Revisiting Random Utility Models

A dissertation presented

by

Hossein Azari Soufiani

to

The School of Engineering and Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Computer Science

Harvard University

Cambridge, Massachusetts

April 2014

Dissertation Advisor:
Professor David C. Parkes

Author:
Hossein Azari Soufiani

Revisiting Random Utility Models

Abstract

This thesis explores extensions of Random Utility Models (RUMs), providing more flexible models and adopting a computational perspective. This includes building new models and understanding their properties such as identifiability and the log concavity of their likelihood functions as well as the development of estimation algorithms.

A special case of RUMs that has received significant attention is the Luce model, for which there are fast inference methods for maximum likelihood estimation. This thesis introduces RUMs including those with exponential family utility distributions, mixture of RUMs, and non-parametric RUMs. Fast inference is achieved through the Monte-Carlo Expectation-Maximization (MC-EM) algorithm. Results on both real-world and simulated data provide support for the ability of these models to better capture heterogeneity in data and for scalable model estimation.

A class of Generalized Method-of-Moments (GMM) algorithms for computing parameters of the Luce model and RUMs is also proposed. The technique is based on breaking full rankings into pairwise comparisons, and then computing parameters that satisfy a set of generalized moment conditions. The conditions for the output of GMM to be unique are identified, leading to a class of pairwise consistent and inconsistent breakings. Theoretical and empirical results show that the algorithms run significantly faster than the classical Minorize-Maximization (MM) and MC-EM approaches, while achieving competitive statistical efficiency.

I propose two preference elicitation scheme for generalized RUMs, in which the utilities can also depend on attributes of agents and alternatives. An empirical study shows that the

proposed elicitation scheme increases the precision of estimation for a given number of queries relative to existing approaches.

Furthermore, a model for differentiated items is developed, where I interpret the data as representing preference orders expressed by a population of agents on items, and each agent and item is associated with attributes. I extend the mixture of RUMs method to this setting, with reversible jump MCMC techniques adopted to estimate the parameters of the model and classify agent types. I develop theoretical conditions that establish the uni-modality of the likelihood function and posterior. Empirical results on real and simulated data provide support for improved model fit relative to single type models and for the scalability of the approach.

Contents

List of Tables

List of Figures

0.1 List of Publications

The following is a list of papers published from the content of this thesis:

[11] H. Azari Soufiani, D. C. Parkes, L. Xia, Random Utility Theory for Social Choice: Theory and Algorithms, In Proceedings of Neural Information Processing systems Foundation NIPS 2012.

[12] H. Azari Soufiani, D. C. Parkes, L. Xia, Preference Elicitation For General Random Utility Models, In Proceedings of Uncertainty and Artificial Intelligence, UAI 2013.

[8] H. Azari Soufiani, H. Diao, Z. Lai, D. C. Parkes, Generalized Random Utility Models with Multiple Types, In Proceedings of Neural Information Processing Systems, NIPS 2013.

[10] H. Azari Soufiani, W. Chen, D. C. Parkes, L. Xia, Generalized Method-of-Moments for Rank Aggregation, In Proceedings of Neural Information Processing Systems, NIPS 2013.

[13] H. Azari Soufiani, D. C. Parkes, L. Xia Computing Parametric Ranking Models via Rank-Breaking, Accepted for publication in ICML 2014

Other published papers during my PhD:

[9] H. Azari Soufiani, D. J. Charles, D. M. Chickering, D. C. Parkes Approximating the Shapley Value via Multi-Issue Decomposition, Accepted for publication in AAMAS2014

[6] H. Azari Soufiani, E.M. Airoldi, Graphlets Decomposition of a Weighted Network, Proc. 15th International Conference on Artificial Intelligence and Statistics AIStats, 54-63, 2012.

Acknowledgments

My PhD work, the majority of which is presented in this thesis, has been a learning journey. The work in this thesis is the result of close collaboration and discussion with many exceptional researchers.

First and foremost, I would like to express my special appreciation and thanks to my advisor Professor David Parkes, who has been a legendary mentor for me in past three years. His unique vision and extraordinary attention directed me in an amazing path of research. David's selflessness and unconditional support in every part of this work made me realize that smartness and intelligence are not enough and taught me how to see the world from a positive perspective and find an opportunity to contribute. I was honored to have him as my advisor and truly enjoyed every moment of working with him. I would also like to acknowledge other members of Harvard's EconCS group which has been an amazing environment for me in the past few years and from whom I learned a lot.

My path through the Harvard Statistics Department has been a pivotal moment in my education and career. I started working on statistical problems with Professor Edoardo Airoldi and I would like to specially thank him for his encouragements and guiding me through a good conduct of research in statistics. Even though not included in this thesis, my work with Edo on Graphlets was my first research paper in my PhD and I learned about statistical approaches and technical presentation through this work.

I was fortunate to learn about probability theory and statistics from Professors Carl Morris and Joe Blitzstein, and I had the great opportunity to learn from other great statisticians at Harvard, such as Professors Don Rubin and Xiao-Li Meng, to name a few. In addition to Edo's effective role in my research, I should acknowledge the members of the Airoldi lab who had a great role in discussing and resolving problems during my research. Among members of this lab, Gábor Csárdi helped with our problems on developing an R package, and Simon Lunagomez helped by familiarizing me with Bayesian thinking and Bayesian experimental design, and methodologies such as reversible-jump MCMC, etc.

I would like to express my thanks to the wide range of researchers and students, with

whom I collaborated. I would like to thank Professor Lirong Xia, at the time a postdoc at at Harvard, who has been a great mentor for me in this thesis and in general in my research career. He had great patience in working with me and our collaboration has been very fruitful. William Chen, an undergraduate student at the time, has helped me and worked with me in developing the StatRank package and building some of the models. William is a to-watch person and he had an important impact on this work. I would also like to thank my other collaborators Hansheng Diao, Zhenyu Lai and Muxi Li.

I was fortunate to have a brilliant PhD committee. David and Edo were my primary and secondary advisors. Professor Ryan Adams has been very helpful along many discussions on my work and I have learned machine learning from him. Professor Greg Lewis has been helping us with the economics aspects of our research and the econometrics direction would not have happened without him. Max Chickering has many interesting works on rank data and experiments and provided very helpful feedback and comments along this work. I was fortunate to work with him for a summer.

I am grateful for receiving very constructive feedback and inspiration from Corinna Cortes, Daryl Pregibon, Sendhil Mullainathan, Devavrat Shah, James Evans and Yiling Chen.

I should thank the Siebel foundation for financially supporting my last year of PhD studies, which gave me flexibility and independence toward the development of this thesis.

I thank SAMSI institute for their support for multiple travels to the research triangle for attending various inspiring workshops.

The School of Engineering and Applied Sciences at Harvard will remain an important part of my career, where I learned a lot, not only about science and engineering but also about other aspects of a good career such as ethical and moral conduct.

I wish to finish this part with special thanks to my family for their ever-lasting effect. Words cannot express how grateful I am to my father, mother, sister and my fiancée Elham who has been standing with me and behind me in this journey. Their love and support is immeasurable.

To my mother and father.

Chapter 1

Introduction

Human behavior is identified with the actions that people take, these actions generally coming about through the choices that we make [5, 43]. As a result of this, understanding human choice is an essential problem and one studied across many fields. As Donagan [43] explains, the problem of understanding choice dates back to Socrates and Aristotle, who viewed choice as being based on wishes and beliefs.

However, it was not until the nineteenth century that we started to develop a quantitative understanding of human choice. Ernst Heinrich Weber, known as one of the founders of experimental psychology, developed a framework, known as Weber's law, to connect psychological events to physical stimulus values that can be measured. These values are supposed to be the backbone of psychological events such as choices. Weber's work emphasizes the existence of a linear physical relation between a stimulus and sensation (such as force and acceleration). Weber's work was continued by his student Gustav Theodor Fechner, leading to a more accurate framework known as Fechner's law. However, the generality of Weber's and Fechner's theories was criticized in the late nineteenth century by William James, who argued that sensation is a rather complex function of multidimensional stimuli.

In his seminal work in the early twentieth century, Thurstone [111] formalized the law of comparative judgment, building from Weber and Fechner's theory. Moreover, he assumed that the stimulus values have a random component which he modeled as a Normal distribution.

Thurstone built different scenarios for the distribution of psychological stimuli which led to the Thurstone's model for pairwise comparisons.

Thurstone continued applying variations of his model to different settings and showed the generality of his model for human choice platforms. Even though Thurstone's model was appealing and explanatory, the estimation approaches were not as flexible as the model itself, and this led to the partial failure of his model in empirical studies.

With the start of the mathematization era in economics from the middle of the twentieth century [42], choice theory also started to grow in the direction of axiomatic models. Von Neumann and Morgenestern formalized the notion of random utilities and the existence of expected utilities that can capture a choice set under reasonable axioms [117].

Along the same lines, Luce provided a choice axiom that led to his model of choice [76, 77]. Luce's axiom led to a model that was easier to fit to data from experimental studies than Thurstone's models and found significant applicability. A pairwise version of Luce's model was proposed by Bradley and Terry [29] for the analysis of data from block design in statistical experimental design. Moreover, Bradley [28, 27] provided the relation between this model and Thurstone's setting. Adams and Messik [2] provided axiomatization for the Thurstone's setting, and Block and Marschak [23] argued for the value of the random component in Thurstone's model.

The relationship between the axiomatic approach of Luce's and Thurstone's model is established in Yellott's work [123]. Yellot shows that Luce's model uniquely satisfies Luce's axiom and Thurstone's comparative law with independent random components.

Even though the axioms provided important support for research in the modeling of choice, the rise in computation power and move toward empirical economics in the late twentieth century brought about a new shift. One revealing comment is from Plackett [99], where he criticizes both Thurstone's and Luce's models for under-parameterizing the space of observations. He proposes an over-parameterized model to overcome this issue along with an estimator for his extended model. Ironically, Plackett later gets his name on Luce's model following a book by Marden [80]. Plackett was concerned with the complexity of data and

the need for more complex models. However, his work seems to be under-noticed because of the simultaneous developments in econometrics mainly by McFadden, on formalizing Random Utility models [83].

In the economics, the psychophysical stimulus values from Thurstone's setting are viewed as utilities and it is assumed that choice makers are maximizing their utilities. McFadden generalized Luce's setting by representing the parameters of Luce's model as a function of the characteristics of alternatives and agents (who make the choice), allowing for more flexibility in capturing complexity in data. The resulting model is called the multinomial logit model (MNL). This direction was very successful since it took advantage of the simplicity in Luce's model and also built an explanatory component into the model that helped with econometrics research [87], earning McFadden a Noble prize (The Sveriges Riksbank Prize in Economic Sciences) of economic sciences in 2000 [86].

McFadden's MNL model was generalized to Nested MNL, generalized extreme value (GEV) models, and the multinomial probit model to overcome some limitations of the MNL model. Furthermore, a mixed multinomial logit model (MMNL) has been shown to be capable of approximating any reasonable RUM model [88]. As explained by McFadden, the extensions to the RUM framework have limits because of computational issues.

RUMs remain a very large set of models, of which only a small fraction are tractable. In McFadden's own words [86]:

Looking back at the development of discrete choice analysis based on the RUM hypothesis, I believe that it has been successful because it emphasized empirical tractability and could address a broad array of policy questions within a framework that allowed results to be linked back to the economic theory of consumer behavior.

Some possibilities for development of the approach have not yet been realized. The RUM foundation for applied choice models has been only lightly exploited. Models have generally conformed to the few basic qualitative constraints that RUM imposes, but have not gone beyond this to explore the structure of consumer preferences or the connections between economic decisions along different dimensions and in different areas. The potentially important role of

perceptions, ranging from classical psychophysical perception of attributes, through psychological shaping of perceptions to reduced dissonance, to mental accounting for times and costs, remains largely unexplored in empirical research on economic choice. Finally, the feedback from the empirical study of choice behavior to the economic theory of the consumer has begun, through behavioral and experimental economics, but is still in its adolescence.

What lies ahead? I believe that the basic RUM theory of decision-making, with a much larger role for experience and information in the formation of perceptions and expression of preferences, and allowance for the use of rules as agents for preferences, can describe most economic choice behavior in markets, surveys, and the laboratory. If so, then this framework can continue for the foreseeable future to form a basis for microeconometric analysis of consumer behavior and the consequences of economic policy.

Even though McFadden expresses hope for research on more complex RUMs, econometrics research in the last decade has mainly focused on the applications of the MMNL model and new estimators for MMNL model extensions based on methods such as the EM algorithm [113, 115, 114]. From the statistical perspective we see a continuing interest in building new estimators for the Luce model such as the minorize-maximize algorithm [67], fixed point estimators for Bradley-Terry Model [100], and rank-centrality algorithm [96].

This motivates the research presented in the present thesis in revisiting the vast set of RUMs from a computational perspective and providing a general framework to estimate and develop inference methods for flexible RUMs that are well suited to choice behavior. Furthermore, this research explores the computational and statistical efficiency trade-offs between different models, and provides a better understanding of the benefits of different estimators.

In terms of new applications, there are many new domains that provide choice data, and the richness of the data is considerably greater than in the classical econometrics setting. The goal is to be able to extend the existing RUM framework to provide a general and powerful methodology that can be used in settings such as crowd-sourcing, online search, and online marketing, in addition to classical econometric applications.

The following provides an overview of the contributions in each chapter. I begin with

presenting a general approach for RUMs, including parametric and non-parametric models where observations can be full rankings or any form of partial ranking on the choice set. An estimator based on the Monte-Carlo-EM (MC-EM) algorithm is developed for general RUMs. Moreover, three different model specifications are studied. The first specification is a RUM with exponential family distributions. The second specification is a mixture of general RUMs, and the third specification adopts a non-parametric joint utility distribution through kernel density estimators on latent utility scores. For each model, theoretical properties such as identifiability and log-concavity of the likelihood functions are studied. Empirical results establish scalability and efficiency on different datasets. Flexible exponential family distributions, such as Normal distribution with a variance parameter, perform better than classic models such as Luce's. Moreover, mixture models provide interpretable groups of agents, and non-parametric models introduce a higher predictive power for applications such as rank completion.

The second chapter pursues a different set of estimators using generalized method of moments (GMM) techniques and builds a theory for estimators defined on pairwise data generated by breaking full-rank observations. This theory includes a new characterization of consistent moment-based estimators results for Luce's model and other RUMs. Empirical results confirm that the GMM approaches are much faster than MC-EM and achieve comparable quality of fit.

The third chapter extends the results in the first chapter to settings where we observe agent and alternative characteristics along with rank data. Furthermore, it provides a method for selecting which agent to elicit ranks from, based on maximizing the gain in expected information. This approach uses the Bayesian experimental design framework. The results show that classical optimality methods such as D-optimality and E-optimality will sometime perform worse than random elicitation. Hence, a new metric is proposed for eliciting rank data, providing better performance in comparison with existing approaches.

The fourth chapter continues the setting in Chapter Three where we observe agent and alternative characteristics and estimate a mixture model based on characteristics on any RUM

extending McFadden's mixture model on MNL model. Identifiability of mixture models is studied, and experimental results demonstrate that a model with multiple types performs better than single-type models.

The final chapter offers some conclusion and suggestions for future work.

The chapters in this thesis have been written to be largely self-contained with minimum cross-references to other chapters.

Chapter 2

Random Utility Theory for Rank Data

2.1 Introduction

A lot of different kinds of data takes the form of rank ordering on alternatives. For examples, rank data from sports competitions, consumption data in markets, elections, meta search and crowd-sourcing applications that use user judgments. Rank data presents an interesting and challenging machine learning problem, because of the factorial size of the rank space. For example, finding an optimal ranking by searching over the whole space of ranking is computationally difficult.

Learning to rank [72] and the adoption of probabilistic models for rank aggregation in social choice [41, 39, 122, 121, 105, 103] are gaining in prominence in recent years. In part, this is due to the explosion of socio-economic platforms, where opinions of users need to be aggregated; e.g., judges in crowd-sourcing contests, or the ranking of movies or user-generated content. Moreover, rank aggregation problems exists in determining the winners of tournaments [67], aggregating search rankings into meta-search results [44], and declaring the winner of an election [51]. Problems of social choice and the aggregation of opinions occur in many other settings as well, for example in peer reviewing and committee work.

In the problem of rank aggregation, we are given ranks over m alternatives from n agents and a single rank order must be selected to be representative of the data. Rank data comes

in many forms. It may consist of full ranks where each observation is a full rank order. It may consist of partial orders, for example when each observation or agent provides ranks on a subset of alternatives (e.g. games, competitions, races) or provides only top preferences out of a set of alternatives (e.g. candidates in elections).

Since Condorcet [37], one approach to rank aggregation has been to formulate rank data as the problem of estimating a true underlying world state (e.g., a true quality ranking of alternatives), where the individual reports are viewed as noisy data in regard to the true state. In this way, the problem can be framed as a problem of inference. Condorcet assumed the existence of a true *ranking* over alternatives, with a agents's preference between any pair of alternatives a and b generated to agree with the true ranking with probability $p > 1/2$ and disagree otherwise.

Condorcet proposed to choose as the outcome of social choice the ranking that maximizes the likelihood of observing the agents' preferences. Later, Kemeny's rule was shown to provide the maximum likelihood estimator (MLE) for this model [124]. But Condorcet's probabilistic model assumes identical and independent distributions on pairwise comparisons. This ignores the strength in agents' preferences (the same probability p is adopted for all pairwise comparisons), and allows for cyclic preferences. In addition, computing the winner through the Kemeny rule is Θ_2^P-complete [62], which is generally thought to be computationally intractable.

