow due to the massive parameter domain. When θ is small, ranking consensus is weak so search algorithms may be pulled into many distinct subspaces of the parameter domain. Speed does not change substantially as R increases.

Accuracy

We measure accuracy of the approximate search algorithms using two metrics: The first is the proportion of simulations in which each algorithm returns the true MLE. However, incorrect estimates may be trivially different from the truth, which leads us to our second metric: The Kendall distance to the true MLE. This measures how far away the estimated ordering of the object quality parameters are from the exact $\hat{\pi}_0$. We compare algorithm accuracy in Figure 3.4 on the basis of these two metrics.

The proportion correct will be 1 and the Kendall distance to the true MLE will be 0 for both the exact algorithms, by definition. For the approximate algorithms, both metrics suggest the order of least to most accurate approximate algorithm is FV, Greedy, and Greedy Local. We point out that even though FV was the fastest algorithm, it exhibits the worst accuracy overall, especially when θ is small. On the other hand, Greedy Local is quite often exactly correct. Accuracy generally improves in all approximate algorithms as R increases, which makes sense given that partial rankings equate to less preference information.

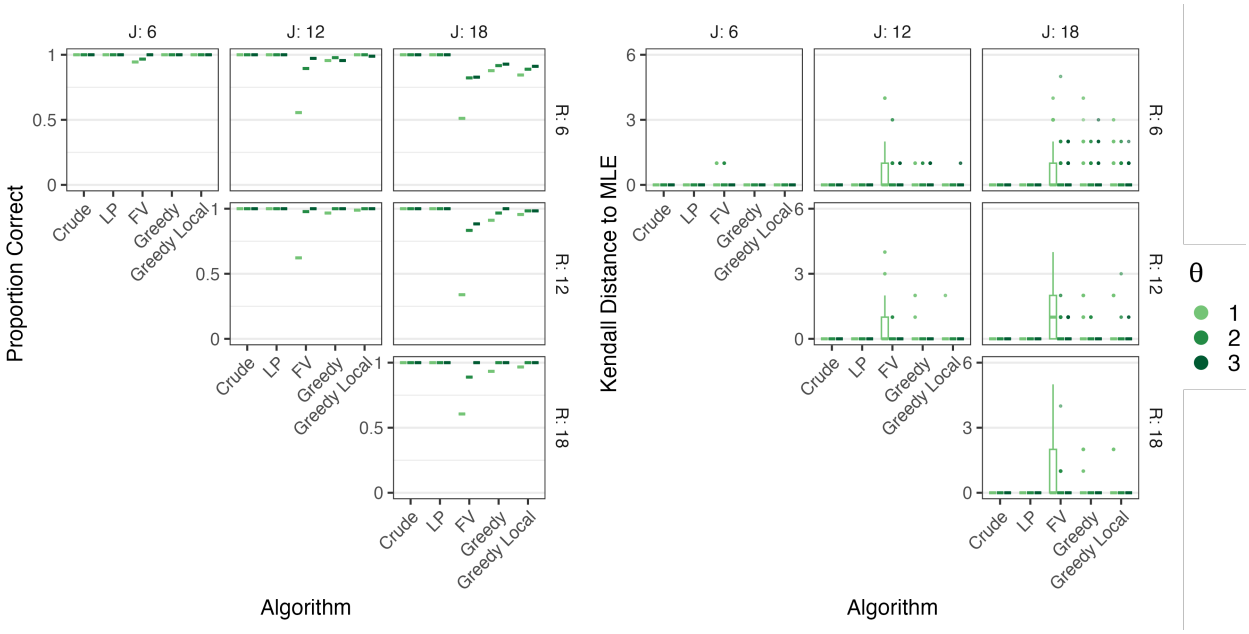

Figure 3.4: Comparison of Mallows-Binomial estimation algorithm accuracy based on the proportion of estimates equal to the true MLE (left) and the Kendall distance to the true MLE (right) across different values of J, R, and θ. Results are aggregated over M and I.

Summary

This section provides insights for practitioners when selecting an estimation algorithm for the Mallows-Binomial model. If exact MLEs are desired, the Crude algorithm is a good choice. When approximations are satisfactory or required due to computational cost, especially when J is large or postulated θ is small, we recommend the Greedy Local algorithm due to its high accuracy, or the FV algorithm for a fast and rough approximation of the consensus ranking.

3.4 Application: Grant Panel Review

We now apply our model to a real data set on grant panel review. After providing an exploratory analysis, we display and interpret estimation results.

3.4.1 Exploratory Analysis

We consider one specific instance of grant panel review conducted by the American Institute of Biological Sciences (AIBS) during Fall 2020, where judges provided both ratings and rankings (Gallo 2020). In the panel, 9 judges discussed 18 proposals. They were allowed to assign ratings between 1.0 and 5.0 in single decimal point increments. Scoring each proposal in turn after an open discussion, judges were asked to provide top-6 partial rankings in private. Ranking ties were not allowed. Since judges discussed every proposal, a proposal not receiving a top-6 ranking was deemed worse than each of the ranked top-6 proposals. With a few exceptions, all judges rated all proposals and ranked their top 6. One judge rated only one proposal and did not provide a ranking; another did not provide a ranking, and a third only provided a top-5 ranking. Based on information from the AIBS, missing data occurred for reasons independent of any characteristics of the proposals, such as child care or family responsibilities as panel review discussions occurred remotely during the Covid-19 pandemic. Thus, we can assume the missing data to be missing completely at random. In this case, estimation using all available (partial or complete) rankings and ratings will not be biased (Little and Rubin 2019). If missingness was due to circumstances related to object quality, for example, one would have to carry out a different treatment of missing data (Little and Rubin 2019). Figure 3.5 summarizes the ratings and rankings received by each proposal.

We observe a variety of scoring and ranking patterns by proposal. For some proposals, all judges gave identical ratings, while for others there was wide disagreement among judges. We notice that proposals with moderate ratings tend to have higher variances than those with generally high or low ratings. These observations suggest that Binomial rating models are reasonable for this data.[2] For rankings, 13 of the 18 proposals were in at least one judge's top-6 ranking. However, Figure 3.5 shows that a smaller subset of proposals were ranked by a majority of the judges (e.g., proposals 1, 6, 7, and 14). Separately, we also measure the consistency between rankings and ratings at the judge level. If the rankings were to

[2]Additional calculations shown in Appendix A.4 indicate that the observed ratings' variance for each proposal are roughly centered around the theoretical variances based on Binomial rating models.

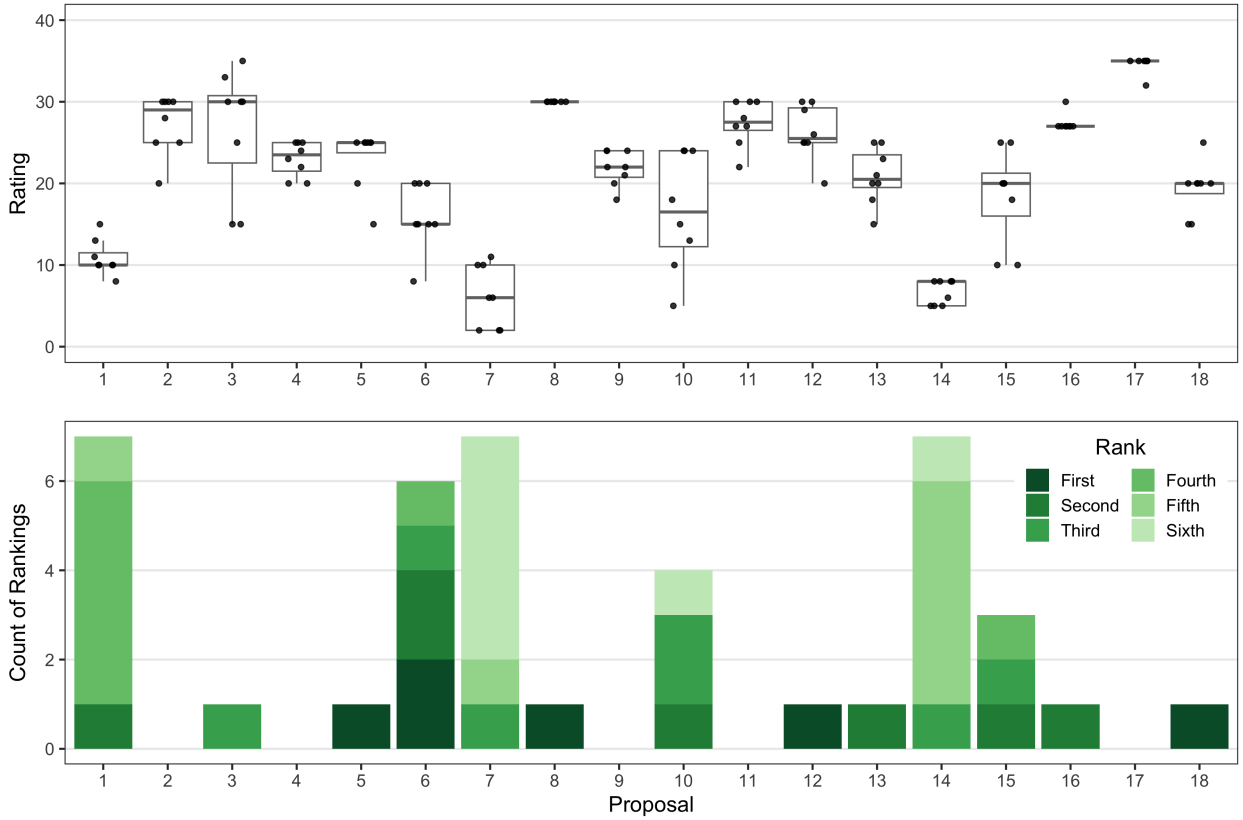

Figure 3.5: Exploratory data analysis of the Fall 2020 AIBS grant panel review data. Ratings (top) displays raw ratings in black and summary boxplots in gray; rankings (bottom) are displayed using stacked bar charts that are colored by rank place.

always align with the order of the ratings, for example, the rankings may be thought of as providing little additional information. To quantify this, we measure the Kendall distance (i.e., the number of pairwise disagreements) between each judge's partial ranking and the implied order of his/her ratings. When a judge assigns equal ratings to any two proposals or does not rank any two proposals, we do not count potential inconsistencies between them. We found the Kendall distances between each judge's partial ranking and rating-implied ranking to be $\{2, 4, 4, 5, 7, 11, 22\}$, ordered from least to greatest. Given that each judge only

provided a top-6 ranking of the proposals, there is substantial discordance between rankings and ratings at the judge level; no judge was internally consistent. We believe this further motivates the use of a combined model for rankings and ratings for this data set.

The AIBS is principally interested in identifying which proposals should receive funding. While thematic and other considerations also contribute to funding decisions, funding agencies rely on peer review to identify which proposals are quality proposals and whether proposals can be ordered or tied in quality. Thus, both estimating proposal quality parameters and identifying a consensus ranking are of interest. Understanding uncertainty in the estimated consensus ranking is key for understanding if objects are of similar quality.

We fit a Mallows-Binomial model to the data, in which $M = 40$, $I = 9$, $J = 18$, and $R = 6$. In doing so, we make note of a few assumptions. First, we assume that each proposal has a true underlying quality. The underlying qualities imply a true ordering of the proposals from best to worst, which we seek to estimate. Second, we assume that the population of judges is homogeneous in its preferences. This may be interpreted as assuming that all judges use the same criteria when ranking or rating and that all variation in ratings and rankings is due to random chance, as opposed to true ideological differences. Third, we assume that all ratings and rankings, even those provided by the same judges, are conditionally independent given the latent true underlying quality of a proposal and the level of consensus strength.

3.4.2 Results

We now present the MLE and the associated bootstrapped 90% confidence intervals of the consensus scale parameter θ and object quality vector p. Confidence intervals are based on 200 bootstrap samples. Table 3.1 contains parameter estimates and Figure 3.6 displays expected ratings and associated confidence intervals overlaid on the judges' observed ratings.

As shown in Figure 3.6, the MLEs of the expected ratings are approximately equal to the means of the observed ratings. However, confidence bands reflect information obtained from both ratings and rankings. For example, proposals 8 and 16 have lower confidence limits that are much lower than the minimum rating they received, which is unusual for a measure of the

Parameter	MLE	90% CI	Parameter	MLE	90% CI
θ	0.529	(0.421,1.124)	p_{10}	0.416	(0.308,0.525)
p_1	0.272	(0.239,0.306)	p_{11}	0.684	(0.642,0.730)
p_2	0.683	(0.626,0.729)	p_{12}	0.656	(0.565,0.711)
p_3	0.666	(0.544,0.766)	p_{13}	0.522	(0.481,0.565)
p_4	0.575	(0.553,0.616)	p_{14}	0.169	(0.150,0.186)
p_5	0.563	(0.511,0.646)	p_{15}	0.463	(0.374,0.541)
p_6	0.400	(0.325,0.453)	p_{16}	0.683	(0.646,0.698)
p_7	0.153	(0.103,0.199)	p_{17}	0.866	(0.850,0.875)
p_8	0.750	(0.711,0.750)	p_{18}	0.484	(0.444,0.541)
p_9	0.563	(0.526,0.588)			

$\hat{\pi}_0 = \{7, 14, 1, 6, 10, 15, 18, 13, 5, 9, 4, 12, 3, 16, 2, 11, 8, 17\}$

Table 3.1: Maximum likelihood estimates of Mallows-Binomial parameters for the Fall 2020 AIBS grant panel review data.

expected (mean) rating. This likely occurs since they were each ranked comparatively better than the ratings they received on average. We also notice that a few proposals share the same MLE of true underlying quality but are strictly ordered (i.e., not tied) in the consensus ranking. For example, proposals 5 and 9 correspond to $\hat{p}_5 = \hat{p}_9 = 0.563$, but proposal 5 is ranked higher than proposal 9 in $\hat{\pi}_0$. In this case, proposal 5 received a marginally worse average rating than 9 but was ranked higher. Thus, the model can capture a difference in ranking while suggesting the true underlying quality is likely nearly identical. See Appendix A.4 for an exploration of model fit.

We display the estimated consensus ranking and associated 90% ranking confidence intervals for each proposal based on the Mallows-Binomial model in Figure 3.7. Additionally, we show results that would be obtained under four separate ranking or rating aggregation models. The first model, *Converted Ratings*, uses ratings and rankings converted into rat-

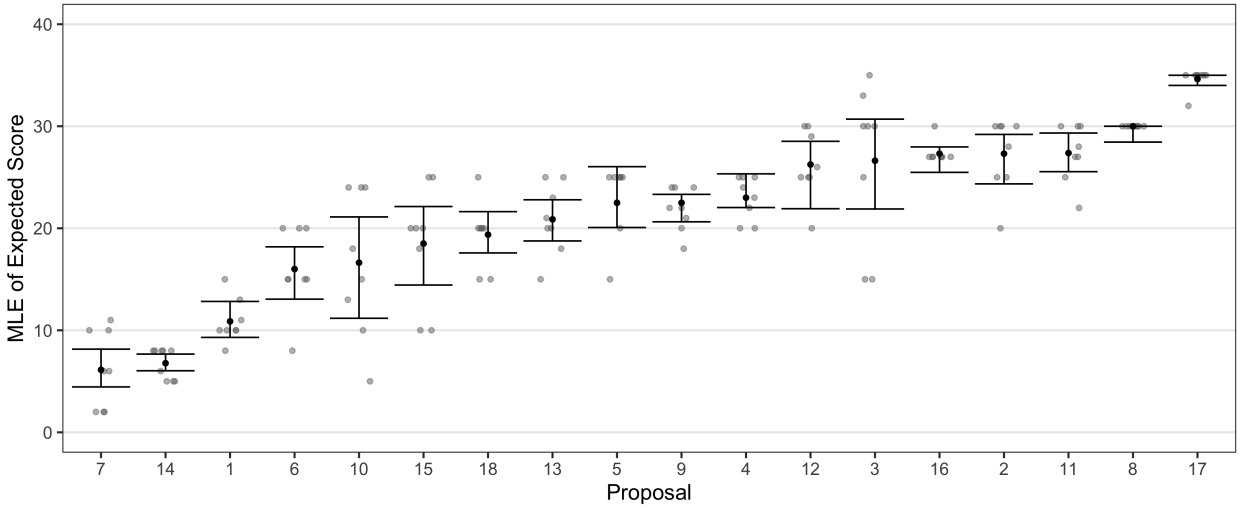

Figure 3.6: Maximum likelihood estimates (MLE) and 90% confidence intervals of expected score (black) overlaid with raw ratings (gray), by proposal. The order of proposals on the x-axis aligns with the MLE of the consensus ranking, $\hat{\pi}_0$.

ings for each judge such that the first-ranked object receives that judge's best rating, the second-ranked object receives that judge's second-best rating, etc., as suggested by Li et al. (2009). Then, the model uses independent Binomial rating distributions for each proposal (no rankings are modeled). The second comparison model, *Only Ratings*, is identical to the first but excludes all rankings. The third comparison model, *Converted Rankings*, uses rankings and ratings converted into rankings for each judge by simple ordering (ties are broken at random). Then, a Mallows distribution is used to model the ranking data. The fourth comparison model, *Only Rankings*, is identical to the third but excludes all ratings. We note that *Converted Ratings* and *Converted Rankings* use all the available data (after conversion) and therefore provide the most direct comparison to the Mallows-Binomial, while *Only Ratings* and *Only Rankings* are limited by the exclusion of certain preference data; none of the comparison methods jointly model the original rankings and ratings. Confidence intervals

for each model are based on 200 bootstrap samples.

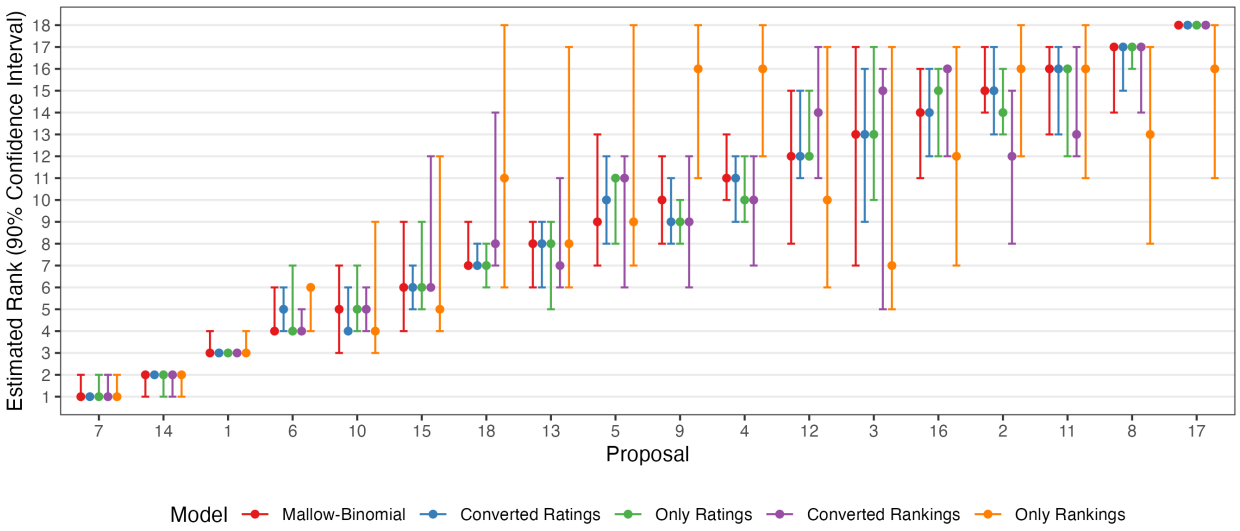

Figure 3.7: Estimated ranks and 90% confidence intervals for the *Mallows-Binomial* model based on ratings and partial rankings and four competing models based on: (1) ratings and rankings converted into ratings in Binomial models (*Converted Ratings*), (2) ratings in Binomial models that exclude rankings (*Only Ratings*), (3) partial rankings and ratings converted into rankings in a Mallows model (*Converted Rankings*), and (4) rankings in a Mallows model that excludes ratings (*Only Rankings*). The order of proposals on the x-axis aligns with the MLE of the consensus ranking in the *Mallows-Binomial* model.

We observe in Figure 3.7 that the Mallows-Binomial model provides a sensible estimated ranking for each proposal: Each proposal has a unique point estimate for rank place and the associated 90% confidence intervals reflect the ratings and ranks it received. For example, proposal 7 was ranked first by 5 of the 7 judges and had the best average rating, but proposal 14 was highly ranked by many judges and received a similarly high average rating. Thus, the 90% confidence intervals of (1,2) for the rank place of proposals 7 and 14 appear appropriate. On the other hand, proposal 3 received the 13th best average rating, which corresponds to

its point estimate for rank place. However, its 90% confidence interval (7,17) for rank place is appropriately wide given its wide range of ratings (minimum 15, maximum 35) and a single fourth-place ranking, which injects uncertainty into the model. In general, confidence intervals are narrow when consensus between ratings and rankings across judges is strong and are wider otherwise.

Results from the Mallows-Binomial model improve upon results from the other models in unique ways. The *Converted Ratings* and *Only Ratings* models provide similar rank place point estimates to the Mallows-Binomial model, but confidence intervals that may be considered inappropriate. *Converted Ratings* provides narrow intervals that reflect an artificially inflated sample size (resulting from combining both original and converted ratings) but does not account for uncertainty arising from converting rankings into ratings. Using only ratings limits the amount of information on judges' perception of proposal quality via rankings, which naturally leads to a loss in precision. However, sometimes the *Only Ratings* model exhibits narrower confidence intervals than the Mallows-Binomial model when ratings are consistent but rankings are not, which still falsely reflects the true combined preferences of the judges. The *Converted Ratings* and *Only Ratings* models do not estimate the consensus scale parameter θ.

Point estimates and confidence intervals from the *Converted Rankings* and *Only Rankings* models differ substantially from those of the Mallows-Binomial. Differences are particularly apparent for proposals ranked in 7th place or worse, as those proposals generally have less data due to the partial rankings collected. The *Converted Rankings* model loses precision compared to the Mallows-Binomial model in the top ranking places, despite having the same number of observations, since ratings converted into rankings via ordering lack information on the strength of the difference in quality between proposals. The *Only Rankings* model has even less precision, since the complete exclusion of ratings and limited information provided by partial rankings constrains inference on the many proposals that were never or rarely ranked and leads to uninformative and insensible rankings. For example, proposals 2, 4, 9, 11, and 17 have near-identical and wide confidence bands as they were never ranked, while

proposals 3, 5, 8, 12, 13, 16, and 18 have even wider confidence bands since they were ranked only by a few judges. Furthermore, the *Converted Rankings* and *Only Rankings* models do not estimate the object quality parameter vector p.

Results from the Mallows-Binomial model allow us to compare proposals with confidence. For example, the model suggests that proposals 7 and 14 are of similarly high quality, and that proposals 1 and 6 are clearly worse than 7 and 14 but uncertain in order between themselves. These types of comparisons may be useful when drawing a funding line at the AIBS. If the AIBS can fund, for example, only 6 proposals, then using 90% marginal confidence intervals by proposal they should fund proposals 7, 14, 1, 6 and select two additional proposals between 10, 15, and 13 (perhaps based on point estimates or a random lottery; see Fang and Casadevall (2016), Roumbanis (2019), and Heyard et al. (2022) for further discussion of partial lotteries in peer review).

3.5 Discussion

In this chapter, we proposed the first unified statistical model for rankings and ratings that does not involve data conversion, the Mallows-Binomial model. We formulated a computationally efficient algorithm to find the exact maximum likelihood estimators of model parameters and demonstrated statistical properties of the model such as bias, consistency, and variance of estimators. This research aligns well with the recommendations from a peer review study at the 2016 Neural Information Processing Systems conference that recommended using both rankings and ratings to gain benefits from each data format (Shah et al. 2018). That study also emphasized the need to design algorithms to efficiently combine ratings and rankings for further guidance on conference submission quality (Shah et al. 2018, p.27).

We applied the Mallows-Binomial model to grant review data which collected both ratings and partial rankings from a panel of judges. The model was used to identify a consensus ranking based on the ratings and partial rankings. The estimated consensus ranking was different from what would be obtained with comparable models for (converted) ratings or

rankings alone. Furthermore, we demonstrated a method to obtain confidence bands of proposal qualities and/or rank places via the bootstrap that can be used to select proposals that are preferred by reviewers with statistical confidence. Confidence bands clearly reflect information from both ratings and rankings provided by the judges.

The proposed model is useful whenever both rankings and ratings for a collection of objects are available. Beyond the example presented here, this may occur in a variety of contexts. For example, relevance of webpages to a search query may be measured from different systems using either numerical metrics (ratings) or ordinal comparisons (rankings). In this example, the object quality parameters would measure both relative and absolute relevance to the search query, and the scale parameter would represent consensus among the systems. If ratings and rankings arise from the same system, rankings may help break ties when ratings are close; if different systems are used to provide different ratings and rankings, using all available data increases estimation precision. In contrast to methods that convert rankings and ratings into data of a single type, the proposed model removes the potential introduction of error by using information from both sources directly. Yet, it allows for using both rankings and ratings to express different types of comparisons and levels of granularity in preferences. Furthermore, because both types of data are incorporated in a statistical model, this allows for uncertainty quantification in the estimation of true underlying quality and strength of consensus when both ratings and rankings are present using standard model-based statistical approaches.

Estimation methods presented in this chapter for the Mallows-Binomial model can be improved or extended upon in a number of ways. Computational efficiency of estimation may be improved via a different heuristic function in the A* algorithm and permit exact estimation of the model in the presence of a large number of objects. Approximate algorithms may be improved to increase accuracy and/or speed. In addition, alternative Bayesian estimation methods may be developed by extending the work of Vitelli et al. (2018) on the Mallows model. Model components may also be generalized: The Beta-Binomial or Poisson distributions may replace the Binomial rating distribution component in our proposed

model if one was interested in accounting for differences in the variance of object ratings among judges or working with rating data with no theoretical maximum value, respectively. Additionally, the Generalized Mallows distribution (Fligner and Verducci 1986) or Infinite Generalized Mallows distribution (Meila et al. 2012) may replace the ranking distribution component of our proposed model. The Bradley-Terry-Luce family of ranking distributions may also replace the ranking distribution component to allow for additional types of ranking data, such as pairwise or groupwise comparisons. Lastly, the model may be considered in a latent class framework to identify the presence of and to measure local consensus among heterogeneous preference groups, e.g., by extending earlier work on the mixture of Mallows distributions (Busse et al. 2007) or Plackett-Luce distributions (Gormley and Murphy 2006; Gormley et al. 2009). In fact, extensions to Bayesian estimation, Bradley-Terry-Luce ranking distributions, and latent class models to estimate preference heterogeneity will all be addressed in the following chapter.

Chapter 4

BAYESIAN CLUSTERING OF PREFERENCES WITH RANKINGS AND RATINGS

This chapter is based on Pearce and Erosheva (2023) and was written in collaboration with Dr. Elena A. Erosheva. This work was supported by the National Science Foundation under Grant No. 2019901.

4.1 Introduction

To our knowledge, the Mallows-Binomial model proposed in Chapter 3 is the first and only joint statistical model for rankings and ratings that does not rely on data conversion. The model shares parameters between Mallows and Binomial distributions for rankings and ratings, respectively, which may be used to perform inference on the absolute and relative qualities of objects. The model does not require judges to provide internally consistent rankings and ratings. However, it has three major drawbacks that limit practical applicability. First, it requires rankings to be either top-r or complete lists of objects. Thus, Mallows-Binomial cannot accommodate pairwise comparisons or cases when different judges have access to different sets of objects (i.e., "separate ballots"). Second, Mallows-Binomial does not allow for heterogeneity. Third, rankings that follow a Mallows distribution do not satisfy Luce's Choice Axiom (Luce 1959), which implies the "independence from irrelevant alternatives" criterion (Marden 1996).

The work in this chapter is motivated by the following three applied settings: (1) assessment of papers to large academic conferences where, for practical reasons, each reviewer only evaluates a small subset of proposals using pairwise or groupwise comparisons; (2) panel review of a small number of grant proposals where reviewers may adhere to distinct ideolo-

gies regarding what type of proposals are preferred; and (3) survey data where heterogeneity and missing data are common. In each, rankings and ratings are of practical use, yet no statistical models exist to estimate group preferences and the associated uncertainty by using rankings and ratings jointly. We describe each setting below in greater detail.

Setting 1: Paper Selection in Large Academic Conferences We first consider quality assessment of papers submitted to large academic conferences, where no single reviewer evaluates all papers. This situation closely mirrors that studied by Liu et al. (2022) in computer science, although so-called "distributed peer review" systems have been observed in astronomy as well (Merrifield and Saari 2009; Patat et al. 2019; Meyer et al. 2022). We suppose a large number of papers are submitted to a conference and only a small subset may be accepted. To make decisions, each paper is assigned a few reviewers and each reviewer is assigned a few papers to review, such that reviewers generally assess overlapping subsets of papers. A simple method for collecting preferences is to ask each reviewer to rate each of their assigned papers on a clearly-defined, discrete "common scale". Then, the papers with the best average ratings are selected for acceptance. However, this approach is suboptimal because (1) ratings may be inconsistent since reviewers may interpret the scale in unique ways (Baumgartner and Steenkamp 2001), and (2) delineating the papers based on average ratings may be impossible or imprecise since average ratings can produce ties or near ties, especially when the scale is coarse or the paper is assessed few times. We note that finding additional reviewers may be impractical. In addition, increasing the granularity of the rating scale may not increase precision, but instead increase noise (Miller 1956; Jones and Loe 2013). We will demonstrate that these problems can be addressed by the introduction of rankings. Both NeurIPS 2016 and ICML 2021 collected rankings from reviewers in addition to ratings (Shah et al. 2018). Still, a principled statistical method by which to incorporate rankings and ratings jointly that could be applicable to such settings does not exist. Because such conferences typically handle high volumes of paper submissions (Shah et al. 2018), we focus only on point estimation of paper quality and do not estimate any potential heterogeneity.

Setting 2: Proposal Selection in Grant Panel Review under Heterogeneity In grant panel review, reviewers evaluate a small number of grant proposals with respect to some criteria such as scientific merit. Often, mean ratings are used to communicate proposal quality for funding decisions, even though they exhibit similar problems to those described in the previous setting. At a time of funding scarcity–for example, R01 research award rates at the National Institutes of Health vary between 10 and 20% (Erosheva et al. 2020)–it is most important to obtain clear and accurate demarcation of proposals at the top. As such, the addition of top-r rankings, in which r is slightly larger than the number of proposals to be funded, may be useful. Top-r rankings provide additional information on the "best" proposals without creating a substantial cognitive burden on reviewers to rank each and every proposal. We studied this setting in Chapter 3 using the frequentist Mallows-Binomial model under the assumption of a single ground-truth ranking of proposals (i.e., no heterogeneity). Lee (2012) studied heterogeneity in peer review, arguing that research commonly overlooks normatively appropriate disagreements among reviewers. Additionally, Mallows-Binomial is unable to account for situations in which not all reviewers assess every proposal due to reviewer burden or conflicts of interest. The model developed in this chapter can handle incomplete or partial rankings and estimates heterogeneous preferences among reviewers and the associated uncertainty, and therefore allows for accurate decision-making at the top of the list.

Setting 3: Modeling of Survey Preference Data under Heterogeneity The third setting relates to the analysis of survey data, where survey respondents express preferences on a collection of items. We study a survey dataset on the sushi preferences of Japanese adults (Kamishima 2003). Respondents provided a complete ranking of ten sushi types and rated them on a 5-point scale. The coarse rating scale leads to frequent ties between items. Furthermore, many respondents rated only a few sushi items, creating a substantial amount of missing data. Given the limited available data, we demonstrate how our proposed model accurately combines information from rankings and incomplete ratings to model preferences

of respondents and identify heterogeneity.

In this chapter, we propose a flexible, joint statistical model for rankings and ratings under heterogeneity: The Bradley-Terry-Luce-Binomial (BTL-Binomial). Using a computationally-efficient Bayesian Mixture of Finite Mixtures (MFM) Miller and Harrison (2018); Frühwirth-Schnatter et al. (2021), we simultaneously estimate both the amount and type of heterogeneity among judges. We develop tools for model interpretation and goodness-of-fit assessment, and illustrate those on real and simulated datasets from the three motivating settings to demonstrate the value and practicality of analyzing preferences jointly with rankings and ratings.

The rest of the chapter is organized as follows. In Section 4.2, we describe the BTL-Binomial MFM approach for jointly modeling rankings and ratings under heterogeneous preference ideologies. Section 4.3 develops tools for Bayesian estimation under fixed and unknown numbers of clusters and proposes tools for model assessment. We use simulated and real data from our three motivating settings to illustrate the proposed model in Section 4.4. We conclude with a discussion in Section 4.5.

4.2 Bradley-Terry-Luce-Binomial Model

Suppose I judges assess J objects. Let $\mathcal{S} = \{1, \ldots, J\}$ be the complete set of objects and $\mathcal{S}_i \subseteq \mathcal{S}$ be the subset assessed by judge i. Let $R_i = |\mathcal{S}_i|$ be the size of $\mathcal{S}_i$. Let $\Pi_i = \{\Pi_i(1) \prec \Pi_i(2) \prec \cdots \prec \Pi_i(r_i)\}$ be judge i's ranking of length $r_i \leq R_i$, such that $\Pi_i(r)$ is the r^{th}-most preferred object by judge i among $\mathcal{S}_i$. Let $X_{ij} \in \{0, 1, \ldots, M\}$ be the rating of judge i to object j, such that 0 is the best and M the worst. This reversed rating scale maintains a symmetry with rankings, in that numerically low ratings correspond to numerically low rankings.

Suppose a judge assesses J objects using a ranking, Π, and ratings, X. Assume Π is of length $R \leq J$, and each rating X_j, $j \in \mathcal{S}$, is an integer between 0 (best) and M (worst). $\mathcal{S}$, R, and M are fixed and known. Under a Bradley-Terry-Luce-Binomial (BTL-Binomial) distribution, their joint probability is given by:

$$P[\Pi = \pi, X = x|p, \theta] = \prod_{r=1}^{R} \frac{\exp(-\theta p_{\pi(r)})}{\sum_{j \in \mathcal{S}} \exp(-\theta p_j) - \sum_{s=1}^{r-1} \exp(-\theta p_{\pi(s)})} \times \prod_{j=1}^{J} \binom{M}{x_j} p_j^{x_j} (1 - p_j)^{M - x_j}$$

$$p = [p_1 \dots p_J]^T \in [0, 1]^J, \theta > 0, \tag{4.1}$$

$$\Pi, X_1, \dots, X_J \text{ are mutually independent.}$$

The BTL-Binomial model combines a BTL ranking distribution parameterized by worth parameters $\omega_j = \exp(-\theta p_j)$ and a Binomial rating distribution for each object, with Binomial probability p_j. As in Chapter 3, we call p the *object quality vector* and θ the *consensus scale parameter*. The parameter p contains the underlying object qualities on the unit interval and appears in both ranking and rating components of the model, thus tying together their estimation to learn preferences. θ measures the strength of ranking consensus.

The Binomial rating parameterization is straightforward and follows the rating parameterization of Chapter 3. The BTL ranking parameterization that sets each ω_j to $\exp(-\theta p_j)$ is new and requires further explanation. Note that small values of p_j correspond to large ω_j, since a small-valued object quality parameter corresponds to a high-quality object, which should thus be ranked highly with greater probability (and vice versa). The parameterization maintains the exponential distance interpretation of the Mallows and Mallows-Binomial models, in that ranking probabilities are determined based on an exponential relationship with rate controlled by θ (Fligner and Verducci (1986); Chapter 3.2). Here, the difference between the underlying qualities of two objects, $p_B - p_A$, is the distance that controls pairwise ranking probabilities. That is because,

$$P[A \prec B] = \frac{\exp(-\theta p_A)}{\exp(-\theta p_A) + \exp(-\theta p_B)} = \frac{1}{1 + \exp(-\theta(p_B - p_A))}.$$

The parameterization also removes the standard identifiability concern of BTL models because the worth parameters are now constrained to the interval $[\exp(-\theta), 1]$ and anchored via the ratings. This claim is made formally in Theorem 8.

Proposition 8 *Let M, J, and R be fixed and positive integers such that $R \leq J$. Then the BTL-Binomial(p, θ) model is identifiable.*

Proof Let $P_{p,\theta}$ denote the probability distribution of ratings X and rankings Π under a BTL-Binomial(p, θ) model. Let $\theta_1, \theta_2 > 0$ and $p_1, p_2 \in [0,1]^J$ such that $P_{\theta_1,p_1} = P_{\theta_2,p_2}$. Given M, the standard Binomial distribution is identifiable. Thus, for $j = 1, \ldots, J$ and arbitrary X_j, we know that $P_{\theta_1,p_1} = P_{\theta_2,p_2}$ if and only if $p_1 = p_2$. Continuing under this assumption, it remains to show that $P_{p_1,\theta_1} = P_{p_1\theta_2} \iff \theta_1 = \theta_2$. Note that,

$$
\begin{aligned}
& P_{p_1,\theta_1} = P_{p_1,\theta_2} \\
& \iff \prod_{r=1}^{R} \frac{e^{-\theta_1 p_{1\Pi(r)}}}{\sum_{j\in\mathcal{S}} e^{-\theta_1 p_{1j}} - \sum_{s=1}^{r-1} e^{-\theta_1 p_{1\Pi(s)}}} = \prod_{r=1}^{R} \frac{e^{-\theta_2 p_{1\Pi(r)}}}{\sum_{j\in\mathcal{S}} e^{-\theta_2 p_{1j}} - \sum_{s=1}^{r-1} e^{-\theta_2 p_{1\Pi(s)}}} \\
& \iff 0 = \sum_{r=1}^{R} \Big[p_{1\Pi(r)}(\theta_2 - \theta_1) + \log\Big(\frac{\sum_{j\in\mathcal{S}} e^{-\theta_2 p_{1j}} - \sum_{s=1}^{r-1} e^{-\theta_2 p_{1\Pi(s)}}}{\sum_{j\in\mathcal{S}} e^{-\theta_1 p_{1j}} - \sum_{s=1}^{r-1} e^{-\theta_1 p_{1\Pi(s)}}} \Big) \Big]
\end{aligned}
$$

which for arbitrary Π will be true only when $\theta_1 = \theta_2$, as desired. ■

We now interpret BTL-Binomial parameters. The vector p reflects object quality, in which values close to 0 indicate relatively high quality and values close to 1 indicate relatively low quality. Comparisons between object quality parameters are also possible. For example, if objects 1 and 2 have quality parameters $p_1 = 0.1$ and $p_2 = 0.11$, respectively, we may assume that the objects are of very similar quality, but that object 1 is slightly better than object 2. If the object qualities are instead $p_1 = 0.1$ and $p_2 = 0.9$, object 1 is clearly better than object 2. Beyond comparisons of specific values, p may be ordered to form a consensus ranking, denoted π_0. For example, if $p = [0.5\ 0.55\ 0.1\ 0.9]$ in a four object system, the consensus ranking is $\pi_0 = 3 \prec 1 \prec 2 \prec 4$. Here, we say objects 1 and 2 are similar in quality, but object 3 is clearly highly quality than object 4. The consensus scale parameter θ is harder to interpret and may be considered a nuisance parameter. Most directly, θ is an input for calculating the probability that some object A is selected over object B in a pairwise tournament. For example, if $p_B - p_A = 0.1$, then the probability that object A is

selected over B is $1/(1+\exp(-0.1\times\theta))$ (see Table 4.1). Higher (lower) values of θ imply rankings among judges will be more (less) similar to each other.