To overcome the first criticism regarding the existence of cycles, a vast literature adopts probabilistic methods to model rank data. Parametric probabilistic modeling of rank data in the form of ranking of alternatives dates back to Thurstone's model [111]. Mosteller elaborated on Thurstone's models for pairwise data [93] and later Bradly, Terry et al. [30] considered analysis of pairwise comparisons between experiments (e.g. control and treatment). Their work was followed by Luce's [77] probabilistic approach for studying individual choice behavior and axiomatic development. The relationship between the axiomatic approach and probabilistic modeling was later established in Yellott's work [123].

Adopting RUMs rules out cyclic preferences, because each agent's outcome corresponds to

an order on real numbers, and it also captures the strength of preference, and thus overcomes the second criticism, by assigning a different parameter to each alternative.

The most important class of probabilistic ranking models are the random utility models (RUMs) [111, 85, 82]. RUMs assume that agents observe latent utilities for each alternative from some joint distribution on utilities. RUMs are statistical methods for rank data, and can be used to infer preferences between alternatives [120, 11]. The systematic study of such models (known as *choice theory*) has been an important topic in psychology and economics since Thurstone's seminal work in 1927 [111], and is well-known as *random utility theory* in economics.

RUMs include the Thurstone and Bradley-Terry models [85, 22]. A popular RUM is Plackett-Luce (P-L) [77, 99], where the random utility terms are generated according to Gumbel distributions with a fixed shape parameter [22, 123]. For P-L, the likelihood function has a simple analytical solution, making MLE inference tractable. P-L has been extensively applied in econometrics [82, 17], and more recently in machine learning and information retrieval (see [72] for an overview). Efficient methods of EM inference [67, 33], and more recently expectation propagation [57], have been developed for P-L and its variants. In application to social choice data, the P-L model has been used to analyze political elections [51, 52, 53, 54].

Although P-L overcomes the two difficulties of the Condorcet-Kemeny approach regarding the existence of cycles and computational hardness, it is still quite restricted, assuming that the random utility terms are distributed as Gumbel, with each alternative characterized by one parameter, the mean of its corresponding distribution. Plackett [99] in his 1975 paper considers parameterizing each alternative with a single parameter a disadvantage. Plackett mentions:

"A disadvantage of both methods (The generalization of the Bradley-Terry model to full ranks and RUM with normal distributions with known variance as noise) is that $r! - 1$ *independent probabilities are expressed in terms of only* $r - 1$ *parameters. In what follows, we construct a saturated model with* $r! - 1$ *parameters, consider the problems of inference which*

arise for unsaturated models in the same class, and apply the results to practical examples."[1]

Even so, the so called Plackett-Luce model has become a commonly used model in dealing with rank data and it is named due to the mention in Plackett's paper to the generalization from Bradley-Terry model, see section 5.6.1 in Marden [81], and despite Placket's criticism.

In fact, RUMs can provide flexible models that can address Plackett's earlier criticism. For example, Stern [108] proposes a new RUM with the Gamma distribution adopted for random utilities. However, because of the computational bottlenecks regarding inference with RUMs, most of the research on parametric models has been focused on Plackett-Luce for full ranks and Thurstone and Bradley-Terry model for pairwise observations [71, 36, 67, 33, 57, 53, 51]. Still, little is known about inference in RUMs beyond P-L. We are not aware of either an analytical solution or an efficient algorithm for MLE inference for one of the most natural models proposed by Thurstone [111], in which utility is Normally distributed.

2.2 Contributions

In this chapter we propose three different extensions to RUMs. The first model considers RUMs in which the random utilities are independently generated from distributions in the *exponential family* (EF) [92]. This extends the P-L model, since the Gumbel distribution with fixed shape parameters belongs to the EF. As an example of this extension, adding a variance parameter for each alternative in Thurstone's Normal model is shown to outperform other methods discussed in the literature such as Luce model.

The improved model performance can be explained through such as flexibility introduced by the variance parameter. One viewpoint is that different groups of agents have different preference behavior; i.e., have a distinct distribution of random utility scores on alternatives. Different preference behaviors can lead to a greater variance in the random utility model. Hence, considering a flexible variance parameter for the random utility of an alternative can capture this difference, and lead to an improved model.

[1] r represents the number of alternatives in Plackett's paper. A saturated model means to adding as many parameters as possible to the model so that the model stays identifiable.

The second model adopts a mixture over RUMs, hypothesizing that rank data is generated in a setting with multiple types of agents. Types are latent groups of agents which can correspond to unobserved characteristics of the agents. The estimated types can be interpretation revealing interesting latent structures in the data.

Mixture models for rank aggregation are appealing for various applications. In social choice [39], multiple types can correspond to different social beliefs among agents (e.g. Democrats and Republicans in US elections). Rank aggregation can also be used in information retrieval [71] and in this case multiple types can be assumed to be generated by using different search engines. In preference aggregation [69], multiple types can capture the personal preferences (e.g. in customers' preferences, some people prefer a product while others do not and capturing the different customer behaviors will help to asses the quality of an aggregation). In rank aggregation problems such as car racing [67] types can model the effect of different conditions for the race (e.g. weather conditions can change the ranking of racers). In ranking data produced from gene expression data [65], different types can correspond to different modes of actions by which treatments affect gene expression. We will outline an application in this direction.

In standard RUMs, the joint distribution on latent utility scores is a product distribution or a mixture of product distributions. This restricts the space of possible random utility models, precluding conditional dependence on the utilities of different alternatives. The third model adapts a nonparametric model that allows flexible densities and correlation between the random utilities on alternatives. Although non-parametric methods are not directly interpretable, non-parametric RUMS (NPRUMs) can unlock new understandings via post-processing and visualization.

We apply the MC-EM algorithm for inference and estimation in all three models. We treat the random utilities as latent variables, and adopt the Expectation Maximization method to estimate parameters. The E-step for this problem is not analytically tractable, and for this we adopt a Monte Carlo approximation. We establish through experiments that the Monte-Carlo error in the E-step is controllable across all models and does not affect inference, as

long as numerical parameterizations are chosen carefully. In addition, the $\epsilon-$ is small, and shrinks along the MC-EM iterations for some models. In addition, for the E-step we suggest a parallelization for the agents and alternatives and a Rao-Blackwellized method, which further increases the scalability of the approach.

MC-EM also extends easily to handle data with partial rank orders.

In the NPRUM case we use a variational version of MC-EM, forgoing distributional assumptions and retaining the correlation structure between utilities. We directly estimate the density function via kernel density estimation (KDE) with a Gaussian kernel, applied to sampled latent utility scores.

The main theoretical contributions in this chapter are Theorem 1 and Theorem 2, which propose conditions under which the log-likelihood function is concave and the set of global maxima solutions is bounded for the *location family*, which are RUMs where the shape of each distribution μ_j is fixed and the only latent variables are the locations, i.e., the means of μ_j's. These results hold for existing special cases, such as the P-L model, and other RUMs where the distributions are chosen from Normal, Laplace and Cauchy. In understanding multimodality for likelihood in the mixture model, we define the new notion of the $\epsilon-$log-concavity of a function. In Theorem 4, we develop conditions for the likelihood function of mixture of RUM models to be $\epsilon-$log-concave.

We evaluate these new RUMs on synthetic data as well as two real-world data-sets: a public election data-set and one involving rank preferences on sushi. The experimental results suggest that the approaches are scalable and provide significantly improved modeling flexibility over existing approaches.

The Luce model performs well on some data-sets (e.g. Election), while RUM with Normal distributions performs well on others (e.g. Sushi). The non-parametric RUM outperforms Luce model and Normal RUMs on all tested data-sets with regard to various metrics because of its flexibility to capture features and describe various types of data. For example, the NPRUM has a better out-of-sample fit in multiple real-world data-sets. It also outperforms existing RUMs in multiple predictive metrics, including predictive log-likelihood, predictive pairwise

preferences, and distribution estimation, and rank completion. Of course, the parametric models are more interpretable than the non-parametric models.

2.3 Related Work

Learning to aggregate full and partial ranks is a well-studied problem [44, 38, 3, 116, 120], and random utility models have been used in economics to model preferences [82].

Mixture models are studied widely in the statistics literature [89], but generally not for rank data. There are multiple issues with mixture models such as non-identifiability and non-uniqueness of maximum likelihood estimators. These issues are difficult in the general case, however, there is an extensive literature on addressing identifiability and uniqueness for special cases [49, 89]. Mixture models are well known to be multi-modal in general, both due to label switching and also non-uniqueness of modes in the equivalence class on the permutations of labels [49]. Gormley et al. [53, 51] apply mixture of Luce model to college application and election data. However, Gormley et al. [53, 51] focus on the Luce model, and do not provide theoretical results in regard to identifiability or uniqueness.

The EM algorithm has been used to learn the *Mallows* model (closely related to the Condorcet's probabilistic model) in Lu et al. [74]. They also introduce a mixture of Mallow models for rank data, and the identified types in their work support the hypothesis of the multiple types of agents. However, their mixture model applies only to the limited case of Mallow's model, and inference appears hard to scale.

2.4 Preliminaries

We define $\mathcal{C} = \{c_1, \ldots, c_m\}$ as the set of m alternatives. Let π denote a permutation of $\{1, \ldots, m\}$, which naturally corresponds to a linear order: $[c_{\pi(1)} \succ c_{\pi(2)} \succ \cdots \succ c_{\pi(m)}]$. Slightly abusing notation, we also use π to denote this linear order.

2.4.1 Random Utility Models

Suppose there is a ground truth utility (or score) associated with each alternative in $\mathcal{C} = \{c_1, \ldots, c_m\}$. These are real-valued parameters, denoted by $\vec{\theta} = (\theta_1, \ldots, \theta_m)$. Given this, an agent independently samples a random utility (U_j) for each alternative c_j with conditional distribution $\mu_j(\cdot|\theta_j)$.

Usually θ_j is the mean of $\mu_j(\cdot|\theta_j)$.[2] Random utility $(U_1, \ldots, U_m)$ generates a distribution on preference orders, as:

$$\Pr(\pi \mid \vec{\theta}) = \Pr(U_{\pi(1)} > U_{\pi(2)} > \ldots > U_{\pi(m)}) \tag{2.1}$$

The preference profile is viewed as *data*, $D = \{\pi^1, \ldots, \pi^n\}$. Given this, the probability (likelihood) of the data given ground truth $\vec{\theta}$ (and for a particular $\vec{\mu}$) is,

$$\Pr(D \mid \vec{\theta}) = \prod_{i=1}^{n} \Pr(\pi^i \mid \vec{\theta}) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \int_{u_{\pi(m)}}^{\infty} .. \int_{u_{\pi(2)}}^{\infty} \prod_{j=1}^{m} \mu_{\pi(j)}(u_{\pi(j)}) du_{\pi(m-j)} \tag{2.2}$$

The generative process is illustrated in Figure 2.1.

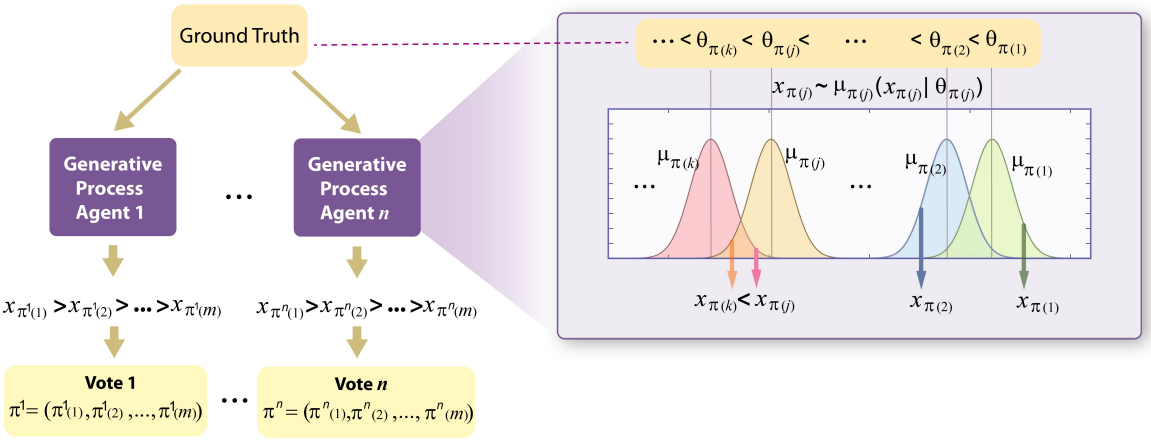


Figure 2.1: The generative process for RUMs.

For the first two proposed extensions, we focus on probabilistic models where each μ_j belongs to the *exponential family (EF)*. The density function for each μ in EF has the following

[2] $\mu_j(\cdot|\theta_j)$ might be parameterized by other parameters, for example variance.

format:

$$\Pr(X = u) = \mu(u) = e^{\eta(\theta)T(u) - A(\theta) + B(u)}, \tag{2.3}$$

where $\eta(\cdot)$ and $A(\cdot)$ are functions of θ, $B(\cdot)$ is a function of u, and $T(u)$ denotes the sufficient statistics for u, which could be multidimensional.

Example 1 (Plackett-Luce as an RUM [11, 22]) *In the RUM, let μ_j's be Gumbel distributions. That is, for alternative $j \in \{1, \ldots, m\}$ we have $\mu_j(u_j|\theta_j) = e^{-(u_j - \theta_j)} e^{-e^{-(u_j - \theta_j)}}$. Then, we have:*

$$Pr(\pi \mid \vec{\lambda}) = \prod_{j=1}^{m} \frac{\lambda_{\pi(j)}}{\sum_{j'=j}^{m} \lambda_{\pi(j')}},$$

where $\eta(\theta_j) = \lambda_j = e^{\theta_j}$, $T(u_j) = -e^{-u_j}$, $B(u_j) = -u_j$ and $A(\theta_j) = -\theta_j$. This gives us the Plackett-Luce model.

The Gumbel distribution with fixed shape parameter belongs to the EF, next example shows that MLE inference under P-L is equivalent to MLE inference for RUMs with an exponential distribution for the inverse profile.

Example 2 *Let π' denote the inverse of π, that is, for every $j \leq m$, $\pi(j) = \pi'(m + 1 - j)$. In RUM, let μ_j's be exponential distributions. That is, for alternative $j \in \{1, \ldots, m\}$ we have $\mu_j(u_j|\theta_j) = e^{-(u_j - \theta_j)} e^{-e^{-(u_j - \theta_j)}}$.*

Likelihood of π given θ under Gumbel is the same as the likelihood of π', which is the inverse of π, given θ under the exponential distribution. Therefore, P-L is equivalent to RUM with exponential distribution for the reverse profile.

Example 3 (Normal Model) *The* Normal model *adopts the Normal distribution for sampling an agent's score on each alternative. For alternative $j \in \{1, \ldots, m\}$, we have,* $\Pr(U_j = u_j \mid \nu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(u_j - \nu_j)^2}{2\sigma_j^2}}$.

One can consider σ_js as known constant or treat them as unknown parameters and estimate them along with ν_js. This is similar to Thurstone's model [111]. For the Normal model the

integral in equation (2.2) is an analytically intractable integral. Harville [59] proposed an approach to approximate it using a Taylor expansion of the likelihood function.

2.4.2 Different Data sets

There are different kinds of ranking data. In this chapter we consider full ranking, sub ranking, and top ranking data:

Definition 1 Full Ranking: *A full ranking has all alternatives $\mathcal{C}$ ranked. We observe the ranking $\pi = [\pi(1) \succ \pi(2) \succ \cdots \succ \pi(m)]$[3], containing all m alternatives.*

Given Pr, the probability for a ranking $\pi = [\pi(1) \succ \pi(2) \succ \cdots \succ \pi(m)]$ (which implies $[u_{\pi(1)} > u_{\pi(2)} > \cdots > u_{\pi(m)}]$) is defined as follows:

$$\Pr(\pi) = \int_{u_{\pi(m)} < \cdots < u_{\pi(1)}} \Pr(\vec{u}_\pi) d\vec{u}_\pi$$

Definition 2 Top Ranking: *A top ranking provides rankings on a proper subset $\mathcal{C}' \subsetneq \mathcal{C}$ with at least two alternatives. All elements of $\mathcal{C}'$ are preferred over the elements of $\mathcal{C}'^c$ (compliment of $\mathcal{C}'$). No information is gained of the preference relationship within the set $\mathcal{C}'^c$.*

In a top ranking we observe the ranking $\pi = [\pi(1) \succ \pi(2) \succ \cdots \succ \pi(m') \succ \{\pi_c\}_{c \in \mathcal{C}'^c}]$, where the set of m' (where $m' < m$) alternatives in $\mathcal{C}'$ are ranked and preferred over the other alternatives in $\mathcal{C}'^c$. We note that this ranking π implies $[u_{\pi(1)} > \cdots > u_{\pi(m')} > \max(\{u_c\}_{c \in \mathcal{C}'^c})]$. The probability of observing such a ranking is:

$$\Pr(\pi) = \int_{\max\left(\{u_c\}_{c \in \mathcal{C}'^c}\right) < u_{\pi(m')} < \cdots < u_{\pi(1)}} \Pr(\vec{u}_\pi) d\vec{u}_\pi$$

This kind of data occurs in elections with many candidates. Agents fill out their top positions with their preferred candidates, and then leave their less desired candidates unranked.

---

[3]WE will use $\pi(i)$ and $c_{\pi(i)}$ exchangeably.

Definition 3 Sub Ranking: *A sub ranking provides full rankings on a proper subset $\mathcal{C}' \subsetneq \mathcal{C}$ with at least two alternatives. However, no information is learned about the alternatives in the set $\mathcal{C}'^c$ or about the relationship between the sets $\mathcal{C}'^c$ and $\mathcal{C}'$.*

In sub ranking data we observe the ranking $\pi = [\pi(1) \succ \pi(2) \succ \cdots \succ \pi(m')]$ on the set of m' (where $m' < m$) alternatives $\mathcal{C}$. We note that this ranking π implies $[u_{\pi(1)} > \cdots > u_{\pi(m')})]$. The probability of observing such a ranking is:

$$\Pr(\pi) = \int_{u_{\pi(m')} < \cdots < u_{\pi(1)}} \Pr(\vec{u}'_\pi) d\vec{u}'_\pi$$

where $\vec{u}'$ is the vector of all $u \in \mathcal{C}'$.

This commonly occurs in race or competition data, where only a subset of the racers and competitors is compared in each ranking.

The integrals for computing the probabilities of rankings are computationally difficult to compute without any distributional assumptions. Yet understanding them is vital to perform inference. We use Monte Carlo methods to estimate probabilities of rank orders and likelihoods of observed data.

2.4.3 Maximum Likelihood Estimator

In the maximum likelihood (MLE) approach to social choice, the preference profile is viewed as *data*, $D = \{\pi^1, \ldots, \pi^n\}$. Given this, the probability (likelihood) of the data given ground truth $\vec{\theta}$ (and for a particular $\vec{\mu}$) is,

$$\Pr(D \mid \vec{\theta}) = \prod_{i=1}^{n} \Pr(\pi^i \mid \vec{\theta}), \tag{2.4}$$

The MLE approach to social choice selects as the winning ranking that which corresponds to the $\vec{\theta}$ that maximizes $\Pr(D \mid \vec{\theta})$. In the case of multiple parameters that maximize the likelihood then the MLE approach returns a set of rankings, one ranking corresponding to each parameterization.

2.5 Three Extensions

2.5.1 Model Extension to Exponential Families

In this section we focus on RUMs in which the random utilities are independently generated with respect to distributions in the *exponential family* (EF) [92].

This extends the P-L model, since the Gumbel distribution with fixed shape parameters belongs to the EF. Our main theoretical contributions are Theorem 1 and Theorem 2, which propose conditions such that the log-likelihood function is concave and the set of global maxima solutions is bounded for the *location family*, which are RUMs where the shape of each distribution μ_j is fixed and the only latent variables are the locations, i.e., the means of μ_j's. These results hold for existing special cases, such as the P-L model, and many other RUMs, for example the ones where each μ_j is chosen from Normal, Gumbel, Laplace and Cauchy.

Global Optimality and Log-Concavity

We provide a condition on distributions that guarantees that the likelihood function (2.2) is log-concave in parameters $\vec{\theta}$. We also provide a condition under which the set of MLE solutions is bounded when any one latent parameter is fixed.

Together, this can guarantees the convergence of algorithms such as gradient descent or EM algorithm approach to a global mode. We focus on the *location family*, which is a subset of RUMs where the shapes of all μ_j's are fixed, and the only parameters are the means of the distributions. For the location family, we can write $U_j = \theta_j + \zeta_j$, where $U_j \sim \mu_j(\cdot|\theta_j)$ and $\zeta_j = U_j - \theta_j$ is a random variable whose mean is 0 and models an agent's *subjective noise*.

The random variables ζ_j's do not need to be identically distributed for all alternatives j; e.g., they can be normal with different fixed variances. We focus on computing solutions $(\vec{\theta})$ to maximize the log-likelihood function,

$$l(\vec{\theta}; D) = \sum_{i=1}^{n} \log \Pr(\pi^i \mid \vec{\theta}) \tag{2.5}$$

Theorem 1 *For the location family, if for every $j \leq m$ the probability density function for ζ_j is log-concave, then $l(\vec{\theta}; D)$ is concave.*

Proof: The theorem is proved by applying the following lemma, which is Theorem 9 in [102].

Lemma 1 *Suppose $g_1(\vec{\theta}, \vec{\zeta}), ..., g_R(\vec{\theta}, \vec{\zeta})$ are concave functions in $\mathbb{R}^{2m}$ where $\vec{\theta}$ is the vector of m parameters and $\vec{\zeta}$ is a vector of m real numbers that are generated according to a distribution whose pdf is logarithmic concave in $\mathbb{R}^m$. Then the following function is log-concave in $\mathbb{R}^m$.*

$$L_i(\vec{\theta}, G) = \Pr(g_1(\vec{\theta}, \vec{\zeta}) \geq 0, ..., g_R(\vec{\theta}, \vec{\zeta}) \geq 0), \quad \vec{\theta} \in \mathbb{R}^m \tag{2.6}$$

To apply Lemma 1, we define a set G^i of function g^i's that is equivalent to an order π^i in the sense of inequalities implied by RUM for π^i and G^i (the joint probability in (2.6) for G^i to be the same as the probity of π^i in RUM with parameters $\vec{\theta}$). Suppose $g^i_r(\vec{\theta}, \vec{\zeta}) = \theta_{\pi^i(r)} + \zeta^i_{\pi^i(r)} - \theta_{\pi^i(r+1)} - \zeta^i_{\pi^i(r+1)}$ for $r = 1, .., m-1$.

Then considering that the length of order π^i is $R+1$, we have:

$$L_i(\vec{\theta}, \pi^i) = L_i(\vec{\theta}, G^i) = \Pr(g^i_1(\vec{\theta}, \vec{\zeta}) \geq 0, ..., g^i_R(\vec{\theta}, \vec{\zeta}) \geq 0), \quad \vec{\theta} \in \mathbb{R}^m \tag{2.7}$$

This is because $g^i_r(\vec{\theta}, \vec{\zeta}) \geq 0$ is equivalent to that in π^i alternative $\pi^i(r)$ is preferred to alternative $\pi^i(r+1)$ in the RUM sense.

To see how this extends to the case where preferences are specified as partial orders, we consider in particular an interpretation where an agent's report for the ranking of m_i alternatives implies that all other alternatives are worse for the agent, in some undefined order. Given this, define $g^i_r(\vec{\theta}, \vec{\zeta}) = \theta_{\pi^i(r)} + \zeta^i_{\pi^i(r)} - \theta_{\pi^i(r+1)} - \zeta^i_{\pi^i(r+1)}$ for $r = 1, .., m_i - 1$ and $g^i_r(\vec{\theta}, \vec{\zeta}) = \theta_{\pi^i(m_i)} + \zeta^i_{\pi^i(m_i)} - \theta_{\pi^i(r+1)} - \zeta^i_{\pi^i(r+1)}$ for $r = m_i, .., m-1$. Considering that $g^i_r(\cdot)$s are linear (hence, concave) and using log concavity of the distributions of $\vec{\zeta^i} = (\zeta^i_1, \zeta^i_2, .., \zeta^i_m)$'s, we can apply Lemma 1 and prove log-concavity of the likelihood function. □

It is not hard to verify that pdfs for Normal and Gumbel are log-concave under reasonable conditions for their parameters, made explicit in the following corollary.

Corollary 1 *For the location family where each ζ_j is a Normal distribution with mean zero and with fixed variance, or Gumbel distribution with mean zeros and fixed shape parameter,*

$l(\vec{\theta}; D)$ is concave. Specifically, the log-likelihood function for P-L is concave.

The concavity of log-likelihood of P-L has been proved [47] using a different technique. Using Fact 3.5. in [104], the set of global maxima solutions to the likelihood function, denoted by S_D, is convex since the likelihood function is log-concave. However, we also need that S_D is bounded, and would further like that it provides one unique order as the estimation for the ground truth.

For P-L, Ford, Jr. [47] proposed the following necessary and sufficient condition for the set of global maxima solutions to be bounded (more precisely, unique) when $\sum_{j=1}^{m} e^{\theta_j} = 1$.

Condition 1 *Given the data D, in every partition of the alternatives $\mathcal{C}$ into two non-empty subsets $\mathcal{C}_1 \cup \mathcal{C}_2$, there exists $c_1 \in \mathcal{C}_1$ and $c_2 \in \mathcal{C}_2$ such that there is at least one ranking in D where $c_1 \succ c_2$.*

Condition 1 is also a necessary and sufficient condition for the set of global maxima solutions S_D to be bounded in location families, when we set one of the values θ_j to be 0 (w.l.o.g., let $\theta_1 = 0$). If we do not bound any parameter, then S_D is unbounded, because for any $\vec{\theta}$, any D, and any number $s \in \mathbb{R}$, $l(\vec{\theta}; D) = l(\vec{\theta} + s; D)$.

Theorem 2 *Suppose we fix $\theta_1 = 0$. Then, the set S_D of global maxima solutions to $l(\theta; D)$ is bounded if and only if the data D satisfies Condition 1.*

Proof: If Condition 1 does not hold, then S_D is unbounded because the parameters for all alternatives in C_1 can be increased simultaneously to improve the log-likelihood. For sufficiency, we use the following lemma.

Lemma 2 *If alternative j is preferred to alternative j' in at least in one ranking then the difference of their mean parameters $\theta_{j'} - \theta_j$ is bounded from above ($\exists Q$ where $\theta_{j'} - \theta_j < Q$) for all the $\vec{\theta}$ that maximize the likelihood function.*

Now consider a directed graph G_D, where the nodes are the alternatives, and there is edge from c_j to $c_{j'}$ if in at least one ranking $c_j \succ c_{j'}$. By Condition 1, for any pair $j \neq j'$, there is a path from c_j to $c_{j'}$ (and conversely, a path from $c_{j'}$ to c_j). To see this, consider building a

path between j and j' by starting from a partition with $\mathcal{C}_1 = \{j\}$ and following an edge from j to j_1 in the graph where j_1 is an alternatives in $\mathcal{C}_2$ for which there must be such an edge, by Condition 1. Consider the partition with $\mathcal{C}_1 = \{j, j_1\}$, and repeat until an edge can be followed to vertex $j' \in \mathcal{C}_2$. It follows from Lemma 2 that for any $\vec{\theta} \in S_D$ we have $|\theta_j - \theta_{j'}| < Qm$, using the telescopic sum of bounded values of the difference of mean parameters along the edges of the path, since the length of the path is no more than m (and tracing the path from j to j' and j' to j), meaning that S_D is bounded. □

Now that we have the log concavity and bounded property, we want conditions under which the bounded convex space of estimated parameters corresponds to a unique order. The next theorem provides a necessary and sufficient condition for all global maxima to correspond to the same order on alternatives. Suppose that we order the alternatives based on estimated θ's (meaning that c_j is ranked higher than $c_{j'}$ iff $\theta_j > \theta_{j'}$).

Theorem 3 *The order over parameters is strict and is the same across all $\vec{\theta} \in S_D$ if, for all $\vec{\theta} \in S_D$ and all alternatives $j \neq j'$, $\theta_j \neq \theta_{j'}$.*

Proof: Suppose for the sake of contradiction there exist two maxima, $\vec{\theta}, \vec{\theta}^* \in S_D$ and a pair of alternatives $j \neq j'$ such that $\theta_j > \theta_{j'}$ and $\theta^*_{j'} > \theta^*_j$. Then, there exists an $\alpha < 1$ such that the jth and j'th components of $\alpha\vec{\theta} + (1-\alpha)\vec{\theta}^*$ are equal, which contradicts the assumption. □

Hence, if there is never a tie in the scores in any $\vec{\theta} \in S_D$, then any vector in S_D will reveal the unique order.

2.5.2 Model Extension to Multiple Type

For this extension, we assume there exists multiple types of agents and we propose a model using mixture of RUMs. Intuitively, we are considering different components each different parameters to represent a different behavior in preference. In other words the probability of a preference is a mixture of a set of RUM models as follows:

$$\Pr(\pi|\Psi) = \sum_{k=1}^{K} \gamma_k \Pr(\pi|\vec{\theta}_k, z_k = 1), \tag{2.8}$$

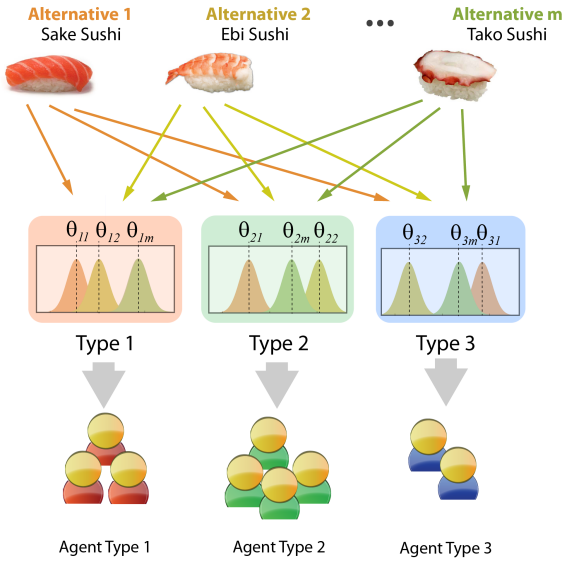


Figure 2.2: The generative process for multiple type RUMs. There are different types of agents with different random utilities for the alternatives.

where $\vec{z}$ indicates the type of the data point and $\Pr(z_k = 1) = \gamma_k$, $\Theta = \{\vec{\theta}_1, .., \vec{\theta}_K\}$ and $\Psi = \{\Theta, \vec{\gamma}\}$. Given this, we have:

$$\Pr(D \mid \Psi) = \prod_{i=1}^{n} \Pr(\pi^i \mid \Psi), \tag{2.9}$$

The generative process is illustrated in Figure 4.1.

Approximate Log concavity

Here, we address the multimodality of mixture models by defining an approximate version of log concavity for the likelihood function (4.2.2) for parameters in Ψ.

We again focus on the *location family* for their generality and canonical form of representation. In order to explore the uniqueness of the solutions for the maximum likelihood estimator, we use Theorem (1) which can provide log concavity conditions for each of the mixture components. However, we need conditions on the log concavity of the mixture likelihood in equation (2.8). As aforementioned, mixture models are multi-modal due to label switching or the shape of likelihood function (the label of mixtures does not matter and we can always switch their labels and get a new mixture model with the same likelihood).

Definition 4 *A function $f(\Psi)$ is called ϵ-log-concave if it can be decomposed into two components where one of them is log-concave and the other one is bounded in logarithm within an $\epsilon \geq 0$ interval. In other words:*

$$f(\Psi) = g(\Psi)e(\Psi),$$

where g is log-concave and for all Ψ in the parameter space, we have $0 \leq \log e(\Psi) \leq \epsilon$.

In the following we prove that mixture models are $\epsilon-$log-concave with some extra constraints on the parameter space.

Theorem 4 *If we have the following constraints:*

1. *For all k, the k-th mixture component $\Pr_k(D|\vec{\theta}_k)$ is log concave in θ_k;*
2. *The prior $\Pr(\vec{z}|\vec{\gamma})$ is not dogmatic (has non-zero values for any z) and it is log concave for $\vec{\gamma}$; and*
3. *Components are diverse, meaning that for two components $k < k'$:*

$$KL(\Pr_k(\pi|\vec{\theta}_k)||\Pr_{k'}(\pi|\vec{\theta}_{k'})) \geq \Delta,$$

where KL is the Kullback-Leibler divergence between two distributions; then

$\Pr(D|\vec{\gamma}, \Theta)$ *is almost surely $\epsilon-$log-concave in $\Psi = \{\vec{\gamma}, \Theta\}$ for a $C > 0$ and*

$$\epsilon = n(K-1)e^{-\Delta/C}$$

Proof: Using Bayes rule and taking the logarithm we have the following:

$$\log \Pr(\pi|\Psi) = \log \Pr(\pi, \vec{z}|\Psi) - \sum_k z_k \log \Pr(z_k = 1|\pi, \Psi)$$

We take the expectation over $\vec{z}$ with respect to the distribution: $\Pr'(\vec{z}|\pi, \Psi^*) = \mathbf{1}(\vec{z} = \arg\max_{\vec{z}} \Pr(\vec{z}|\pi, \Psi^*))$ (for a Ψ^* which is consistent with assumption 3) from both sides of the above equation, and obtain:

$$\log \Pr(\pi|\Psi) = E_{\vec{z}}\{\log \Pr(\pi, \vec{z}|\Psi)|\pi, \Psi^*\} - \sum_{k=1}^{K} \Pr'(z_k = 1|\pi, \Psi^*) \log \Pr(z_k = 1|\pi, \Psi)$$

The concavity of $\log \Pr(\pi, \vec{z}|\Psi) = \log \Pr(\pi|\vec{z}, \Psi) + \log \Pr(\vec{z}|\vec{\gamma})$ is now a direct result of concavity of $\log \Pr(\pi|\vec{z}, \Psi) = \log \Pr(\pi|\theta_k)$ for $(z_k = 1)$ from Theorem 1 and concavity of $\log \Pr(\vec{z}|\vec{\gamma})$ (from assumption 2). Hence, the term $E_{\vec{z}}\{\log \Pr(\pi, \vec{z}|\Psi)|\pi, \Psi^*\}$ is concave as well. In the following we show that the absolute value of the term,

$$H(\Psi|\pi, \Psi^*) = -\sum_{k=1}^{K} \Pr'(z_k = 1|\pi, \Psi^*) \log \Pr(z_k = 1|\pi, \Psi) \tag{2.10}$$

is bounded in an epsilon interval if the components satisfy the proposed constraint in assumption 1.