θ	1	5	10	20	40
$P[A \prec B \mid \theta, p_B - p_A = 0.1]$	0.525	0.622	0.731	0.881	0.982

Table 4.1: Pairwise ranking probabilities given $p_B - p_A = 0.1$ under various θ.

4.2.1 MFM Approach to Estimating Heterogeneity

Heterogeneous Preferences

Standard preference models rely on the assumption that each object has a single, true underlying quality. This assumption is inappropriate when judges exhibit heterogeneity. For example, voters of different political parties may have diverging opinions of candidates. Another example arises in peer review, where reviewers may adhere to distinct ideologies for what constitutes promising research based on their background and training (Lee 2012). In such situations, we say the judges exhibit heterogeneous preference ideologies.

A latent class mixture model can be used to capture heterogeneous preference ideologies. Mixture models have been used in the context of both Mallows (Busse et al. 2007; Ali et al. 2010; Meila and Chen 2012; Liu and Moitra 2018) and BTL distributions (Gormley and Murphy 2006; Gormley et al. 2009; Mollica and Tardella 2017). To the best of our knowledge, no joint statistical model for rankings and ratings under heterogeneity exists. Latent class preference models generally assume there exist K preference ideologies and that each judge adheres to precisely one. Latent classes represent the preference ideologies, such that each class $k \in \{1, \ldots, K\}$ has its own set of parameters. We let $Z_i = k$ denote judge i's class.

The true number of preference classes, K, is often unknown and must be identified or estimated. Most of the literature on heterogeneous preferences fits separate models un-

der various choices of K and selects the best-fitting or most parsimonious model via some goodness-of-fit criteria (Gormley and Murphy 2006; Mollica and Tardella 2017). However, it is also possible to estimate K probabilistically. A vast Bayesian literature exists regarding these models (Antoniak 1974; Richardson and Green 1997; Pitman and Yor 1997; Nobile 2004; McCullagh and Yang 2008; Miller and Harrison 2018; Frühwirth-Schnatter et al. 2021). We elect to use a Mixture of Finite Mixtures (MFM) approach. MFMs can be described as Bayesian latent class mixture models in which the number of classes itself is a random variable and assigned a prior. In their most general form, MFMs are easily interpretable and consistent for the true number of classes as the sample size grows (Miller and Harrison 2018).

BTL-Binomial MFM Model

We now propose a joint statistical model for rankings and ratings under heterogeneity. Under the BTL-Binomial MFM model, the observed preference data Π and X are assumed to arise from the following generative model:

$$
\begin{aligned}
K &\sim f_K(\cdot) && f_K \text{ is a pmf on } \{1, 2, \dots\} \\
\gamma &\sim f_\gamma(\cdot) && f_\gamma \text{ is a pdf on } \mathbb{R}^+ \\
\pi | K, \gamma &\sim \text{Dirichlet}_K(\gamma, \dots, \gamma) && \\
(p_k, \theta_k) &\overset{iid}{\sim} f_{p,\theta}(\cdot) && f_{p,\theta} \text{ is a pdf on } [0,1]^J \times \mathbb{R}^+ \\
Z_i | \pi &\overset{iid}{\sim} \text{Categorical}(\pi_1, \dots, \pi_K) && \\
\Pi_i, X_i | Z_i = k, p, \theta &\overset{ind.}{\sim} \text{BTL-Binomial}(p_k, \theta_k) &&
\end{aligned}
\tag{4.2}
$$

We briefly interpret the generative model in Equation 4.2. The *number of heterogeneous preference ideologies*, K, is drawn from a prior. Independently, a *concentration parameter* $\gamma > 0$ is drawn from a hyperprior, where γ controls the concentration of class weights between sparsity (few classes have substantial weight) and equality (all classes have equal weight). The *class weights*, $\pi = (\pi_1, \dots, \pi_K)$ are then drawn from a symmetric Dirichlet prior. Given

K, *class-specific preference parameters* (p_k, θ_k) are drawn from a prior for each class k. After drawing the class label Z_i for each judge i, their ranking and ratings (Π_i, X_i) are drawn from a BTL-Binomial distribution with class-specific parameters.

4.2.2 Prior Selection

Table 4.2 summarizes model priors and hyperpriors. Following an example in Frühwirth-Schnatter et al. (2021), we assign K a shifted Poisson prior such that $K - 1 \sim \text{Poisson}(\lambda)$. The shifted Poisson gives mass only to positive integers. Next, we note that γ controls the homogeneity in class size. We assign γ a $\text{Gamma}(\xi_1, \xi_2)$ hyperprior, as suggested in the Dirichlet Process Mixture (DPM) (Escobar and West 1995; Jara et al. 2007) and MFM (Miller and Harrison 2018) literatures. When γ is small, we expect some large, small, or even empty classes; when γ is large the classes are expected to be roughly uniform in size. We assign π (*class weights*), a symmetric Dirichlet prior with concentration parameter γ. This so-called "static" MFM is simple and common (Frühwirth-Schnatter et al. 2021). We assign the BTL-Binomial parameters p_{jk} i.i.d. $\text{Beta}(a, b)$ priors, which are not conjugate but simplify the posterior given Binomial ratings. We assign θ_k i.i.d. $\text{Gamma}(\gamma_1, \gamma_2)$ priors.

Parameter	Interpretation	Prior
K	Number of Ideology Classes	$\text{Poisson}(K - 1 \mid \lambda)$
γ	Dirichlet Concentration Parameter	$\text{Gamma}(\gamma \mid \xi_1, \xi_2)$
π	Class Weights	$\text{Dirichlet}_K(\pi \mid \gamma, \ldots, \gamma)$
p_k, θ_k	BTL-Binomial Preference Parameters	$\prod_{j=1}^{J} \text{Beta}(p_{jk} \mid a, b) \times \text{Gamma}(\theta_k \mid \gamma_1, \gamma_2)$

Table 4.2: Priors for the BTL-Binomial MFM Model.

Hyperparameter settings may be highly influential. λ influences the prior expectation on K, such that $E_\lambda[K] = \lambda + 1$. However, the Dirichlet concentration parameter γ allows for unequal weights between classes, thus influencing the number of non-empty classes, $K^+ \leq K$.

Values of γ close to 0 allow for parsimony in the case of no heterogeneity; values greater than 1 give high probability to $K^+ = K$. The conditional prior density on K^+ given γ can be calculated exactly based on the work of Greve et al. (2022) and Frühwirth-Schnatter et al. (2021), as implemented in the `R` package `fipp` (Greve 2021). Selection of λ is highly dependent on context; we suggest choosing the Gamma hyperparameters ξ_1, ξ_2 to provide density to $\gamma \in [0, 3]$, which corresponds to substantial probability that K^+ may be any integer between 1 and K. For the Beta hyperparameters, $a = b = 1$ leads to a proper and minimally informative Uniform prior. Instead, selecting a and b via an empirical Bayes approach based on the observed ratings may improve estimation efficiency. For the Gamma hyperparameters on θ, setting $\gamma_1 = 1, \gamma_2 = 0$ leads to a flat but improper prior. We suggest choosing values to provide substantial density in the region $\theta \in [5, 35]$, which corresponds to varying but reasonable levels of consensus.

4.3 Bayesian Estimation

4.3.1 Estimation of BTL-Binomial MFM Model

Until recently, MFM models have been computationally challenging to estimate due to difficulties associated with reversible jump MCMC (RJMCMC), the primary available estimation tool (Nobile 2004; Phillips and Smith 1996; Richardson and Green 1997; McCullagh and Yang 2008). Miller and Harrison (2018) proved theoretical connections between MFM and DPM models, thus expanding the toolkit and improving speed. Subsequently, Frühwirth-Schnatter et al. (2021) proposed the "telescoping sampler" which drastically lowered the computational burden of fitting MFM models. Their work cleverly decomposes the total number of latent classes, K, from the number of non-empty classes, K^+. Separating these quantities permits a simple Gibbs-type sampler (e.g., no RJMCMC) that is similar in form to those for Bayesian mixture models with fixed K. We adapt the telescoping sampler for the BTL-Binomial MFM model, which is presented in Algorithm 1. Further details of the algorithm can be found in Appendix B.1.

Algorithm 1 Telescoping Sampler for BTL-Binomial MFM Model

Algorithm Parameters: $B^{\text{Gibbs}}, B^{\text{MH}}, \sigma_p^2, \sigma_\theta^2, \sigma_\gamma^2$

1. Initialize: Select starting values for K, γ, π_k, p_k, and θ_k for $k \in \{1, \ldots, K\}$, at random.

2. Gibbs Iterations: Repeat B_{Gibbs} times:

   (a) Update Z and K^+:

      i. Sample Z_i, $i = 1, \ldots, I$, using

$$P[Z_i = k | \pi, p, \theta] = \frac{\pi_k P[X_i, \Pi_i | Z_i = k, p_k, \theta_k]}{\sum_{k'=1}^{K} \pi_{k'} P[X_i, \Pi_i | Z_i = k', p_{k'}, \theta_{k'}]}$$

      ii. Calculate $N_k = \sum_i I\{Z_i = k\}$ and $K^+ = \sum_{k=1}^{K} I\{N_k > 0\}$, where $I\{\cdot\}$ is the indicator function. Relabel the classes such that the first K^+ are non-empty.

   (b) Update (Non-Empty) Class Parameters: For $k = 1, \ldots, K^+$, repeat B^{MH} times:

      i. Sample each p_{jk}, $j = 1, \ldots, J$, via random-walk Metropolis-Hastings (RWMH) with proposal distribution Normal(p_{jk}, σ_p^2).

      ii. Sample θ_k via RWMH with proposal distribution Normal$(\theta_k, \sigma_\theta^2)$.

   (c) Update K and γ:

      i. Sample $K | K^+, \gamma$ such that $K \geq K^+$ using

$$P[K | K^+, \gamma] \propto f_K(K) \frac{K!}{(K - K^+)!} \frac{\Gamma(\gamma K)}{\Gamma(I + \gamma K)}$$

      ii. Sample $\gamma | Z, K$ via RWMH with proposal distribution Normal$(\gamma, \sigma_\gamma^2)$ using

$$P[\gamma | Z, K] \propto f_\gamma(\gamma) \frac{\Gamma(\gamma K)}{\Gamma(I + \gamma K)} \prod_{k=1}^{K^+} \frac{\Gamma(N_k + \gamma)}{\Gamma(\gamma)}$$

   (d) Update Empty Classes and π:

      i. If $K > K^+$, sample (p_k, θ_k) directly from its prior for $k = K^+ + 1, \ldots, K$.

      ii. Sample $\pi | K, \gamma, Z \propto$ Dirichlet$(\gamma + N_1, \ldots, \gamma + N_K)$.

4.3.2 Estimation Under a Fixed Number of Clusters

The number of clusters, K, may alternatively be considered known and fixed. In such contexts, we may assume the following generative model, which we call the Bayesian BTL-Binomial latent class mixture model:

$$\begin{aligned} \gamma &\sim f_\gamma(\cdot) && f_\gamma \text{ is a p.d.f. on } \mathbb{R}^+ \\ \pi|\gamma &\sim \text{Dirichlet}_K(\gamma, \ldots, \gamma) \\ (p_k, \theta_k) &\overset{iid}{\sim} f_{p,\theta}(\cdot) && f_{p,\theta} \text{ is a p.d.f. on } [0,1]^J \times \mathbb{R}^+ \\ Z_i|\pi &\overset{iid}{\sim} \text{Categorical}(\pi_1, \ldots, \pi_K) \\ \Pi_i, X_i|Z_i = k, p, \theta &\overset{ind.}{\sim} \text{BTL-Binomial}(p_k, \theta_k) \end{aligned} \tag{4.3}$$

Note that Equation 4.3 is identical to Equation 4.2 for the BTL-Binomial MFM model, less the initial sampling of K. Furthermore, we assume the same priors f_γ and $f_{p,\theta}$.

We estimate the Bayesian BTL-Binomial latent class mixture model using the Gibbs sampler proposed in Algorithm 2. Further details of the algorithm can be found in Appendix B.1.

4.3.3 Maximum A Posteriori Estimation Under a Fixed Number of Clusters

Third, we note that maximum *a posteriori* (MAP) estimates may be desired in the Bayesian BTL-Binomial latent class mixture model presented in Equation 4.3 for the purpose of point estimation, decision-making, or prediction. Therefore, we propose an Expectation-Maximization (EM) algorithm in Algorithm 3. Further details of the algorithm, including a discussion of how to select priors which align MAP estimators with frequentist maximum likelihood estimators (MLE), can be found in Appendix B.1.

Algorithm 2 Gibbs Sampler for BTL-Binomial Latent Class Mixture Model

Algorithm Parameters: $B^{\text{Gibbs}}, B^{\text{MH}}, \sigma_p^2, \sigma_\theta^2, \sigma_\gamma^2$

1. Initialize: Select starting values for γ, π_k, p_k, and θ_k for $k \in \{1, \ldots, K\}$, at random.

2. Gibbs Iterations: Repeat B_{Gibbs} times:

   (a) Update Z: Sample Z_i, $i = 1, \ldots, I$, using

   $$P[Z_i = k|\pi, p, \theta] = \frac{\pi_k P[X_i, \Pi_i|Z_i = k, p_k, \theta_k]}{\sum_{k'=1}^{K} \pi_{k'} P[X_i, \Pi_i|Z_i = k', p_{k'}, \theta_{k'}]}$$

   (b) Update Class Parameters: For $k = 1, \ldots, K$, repeat B^{MH} times:

      i. Sample each p_{jk}, $j = 1, \ldots, J$, via random-walk Metropolis-Hastings (RWMH) with proposal distribution Normal(p_{jk}, σ_p^2).

      ii. Sample θ_k via RWMH with proposal distribution Normal$(\theta_k, \sigma_\theta^2)$.

   (c) Update γ: Sample $\gamma|\pi$ via RWMH with proposal distribution Normal$(\gamma, \sigma_\gamma^2)$ using

   $$P[\gamma|\pi] \propto f_\gamma(\gamma) \prod_{k=1}^{K} \pi_k^{\gamma - 1}$$

   (d) Update π: Sample $\pi|\gamma, Z \propto$ Dirichlet$(\gamma + N_1, \ldots, \gamma + N_K)$, where $N_k = \sum_i I\{Z_i = k\}$.

Algorithm 3 MAP Estimation via EM in the BTL-Binomial Latent Class Mixture Model
Algorithm Parameters: $tol > 0$

1. Initialize: Select starting values for γ, π_k, p_k, and θ_k for $k \in \{1, \ldots, K\}$, at random.

2. EM Iterations: Repeat until convergence, which we define as when the absolute difference in the model log likelihood between iterations is below the prespecified tol:

   (a) E-Step: For each unique pair (i, k), $i = 1, \ldots, I$ and $k = 1, \ldots, K$, calculate

   $$\hat{z}_{ik} \equiv \mathbb{E}_{\pi,p,\theta}[z_{ik}|\Pi_i, X_i] = \frac{\pi_k P[X_i, \Pi_i|Z_i = k, p_k, \theta_k]}{\sum_{k'=1}^{K} \pi_{k'} P[X_i, \Pi_i|Z_i = k', p_{k'}, \theta_{k'}]}$$

   (b) M-Step: Maximize the unknown parameters (π, γ, p, θ) sequentially:

      i. Update π according to,

      $$\pi = \arg\max_{\pi} \Big(E_{Z|X}\big[\log \mathcal{L}(\Pi, X, Z|\pi, p, \theta)\big] + \log f(\pi|\gamma) \Big) \text{ subject to } \sum_{k=1}^{K} \pi_k = 1$$

      $$\implies \pi_k = \frac{\gamma - 1 + \sum_{i=1}^{I} \hat{z}_{ik}}{K\gamma - K + I}$$

      ii. Update γ via numerical optimization according to,

      $$\gamma = \arg\max_{\gamma} \Big(\log f(\pi|\gamma) + \log f(\gamma) \Big)$$

      iii. Update each (p_k, θ_k) via numerical optimization according to,

      $$\begin{aligned}(p_k, \theta_k) &= \arg\max_{p_k, \theta_k} \Big(E_{Z|X}\big[\log \mathcal{L}(\Pi, X, Z|\pi, p, \theta)\big] + \log f(p_k, \theta_k) \Big) \\ &= \arg\max_{p_k, \theta_k} \Big(\sum_{i=1}^{I} \hat{z}_{ik} \log\big(\text{BTL-Binomial}(\Pi_i, X_i|p_k, \theta_k)\big) + \log f(p_k, \theta_k) \Big)\end{aligned}$$

4.3.4 Model Assessment

In both the BTL-Binomial MFM and related BTL-Binomial Latent Class Mixture Model, we assess mixing and convergence by examining trace plots of quantities which are invariant to label-switching (Stephens 2000). We follow the recommendation of Frühwirth-Schnatter et al. (2021) to examine trace plots of K^+ and π (ordered by mean). We also examine trace plots of K and γ. If class-specific parameters suffer from label-switching, one may apply the algorithm of Stephens (2000) to the posterior samples.

We also examine goodness of fit by comparing the observed and posterior predictive distributions of three types of statistics: (1) rating mean, by object; (2) rating variance, by object, and (3) pairwise probability that object A is ranked above object B, for each pair of objects (A, B). Under a well-fitting model, the observed and posterior predicted statistics should be similar. We assess similarity via visual inspection.

4.4 Applications in Peer Review and Survey Data

We apply the BTL-Binomial model to three motivating examples: paper selection in large academic conferences under sparsity of comparisons, proposal selection in grant panel review under heterogeneity, and modeling of survey data under heterogeneity.

4.4.1 Setting 1: Paper Selection in Large Academic Conferences

Our first application is to the paper selection process in large and highly competitive academic conferences. These conferences typically handle high volumes of paper submissions, and thus reviews are dispersed among many reviewers. Here, we simulate reviews in the form of ratings and rankings and use them to estimate proposal quality via the BTL-Binomial model. We focus on point estimation under the assumption of a single ideology among reviewers (i.e., fixed $K = 1$). Beyond self-selection of papers into research areas and the timing restrictions of organizers, estimating heterogeneity in this context may be particularly noisy given the limited amount of data available from each reviewer.

Simulation Setup

Our simulation study is loosely based on that of Liu et al. (2022), who proposed an algorithm to integrate rankings into ratings for the paper selection process used by the International Conference on Learning Representations 2017 (Kang et al. 2018). In our study, we simulate a conference that has recruited $I = 50$ reviewers to assess $J = 50$ papers. Each reviewer i cannot possibly assess every paper, so instead each is assigned a subset, $\mathcal{S}_i$, at random such that each paper receives an equal number of reviews. Specifically, $|\mathcal{S}_i| = R_i = R$ and each paper receives R reviews since $I = J$. Reviewer i first provides ratings $X_{ij}, j \in \mathcal{S}_i$. Ratings are integers between 0 (exemplary) and M (poor). Reviewers do not rate the other papers. Second, reviewer i provides a top-4 ranking, Π_i, of their favorite papers among those assigned, without ties. We assume that a reviewer deems their "unranked" papers (i.e., $\{j \in \mathcal{S}_i | j \notin \Pi_i\}$) worse than those which were ranked. However, no information can be gleaned from reviewer i for papers not in $\mathcal{S}_i$.

We generate ratings and top-4 rankings from a BTL-Binomial distribution. To capture different amounts of data and noise, we consider all combinations of the following values: (1) $R \in \{4, 8, 12, 24\}$. Small R signifies less work for each reviewer and provides less preference data. (2) $M \in \{4, 9\}$ for a 5- or 10-point rating scale, respectively. Small M increases the coarseness of ratings and thus the probability of ties. (3) $\theta \in \{1, 5, 10, 20, 40\}$. Large θ implies more consensus in rankings; see Table 4.1. In each simulation scenario, we draw $p_j \sim \text{Beta}(1, 1) \overset{d}{=} \text{Uniform}[0, 1]$, and then fit a BTL-Binomial model with a single latent class to the data using hyperparameters $a = 1, b = 1, \gamma_1 = 5$, and $\gamma_2 = 0.25$, which were chosen to be diffuse. Each scenario is replicated 100 times.

Results

We now demonstrate the model's ability to accurately estimate the true overall ranking of papers, π_0, and improve estimation of π_0 in comparison to the standard paper selection method based solely on ratings. That method is to order the papers by their mean rating and

break ties on the basis of another reviewer (Shah et al. 2018). In the absence of additional reviewers, we use random tie-breaking. We let $\hat{\pi}_0^{\text{BTLB}}$ be the BTL-Binomial MAP estimate of π_0, determined by ordering papers based on $\hat{p}$. Similarly, we let $\hat{\pi}_0^{\text{X}}$ be the ratings-only MAP estimate of π_0, determined by ordering papers based on mean ratings. The accuracy of MAP estimates $(\hat{p}, \hat{\theta})$ is shown in Appendix B.2. In each estimated ranking, ties are broken at random.[1] To measure the inaccuracy of each model, we calculate the percentage of object pairs in which the model incorrectly identifies the true order of the objects, π_0. This is equivalent to a normalized Kendall's τ distance between π_0 and a model estimate $\hat{\pi}_0$. We plot the mean inaccuracy across simulations for each combination of R, M, and θ from the BTL-Binomial and ratings-only models in Figure 4.1.

The BTL-Binomial model outperforms the standard ratings-only model on average for every combination of M, R, and θ. The largest improvement in estimation accuracy with the BTL-Binomial over the standard ratings-only model occurs when R and/or M is small, which are precisely the settings of the utmost interest for large academic conferences. These results should be intuitive: When R is small, each paper receives few assessments and thus the additional information from rankings is highly beneficial to accurate preference modeling.[2] When M is small, ties will be common in ratings and lead to haphazard estimation based on random tie-breaking; rankings help to accurately break those ties. We also notice that as θ increases, so does the accuracy of the BTL-Binomial model. This is because higher θ means that rankings will be more adherent to the true ranking π_0 on average, and thus will provide less noisy information for accurate modeling of paper quality.

We have demonstrated that the BTL-Binomial model leads to more accurate decision-making using rankings and ratings in academic conference paper selection under realistic

[1] Ties are uncommon but possible in the BTL-Binomial model. For example, if papers A and B receive identical ratings and are assigned the ranking $A \prec B$ by half of their reviewers and $B \prec A$ by the other half, they cannot be distinguished.

[2] The present simulation study does not allow for disentangling the relative effects of R and I on the estimation accuracy of the BTL-Binomial model. See Appendix B.2 for an additional simulation study in which R and I vary such that the total number of assessments, $I \times R$, remains fixed.

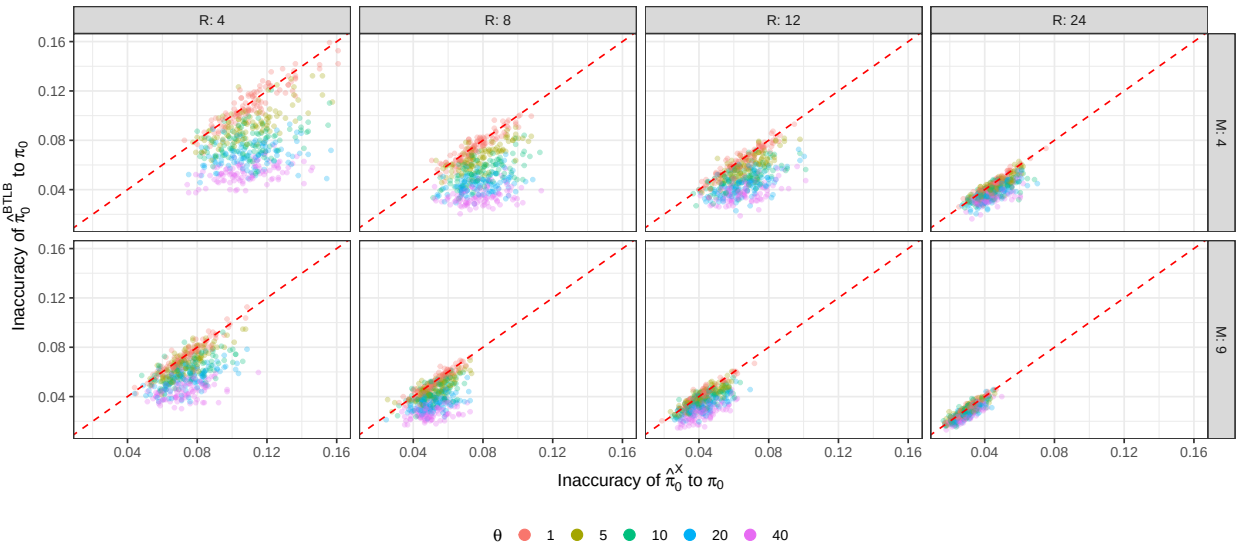

Figure 4.1: Scatterplots of the mean inaccuracy across 100 simulations of estimated $\hat{\pi}_0^{BTLB}$ (BTL-Binomial model) and $\hat{\pi}_0^{\mathrm{X}}$ (standard ratings-only model) to the true ranking of papers π_0 under various combinations of R, M, and θ.

review settings. A key benefit is that reviewers need not assess many papers or greatly increase their workload. In fact, even with a coarse 5-point rating scale and top-4 rankings of a small number of papers, quality assessments may be made based on rankings and ratings with greater accuracy in comparison to the standard mean-ratings model. Furthermore, there becomes little need for random or subjective tie-breaking, making the work for data aggregators and conference chairs both easier and more objective.

4.4.2 Setting 2: Proposal Selection in Grant Panel Review Under Heterogeneity

Our second application is to grant panel review administered by the American Institute of Biological Sciences (AIBS) during the 2021 season (Gallo 2023). The AIBS issues a call for funding, recruits a panel of qualified reviewers, and administers the peer review process. Prior

to panel discussion, reviewers are given access to the grant proposals, although they do not necessarily read each of them in detail. During panel discussion, each proposal is discussed in turn and each reviewer provides a rating which reflects the overall scientific merit of each proposal using the numbers between 1 (*excellent*) and 5 (*poor*) in single decimal point increments (which we transform to the integers between 0 and $M = 40$). After discussion, each reviewer provides a top-6 ranking of their overall preferred proposals. The ranking is not required to align with the reviewer's ratings. The AIBS would like to know if there are distinct preference groups among reviewers, what those preferences are, and know how much uncertainty exists in the estimated proposal quality assessments.

Some rankings and ratings are missing due to conflicts of interest and other reasons unrelated to proposal quality (e.g., intermittent distractions, internet connectivity issues, lack of qualification to accurately review). For missing ratings, we simply remove the corresponding Binomial components from the likelihood. If a reviewer does not provide a ranking, we remove the corresponding BTL components from the likelihood. If reviewer i does not rank proposals due to a conflict of interest, those proposals are removed from his/her set of proposals, $\mathcal{S}_i$. On the basis of Luce's Choice Axiom, estimation of model parameters corresponding to proposals with which a reviewer has a conflict of interest is not affected.

Exploratory Analyses

We study a panel with $I = 17$ reviewers and $J = 25$ proposals. Figure 4.2 displays boxplots of ratings and bar charts of rankings, by proposal. Proposals are numbered at random but ordered on the basis of their mean rating. Boxplots show observed ratings (after transformation to the integer scale) by proposal. The mean and variance of ratings given to each proposal highly vary. For rankings, 14 of the 25 proposals are included in at least one judge's top-6 ranking. Although there is no wide agreement between judges on the basis of rankings, certain proposals nonetheless appear often in the top. For example, proposals 6, 8, 18, and 19 are frequently assigned top-4 ranks (perhaps not coincidentally, these proposals have the best mean ratings).

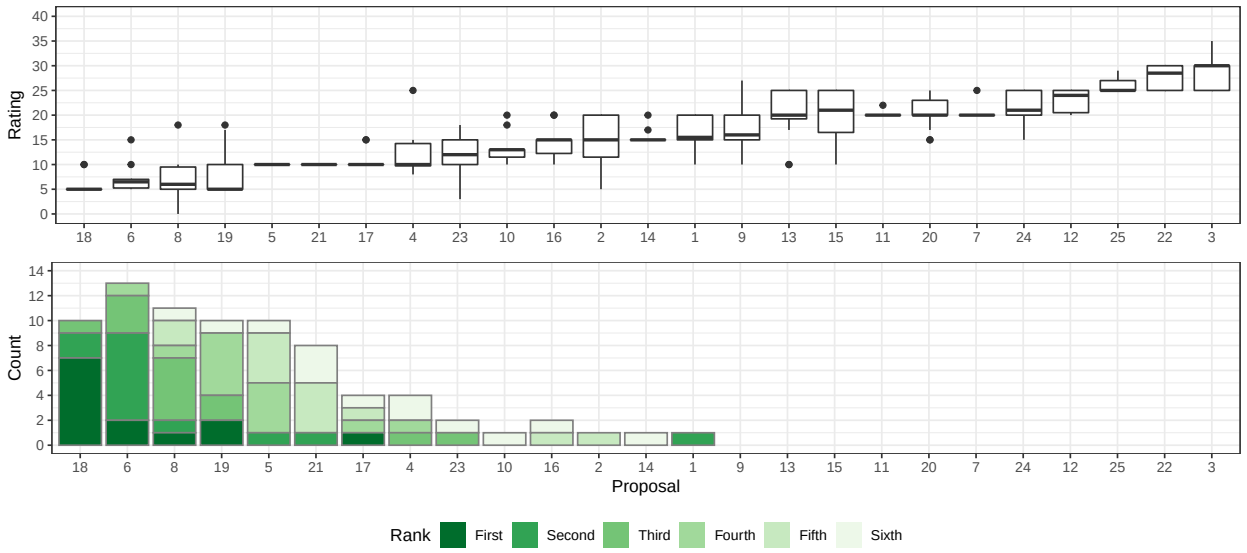

Figure 4.2: Boxplots of ratings (*top*) and stacked bar charts of ranks (*bottom*) by proposal.

To explore the internal consistency of ratings and rankings, we calculate Kendall's τ distance[3] between the ranking of each judge with the ranking induced by the order of his/her ratings. Internal inconsistency between quality assessments via ratings and rankings by the same judge was common: For the 13 judges who provided rankings, the Kendall τ distances are $\{0, 1, 1, 2, 2, 2, 2, 3, 6, 7, 9, 21, 26\}$. This means that only one judge was internally consistent, even under partial rankings, and two judges provided rankings substantially different from those induced by their ratings (judges 8 and 12). This aligns with psychological research that the two assessments rely on different cognitive processes (Goffin and Olson 2011). Furthermore, it suggests that rankings can provide additional and unique information.

[3]As stated in Chapter 3, the Kendall τ distance is equivalent to the number of pairs of proposals where the two rankings differ in their ordering. We break ties in favor of the judge and do not consider pairs in which an ordering between proposals cannot be inferred from the available data (e.g., a missing rating).

Estimation and Results

We now fit a BTL-Binomial MFM model to the AIBS data. We set priors as follows: Given the small sample size, we choose $\lambda = 1$, $\xi_1 = 2$, and $\xi_2 = 3$ to assign prior weight primarily to $K^+ \in \{1, 2, 3\}$. We set $a = 2.50$ and $b = 3.77$ using an empirical Bayes approach. We set $\gamma_1 = 10$ and $\gamma_2 = 0.5$ to provide substantial weight to values of $\theta \in [5, 35]$. Further information on model estimation and assessment is provided in Appendix B.2.

Figure 4.3 displays model results. The top-left panel displays the posterior of K^+, which provides strong evidence for a 2-class model. Thus, we display results conditional on $K^+ = 2$ in what follows. The top-right panel displays estimated class membership probabilities by judge. Class 1 includes 16 judges and class 2 includes just one. The bottom panel displays the posteriors of preference parameters for each proposal and class. Class 1 prefers proposals 18, 6, 8, and 19, in that order, and exhibits relatively strong consensus. Alternatively, class

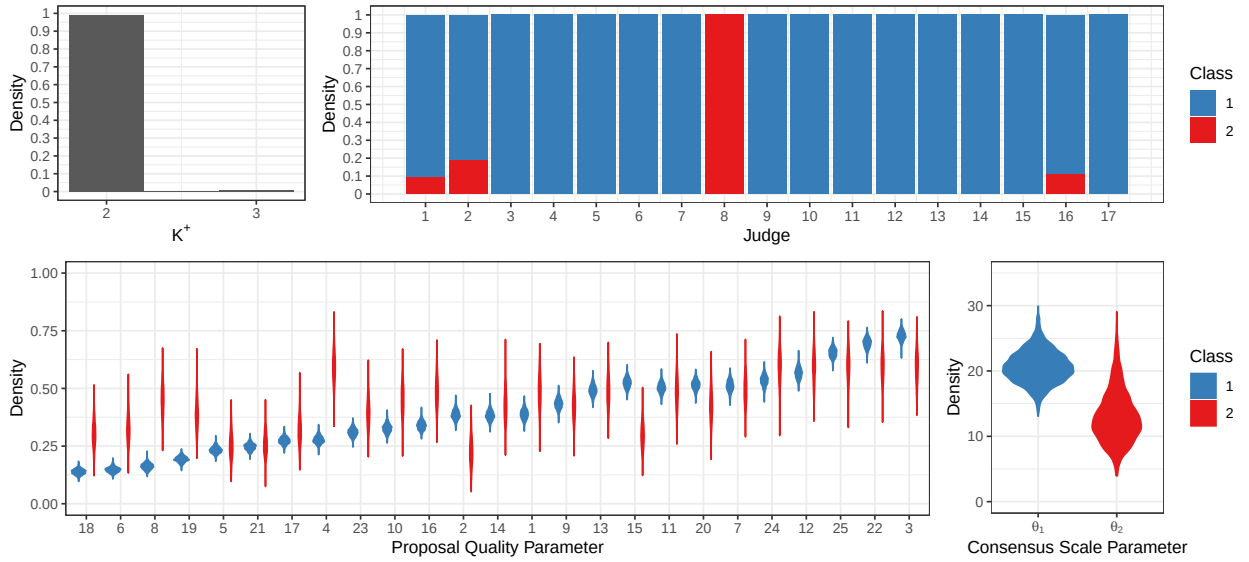

Figure 4.3: Posterior summaries of K^+ (*top-left*); class membership probabilities given $K^+ = 2$ (*top-right*); and class-specific preference parameters given $K^+ = 2$ (*bottom*).

2 largely reflects judge 8. The high levels of uncertainty in class 2's preference parameters reflects that it comprises a single reviewer. Given the "outlier" judge, the funding agency may wish to consider what made judge 8 provide such unique preferences and decide if those warrant separate consideration, or if the results from class 1 should be considered alone in making funding decisions.

4.4.3 Setting 3: Modeling Survey Preference Data Under Heterogeneity

Our final application is to survey data on the sushi preferences of $I = 5,000$ Japanese adults (Kamishima 2003). Both rankings and ratings were collected as part of the survey. Respondents were first asked to rank a collection of $J = 10$ sushi types from best to worst. The sushi types were fatty tuna, tuna, shrimp, tuna roll, sea eel, salmon roe, squid, egg, sea urchin, and cucumber roll. Each respondent was shown the same collection of sushi types and provided a complete ranking. Second, the respondents were asked to rate the sushi types using a 5-point integer scale at will, coded from 0 (best) to $M = 4$ (worst). Each respondent generally provided only a few ratings. Our goal is to probabilistically model the amount and type of heterogeneity in sushi preferences among survey respondents.

The present survey dataset is uniquely positioned for analysis via a joint ranking and rating model: While rankings are complete, they lack granularity; the ratings provide granularity but have a very high rate of missingness. For the purpose of understanding heterogeneity among judges, using both rankings and ratings may be especially helpful for accurate inference on preferences.

Exploratory Analyses

Figure 4.4 displays ratings and rankings by sushi type. For ratings, most sushi types have unimodal distributions with right skew. However, cucumber roll has a unimodal distribution centered at the middle rating while sea urchin has a bimodal distribution with peaks at the best and worst ratings. We can also see relative differences in the number of ratings each sushi type received. For example, the lower density of points for tuna roll and cucumber roll implies

fewer respondents rated these types. 74.22% of ratings are missing. For rankings, fatty tuna was ranked first by approximately one-third of respondents, while cucumber roll was ranked last by approximately one-third. Ties were not allowed in rankings. Consequently, we observe more demarcation in ranking distributions than in ratings. For example, fatty tuna and tuna have similar rating distributions, but far more respondents ranked fatty tuna in first place than tuna.