Using a concentration inequality we show that for every π almost surely there exists a k such that for any $k' \neq k$, there exists a fixed constant C such that we have:

$$\frac{\Pr(z_k = 1|\pi, \Psi)}{\Pr(z_{k'} = 1|\pi, \Psi)} \geq e^{\Delta/C} \frac{\Pr(z_k = 1|\vec{\gamma})}{\Pr(z_{k'} = 1|\vec{\gamma})}$$

Then if there is no switching between Ψ and Ψ^*, (meaning if $\Pr(z_k = 1|\pi, \Psi^*)$ is close to 1, then $\Pr(z_k = 1|\pi, \Psi)$ is close to 1 as well), and by assuming a uniform prior for type memberships WLOG, we have:

$$H(\Psi|\pi, \Psi^*) \leq (K - 1)e^{-\Delta/C},$$

This provides the decomposition leading to the ϵ-log-concavity of the function $\Pr(\pi|\Psi)$. $\square$

Even though the $\epsilon-$log-concavity of the likelihood function in (2.9) does not directly lead to uniqueness of MLE, when ϵ is very small the log likelihood function will have a maximum that can be reached by EM algorithms that are able to skip any local optima that have ϵ depth.

We will illustrate some empirical results on behavior of ϵ for the likelihood function in the empirical studies on the data sets we are using.

We have computed ϵ from equation (2.10) and the average log likelihood function values for both the sushi data along 20 iterations of our MC-EM algorithm. The values are plotted in Figure 2.3 for the iterations revealing the shrinking behavior of the ϵ (clustering quality) while the average log likelihood converges. Clustering quality is computed from the MC-E

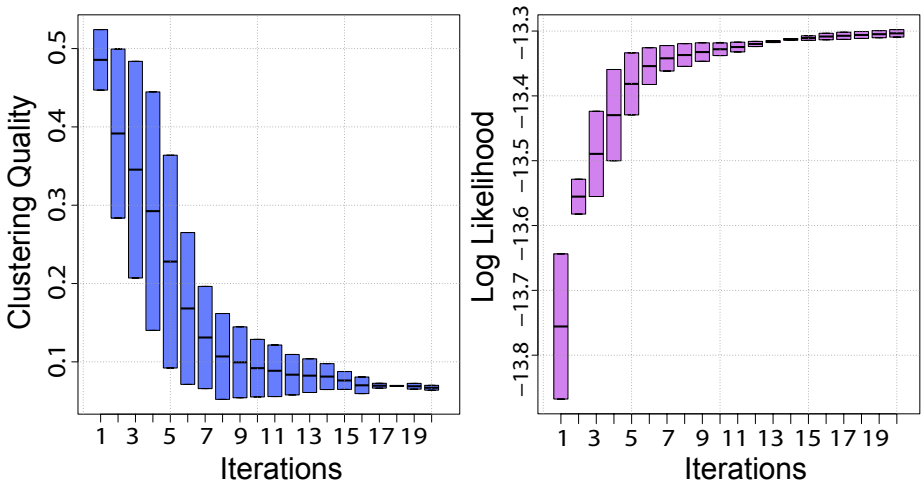


Figure 2.3: Convergence of the MCEM algorithm for the average log likelihood in the right panel and the ϵ for the $\epsilon-$log-concavity in the left panel The lower the ϵ the better the quality of clustering. Both of the plots are for the 2NFV.

step in every iteration of the algorithm and it corresponds to the Kullback Leibler divergence between components of the mixture model as shown in Theorem 4. More distinguishable components lead to smaller (improved) clustering quality.

We can show that Condition 1 is also a necessary and sufficient condition for the set of global maxima solutions to be bounded in each component of the mixture for location families. Here we need to set one of the values θ_{kj} for each component k to be 0 (w.l.o.g., let $\theta_{k1} = 0$).

Theorem 5 *Suppose we fix $\theta_{k1} = 0$ for all of the components and γ_ks are all non-zero. Then, the parameters providing a maxima solution to $l(\Psi; D)$ are bounded if and only if the data D satisfies Condition 1.*

Proof: The proof for the above theorem follows from the boundedness result for each component. □

2.5.3 Model Extension to Non-parametric settings

As the third extension, we propose non-parametric random utility model, with a non-parametric joint distribution on random utilities.

We impose restrictions on the non-parametric distribution using kernel density estimators (KDE) with Normal kernels [106, 97]. The samples for KDE are generated from MC-E step

of the algorithm. Hence, our NPRUM will be continuous with smoothness imposed by the bandwidth $h > 0$. Specifically, given a set of sample utilities u_{ij} for a specific alternative j, we estimate $\Pr_j$, the marginal utility distribution of alternative j, as:

$$\Pr_j(x) \propto \begin{cases} 0 & \text{if } x \notin (0,1) \\ \sum_i \phi_h(x - u_{ij}) & \text{if } x \in (0,1) \end{cases}$$

where $\phi_h(x) \propto \exp\{-\frac{x^2}{2h^2}\}$, the density function of kernel $\mathcal{N}(0, h^2)$. $\Pr_j(x)$ is rescaled to integrate to 1. To store the function, we evaluate the $\Pr_j(x)$ on a set of evenly-spaced evaluation points $x \in \{0, 1/d, 2/d, .., 1\}$ for a d which indicates the resolution of our non-parametric densities. As shown in Figure 2.4, a larger h leads to more smoothing of the resulting marginal utility distribution.

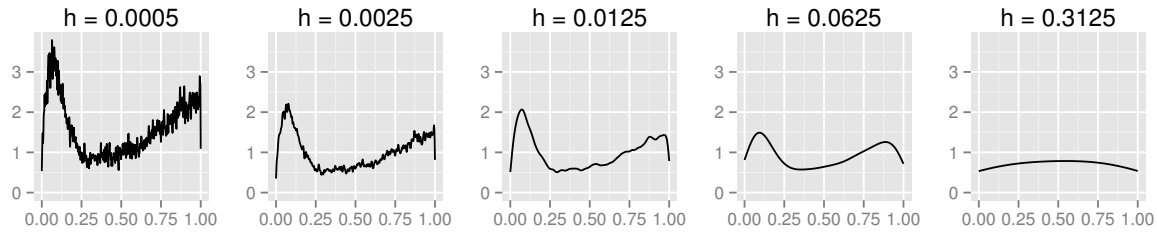


Figure 2.4: Sample KDE. If h is too low, there are spurious artifacts. If h is too high, it drowns out the features of the distribution.

We want a bounded range for our random utilities in order to fix the effect of h, and prevent the need to consider positive affine transformations of our RUMs. Picking a set of evaluation points for the KDE is simpler when the support of distributions is finite. The specific bounded interval $[0, 1]$ is chosen for simplicity.

2.6 Maximum Likelihood Estimator

We propose a novel application of MC-EM to estimate all three proposed models. We treat the random utilities ($\vec{U}$) and the type variables (Z) as latent variables, and adopt the Expectation Maximization (EM) method to estimate the utility distributions. The E-step for this problem is not analytically tractable, and for this we adopt a Monte Carlo approximation.

We generally assume that the data provides total orders on alternatives from agents, but comment on how to extend the method and theory to the case where the input preferences are *partial* orders.

2.6.1 EM algorithm for Latent Space Models

Computing the likelihood involves a multidimensional integral and hence direct optimization of the likelihood function is intractable. Thus, we use an MC-EM algorithm. The EM algorithm determines the MLE joint distribution $\Pr^*(\vec{U})$ iteratively and it is composed of iterations on an E-step and an M-step. Given $\Pr^t(\vec{U})$ from the previous iteration, we perform the following on each iteration $t+1$:

$$\text{E-step}: \quad Q(\Pr, \Pr^t) = E_{\vec{U}}\left\{\log \prod_{i=1}^{n} \Pr(\vec{u}_i, \pi_i) \mid D, \Pr^t\right\}$$

$$\text{M-step}: \quad \Pr^{t+1} \in \arg\max_{\Pr} Q(\Pr, \Pr^t)$$

2.6.2 MC-EM for Exponential Family RUM

In this section, we propose an MC-EM algorithm for MLE inference for RUMs where every μ_j belongs to the EF.[4] The EM algorithm determines the MLE parameters $\vec{\theta}$ iteratively, and proceeds as follows. In each iteration $t+1$, given parameters $\vec{\theta}^t$ from the previous iteration, the algorithm is composed of an E-step and an M-step. For the E-step, for any given $\vec{\theta} = (\theta_1, \ldots, \theta_m)$, we compute the conditional expectation of the complete-data log-likelihood (latent variables $\vec{x}$ and data D), where the latent variables $\vec{x}$ are distributed according to data D and parameters $\vec{\theta}^t$ from the last iteration.

For the M-step, we optimize $\vec{\theta}$ to maximize the expected log-likelihood computed in the

[4] Our algorithm can be naturally extended to compute a maximum *a posteriori* probability (MAP) estimate, when we have a prior over the parameters $\vec{\theta}$. Still, it might be hard to motivate the imposition of a prior on parameters in some application such as social choice domains.

E-step, and use it as the input $\vec{\theta}^{t+1}$ for the next iteration:

$$\text{E-Step}: \quad Q(\vec{\theta}, \vec{\theta}^t) = E_{\vec{U}} \left\{ \log \prod_{i=1}^{n} \Pr(\vec{u}^i, \pi^i \mid \vec{\theta}) \mid D, \vec{\theta}^t \right\}$$

$$\text{M-step}: \quad \vec{\theta}^{t+1} \in \arg\max_{\vec{\theta}} Q(\vec{\theta}, \vec{\theta}^t)$$

Monte Carlo E-step by Gibbs sampler

The E-step can be simplified using (2.3) as follows:

$$E_{\vec{U}}\{\log \prod_{i=1}^{n} \Pr(\vec{u}^i, \pi^i \mid \vec{\theta}) \mid D, \vec{\theta}^t\} = E_{\vec{U}}\{\log \prod_{i=1}^{n} \Pr(\vec{u}^i \mid \vec{\theta}) \Pr(\pi^i | \vec{u}^i) \mid D, \vec{\theta}^t\}$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} E_{U_j^i}\{\log \mu_j(u_j^i|\theta_j) \mid \pi^i, \vec{\theta}^t\} = \sum_{i=1}^{n} \sum_{j=1}^{m} (\eta(\theta_j) E_{U_j^i}\{T(u_j^i) \mid \pi^i, \vec{\theta}^t\} - A(\theta_j) + W,$$

where $W = E_{U_j^i}\{B(u_j^i) \mid \pi^i, \vec{\theta}^t\}$ only depends on $\vec{\theta}_t$ and D (not on $\vec{\theta}$), which means that it can be treated as a constant in the M-step.

Hence, in the E-step we only need to compute $S_j^{i,t+1} = E_{U_j^i}\{T(u_j^i) \mid \pi^i, \vec{\theta}^t\}$ where $T(u_j^i)$ is the sufficient statistic for the parameter θ_j in the model. We are not aware of an analytical solution for $E_{U_j^i}\{T(u_j^i) \mid \pi^i, \vec{\theta}^t\}$. However, we can use a Monte Carlo approximation, which involves sampling $\vec{x}^i$ from the distribution $\Pr(\vec{u}^i \mid \pi^i, \vec{\theta}^t)$ using a Gibbs sampler, and then approximates $S_j^{i,t+1}$ by $\frac{1}{N} \sum_{k=1}^{N} T(u_j^{i,k})$ where N is the number of samples in the Gibbs sampler.

In each step of our Gibbs sampler for agent i, we randomly choose a position j in π^i and sample $x^i_{\pi^i(j)}$ according to a *TruncatedEF* distribution $\Pr(\cdot \mid u_{\pi^i(-j)}, \vec{\theta}^t, \pi^i)$, where $u_{\pi^i(-j)} = (u_{\pi^i(1)}, \ldots, u_{\pi^i(j-1)}, u_{\pi^i(j+1)}, \ldots, u_{\pi^i(m)})$. The TruncatedEF is obtained by truncating the tails of $\mu_{\pi^i(j)}(\cdot|\theta^t_{\pi^i(j)})$ at $u_{\pi^i(j-1)}$ and $u_{\pi^i(j+1)}$, respectively. For example, a truncated normal distribution is illustrated in Figure 2.5.

Rao-Blackwellized: To further improve the Gibbs sampler, we use Rao-Blackwellized [32] estimation using $E\{T(u_j^{i,k}) \mid u_{-j}^{i,k}, \pi^i, \vec{\theta}^t\}$ instead of the sample $x_j^{i,k}$, where $u_{-j}^{i,k}$ is all of $\vec{u}^{i,k}$ except for $u_j^{i,k}$. Finally, we estimate $E\{T(u_j^{i,k}) \mid u_{-j}^{i,k}, \pi^i, \vec{\theta}^t\}$ in each step of the Gibbs sampler

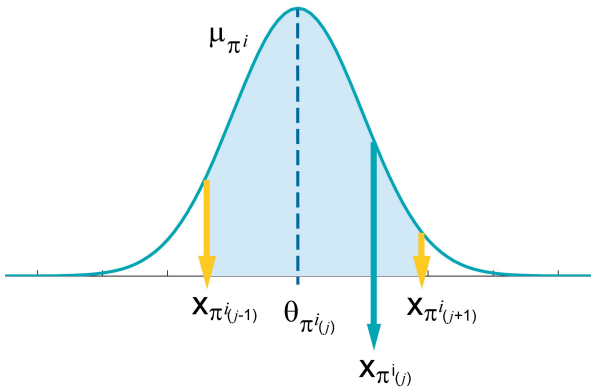


Figure 2.5: Sampling from a truncated Normal distribution.

using M samples as $S_j^{i,t+1} \simeq \frac{1}{N}\sum_{k=1}^{N} E\{T(u_j^{i,k}) \mid u_{-j}^k, \pi^i, \vec{\theta}^t\} \simeq \frac{1}{NM}\sum_{k=1}^{N}\sum_{l=1}^{M} T(u_j^{i^l,k})$, where $u_j^{i^l,k} \sim \Pr(u_j^{i^l,k} \mid u_{-j}^{i,k}, \pi^i, \vec{\theta})$. Rao-Blackwellization reduces the variance of the estimator because of conditioning and expectation in $E\{T(u_j^{i,k}) \mid u_{-j}^{i,k}, \pi^i, \vec{\theta}^t\}$.

M-step

In the E-step we have (approximately) computed $S_j^{i,t+1}$. In the M-step we compute $\vec{\theta}^{t+1}$ to maximize $\sum_{i=1}^{n}\sum_{j=1}^{m}(\eta(\theta_j)E_{U_j^i}\{T(u_j^i) \mid \pi^i, \vec{\theta}^t\} - A(\theta_j) + E_{U_j^i}\{B(u_j^i) \mid \pi^i, \vec{\theta}^t\})$. Equivalently, we compute θ_j^{t+1} for each $j \le m$ separately to maximize $\sum_{i=1}^{n}\{\eta(\theta_j)E_{U_j^i}\{T(u_j^i) \mid \pi^i, \vec{\theta}^t\} - A(\theta_j)\} = \eta(\theta_j)\sum_{i=1}^{n} S_j^{i,t+1} - nA(\theta_j)$. For the case of the normal distribution with fixed variance, where $\eta(\theta_j) = \theta_j$ and $A(\theta_j) = (\theta_j)^2$, we have $\theta_j^{t+1} = \frac{1}{n}\sum_{i=1}^{n} S_j^{i,t+1}$. The algorithm is illustrated in Figure 2.6. Theorem 1 and Theorem 2 guarantee the convergence of MC-EM

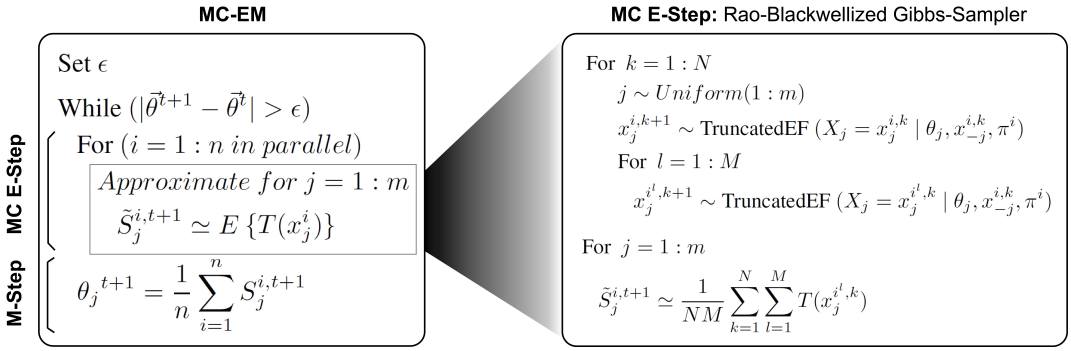


Figure 2.6: The MC-EM algorithm for normal distribution.

for an exact E-step. In order to control the error of approximation in the MC-E step we can

increase the number of samples with the iterations, in order to decrease the error in Monte Carlo step [119].

2.6.3 MC-EM for the Multiple Type RUM

The E-step can be simplified using (2.3) as follows:

$$E_{\vec{U},\vec{Z}}\left\{\log\prod_{i=1}^{n}\Pr(\vec{x}^i,\vec{z}^k,\pi^i\mid\Psi)\mid D,\Psi^t\right\}=\sum_{i,j,k}[E_{Z^i}\{\mathbf{1}(Z^i=k)|\ \pi^i,\vec{\Theta}^t\}\log\gamma_k$$
$$+E_{U_j^i,Z^i}\{\mathbf{1}(Z=k)\log\mu_j(u_j^i|\theta_{kj})\mid\pi^i,{\theta_{kj}}^t\}]$$

And we define the ESTEP functions as following,

$$ESTEP1_{ijk}^t(\theta_{kj})=\eta(\theta_{kj})E_{U_j^i,Z^i}\{\mathbf{1}(Z=k)T(u_j^i)\mid\pi^i,\vec{\theta_{kj}}^t\}-A(\theta_{kj})+E_{U_j^i,Z^i}\{\mathbf{1}(Z=k)B(u_j^i)\mid\pi^i,\vec{\theta_{kj}}^t\},$$
$$ESTEP2_{ik}^t=E_{Z^i}\{\mathbf{1}(Z^i=k)|\ \pi^i,\Theta^t\}$$

where $E_{U_j^i}\{B(u_j^i)\mid\pi^i,\Theta^t\}$ only depends on Θ^t and D (not on $\vec{\theta}$), which means that it can be treated as a constant in the M-step.

Hence, in the E-step we only need to compute $S_{j,k}^{i,t+1}=E_{U_j^i}\{\mathbf{1}(Z=k)T(u_j^i)\mid\pi^i,\vec{\theta_{kj}}^t\}$ where $T(u_j^i)$ is the sufficient statistic for the parameter θ_j in the model.

2.6.4 MC-EM for Non-parametric RUM

E-step:

The E-step draws from the joint utility distribution conditional on observed rank data. Drawing directly from the joint density is intractable, so we rely on Monte-Carlo methods. We want to sample a vector of utility observations for each observation (agent), conditional on their observed rank preference. Sampling the whole vector simultaneously is difficult, so we adopt a Gibbs method to sample each utility sequentially. Conditioning each sample on the rank order and other utilities, we sample from the alternative's utility distribution. In the case of full ranks, the rank order and other utilities imposes the following restriction on

Algorithm 1 MC-EM Algorithm for multiple type RUMs with Normal random utility distributions

Initialize: Ψ^0 and $u_{k,j}^{i,0}$
Variables: T_1, T_2
  for $t_1 = 0 : T_1$ **do**
    MC E-Step:
    set $T_2 = 3000 + 300 * T_1$
    for $t_2 = 0 : T_2$ **do**
      $i \sim \text{Uniform}(1:n)$
      $z^i \sim \Pr(Z|U^{t_2}, \Psi^{t_1}) = \frac{\gamma_{z^i}^{t_1} \Pr(u^{t_2-1}|Z=z^i,\Theta^{t_1})}{\sum_k \gamma_k^{t_1} \Pr(u^{t_2-1}|Z=k,\Theta^{t_1})}$
      $k = z^i$
      for $j = 1 : m$ **do**
        $x_j^{i,t_2} \sim \text{TruncatedEF}(X|x_{k,-j}^{i,t_2}, \pi^i, \theta_{kj}^{t_1})$
      end for
    end for
    for $k = 1 : K$ and $j = 1 : m$ **do**
      $S_{j,k}^{i,t_1+1} = \frac{1}{T_2} \sum_{t_2=1}^{T_2} \mathbf{1}(z^i = k) T(u_j^{i,t_2})$
    end for
    M-Step:
    for $k = 1 : K$ and $j = 1 : m$ **do**
      $\theta_{kj}^{t_1} = \frac{1}{n} \sum_i S_{j,k}^{i,t_1}$, $\gamma_k^{t_1+1} = \frac{\sum_i ESTEP2_{ik}^{t_1}}{\sum_k \sum_i ESTEP2_{ik}^{t_1}}$
    end for
  end for

utility orderings:

$$u_{\pi(j)} \in \begin{cases} (0, u_{\pi_i(j+1)}) & \text{if } j = 1 \\ \left(u_{\pi_i(j-1)}, u_{\pi_i(j+1)}\right) & \text{if } 1 < j < m \\ \left(u_{\pi_i(j-1)}, 1\right) & \text{if } j = m \end{cases}$$

In the case of partial ranks, we can modify the restriction in a way any observation, any alternative ranked above another must also have a higher utility.

Within the Gibbs sampler, we use slice sampling [95] to sample latent utilities. Tarlow et al. [110] argues slice sampling is well suited for sampling latent variables in MC-EM. We rely on Neal's implementation of his slice sampler [94], and leave a more detailed explanation of this method to Neal [95].

M-step:

The M-step estimates the non-parametric joint density over the utilities using kernel density estimation, assuming Normal kernels with a bandwidth h. However, KDE on many dimensions is intractable as the number of evaluation points grows exponentially with the m. Therefore, we adopt a variational method and estimate the joint distribution as a product distribution $\hat{\Pr}(\vec{u}) = \prod_j \Pr_j(\vec{u})$.

The variational M-step can be done for each of the marginal distributions separately. We note that even though the M-step uses the marginal distributions for inference, the output of the MC-EM algorithm keeps the correlation structure.

Algorithm: From the output of the MC-EM algorithm, we construct the joint distribution over utilities using the KDE. This joint distribution is easy to sample from, as we can draw a random $\vec{u}_i$ and a corresponding value from the kernel associated with the point. See Algorithm 2 for a summary.

2.7 Experimental Results

We evaluate our methods on the datasets in Table 2.1. Via experiments, we compare the ability of various RUMs to:

Algorithm 2 MC-EM algorithm for NPRUM

$t \leftarrow 0$
repeat
  (Variational MCMC E-step)
  for all agents i **do**
    repeat
      for all alternatives j **do**
        $u_{ij}^{t+1} \leftarrow$ slice sample from $\Pr_j^t(u_{ij}|u_{i(-j)}, \pi_i)$
      end for
    until Gibbs convergence
  end for
  (Variational M-step)
  for all alternatives j **do**
    (KDE estimation of $\Pr_j$)
    $\Pr'_j(x) \leftarrow \mathbb{I}_{x\in(0,1)} \sum_i \exp\left\{\frac{-(x-u_{ij}^{t+1})^2}{2h^2}\right\}$
    $\Pr_j^{t+1}(x) \leftarrow \Pr'_j(x) / \int_0^1 \Pr'_j(x)dx$
  end for
  $t \leftarrow t+1$
until Convergence of all $\Pr_j^t$
return Joint KDE on the $n \times m$ matrix of latent u_{ij}

1. Capture heterogeneity and correlation of alternatives in rank data

2. Predict out-of-sample data and pairwise matrices

3. Complete ranks

Table 2.1: Our datasets. † denotes a subset of the full data

	Rank Type	m	n
Election [112]	Top Partial	10	380
Nascar [67]	Sub Partial	7†	36
Sushi [69]	Full	10	5000

We used log likelihood for test data as well as total variation distance and mean squared error for the metrics of prediction power.

Simulations have been performed in R on an i5 3.30GHz Intel(R). We contribute the R package StatRank [7] for existing methods.

2.7.1 Capturing heterogeneity and correlation

Heterogeneity: The heterogeneity of the utility distribution for an alternative represents diversity of opinion. To understand this heterogeneity, we fit various RUMs to 5000 data points of the Sushi data and plot the estimated marginal utility distributions for five alternatives in Figure 2.7.

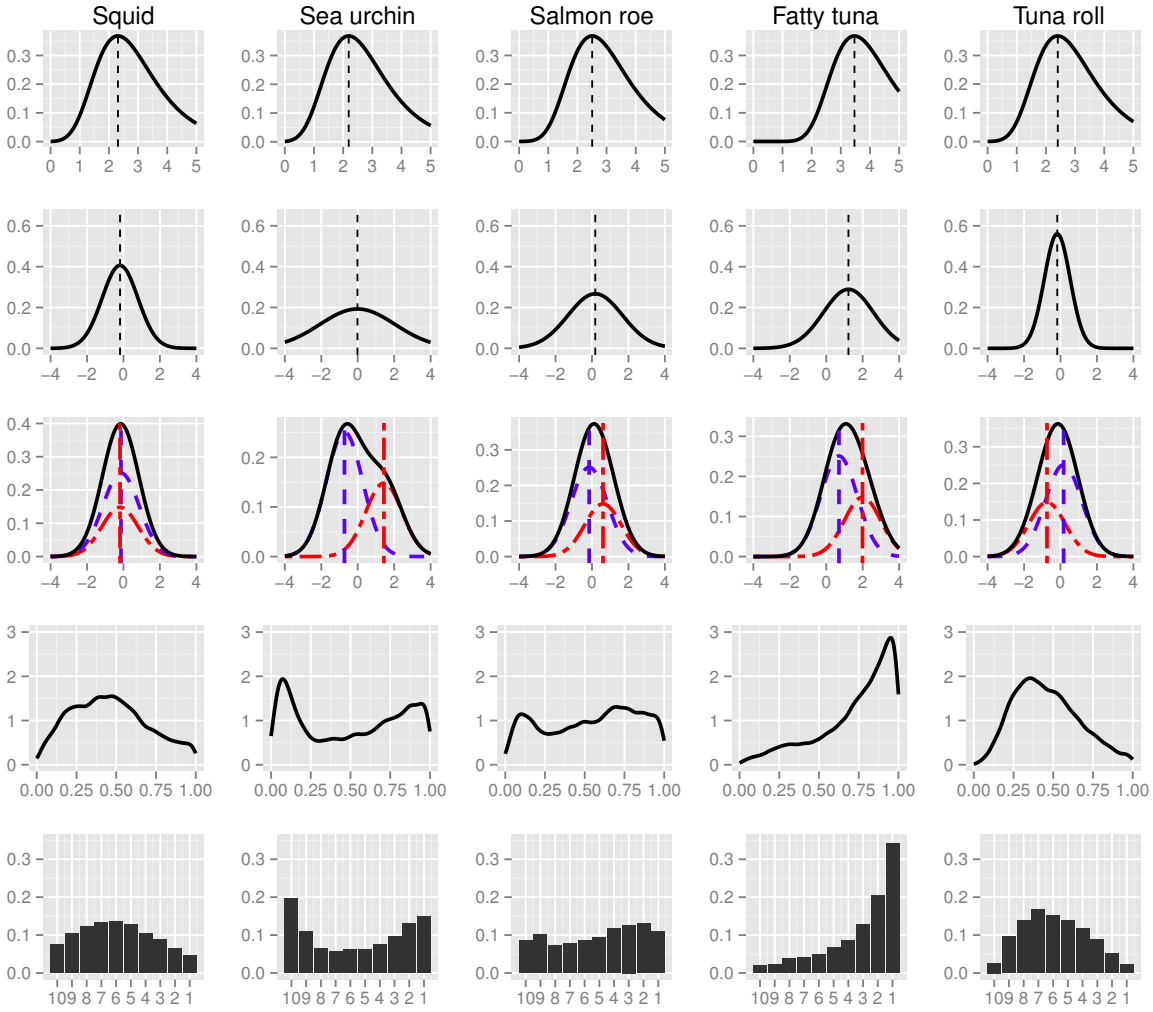


Figure 2.7: (top to bottom) Plackett-Luce RUM, Normal different variances (DV) RUM, 2x Normal fixed variances (FV) RUM (variance is fixed to 1), NPRUM, Empirical distribution of the sushi dataset. The x-axis denotes the utilities and the y-axis denotes the densities.

Generally, the richer the possible space of models, the richer the data sets that a RUM can encode. With more parameters, a model can go beyond capturing only the location parameter

of utilities (e.g. go beyond Gumbel). A model can also capture multi-modality and differing variances across alternatives. The most notable example is the utility of the sea urchin sushi in Figure 2.7.

Comparing the empirical distribution and NPRUM within Figure 2.7, we notice that the estimated utility distributions are very similar to the empirical rank distributions. As mentioned in Section 2.5.3, we know that the empirical rank distribution given all observations (agents) can be a good approximation of a possible random utility distribution.

Utility Correlation: A key benefit of NPRUM over existing RUM methods and the two extensions is NPRUM's ability to capture the correlation structure between utilities. Figure 2.8 illustrates this correlation structure for two pairs of sushi. We believe the two modes in the joint distribution of salmon roe and sea urchin utility correspond to two different types of agents. One type ranks both high, while the other ranks both low. Similarly, we see agents that tend to like fatty tuna tend to dislike cucumber roll sushi. Modeling correlation allows to better understand agents' taste preferences, and will assist in rank completion.

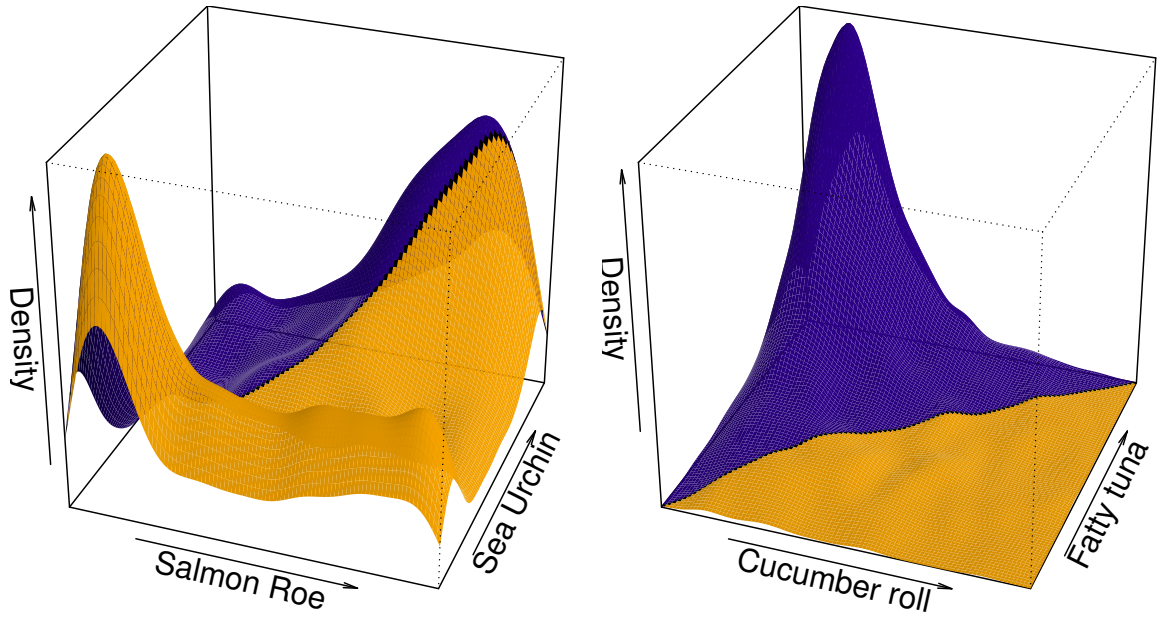


Figure 2.8: Joint distribution for two sets of positively correlated (salmon roe and sea urchin) and negatively correlated (cucumber roll and fatty tuna) sushi. The orange region represents the preference of salmon roe over sea urchin or cucumber roll over fatty tuna, respectively.

2.7.2 Rank distribution prediction via smoothing

As discussed in Section 2.5.3, estimating the rank distribution of rank data has been performed for small n. However, rank distribution can be useful in many contexts. For example, we might want to estimate *What will be the demand for this sushi?* using the rank distribution as non parametric approach.