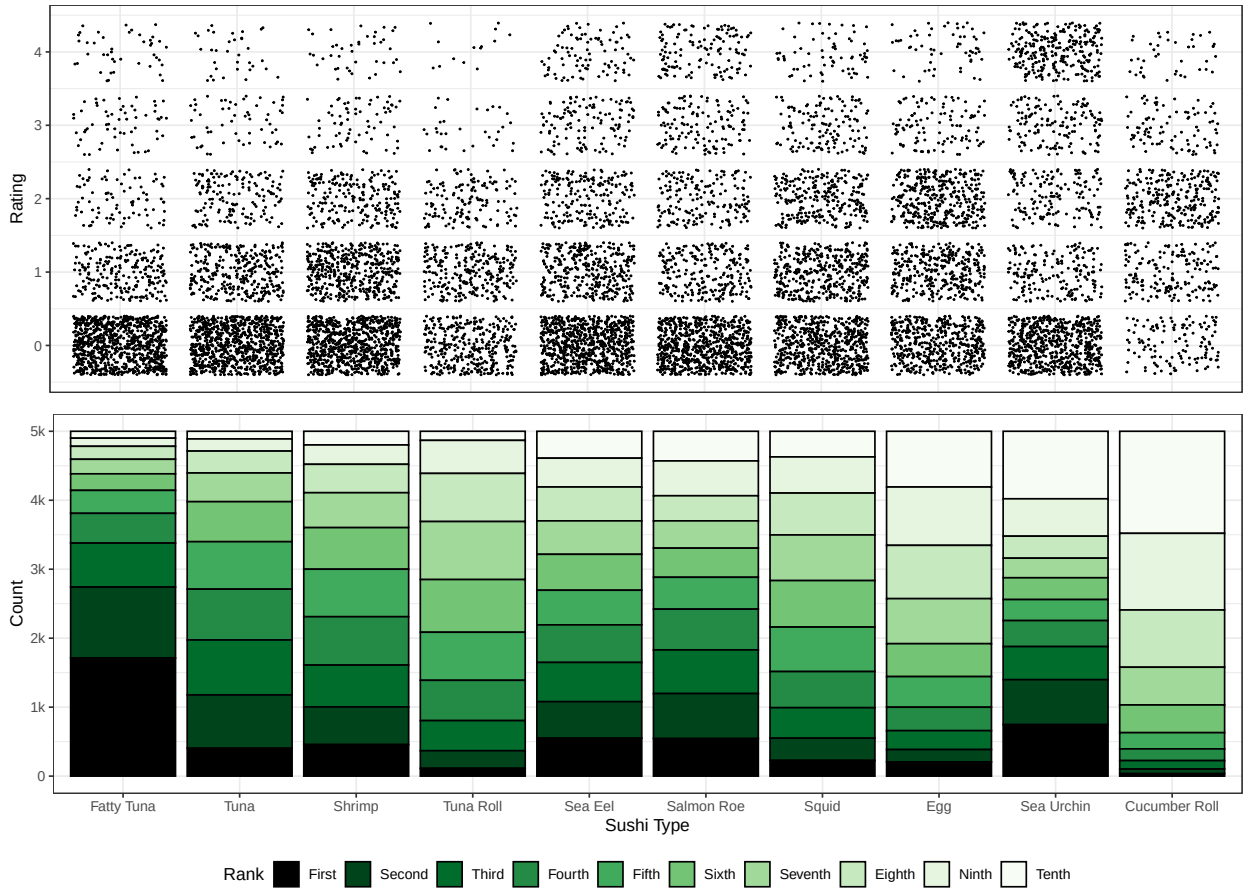

Figure 4.4: Ratings (*top*) and stacked bar charts of rankings (*bottom*) by sushi type.

Estimation and Results

We now fit a BTL-Binomial MFM model to the sushi data. We set priors as follows: To aid interpretability given the large sample size, we assign prior weight to a moderate number of classes using $\lambda = 7$, $\xi_1 = 3$, and $\xi_2 = 1$. We set $a = 0.26$ and $b = 0.77$ using an empirical Bayes approach. We set $\gamma_1 = 20$ and $\gamma_2 = 1$ to provide substantial weight to $\theta \in [10, 30]$. Further information on model estimation and assessment is provided in Appendix B.2.

Results show high posterior probability on $K^+ = 9$, indicating that 9 heterogeneous preference classes exist among the respondents. Table 4.3 summarizes the estimated classes conditional on $K^+ = 9$. Each row contains the posterior mean population proportion $\hat{\pi}_k$, top-3 sushi preferences, and posterior mean consensus scale parameter $\hat{\theta}_k$ for each estimated class (rows ordered by $\hat{\pi}_k$). Classes 1 and 2 are the largest and exhibit the highest consensus. Thus, these classes may reflect reasonably homogeneous plurality classes. The remaining classes represent smaller proportions of the survey respondents and exhibit relatively weak consensus. Still, we may think of these classes as representing subgroups in the population that are present but less well-defined. Additional results can be found in Appendix B.2.

Class	$\hat{\pi}_k$	Top-3 Sushi Preferences	$\hat{\theta}_k$
1	0.21	Sea Urchin, Fatty Tuna, Salmon Roe	10.02
2	0.15	Fatty Tuna, Tuna, Shrimp	16.53
3	0.11	Sea Eel, Fatty Tuna, Tuna	5.78
4	0.10	Salmon Roe, Fatty Tuna, Tuna	4.60
5	0.10	Fatty Tuna, Tuna, Shrimp	4.88
6	0.10	Squid, Shrimp, Tuna	3.33
7	0.09	Sea Eel, Salmon Roe, Shrimp	3.84
8	0.08	Sea Urchin, Fatty Tuna, Salmon Roe	7.71
9	0.05	Egg, Shrimp, Sea Eel	2.71

Table 4.3: Posterior summaries of preference classes conditional on $K^+ = 9$.

4.5 Discussion

In this chapter, we propose a statistical model for joint analysis of rankings and ratings under heterogeneity, the BTL-Binomial MFM model. The model is quite flexible in several important ways. First, it allows for analyzing ranking preference data of various types—pairwise comparisons, partial rankings, as well as complete rankings—jointly with ratings. Second, the model allows for incomplete designs (or structural missingness) where each judge by design would only be assigned to review certain objects and not others. Such "separate ballots" could arise from either conflicts of interest or selective assignments to reflect expertise or manage judges' workloads. The model can also accommodate missing at random data where rankings or ratings are missing due to circumstances unrelated to the quality of objects being assessed, such as reviewer fatigue or cases when subsets of reviewers only rate or only rank the objects. Third, the model allows each reviewer's ranking to be inconsistent with their own ratings, which often happens in practice. Fourth, the BTL-Binomial satisfies Luce's Choice Axiom and the related independence from irrelevant alternatives criterion which is a desirable property in Social Choice Theory (Arrow 1950; Luce 1977). Fifth, the model does not assume a specific number of latent heterogeneous preference groups among the judges (i.e., the amount of heterogeneity), but instead simultaneously estimates both the number of heterogeneous ideologies and the specific preferences of each group, as well as the associated uncertainty.

The BTL-Binomial MFM model makes few parametric assumptions on the reviewers, objects, and data. The model assumes that, for reviewers in each ideological class, proposals have a true underlying quality that can be measured on the unit interval. Rankings and ratings must reflect random deviations from the assumed truth. Rankings must arise from the Bradley-Terry-Luce (BTL) family of ranking distributions. We consider this assumption nonrestrictive, as the BTL family allows for a variety of ordinal data types and has been used in a large variety of application areas (see Chapter 2.1) On the other hand, ratings are assumed to be Binomial. Integer-valued ratings are appropriate whenever they arise

from an ordinal and equally-spaced set with minimum and maximum allowable values. The mean-variance relationship imposed by the Binomial can be tested for validity after model estimation (e.g., Appendix B.2). If this assumption is not met, it may be indicative of an incorrectly estimated number of heterogeneous preference groups. We have proposed sensible goodness-of-fit criteria for assessing the parametric assumptions imposed by the model.

We fit the model by adapting the telescoping sampler of Frühwirth-Schnatter et al. (2021), which provides computationally efficient Bayesian estimation and a natural approach to uncertainty quantification. Furthermore, we provide algorithms for posterior sampling and MAP estimation under a fixed and pre-specified number of heterogeneous preference ideology classes, K.

The MFM approach was chosen to estimate heterogeneity in preferences among judges. We find the Bayesian framework of MFM models to be attractive for three principal reasons. First, Bayesian estimation allows for the incorporation of prior knowledge into the estimation procedure. This is useful in the case of limited preference data, which is common to many applications. When no prior knowledge is available, flat and/or minimally informative priors are available for all model parameters. Second, Bayesian estimation provides a unifying framework for obtaining uncertainty estimates, which is a key component of our work. Third, the proposed estimation procedure may actually reduce computation time when compared to frequentist procedures, due to the fact that K is estimated simultaneously with model parameters and therefore removes the need to repeatedly fit models with different values of K. Additionally, analytic uncertainty results are unavailable for the BTL-Binomial model (and the related Mallows-Binomial model) in the frequentist setting and therefore require the bootstrap, which can be extremely computationally burdensome (see Chapter 3.2.2). This is avoided by the present Bayesian approach where uncertainty estimation is natural. We note that another reasonable approach would have been to use a Dirichlet Process Mixture (Escobar and West 1995). However, it may be computationally difficult to estimate in high dimensions due to estimation via reversible jump MCMC and is inconsistent for the true number of latent classes (Miller and Harrison 2018). Given that we are interested not only

in density estimation, but also in the latent classes themselves, the MFM approach is more suitable.

Two models for rankings and ratings may be directly compared to the BTL-Binomial MFM. First, the Mallows-Binomial proposed in Chapter 3 is a joint statistical model for rankings and ratings that combines a Mallows ranking distribution with independent Binomial rating distributions. Unlike the BTL-Binomial MFM, the Mallows-Binomial allows only for partial or complete rankings, does not satisfy Luce's Choice Axiom, cannot easily handle separate ballots, and cannot estimate heterogeneity directly. Furthermore, Mallows-Binomial is estimated in a frequentist framework, which is computationally slow when the number of objects is large or when uncertainty estimates are desired, which requires the bootstrap. As a result, the BTL-Binomial MFM is much more flexible while still providing a unified and statistical approach to preference modeling with rankings and ratings. (A more thorough comparison between the Mallows-Binomial and BTL-Binomial models is provided in Chapter 5.) A second comparable work is Liu et al. (2022), which proposes a non-parametric algorithm for integrating rankings into ratings. Their algorithm is not statistical and does not yield a preference ordering, but instead returns a "de-quantized" score for each judge and object in a fully data-driven approach. De-quantized scores may be useful and practical for decision-making, but do not allow for estimation of (heterogeneous) preferences or their inherent uncertainty. Liu et al. (2022) also assume internal consistency between rankings and ratings, which is not practical in the motivational settings described herein.

To demonstrate the utility and benefits of modeling preferences with both rankings and ratings, we analyzed three datasets motivated by real-world settings. The first was a paper selection dataset from a hypothetical large and highly competitive academic conference, inspired by current paper review procedures in computer science conferences. Using these simulated data, we showed that incorporating rankings into the traditional rating system improves accuracy of paper selection and reduces the frequency of ties among estimated paper qualities. These benefits exist even when a coarse 5-point rating scale is employed, and are especially apparent when each reviewer can only assess a few papers. While our

simulation studies show that similar benefits can be achieved by increasing the number of papers assessed by each judge, our proposal of incorporating top-4 rankings into the analysis achieves the same benefits with little additional cognitive burden on reviewers. Thus, our analyses show that collecting top-4 rankings and using the BTL-Binomial model for estimating paper quality will make the work of paper selection by computer science conference area chairs both easier and more objective.

The second example concerned a smaller-scale case of peer review, in which a panel assessed grant proposals using rankings and ratings. In this setting, the ability of the BTL-Binomial MFM model to flexibly handle a variety of realistic complexities was demonstrated. These include missing rankings and ratings due to conflicts of interest, reviewer fatigue, and logistical difficulties during the review panel; inconsistency between ratings and rankings at the reviewer level; and potential heterogeneity in preferences among reviewers. We demonstrated how the BTL-Binomial MFM model naturally handles missing data and inconsistent rankings and ratings through a flexible model formulation and simultaneously estimates the number of heterogeneous preference groups among the reviewers with the overall preferences and level of consensus in each group. We identified two heterogeneous preference groups: a dominant collection of reviewers and an "outlier" reviewer whose opinions may have otherwise unduly influenced the panel decision. However, our model could capture different types of heterogeneity, such as potential reviewer preferences for basic versus translational science that have been noted previously (Lee et al. 2013; Smith 2021; Erosheva et al. 2020; Kaatz et al. 2014; Helmer et al. 2017; Marsh et al. 2008). Model results may be used to communicate uncertainty in peer review quality assessment which is important for funding decisions; see Gallo et al. (2023) for an illustration of decision-making with rankings and ratings on another peer review dataset.

The third example relates to the analysis of survey preference data, in which Japanese adults were asked to rate and rank common sushi types. In this setting, the model successfully combines rankings (which are complete but lack granularity) with ratings (which provide granularity but with a very high rate of missingness). Previous work on preference surveys has

modeled heterogeneity among respondents using rankings (e.g., Gormley and Murphy (2010); Mollica and Tardella (2017); Wang et al. (2017)) and ratings (e.g., Patterson et al. (2002); Morey et al. (2008); Breffle et al. (2011)). However, the number of mixture components is usually selected in advance or via goodness-of-fit statistics. Here, we estimate the number of heterogeneous preference groups, as well as their population proportion, preferences, and level of consensus concurrently.

Additional research is needed on joint models for rankings and ratings. As noted previously, the specific parametric form of the rating model based on the Binomial distribution implies that ratings arise from a discrete, ordinal, finite, and equally-spaced set, and furthermore imposes a specific mean-variance relationship on the ratings. These assumptions may not be valid in some contexts, and extensions of our model to modify or relax these parametric assumptions may be useful. Furthermore, the relative influence of rankings and ratings in this model depends on the amount of available data of each type. In some cases, the ability to weight the importance of rankings and ratings during estimation may be important to some practitioners, particularly in contexts where either rankings or ratings are thought to be more relevant. In addition, the BTL-Binomial model also does not incorporate covariates or predictors, which may be of interest. The model can be extended to include covariates in a similar fashion as in BTL models (Tkachenko and Lauw 2016; Chapman and Staelin 1982; Cheng et al. 2010; Schauberger and Tutz 2017; Schäfer and Hüllermeier 2018).

Incorporating rankings into existing decision-making processes or analyses that currently use only ratings, or vice versa, has certain benefits. From a psychological or psychometric perspective, rankings force demarcation and make explicit comparisons but are coarse and impose high cognitive load on the judges. On the other hand, ratings may provide granularity and allow for ties yet may be highly subjective or inconsistent. The BTL-Binomial MFM model provides a principled Bayesian approach for analyzing various types of ranking and rating data jointly, can account for separate ballots, allows for data missing at random, and imposes minimal parametric assumptions. Furthermore, the model estimates the amount and type of heterogeneity among reviewers concurrently. Examples from three different contexts

demonstrate practical applicability of this model for learning preferences. For large-scale conference review, the model provides a mechanism for tie-breaking similar-quality proposals without requiring reviewers to consider more than a few papers. In small-scale panel review, the model successfully identifies an "outlier reviewer" and estimates the preferences of the dominant subgroup for decision-making. In survey sushi data, the model estimates heterogeneous groups of respondents with their distinct preferences, even in the presence of substantial missingness. Overall, we find the BTL-Binomial MFM to be useful and efficient in estimating heterogeneous preferences from rankings and ratings jointly.

Chapter 5

A COMPARISON OF JOINT MODELS FOR RANKINGS AND RATINGS

This chapter presents material that is largely original to this dissertation. It was written in conversation with Dr. Elena A. Erosheva and was supported by the National Science Foundation under Grant No. 2019901.

5.1 Introduction

Chapters 3 and 4 present two joint distributions for rankings and ratings: The Mallows-Binomial and Bradley-Terry-Luce-Binomial. For convenience, we will refer to these distributions as MB and BTLB, respectively. To the best of our knowledge, these are the first distributions for modeling preferences from ordinal and cardinal data jointly in a statistical framework without data conversion.

We begin by describing similarities between the MB and BTLB distributions. In both, we suppose I judges assess J objects. We assume that each object j has a latent quality, $p_j \in [0, 1]$, which represents both the absolute and relative qualities of the collection of objects. Furthermore, we assume there exists some true level of the consensus strength in the population, $\theta > 0$. Conditional on p and θ, ordinal and cardinal preference data arise independently. Shared parameters between ordinal and cardinal data-generating mechanisms tie estimation of overall preferences. In both models, cardinal preference data arises in the form of integer ratings via the Binomial distribution. For ordinal data, both models employ an exponential distance model. Exponential distance models have a long history in ordinal data (e.g., Mallows (1957); Feigin and Cohen (1978); Fligner and Verducci (1986); Critchlow et al. (1991); Mandhani and Meila (2009); Meila and Bao (2010); Meila et al. (2012)), which

we continue.

However, MB and BTLB employ different exponential distance distributions for ordinal data. As suggested by their names, MB employs the Mallows distribution, while BTLB employs the Bradley-Terry-Luce family of distributions. These distributions impose different assumptions and properties on the joint models.

The remainder of this chapter is organized as follows. The next four subsections expound similarities and differences between the MB and BTLB distributions with respect to: the types of data and data missingness permitted by each model (Section 5.2), the assumptions imposed by each model and how those relate to different axioms of Social Choice Theory (Section 5.3), the interpretation of model parameters and results (Section 5.4), and computational considerations of fitting the models to data under frequentist and Bayesian frameworks (Section 5.5). Section 5.6 concludes the chapter with a brief discussion on selecting a model in practice.

5.2 Data Types and Missingness

Both MB and BTLB jointly model cardinal and ordinal data, but the precise data types permitted by each model vary. Relatedly, the distributions differ in the types of missing data they may account for. In practice, these differences affect which distribution is most appropriate given cardinal and ordinal preference data. We describe these differences below.

Both MB and BTLB distributions assume a Binomial model for cardinal data, which capture cardinal preferences in the form of integer ratings between 0 and some known maximum rating, M. We use the convention that low ratings indicate high-quality objects. This convention maintains a symmetry with rankings, in which low rank places correspond to highly-preferred objects. Whenever ratings arise from a discrete, finite, and equally-spaced set, they can be converted via linear transformation, $f : x \to x'$, into the integer form required by each model. For example, suppose the original ratings, x, are collected on a numerical scale from 1 to 5 in single decimal increments, where 5 indicates the top rating. The original 41-point scale may be linearly transformed into the reverse integer scale required by

the MB and BTLB distribution via $f(x) = 50 - 10x$, such that $f(1) = 40$ and $f(5) = 0$.

Regarding ordinal data, MB may be used for complete and partial rankings. Thus, given a collection of J objects, MB permits rankings where all J objects are ranked from best to worst, or top-R rankings, $R \leq J$, such as a top-3 ranking from a collection of 10 objects. It is assumed that judges assess the complete collection of objects when creating these rankings, i.e., objects which are not included in a partial ranking are deemed worse than those which are included. BTLB similarly permits complete and partial rankings, but also allows for pairwise and groupwise comparisons. In pairwise or groupwise comparisons, no information may be gleaned on a judge's preferences regarding objects that were not included in their ranking, since those objects were simply not considered.

5.2.1 Separate Ballots

Related to the inclusion of pairwise or groupwise comparison data is the case of *separate ballots*. Separate ballots arise whenever different judges have access to different subsets of objects when expressing preferences. For example, separate ballots occur when judges are not allowed to assess certain proposals due to a conflict of interest. Another example of separate ballots is *distributed peer review*, where perhaps 50 reviewers assess 50 proposals, but each reviewer is randomly assigned only $R = 10$ proposals to read and assess. Pairwise and groupwise comparisons are examples of separate ballots by construction. Since the BTLB distribution permits pairwise and groupwise comparisons, it is thus able to easily capture ordinal preference data arising under separate ballots (Chapter 4 contains additional details on how this is represented in its model likelihood).

Alternatively, pairwise and groupwise comparisons may be thought of complete rankings made under separate ballots. For example, suppose objects A, B, C, D, E, and F represent the complete collection of objects and some specific judge only has access to the first three. If she provides the ranking $\{A \prec B \prec C\}$, we may consider her ranking as either a groupwise comparison among objects A, B, and C, or alternatively as a complete ranking under her "separate ballot." If we operate under the latter assumption, we may ask: Would it be

possible to model such pairwise or groupwise comparison data using the MB distribution?

The Mallows ranking distribution was not designed to handle separate ballots, resulting in substantial complications for using the MB distribution in the presence of pairwise and groupwise comparison data. Still, it is theoretically possible under minor modifications to the model likelihood and corresponding changes to the interpretation of model parameters (Lu and Boutilier 2011). These changes are best demonstrated via example, which we now provide.

Suppose there exists a collection of $J = 6$ objects, $\{A, B, C, D, E, F\}$ and $I = 2$ judges to assess them, Elena and Michael. Furthermore, suppose Elena has access only to objects A, B, and C, while Michael has access to C, D, E, and F. We assume Elena and Michael express preferences as stated in Table 5.1. In this example, Elena and Michael each provide

Judge	Ballot	Ranking	Ratings
Elena	A, B, C	$\{A \prec B \prec C\}$	$A = 1, B = 2, C = 4$
Michael	C, D, E, F	$\{C \prec D \prec E \prec F\}$	$C = 5, D = 6, E = 6, F = 7$

Table 5.1: Example preference data arising under separate ballots.

internally consistent preferences. Their preferences also align with each other, in that we can use the observed ratings and rankings to reasonably estimate that $A \prec B \prec C \prec D \prec E \prec F$. Statistical models, like MB and BTLB, could additionally allow us to quantify the uncertainty inherent in estimated preferences. Under separate ballots, it is not obvious how to use the MB for estimating preferences with uncertainty in this setting.

If we consider Elena's perspective alone, we may modify the probability of observing her preferences under a MB distribution as follows:

$$\begin{aligned}\Pr_{p,\theta}\big[\Pi = \{A \prec B \prec C\}, X = [1, 2, 4]\big] \\ = \frac{e^{-\theta d_K(\Pi, \text{Order}(p_A, p_B, p_C))}}{\psi_{R=3,J=3}(\theta)} \times \prod_{j=A,B,C} \binom{M}{X_j} p_j^{X_j} (1 - p_j)^{M - p_j}.\end{aligned} \tag{5.1}$$

Similarly, the probability of Michael's preferences may be written:

$$\begin{aligned}\Pr_{p,\theta}\big[\Pi = \{C \prec D \prec E \prec F\}, X = [5,6,6,7]\big] \\ = \frac{e^{-\theta d_K(\Pi, \text{Order}(p_C, p_D, p_E, p_F)))}}{\psi_{R=4,J=4}(\theta)} \times \prod_{j=C,D,E,F} \binom{M}{X_j} p_j^{X_j}(1-p_j)^{M-p_j}.\end{aligned} \tag{5.2}$$

Note that the likelihoods in Equations 5.1 and 5.2 share the parameters θ and p_C, but Elena's likelihood alone contains p_A and p_B, while Michael's likelihood alone contains p_D, p_E, and p_F.

These two data likelihoods could be straightforwardly multiplied, and then parameter estimation could be performed via maximum likelihood. While theoretically possible, the practitioner should keep in mind two important considerations: First, it was noted in Chapter 3 how interpretation of the consensus scale parameter θ in the MB distribution depends upon both the number of objects, J and the size of the rankings, R (this dependence of interpretation does not occur in the BTLB distribution). Since R and J differ for Elena and Michael, interpretation of a single θ must be treated with care. One may consider θ under an interpretation of marginalization, i.e., as the level of consensus averaging out the differing sizes of separate ballots and ranking lengths. Second, the value of θ may affect both point estimates and estimated uncertainty of the object quality vector p in MB models, particularly in cases where rankings and ratings do not align. Thus, one should be aware that the presence of separate ballots may affect the estimates of p in addition to θ.

In summary, it is theoretically possible to fit an MB distribution in the presence of pairwise or groupwise ordinal preference data, but special care must be taken when interpreting or using model results. In general, the BTLB distribution is more appropriate for modeling ordinal and cardinal data arising under separate ballots. However, if one still elects to model preferences made under separate ballots with the MB distribution (perhaps to align with other assumptions imposed by MB), we recommend additionally fitting the BTLB distribution and comparing the resulting estimates as a check for robustness.

5.2.2 Missing Data

Both MB and BTLB handle missing preference data similarly. Consider preference data that is entirely missing due to reasons unrelated to the quality of the objects. Chapters 3 and 4 included examples from grant panel review in which reviewers did not rate individual proposals based on reasons such as conflicts of interest, intermittent connectivity issues, or lack of expertise. In such cases, the missing at random (MAR) assumption is appropriate. MAR occurs when missingness is not random, but is instead fully explained by variables for which there is complete information (Little and Rubin 2019). Under the MAR assumption, the rating and ranking components of the MB and BTLB distributions may simply be removed and estimation based on the available data will not be biased. If missingness was due to circumstances related to object quality, this case would no longer be considered as MAR and one would need to carry out a different treatment of missing data (Little and Rubin 2019). We consider the scenario where data are missing not at random to be beyond the scope of this dissertation.

5.2.3 Ties

In some preference data applications, judges are allowed to explicitly express ties in ordinal preferences. For example, the ranking $\{A \prec B = C \prec D\}$ suggests that objects B and C are equally preferred by a judge. Neither MB nor BTLB distribution permits these types of ordinal preferences.[1] The expression of ties in partial rankings is more subtle. In partial rankings, unranked objects are deemed worse than those which are ranked by the model, but specific preferences among the unranked objects are unknown. Although this may be considered an expression of a preference tie, it is more suitably considered as a case of missing data on the precise preference ordering among the unranked objects. The underlying

[1]Previous authors have proposed variations of the Mallows (Adkins and Fligner 1998; Brancotte et al. 2015; Zhu et al. 2019) and BTL (Rao and Kupper 1967; Davidson 1970; Cattelan et al. 2013; Tutz and Schauberger 2015; Sawadogo et al. 2017) ranking distributions which allow for ties. These variations often involve addition parameters and assumptions, and for simplicity are not considered in this dissertation.

constructions of the Mallows and BTL distributions allow estimation to be unaffected by partial rankings.

5.3 Assumptions

The MB and BTLB distributions share many assumptions on the underlying structure of the ordinal and cardinal preference data, yet some differences exist. Differences largely arise from the distinct ordinal data distributions used by the models. In this subsection, we explore the assumptions made by the two models.

We begin by specifying assumptions (or lack thereof) that are shared by the MB and BTLB distributions:

1. *Homogeneity:* Each object is assumed to have a single, true underlying quality which can be represented on the unit interval. The true quality is a reflection of both the object's inherent quality and its quality relative to other objects in the collection. By extension, the distributions assume that judges are homogeneous in their preferences. That is, the judges express preferences which are reflections of the true underlying qualities of the objects, subject to error.[2]

2. *Conditional independence:* Conditional on the true model parameters, both models assume that the observed ratings and rankings are conditionally independent. This assumption implies that judges need not be internally consistent when expressing ratings and rankings. Not requiring internal consistency makes the models more flexible and increases the situations in which either can be used since inconsistency in common in practice. However, the conditional independence assumption means that the MB and BTLB distributions may not be appropriate when rankings and ratings are required to be internally consistent.

[2]If judges exhibit non-homogeneous preferences, a latent class mixture model may be applied to either model to capture the heterogeneity. This technique was applied in Chapter 4 to the BTLB distribution, but could be straightforwardly applied to the MB distribution as well.

3. *No requirement of equal data:* Neither model assumes that there are equal amounts of ordinal or cardinal preference data, either overall or at the level of each judge. As a result, either model may be used in cases where certain judges provide ratings and others provide rankings.

4. *Error structure of ratings:* Both models assume that the ratings for each object may be modeled via a Binomial distribution. This parametric model imposes a specific mean-variance relationship on the ratings, as described in Chapter 3. When using either model, this assumption can easily be checked based on comparing theoretical and observed means and variances of the ratings for each object.

Next, we specify three major differences in assumptions between the MB and BTLB distributions:

1. *Independence from Irrelevant Alternative (IIA):* The BTLB distribution employs the Bradley-Terry-Luce family of distributions for rankings, which is based on Luce's Axiom of Choice. This axiom implies the IIA criterion (and is, in fact, a slightly stronger criterion). IIA is often considered a desirable property in Social Choice Theory (Arrow 1950), but is not always realistic. In contrast, the MB distribution employs the Mallows distribution for rankings, which does not satisfy or require IIA (Marden (1996); proof provided in Chapter 2.4.1).

2. *Stagewise ranking:* Related to IIA, the BTL family of distributions employed by BTLB can be interpreted as a stagewise selection distribution for rankings, which is not true for the Mallows model. Instead, the Mallows assigns probability to a ranking based on its distance to the overall ranking. As such, we may consider it as a "holistic" ranking creation process, in which a judge forms his/her ranking all at once while attempting to balance the relative orders between objects simultaneously.

3. *Error structure of rankings:* The error structure of rankings in both models depends

upon the scale parameter θ. However, the precise error structure varies between the models based on their unique parameterizations. In the MB distribution, θ controls the probability that a ranking of a given Kendall distance to the true overall ranking is drawn; all rankings of a given Kendall distance to the true overall ranking have the same probability. Thus, the MB model's error structure depends only on distance, and not on the order of specific objects in the ranking. In the BTLB distribution, θ instead controls the probability that each object is selected over another at each stage of the ranking process based on the differences in their quality parameters, p. That is, the probability of observing rankings depends on both θ and the continuous values of p, instead of simply on θ and the discrete order of p.

5.4 Parameter Interpretation

Both the MB and BTLB distributions are parameterized by the vector-valued object quality parameter, p, and the consensus scale parameter, θ. However, these parameters should not all be interpreted identically. In this section, we describe the similarities and differences between parameter interpretation in the MB and BTLB distributions.

First, we note that the parameter p may be interpreted identically in both models with respect to object quality. That is, the value in p corresponding to each object represents the absolute and relative qualities among objects on the unit interval in the MB and BTLB distributions. Furthermore, the order of the values of p from least to greatest represents the overall ranking, which is often of interest. This vector-valued parameter is often of highest interest in applications, and its shared meaning in both distributions facilitates comparison of results.

The consensus scale parameter, θ, does not have identical interpretation in each model. In the MB distribution, θ controls the probability that a ranking of a given distance to the overall ranking is drawn (see previous section for details). Alternatively, in the BTLB distribution θ controls the probability of stagewise object selection in tandem with the object quality parameters for objects under consideration. One may also use θ in the BTLB distribution

to calculate pairwise selection probabilities. For example, if $p_A = .4$, $p_B = .6$, and $\theta = 5$, the probability that A is selected above B in a pairwise tournament is,

$$\Pr\big[A \prec B\big] = \frac{e^{-\theta p_A}}{e^{-\theta p_A} + e^{-\theta p_B}} = \frac{e^{-2}}{e^{-2} + e^{-3}} \approx 0.73. \tag{5.3}$$

This type of probability cannot be easily calculated from the MB distribution (or the related Mallows ranking model). Due to the differing interpretations of θ in each distribution, specific values should not be compared.

Despite the differing specific interpretations of θ in the MB and BTLB distributions, each model retains the interpretation that low values of θ represent weak consensus in rankings among judges, and that high values of θ indicate strong consensus in rankings among judges. What precisely constitutes a high or low value of θ depends on the model and type of available data. To gain intuition for the meaning of a specific value of θ in the context of a model, examining the predictive or posterior predictive distribution may be useful.

5.5 Computation

Last, we consider the computational cost of estimation in each distribution. We note that computational cost depends on the framework in which each model is estimated, and that MB was proposed in a frequentist setting and BTLB was proposed in a Bayesian setting that incorporated latent classes. Still, we discuss various qualities of each distribution that affect computational cost during estimation.

We first consider estimation in the frequentist setting. As described in Chapter 3, frequentist estimation of parameters in the MB distribution is NP-hard. The challenge arises from the discrete nature of the model's overall ranking parameter. Although an efficient, exact estimation algorithm was proposed (as well as fast approximate algorithms), the computational cost associated with estimation may be substantial, especially in the presence of a large number of objects, a large number of judges, or weak consensus among judges. In contrast, the BTLB distribution relies solely on continuous parameters and thus frequentist estimation is not NP-hard. However, Hunter et al. (2004) identified challenges with ensuring

exact frequentist estimation of BTL distribution parameters, which contributed to the BTLB distribution being proposed in a Bayesian framework.[3] We observe that the estimation of judge heterogeneity via a latent class mixture model is important to many preference data applications. Estimation of such a model for the MB distribution in the frequentist setting would likely be intractably slow, and as such was not considered when proposing the model. This is not a problem in the BTLB model, as discussed in Chapter 4.

Estimation in the Bayesian setting does not vary substantially between models with respect to computational cost. Unfortunately, neither model has conjugate priors available. Thus, estimation via Markov chain Monte Carlo (MCMC) is required. Although no Bayesian estimation procedure has been proposed for MB, a similar MCMC method to that seen in Chapter 4.3.2 for the BTLB distribution should suffice. We note that the algorithms presented in Chapter 4.3 occasionally exhibit difficulty with mixing and convergence, and thus it is important to check for such properties during Bayesian estimation of the MB distribution. Bayesian estimation of both the MB and BTLB distributions may require reasonable computational power and statistical knowledge to confirm accurate results.

5.6 Discussion

In this chapter, we have explored similarities and differences between the Mallows-Binomial (MB) and Bradley-Terry-Luce-Binomial (BTLB) distributions. To organize our discussion, we have considered four main axes of comparison: First, we considered the types of data and missingness allowed by each model. Although each model can handle identical types of ratings and generally treats missing data identically, BTLB flexibly allows for a wide variety of ordinal data types (including those arising under separate ballots), while MB is largely limited to partial and complete rankings. Second, we described the varying assumptions of each model. In their simplest form, both models assume homogeneity, conditional independence between rankings and ratings, and the same distributional form for ratings; neither

[3]Chapter 4 does in fact propose an algorithm for maximum *a posteriori* estimation of the BTLB distribution and discuses priors which align MAP estimates with the frequentist MLE.

assumes judges provide equal amounts of ordinal or cardinal preference data. However, the models differ based on their distributional form for rankings, in which the BTLB model assumes a stagewise ranking process and implies the IIA criterion. Third, we described how parameters may be interpreted differently in each model. Although both estimate object quality parameters and consequently, the overall ranking, the interpretation of θ is model- and data-dependent, and thus should not be compared between models. Fourth, we considered the computational complexity of estimating each model in both frequentist and Bayesian frameworks, identifying potential challenges in either case.

In the presence of ordinal and cardinal data, one should carefully consider the varying requirements and assumptions of MB and BTLB before selecting a model. To use either model, it is important to consider if ratings can be appropriately modeled via independent Binomial distributions (perhaps after linear transformation to a finite set of integers). If so, carefully weighing the distinct assumptions of the BTL and Mallows ranking distributions is important: Were judges likely to adhere to the IIA criterion? Did judges employ a stagewise or holistic approach to forming ordinal preferences? Did the data collection mechanism impose separate ballots? Model selection may also depend upon the types of results that are important to practitioners, such as the ability to estimate pairwise selection probabilities in the BTLB distribution. Last, computational considerations are often of practical importance and may influence whether a model is estimated in a frequentist or Bayesian framework.

Chapter 6

BAYESIAN RANK-CLUSTERING

This work was written in collaboration with Dr. Elena A. Erosheva and was supported by the National Science Foundation under Grant No. 2019901. We would like to acknowledge Drs. T. Brendan Murphy and I. Claire Gormley for inspiring conversations on rank data modeling during the early stages of this work.

6.1 Introduction

In a traditional analysis of ordinal data, we assume a group of I judges assess J objects by providing ordinal preferences, Π. Each judge's ordinal preferences, Π_i, may be a partial ranking, complete ranking, pairwise comparison, groupwise comparison, or mixture thereof. Then, a model uses the observed preference data to estimate the overall *rank* of each object at the population level. Most analyses, including those from all statistical models reviewed and introduced previously in this dissertation, derive or estimate the rank of each object such that each object receives a unique rank. Thus, they obtain an *overall ranking* that orders all objects from best to worst on the basis of their estimated rank.[1] Analyses of this kind are performed in diverse settings, such as to rank candidates in an election using ranked choice votes, sports teams in a league using individual game outcomes, or graduate programs in an annual list using preference rankings elicited in a survey of academics.

However, requiring estimated ranks to be unique is not always useful or appropriate: When objects are indistinguishable in quality or ability, they may be more correctly considered as *rank-clustered*, i.e., identical in rank. Estimating overall orderings with rank-clusters,

[1]The overall ranking is sometimes referred to as a *social order* or *consensus ranking* in the philosophy or computer science literatures, respectively (see Chapter 2.2). However, since the population may not truly exhibit a social ordering or consensus, we instead use the phrase *overall ranking* in this chapter.

when appropriate, may therefore improve interpretability and aid accurate inference, prediction, and decision-making.

Limited methods exist for estimating rank-clusters with ordinal preference data. Many of these are based on *parameter fusion*, which is the process of simultaneously estimating parameter values and groups of parameters that should be set equal in value (i.e., "fusing" parameters together). Masarotto and Varin (2012) analyze pairwise comparison data from sports tournaments with techniques from parameter fusion under the Bradley-Terry model. The Bradley-Terry model is parameterized by the vector $\omega \in \mathcal{R}^J_{>0}$, in which each ω_j corresponds to the *worth* of object j (see Chapter 2.4.1 for more details on the parameterization and interpretation of Bradley-Terry models for ordinal data). They estimate an overall ranking of teams with rank-clusters by applying the frequentist *fused lasso* (Tibshirani et al. 2005), in which the absolute difference between every pair of worth parameters is penalized after some data-driven normalization. In this approach, the fused parameters are made equal and thus create a rank-cluster among their corresponding objects. The approach of Masarotto and Varin (2012) was applied to additional datasets in sports (Tutz and Schauberger 2015) and academic journal rankings (Varin et al. 2016; Vana et al. 2016). Jeon and Choi (2018) argue that shrinkage methods like those proposed by Masarotto and Varin (2012) and Tutz and Schauberger (2015) were developed specifically for pairwise comparisons, and thus have inappropriate penalty functions for application to richer kinds of ordinal data like partial or complete rankings. As a result, Jeon and Choi (2018) proposed a modified regularization penalty that may be applied to partial or complete rankings under the more general Plackett-Luce(ω) model.

The parameter fusion methods described in the previous paragraph exhibit five distinct disadvantages: First, maximum likelihood estimation of Bradley-Terry-Luce models, even in their simplest forms, often suffers from numerical instability and slow computational speed. As a result, numerous authors have proposed complex algorithms to improve estimation accuracy or speed (Hunter et al. 2004; Maystre and Grossglauser 2015; Turner et al. 2020; Nguyen and Zhang 2023). Second, uncertainty quantification is challenging and theoretically

tenuous in lasso-based methods (Tibshirani 1996; Fan and Li 2001). Third, lasso penalty parameters may be difficult to select, requiring data-driven or *ad hoc* techniques (Tibshirani 1996; Masarotto and Varin 2012). Thus, interpretation of the resulting parameter estimates and associated uncertainty is reliant on the specific choice of penalty parameter. Fourth, prior knowledge on the amount and size of rank-clusters cannot be directly incorporated into the frequentist framework: Although the penalty parameter influences estimation of rank-clusters, the specific meaning of various possible choices is not directly interpretable in advance. Fifth, to our knowledge only Jeon and Choi (2018) have formulated a parameter fusion method for ordinal data types other than simple pairwise comparisons. As a result, only their method is available to estimate rank-clusters from complete or partial ranking data.