The empirical rank distribution is not a good estimate for the true rank distribution because of noise. Instead, we smooth out the noise by fitting a RUM. After fitting the RUM, we recreate ranked data by drawing a large number of samples from the model. As we see by comparing the top and bottom rows of Figure 2.9 with the actual rank distribution in Figure 2.7, the smoothed data is a better estimate of the rank distribution. In order to

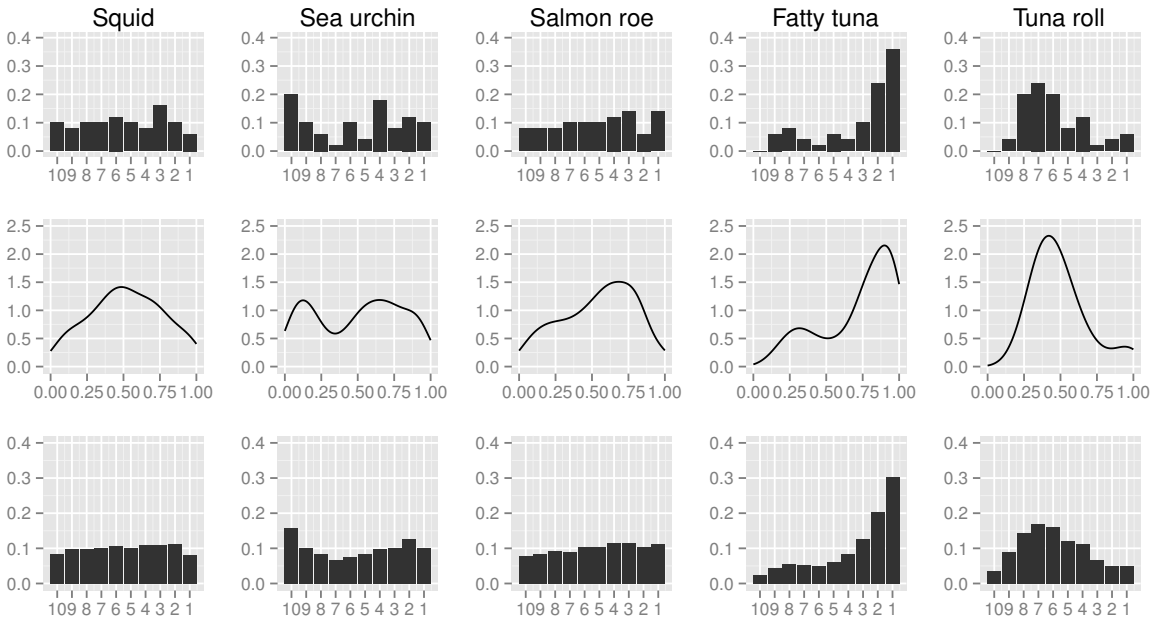


Figure 2.9: (top) Empirical rank distribution of first 50 sushi agents. (middle) NPRUM fit on first 50 sushi agents. (bottom) rank distribution of 5000 *simulated* agents drawn from NPRUM fit on first 50 sushi agents.

explore this concretely, we compare NPRUM with the following other RUMs in their ability to estimate rank distributions:

- **Empirical:** Unsmoothed rank distribution as a baseline.

- **Plackett-Luce:** Gumbel RUM
- **2 x Normal Fixed Variance (FV):** Each agent is in one of two "types" with a certain probability. The two types each have a different multivariate normal distribution (with covariance matrix $\mathbb{I}$) for the joint utility density.
- **Normal Different Variance (DV):** The alternatives each have independent normally-distributed utilities with different variances.

We measure the success of smoothing by comparing the smoothed rank distribution from a random $n = 50$ or 100 agents from the sushi dataset with the rank distribution of the remaining $5000 - n$. We use total variation distance (TVD) between the rank distributions as our metric, with

$$\delta(P, Q) = \frac{1}{2}||P - Q||_1$$

where Q is the smoothed rank distribution of the original n agents, and P is the rank distribution of the remaining $5000 - n$ agents. We present the results of this experiment in Figure 2.10. In the case where $n = 50$, a bandwidth of 0.12 outperforms all other bandwidths

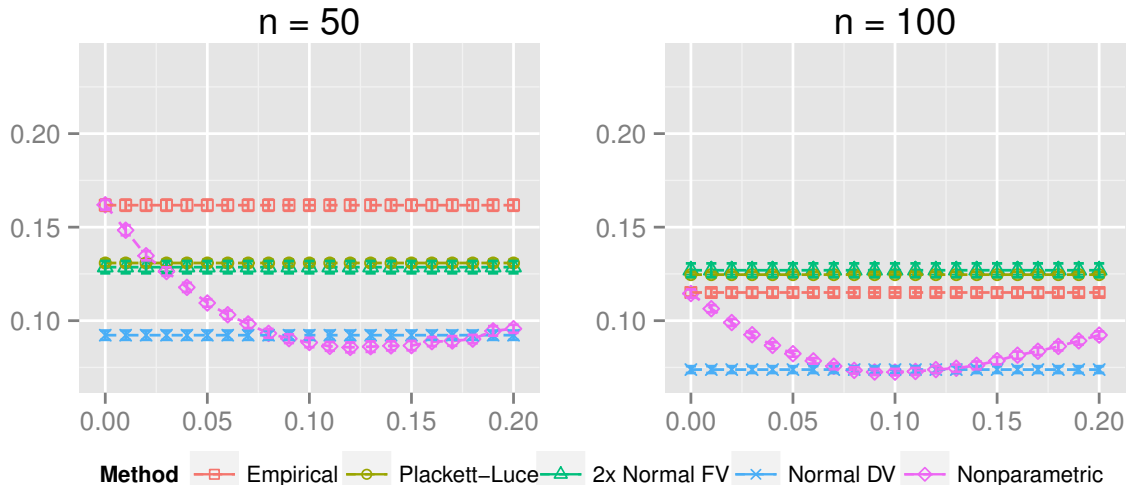


Figure 2.10: Rank distribution prediction performance. x-axis is bandwidth (h). y-axis is TVD. 75 repetitions are done for each data point. Error bars represent 95% confidence intervals. n represents the number of agents for which rank distribution was smoothed.

with a TVD of 0.0856 ± 0.0026 (95% interval), which is 7% and 47% less than the TVDs of

Normal RUM and empirical, respectively. Non-parametric RUM for $n = 50$ outperforms all other RUMs with statistical significance ($\alpha = 0.05$) at bandwidths $h \in \{0.10, \ldots, 0.16\}$. We see that NPRUM's advantage is more pronounced with a smaller n.

2.7.3 RUM Comparison Results

To compare predictive and estimation capabilities of various RUM models, we establish two metrics. The first metric, average log-likelihood, evaluates both in-sample and out-of-sample fit. The second metric measures error in estimating the pairwise matrix P. In this matrix, $p_{ij}(\forall i \neq j)$ is the probability that alternative i is preferred over alternative j. We use the same procedure from the previous section involving TVD on the pairwise matrices.

We compare Plackett-Luce, Normal FV, 2x Normal FV, and Normal DV, to the proposed non-parametric RUM. For the pairwise matrix metrics, we also include the error metrics for the "Empirical" model, where the model matrix is exactly the preference matrix of the training dataset. We run each model and dataset pair for 20 repetitions and 20 iterations each,[5] and report the mean and standard error for each metric. Our results are shown in Table 2.2.

Table 2.2: (top) Average log likelihood. (bottom) Total variation distance between pairwise matrices. Numbers in bold are significantly better than other methods. * means that the method does not converge

Method	Election Test	Nascar Test	Sushi Test
Plackett-Luce	**-5.98 (3e-02)**	-4.43 (5e-02)	-14.37 (1e-02)
Normal FV	-7.44 (3e-02)	-6.89 (3e-01)	-14.06 (1e-02)
2 x NormalFV	-8.41 (3e-02)	**-4.17 (3e-02)**	-14.21 (2e-02)
Normal DV	-7.66 (2e-02)	*	**-13.96 (1e-02)**
Plackett-Luce	14.51 (6e-02)	5.83 (3e-02)	4.35 (3e-02)
Normal FV	6.16 (4e-02)	3.07 (2e-02)	5.85 (4e-02)
2 x NormalFV	5.64 (5e-02)	**2.80 (2e-02)**	4.94 (4e-02)
Normal DV	5.27 (7e-02)	*	5.29 (6e-02)
Empirical	**4.68 (4e-02)**	3.19 (2e-02)	**3.86 (3e-02)**

We note that the non-parametric outperforms the parametric RUMs on every out-of-sample metric for all of the data-sets. In the Sushi data, Normal DV outperforms NPRUM on

[5]Converging methods need fewer than 10 iterations. We chose to run 20 iterations for all methods to have a fair time comparison.

Table 2.3: Runtime (seconds). Numbers in bold are significantly better than other methods. * means the method does not converge.

Method	Election	Nascar	Sushi
Plackett-Luce	28390 (2e+02)	930 (8e+00)	**150 (1e+00)**
Normal FV	28570 (1e+02)	920 (3e+00)	13680 (7e+01)
2x Normal FV	39120 (2e+02)	1910 (9e+00)	22280 (1e+02)
Normal DV	27570 (1e+02)	*	13610 (7e+01)
NP (h = .11)	**210 (1e+00)**	**60 (3e-01)**	180 (8e-01)

in-sample log-likelihood but NPRUM outperforms Normal DV on out-of-sample log-likelihood, which is evidence that Normal DV may have overfit to the training set. In the same data, the same behavior is evident when comparing 2x Normal FV to Normal FV. 2x Normal FV outperforms in training but not in the test set.

The non-parametric method takes significantly less time than any other method on any given data-set (with the exception of PL on Sushi). Estimation of parameters for PL model for Nascar and Sushi data was done with the MM algorithm [67] which is faster than the general MC-EM algorithm.

We have additional experimental results with more RUMs and more data-sets.

2.7.4 Rank Completion

We can apply the propose RUMs to rank completion, a recommendation problem where we may want to predict the full rankings for an agent given observed partial rankings.

We design an experiment where given an agent's top-ranked sushi, we predict the agent's second-ranked sushi. From the n-agent training set, we estimate the conditional distribution $\Pr(\pi(2)|\pi(1))$ for each first-ranked distribution. We calculate the TVD between this predicted conditional distribution and the actual conditional distribution on the $5000-n$ agents used as test data. We take the average of the conditional TVDs as our performance metric, weighted by the frequency of each first-ranked alternative.

We show in Figure 2.11 the performance of the existing RUM methods at this rank completion problem. Interestingly, we note that the parametric RUMs barely improve when we increase sample size from $n = 50$ to $n = 100$. NPRUM's advantage widens with more data

because NPRUM is the only existing RUM able to capture correlation, which is vital for rank completion.

Normal DV does not capture correlation, we believe the flexible variance structure is the reason for good performance. Our rank completion question can be generalized to answer a wide variety of recommendation and customization questions.

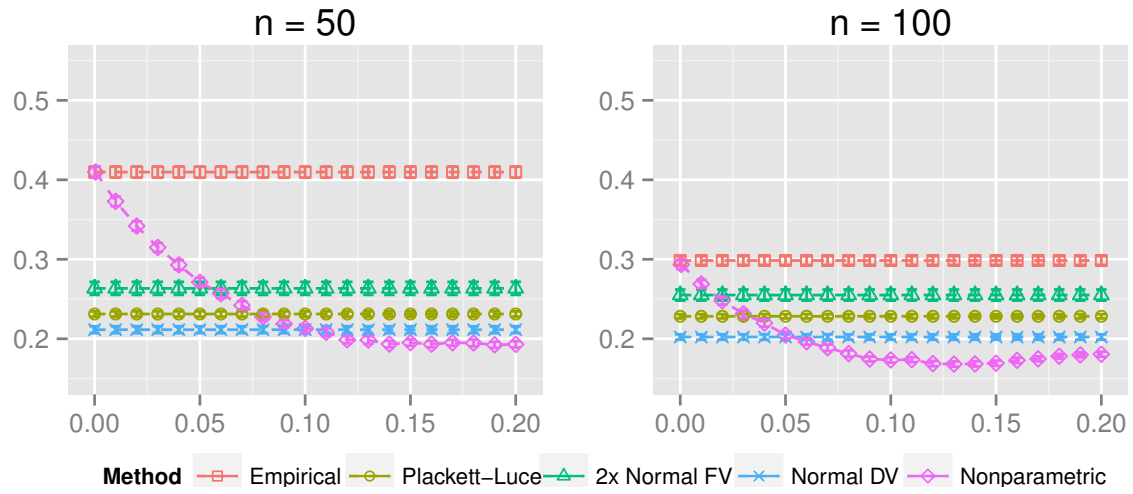


Figure 2.11: Rank completion performance. x-axis is bandwidth (h). y-axis is weighted mean TVD. 100 repetitions are done for each data point. Error bars represent 95% confidence intervals. n represents the number of agents used as training for rank completion.

2.8 Discussion

Here we discuss the advantages and disadvantages of different extensions of RUMs. Naturally, increasing the model complexity and relaxing assumptions comes with challenges involving estimation and inference.

2.8.1 Distributional assumptions (inductive bias)

NPRUM's weak assumptions (weak inductive bias) regarding continuity and smoothing make it more generally applicable than RUMs with distributional and independence assumptions. However, assumptions are useful in certain settings. For example, PL outperforms Normal for

Election data, but Normal outperforms PL for Sushi data (see Table 2.2). Hence, one model's assumptions may be more correct than others. However, NPRUM outperforms both PL and Normal in our data-sets, indicating that NPRUM's weak assumptions work better than the strong ones of PL and Normal RUM.

2.8.2 Estimation

The MC-EM algorithm is used for all methods. We compare the complexity of MC-EM in the parametric and non-parametric settings.

Time Complexity: In the E-step, sampling from truncated the parametric and non-parametric distributions can be accomplished via similar techniques. This leads to similar run time. We believe that our implementation of MC-EM for NPRUM is more efficient, leading to better running times in comparison to existing methods (Table 2.2).

In the M-step, fitting utility densities for Exponential Family distributions [92] is simple because of the relationship between the sufficient statistics and the MLE parameters. Fitting the non-parametric model is more difficult as it requires kernel density estimation, a choice of kernel (fixed at Gaussian for this chapter), and a bandwidth. Identifying the distribution in the M-step of a parametric RUM is $O(mn)$, while identifying the KDE in the M-step of NPRUM is $O(dmn)$, where d is the number of evaluation points we want in a dimension. d can be a large constant.

Space Complexity: Representing a parametric RUM is storing m location parameters for a Plackett-Luce model or $2m$ parameters for a Normal model. The non-parametric model needs to be represented by the original vectors of utilities from the agents, which is proportional in size to the data.

This leads us to conclude that parametric RUMs are $O(m)$ in space complexity while NPRUMs are $O(mn)$. The other option for representing NPRUM, storing values of the density function on a lattice grid, quickly becomes unfeasible with many alternatives (exponential in m, leading to curse of dimensionality).

2.8.3 Inference

Tasks such as identifying the maximal posterior probability ranking and specifying the distribution over ranks are intractable because of the $m!$ size of the permutation space.

However, distributional assumptions such as independence from irrelevant alternative (IIA) in the Luce model allows maximal posterior probability rankings and distributions over ranks to be found easily. Pairwise preferences are also found easily in Normal and PL RUMs.

For NPRUM, we must rely on Monte-Carlo and resampling methods to perform these inferential tasks. Integration and summarizing properties of multivariate kernel density estimates is difficult, but sampling from multivariate kernel density estimates is easy.

2.9 Conclusions

This chapter describes a framework to estimate extensions of classical RUM models. We provide three extensions to establish effectiveness of the method. The extensions are designed to capture different aspects of the data such as heterogeneity, multiple types in the data and nonparametric representation of distribution on rankings.

Our work is a comprehensive study of various RUMs with different evaluation metrics. This evaluation has been done for multiple predictive metrics, including rank position distribution prediction, out-of-sample average log-likelihood, and rank completion. We find that RUMs are flexible enough to capture the best features in every setting, leading to superior performance against existing RUMs for description, interpretation and prediction. The parametric extensions, on the other hand, provide a descriptive model with interpretation of the parameters for the data.

We provide an application to rank data, where we can complete an agent's partial ranks (useful for recommendation systems). NPRUM outperforms existing RUMs in rank completion. This is a result of a more expressive latent utility model that accounts for features such as correlation, which is imperative in any rank model that seeks to complete ranks given

partial data. We have better empirical results compared to prior work on clustering and other RUMs.

For future work, we think it is interesting to look to adopt regularization when fitting a model, in order to insist that, for each data point, the data point is assigned with high probability to a particular component.

Chapter 3

Generalized Method of Moments Estimators for RUMs

3.1 Introduction

In many applications, we need to aggregate the preferences of agents[1] over a set of alternatives to produce a joint ranking. For example, in systems for ranking the quality of products, restaurants, or other services, we can generate an aggregate rank through feedback from individual users. This idea of *rank aggregation* also plays an important role in multi-agent systems, meta-search engines [44], belief merging [46], crowdsourcing [79], and many other e-commerce applications.

A standard approach towards rank aggregation is to treat input rankings as data generated from a probabilistic model, and then learn the MLE of the input data. As described in Chapter 1 , this idea has been explored in both the machine learning community and the (computational) social choice community. The most popular statistical models are the Bradley-Terry-Luce model (BTL for short) [29, 77], the Plackett-Luce model (PL for short) [99, 77], the random utility model [111], and the Mallows (Condorcet) model [78, 37]. In machine learning, researchers have focused on designing efficient algorithms to estimate parameters

[1]We will consider that ranks are generated from agents, but the approach is applicable to any rank data.

for popular models; e.g. [67, 74, 11]. This line of research is sometimes referred to as *learning to rank* [72].

However, for many parametric ranking models the MLE is hard to compute. For example, computing MLE for the Mallows models is $\mathsf{P}_{\|}^{\mathsf{NP}}$-complete [62]. Among the *Random Utility Models (RUMs)*, only the *Plackett-Luce (PL)* model [99, 77] is known to have an analytical solution to the likelihood function. Some previous work has focused on computing specific parametric ranking models. For example, Hunter [67] propose a Minorize-Maximization (MM) algorithm for MLE in the PL model. In the former chapter we proposed a Monte-Carlo Expectation-Maximization (MC-EM) algorithm to compute MLE for a general class of RUMs. While this extends the computational reach to more expressive RUMs beyond PL, the running time may still be too large for data sets of practical interest.

Recently, Negahban et al. [96] proposed a rank aggregation algorithm, called *Rank Centrality* (RC), based on computing the stationary distribution of a Markov chain whose transition matrix is defined according to the data (pairwise comparisons among alternatives). The authors describe the approach as being model independent, and prove that for data generated according to BTL, the output of RC converges to the ground truth, and the performance of RC is almost identical to the performance of MLE for BTL. Moreover, they characterized the convergence rate and showed experimental comparisons. However, their method is used for pairwise rank data and can not be applied to full ranks.

Another alternative to MLE is to adopt a *Generalized Method of Moments (GMM)* algorithm for estimation.We introduce the idea of *rank-breaking* as a way to apply GMM to full ranking data. In rank-breaking, each ranking in the data is decomposed into a subset of pairwise comparisons, to which GMM is then applied; e.g., for example we might take the statistics used for GMM as a count of all pairs of alternatives that appear in first position and second position, or we can consider all possible pairs of positions (this is called *full* breaking).

Rank breaking is of interest because it can allow for estimation methods that are considerably quicker than MLE. We fully characterize conditions for a breaking to provide a GMM that is consistent for PL. Consistency is a desired statistical property that says as the size of

data generated according to a model within the class assumed by the estimator grows without bound, the output of the estimator converges to the true parameters. We answer the question about how to extend rank-breaking to other parametric ranking models beyond PL as well.

Finding consistent, partial breakings is interesting because computing the statistics that are used for GMM becomes the bottleneck as the size of datasets grows.

3.2 Our Contributions

The main contribution is to introduce a class of GMMs for parameter estimation in RUMs. As a summary, we explore the idea of breaking for a general set of distributions and we address these questions. For the first question we propose a GMM algorithm (Algorithm 4) for any model in the *location family* of RUMs, which includes PL and Normal-RUM and develop a general condition for when the breaking will provide a consistent estimator. We provide a trichotomy theorem that characterizes what is required for *single-edge breakings*, which are simple breakings with only a particular pair of rank positions, to be consistent.

The proposed algorithms first *break* full rankings into pairwise comparisons, and then solve the generalized moment conditions to find the parameters. Each GMM is characterized by a way of breaking full rankings. We characterize conditions for the output of the algorithm to be unique, and obtain characterizations about which method of breaking leads to a consistent GMM. Specifically, *full breaking* (which uses all pairwise comparisons in the ranking) is consistent for all RUMs, but *adjacent breaking* (which only uses pairwise comparisons in adjacent positions) is inconsistent for PL model. Full breaking is the only consistent approach for models with mean-symmetric utility distributions. In addition, we fully characterize the class of consistent breakings for the widely studied PL model, and establish that the natural approach of adjacent breaking is not consistent.

We characterize the computational complexity of our GMMs, and show that the asymptotic complexity is better than for the classical Minorize-Maximization (MM) algorithm for MLE in the PL model [67].

We first reveal a new and natural connection between the RC algorithm [96] and the BTL

model by showing that RC algorithm can be interpreted as a GMM estimator applied to the BTL model. Technically our technique in the case of pairwise ranking is related to the random walk approach [96]. However, we note that our algorithms aggregate *full rankings* under PL and RUMs in general, while the RC algorithm aggregates *pairwise comparisons* solely for the BTL model. Therefore, it is quite hard to directly compare our GMMs and RC fairly since they are designed for different types of data. Moreover, by taking a GMM point of view, we prove the consistency of our algorithms on top of theories for GMMs, while Negahban et al. proved the consistency of RC directly.

We compare statistical efficiency and running time of proposed methods experimentally using both synthetic and real-world data. All GMMs run much quicker than the MM algorithm and MC-EM algorithm. For the synthetic data, we observe that many consistent GMMs converge as quick as the MM algorithm, while there exists a clear tradeoff between computational complexity and statistical efficiency among consistent GMMs for PL model.

3.3 Preliminaries

Let $\mathcal{A} = \{a_1, \ldots, a_m\}$ denote the set of alternatives. Let $D_r = (d_1, \ldots, d_n)$ denote the data, where each d_j is a full ranking over $\mathcal{A}$. Let $\mathcal{L}(\mathcal{A})$ denote the set of all full rankings (that is, all antisymmetry, transitive, and complete binary relationships) over $\mathcal{A}$. For any $d \in \mathcal{L}(\mathcal{A})$ and any pair of alternatives a, a', we $a \succ_d a'$ if and only if a is preferred to a' in d, i.e., $(a, a') \in d$. In a *parametric ranking model* $\mathcal{M}_r$, we let $\Omega \subseteq \mathbb{R}^s$ denote the parameter space and for any $\vec{\gamma} \in \Omega$, let $\Pr_{\mathcal{M}_r}(\cdot|\vec{\gamma})$ denote a distribution over $\mathcal{L}(\mathcal{A})$. Sometimes the subscript in $\Pr_{\mathcal{M}_r}$ is omitted when it does not cause confusion.

3.3.1 Random Utility Models (RUMs)

In a RUM, each alternative a is characterized by a utility distribution μ_a, parameterized by a vector $\vec{\gamma}_a$. Given any ground truth $\vec{\gamma} = (\vec{\gamma}_1, \ldots, \vec{\gamma}_m)$, an agent generates a full ranking over $\mathcal{A}$ in the following way: she independently samples a random utility U_j for each alternative a_j with conditional distribution $\Pr_a(\cdot|\vec{\gamma}_a)$, then ranks the alternatives according to their respec-

tive perceived utilities, such that she prefers a to a' if and only if $U_a > U_{a'}$.[2] The probability for a ranking d is the following, where $d(j)$ is the index of the alternative ranked in the jth position:

$$\Pr(d|\vec{\gamma}) = \Pr(U_{d(1)} > U_{d(2)} > \ldots > U_{d(m)})$$

In this chapter, the *location family* refers to the class of RUMs where each distribution is only parameterized by its mean. In other words, the shapes of utility distributions are fixed, though they are not necessarily identical for each alternative. A *homogeneous location family* is a location family where the shapes of the distributions are identical.[3] We study homogeneous location families with the following distributions:

• Gumbel distribution with $\lambda = 1$, whose PDF is $\Pr_G(x) = e^{-x}e^{-e^{-x}}$: the corresponding homogeneous location family is PL.

The PL model is a parametric model where each alternative c_i is parameterized by $\gamma_i \in (0, 1)$, such that $\sum_{i=1}^{m} \gamma_i = 1$. Let $\vec{\gamma} = (\gamma_1, \ldots, \gamma_m)$ and Ω denote the parameter space. Let $\bar{\Omega}$ denote the closure of Ω. That is, $\bar{\Omega} = \{\vec{\gamma} : \forall i, \gamma_i \geq 0 \text{ and } \sum_{i=1}^{m} \gamma_i = 1\}$. Given $\vec{\gamma}^* \in \Omega$, the probability for a ranking $d = [c_{i_1} \succ c_{i_2} \succ \cdots \succ c_{i_m}]$ is defined as follows,

$$\Pr_{\text{PL}}(d|\vec{\gamma}) = \frac{\gamma_{i_1}}{\sum_{l=1}^{m} \gamma_{i_l}} \times \frac{\gamma_{i_2}}{\sum_{l=2}^{m} \gamma_{i_l}} \times \cdots \times \frac{\gamma_{i_{m-1}}}{\gamma_{i_{m-1}} + \gamma_{i_m}}$$

In the BTL model, the data is composed of pairwise comparisons instead of rankings, and the model is parameterized in the same way as PL, such that $\Pr_{\text{BTL}}(a_{i_1} \succ a_{i_2}|\vec{\gamma}) = \frac{\gamma_{i_1}}{\gamma_{i_1} + \gamma_{i_2}}$. BTL can be thought of as a special case of PL via marginalization, since $\Pr_{\text{BTL}}(a_{i_1} \succ a_{i_2}|\vec{\gamma}) = \sum_{d:a_{i_1} \succ a_{i_2}} \Pr_{\text{PL}}(d|\vec{\gamma})$. In the rest of the chapter, we denote $\Pr = \Pr_{\text{PL}}$.

• Flipped Gumbel distribution: the PDF is $\Pr_G(-x)$, where $\Pr_G$ is the PDF of the Gumbel distribution with $\lambda = 1$. Fliped Gumbel is not the same as the Gumbel distribution. However it can be seen as a Gumbel distribution case where the smaller the x the better the alternative in ranking (e.g. a latent space x can be the time each horse takes to finish the race in a horse

[2] We ignore the case of ties where $U_a = U_{a'}$ since this happens with negligible probability for popular utility distributions.

[3] In this chapter we will use $\Pr(d|\vec{\gamma})$ and $\Pr(d)$ exchangeably.

race competition).

• Normal distribution: no analytic solution to the likelihood function is known. The MC-EM algorithm proposed for this case is accurate however, we propose a quick algorithm for this case.

3.3.2 Generalized Method-of-Moments

The *Generalized Method-of-Moments (GMM)* provides a wide class of algorithms for parameter estimation. In GMM, we are given a parametric model whose parametric space is $\Omega \subseteq \mathbb{R}^s$, an infinite series of $q \times q$ matrices $\mathcal{W} = \{W_n : n \geq 1\}$, and a column-vector-valued function $g(d, \vec{\gamma}) \in \mathbb{R}^q$.

For any vector $\vec{h} \in \mathbb{R}^q$ and any $q \times q$ matrix W, let $\|\vec{h}\|_W = (\vec{h})^T W \vec{h}$. For any data D_r, let $g(D_r, \vec{\gamma}) = \frac{1}{n} \sum_{d \in D_r} g(d, \vec{\gamma})$. The GMM method computes parameters $\vec{\gamma}' \in \Omega$ that minimize $\|g(D_r, \vec{\gamma}')\|_{W_n}$:

$$\begin{aligned} &\mathrm{GMM}_g(D_r, \mathcal{W}) = \\ &\qquad \{\vec{\gamma}' \in \Omega : \|g(D_r, \vec{\gamma}')\|_{W_n} = \inf_{\vec{\gamma} \in \Omega} \|g(D_r, \vec{\gamma})\|_{W_n}\} \end{aligned} \tag{3.1}$$

Since Ω may not be compact (as in PL), the set of parameters $\mathrm{GMM}_g(D_r, \mathcal{W})$ can be empty. A GMM is *consistent* if and only if for any $\vec{\gamma}^* \in \Omega$, $\mathrm{GMM}_g(D_r, \mathcal{W})$ converges in probability to $\vec{\gamma}^*$ as $n \to \infty$ when the data is drawn i.i.d. given $\vec{\gamma}^*$.

In this chapter, we let $W_n = I$ for all n. Let $\| \cdot \|_2$ denote the L-2 norm. Equation (3.1) becomes

$$\mathrm{GMM}_g(D_r) = \{\vec{\gamma}' \in \Omega : \|g(D_r, \vec{\gamma}')\|_2 = \inf_{\vec{\gamma} \in \Omega} \|g(D_r, \vec{\gamma})\|_2\} \tag{3.2}$$

It is well-known that $\mathrm{GMM}_g(D, \mathcal{W})$ is consistent if it satisfies some regularity conditions plus the following condition [58]:

Condition 2 *$E_{d|\vec{\gamma}^*}[g(d, \vec{\gamma})] = 0$ if and only if $\vec{\gamma} = \vec{\gamma}^*$.*

Example 1 MLE as a consistent GMM: *Suppose the likelihood function is twice-differentiable, then the MLE is a consistent GMM where $g(d, \vec{\gamma}) = \bigtriangledown_{\vec{\gamma}} \log \Pr(d|\vec{\gamma})$ and $W_n = I$.*

Example 2 *Negahban et al. [96] proposed the* Rank Centrality (RC) *algorithm that aggregates pairwise comparisons* $D_P = \{Y_1, \ldots, Y_n\}$.[4] *Let* a_{ij} *denote the number of* $c_i \succ c_j$ *in* D_P *and it is assumed that for any* $i \neq j$, $a_{ij} + a_{ji} = k$. *Let* d_{max} *denote the maximum pairwise defeats for an alternative. RC first computes the following* $m \times m$ *column stochastic matrix:*

$$P_{RC}(D_P)_{ij} = \begin{cases} a_{ij}/(kd_{max}) & \text{if } i \neq j \\ 1 - \sum_{l \neq i} a_{li}/(kd_{max}) & \text{if } i = j \end{cases}$$

Then, RC computes $(P_{RC}(D_P))^T$*'s stationary distribution* $\vec{\gamma}$ *as the output.*

Let $X^{c_i \succ c_j}(Y) = \begin{cases} 1 & \text{if } Y = [c_i \succ c_j] \\ 0 & \text{otherwise} \end{cases}$ *and* $P^*_{RC}(Y) = \begin{cases} X^{c_i \succ c_j} & \text{if } i \neq j \\ -\sum_{l \neq i} X^{c_l \succ c_i} & \text{if } i = j \end{cases}$. *Let* $g_{RC}(d, \vec{\gamma}) = P^*_{RC}(d) \cdot \vec{\gamma}$. *It is not hard to check that the output of RC is the output of* $GMM_{g_{RC}}$. *Moreover,* $GMM_{g_{RC}}$ *satisfies Condition 2 under the BTL model, and as we will show later in Corollary 4,* $GMM_{g_{RC}}$ *is consistent for BTL.*

3.4 Breakings

A *rank-breaking* (*breaking* for short) B_G is defined as a function $\mathcal{L}(\mathcal{A}) \to 2^{\{a \succ a' : a, a' \in \mathcal{A}\}}$ that is represented by an undirected graph G. The vertices of G correspond to the m *positions* in a full ranking. For any full ranking $d = [a_{i_1} \succ a_{i_2} \succ \cdots \succ a_{i_m}]$, $B_G(d) = \{a_{i_j} \succ a_{i_l} : a_{i_j} \succ_d a_{i_l}$ and $\{j, l\} \in G\}$. That is, B_G breaks d into pairwise comparisons for all pairs of alternatives at position j and l such that $\{j, l\}$ is an edge in G. If G only contains a single edge, then B_G is called a *single-edge breaking*.[5]

We extend the B_G definition to apply to data D, so for any data D_r composed of full rankings, we let $B_G(D_r) = \bigcup_{d \in D_r} B_G(d)$ where the union is in multiset sense.

Intuitively, a breaking is an undirected graph over the m positions in a ranking, such that for any full ranking d, the pairwise comparisons between alternatives in the ith position and

[4]The BTL model applied in [96] is slightly different from our model. Therefore, in this example we adopt an equivalent description of the RC algorithm.

[5]The direction is implicit in graph G; e.g., edge 2-4 will only ever generate a count for the alternative in position 2 being ahead of that in position 4. It doesn't also include a count for the one in position 4 being behind the one in position 2.

jth position are counted to construct $P_G(d)$ if and only if $\{i,j\} \in G$.

Definition 1 *A* breaking *is a non-empty undirected graph G whose vertices are $\{1,\ldots,m\}$. Given any breaking G, any full ranking d over $\mathcal{C}$, and any $c_i, c_j \in \mathcal{C}$, we let*

- $X_G^{c_i \succ c_j}(d) = \begin{cases} 1 & \{\text{Pos}(c_i,d), \text{Pos}(c_j,d)\} \in G \text{ and } c_i \succ_d c_j \\ 0 & \text{otherwise} \end{cases}$, *where* $\text{Pos}(c_i,d)$ *is the position of c_i in d.*
- $P_G(d)$ *be an $m \times m$ matrix where* $P_G(d)_{ij} = \begin{cases} X_G^{c_i \succ c_j}(d) & \text{if } i \neq j \\ -\sum_{l \neq i} X_G^{c_l \succ c_i}(d) & \text{if } i = j \end{cases}$
- $g_G(d, \vec{\gamma}) = P_G(d) \cdot \vec{\gamma}$
- *$\text{GMM}_G(D)$ be the GMM method that solves Equation (3.1) for g_G and $W_n = I$.*[6]

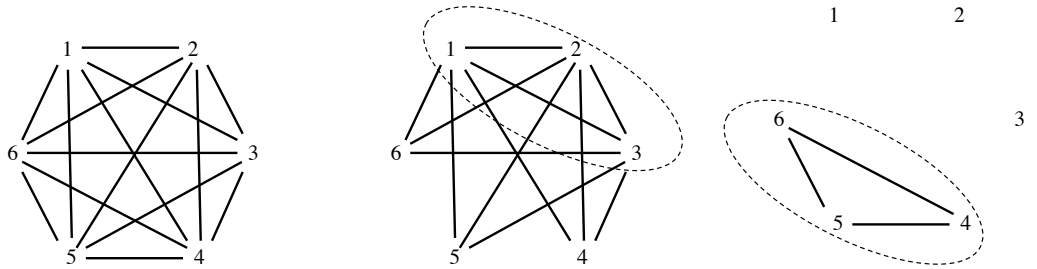
In this chapter, we focus on the following breakings, illustrated in Figure 3.1.

- **Full breaking**: G_F is the complete graph. Example 3 is the GMM with full breaking.
- **Top-k breaking**: for any $k \leq m$, $G_T^k = \{\{i,j\} : i \leq k, j \neq i\}$.
- **Bottom-k breaking**: for any $k \geq 2$, $G_B^k = \{\{i,j\} : i,j \geq m+1-k, j \neq i\}$.[7]
- **Adjacent breaking**: $G_A = \{\{1,2\},\{2,3\},\ldots,\{m-1,m\}\}$.
- **Position-k breaking**: for any $k \geq 2$, $G_P^k = \{\{k,i\} : i \neq k\}$. Intuitively, the full breaking contains all the pairwise comparisons that can be extracted from each agent's full rank information in the ranking; the top-k breaking contains all pairwise comparisons that can be extracted from the rank provided by an agent when she only reveals her top k alternatives and the ranking among them; the bottom-k breaking can be computed when an agent only reveals her bottom k alternatives and the ranking among them; and the position-k breaking can be computed when the agent only reveals the alternative that is ranked at the kth position and the set of alternatives ranked in lower positions.