Many of these disadvantages may be addressed using Bayesian methods. For example, Bayesian counterparts of the lasso and fused lasso (specifically, the Bayesian lasso (Park and Casella 2008) and Bayesian fused lasso (Casella et al. 2010)) allow for valid uncertainty quantification and the incorporation of prior knowledge via the selection of interpretable hyperparameters (and not penalty parameters). Additional Bayesian methods for shrinkage and variable selection may also be useful for rank-clustering with ordinal data, but have not yet been studied in that context.

Bayesian methods for shrinkage and variable selection can be broadly categorized into two classes: Continuous shrinkage and spike-and-slab priors. Examples of continuous shrinkage priors include the Bayesian Lasso (i.e., Laplace; Park and Casella (2008)), Bayesian Fused Lasso (Casella et al. 2010), t-distribution (Song and Cheng 2020), Normal-Exponential-Gamma (Griffin and Brown 2005; Shimamura et al. 2019), Normal-Gamma (Griffin and Brown 2010), horseshoe (Carvalho et al. 2010), and Dirichlet-Laplace (Bhattacharya et al. 2015). No continuous shrinkage priors place positive probability on coefficients (or their differences) being precisely zero, which would encourage sparsity. Thus, parameter fusion must be performed via thresholding the posterior distribution, which is often ad-hoc (Porwal and Rodriguez 2021). This leads us to consider priors of the latter class. Spike-and-slab

priors (Mitchell and Beauchamp 1988; George and McCulloch 1997; Ishwaran and Rao 2005) assign weight to both a point-mass at 0 ("spike") and a continuous density function ("slab"). Although the specific formulations of these priors vary, they allow us to estimate which parameters are precisely zero in a probabilistic framework. Thus, spike-and-slab priors can be used for parameter fusion and uncertainty quantification. However, we are aware of only one existing modification of this prior class for parameter fusion: Wu et al. (2021) apply spike-and-slab to the differences in successive parameters in the linear regression setting. In their method, the order of parameters from least to greatest in coefficient value must be known in advance (as in the fused lasso). This is not practical in the ordinal data setting because the parameter order is equivalent to the overall ranking, whose estimation is a primary goal in the canonical preference data problem. Thus, no Bayesian parameter fusions methods exist which may be directly applied to ordinal data analyses with rank-clustering.

In this chapter, we propose a Bayesian method for ordinal data analysis that estimates an overall ranking of objects with rank-clusters. We use the Bradley-Terry-Luce family of distributions, which allows us to analyze ordinal preferences from pairwise comparisons, partial rankings, complete rankings, and comparisons made under separate ballots. We propose a novel spike-and-slab prior that uses partitions to induce rank-clusters. The model does not require the parameter order nor the number or size of rank-clusters to be known in advance. Instead, these quantities are treated as random variables and estimated simultaneously. We develop a computationally-efficient Gibbs sampler for estimation and apply the model to real and simulated data.

The rest of this chapter is organized as follows. We propose the Partition-based Spike-and-Slab Fusion prior in Section 6.2 and apply it to a Bradley-Terry-Luce model for ordinal data in Section 6.3. We develop a computationally-efficient Gibbs sampler based on reversible jump Markov chain Monte Carlo in Section 6.4 and demonstrate its accuracy on simulated data. Section 6.5 applies the model to ranked choice voting data from the 2021 Minneapolis mayoral election. We conclude with a brief discussion in Section 6.6.

6.2 Partition-based Spike-and-Slab Fusion Prior

Suppose data are drawn exchangeably from a model, $\mathcal{M}$, parameterized by the vector ω. We suppose ω is of length J and let each $\omega_j \in \Omega$, $\Omega \subseteq \mathbb{R}$. Our goal is to estimate ω under the belief that some pairs or groups of parameters in ω may be clustered (i.e., *fused*). We say that two parameters $m, n \in \{1, \ldots, J\}$, $m \neq n$, are clustered precisely when $\omega_m = \omega_n$. Clustered parameters may take on any value in their domain, Ω.

Before specifying the prior, we provide some notation on partitions. A partition of an object set $\mathcal{J} = \{1, 2, \ldots, J\}$ is a collection $g = \{C(1), C(2), \ldots, C(K)\}$ of K disjoint nonempty subsets (henceforth referred to as "clusters") of $\mathcal{J}$ such that their union forms $\mathcal{J}$. Let $C^{-1}(j)$ represent the cluster that contains object $j \in \mathcal{J}$. We let $S(k) = \left|\{C(k)\}\right|$ be the size of the subset $C(k)$, and denote by K the number of clusters in g. To emphasize dependence on g, we often write K_g, $C_g(k)$, etc. Lastly, we let $\mathcal{G}$ represent the collection of all partitions g of $\mathcal{J}$, and let $\mathcal{G}_k = \{g \in \mathcal{G} : K_g = k\}$.

We are now ready to specify the Partition-based Spike-and-Slab Fusion (PSSF) prior. Under PSSF, ω is assumed to be generated via the following hierarchical model:

$$\begin{aligned} G &\sim f_G \\ \nu_k | G = g &\overset{iid}{\sim} f_\nu && k = 1, 2, \ldots, K_g \\ \omega_j &= \nu_{C_g^{-1}(j)} && j \in \mathcal{J} \end{aligned} \tag{6.1}$$

In Equation 6.1, $f_G(\cdot)$ is a probability mass function on $\mathcal{G}$ and $f_\nu(\cdot)$ is a probability density function on Ω. In words, the prior generates a partition g, and then assigns a unique value ν_k to each cluster $C(k) \in g$. Last, each parameter in ω is assigned the value of ν corresponding to its cluster in g.

As an example, suppose $\mathcal{J} = \{1, 2, 3\}$ and we draw $g = \{C(1), C(2)\}$ such that $C(1) = \{2\}$ and $C(2) = \{1, 3\}$, and draw $\nu = [5, 10]$. Then, $\omega = [10, 5, 10]$ because,

$$\omega_1 = \nu_{C_g^{-1}(1)} = \nu_2 = 10,$$
$$\omega_2 = \nu_{C_g^{-1}(2)} = \nu_1 = 5, \text{ and}$$
$$\omega_3 = \nu_{C_g^{-1}(3)} = \nu_2 = 10.$$

6.2.1 Marginal Prior Probabilities

A useful feature of the PSSF prior is that, regardless of f_G, the marginal distribution of each ω_j follows f_ν. This is because,

$$P[\omega_j] = \sum_{k=1}^{J} P[\nu_k | j \in C(k)] P[j \in C(k)] \tag{6.2}$$
$$= P[\nu_1] \sum_{k=1}^{J} P[j \in C(k)] \tag{6.3}$$
$$= f_\nu(\cdot). \tag{6.4}$$

Equation 6.2 holds as there cannot be more than J clusters and each object belongs to a cluster, Equation 6.3 holds by the exchangeability of ν_k, and Equation 6.4 holds since $P[\nu_1] = f_\nu(\cdot)$ by definition and the Law of Total Probability.

6.2.2 Relationship to Spike-and-Slab

We have not yet explained the proposed PSSF prior's relationship to the spike-and-slab. It is easiest to understand their connection by considering the joint prior distribution on two arbitrary component parameters, ω_m and ω_n, such that $m \neq n$. Due to the partitioning structure of parameters in the PSSF prior, there is prior probability associated with a parameter cluster. Thus, their joint prior distribution contains a "spike" component along the line $\omega_m = \omega_n$, with density of that line determined by f_ν. Oppositely, given $\omega_m \neq \omega_n$ their joint prior distribution reflects independent draws from f_ν.

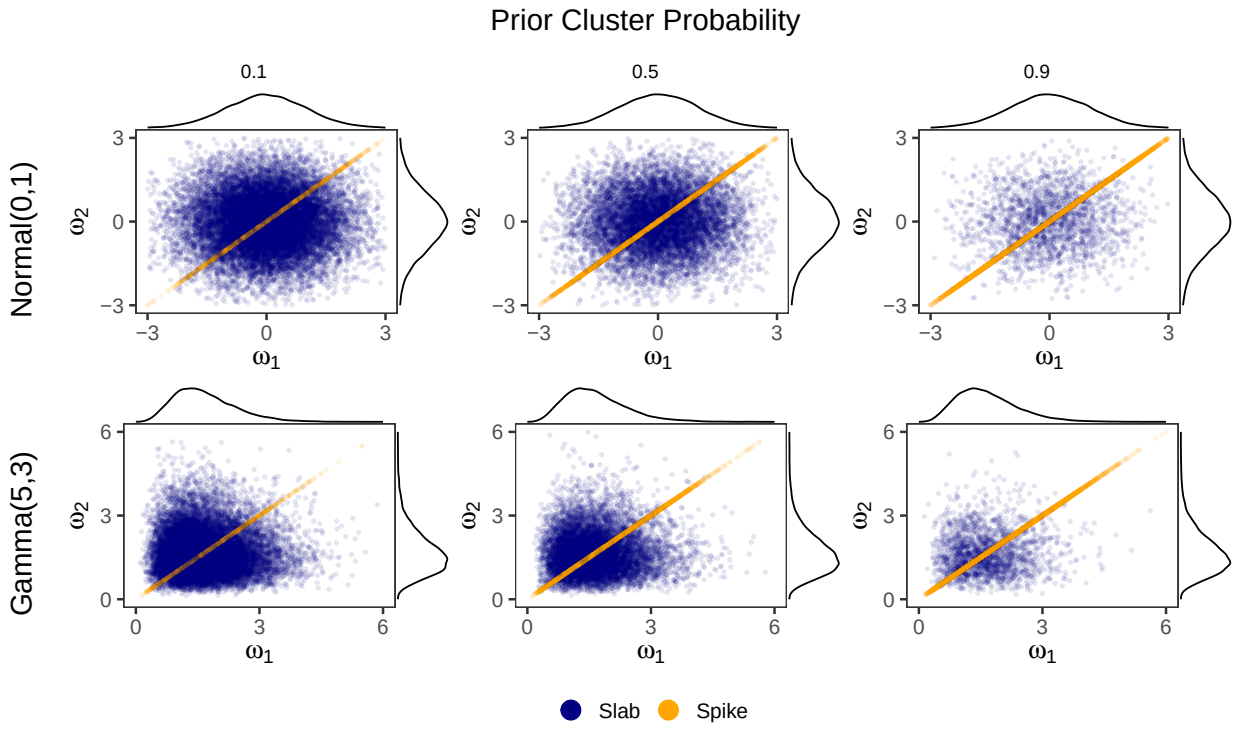

Figure 6.1: Joint distribution of (ω_1, ω_2) under the PSSF prior with varying combinations of f_G and f_ν. In all cases, $\mathcal{J} = \{1, 2\}$, and plots show 20,000 sampled values with marginal density estimates along the axes. Rows correspond to the choice of f_ν and columns to f_G.

Figure 6.1 gives examples of the PSSF prior under varying choices of f_G and f_ν. In all panels, we let $\mathcal{J} = \{1, 2\}$ and display the joint prior distribution of (ω_1, ω_2). In this setting, there are only two unique partitions, $g = \{1, 1\}$ and $g = \{1, 2\}$. Thus, we specify the prior f_G by stating the so-called "cluster probability," i.e., the probability that $g = \{1, 1\}$. Columns correspond to cluster probabilities $0.1, 0.5$, and 0.9, respectively. Rows correspond to $f_\nu = \text{Normal}(0, 1)$ and $\text{Gamma}(5, 3)$, respectively. We notice that as the cluster probability increases, so does the density of points in the spike component. Regardless of f_G, marginal distributions of each parameter follow f_ν. The marginal relationships seen in Figure 6.1 hold

identically even as $\mathcal{J}$ grows.

Additionally, we display the difference between parameters, $\omega_2 - \omega_1$, across different scenarios in Figure 6.2. The rows and columns are identical to that from Figure 6.1 and make clear the PSSF prior's relationship with the traditional spike-and-slab, which has a spike component at 0 and a background slab density.

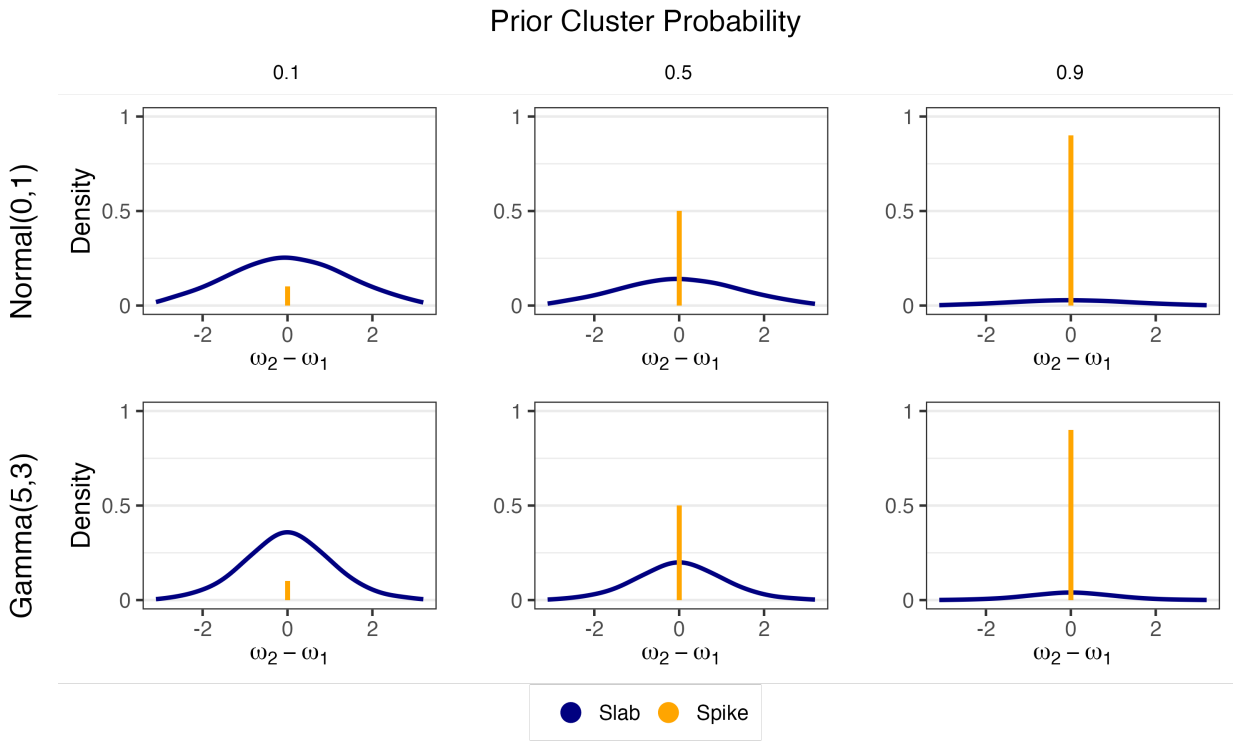

Figure 6.2: Distribution of $\omega_2 - \omega_1$ under the PSSF prior with varying combinations of f_G and f_ν. In all cases, $\mathcal{J} = \{1, 2\}$. Rows correspond to the choice of f_ν and columns to f_G.

6.3 Rank-Clustered Bradley-Terry-Luce Model

We now introduce the Rank-Clustered Bradley-Terry-Luce model for ordinal data using the notation of previous chapters. As a reminder, let I be the number of judges who assess J

objects. Let Π_i represent the ordinal preferences provided by judge i, which may be a partial ranking, complete ranking, pairwise comparison, or groupwise comparison. Let $R_i = |\Pi_i|$ and $\mathcal{S}_i$ denote the objects in $\mathcal{J}$ considered by judge i, such that $\mathcal{S}_i \subseteq \mathcal{J}$. R_i and $\mathcal{S}_i$ are assumed known.

Under the *Rank-Clustered Bradley-Terry-Luce* (BTL) model, we assume ordinal data is generated via the following Bayesian model:

$$\begin{aligned} \omega &\sim \text{PSSF}(f_G \propto \text{Poisson}(K_g|\lambda), f_\nu = \text{Gamma}(\nu_k|a_\gamma, b_\gamma)) \\ \Pi_i|\omega &\overset{iid}{\sim} \text{Bradley-Terry-Luce}(\omega|\mathcal{S}_i, R_i) \qquad\qquad i = 1, \ldots, I \end{aligned} \tag{6.5}$$

Rank-Clustered BTL applies the proposed PSSF prior under specific choices of f_G and $f\nu$ to the BTL family of distributions for ordinal data. We emphasize that the model does not pre-specify the number of clusters, a specific rank-clustering structure, or the order of objects. These are treated as random variables and estimated simultaneously.

6.3.1 Prior Selection

We now discuss the selection of priors and hyperparameters. We set f_G according to

$$f_G(g) \propto \text{Poisson}(K_g|\lambda). \tag{6.6}$$

In words, the prior probability of drawing a specific partition g depends only on how many unique clusters, K_g, it contains. Thus, every partition with the same K_g has equal prior probability. As a consequence, cluster sizes do not explicitly impact the prior probability of each g. Still, there is an implicit connection between cluster size and K_g. For example, if $K_g = J$, every cluster must be a singleton. In this setup, one could set $\lambda \approx 1$ to encourage rank-clustering, or $\lambda \approx J$ to discourage rank-clustering. Next, we set f_ν according to

$$f_\nu(\nu_k) = \text{Gamma}(\nu_k|a_\gamma, b_\gamma). \tag{6.7}$$

This Gamma prior has been used in Bayesian estimation of BTL models as it allows for closed-form Gibbs sampling via data augmentation (Caron and Doucet 2012; Mollica and

Tardella 2017). The hyperparameters a_γ and b_γ control the prior distribution on the worth parameters. Since ω is invariant to multiplicative transformations, a_γ and b_γ are generally non-influential. Still, the ratios between worth parameters could become very large when one object is strongly preferred over another. Thus, (a_γ, b_γ) should be chosen to give some density to values near 0 to allow for such extreme ratios.

6.4 Bayesian Estimation

In this section, we develop a Gibbs sampler for Bayesian estimation of Rank-Clustered BTL models and provide simulations to demonstrate its performance in different regimes.

6.4.1 Gibbs Sampler

Equation 6.1 defines ω by the pair (ν, g). Thus, to estimate ω, we sample from the joint posterior distribution of (ν, g). We do so using a reversible jump Markov chain Monte Carlo (RJMCMC) Gibbs sampler that alternates between updating g and ν via their full conditionals after data augmentation. The sampler is summarized in Algorithm 4.

Algorithm 4 Gibbs sampler for Rank-Clustered Bradley-Terry-Luce models

1. Initialize $g^{(0)}, \nu^{(0)}$ at random, ensuring that $|\nu^{(0)}| = K_{g^{(0)}}$.

2. For $t = 1, 2, \ldots, T_1$,

   (a) Sample $g^{(t)}$ via its full conditional using RJMCMC in order to traverse the space of partitions of varying numbers of clusters.

   (b) Sample $\nu^{(t)}$ via its full conditional T_2 times, which is possible via closed-form Gibbs sampling with data augmentation.

Based on our experience fitting Rank-Clustered BTL models to real and simulated data, we recommend initializing $g^{(0)} = \{1, 2, \ldots, J\}$ (and thus $K_{g^{(0)}} = J$) as it allows rank-clusters

to be formed during the estimation process (as opposed to being imposed by the analyst during initialization). For Step 2, T_1 should be sufficiently large to allow for convergence of the MCMC chain, although specific choices are context-dependent. Step 2(a) performs RJMCMC on clusters of objects. Since RJMCMC can be slow to converge in high dimensions, it is important to run multiple chains and assess for mixing and convergence (Gelman et al. 2013). Step 2(b) relies on a closed-form Gibbs sampler. We find $T_2 \leq 5$ is usually sufficient for posterior sampling.

Details of Step 2(a)

We now detail Step 2(a), which proposes a new partition g' based on the current partition g. Since (g, ν) are intricately tied, ν must simultaneously be updated to an appropriate ν'. The sampling of discrete partitions is challenging to perform efficiently. In a seminal paper on RJMCMC, Green (1995) provided a method for sampling partitions. We adapt that work for the Rank-Clustered BTL model.

Following Green (1995), we only propose g' which are slight modifications of g: Precisely, we allow only for 'births' splitting one cluster into two, or 'deaths' merging two clusters into one. Since all partitions have positive probability, this process is irreducible, as required. There is no need to propose g' that shuffle the partitions but maintain the number of clusters, as these partitions may be obtained by successive birth and death moves.

Births are attempted with probability $b_g = 0.5$.[2] In this case, we select a cluster k at random among those with at least two objects. The cluster is split "binomially", meaning that each object is placed independently into one of the "child" subgroups, k_1 or k_2, with equal probability, conditional on each subgroup ultimately containing at least one object. Deaths are attempted with probability $d_g = 1 - b_g = 0.5$. In a death, two adjacent clusters are merged at random. Adjacency means that $\nexists k : \nu_k \in (\nu_{k_1}, \nu_{k_2})$.

Births and deaths require updating ν by increasing or decreasing its dimension by 1,

[2]One could specify an alternative $b_g \in (0, 1)$ or make b_g a function of K_g (as in Green (1995)). For simplicity, we fix $b_g = 0.5$.

respectively. In a birth, we split a cluster's worth ν_k into (ν'_{k_1}, ν'_{k_2}) using,

$$\nu'_{k_1} = u\nu_k, \quad \nu'_{k_2} = u^{-1}\nu_k, \tag{6.8}$$

where $u \sim \text{Unif}(0.5, 1.5)$. The corresponding death solves these equations simultaneously:

$$\nu_k = \sqrt{\nu'_{k_1}\nu'_{k_2}}. \tag{6.9}$$

For reversibility, we automatically reject proposed births where ν'_{k_1}, ν'_{k_2} are not adjacent.

Per Green (1995), the Metropolis-Hastings probabilities for a birth and death, respectively, are $\min(1, A)$ and $\min(1, A^{-1})$, where

$$A = \frac{P(\nu', g'|\Pi)}{P(\nu, g|\Pi)} \times \frac{q(\nu, g|\nu', g')}{q(\nu', g'|\nu, g)P(u)} \times \left|\frac{\partial(\nu'_{k_1}, \nu'_{k_2})}{\partial(u, \nu_k)}\right|, \tag{6.10}$$

where $q(\nu', g'|\nu, g)$ is the transition probability of sampling (ν', g') given current parameter set (ν, g). We now calculate each term in A. First,

$$\begin{aligned}
\frac{P(\nu', g'|\Pi)}{P(\nu, g|\Pi)} &= \frac{P(\Pi|\nu', g')P[\nu'|g']P[g']}{\sum_{g''}\int_{\nu''} P(\Pi|\nu'', g'')P[\nu''|g'']d\nu''P[g'']} \frac{\sum_{g''}\int_{\nu''} P(\Pi|\nu'', g'')P[\nu''|g'']d\nu''P[g'']}{P(\Pi|\nu, g)P[\nu|g]P[g]} \\
&= \frac{P(\Pi|\nu', g')P[\nu'|g']P[g']}{P(\Pi|\nu, g)P[\nu|g]P[g]} \\
&= \frac{P(\Pi|\nu', g')}{P(\Pi|\nu, g)} \times \frac{\text{Gamma}(\nu'_{k_1}|a_\gamma, b_\gamma)\text{Gamma}(\nu'_{k_2}|a_\gamma, b_\gamma)}{\text{Gamma}(\nu_k|a_\gamma, b_\gamma)} \times \frac{P[g']}{P[g]},
\end{aligned} \tag{6.11}$$

where $P(\Pi|\nu, g)$ and $P[g]$ are defined by Equation 6.5. Second,

$$\begin{aligned}
\frac{q(\nu, g|\nu', g')}{q(\nu', g'|\nu, g)P(u)} &= \frac{d_{g'} \times \frac{1}{K_{g'}-1}}{\left(b_g \times \frac{1}{\#\{l:S_l(g)\geq 2\}} \times \frac{2}{2^{S_g(k)}-2}\right)\left(\frac{1}{1.5-0.5}\right)} \\
&= \frac{d_{g'}\#\{l : S_g(l) \geq 2\}(2^{S_g(k)-1} - 1)}{b_g(K_{g'} - 1)}
\end{aligned} \tag{6.12}$$

The numerator in Equation 6.12 is the death probability, $d_{g'}$, times the probability of selecting a pair of adjacent partitions given $K_{g'}$ total partitions after a split (there are $K_{g'} - 1$ such pairs). The denominator is the birth probability, b_g, times the probability of selecting a specific cluster k among those with at least two members. This term also includes the

probability of dividing the $S_g(k)$ objects in cluster k into two non-empty subsets. There are $(2^{S_g(k)} - 2)/2$ such subsets, since there are $2^{S_g(k)}$ total possible partitions, two empty partitions, and two ways to obtain each two-way split. Third and last,

$$\begin{aligned}\left|\frac{\partial(\nu'_{k_1}, \nu'_{k_2})}{\partial(u, \nu_k)}\right| &= \left| \begin{bmatrix} \frac{\partial}{\partial u}\nu'_{k_1} & \frac{\partial}{\partial \nu_k}\nu'_{k_1} \\ \frac{\partial}{\partial u}\nu'_{k_2} & \frac{\partial}{\partial \nu_k}\nu'_{k_2} \end{bmatrix} \right| = \left| \begin{bmatrix} \frac{\partial}{\partial u}u\nu_k & \frac{\partial}{\partial \nu_k}u\nu_k \\ \frac{\partial}{\partial u}\nu_k/u & \frac{\partial}{\partial \nu_k}\nu_k/u \end{bmatrix} \right| = \left| \begin{bmatrix} \nu_k & u \\ -\nu_k/u^2 & 1/u \end{bmatrix} \right| \\ &= \frac{2\nu_k}{u}. \end{aligned} \tag{6.13}$$

Details of Step 2(b)

To update ν conditional on a partition g and our data, Π, we turn to a clever trick for Bayesian estimation of Plackett-Luce models proposed by Caron and Doucet (2012). Here, we adapt their trick to account for the more general BTL family of distributions and rank-clustering. Let $Y = \{Y_{ir}\}$ be a collection of independent random variables, $i = 1, \ldots, I$ and $r = 1, \ldots, R_i$, sampled according to

$$Y_{ir} \sim \text{Exponential}\Big(\sum_{j \in \mathcal{S}_i} \nu_{g^{-1}(j)} - \sum_{s=0}^{r-1} \nu_{g^{-1}(\pi_i(s))}\Big). \tag{6.14}$$

The exponential rates are precisely the denominator terms from BTL densities that are burdensome to calculate. Conditional on Y, the full conditional probability $P[\nu|Y, \Pi, g]$ is,

$$\begin{aligned} P[\nu|Y, \Pi, g] \propto& P[Y|\Pi, g, \nu]P[\Pi|g, \nu]P[g|\nu]P[\nu] \\ \propto& P[Y|\Pi, g, \nu]P[\Pi|g, \nu]P[\nu] \\ =& \prod_{i=1}^{I}\prod_{r=1}^{R_i}\Big(\sum_{j \in \mathcal{S}_i} \nu_{g^{-1}(j)} - \sum_{s=0}^{r-1} \nu_{g^{-1}(\pi_i(s))}\Big) e^{-y_{ir}\left(\sum_{j \in \mathcal{S}_i} \nu_{g^{-1}(j)} - \sum_{s=0}^{r-1} \nu_{g^{-1}(\pi_i(s))}\right)} \times \\ & \prod_{i=1}^{I}\prod_{r=1}^{R_i} \frac{\nu_{g^{-1}(\pi_i(r))}}{\sum_{j \in \mathcal{S}_i} \nu_{g^{-1}(j)} - \sum_{s=0}^{r-1} \nu_{g^{-1}(\pi_i(s))}} \times \prod_{k=1}^{K} \nu_k^{a_\gamma - 1} e^{-b_\gamma \nu_k} \\ =& \prod_{i=1}^{I}\prod_{r=1}^{R_i} \nu_{g^{-1}(\pi_i(r))} e^{-y_{ir}\left(\sum_{j \in \mathcal{S}_i} \nu_{g^{-1}(j)} - \sum_{s=0}^{r-1} \nu_{g^{-1}(\pi_i(s))}\right)} \times \prod_{k=1}^{K} \nu_k^{a_\gamma - 1} e^{-b_\gamma \nu_k} \end{aligned} \tag{6.15}$$

Given these cancellations, we notice a closed-form expression for the posterior:

$$
\begin{aligned}
P[\nu|Y,\Pi,g] &\propto \prod_{i=1}^{I}\prod_{k=1}^{K} \nu_k^{c_{ki}} e^{-\nu_k \sum_{r=1}^{R_i} y_{ir}\delta_{irk}} \times \prod_{k=1}^{K} \nu_k^{a_\gamma - 1} e^{-b_\gamma \nu_k} \\
&= \prod_{k=1}^{K} \nu_k^{a_\gamma + \sum_{i=1}^{I} c_{ki} - 1} \; e^{-\nu_k (b_\gamma + \sum_{i=1}^{I}\sum_{r=1}^{R_i} y_{ir}\delta_{irk})} \\
&\propto \prod_{k=1}^{K} \text{Gamma}\Big(\nu_k \;\Big|\; a_\gamma + \sum_{i=1}^{I} c_{ki}, b_\gamma + \sum_{i=1}^{I}\sum_{r=1}^{R_i} y_{ir}\delta_{irk}\Big) \qquad (6.16)
\end{aligned}
$$

where

$$c_{ki} = \Big|\big\{j : j \in \pi_i, g^{-1}(j) = k\big\}\Big| \qquad (6.17)$$

$$\delta_{irk} = \Big|\big\{j : j \in \mathcal{S}_i, j \notin \{\pi_i(1), \ldots, \pi_i(r-1)\}, g^{-1}(j) = k\big\}\Big|. \qquad (6.18)$$

Thus, we can sample ν from a closed-form Gamma distribution after augmentation of the conditioning data Π and random variable g with Y.

6.4.2 Numerical Simulation

We now demonstrate accurate estimation of worth parameters and rank-clusters via a Rank-Clustered BTL model in a numerical simulation. We assume there are $J = 12$ objects which form K=3, 6, 9, or 12 rank-clusters. When $K = J = 12$, every object is independent; there are only singleton rank-clusters. In the true worth parameter vector, ω_0, rank-clustered objects have identical values and successive rank-clusters are separated in value by a factor of 4 (see Table 6.1 for specific values). Fourfold increases create strong but not absolute separation between objects: For demonstration, in a pairwise tournament between an object with $\omega_1 = 1$ and $\omega_2 = 4$, the probability of selecting object 2 is,

$$P[2 \prec 1 | \omega_1 = 1, \omega_2 = 4] = \frac{\omega_2}{\omega_1 + \omega_2} = \frac{4}{4+1} = 0.8.$$

We also vary the Poisson hyperparameter on the number of rank-clusters, $\lambda \in \{4, 8, 12\}$, which encourages rank-clustering to different extents and allows us to measure robustness

of results when λ is somewhat misspecified. Finally, we vary the number of judges $I \in \{50, 200, 800\}$. For each combination of I, ω_0, and λ, we generate 10 independent datasets and fit a Rank-Clustered Bradley-Terry-Luce distribution to each. We set $a_\gamma = 5$ and $b_\gamma = 3$. We set $T_1 = 5,000$ and $T_2 = 4$ to obtain 25,000 posterior samples in each MCMC chain and burn-out the first 15,000. For identifiability, all posterior estimates of ω_0 are normalized *post-hoc* such that $\sum_j \omega_{0j} = 1$.

Setting:	ω_0
$K = 3$	{1,1,1,1, 4,4,4,4, 16,16,16,16}
$K = 6$	{1,1, 4,4, 16,16, 64,64, 256,256, 1024,1024}
$K = 9$	{1,1, 4, 16, 64,64, 256, 1024, 4096,4096, 16384, 65536}
$K = 12$	{1, 4, 16, 64, 256, 1024, 4096, 16384, 65536, 262144, 1048576, 4194304}

Table 6.1: Simulation settings for ω_0 under varying numbers of true rank-clusters, K.

We first examine the accuracy of estimation for ω_0 across our distinct simulation settings. Figure 6.3 displays boxplots of mean absolute error (MAE) for ω_0 by number of judges I, true number of rank-clusters K, and the choice of hyperparameter λ. In general, estimation is quite accurate. We see that for any specific combination of K and λ, MAE decreases as I increases. Estimation error is higher when K is large and I is small, most likely the result of error estimating a complex rank-clustering structure.

Figure 6.4 displays the mean posterior probability of rank-clustering across object pairs which are truly rank-clustered (blue) or independent (gold) in ω_0. Results are further separated by the number of judges, I, true number of clusters, K, and hyperparameter λ. Overall, the model correctly identifies rank-clustered and independent object pairs. For rank-clustered pairs, accuracy increases as the number of judges I increases. Accuracy is generally best when hyperparameter $\lambda \approx K$, which occurs when prior belief regarding the number of clusters is approximately correct. The posterior probability of rank-clustering independent object pairs is near 0 in all simulations, indicating excellent accuracy.

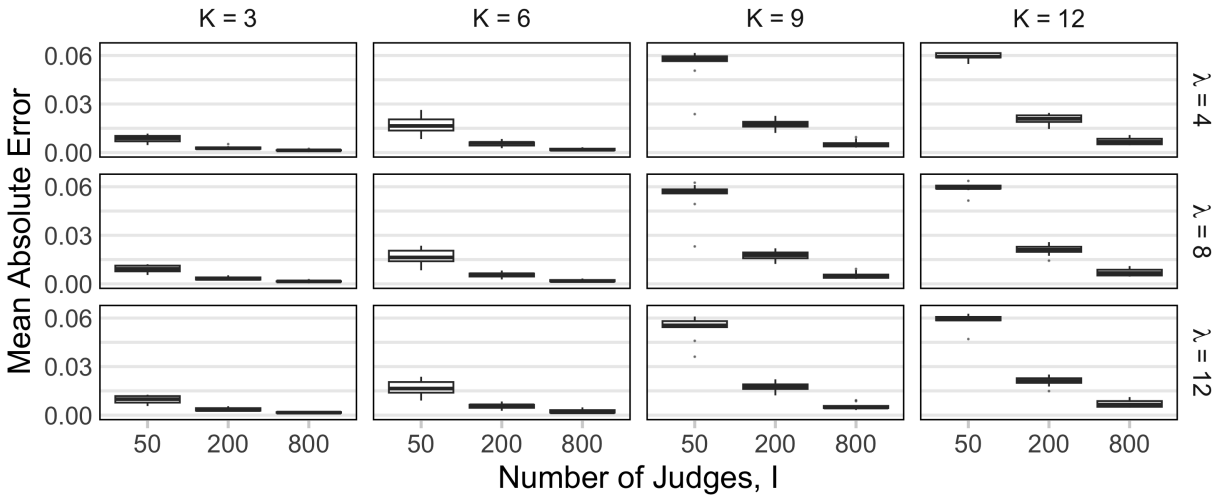

Figure 6.3: Boxplots of mean absolute error for ω_0 across combinations of the number of judges I, true number of rank-clusters K, and hyperparameter λ.

Figure 6.4: Boxplots of the mean posterior probability of rank-clustering object pairs which are truly rank-clustered (blue) or independent (gold), across combinations of I, K, and λ.

6.5 Application: 2021 Minneapolis Mayoral Election

We now analyze real voting data from the 2021 mayoral election in Minneapolis, Minnesota (Minneapolis Elections and Voter Services 2021). This election included 17 candidates (excluding write-ins and one who received no votes) and asked voters to rank their top-3 choices, in order. This mayoral election was the first after the 2020 murder of George Floyd in Minneapolis, leading to a contentious campaign. We want to estimate the overall preferences of Minneapolis voters regarding mayoral candidates and learn which candidates, if any, are rank-clustered at the population level. We expect there may be some rank-clustering based on the political ideologies and experiences of the candidates.

This dataset is well-suited to be studied by a Rank-Clustered BTL. Local elections in Minneapolis use ranked choice voting, which provides ample information for modeling relative preferences among the candidates. Clustering candidates may be of interest to political scientists or local political organizations for the purpose of understanding voter preferences (Gunther and Diamond 2003; Dimock et al. 2014).

6.5.1 Exploratory Analysis

A total of 145,337 votes were cast in this election. For our analysis, we randomly sample 1000 valid votes for modeling, which we treat as a sample of preferences from the population of Minneapolis voters. Figure 6.5 displays stacked bar charts of the sampled votes by rank level for each candidate. Candidates are ordered by their final placement according to the official ranked choice voting algorithm. The incumbent, Jacob Frey, receives the largest share of first place votes, although Kate Knuth and Sheila Nezhad also receive substantial support. The remaining candidates receive comparatively few votes. Most candidates are associated with the Democratic-Farmer-Labor (DFL) party, which is affiliated with the national Democratic Party. Laverne Turner and Bob "Again" Carney Jr. are the only Republicans (GOP) in the race. The remaining candidates represent Grassroots–Legalize Cannabis (GLC), Libertarian (LIB), Socialist Workers Party (SWP), For the People Party (FPP), Independence (INC),

Independent (IND), and Humanitarian–Community Party (HCP).

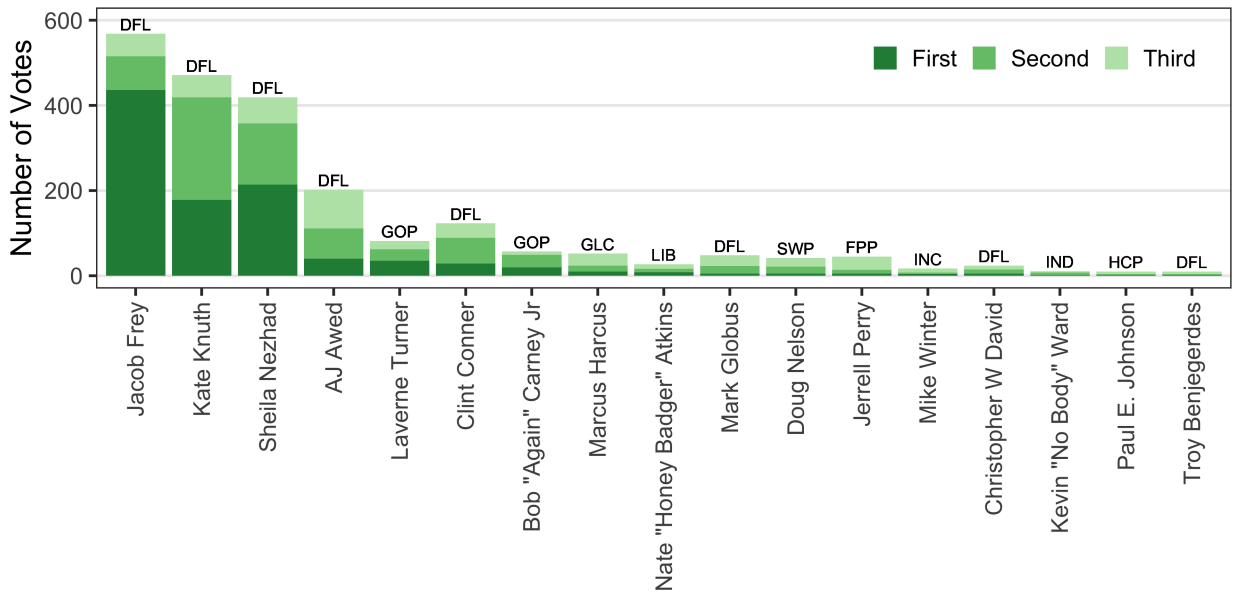

Figure 6.5: Number of votes by rank level and candidate. Candidates are ordered by their position in the official ranked choice election. Acronyms on the tops of bars represent each candidate's political party.

6.5.2 Results

We now fit a Rank-Clustered BTL to the 2021 Minneapolis mayoral election data. We set $a_\gamma = 5$, $b_\gamma = 3$, and $\lambda = 2$ to encourage a small number of rank-clusters.[3] We run 10 independent chains, each with $T_1 = 4000$ and $T_2 = 4$, for a total of 20,000 iterations per chain. After removing the first half of each chain as burn-in, we merge the chains and view results on the combined posterior samples.[4]

[3]See Appendix C for a visualization of the prior with respect to partitions.

[4]We also estimate models under alternative choices of λ to assess robustness. We find largely similar results. See Appendix C for details.

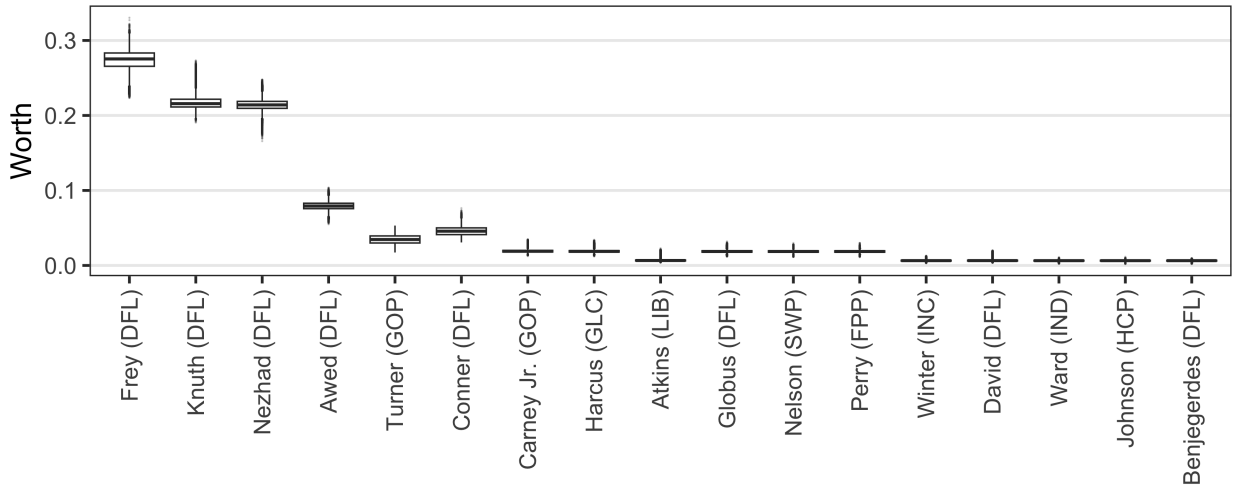

Figure 6.6: Posterior boxplots of normalized worth, ω, by candidate. Results marginalize out the estimated clustering structure.

Figure 6.6 displays posterior distributions of worth parameters, ω, by candidate. Each posterior sample of ω has been normalized to sum to 1 for identifiability and ease of interpretation. The candidates are again ordered by their position in the official ranked choice election. We see that Jacob Frey does not have majority support, but represents a plurality of voters with an estimated 28% of voters selecting him as their first choice according to the model. We also notice that adjacent candidates according to the ranked choice election do not necessarily have posterior worth distributions which follow that order. These changes in candidate order from the ranked choice election may occur due to the imposed rank-clustering structure of the model and the BTL model's consideration of all voting data (as opposed to the ranked choice voting algorithm, which considers only first-place votes until a candidate is "knocked out"). For example, Turner is ranked worse than Conner based on posterior worth, despite ranking better in the official election. However, Turner received far fewer overall votes than Conner, despite receiving slightly more first-place votes.

We now look at results from the perspective of estimated rank-clustering. Figure 6.7 displays a matrix of posterior probabilities that each pair of candidates is rank-clustered. In the matrix, the color of the $(m, n)^{\text{th}}$ entry corresponds to the probability that candidates m and n are rank-clustered. Candidates are ordered by their estimated rank in the Rank-Clustered BTL model, which is defined according to the order of posterior median estimates of worth, ω. The order of candidates according to the model does not match that of the official ranked choice voting algorithm. Cluster 1 consists of Jacob Frey, the winner and incumbent. Cluster 2 consists of Kate Knuth and Sheila Nezhad, both non-incumbent DFL candidates with substantial political backing and fundraising. There is some evidence that Knuth could be rank-clustered with Frey, which aligns with the ranked choice election results in which Knuth advanced to the final round and Nezhad was knocked out. Clusters 3 and 4 consist of AJ Awed and Clint Conner, respectively, who are each DFL candidates with less experience and political backing than candidates in previous rank-clusters. Cluster 5 consists of Laverne Turner, a GOP candidate, and Cluster 6 consists of Carney Jr., Harcus, Globus, and Nelson. There is some evidence that Turner should be merged into Cluster 6, and even stronger evidence for merging her with Carney Jr. and Harcus. This is notable because Turner and Carney Jr. are the only Republican candidates, and Harcus represents the Grassroots–Legalize Cannabis Party which shares many ideological viewpoints with the GOP. Last, Cluster 7 consists of 6 candidates with minimal support and liberal/independent ideologies.

Figure 6.8 compares point estimates of rank for each candidate among four methods. The first and second rows display assigned ranks from ranked choice and "first-past-the-post" (FPP) election procedures, respectively.[5] To calculate FPP ranks, we order the candidates by the number of first place votes he/she received (ignoring all second and third place votes). If an actual election of this kind had occurred, the results may be different based on the differing voter strategies encouraged by ranked choice and FPP elections. The third and

[5]Since ranked choice and FPP elections use deterministic procedures, we use the term "assigned rank".

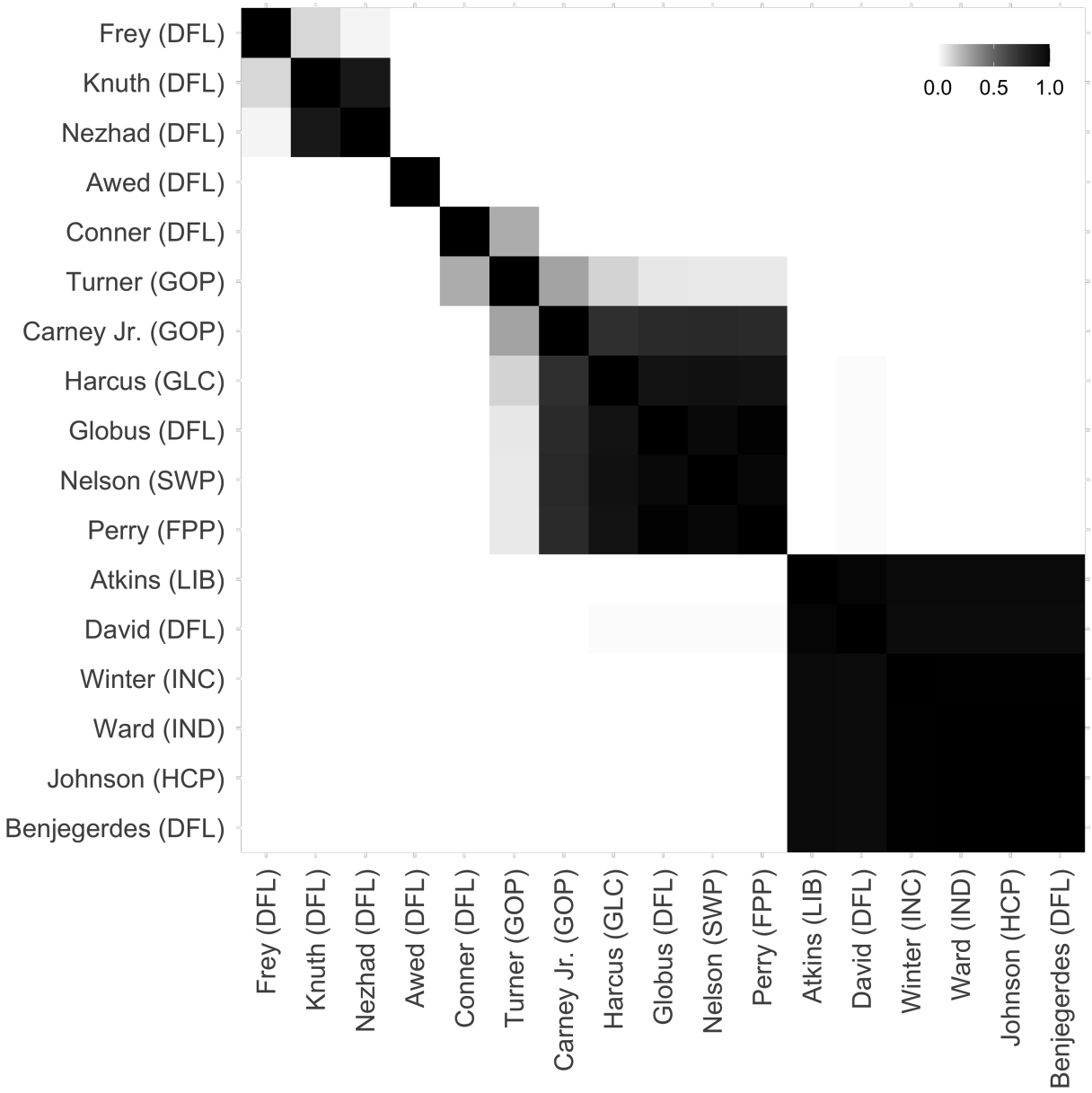

Figure 6.7: Clustering matrix showing the posterior probability that each pair of candidates is clustered. Candidates are ordered by their estimated rank according to the Rank-Clustered BTL model.

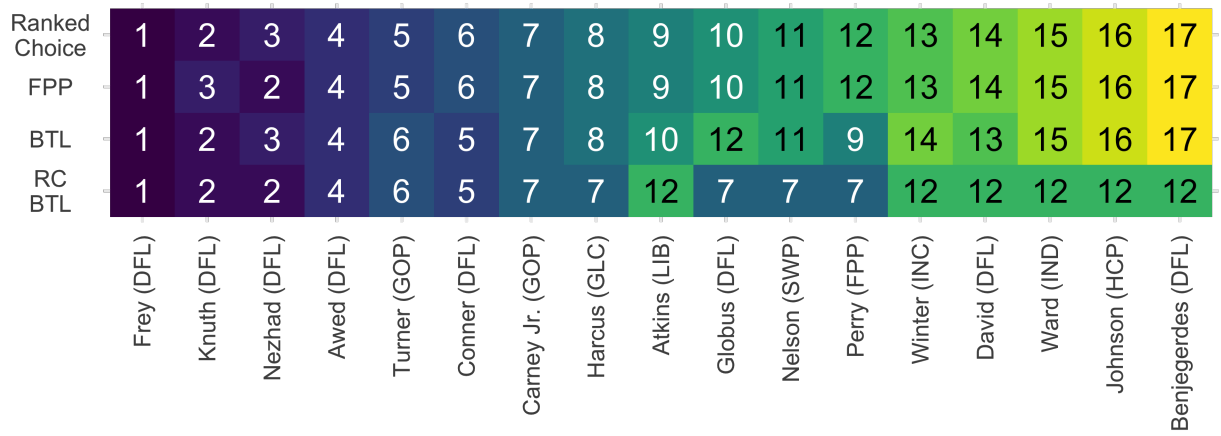

Figure 6.8: Comparison of estimated rank for each candidate across four aggregation methods: Ranked Choice, First-Past-the-Post (FPP), BTL, and Rank-Clustered BTL (RC BTL). Candidates are ordered by their rank in the actual ranked choice election.

fourth rows display estimated ranks from a standard Bayesian BTL and our Rank-Clustered BTL, respectively. Estimated ranks are based on maximum *a posteriori* estimates. Frey wins the election in all methods. The two deterministic algorithms, Ranked Choice and FPP, switch the second and third place candidates but otherwise provide identical results. The BTL and Rank-Clustered BTL models roughly reflect the deterministic algorithms, although we notice some swaps in candidate ranks which may be attributed to differences between first place and second or third place votes. For example, Conner received fewer first place votes than Turner, but far more second and third place votes. As a result, deterministic algorithms rank Turner above Conner, while the BTL model takes into account the additional preference information and ranks Conner above Turner. The Rank-Clustered BTL additionally clusters candidates with similar levels of support. Most notably, Knuth and Nezhad are rank-clustered. Additionally, the eccentric third-party candidate Nate "Honey Badger" Atkins is estimated to be in the last place rank-cluster despite a modest number of

first place votes. This may be attributed to his limited number of second and third place votes, suggesting that overall he is less preferred than Globus, Nelson, and Perry (who were ranked worse than Atkins in the deterministic methods).

6.6 Discussion

In this chapter, we proposed the Rank-Clustered Bradley-Terry-Luce model for estimating an overall ranking with rank-clusters. The model employs the Bradley-Terry-Luce (BTL) family of distributions for ordinal data and estimates model parameters in a Bayesian framework under the proposed Partition-based Spike-and-Slab (PSSF) prior. In a simulation study, we demonstrated the model's ability to accurately estimate an overall ranking of objects that rank-clusters objects that are identical in quality, and successfully distinguishes objects that are not. Furthermore, we showed that estimation error decreases as the number of data points increases. When applied to voting data from the 2021 Minneapolis mayoral election, the model estimated the overall preferences of voters while simultaneously estimating rank-clusters of candidates. Thus, the model is a useful tool for understanding preferences via rankings.

To our knowledge, the PSSF prior is only the second spike-and-slab based prior for parameter fusion, after Wu et al. (2021), whose prior was developed for regression and requires a known parameter order. Visual inspection of the prior distribution makes obvious its connection to spike-and-slab: "spike" components correspond to parameter clusters and "slab" components correspond to independent parameters. Estimation of parameters under this model requires reversible jump MCMC. We proposed a computationally efficient Gibbs sampler for estimation based on the seminal work of Green (1995).

A useful benefit of estimating parameter values and clusters in a single Bayesian framework is the avoidance of *selective inference* (Taylor and Tibshirani 2015) or more colloquially, *double dipping* (Kriegeskorte et al. 2009). Selective inference occurs in frequentist analyses when the same data is used twice in the process of model selection and/or estimation, e.g., to estimate some latent structure underlying the data and subsequently to estimate parameters

conditional on that estimated structure. In our context, selective inference would occur if ordinal preference data was used first to identify rank-clusters and then used again to estimate worth parameter values conditional on those clusters. Selective inference often leads to invalid inference in part because uncertainty regarding the estimated clustering structure is not taken into account. However, Rank-Clustered Bradley-Terry-Luce models do not perform selective inference because parameter values and rank-clusters are estimated simultaneously. As such, our parameter estimates incorporate uncertainty across the posterior distributions of both the rank-clustering structure and the specific parameter values.

Results from Rank-Clustered BTL models are useful in a variety of contexts. As noted in other fusion literatures on rankings, estimated overall rankings may be easier to understand and interpret when rank-clusters of objects are identified, as rank-clusters lead to fewer rank levels of objects to distinguish (Masarotto and Varin 2012). In contexts where model results are used for prediction, such as in sports, estimating rank-clusters could improve predictive accuracy (Tutz and Schauberger 2015). Similarly, estimating rank-clusters is important in the context of decision-making: In peer review, for example, rank-clusters can be beneficial for communicating uncertainty in the assessment of preferences and for better transparency in funding decisions. We might imagine a scenario where a government agency is only able to fund two grants, however, two grant proposals are rank-clustered in second place. In this case, rank-clustering can be used to communicate uncertainty in the relative quality of the top proposals. A potential danger is that under this uncertainty, decision makers may be tempted to resort to unfair tie-breaking methods, e.g., selecting the proposal with the most famous author. Instead, tie-breaking should occur based on a fairer or more principled method, such as a partial lottery (Fang and Casadevall 2016; Roumbanis 2019; Heyard et al. 2022).

We list a few possible directions for future research. In this chapter, the model was only used to analyze partial and complete rankings. Analysis of other forms of ordinal preference data allowed by the BTL distribution, such as pairwise comparisons or rankings made under separate ballots, require further study, including a comparison of rank-clustering accuracy

with existing frequentist methods available for analyzing pairwise comparison data (e.g., Masarotto and Varin (2012)). In the presence of such data, it will also be important to study how estimation is impacted based on the level of interconnectedness in the objects assessed by the judges (e.g., if separate groups of judges assess completely distinct sets of objects). We expect the Rank-Clustered BTL model to perform well on pairwise comparison data based on our analysis of top-3 voter preferences in Section 6.5. However, simulation studies and applications to real data are needed to confirm these expectations.

Moreover, the PSSF prior could be applied in other settings. In the field of preference learning, the PSSF prior could be applied to BTL distributions that incorporate covariates (e.g., Gormley and Murphy (2010); Chapman and Staelin (1982)). In that case, the prior could require modification to allow for covariate parameter estimation in addition to rank-clustering. Also, the PSSF prior may be used in regression for variable fusion, and its performance may be compared to other existing Bayesian variable fusion methods (e.g., Casella et al. (2010); Song and Cheng (2020); Shimamura et al. (2019)).

Chapter 7

DISCUSSION

7.1 Contributions

In this dissertation, we have provided an overview of deterministic and statistical methods for the analysis of preference data and proposed three new methods for preference analysis driven by applications in the social sciences. We expound our primary contributions below.

In Chapter 3, we proposed the first joint statistical model for rankings and ratings without data conversion, the Mallows-Binomial. Mallows-Binomial uses shared parameters between ranking and rating components to estimate the preferences of a group or population regarding a collection of objects. The model was proposed in a frequentist framework and estimated via a computationally efficient A* search algorithm. In Chapter 4, we proposed a second joint statistical model for rankings and ratings, the Bradley-Terry-Luce-Binomial (BTL-Binomial). BTL-Binomial allows for additional types of ordinal preference data beyond standard rankings, such as pairwise comparisons and rankings made under separate ballots. Furthermore, BTL-Binomial was proposed in a Bayesian Mixture of Finite Mixtures framework, which allows for the estimation of both the number of heterogeneous preference ideologies in a population and the specific preference ideologies of each group. Chapter 5 compared and contrasted the various assumptions, properties, and practical considerations of using each model, providing insights on model selection in the presence of both ordinal and cardinal preference data. Through simulation studies, toy examples, and real data analyses in Chapters 3-5, both the Mallows-Binomial and BTL-Binomial were demonstrated to sensibly combine the distinct and complementary preference information provided by rankings and ratings. Furthermore, we analyzed datasets to demonstrate the models' ability to account for missing data, estimate potentially heterogeneous preferences, and aid accurate decision-

making in a number of contexts, including large-scale distributed peer review, small-scale grant panel review (with and without heterogeneity), and survey preference data.

Chapter 6 proposed a novel methodology for analyzing ordinal preference data with rank-clusters. In traditional ranking data analyses, an overall ranking of objects is estimated based on the observed data. However, if some objects are thought to be equal or indistinguishable in quality, they may be more correctly considered rank-clustered, i.e., identical in population-level rank. Existing rank-clustering methods rely on frequentist penalty-based techniques and primarily allow for analyzing only pairwise comparison data, thus limiting their use. To address this problem, we proposed the Rank-Clustered BTL model, which applies a cluster-inducing prior to the BTL family of ranking distributions. Our Bayesian model has interpretable prior parameters, is efficiently estimated via Gibbs sampling, and allows for the estimation of uncertainty in a unified approach. We applied the model to a sample of voting data from the 2021 Minneapolis mayoral election and demonstrated its ability to cluster similarly-preferred candidates along interpretable dimensions.

7.2 Discussion and Future Work

The methodologies and data analyses presented in this dissertation relate to a variety of existing literatures in the natural sciences, social sciences, and humanities. At their root, each of our proposed methods aggregates observed preference data into summary statistic(s) of preferences. This task is central to the philosophical literature on Social Choice Theory (described in Chapter 2.3). Social Choice theorists have been describing desirable properties of social choice and proposing methods to satisfy them since the Marquis de Condorcet first noticed the existence of cyclic preferences in the late 18th century (de Condorcet 1785). In this work, we have noted qualities satisfied by our methods, such as the Independence from Irrelevant Alternatives (IIA) criterion satisfied by the BTL family of ranking distributions (but not the Mallows), and the non-dictatorship and unrestricted domain satisfied by all our methods. Still, additional research is needed to connect statistical preference models, including, but not limited to, the models proposed in this dissertation with axioms of social

choice. There is some modern literature connecting deterministic preference aggregation algorithms to social choice axioms (Hammond 2007; Mixon Jr and King 2012; Medcalfe 2018; Boudreau et al. 2018; Nguyen et al. 2021), but only limited research connects Social Choice axioms to statistical preference models (e.g., Marden (1996); Nagaraja and Sanders (2020); Sanders et al. (2022)).

Chapters 4 and 6 draw on methods from model-based clustering. Model-based clustering refers to the statistical estimation of clusters using a probability model, as opposed to deterministic algorithms or methods based on heuristics (Bouveyron et al. 2019). We have performed two distinct kinds of clustering: First, the proposed BTL-Binomial Mixture of Finite Mixtures (MFM) model clusters *judges* into heterogeneous preference ideologies based on their observed preference data. The MFM approach extends traditional latent class analyses (Lazarsfeld 1950), in which the number of clusters is selected via an ad-hoc heuristic or goodness-of-fit statistic and then estimates cluster labels and parameters using traditional statistical techniques like maximum likelihood (Sinha et al. 2021). Instead, MFM is a Bayesian framework that treats both the number of latent classes and the associated class-specific parameters as random variables, and estimates both simultaneously (Miller and Harrison 2018). Second, the Rank-Clustered BTL clusters *objects* in their rank based on their overall quality or support in the population of judges. We emphasize that the Rank-Clustered BTL model does not cluster judges, but instead clusters model parameters that represent the quality of each object. Parameters are considered clustered when their values are made identical, which results in the formation of a rank-cluster among their corresponding objects. Our rank-clustering technique draws on the statistical model selection literature, such as the fused lasso (Tibshirani 1996) and spike-and-slab priors (Mitchell and Beauchamp 1988). Here, we propose the Partition-based Spike-and-Slab Fusion (PSSF) prior that induces parameter clustering, and hence, rank-clustering when applied to the BTL distribution. Although our Rank-Clustered BTL model was shown to successfully estimate rank-clusters and independent objects on both real and simulated data, additional research is required. Theoretical properties of the model and prior, such as asymptotic convergence

rates and consistency, should be investigated. Comparisons of results should also be made with existing frequentist methods for rank-clustering. Last, we note that all models proposed in this dissertation could be extended to allow for co-clustering. Co-clustering refers to the simultaneous clustering of observations and variables (Hartigan 1972), which in our preference context would mean clustering both judges and objects. Co-clustering could be accomplished by applying a modified PSSF prior alongside a Bayesian latent class model.

We now turn to future work regarding our joint models for rankings and ratings. First, each model uses the Binomial distribution to model ratings. However, the Binomial distribution imposes certain assumptions on the data. That is, ratings must arise from a discrete, finite, and equally-space set, and follow a specific mean-variance relationship. The rating component of each model could be replaced with other distributions. For example, a Beta-Binomial distribution when the Binomial mean-variance relationship does not hold or a Truncated Normal for continuous ratings. Second, the relative weights of ratings and rankings in our joint models could be directly specified by practitioners or estimated. Currently, no direct weighting occurs, although data missingness and data types may impact the influence of each model component during estimation. For example, more granular rating scales imply that ratings are more precise and contain more information. However, this precision may not be meaningful in practice (Miller 1956; Jones and Loe 2013), and practitioners may wish to down weight ratings during estimation as a result. Still, how to impose, interpret, and modify weights between ratings and rankings is unclear. Future research could formulate methods for direct weighting of rating and ranking model components in the presence of distinct levels of missing data and the available data formats. Third, both models could be modified to incorporate covariates, perhaps with respect to judge or object characteristics. Incorporating covariates may be useful for hypothesis testing. In the context of peer review, object-level covariates would allow for testing if significant disparities exist in the perceived quality of proposals based on the proposer's gender or race. Fourth, the Mallows-Binomial model could be estimated in a Bayesian framework and under preference heterogeneity via a latent class model. A Bayesian approach would reduce the substantial burden of frequentist

estimation, an NP-hard problem that is especially difficult when the number of objects is large.

Mallows-Binomial and BTL-Binomial joint models for rankings and ratings may be useful for estimating preferences in social science disciplines beyond those presented in this dissertation. As discussed in Chapter 2.1.3, psychologists and psychometricians have long been proposing measurement approaches that draw on the distinct and complementary properties of ratings and rankings (e.g., Smith and Kendall (1963); Sung and Wu (2018)). Unfortunately, these methods require elaborate and time-consuming data collection mechanisms. In contrast, our joint models allow for a variety of standard ordinal and cardinal data types and have estimation tools available under multiple frameworks. Our models could allow for better understanding of preferences in a plethora of social science disciplines: Political scientists may obtain fine-grained estimates of political preferences among voters when asked to assess potential candidates with ratings and rankings (Campbell and Cowley 2014), sociologists could measure absolute and relative support for certain opinions or attitudes by asking study participants to rate and rank pairs of statements from a larger collection, educators may measure the effectiveness of teaching strategies by asking students to compare them using numerical scores and ordinal comparisons, and health scientists may transparently rank health systems across geographic regions using existing ordinal and cardinal data on patient outcomes (Remington et al. 2015; Schütte et al. 2018). These examples barely scratch the surface of settings where our joint models may provide more accurate summaries of preferences or quality assessments among members of a group or population.

Future work is needed on the Rank-Clustered BTL model as well. With potential modification to the PSSF prior, the Rank-Clustered BTL model may be extended to account for ties in observed preferences (e.g., a tie in pairwise comparison can be used to represent tied matches when modeling sports tournaments (Tutz and Schauberger 2015)) or covariates corresponding to the objects being assessed. The PSSF prior may also be considered in the context of model selection in regression, which is a more traditional home for spike-and-slab priors. The Rank-Clustered BTL model may also be useful in social science settings beyond

political preferences. For example, the model may be used to estimate an overall ranking of graduate programs where approximate ties in quality are stated, an overall ranking of hospitals based on patient outcomes that clusters hospitals into relative categories such as "better outcomes" or "worse outcomes", or an overall ranking of grant proposals that estimates if proposals near the funding line are truly different in quality or not.

Statistical preference analysis estimates and summarizes population-level preferences with uncertainty. This dissertation has proposed three novel methods for statistical preference analysis, driven by social science settings which previously had few or no principled methods. This work improves our understanding of statistical inference on preferences, which in turn aids informed decision-making and opens the door for future developments in the statistical analysis of preferences.

BIBLIOGRAPHY

Adkins, L. and M. Fligner (1998). A non-iterative procedure for maximum likelihood estimation of the parameters of Mallows' model based on partial rankings. *Communications in Statistics–Theory and Methods 27*(9), 2199–2220.

Ailon, N. (2010). Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica 57*(2), 284–300.

Ali, A. and M. Meilă (2012). Experiments with Kemeny ranking: What works when? *Mathematical Social Sciences 64*(1), 28–40.

Ali, A., T. B. Murphy, M. Meila, and H. Chen (2010). Preferences in college applications–a nonparametric Bayesian analysis of top-10 rankings. In *NIPS Workshop on Computational Social Science and the Wisdom of Crowds*.

Alwin, D. F. and J. A. Krosnick (1985). The measurement of values in surveys: A comparison of ratings and rankings. *Public Opinion Quarterly 49*(4), 535–552.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics 2*(6), 1152–1174.

Armstrong, J. S. (1997). Peer review for journals: Evidence on quality control, fairness, and innovation. *Science and Engineering Ethics 3*(1), 63–84.

Arrow, K. J. (1950). A difficulty in the concept of social welfare. *Journal of Political Economy 58*(4), 328–346.

Ayer, M., H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman (1955). An empirical dis-

tribution function for sampling with incomplete information. *The Annals of Mathematical Statistics 26*(4), 641–647.

Balog, K., Y. Fang, M. De Rijke, P. Serdyukov, and L. Si (2012). Expertise retrieval. *Foundations and Trends in Information Retrieval 6*(2–3), 127–256.

Barlow, R. E. and H. D. Brunk (1972). The isotonic regression problem and its dual. *Journal of the American Statistical Association 67*(337), 140–147.

Baumgartner, H. and J.-B. E. Steenkamp (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research 38*(2), 143–156.

Belkin, N. J., P. Kantor, E. A. Fox, and J. A. Shaw (1995). Combining the evidence of multiple query representations for information retrieval. *Information Processing & Management 31*(3), 431–448.

Bhamidipati, N. L. and S. K. Pal (2008). Comparing scores intended for ranking. *IEEE Transactions on Knowledge and Data Engineering 21*(1), 21–34.

Bhattacharya, A., D. Pati, N. S. Pillai, and D. B. Dunson (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association 110*(512), 1479–1490.

Bickel, P. J. and D. A. Freedman (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics 9*(6), 1196–1217.

Biernat, M. (1995). The shifting standards model: Implications of stereotype accuracy for social judgment. In *Stereotype Accuracy: Toward Appreciating Group Differences.*, pp. 87–114. American Psychological Association.

Biernat, M. (2003). Toward a broader view of social stereotyping. *American Psychologist 58*(12), 1019.

Biernat, M., E. C. Collins, I. Katzarska-Miller, and E. R. Thompson (2009). Race-based shifting standards and racial discrimination. *Personality and Social Psychology Bulletin 35*(1), 16–28.

Biernat, M. and D. Kobrynowicz (1997). Gender-and race-based standards of competence: Lower minimum standards but higher ability standards for devalued groups. *Journal of Personality and Social Psychology 72*(3), 544.

Biernat, M. and T. K. Vescio (2002). She swings, she hits, she's great, she's benched: Implications of gender-based shifting standards for judgment and behavior. *Personality and Social Psychology Bulletin 28*(1), 66–77.

Biernat, M., T. K. Vescio, and M. Manis (1998). Judging and behaving toward members of stereotyped groups: A shifting standards perspective. In C. Sedikides, J. Schopler, and C. A. Insko (Eds.), *Intergroup Cognition and Intergroup Behavior*, pp. 151–175. Hillsdale, NJ: Erlbaum.

Bornmann, L., R. Mutz, and H.-D. Daniel (2007). Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics 1*(3), 226–238.

Boudreau, J., J. Ehrlich, M. F. Raza, and S. Sanders (2018). The likelihood of social choice violations in rank sum scoring: Algorithms and evidence from NCAA cross country running. *Public Choice 174*(3), 219–238.

Boudreau, J. W. and S. Sanders (2015). Choosing "flawed" aggregation rules: The benefit of social choice violations in a league that values competitive balance. *Economics Letters 137*, 106–108.

Bouveyron, C., G. Celeux, T. B. Murphy, and A. E. Raftery (2019). *Model-based Clustering: Basic Ideas*, pp. 15–78. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Bradley, R. A. and M. E. Terry (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika 39*(3/4), 324–345.

Brancotte, B., B. Yang, G. Blin, S. Cohen-Boulakia, A. Denise, and S. Hamel (2015, July). Rank aggregation with ties: Experiments and analysis. *Proceedings of the VLDB Endowment (PVLDB) 8*(11), 1202–1213.

Breffle, W. S., E. R. Morey, and J. A. Thacher (2011). A joint latent-class model: Combining Likert-scale preference statements with choice data to harvest preference heterogeneity. *Environmental and Resource Economics 50*(1), 83–110.

Brezis, E. S. and A. Birukou (2020). Arbitrariness in the peer review process. *Scientometrics 123*(1), 393–411.

Busa-Fekete, R., D. Fotakis, B. Szorenyi, and E. Zampetakis (2021). Identity testing for Mallows model. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems*, Volume 34, pp. 23179–23190. Curran Associates, Inc.

Busse, L. M., P. Orbanz, and J. M. Buhmann (2007). Cluster analysis of heterogeneous rank data. In *Proceedings of the 24th International Conference on Machine Learning*, pp. 113–120.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing 16*(5), 1190–1208.

Campbell, R. and P. Cowley (2014). What voters want: Reactions to candidate characteristics in a survey experiment. *Political Studies 62*(4), 745–765.

Caron, F. and A. Doucet (2012). Efficient Bayesian inference for generalized Bradley–Terry models. *Journal of Computational and Graphical Statistics 21*(1), 174–196.

Caron, F., Y. W. Teh, T. B. Murphy, et al. (2014). Bayesian nonparametric Plackett–Luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics 8*(2), 1145–1181.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Casella, G., M. Ghosh, J. Gill, and M. Kyung (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis 5*(2), 369 – 411.

Cattelan, M., C. Varin, and D. Firth (2013). Dynamic Bradley–Terry modelling of sports tournaments. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 62*(1), 135–150.

Chapman, R. G. and R. Staelin (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research 19*(3), 288–301.

Cheng, W., E. Hüllermeier, and K. J. Dembczynski (2010). Label ranking methods based on the Plackett–Luce model. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 215–222.

Chierichetti, F., A. Dasgupta, S. Haddadan, R. Kumar, and S. Lattanzi (2018). Mallows models for top-k lists. *Advances in Neural Information Processing Systems 31*, 4382–4392.

Cohen, W. W., R. E. Schapire, and Y. Singer (1999). Learning to order things. *Journal of Artificial Intelligence Research 10*, 243–270.

Collas, F. and E. Irurozki (2021). Concentric mixtures of Mallows models for top-k rankings: sampling and identifiability. In *International Conference on Machine Learning*, pp. 2079–2088. PMLR.

Conitzer, V., A. Davenport, and J. Kalagnanam (2006). Improved bounds for computing Kemeny rankings. In *AAAI*, Volume 6, pp. 620–626.

Crispino, M. and I. Antoniano-Villalobos (2022). Informative priors for the consensus ranking in the Bayesian Mallows model. *Bayesian Analysis 1*(1), 1–24.

Critchlow, D. E., M. A. Fligner, and J. S. Verducci (1991). Probability models on rankings. *Journal of Mathematical Psychology 35*(3), 294–318.

Dahl, D. B., D. J. Johnson, and P. Müller (2021). *salso: Search Algorithms and Loss Functions for Bayesian Clustering*. R package version 0.3.0.

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.

de Borda, J.-C. (1781). Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences 12*.

de Chiusole, D. and L. Stefanutti (2011). Rating, ranking, or both? A joint application of two probabilistic models for the measurement of values. *TPM-Testing, Psychometrics, Methodology in Applied Psychology 18*(1), 49–60.

de Condorcet, N. (1785). *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix.* de l'Imprimerie Royale.

Dimock, M., C. Doherty, J. Kiley, and V. Krishnamurthy (2014). Beyond red vs. blue: The political typology. *Pew Research Center*.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics 7*(1), 1 – 26.

Efron, B. and R. J. Tibshirani (1994). *An Introduction to the Bootstrap*. CRC Press.

Egidi, L. and N. Torelli (2021). Comparing goal-based and result-based approaches in modelling football outcomes. *Social Indicators Research 156*(2), 801–813.

Emerson, J. W. and T. B. Arnold (2011). Statistical sleuthing by leveraging human nature: A study of Olympic figure skating. *The American Statistician 65*(3), 143–148.

Erosheva, E. A., S. Grant, M.-C. Chen, M. D. Lindner, R. K. Nakamura, and C. J. Lee (2020). NIH peer review: Criterion scores completely account for racial disparities in overall impact scores. *Science Advances 6*(23), eaaz4868.

Erosheva, E. A., P. Martinková, and C. J. Lee (2021). When zero may not be zero: A cautionary note on the use of inter-rater reliability in evaluating grant peer review. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 184*(3), 904–919.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Fagin, R. (2002). Combining fuzzy information: An overview. *ACM SIGMOD Record 31*(2), 109–118.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Fang, F. C. and A. Casadevall (2016). Research funding: The case for a modified lottery. *mBio 7*(2), e00422.

Feather, N. T. (1973). The measurement of values: Effects of different assessment procedures. *Australian Journal of Psychology 25*(3), 221–231.

Feigin, P. D. and A. Cohen (1978). On a model for concordance between judges. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 40*(2), 203–213.

Fligner, M. A. and J. S. Verducci (1986). Distance based ranking models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 48*(3), 359–369.

Fligner, M. A. and J. S. Verducci (1988). Multistage ranking models. *Journal of the American Statistical Association 83*(403), 892–901.

Frühwirth-Schnatter, S., G. Malsiner-Walli, and B. Grün (2021). Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis 16*(4), 1279–1307.

Fürnkranz, J. and E. Huellermeier (2010). *Preference learning*. Berlin & Heidelberg: Springer.

Gallo, S. (2020). Grant panel review data with rankings and scores. https://doi.org/10.6084/m9.figshare.14828916.v1.

Gallo, S. (2023). Grant peer review scoring ranking data A2. https://doi.org/10.6084/m9.figshare.19692223.v1.

Gallo, S. A., M. Pearce, C. J. Lee, and E. A. Erosheva (2023). A new approach to peer review assessments: Score, then rank. *Research Integrity and Peer Review*.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 339–373.

Ginther, D. K., J. Basner, U. Jensen, J. Schnell, R. Kington, and W. T. Schaffer (2018). Publications as predictors of racial and ethnic differences in NIH research awards. *PLoS One 13*(11), e0205929.

Goffin, R. D., R. B. Jelley, D. M. Powell, and N. G. Johnston (2009). Taking advantage of social comparisons in performance appraisal: The relative percentile method. *Human Resource Management 48*(2), 251–268.

Goffin, R. D. and J. M. Olson (2011). Is it all relative? Comparative judgments and the possible improvement of self-ratings and ratings of others. *Perspectives on Psychological Science 6*(1), 48–60.

Gormley, I. C. and T. B. Murphy (2006). Analysis of Irish third-level college applications data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 169*(2), 361–379.

Gormley, I. C. and T. B. Murphy (2008). Exploring voting blocs within the Irish electorate: A mixture modeling approach. *Journal of the American Statistical Association 103*(483), 1014–1027.

Gormley, I. C. and T. B. Murphy (2010). Clustering ranked preference data using sociodemographic covariates. In *Choice Modelling: The State-of-the-Art and the State-of-Practice*. Emerald Group Publishing Limited.

Gormley, I. C., T. B. Murphy, et al. (2009). A grade of membership model for rank data. *Bayesian Analysis 4*(2), 265–295.

Grant, J., S. Burden, and G. Breen (1997). No evidence of sexism in peer review. *Nature 390*(6659), 438–438.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*(4), 711–732.

Greve, J. (2021). *fipp: Induced Priors in Bayesian Mixture Models*. R package version 1.0.0.

Greve, J., B. Grün, G. Malsiner-Walli, and S. Frühwirth-Schnatter (2022). Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis. *Australian & New Zealand Journal of Statistics 64*(2), 205–229.

Griffin, D. and L. Brenner (2004). Perspectives on probability judgment calibration. *Blackwell Handbook of Judgment and Decision Making 199*, 158–177.

Griffin, J. and P. Brown (2005). Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick.

Griffin, J. E. and P. J. Brown (2010). Inference with Normal-Gamma prior distributions in regression problems. *Bayesian Analysis 5*(1), 171–188.

Guiver, J. and E. Snelson (2009). Bayesian inference for Plackett–Luce ranking models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 377–384.

Gunther, R. and L. Diamond (2003). Species of political parties: A new typology. *Party Politics 9*(2), 167–199.

Hall, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media.

Hammond, T. H. (2007). Rank injustice?: How the scoring method for cross-country running competitions violates major social choice principles. *Public Choice 133*(3), 359–375.

Hart, P. E., N. J. Nilsson, and B. Raphael (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics 4*(2), 100–107.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association 67*(337), 123–129.

He, J. (2022). A central limit theorem for descents of a Mallows permutation and its inverse. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, Volume 58, pp. 667–694. Institut Henri Poincaré.

Helmer, M., M. Schottdorf, A. Neef, and D. Battaglia (2017). Gender bias in scholarly peer review. *Elife 6*, e21718.

Heyard, R., M. Ott, G. Salanti, and M. Egger (2022). Rethinking the funding line at the Swiss national science foundation: Bayesian ranking and lottery. *Statistics and Public Policy 9*(1), 110–121.

Hochbaum, D. S. and E. Moreno-Centeno (2021). Joint aggregation of cardinal and ordinal evaluations with an application to a student paper competition. *arXiv preprint arXiv:2101.04765*.

Hsu, D. F. and I. Taksa (2005). Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval 8*(3), 449–480.

Hug, S. E. and M. Ochsner (2022). Do peers share the same criteria for assessing grant applications? *Research Evaluation 31*(1), 104–117.

Hunter, D. R. et al. (2004). MM algorithms for generalized Bradley–Terry models. *The Annals of Statistics 32*(1), 384–406.

Ishwaran, H. and J. S. Rao (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics 33*(2), 730–773.

Jara, A., M. J. Garcia-Zattera, and E. Lesaffre (2007). A Dirichlet process mixture model for the analysis of correlated binary responses. *Computational Statistics & Data Analysis 51*(11), 5402–5415.

Jeon, J.-J. and H. Choi (2018). The sparse Luce model. *Applied Intelligence 48*, 1953–1964.

Johnson, V. E. (2008). Statistical analysis of the National Institutes of Health peer review system. *Proceedings of the National Academy of Sciences 105*(32), 11076–11080.

Jones, W. P. and S. A. Loe (2013). Optimal number of questionnaire response categories: More may not be better. *Sage Open 3*(2), 2158244013489691.

Kaatz, A., B. Gutierrez, and M. Carnes (2014). Threats to objectivity in peer review: The case of gender. *Trends in Pharmacological Sciences 35*(8), 371–373.

Kamishima, T. (2003). Nantonac collaborative filtering: Recommendation based on order responses. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 583–588.

Kang, D., W. Ammar, B. Dalvi, M. Van Zuylen, S. Kohlmeier, E. Hovy, and R. Schwartz (2018). A dataset of peer reviews (peerread): Collection, insights and NLP applications. *arXiv preprint arXiv:1804.09635*.

Kemeny, J. G. (1959). Mathematics without numbers. *Daedalus 88*(4), 577–591.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika 30*(1/2), 81–93.

Kim, M., F. Farnoud, and O. Milenkovic (2015). Hydra: Gene prioritization via hybrid distance-score rank aggregation. *Bioinformatics 31*(7), 1034–1043.

Kriegeskorte, N., W. K. Simmons, P. S. Bellgowan, and C. I. Baker (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience 12*(5), 535–540.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, 362–412.

Lee, C. J. (2012). A Kuhnian critique of psychometric research on peer review. *Philosophy of Science 79*(5), 859–870.

Lee, C. J., S. Grant, and E. A. Erosheva (2020). Alternative grant models might perpetuate Black–white funding gaps. *The Lancet 396*(10256), 955–956.

Lee, C. J., C. R. Sugimoto, G. Zhang, and B. Cronin (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology 64*(1), 2–17.

Lee, J. H. (1997). Analyses of multiple evidence combination. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–276.

Li, Y., D. F. Hsu, and S. M. Chung (2009). Combining multiple feature selection methods for text categorization by using rank-score characteristics. In *2009 21st IEEE International Conference on Tools with Artificial Intelligence*, pp. 508–517. IEEE.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*.

Little, R. J. and D. B. Rubin (2019). *Statistical Analysis with Missing Data*, Volume 793. John Wiley & Sons.

Liu, A. and A. Moitra (2018). Efficiently learning mixtures of Mallows models. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 627–638. IEEE.

Liu, A., Z. Zhao, C. Liao, P. Lu, and L. Xia (2019). Learning Plackett–Luce mixtures from partial preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 33, pp. 4328–4335.

Liu, Q., A. H. Reiner, A. Frigessi, and I. Scheel (2019). Diverse personalized recommendations with uncertainty from implicit preference data with the Bayesian Mallows model. *Knowledge-Based Systems 186*, 104960.

Liu, Y., Y. Xu, N. B. Shah, and A. Singh (2022). Integrating rankings into quantized scores in peer review. *arXiv preprint arXiv:2204.03505*.

Lu, T. and C. Boutilier (2011). Learning Mallows models with pairwise preferences. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 145–152.

Luce, R. D. (1959). *Individual Choice Behavior*. John Wiley and Sons, Inc.

Luce, R. D. (1977). The Choice Axiom after twenty years. *Journal of Mathematical Psychology 15*(3), 215–233.

Macdonald, C. and I. Ounis (2009). Searching for expertise: Experiments with the voting model. *The Computer Journal 52*(7), 729–748.

Mallard, G., M. Lamont, and J. Guetzkow (2009). Fairness as appropriateness: Negotiating epistemological differences in peer review. *Science, Technology, & Human Values 34*(5), 573–606.

Mallows, C. L. (1957). Non-null ranking models. I. *Biometrika 44*(1/2), 114–130.

Mandhani, B. and M. Meila (2009). Tractable search for learning exponential models of rankings. In *Artificial Intelligence and Statistics*, pp. 392–399. PMLR.

Marden, J. I. (1996). *Analyzing and Modeling Rank Data*. CRC Press.

Marsh, H. W., U. W. Jayasinghe, and N. W. Bond (2008). Improving the peer-review process for grant applications: Reliability, validity, bias, and generalizability. *American Psychologist 63*(3), 160.

Masarotto, G. and C. Varin (2012). The ranking lasso and its application to sport tournaments. *The Annals of Applied Statistics*, 1949–1970.

Maystre, L. and M. Grossglauser (2015). Fast and accurate inference of Plackett–Luce models. *Advances in Neural Information Processing Systems 28*.

McCullagh, P. and J. Yang (2008). How many clusters? *Bayesian Analysis 3*(1), 101–120.

McGraw, K. O. and S. P. Wong (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods 1*(1), 30.

Medcalfe, S. (2018). Economic well-being in US metropolitan statistical areas. *Social Indicators Research 139*(3), 1147–1167.

Meila, M. and L. Bao (2010). An exponential model for infinite rankings. *Journal of Machine Learning Research 11*, 3481–3518.

Meila, M. and H. Chen (2012). Dirichlet process mixtures of generalized Mallows models. *arXiv preprint arXiv:1203.3496*.

Meila, M., K. Phadnis, A. Patterson, and J. A. Bilmes (2012). Consensus ranking under the exponential model. *arXiv preprint arXiv:1206.5265*.

Merrifield, M. R. and D. G. Saari (2009). Telescope time without tears: A distributed approach to peer review. *Astronomy & Geophysics 50*(4), 4–16.

Meyer, J. D., A. Corvillón, J. M. Carpenter, A. L. Plunkett, R. Kurowski, A. Chalevin, J. Bruenker, D.-C. Kim, and E. Macías (2022). Analysis of the ALMA cycle 8 distributed peer review process. *arXiv preprint arXiv:2204.05390*.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*(2), 81.

Miller, J. W. and M. T. Harrison (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association 113*(521), 340–356.

Minneapolis Elections and Voter Services (2021). 2021 mayor results. https://vote.minneapolismn.gov/results-data/election-results/2021/mayor/.

Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association 83*(404), 1023–1032.

Mixon Jr, F. G. and E. W. King (2012). Social choice theory in 10,000 meters: Examining independence and transitivity in the NCAA cross-country championships. *The American Economist 57*(1), 32–41.

Mollica, C. and L. Tardella (2017). Bayesian Plackett–Luce mixture models for partially ranked data. *Psychometrika 82*(2), 442–458.

Morey, E., M. Thiene, M. De Salvo, and G. Signorello (2008). Using attitudinal data to identify latent classes that vary in their preference for landscape preservation. *Ecological Economics 68*(1-2), 536–546.

Munzel, U. and B. Bandelow (1998). The use of parametric vs. nonparametric tests in the statistical evaluation of rating scales. *Pharmacopsychiatry 31*(06), 222–224.

Murphy, T. B. and D. Martin (2003). Mixtures of distance-based models for ranking data. *Computational Statistics & Data Analysis 41*(3-4), 645–655.

Nagaraja, H. N. and S. Sanders (2020). The aggregation paradox for statistical rankings and nonparametric tests. *PLoS One 15*(3), e0228627.

National Institutes of Health (2021). Peer review. https://grants.nih.gov/grants/peer-review.htm.

Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, 1161–1167.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics 4*, 2111–2245.

Nguyen, D. and A. Y. Zhang (2023). Efficient and accurate learning of mixtures of Plackett–Luce models. *arXiv preprint arXiv:2302.05343*.

Nguyen, Q., H. Butler, and G. J. Matthews (2021). An examination of sport climbing's competition format and scoring system. *arXiv preprint arXiv:2111.05310*.

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics 32*(5), 2044–2073.

Ovadia, S. (2004). Ratings and rankings: Reconsidering the structure of values and their measurement. *International Journal of Social Research Methodology 7*(5), 403–414.

Park, T. and G. Casella (2008). The Bayesian lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Patat, F., W. Kerzendorf, D. Bordelon, G. Van de Ven, and T. Pritchard (2019). The distributed peer review experiment. *The Messenger 177*, 3–13.

Patterson, B. H., C. M. Dayton, and B. I. Graubard (2002). Latent class analysis of complex sample survey data: Application to dietary data. *Journal of the American Statistical Association 97*(459), 721–741.

Pearce, M. and E. A. Erosheva (2022a). On the validity of bootstrap uncertainty estimates in the Mallows-Binomial model. *arXiv preprint arXiv:2206.12365*.

Pearce, M. and E. A. Erosheva (2022b). A unified statistical learning model for rankings and scores with application to grant panel review. *Journal of Machine Learning Research 23*(210).

Pearce, M. and E. A. Erosheva (2023). Modeling preferences: A Bayesian mixture of finite mixtures for rankings and ratings. *arXiv preprint arXiv:2301.09755*.

Phillips, D. B. and A. F. Smith (1996). Bayesian model comparison via jump diffusions. *Markov Chain Monte Carlo in Practice 215*, 239.

Pier, E. L., M. Brauer, A. Filut, A. Kaatz, J. Raclaw, M. J. Nathan, C. E. Ford, and M. Carnes (2018). Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences 115*(12), 2952–2957.

Pitman, J. and M. Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 855–900.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 24*(2), 193–202.

Politis, D. N. (1998). Computer-intensive methods in statistical analysis. *IEEE Signal Processing Magazine 15*(1), 39–55.

Porwal, A. and A. Rodriguez (2021). Laplace power-expected-posterior priors for generalized linear models with applications to logistic regression. *arXiv preprint arXiv:2112.02524*.

Poston, R. S. (2008). Using and fixing biased rating schemes. *Communications of the ACM 51*(9), 105–109.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rao, P. and L. L. Kupper (1967). Ties in paired-comparison experiments: A generalization of the Bradley–Terry model. *Journal of the American Statistical Association 62*(317), 194–204.

Remington, P. L., B. B. Catlin, and K. P. Gennuso (2015). The county health rankings: Rationale and methods. *Population Health Metrics 13*(1), 1–12.

Resnik, D. B. and S. A. Elmore (2016). Ensuring the quality, fairness, and integrity of journal peer review: A possible role of editors. *Science and Engineering Ethics 22*(1), 169–188.

Richardson, S. and P. J. Green (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.

Robertson, T. (1988). Order restricted statistical inference. Technical report.

Roumbanis, L. (2019). Peer review or lottery? A critical analysis of two different forms of decision-making mechanisms for allocation of research grants. *Science, Technology, & Human Values 44*(6), 994–1019.

Russell, P. A. and C. D. Gray (1994). Ranking or rating? Some data and their implications for the measurement of evaluative response. *British Journal of Psychology 85*(1), 79–92.

Salomon, J. A. (2003). Reconsidering the use of rankings in the valuation of health states: A model for estimating cardinal values from ordinal data. *Population Health Metrics 1*(1), 1–12.

Sanders, S., J. Ehrlich, and J. Boudreau (2022). Rule selection invariance as a robustness check in collective choice and nonparametric statistical settings. *Public Choice*, 1–20.

Sawadogo, A., S. Dossou-Gbété, and D. Lafon (2017). Ties in one block comparison experiments: A generalization of the Mallows–Bradley–Terry ranking model. *Journal of Applied Statistics 44*(14), 2621–2644.

Schäfer, D. and E. Hüllermeier (2018). Dyad ranking using Plackett–Luce models based on joint feature representations. *Machine Learning 107*(5), 903–941.

Schauberger, G. and G. Tutz (2017). Subject-specific modelling of paired comparison data: A lasso-type penalty approach. *Statistical Modelling 17*(3), 223–243.

Schütte, S., P. N. M. Acevedo, and A. Flahault (2018). Health systems around the world–a comparison of existing health system rankings. *Journal of Global Health 8*(1).

Sen, A. (1986). Social choice theory. *Handbook of Mathematical Economics 3*, 1073–1181.

Shah, N. B., B. Tabibian, K. Muandet, I. Guyon, and U. Von Luxburg (2018). Design and analysis of the NIPS 2016 review process. *Journal of Machine Learning Research*.

Shimamura, K., M. Ueki, S. Kawano, and S. Konishi (2019). Bayesian generalized fused lasso modeling via NEG distribution. *Communications in Statistics–Theory and Methods 48*(16), 4132–4153.

Sinha, P., C. S. Calfee, and K. L. Delucchi (2021). Practitioner's guide to latent class analysis: Methodological considerations and common pitfalls. *Critical Care Medicine 49*(1), e63.

Smith, E. M. (2021). Reimagining the peer-review system for translational health science journals. *Clinical and Translational Science 14*(4), 1210–1221.

Smith, P. C. and L. M. Kendall (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology 47*(2), 149.

Somers, T., N. R. Lawrance, and G. A. Hollinger (2017). Efficient learning of trajectory preferences using combined ratings and rankings. In *Proc. Robotics: Science and Systems Conference Workshop on Mathematical Models, Algorithms, and Human-Robot Interaction (RSS), Boston, MA*.

Song, Q. and G. Cheng (2020). Bayesian fusion estimation via t shrinkage. *Sankhya A 82*(2), 353–385.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 795–809.

Su, W. J. (2022). A truthful owner-assisted scoring mechanism. *arXiv preprint arXiv:2206.08149*.

Sung, Y.-T. and J.-S. Wu (2018). The visual analogue scale for rating, ranking and paired-comparison (VAS-RRP): A new technique for psychological measurement. *Behavior Research Methods 50*(4), 1694–1715.

Tang, W. (2019). Mallows ranking models: Maximum likelihood estimate and regeneration. In *International Conference on Machine Learning*, pp. 6125–6134. PMLR.

Tay, W., X. Zhang, and S. Karimi (2020). Beyond mean rating: Probabilistic aggregation of star ratings based on helpfulness. *Journal of the Association for Information Science and Technology 71*(7), 784–799.

Taylor, J. and R. J. Tibshirani (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences 112*(25), 7629–7634.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review 34*(4), 273.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*(1), 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(1), 91–108.

Tkachenko, M. and H. W. Lauw (2016). Plackett–Luce regression mixture model for heterogeneous rankings. In *Proceedings of the 25th ACM International Conference on Information and Knowledge Management*, pp. 237–246.

Turner, H. L., J. van Etten, D. Firth, and I. Kosmidis (2020). Modelling rankings in R: The PlackettLuce package. *Computational Statistics 35*(3), 1027–1057.

Tutz, G. and G. Schauberger (2015). Extended ordered paired comparison models with application to football data from German Bundesliga. *Advances in Statistical Analysis 99*, 209–227.

van der Vaart, A. W. (2000). *Asymptotic Statistics*, Volume 3. Cambridge University Press.

van Herk, H. and M. van de Velden (2007). Insight into the relative merits of rating and ranking in a cross-national context using three-way correspondence analysis. *Food Quality and Preference 18*(8), 1096–1105.

Vana, L., R. Hochreiter, and K. Hornik (2016). Computing a journal meta-ranking using paired comparisons and adaptive lasso estimators. *Scientometrics 106*, 229–251.

Varin, C., M. Cattelan, and D. Firth (2016). Statistical modelling of citation exchange between statistics journals. *Journal of the Royal Statistical Society, Series A(Statistics in Society) 179*(1), 1.

Venkatraghavan, V., E. E. Bron, W. J. Niessen, S. Klein, A. D. N. Initiative, et al. (2019). Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage 186*, 518–532.

Vitelli, V., Ø. Sørensen, M. Crispino, A. Frigessi Di Rattalma, and E. Arjas (2018). Probabilistic preference learning with the Mallows rank model. *Journal of Machine Learning Research 18*(158), 1–49.

Wang, J. and N. B. Shah (2018). Your 2 is my 1, your 3 is my 9: Handling arbitrary miscalibrations in ratings. *arXiv preprint arXiv:1806.05085*.

Wang, J. and N. B. Shah (2020). Ranking and rating rankings and ratings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 34, pp. 13704–13707.

Wang, Y. S., R. L. Matsueda, E. A. Erosheva, et al. (2017). A variational EM method for mixed membership models with multivariate rank data: An analysis of public policy preferences. *The Annals of Applied Statistics 11*(3), 1452–1480.

Wenneras, C. and A. Wold (2010). *Nepotism and Sexism in Peer-Review*. Routledge.

Wu, S., K. Shimamura, K. Yoshikawa, K. Murayama, and S. Kawano (2021). Variable fusion for Bayesian linear regression via spike-and-slab priors. In *Intelligent Decision Technologies: Proceedings of the 13th KES-IDT 2021 Conference*, pp. 491–501. Springer.

Xu, Y., S. Balakrishnan, A. Singh, and A. Dubrawski (2020). Regression with comparisons: Escaping the curse of dimensionality with ordinal information. *Journal of Machine Learning Research 21*(1), 6480–6533.

Young, H. P. (1988). Condorcet's theory of voting. *American Political Science Review 82*(4), 1231–1244.

Zeng, S. and J. Shen (2020). Learning from the crowd with pairwise comparison. *arXiv preprint arXiv:2011.01104*.

Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift 29*(1), 436–460.

Zhang, X., X. Zhang, P.-L. Loh, and Y. Liang (2022). On the identifiability of mixtures of ranking models. *arXiv preprint arXiv:2201.13132*.

Zhao, Z., P. Piech, and L. Xia (2016). Learning mixtures of Plackett–Luce models. In *International Conference on Machine Learning*, pp. 2906–2914. PMLR.

Zhao, Z. and L. Xia (2019). Learning mixtures of Plackett–Luce models from structured partial orders. *Advances in Neural Information Processing Systems 32*.

Zhu, Z., X. Wang, and S. Qiu (2019). Comparing rank aggregation methods based on Mallows model. In *3rd International Conference on Mechatronics Engineering and Information Technology (ICMEIT 2019)*, pp. 609–616. Atlantis Press.

Appendix A

This is the Appendix to Chapter 3. It includes an exploration of estimation bias, proofs of asymptotic consistency of the maximum likelihood estimators and asymptotic validity of bootstrapped standard errors for the Mallows-Binomial model, and additional results from the AIBS grant panel review application from Chapter 3.4.

A.1 Bias in Mallows-Binomial Maximum Likelihood Estimators

In this section, we explore bias of the Mallows-Binomial maximum likelihood estimates (MLE), $(\hat{p}, \hat{\theta})$, for parameters (p_0, θ_0). We start with a simple example of bias before providing theoretical results.

A.1.1 Example Demonstrating Bias

Consider the following example: Let $I = 1$, $J = R = 3$, and $M = 1$. Then, there are 48 unique possible preference observations. Specifically, they are the combinations of $X_1 \in \{0, 1\}$, $X_2 \in \{0, 1\}$, $X_3 \in \{0, 1\}$, and $\Pi \in \{1 \prec 2 \prec 3, 1 \prec 3 \prec 2, 2 \prec 1 \prec 3, 2 \prec 1 \prec 3, 3 \prec 1 \prec 2, 3 \prec 2 \prec 1\}$. We enumerate each possible observation and state its associated MLE in Table A.1.

Now suppose $p_0 = [0.1, 0.4, 0.9]^T$ and $\theta_0 = 1$. Then, the bias of each MLE is,

$$\begin{aligned}
\text{Bias}(\hat{p}_1) &= E_{P_{p_0,\theta_0}}[\hat{p}_1] - p_{01} \approx 0.1419 - 0.1 = 0.0419 \\
\text{Bias}(\hat{p}_2) &= E_{P_{p_0,\theta_0}}[\hat{p}_2] - p_{02} \approx 0.4192 - 0.4 = 0.0192 \\
\text{Bias}(\hat{p}_3) &= E_{P_{p_0,\theta_0}}[\hat{p}_3] - p_{03} \approx 0.8390 - 0.9 = -0.0610 \\
\text{Bias}(\hat{\theta}) &= E_{P_{p_0,\theta_0}}[\hat{\theta}] - \theta_0 = \infty - 1 = \infty
\end{aligned}$$

Thus, each MLE in this example is biased.

Observed Data				Associated MLE				Observed Data				Associated MLE			
Π	X_1	X_2	X_3	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{\theta}$	Π	X_1	X_2	X_3	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{\theta}$
$1 \prec 3 \prec 2$	0	0	0	0	0	0	∞	$2 \prec 1 \prec 3$	0	0	0	0	0	0	∞
$2 \prec 1 \prec 3$	1	0	0	0.5	0	0.5	∞	$3 \prec 2 \prec 1$	1	0	0	1	0	0	∞
$3 \prec 2 \prec 1$	0	1	0	0.5	0.5	0	∞	$1 \prec 3 \prec 2$	0	1	0	0	1	0	∞
$1 \prec 3 \prec 2$	1	1	0	0.5	1	0.5	∞	$2 \prec 1 \prec 3$	1	1	0	1	1	0	0
$2 \prec 1 \prec 3$	0	0	1	0	0	1	∞	$3 \prec 2 \prec 1$	0	0	1	0	0	1	0
$3 \prec 2 \prec 1$	1	0	1	1	0.5	0.5	∞	$1 \prec 3 \prec 2$	1	0	1	1	0	1	0
$1 \prec 3 \prec 2$	0	1	1	0	1	1	∞	$2 \prec 1 \prec 3$	0	1	1	0.5	0.5	1	∞
$2 \prec 1 \prec 3$	1	1	1	1	1	1	∞	$3 \prec 2 \prec 1$	1	1	1	1	1	1	∞
$1 \prec 2 \prec 3$	0	0	0	0	0	0	∞	$3 \prec 2 \prec 1$	0	0	0	0	0	0	∞
$3 \prec 1 \prec 2$	1	0	0	0.5	0.5	0	∞	$1 \prec 3 \prec 2$	1	0	0	1	0	0	0
$2 \prec 3 \prec 1$	0	1	0	0	1	0	0	$2 \prec 1 \prec 3$	0	1	0	0	1	0	0
$1 \prec 2 \prec 3$	1	1	0	1	1	0	0	$3 \prec 2 \prec 1$	1	1	0	1	1	0	∞
$3 \prec 1 \prec 2$	0	0	1	0	0	1	0	$1 \prec 3 \prec 2$	0	0	1	0	0.5	0.5	∞
$2 \prec 3 \prec 1$	1	0	1	1	0	1	∞	$2 \prec 1 \prec 3$	1	0	1	1	0	1	∞
$1 \prec 2 \prec 3$	0	1	1	0	1	1	∞	$3 \prec 2 \prec 1$	0	1	1	0	1	1	0
$3 \prec 1 \prec 2$	1	1	1	1	1	1	∞	$1 \prec 3 \prec 2$	1	1	1	1	1	1	∞
$2 \prec 3 \prec 1$	0	0	0	0	0	0	∞	$3 \prec 1 \prec 2$	0	0	0	0	0	0	∞
$1 \prec 2 \prec 3$	1	0	0	1	0	0	0	$2 \prec 3 \prec 1$	1	0	0	1	0	0	∞
$3 \prec 1 \prec 2$	0	1	0	0	1	0	∞	$1 \prec 2 \prec 3$	0	1	0	0	0.5	0.5	∞
$2 \prec 3 \prec 1$	1	1	0	1	0.5	0.5	∞	$3 \prec 1 \prec 2$	1	1	0	1	1	0	∞
$1 \prec 2 \prec 3$	0	0	1	0	0	1	∞	$2 \prec 3 \prec 1$	0	0	1	0.5	0	0.5	∞
$3 \prec 1 \prec 2$	1	0	1	1	0	1	0	$1 \prec 2 \prec 3$	1	0	1	0.5	0.5	1	∞
$2 \prec 3 \prec 1$	0	1	1	0	1	1	0	$3 \prec 1 \prec 2$	0	1	1	0.5	1	0.5	∞
$1 \prec 2 \prec 3$	1	1	1	1	1	1	∞	$2 \prec 3 \prec 1$	1	1	1	1	1	1	∞

Table A.1: Enumerated observations and associated MLEs in a toy example demonstrating bias of Mallows-Binomial maximum likelihood estimators.

A.1.2 Theoretical Results on Bias

We begin by proving $(\hat{p}, \hat{\theta})$ is biased when π_0 is known: Assume π_0 is fixed and known. Then,

$$\begin{aligned}(\hat{p}, \hat{\theta})|\pi_0 &= \underset{p,\theta|\text{Order}(p)=\pi_0}{\arg\max} \prod_{i=1}^{I} \frac{e^{-\theta d_K(\pi_i, \text{Order}(p))}}{\psi_{R,J}(\theta)} \prod_{j=1}^{J} p_j^{x_{ij}} (1-p_j)^{M-x_{ij}} \\ &= \underset{p,\theta|\text{Order}(p)=\pi_0}{\arg\max} -\theta\Big[\sum_{i=1}^{I} d_K(\pi_i, \pi_0)\Big] - I \log \psi_{R,J}(\theta) + \\ &\qquad \sum_{j=1}^{J} \Big[\sum_{i=1}^{I} x_{ij} \log p_j + (IM - \sum_{i=1}^{I} x_{ij}) \log(1-p_j)\Big] \end{aligned} \tag{A.1}$$

p and θ factor in Equation A.1. Thus,

$$\begin{aligned}\hat{\theta}|\pi_0 &= \underset{\theta|\pi_0}{\arg\max} -\theta\Big[\sum_{i=1}^{I} d(\pi_i, \pi_0)\Big] - I \log \psi_R(\theta) \\ \hat{p}|\pi_0 &= \underset{p|\text{Order}(p)=\pi_0}{\arg\max} \sum_{j=1}^{J} \Big[\overline{x}_j \log p_j + (M - \overline{x}_j) \log(1-p_j)\Big] \end{aligned}$$

Tang (2019) proved that $E[\hat{\theta}|\pi_0] > \theta$. So when π_0 is known, $\hat{\theta}$ is biased upward. Regarding $\hat{p}$, the problem is now precisely an *isotonic regression* problem for Binomial probabilities. It was shown in Ayer et al. (1955) and Barlow and Brunk (1972) that

$$\hat{p}_j|\pi_0 = \begin{cases} \overline{x}_j/M & \overline{x}_1, \ldots, \overline{x}_{j-1} \le \overline{x}_j \le \overline{x}_{j+1}, \ldots, \overline{x}_J \\ (\sum_{i \in A_j} \overline{x}_i)/M, & \text{otherwise} \end{cases}$$

where A_j is the intersection of the lower and upper sets of π_0 that include j. Robertson (1988) proved that $E[\hat{p}_j|\pi_0] \neq p_j$ in generality, i.e., the parameters p_j, $j = 1, \ldots, J$ are biased. The direction of the bias may vary between each p_j. Thus, $(\hat{p}, \hat{\theta})|\pi_0$ is biased.

We would now like to prove that the MLEs are biased even when π_0 is not known. This is a substantially more challenging problem due to the interconnectedness of $\hat{p}$ and $\hat{\theta}$ during estimation. A complete proof is left open. However, the previous counterexample demonstrates that bias is present in at least some situations. At the same time, simulations in Chapter 3.2.2 demonstrate the bias is often minimal.

A.2 Consistency of Mallows-Binomial Maximum Likelihood Estimators

In this section, we prove consistency of the Mallows-Binomial MLE.

Proposition 2 *Let M, J, and R be fixed and positive integers such that $R \leq J$. Let $\theta_0 \in (0, \infty)$ and $p_0 \in (0,1)^J$. Let $(X, \Pi)_I$ denote a sample of I independent and identically distributed samples from a Mallows-Binomial(p_0, θ_0) distribution, and $(\hat{p}, \hat{\theta})_I$ be the maximum likelihood estimators based on that sample. Then, $(\hat{p}, \hat{\theta})_I \xrightarrow{p} (p_0, \theta_0)$.*

Proof Assume that each true p_{0j}, $j = 1, \ldots, J$ lie within in the unit interval and are each bounded away from 0 and 1. Furthermore, assume that the true θ_0 is less than J and is bounded greater than 0. We note that the restrictions on p_0 ensure each proposal may receive any integer rating between 0 and M with positive probability. Furthermore, the restrictions on θ_0 ensure there is not a complete lack of consensus (as if θ_0 were equal to 0) and does not substantially impact situations of near-complete consensus (when $\theta_0 \geq J$, it is near certain that all rankings are identical and match the true π_0 regardless of how large θ_0 is). Under these assumptions, the unknown parameters live in a compact space. We denote this space by Θ.

We define a few quantities. First, let $\ell_{p,\theta}(X_i, \Pi_i)$ be the log likelihood of data (X_i, Π_i) under a Mallows-Binomial(p, θ) distribution, less the normalizing constant. Specifically,

$$\begin{aligned} \ell_{p,\theta}(X_i, \Pi_i) = &- \theta d_{R,J}\big(\Pi_i, \text{Order}(p)\big) - \log \psi_{R,J}(\theta) \\ &+ \sum_{j=1}^{J} \left[X_{ij} \log p_j + (M - X_{ij}) \log(1 - p_j) \right] \end{aligned} \tag{A.2}$$

Note that $\ell(p, \theta)(\cdot, \cdot)$ is not continuous in p. Discontinuities may exist when $p_j = p_k$, $j \neq k$. Furthermore, note that $d_{R,J}(\Pi_i, \text{Order}(p)) \in \{0, 1, \ldots, J(J-1)/2\}$ and $\psi_{R,J}(\theta) \in (1, J!)$. Continuing on, we define,

$$M_I(p, \theta) = \frac{1}{I} \sum_{i=1}^{I} \ell_{p,\theta}(X_i, \Pi_i) \tag{A.3}$$

$$M(p, \theta) = E[\ell_{p,\theta}(X_1, \Pi_1)]. \tag{A.4}$$

Note that

$$M_I(p,\theta) = -\theta\overline{d_{R,J}}(\Pi, \text{Order}(p)) - \log\psi_{R,J}(\theta) + \sum_{j=1}^{J}[\overline{X}_j \log p_j + (M - \overline{X}_j)\log(1-p_j)]$$

by definition and

$$\begin{aligned} M(p,\theta) = &-\theta\Big[\frac{Re^{-\theta}}{1-e^{-\theta}} - \sum_{j=J-R+1}^{J}\frac{je^{-j\theta}}{1-e^{-j\theta}}\Big] - \log\psi_{R,J}(\theta) \\ &+ \sum_{j=1}^{J}\Big[Mp_j\log p_j + (M - Mp_j)\log(1-p_j)\Big] \end{aligned}$$

since,

$$\mathrm{E}[X_j] = Mp_j, \quad \mathrm{Var}[X_j] = Mp_j(1-p_j) \tag{A.5}$$

$$\mathrm{E}[d_{R,J}] = \frac{Re^{-\theta}}{1-e^{-\theta}} - \sum_{j=J-R+1}^{J}\frac{je^{-j\theta}}{1-e^{-j\theta}} \tag{A.6}$$

$$\mathrm{Var}[d_{R,J}] = \frac{Re^{-\theta}}{(1-e^{-\theta})^2} - \sum_{j=J-R+1}^{J}\frac{j^2e^{-j\theta}}{(1-e^{-j\theta})^2} \tag{A.7}$$

Equation A.5 is a standard result for Binomial random variables and Equations A.6 and A.7 follow directly from Fligner and Verducci (1986).

We are now ready to prove the consistency of the maximum likelihood estimators. We do so using Theorem 5.7 of van der Vaart (2000). Specifically, we must prove (1) uniform consistency of $M_I(p,\theta)$ to $M(p,\theta)$, (2) the true parameter values are well-separated in $M(p,\theta)$, and (3) that $M_I(\hat{p}_I, \hat{\theta}_I) \geq M_I(p_0,\theta_0) - o_p(1)$ as $I \to \infty$. Then, the MLE is consistent.

Condition 1 Uniform consistency is proven via Corollary 2.2 of Newey (1991). Specifically, under the assumption that the true parameters live in the compact space Θ, we must prove that M_I converges pointwise to M and that $\forall(p,\theta),(p',\theta') \in \Theta$, $|M_I(p',\theta') - M_I(p,\theta)| \leq O_p(1)||(p',\theta'),(p,\theta)||_1$.

We start with pointwise convergence: Let $(p,\theta) \in \Theta$ be arbitrary but fixed. Then,

$$
\begin{aligned}
M_I(p,\theta) - M(p,\theta) =& -\theta(\overline{d}_{R,J} - \mathrm{E}[\overline{d}_{R,J}]) - \log\frac{\psi(\theta)}{\psi(\theta)} \\
&+ \sum_{j=1}^{J} \log p_j(\overline{X}_j - \mathrm{E}[\overline{X}_j]) + \log(1-p_j)(\mathrm{E}[\overline{X}_j] - \overline{X}_j) \\
=& -\theta o_p(1) + \sum_{j=1}^{J} \log p_j o_p(1) + \log(1-p_j) o_p(1) \\
=& o_p(1).
\end{aligned}
$$

Thus, M_I converges pointwise to M. Next, let $(p,\theta),(p',\theta') \in \Theta$ be arbitrary but fixed. Then,

$$
\begin{aligned}
|M_I(p',\theta') - M_I(p,\theta)| =& \Big|\theta'\overline{D}(\Pi, \mathrm{Order}(p')) - \theta\overline{D}(\Pi, \mathrm{Order}(p')) + \log\frac{\psi(\theta')}{\psi(\theta)} \\
&+ \sum_{j=1}^{J} \Big[\overline{X}_j \log\frac{p_j}{p'_j} + (M - \overline{X}_j)\log\frac{1-p_j}{1-p'_j}\Big]\Big| \qquad \text{(A.8)} \\
\leq& \underbrace{\big|\theta'\overline{D}(\Pi, \mathrm{Order}(p')) - \theta\overline{D}(\Pi, \mathrm{Order}(p'))\big|}_{\text{Term 1}} + \underbrace{\big|\log\frac{\psi(\theta')}{\psi(\theta)}\big|}_{\text{Term 2}} \\
&+ \underbrace{\sum_{j=1}^{J} \Big|\overline{X}_j \log\frac{p_j}{p'_j} + (M - \overline{X}_j)\log\frac{1-p_j}{1-p'_j}\Big|}_{\text{Term 3}} \qquad \text{(A.9)}
\end{aligned}
$$

where Equation A.8 holds by definition and Equation A.9 by the triangle inequality. We investigate each numbered term sequentially. Starting with Term 1,

Term 1

$$
\begin{aligned}
&= \Big|\theta' \overline{D}(\Pi, \text{Order}(p')) - \theta \overline{D}(\Pi, \text{Order}(p'))\Big| \\
&= \Big|\theta' \Big(\overline{D}(\Pi, \text{Order}(p')) - \text{E}[\overline{D}(\Pi, \text{Order}(p'))] + \text{E}[\overline{D}(\Pi, \text{Order}(p'))] - \text{E}[\overline{D}(\Pi, \text{Order}(p_0))]\Big) - \\
&\quad \theta \Big(\overline{D}(\Pi, \text{Order}(p')) - \text{E}[\overline{D}(\Pi, \text{Order}(p'))] + \text{E}[\overline{D}(\Pi, \text{Order}(p'))] - \text{E}[\overline{D}(\Pi, \text{Order}(p_0))]\Big) + \\
&\quad (\theta' - \theta) E[\overline{D}(\Pi, \text{Order}(p_0))]\Big| \\
&= \Big|\theta' \big(o_p(1) + O_p(1)\big) - \theta \big(o_p(1) + O_p(1)\big) + \big(\theta' - \theta\big) O_p(1)\Big| \\
&= \Big|(\theta' - \theta)\Big| O_p(1).
\end{aligned}
$$

Next,

$$\text{Term } 2 = |\log \psi(\theta') - \log \psi(\theta)| \leq C_1 |\theta' - \theta| = O_p(1)|\theta' - \theta|,$$

since $\log \psi(\theta)$ is a continuous function defined on a compact range and therefore must have a maximum and minimum slope over that range. Similarly,

$$
\begin{aligned}
\text{Term } 3 &= \sum_{j=1}^{J} \big|\overline{X}_j (\log p_j - \log p'_j) + (M - \overline{X}_j)(\log(1 - p_j) - \log(1 - p'_j))\big| \\
&\leq \sum_{j=1}^{J} \overline{X}_j C_2 \big|p_j - p'_j\big| + (M - \overline{X}_j) C_3 \big|p_j - p'_j\big| \\
&= \sum_{j=1}^{J} \big|p' - p\big| O_p(1).
\end{aligned}
$$

since $\log(p_j)$ and $\log(1 - p_j)$ are also continuous functions on the compact range. Combining these results for Terms 1, 2, and 3, we have

$$
\begin{aligned}
|M_I(p', \theta') - M_I(p, \theta)| =& O_p(1)|\theta' - \theta| + O_p(1)|\theta' - \theta| + \sum_{j=1}^{J} O_p(1)|p'_j - p_j| \\
=& O_p(1)||(p', \theta'), (p, \theta)||_1. \qquad \text{(A.10)}
\end{aligned}
$$

Therefore, the estimator is uniformly consistent.

Condition 2 The well-separation condition on $M(p_0, \theta_0)$ is straightforward to prove using Lemma 2.2 of Newey and McFadden (1994): Since the model is identified (as proved via Proposition 1) and $\mathrm{E}|\ell_{p,\theta}(X, \Pi)| < \infty \; \forall p, \theta \in \Theta$ due to the constrained domain of each X_{ij} and Π_i regardless of $p, \theta \in \Theta$, we have the desired result.

Condition 3 The third condition that the sequence $(\hat{p}_I, \hat{\theta}_I)$ as $I \to \infty$ satisfies $M_I(\hat{p}_I, \hat{\theta}_I) \geq M_I(p_0, \theta_0) - o_p(1)$ is a standard property of the MLE.

Thus, according to the conditions of Theorem 5.7 in van der Vaart (2000), $(\hat{p}_I, \hat{\theta}_I)$ is consistent for (p_0, θ_0).

■

A.3 Asymptotic Validity of Bootstrapped Standard Errors

In this section, we prove the asymptotic validity of bootstrapped standard errors in the Mallows-Binomial model. This work was adapted from Pearce and Erosheva (2022a).

A.3.1 Background on the Nonparametric Bootstrap

The bootstrap is a very general tool for uncertainty quantification that was first proposed in Efron (1979). The nonparametric bootstrap is used to estimate the inherent uncertainty of an estimator without making assumptions on its distributional form. Given n i.i.d. observations, $X = (X_1, \ldots, X_n)$, the nonparametric bootstrap is performed using the following steps:

1. Re-sample n observations with replacement from the original dataset B times, for large B. Denote each bootstrap sample X^b, $b = 1, \ldots, B$.

2. Estimate the unknown statistic(s) of interest, θ, separately using each bootstrap sample. Denote the estimates from each bootstrap sample $\hat{\theta}^b$, $b = 1, \ldots, B$.

3. Form an empirical distribution for $\hat{\theta}$ using the values $\hat{\theta}^b$, $b = 1, \ldots, B$.

Quantiles from the empirical distribution of the unknown statistic(s) of interest may be used for the purpose of creating confidence regions.

Despite its wide applicability, the nonparametric bootstrap does not always yield asymptotically valid confidence intervals. A canonical example in which the bootstrap fails is the estimation of the unknown parameter θ given i.i.d. samples from a Uniform$(0, \theta)$ distribution. Here, the MLE of θ is the maximum order statistic, which has a limiting exponential distribution. The bootstrap empirical distribution, however, will not be able to replicate the asymptotic distribution given the fixed sample, in which the bootstrap estimates of θ will always be less than or equal to the full-sample maximum order statistic (Bickel and Freedman 1981). Another canonical example in which the bootstrap fails is in estimating the location parameter of a Cauchy distribution. In this setting, the MLE is the sample mean which is itself Cauchy distributed and therefore has infinite variance. As a result, the bootstrap estimator behaves poorly even given large samples (Politis 1998).

The Mallows-Binomial likelihood has an unusual form that makes the asymptotic validity of bootstrap uncertainty for the MLE unclear. Specifically, the model is parameterized by continuous parameters whose discrete order impacts the likelihood. That is, the likelihood contains both continuous and discrete components; discontinuities may exist whenever the order of certain parameters change. Thus, frequentist estimation of the Mallows-Binomial model is both a continuous and discrete problem. In the absence of theoretical results regarding the asymptotic distribution of the maximum likelihood estimators, the validity of the nonparametric bootstrap is unclear.

A.3.2 Proof

We now prove the asymptotic validity of bootstrapped standard errors in the Mallows-Binomial model.

Proposition 3 *Let M, J, and R be fixed and positive integers such that $R \leq J$. Let*

$(X, \Pi)_I$ denote a sample of I independent and identically distributed samples from a Mallows-Binomial(p_0, θ_0) distribution. Then, the nonparametric bootstrap estimates of standard errors for $(\hat{p}, \hat{\theta})$ based on $(X, \Pi)_I$ are parameter-wise asymptotically valid as $I \to \infty$.

Proof We begin with a few preliminaries. First, let $\pi_{p_0} = \text{Order}(p_0)$, i.e., the true consensus ranking. Then, note that the joint likelihood of the observed data $(X, \Pi)_I$ can be written,

$$\begin{aligned}\mathcal{L}\Big((X,\Pi)_I|p_0,\theta_0\Big) &= \prod_{i=1}^{I}\left(\frac{e^{-\theta_0 d(\pi_i,\pi_{p_0})}}{\psi(\theta_0)} \times \prod_{j=1}^{J}\binom{M}{x_{ij}} p_{0j}^{x_{ij}}(1-p_{0j})^{M-x_{ij}}\right) \\ &= \frac{e^{-\theta_0 \sum_{i=1}^{I} d(\pi_i,\pi_{p_0})}}{\psi(\theta_0)^I} \times \prod_{j=1}^{J}\left(\prod_{i=1}^{I}\binom{M}{x_{ij}}\right) p_{0j}^{\sum_{i=1}^{I} x_{ij}}(1-p_{0j})^{IM-\sum_{i=1}^{I} x_{ij}},\end{aligned} \tag{A.11}$$

where $d(\cdot,\cdot)$ is the Kendall's τ distance between the two rankings and

$$\psi(\theta) = \prod_{j=1}^{J}\frac{1-e^{-j\theta}}{1-e^{-\theta}}. \tag{A.12}$$

Thus, the MLE $(\hat{p}, \hat{\theta})$ can be expressed according to,

$$\begin{aligned}(\hat{p},\hat{\theta}) &= \arg\max_{p,\theta}\Bigg[-\theta\sum_{i=1}^{I} d(\pi_i,\pi_p) - I\log\psi(\theta) \\ &\qquad + \sum_{j=1}^{J}(\sum_{i=1}^{I} x_{ij})\log(p_j) + (IM - \sum_{i=1}^{I} x_{ij})\log(1-p_j)\Bigg] \\ &= \arg\max_{p,\theta}\Bigg[-\theta\bar{D}(\pi,\pi_p) - \log\psi(\theta) + \sum_{j=1}^{J}\Big(\bar{x_j}\log(p_j) + (M-\bar{x_j})\log(1-p_j)\Big)\Bigg] \\ &\equiv \arg\max_{p,\theta}\Big[f_{(X,\Pi)_I}(p,\theta)\Big],\end{aligned} \tag{A.13}$$

where

$$\bar{D}(\pi,\pi_p) = I^{-1}\sum_i d(\pi_i,\pi_p) \tag{A.14}$$

$$\bar{x}_j = I^{-1}\sum_i x_{ij}. \tag{A.15}$$

Now, assume that $p_{0j} \neq p_{0k}$ whenever $j \neq k$, each $p_{0j} \in (a, b) \subset [0, 1]$, and $\theta \in (c, d) \subset [0, \infty)$. Under these conditions, Proposition 2 proves that $(\hat{p}, \hat{\theta})$ is consistent for (p_0, θ_0) as $I \to \infty$.

A sufficient condition for bootstrap validity is local asymptotic normality of the MLE (Hall 2013; Bickel and Freedman 1981). As such, we show that the $(J+1)$-dimensional MLE $(\hat{p}, \hat{\theta})$ is coordinate-wise, locally asymptotically normal.

Local Asymptotic Normality of $\hat{p}_j$: We begin by considering each $\hat{p}_j$, $j = 1, \ldots, J$. Note that $\hat{p}_j$ is the solution to the following equation:

$$0 = \frac{\partial}{\partial p_j} f_{(X,\Pi)_I}(p, \theta) \tag{A.16}$$

$$= -\theta \Big[\frac{\partial}{\partial p_j} \bar{D}(\pi, \pi_p) \Big] + \frac{\bar{x}_j}{p_j} - \frac{M - \bar{x}_j}{1 - p_j} \tag{A.17}$$

A key challenge in calculating $\hat{p}_j$ is the derivative $\frac{\partial}{\partial p_j} \bar{D}(\pi, \pi_p)$, which is a function of the J-dimensional vector p. However, as long as each $p_k \neq p_j$ whenever $k \neq j$, then π_p will remain constant in small perturbations around p_j and the derivative in p_j will thus be 0.

Generalizing from p_j to p, we require there to exist an ϵ-ball around the J-dimensional vector p such that $\text{Order}(p) = \pi_p$ remains constant for all p' in the ball. In such cases, we have $\frac{\partial}{\partial p_j} \bar{D}(\pi, \pi_p) = 0$ and thus $\hat{p}_j$ is defined by the standard Binomial MLE, $\bar{x}_j / M$, which is asymptotically normal as it is a function of the mean of i.i.d. random variables.

Specifically, this means that in a local region defined by the order of $\hat{p}$, we have the standard Binomial result,

$$\sqrt{I}\Big(\hat{p}_j - p_{0j}\Big) \xrightarrow{d} N\Big(0, \frac{p_{0j}(1 - p_{0j})}{M}\Big), \tag{A.18}$$

which establishes the coordinate-wise local asymptotic normality of $\hat{p}_j$, $j = 1, \ldots, J$.

Local Asymptotic Normality of $\hat{\theta}$: We now show that $\hat{\theta}$ is coordinate-wise a locally asymptotically normal estimator of θ_0. Note that $\hat{\theta}$ is the solution to the following equation:

$$0 = \frac{\partial}{\partial\theta} f_{(X,\Pi)_I}(p,\theta) \tag{A.19}$$

$$= -\bar{D}(\pi,\pi_p) - \frac{\psi'(\theta)}{\psi(\theta)} \tag{A.20}$$

$$\implies \bar{D}(\pi,\pi_p) = -\frac{\psi'(\theta)}{\psi(\theta)}. \tag{A.21}$$

For simplicity, we define $\kappa(\theta) = -\psi'(\theta)/\psi(\theta)$. Thus, we have

$$\hat{\theta} = \kappa^{-1}(\bar{D}(\pi,\pi_p)). \tag{A.22}$$

No simple expression for κ^{-1} exists. However, it can be seen from Equation A.12 that when $J \geq 2$ and for any $\theta > 0$, ψ is a continuous and positive function with a continuous and strictly negative first derivative and a continuous and strictly positive second derivative. Thus, $\kappa(\cdot)$ is a smooth and positive function. Furthermore, κ is monotone decreasing (Fligner and Verducci 1986). As a result, its inverse $\kappa^{-1}(\cdot)$ is well-defined and so is $\hat{\theta}$ given $\bar{D}(\pi,\pi_p)$.

For reasons which will be made clear later, we also write out an expression for $\kappa(\theta)$:

$$\kappa(\theta) = -\frac{\psi'(\theta)}{\psi(\theta)} = -\frac{\frac{(1-e^{-\theta})^J(\frac{\partial}{\partial\theta}\prod_{j=1}^J 1-e^{-j\theta}) - (\prod_{j=1}^J 1-e^{-j\theta})Je^{-\theta}(1-e^{-\theta})^{J-1}}{(1-e^{-\theta})^{2J}}}{\frac{\prod_{j=1}^J 1-e^{-j\theta}}{(1-e^{-\theta})^J}} \tag{A.23}$$

$$= -\frac{(\frac{\partial}{\partial\theta}\prod_{j=1}^J 1-e^{-j\theta}) - (\prod_{j=1}^J 1-e^{-j\theta})Je^{-\theta}/(1-e^{-\theta})}{\prod_{j=1}^J 1-e^{-j\theta}} \tag{A.24}$$

$$= \frac{Je^{-\theta}}{1-e^{-\theta}} - \frac{(\frac{\partial}{\partial\theta}\prod_{j=1}^J 1-e^{-j\theta})}{\prod_{j=1}^J 1-e^{-j\theta}} \tag{A.25}$$

$$= \frac{Je^{-\theta}}{1-e^{-\theta}} - \frac{\partial}{\partial\theta}\log(\prod_{j=1}^J 1-e^{-j\theta}) \tag{A.26}$$

$$= \frac{Je^{-\theta}}{1-e^{-\theta}} - \sum_{j=1}^J \frac{je^{-j\theta}}{1-e^{-j\theta}} \tag{A.27}$$

Next, note that $\bar{D}(\pi,\pi)$ is a random variable given a fixed consensus ranking π, due to the randomness in the collection of rankings π. According to Fligner and Verducci (1986),

$\bar{D}(\pi, \pi_{p_0})$ is asymptotically normal with mean and variance depending on the true θ_0. Specifically,

$$\mu_{\theta_0} \equiv E_{\theta_0}[D(\pi_i, \pi_{p_0})] = \frac{Je^{-\theta_0}}{1-e^{-\theta_0}} - \sum_{j=1}^{J} \frac{je^{-j\theta_0}}{1-e^{-j\theta_0}} \tag{A.28}$$

$$\sigma^2_{\theta_0} \equiv Var_{\theta_0}[D(\pi_i, \pi_{p_0})] = \frac{Je^{-\theta_0}}{(1-e^{-\theta_0})^2} - \sum_{j=1}^{J} \frac{j^2e^{-j\theta_0}}{(1-e^{-j\theta_0})^2} \tag{A.29}$$

$$\implies \sqrt{I}(\bar{D}(\pi, \pi_{p_0}) - \mu_{\theta_0}) \xrightarrow{d} N(0, \sigma^2_{\theta_0}) \tag{A.30}$$

Interestingly, we see from comparing Equations A.27 and A.28 that $\mu_{\theta_0} = \kappa(\theta_0)$, which implies $\kappa^{-1}(\mu_{\theta_0}) = \theta_0$. Since $\kappa^{-1\prime}$ is a real-valued function that does not equal 0, by the Delta method,

$$\sqrt{I}\Big(\kappa^{-1}(\bar{D}(\pi, \pi_{p_0})) - \kappa^{-1}(\mu_{\theta_0})\Big) \xrightarrow{d} N\Big(0, \sigma^2_{\theta_0}[\kappa^{-1\prime}(\mu_{\theta_0})]^2\Big) \tag{A.31}$$

$$\implies \sqrt{I}\Big(\hat{\theta} - \theta_0\Big) \xrightarrow{d} N\Big(0, \sigma^2_{\theta_0}[\kappa^{-1\prime}(\mu_{\theta_0})]^2\Big) \tag{A.32}$$

Although the asymptotic variance cannot be written in closed-form expression, it is positive and finite. Therefore, in a local area of π_{p_0}, $\hat{\theta}$ is coordinate-wise an asymptotically normal estimator of θ.

Asymptotic Normality of $(\hat{p}, \hat{\theta})$: We have now proven that in a local neighborhood of the true consensus ranking π_{p_0}, the estimators $\hat{p}_j$, $j = 1, \ldots, J$ and $\hat{\theta}$ are coordinate-wise asymptotically normal. As stated, the MLE is consistent as $I \to \infty$. Thus, $\hat{p} \xrightarrow{p} p_0$ and $\hat{\theta} \xrightarrow{p} \theta_0$. As a result, as $I \to \infty$ the MLE $\pi_{\hat{p}}$ will be in a local neighborhood of the true π_{p_0} with probability tending to 1, and the MLE $(\hat{p}, \hat{\theta})$ will be coordinate-wise an asymptotically normal estimator of (p_0, θ_0) in that local neighborhood. ■

A.3.3 Additional Notes on Bootstrap Validity

This work proves coordinate-wise, local asymptotic normality of the vector-valued MLE. Thus, only asymptotically accurate marginal coverage in guaranteed. As a result, boot-

strapped confidence intervals may suffer from either overcoverage or undercoverage when applied jointly. To understand why, we can think of the marginal confidence intervals jointly creating a confidence region that is a $(J+1)$-dimensional hypercube, as opposed to a $(J+1)$-dimensional confidence ellipse that could be created via a true joint analysis. Overcoverage may occur if the confidence hypercube contains as a subset the (theoretical) confidence ellipse. But if each coordinate of the confidence hypercube is independent, joint coverage becomes a multiple testing problem and may result in undercoverage. That said, the present results do not preclude the possibility that bootstrap uncertainty estimates provide asymptotically correct coverage in the joint setting. We find no evidence to suggest asymptotically correct joint intervals are invalid. Further research may demonstrate proper coverage in the joint setting by establishing joint asymptotic normality of the $(J+1)$-dimensional MLE in a local neighborhood of π_{p_0}.

A.4 Additional Application Results from Section 3.4

A.4.1 Mean-Variance

Figure A.1 displays the relationship between the sample mean and variance of real grant panel review ratings to test the appropriateness of the Binomial rating model. We find that the sample variances are roughly centered around their theoretical values, and thus suggest that the Binomial rating model is appropriate.

A.4.2 Kendall Distances

We also examine model fit by calculating the Kendall distance between each judge's partial ranking with the estimated consensus ranking. In the right panel, we notice that for most judges, their partial ranking is mostly aligned with the MLE of the consensus ranking. The outlier in the right panel of Figure A.2 corresponds to a judge who assigned top-6 rankings to three proposals with comparatively poor ratings (proposals 3, 8, and 16).

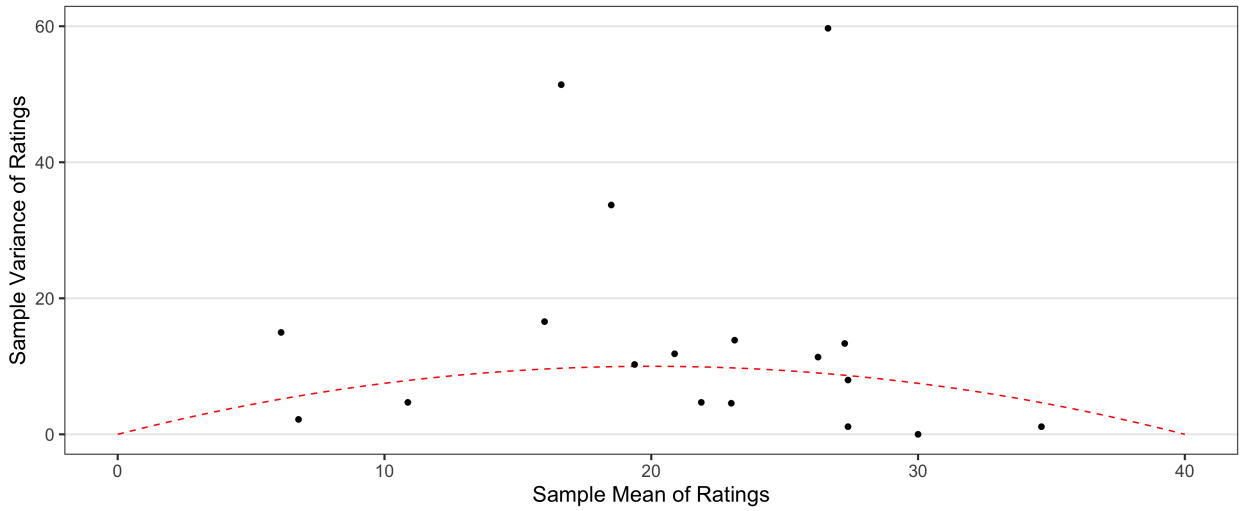

Figure A.1: Sample variance vs. sample mean of ratings by proposal (black circles) and the theoretical mean-variance relationship in a Binomial distribution (red dotted line).

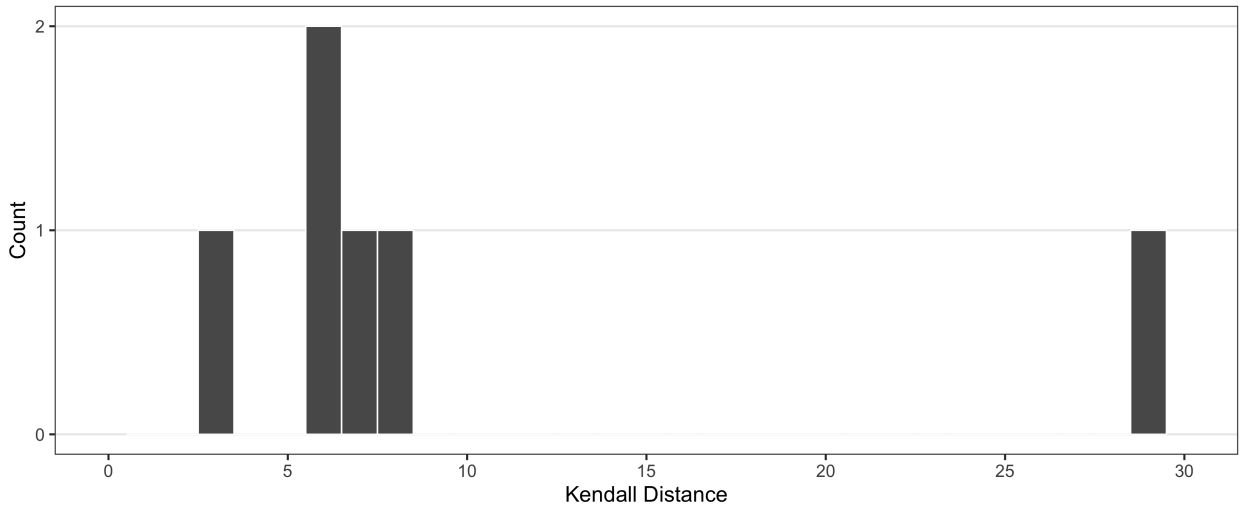

Figure A.2: Histogram of the Kendall distance between each judge's partial ranking with the estimated consensus ranking.

Appendix B

This is the Appendix to Chapter 4. It includes additional details on BTL-Binomial estimation algorithms (under the MFM framework, fixed K, and MAP estimation under fixed K) and further material from the three applications studied in this chapter.

B.1 Additional Details on BTL-Binomial Estimation Algorithms

B.1.1 Algorithm 1: Telescoping Sampler for BTL-Binomial MFM Model

- **Selection of Algorithm Parameters:** B^{Gibbs} is the number of times steps 2(a)-(d) will be repeated, which corresponds to the number of unique times Z, K^+, K, γ, and π are re-sampled. B^{MH} is the number of times within each Gibbs iteration that each set of class parameters (p_k, θ_k) will be re-sampled using Metropolis-Hastings. Setting $B^{\text{MH}} > 1$ often improves algorithm efficiency since class labels tend to stabilize over the course of a chain and hence need not be resampled each time that the class parameters are updated. K_{start} is the initial number of classes, which may be chosen to be any integer between 1 and I. We find that setting $K_{\text{start}} = I$ often yields quickly converging chains but comes with a computational cost when I is large in the first few Gibbs iterations as classes collapse. Lastly, the Metropolis-Hastings proposal variance parameters $\sigma_p^2, \sigma_\theta^2$, and σ_γ^2 should be chosen such that the acceptance probabilities are reasonably efficient (see Gelman et al. (2013) for further information on the efficiency of Metropolis-Hastings). Tuning these parameters after a short initial test run can be useful.

- **Step 1 (Initialization):** One may initialize γ, π, p, and θ by sampling these quantities from their prior distributions. For greater efficiency, one may also initialize using MAP

estimates for a prespecified choice of K. If multiple chains are run, they should have unique initializers.

- **Step 2(b) (Update Non-Empty Class Parameters):** Due to the assumed independence structure of observation from each class, the parameters from each of the non-empty classes can be updated independently (and thus in a parallel manner to for computational efficiency, if desired). We propose sequential updated of each of the $J+1$ specific parameters in (p_k, θ_k) for simplicity.

- **Step 2(c) (Update K and γ):** The probabilities found in Step 2(c) are taken directly from Frühwirth-Schnatter et al. (2021), Algorithm 3, which includes details of their derivation.

B.1.2 Algorithm 2: Gibbs Sampler for BTL-Binomial Latent Class Mixture Model

- **Selection of Algorithm Parameters and Initializers:** See discussion in Appendix Section B.1.1 for advice on selecting the algorithm parameters specific to Algorithm 2.

B.1.3 Algorithm 3: MAP Estimation via EM in the BTL-Binomial Latent Class Mixture Model

- **Data Log Likelihood and Objective Function:** To aid intuition for the EM algorithm presented for MAP estimation, we state the log likelihood after augmentation with latent classes and the objective function we seek to maximize: Assume the model presented in Equation 4.3. Let $z_{ik} = I\{Z_i = k\}$, where $I(\cdot)$ is the indicator function. Then, the log likelihood of the preference data augmented with class indicators Z can be written,

$$\mathcal{L}(\Pi, X, Z|\pi, p, \theta) \tag{B.1}$$
$$= \prod_{i=1}^{I}\prod_{k=1}^{K}\left(\pi_k \times \prod_{r=1}^{R}\frac{e^{-\theta_k p_{\pi_i(r)k}}}{\sum_{j\in\mathcal{S}} e^{-\theta_k p_{jk}} - \sum_{s=1}^{r-1} e^{-\theta_k p_{\pi_i(s)k}}} \times \prod_{j=1}^{J}\binom{M}{x_{ij}} p_{jk}^{x_{ij}}(1-p_{jk})^{M-x_{ij}}\right)^{z_{ik}}.$$

We seek MAP estimates $(\hat{\gamma}, \hat{\pi}, \hat{p}, \hat{\theta})$ defined as,

$$\hat{\gamma}, \hat{\pi}, \hat{p}, \hat{\theta} = \underset{\gamma,\pi,p,\theta}{\arg\max} \Big(\log \mathcal{L}(\Pi, X, Z|\pi, p, \theta) + \log f(\gamma, \pi, p, \theta) \Big), \tag{B.2}$$

where $f(\gamma, \pi, p, \theta)$ specifies the joint density of the prior distributions on γ, π, p, and θ.

- **Selection of Algorithm Parameters:** Algorithm 3 includes a single algorithm parameter, the convergence tolerance, *tol*. The tolerance should be a small, positive constant such as 10^{-3}. Alternatively, one may specify *tol* as a percentage and repeat the EM iterations until the absolute change in the model log likelihood between successive iterations falls less than the pre-specified tolerance.

- **Step 2(b)(ii) (Update γ):** At each stage of the EM algorithm, the current estimate of γ can be found via univariate numerical optimization. We use the function `optimize` in base `R` (R Core Team 2022).

- **Step 2(b)(iii) (Update (p_k, θ_k)):** At each stage of the EM algorithm, the current estimates of (p_k, θ_k) can be found independently for each k via multivariate numerical optimization. We use the function `optim` in base `R` with the method L-BFGS (R Core Team 2022; Byrd et al. 1995).

- **Obtaining Frequentist Maximum Likelihood Estimators:** To obtain frequentist maximum likelihood estimators, hyperparameters are available for each prior distribution. Setting $\xi_1 = 1$ and $\xi_2 = 0$ yields a flat but improper prior on γ. Alternatively, one may set $\gamma = 1$ and remove step 2(b)(ii). Setting $a = b = 1$ yields a flat and proper prior on each p_{jk}, and setting $\gamma_1 = 1$ and $\gamma_2 = 0$ yields a flat but improper prior on each θ_k.

B.2 Additional Application Results from Section 4.4

B.2.1 Paper Selection in Large Academic Conferences

Bias and Consistency of MAP Estimates We present the bias and consistency of MAP estimates in order to demonstrate good statistical properties of the BTL-Binomial model. We examine bias in Figure B.1 and consistency in Figure B.2.

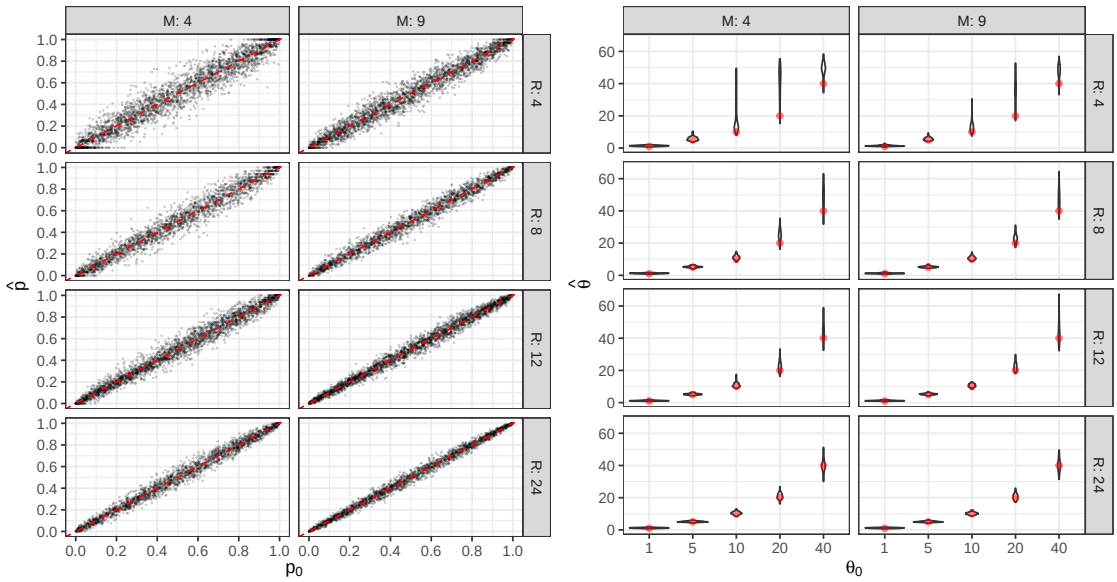

Figure B.1: Scatterplot of MAP estimates $\hat{p}$ against true p_0 (left) and violin plots of MAP estimates $\hat{\theta}$ against true θ_0 (right) under various M and R.

We observe in the left panels of Figures B.1 and B.2 that estimation of p appears to be unbiased and consistent in both M and R, regardless of θ_0. Estimation of p is central for understanding the quality of each paper and consensus ranking of the reviewers. Thus, the apparent unbiasedness and consistency of $\hat{p}$ suggest accurate estimation of group preferences regardless of M, R, and θ_0. For θ, the right panels demonstrate potentially biased estimation. This is unsurprising given a similar result for the corresponding parameter in the related Mallows-Binomial model from Chapter 3. Estimation accuracy appears worst when θ_0 is

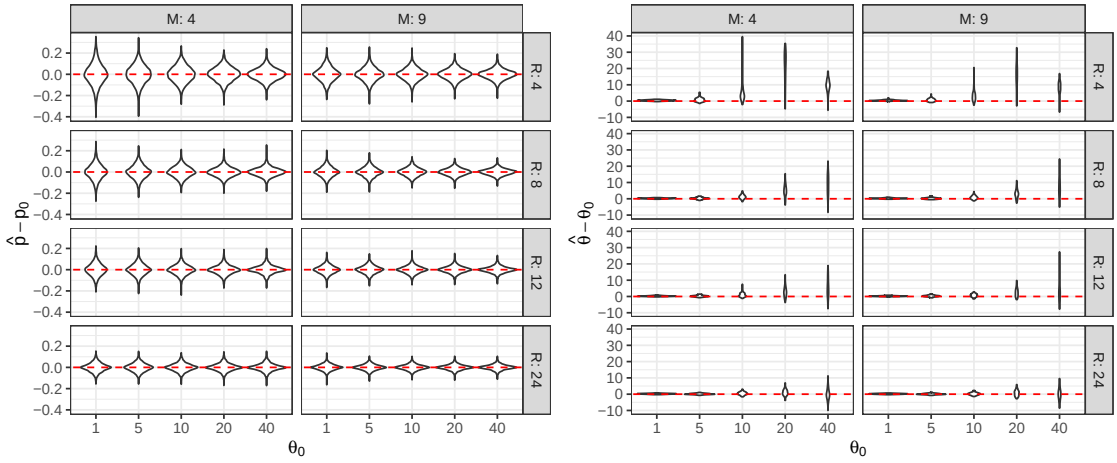

Figure B.2: Violin plots of estimation error for $\hat{p}$, calculated $\hat{p} - p_0$ (left) and estimation error for $\hat{\theta}$, calculated $\hat{\theta} - \theta_0$ (right) under various M and R.

very large, which often leads to perfect uniformity of observed rankings. In such cases, estimation of θ is most difficult; estimation is more accurate when θ is small. We notice that $\hat{\theta}$ appears consistent in R but not M, which makes sense given that M only relates to the rating scale while θ is only applicable to ranking consistency. Although the potentially biased estimation of θ is disappointing, it may not have much practical impact in the present setting since it corresponds to the strength of consensus and not the relative or ordered preferences of the group, which are paramount for deciding which papers to accept to the conference.

Additional Simulation to Study the Relative Effects of R and I The simulation study from Chapter 4.4.1 demonstrated the accuracy of the BTL-Binomial model for fixed I and J as R, M, and θ varied. One conclusion from the study was that estimation accuracy increases as the number of papers reviewed by each judge, R, increases. This is an intuitive result since increasing R represents an increased sample size. However, we note that increasing the number of reviewers, I, also increases the sample size. Since I was kept fixed in the

simulation study, we cannot use the study to disentangle the relative effects of R and I.

We now conduct an additional simulation to study if and how accuracy of the BTL-Binomial model changes as R and I vary such that the total number of paper assessments, $I \times R$, is fixed. Specifically, we let $J = 50$ and fix $R \times I = 600$, and vary $I \in \{15, 25, 50, 100, 200, 300\}$ such that $R = 600/I$. Additionally, we vary $M \in \{4, 9\}$ and $\theta \in \{1, 5, 10, 20, 40\}$. In each simulation scenario, we draw $p_j \sim \text{Beta}(1, 1) \stackrel{d}{=} \text{Uniform}[0, 1]$, and subsequently draw data from a BTL-Binomial model with a single latent class. Unlike in the original simulation study, judges rank all R objects they are assigned. Then, we fit a BTL-Binomial model to the data using hyperparameters $a = 1, b = 1, \gamma_1 = 5, \gamma_2 = 0.25$, which were chosen to be diffuse. Each scenario is replicated 200 times.

First, we present the bias and consistency of MAP estimates across simulation scenarios in Figures B.3 and B.4, respectively. We notice similar patterns in estimation bias and

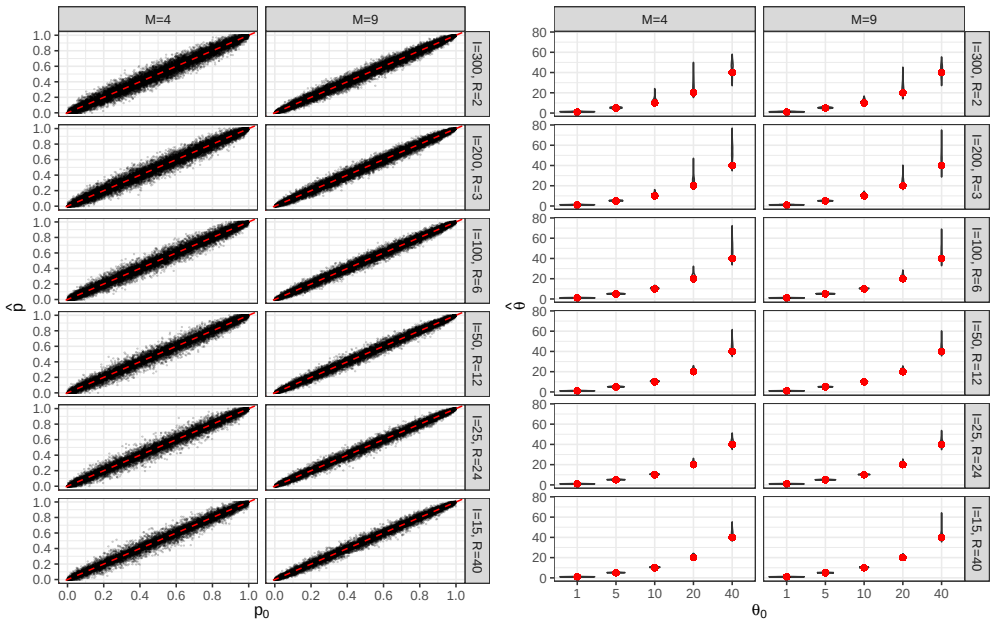

Figure B.3: Scatterplots of MAP estimates $\hat{p}$ against true p_0 (left) and violin plots of MAP estimates $\hat{\theta}$ against true θ_0 (right) under various I, R, and M such that $I \times R$ is fixed.

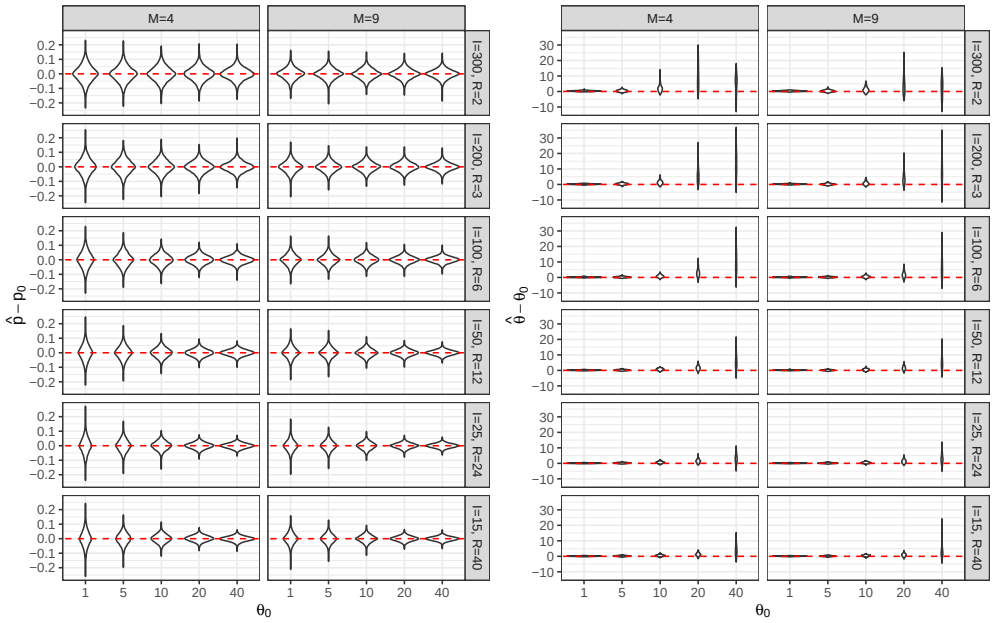

Figure B.4: Violin plots of estimation error for $\hat{p}$, calculated $\hat{p} - p_0$ (left) and estimation error for $\hat{\theta}$, calculated $\hat{\theta} - \theta_0$ (right) under various I, R, and M such that $I \times R$ is fixed.

consistency as in Figures B.1 and B.2 from the original simulation study. Regarding the relationship between R and I with respect to estimation accuracy, we notice that parameter estimates appear more accurate as R increases, even with commensurate decreases in I. The result is especially pronounced for the object quality parameter estimates, $\hat{p}$.

Second, we present the mean inaccuracy of the BTL-Binomial and ratings-only model when estimating the true consensus ranking of objects, π_0 in Figure B.5. The plot was created identically to Figure 4.1; see Section 4.4.1 for details. Again, we observe that the accuracy of the BTL-Binomial model tends to increase as R increases, even with commensurate decreases in I.

In conclusion, we find that given a fixed number of assessments across all judges and proposals, the accuracy of the BTL-Binomial model when estimating p, θ, and π_0 increases as the number of objects assessed by each judge, R, increases.

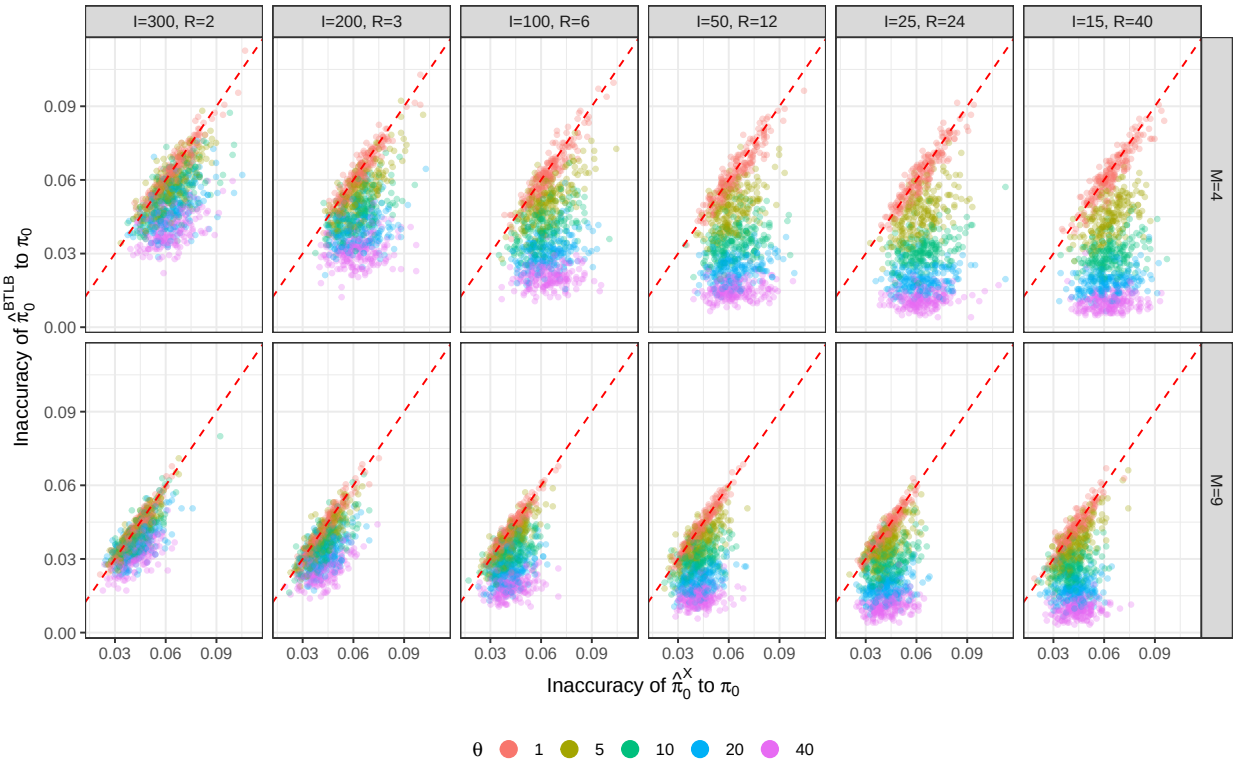

Figure B.5: Scatterplots of the mean inaccuracy across 100 simulations of estimated $\hat{\pi}_0^{BTLB}$ (BTL-Binomial model) and $\hat{\pi}_0^{\mathrm{X}}$ (standard ratings-only model) to the true ranking of papers π_0 under various I, R, and M such that $I \times R$ is fixed.

B.2.2 Proposal Selection in Grant Panel Review under Heterogeneity

Hyperparameter and Algorithm Parameter Settings Given the small sample size, we assign prior weight primarily to $K^+ \in \{1, 2, 3\}$. Thus, we choose $\lambda = 1$, $\xi_1 = 2$, and $\xi_2 = 3$. The effect of these choices on the prior distribution of K^+ can be seen in Figure B.6. We set $a = 2.50$ and $b = 3.77$ using an empirical Bayes approach, in which we fit a Beta distribution to the observed ratings after normalization to the unit interval based on maximum moment estimators of the first two moments. We set $\gamma_1 = 10$ and $\gamma_2 = 0.5$ to provide substantial

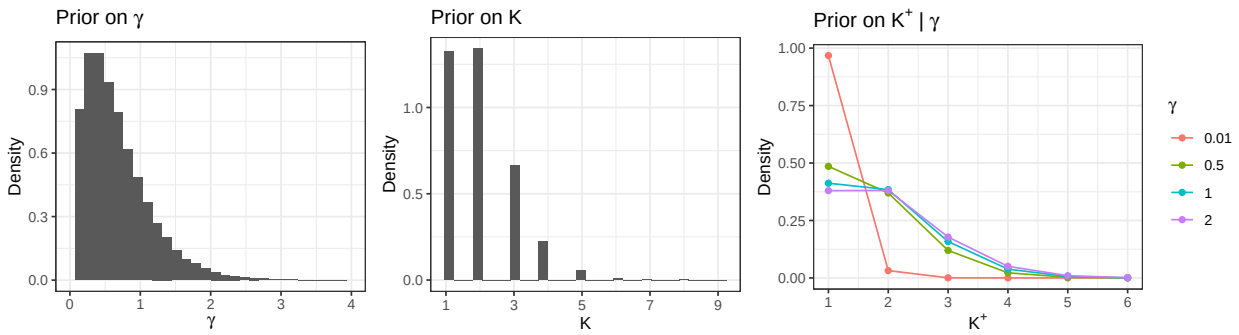

Figure B.6: Prior distribution on γ, K, and K^+ given $\lambda = 1$ and $\gamma \in \{0.01, 0.5, 1, 2\}$.

weight to values of $\theta \in [5, 35]$. We carry out Algorithm 1 with $B^{\text{Gibbs}} = 1000$ and $B^{\text{MH}} = 10$, $K_{\text{start}} = I = 17$, $\sigma_p^2 = 0.05$, $\sigma_\theta^2 = 3$, and $\sigma_\gamma^2 = .5$. The first half of the total 10,000 iterations were removed as burn-in.

Goodness-of-Fit and Trace Plots Figures B.7, B.8, and B.9 display goodness-of-fit and trace plots for Setting 2. We find the results to be satisfactory.

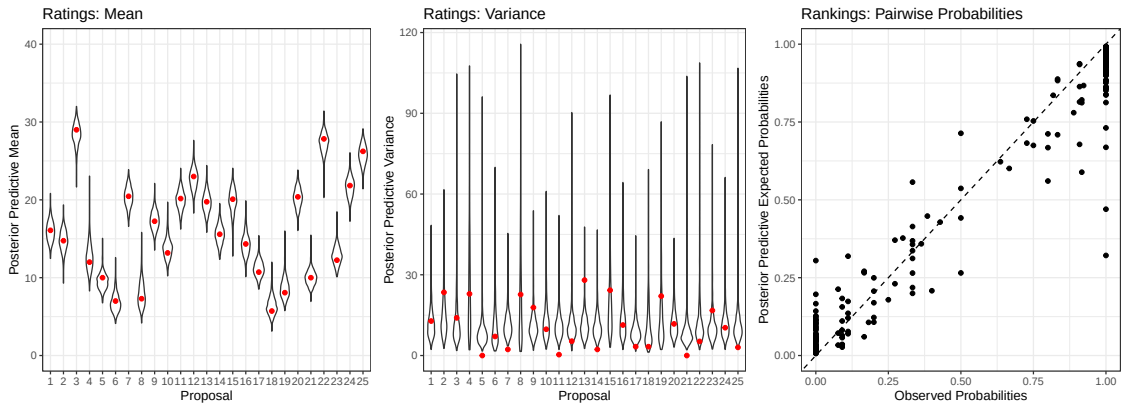

Figure B.7: BTL-Binomial MFM goodness-of-fit. Red dots represent the observed mean (left) or variance (center).

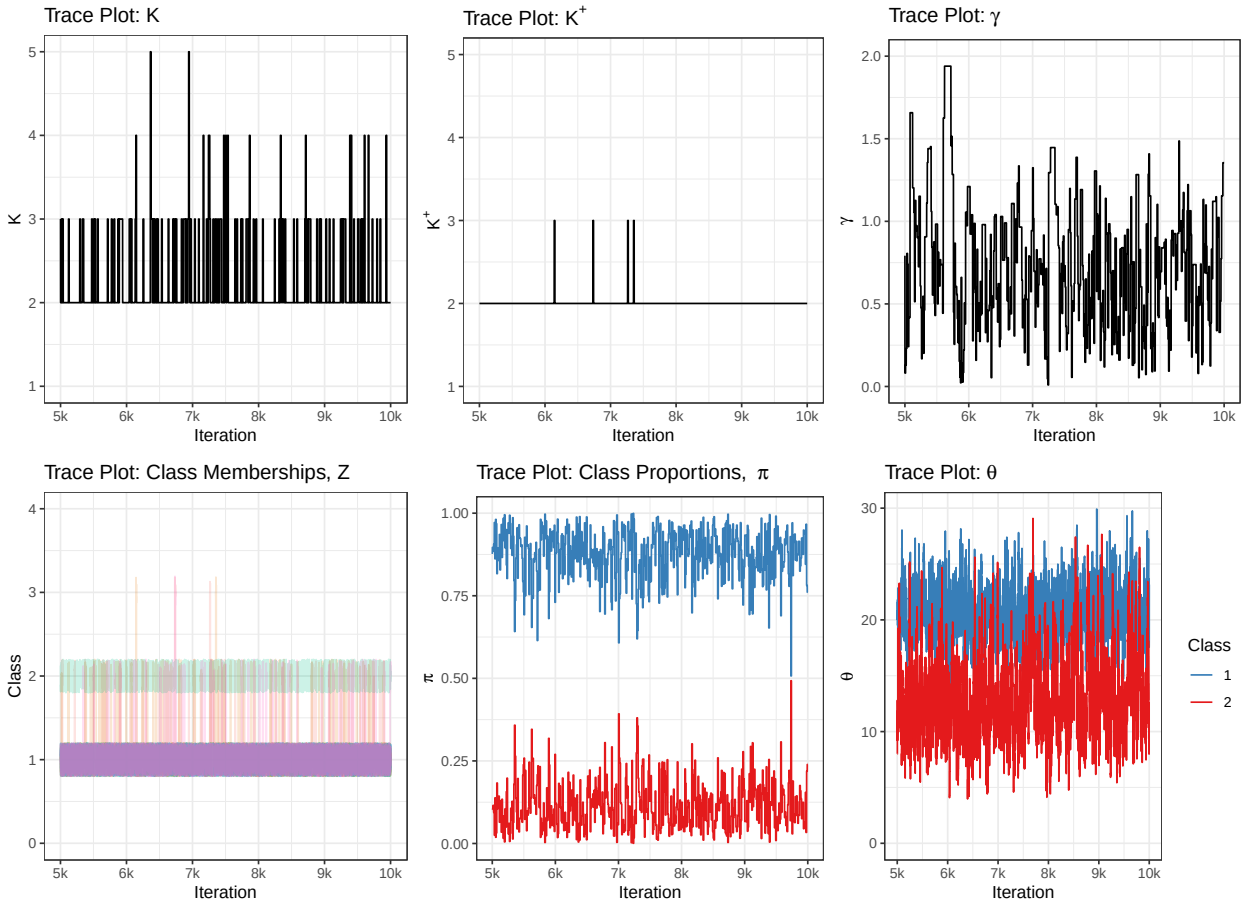

Figure B.8: Trace plots of K, K^+, γ, Z, π, and θ

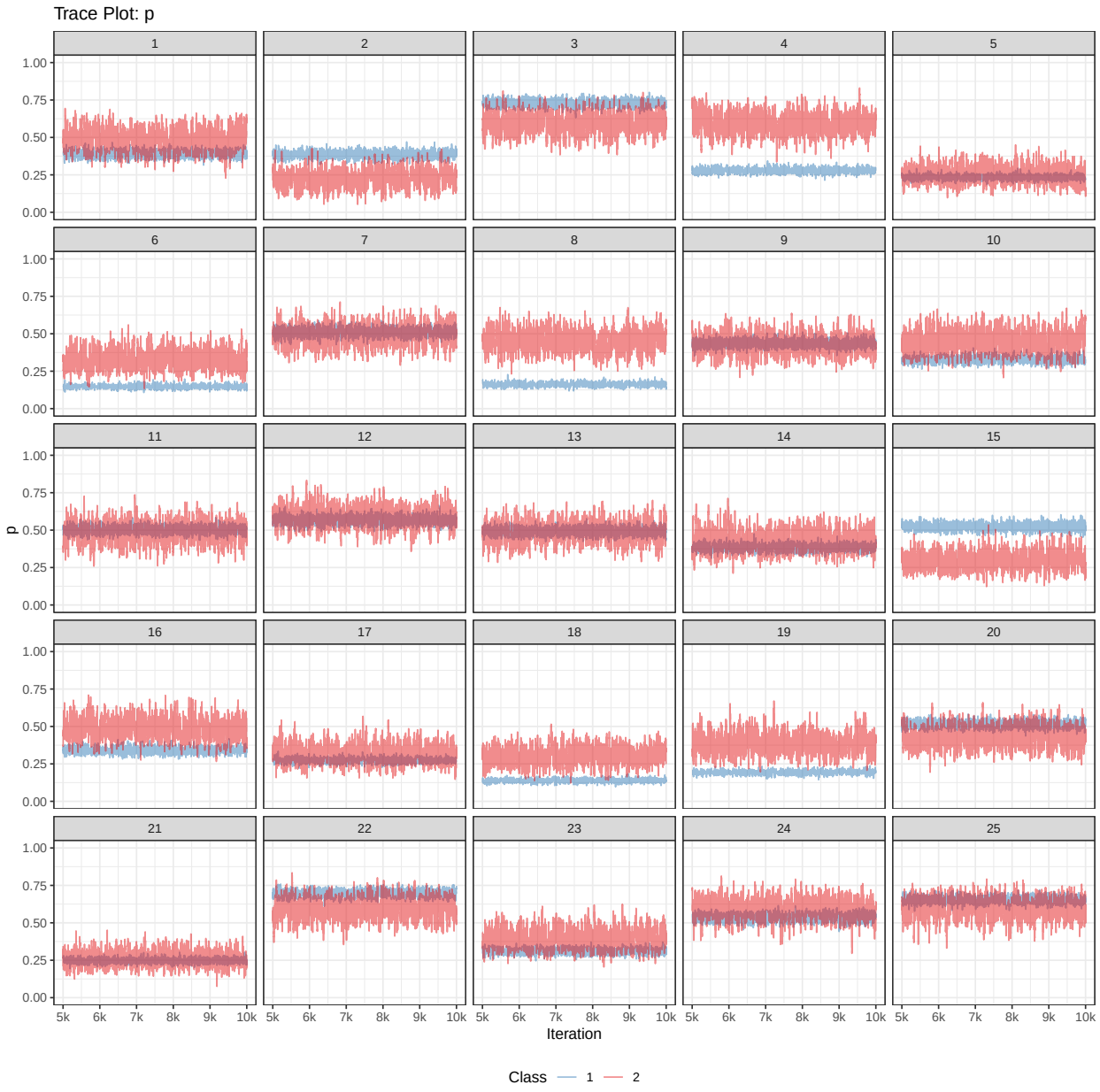

Figure B.9: Trace plots of p

B.2.3 Modeling Complex Survey Data under Heterogeneity

Hyperparameter and Algorithm Parameter Settings To aid interpretability given the large sample size, we assign prior weight primarily to $K^+ \in \{5, \ldots, 10\}$ using $\lambda = 7$, $\xi_1 = 3$, and $\xi_2 = 1$. The effect of these choices on the prior distribution of γ, K, and $K+$ can be seen in Figure B.10. We set $a = 0.26$ and $b = 0.77$ using an empirical Bayes approach, in which we fit a Beta distribution to the observed ratings after normalization to the unit interval based on maximum moment estimators of the first two moments. We set $\gamma_1 = 20$ and $\gamma_2 = 1$ to provide substantial weight to values of $\theta \in [10, 30]$. We carry out Algorithm 1 with $B^{\text{Gibbs}} = 35000$ and $B^{\text{MH}} = 2$, $K_{\text{start}} = 1$, $\sigma_p^2 = 0.05$, $\sigma_\theta^2 = 2$, and $\sigma_\gamma^2 = 2$. The first 20% of the 35,000 iterations were removed as burn-in.

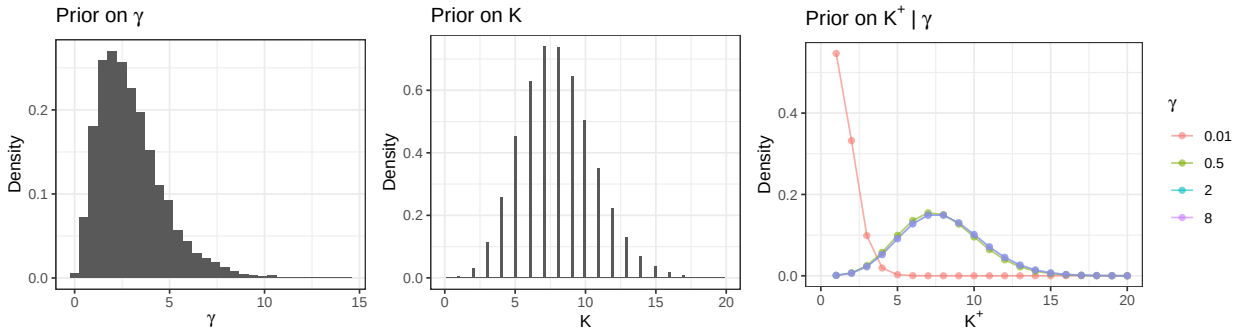

Figure B.10: Prior distribution on γ, K, and K^+ given γ.

Further Estimation Results Figure B.11 displays posterior summaries of K^+, γ, and π, where classes are ordered by size. We see that an 9-class model has very high posterior probability, which leads us to present results conditional on $K^+ = 9$.

Figure B.12 displays posterior probabilities of shared class membership across survey respondents. We do not display class membership probabilities by respondent due to the very large number of respondents. On the x- and y-axes are survey respondents, with order

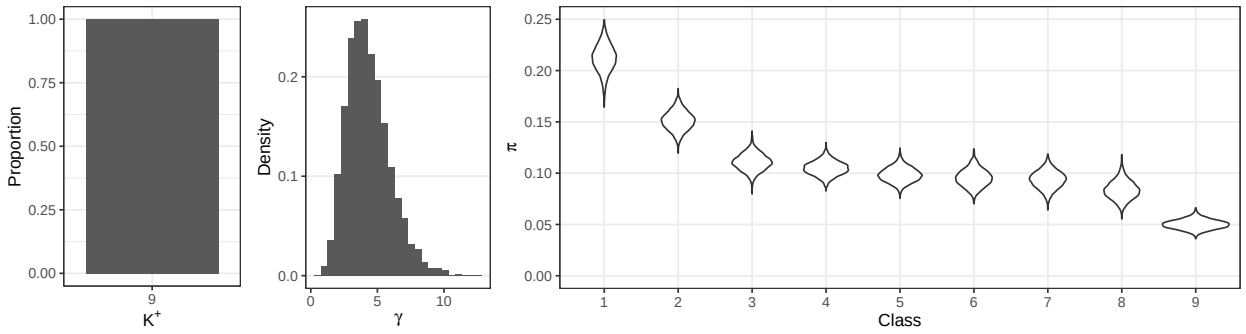

Figure B.11: Posterior distributions of $K+$ (*left*), γ (*center*), and π (*right*).

determined to keep similarly-clustered respondents together, as determined by the `salso` package (Dahl et al. 2021). The color indicates cluster similarity via the posterior probability of shared class membership (white represents low probability; black represents high probability). We notice high within-class homogeneity with respect to clustering probability and relatively strong heterogeneity between classes. We take this as evidence that the algorithm is successful at distinguishing heterogeneous groups.

Goodness-of-Fit and Trace Plots Figures B.13 and B.14 display goodness-of-fit and trace plots for Setting 3. We notice that posterior means for ratings and pairwise probabilities for rankings appear satisfactory. The posterior predictive ratings variance appears to be low in comparison to the observed ratings variance, which is likely a result of providing strong prior probability to a relatively small number of clusters. Given the focus on interpretability, we find the results to be satisfactory.

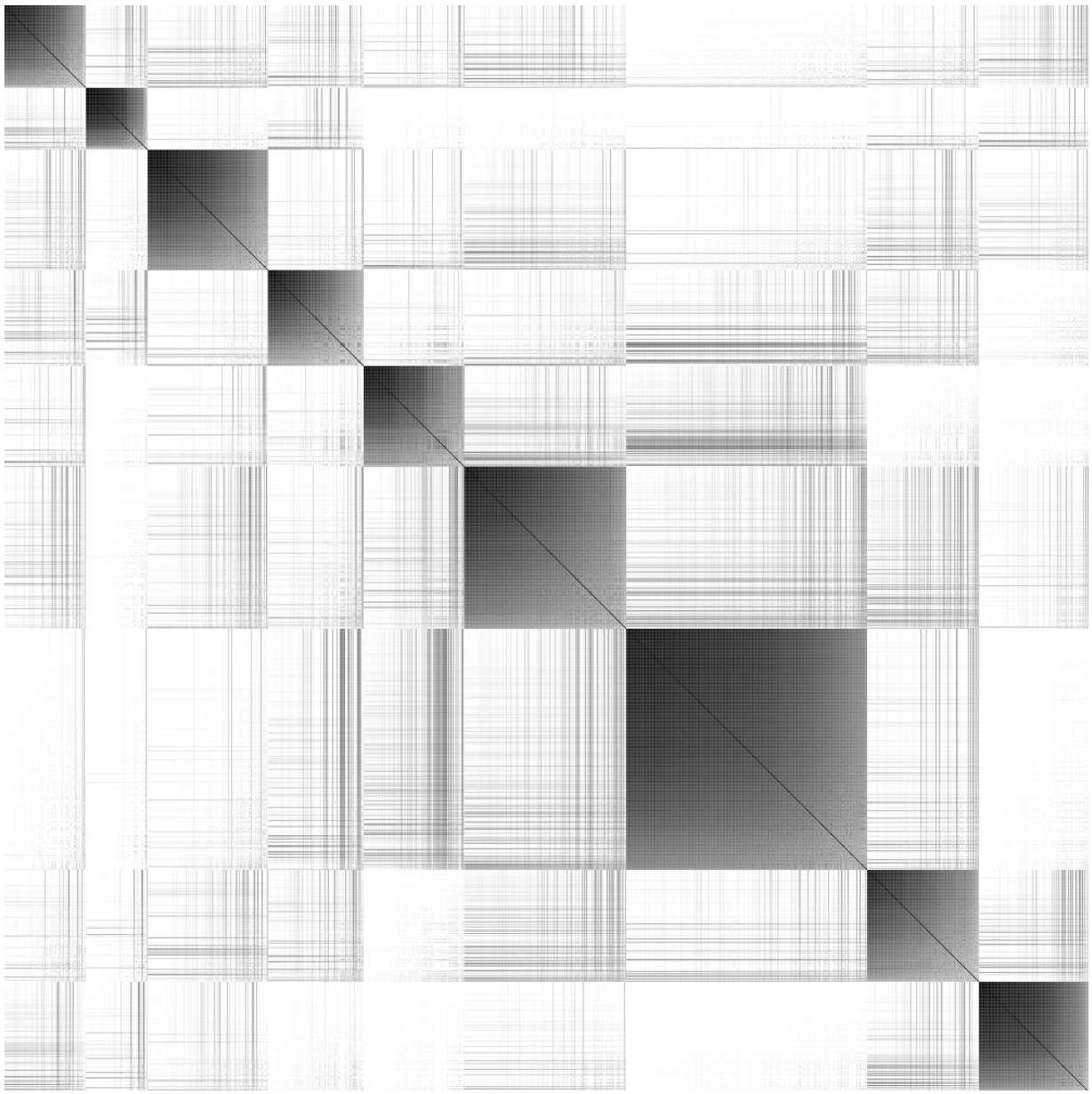

Figure B.12: Similarity matrix of survey respondents by shared class membership. Survey respondents are on the x- and y-axes, ordered to display clusters cohesively. White (black) represents low (high) posterior probability that two respondents are in the same class.

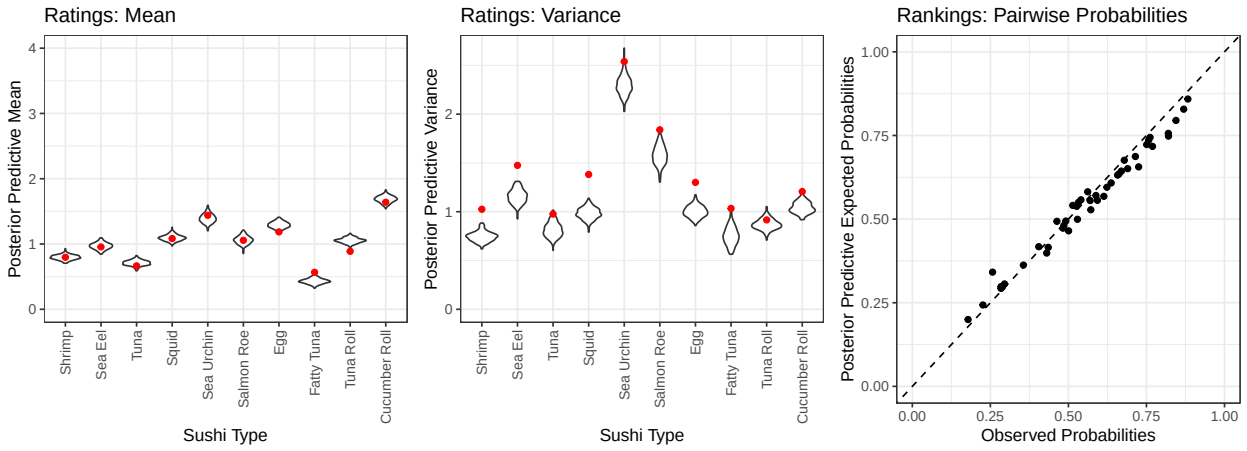

Figure B.13: BTL-Binomial MFM goodness-of-fit. Red dots represent the observed posterior mean (left) or variance (center).

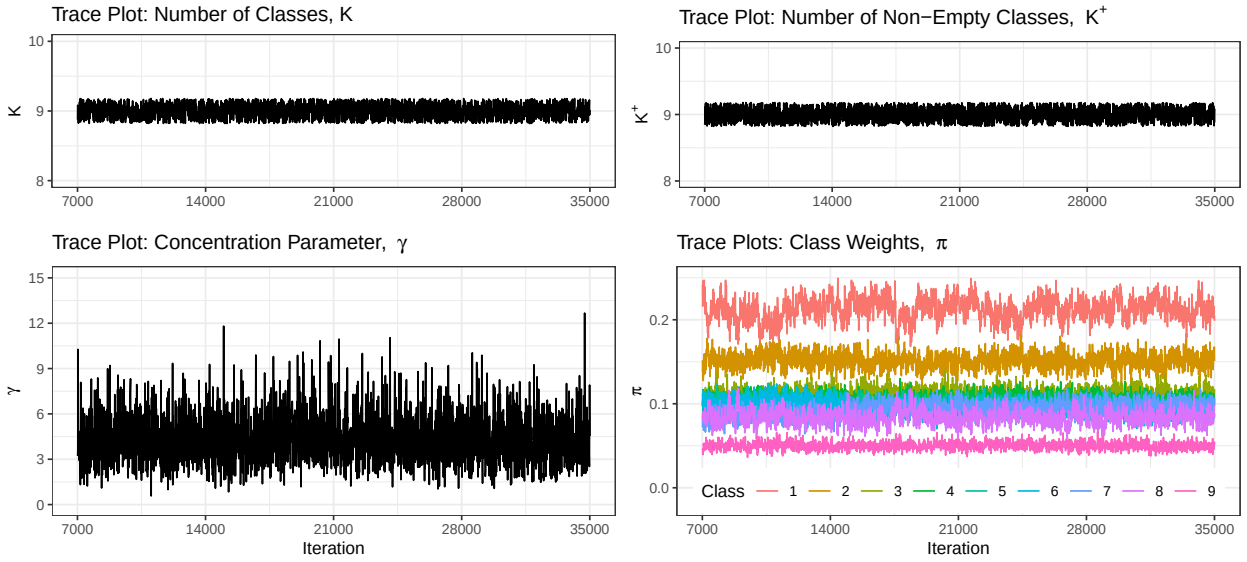

Figure B.14: Trace plots of K, K^+, γ, and π

Appendix C

This is the Appendix to Chapter 6. It includes additional results from our analysis of the 2021 Minneapolis mayoral election data.

C.1 Additional Application Results from Section 6.5

First, we visualize the prior density given to each partition, g, based on the $J = 17$ candidates and choice of Poisson hyperparameter $\lambda = 2$ in Figure C.1. We note that the prior gives some density to all partitions, regardless of K, even though there are significantly more partitions g with moderate K_g than those with K_g near 1 or 17.

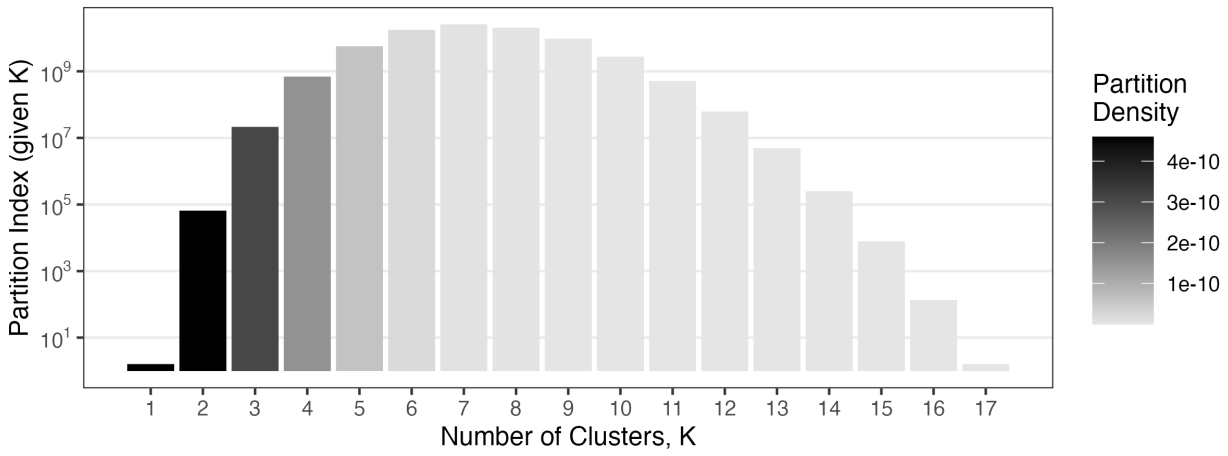

Figure C.1: Prior density on partitions in the 2021 Minneapolis mayoral election data.

Second, we display trace plots of K and ω across 10 independent MCMC chains and $\lambda \in \{1, 2, 4\}$ in Figures C.2 and C.3, respectively. By visual inspection, the chains seem

to have mixed and converged. Furthermore, results look very similar across values of λ, indicating robustness to our choice of prior.

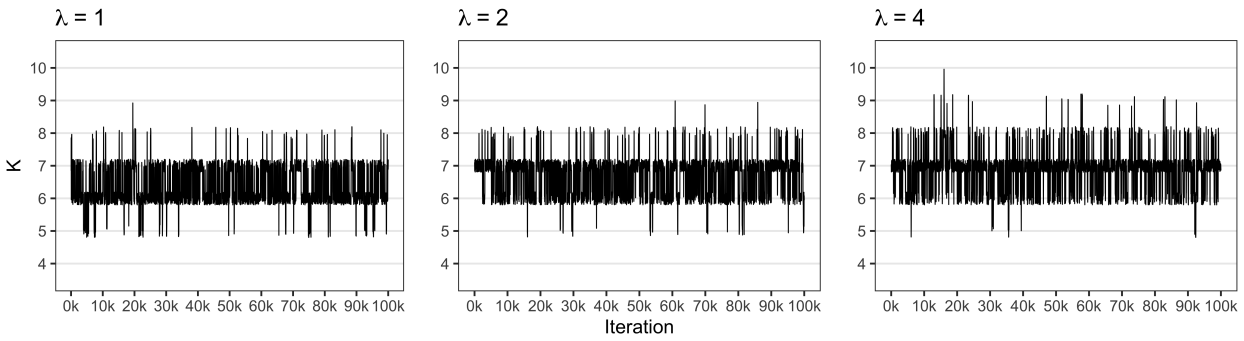

Figure C.2: Trace plots of the number of candidate clusters, K, by $\lambda \in \{1, 2, 4\}$. Every successive 10k iterations represents the final 10k iterations of an independent MCMC chain.

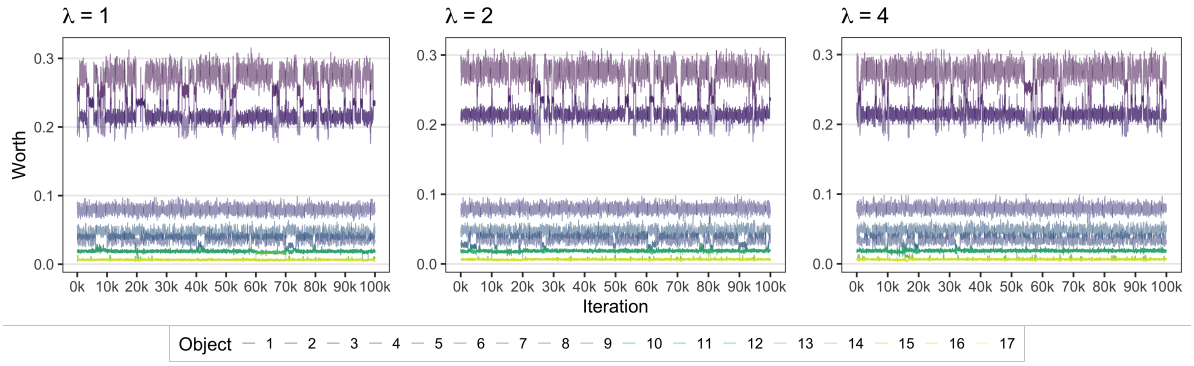

Figure C.3: Trace plots of normalized candidate worth parameters, ω, by $\lambda \in \{1, 2, 4\}$. Every successive 10k iterations represents the final 10k iterations of an independent MCMC chain.

Third, we view the candidate clustering matrices by λ. Again, the results are quite similar across alternative λ, indicating robustness to our choice of prior.

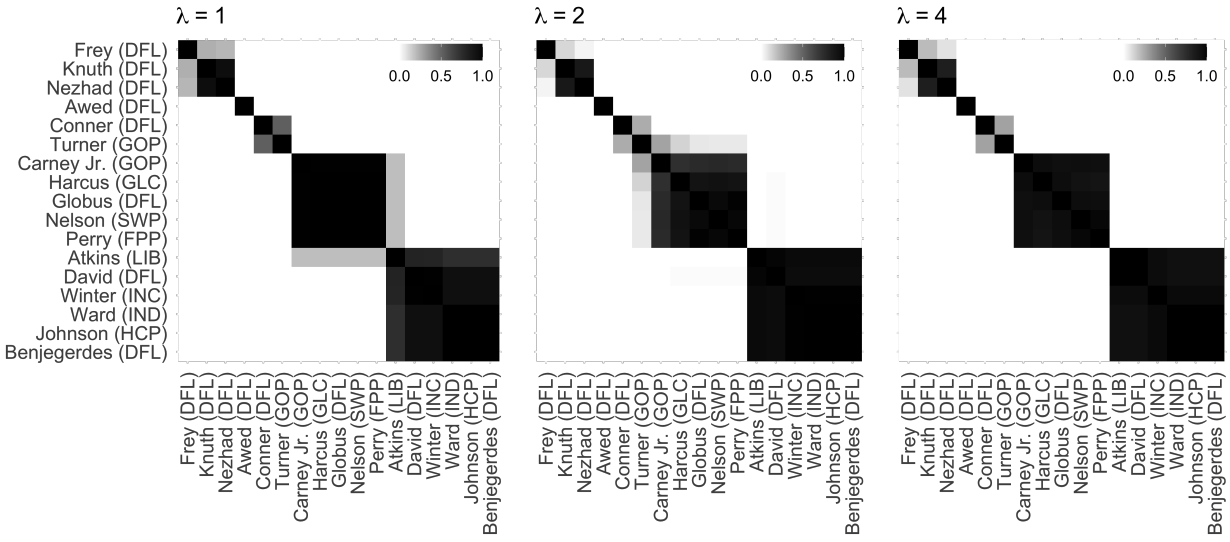

Figure C.4: Posterior candidate clustering matrices by $\lambda \in \{1, 2, 4\}$

VITA

Michael Pearce is a proud Minnesotan who enjoys winter activities, choral singing, and hotdish (*uff da!*). He earned a Bachelor of Arts in mathematics from St. Olaf College before beginning his doctoral studies at the University of Washington. He has worked in private industry as a statistical consultant at Deloitte and an applied statistician at Boeing. Additional information can be found on his website, https://pearce790.github.io.