We note that $G_T^m = G_B^m = G_F$, $G_T^1 = G_P^1$, and for any $k \leq m-1$, $G_T^k \cup G_B^{m-k} = G_F$, and $G_T^k = \bigcup_{l=1}^k G_P^l$. We are now ready to present our GMM algorithm (Algorithm 3) parameterized by a breaking G.

[6] To simplify notation, we use GMM_G instead of GMM_{g_G}.

[7] We need $k \geq 2$ since G_B^k is empty.



(a) Full breaking. (b) Top-3 breaking. (c) Bottom-3 breaking.

(d) Adjacent breaking. (e) Position-2 breaking.

Figure 3.1: Example breakings for $m = 6$.

3.5 Generalized Method-of-Moments for the Plakett-Luce model

In this section we introduce our GMMs for rank aggregation under PL model. The PL model is a good model to start with because of its simplicity and wide application. In our methods, $q = m$, $W_n = I$ and g is linear in $\vec{\gamma}$. We start with a simple special case to illustrate the idea.

Example 3 *For any full ranking d over $\mathcal{C}$, we let*

- $X^{c_i \succ c_j}(d) = \begin{cases} 1 & c_i \succ_d c_j \\ 0 & \text{otherwise} \end{cases}$
- *$P(d)$ be an $m \times m$ matrix where* $P(d)_{ij} = \begin{cases} X^{c_i \succ c_j}(d) & \text{if } i \neq j \\ -\sum_{l \neq i} X^{c_l \succ c_i}(d) & \text{if } i = j \end{cases}$
- $g_F(d, \vec{\gamma}) = P(d) \cdot \vec{\gamma}$ *and* $P(D) = \frac{1}{n} \sum_{d \in D} P(d)$

For example, let $m = 3$, $D = \{[c_1 \succ c_2 \succ c_3], [c_2 \succ c_3 \succ c_1]\}$. Then $P(D) = \begin{bmatrix} -1 & 1/2 & 1/2 \\ 1/2 & -1/2 & 1 \\ 1/2 & 0 & -3/2 \end{bmatrix}$. *The corresponding GMM seeks to minimize $\|P(D) \cdot \vec{\gamma}\|_2^2$ for $\vec{\gamma} \in \Omega$.*

It is not hard to verify that $(E_{d|\vec{\gamma}^*}[P(d)])_{ij} = \begin{cases} \frac{\gamma_i^*}{\gamma_i^*+\gamma_j^*} & \text{if } i \neq j \\ -\sum_{l\neq i} \frac{\gamma_l^*}{\gamma_i^*+\gamma_l^*} & \text{if } i = j \end{cases}$*, which means that* $E_{d|\vec{\gamma}^*}[g_F(d,\vec{\gamma}^*)] = E_{d|\vec{\gamma}^*}[P(d)]\cdot\vec{\gamma}^* = 0$*. It is not hard to verify that* $\vec{\gamma}^*$ *is the only solution to* $E_{d|\vec{\gamma}^*}[g_F(d,\vec{\gamma})] = 0$*. Therefore,* GMM_{g_F} *satisfies Condition 2. Moreover, we will show in Corollary 3 that* GMM_{g_F} *is consistent for PL model.*

In the above example, we count all pairwise comparisons in a full ranking d to build $P(d)$, and define $g = P(D)\cdot\vec{\gamma}$ to be linear in $\vec{\gamma}$. As aforementioned, we may consider some subset of pairwise comparisons, which leads to the definition of the class of GMMs based on the notion of *breakings*.

Algorithm 3 $\text{GMM}_G(D)$

A breaking G and data $D = \{d_1, \ldots, d_n\}$ composed of full rankings. Estimation $\text{GMM}_G(D)$ of parameters under PL. Compute $P_G(D) = \frac{1}{n}\sum_{d\in D} P_G(d)$ in Definition 2. Compute $\text{GMM}_G(D)$ according to (3.1). Return $\text{GMM}_G(D)$.

Theorem 6 *For any breaking* G *and any data* D*, there exists* $\vec{\gamma} \in \bar{\Omega}$ *such that* $P_G(D)\cdot\vec{\gamma} = 0$.

Theorem 6 implies that in Equation (3.1), $\inf_{\vec{\gamma}\in\Omega} g(D,\vec{\gamma})^T W_n g(D,\vec{\gamma})\} = 0$. Therefore, Step 3 can be replaced by: **3*** Let $\text{GMM}_G = \{\vec{\gamma} \in \Omega : P_G(D)\cdot\vec{\gamma} = 0\}$.

3.5.1 Uniqueness of Solution

It is possible that for some data D, $\text{GMM}_G(D)$ is empty or non-unique. Our next theorem characterizes conditions for $|\text{GMM}_G(D)| = 1$ and $|\text{GMM}_G(D)| \neq \emptyset$. A Markov chain (row stochastic matrix) M is *irreducible*, if any state can be reached from any other state. That is, M only has one communicating class.

Theorem 7 *Among the following three conditions,* 1 *and* 2 *are equivalent for any breaking* G *and any data* D*. Moreover, conditions* 1 *and* 2 *are equivalent to condition* 3 *if and only if* G *is connected.*

1. $(I + P_G(D)/m)^T$ *is irreducible.*
2. $|GMM_G(D)| = 1$.

3. $GMM_G(D) \neq \emptyset$.

Corollary 1 *For the full breaking, adjacent breaking, and any top-k breaking, the three statements in Theorem 7 are equivalent for any data D. For any position-k (with $k \geq 2$) and any bottom-k (with $k \leq m-1$), 1 and 2 are not equivalent to 3 for some data D.*

Ford, Jr. [47] identified a necessary and sufficient condition on data D for the MLE under PL to be unique, which is equivalent to condition 1 in Theorem 7. Therefore, we have the following corollary.

Corollary 2 *For the full breaking G_F, $|GMM_{G_F}(D)| = 1$ if and only if $|MLE_{PL}(D)| = 1$.*

3.5.2 Consistency

We say a breaking G is *pairwise consistent* (for RUMs), if GMM_G is consistent.

Theorem 8 *A breaking G is pairwise consistent if and only if $E_{d|\vec{\gamma}^*}[g(d, \vec{\gamma}^*)] = 0$, which is equivalent to the following equalities:*

$$\text{for all } i \neq j, \qquad \frac{\Pr(c_i \succ c_j | \{\text{Pos}(c_i, d), \text{Pos}(c_j, d)\} \in G)}{\Pr(c_j \succ c_i | \{\text{Pos}(c_i), \text{Pos}(c_j)\} \in G)} = \frac{\gamma_i^*}{\gamma_j^*}. \tag{3.3}$$

Theorem 9 *Let G_1, G_2 be a pair of pairwise consistent breakings.*

1. *If $G_1 \cap G_2 = \emptyset$, then $G_1 \cup G_2$ is also consistent.*
2. *If $G_1 \subsetneq G_2$ and $(G_2 \setminus G_1) \neq \emptyset$, then $(G_2 \setminus G_1)$ is also consistent.*

Continuing, we show that position-k breakings are pairwise consistent, then use this and Theorem 9 as building blocks to prove additional consistency results.

Proposition 1 *For any $k \geq 1$, the position-k breaking G_P^k is pairwise consistent.*

We recall that $G_T^k = \bigcup_{l=1}^k G_P^l$, $G_F = G_T^m$, and $G_B^k = G_F \setminus G_T^{m-k}$. Therefore, we have the following corollary.

Corollary 3 *The full breaking G_F is pairwise consistent; for any k, G_T^k is pairwise consistent, and for any $k \geq 2$, G_B^k is pairwise consistent.*

Theorem 10 *Adjacent breaking G_A is pairwise consistent if and only if all components in $\bar{\gamma}^*$ are the same.*

Lastly, the technique developed in this section can also provide an independent proof that the RC algorithm is consistent for BTL, which is implied by the main theorem in [96]:

Corollary 4 *[96] The RC algorithm is consistent for BTL.*

RC is equivalent to $GMM_{g_{RC}}$, which satisfies Condition 1. By checking conditions that are analogues to those in the proof of Theorem 11, we can prove that $GMM_{g_{RC}}$ is consistent for BTL.

For the case of PL model, the results in this section suggest that if we want to learn the parameters of PL, we should use pairwise consistent breakings, including full breaking, top-k breakings, bottom-k breakings, and position-k breakings. The adjacent breaking seems quite natural, but it is not pairwise consistent, thus will not provide a good estimate to the parameters of PL. This will also be verified by experimental results in Section 4.6. We will provide results on GMM for some other cases of RUMs as well.

3.5.3 Complexity

Proposition 2 *The computational complexity of the MM algorithm for PL [67] and our GMMs are listed below.*

- **MM:** $O(m^3 n)$ *per iteration.*
- **GMM (Algorithm 3) with full breaking:** $O(m^2 n + m^{2.376})$*, with* $O(m^2 n)$ *for breaking and* $O(m^{2.376})$ *for computing step* 2^* *in Algorithm 3 (matrix inversion).*
- **GMM with adjacent breaking:** $O(mn + m^{2.376})$*, with* $O(mn)$ *for breaking and* $O(m^{2.376})$ *for computing step* 2^* *in Algorithm 3.*
- **GMM with top-k breaking:** $O((m+k)kn+m^{2.376})$*, with* $O((m+k)kn)$ *for breaking and* $O(m^{2.376})$ *for computing step* 2^* *in Algorithm 3.*

It follows that the asymptotic complexity of the GMM algorithms is better than for the classical MM algorithm. In particular, the GMM with adjacent breaking and top-k breaking

for constant k's are the quickest. However, we recall that the GMM with adjacent breaking is not consistent, while the other algorithms are consistent. We would expect that as data size grows, the GMM with adjacent breaking will provide a relatively poor estimation to $\vec{\gamma}^*$ compared to the other methods.

Moreover in the statistical setting in order to gain consistency we need regimes that $m << n$ and large ns are going to lead to major computational bottlenecks. All the above algorithms (MM and different GMMs) have linear complexity in n, hence, the coefficient for n is essential in determining the tradeoffs between these methods. As it can be seen above the coefficient for n is linear in m for top-k breaking and quadratic for full breaking while it is cubic in m for the MM algorithm. This difference is illustrated through experiments in Figure 4.5.

3.6 A GMM Algorithm for the Location Family of RUM

We recall that in the location family, each utility distribution has only one parameter (its mean). Therefore, we can write $\vec{\gamma} = (\gamma_1, \ldots, \gamma_m)$, where for any $i \leq m$, γ_i is the mean parameter of the utility distribution for a_i. W.l.o.g. let $\gamma_m = 0$.

To specify the GMM, it suffices to specify the moment conditions. Given a parametric ranking model $\mathcal{M}_r$ in the location family, for any two alternatives $a \neq a'$, any $\vec{\gamma} \in \Omega$, and any breaking B_G, we let $f_G^{aa'}(\vec{\gamma})$ denote the probability that given $\vec{\gamma}$, $a \succ a'$ in $B_G(d)$. That is, $f_G^{aa'}(\vec{\gamma}) = \Pr_{\mathcal{M}_r}(a \succ a' \in B_G(d)|\vec{\gamma})$. When $G = G_F$, that is, G is the complete graph, we use shorthand notation $f^{aa'} = f_G^{aa'}$. Since the perceived utilities are generated independently, $f^{aa'}$ is a function of $\gamma_a - \gamma_{a'}$. Therefore, we sometimes write $f^{aa'}(\gamma_a - \gamma_{a'})$. We note that in general $f_G^{aa'}$ may depend on other components of $\vec{\gamma}$.

Definition 2 *Given any breaking B_G, any $d \in \mathcal{L}(\mathcal{A})$, and any $a, a' \in \mathcal{A}$, we let:*

- $X_G^{a \succ a'}(d) = \begin{cases} 1 & a \succ a' \in B_G(d) \\ 0 & \textit{otherwise} \end{cases}$ *, and*
- $X_G^{a \succ a'}(D_r) = \frac{1}{n} \sum_{d \in D_r} X_G^{a \succ a'}(d)$

In words, $X_G^{a\succ a'}(D_r)$ is the normalized frequency of times that alternative a is preferred to alternative a' (i.e., $a \succ a'$). By definition, $E[X_G^{a\succ a'}(d)] = f_G^{aa'}$. We now present the moment conditions used in our algorithm, and then comment on why we do not use other seemingly more natural ones. Our moment conditions are: for $a \neq a'$,

$$g_G^{aa'}(d, \vec{\gamma}) = X_G^{a\succ a'}(d) \times f^{a'a}(\vec{\gamma}) - X_G^{a'\succ a}(d) \times f^{aa'}(\vec{\gamma}) \tag{3.4}$$

We are now ready to present our algorithm as Algorithm 4. We note that in (3.4) we use $f^{aa'}$

Algorithm 4 $\text{GMM}_G(D_r)$

For all a, a', compute $X_G^{a\succ a'}(D_r)$.
Compute $\text{GMM}_G(D_r)$ according to (3.2) using the moment conditions in (3.4) (e.g. using gradient descent).
return $\text{GMM}_G(D_r)$.

and $f^{a'a}$ instead of $f_G^{aa'}$ and $f_G^{a'a}$. Therefore it is not immediately clear whether the moment conditions equal to 0 in expectation for a graph G that is not the complete graph. The next definition provides a condition used to guarantee that when a pairwise consistent breaking G is used in Algorithm 4, the moment conditions (3.4) equal to 0 in expectation.

Definition 3 *A breaking B_G is* consistent *for a location family RUM, if G has at least one edge and for any $\vec{\gamma}$ and any $a \neq a'$,*[8]

$$\frac{f_G^{aa'}(\vec{\gamma})}{f_G^{a'a}(\vec{\gamma})} = \frac{f^{aa'}(\vec{\gamma})}{f^{a'a}(\vec{\gamma})}$$

Where,

$$\frac{f^{aa'}(\vec{\gamma})}{f^{a'a}(\vec{\gamma})} = \frac{\Pr_{\mathcal{M}_r}(a \succ a'|\vec{\gamma})}{\Pr_{\mathcal{M}_r}(a' \succ a|\vec{\gamma})}$$

We will be interested in understanding when breakings are consistent. By definition, the full breaking is consistent. Let CDF_a denote the CDF of $\Pr_a(\cdot|0)$. For the location family we

[8] The definition of pairwise consistent breakings is more general than the definition in [8], which was defined only for PL.

have:

$$f^{aa'}(\vec{\gamma}) = f^{aa'}(\gamma_a - \gamma_{a'}) = \int_{-\infty}^{\infty} \Pr_{a'}(y)(1 - \mathrm{CDF}_a(y - \gamma_a + \gamma_{a'}))dy \tag{3.5}$$

We have the following proposition for $f^{aa'}(\gamma_a - \gamma_{a'})$.

Proposition 3 *For any model in the location family where each utility distribution has support $(-\infty, \infty)$, $f^{aa'}$ is monotonic increasing (as a function of $\gamma_a - \gamma_{a'}$) on $(-\infty, \infty)$ with $\lim_{x\to-\infty} f^{aa'}(x) = 0$ and $\lim_{x\to\infty} f^{aa'}(x) = 1$. Moreover, if $\Pr_a$ and $\Pr_{a'}$ are continuous then $f^{aa'}$ is continuously differentiable with $f^{aa'}(x)' = \int_{-\infty}^{\infty} \Pr_{a'}(y) \Pr_a(y - x)dy$.*

Theorem 11 *For any model in the location family with (possibly) inhomogeneous distributions and any pairwise consistent breaking B_G, if the PDF of every utility distribution is continuous, then Algorithm 4 is consistent.*

Proof: We prove the theorem by verifying the conditions in Theorem 2.1 in [58].

Assumption 2.1: The distribution on D is stationary and ergodic. This holds because in any RUM, data in D are generated i.i.d.

Assumption 2.2: Ω is a separable metric space. Since $\mathbb{R}^m$ is a metric separable space and Ω is an subset of $\mathbb{R}^m$, Ω is also separable.

Assumption 2.3: $g_G^{aa'}(\cdot, \vec{\gamma})$ is Borel measurable for any $a \neq a'$ and each $\vec{\gamma} \in \Omega$ and $g_G^{aa'}(d, \cdot)$ is continuous on Ω for each d. Since the domain of $g_G^{aa'}(\cdot, \vec{\gamma})$ is discrete, $g_G^{aa'}(\cdot, \vec{\gamma})$ is continuous, which means that $g_G^{aa'}(\cdot, \vec{\gamma})$ is Borel measurable. We note that $g_G^{aa'}(d, \cdot)$ is linear in $f^{aa'}(\vec{\gamma})$ and by Proposition 3, $f^{aa'}$ is continuous in $\vec{\gamma}$.

Assumption 2.4: $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma})]$ exists and is finite for all $\vec{\gamma} \in \Omega$, and $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma}^*)] = 0$. The former is because $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma})]$ is linear in $f^{aa'}(\vec{\gamma})$ and by Proposition 3, $f^{aa'}(\Omega)$ is bounded above by 1. The second part holds because $E_{d|\vec{\gamma}^*}[X_G^{a \succ a'}(d)] = f_G^{aa'}(\vec{\gamma}^*)$, which means that $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d, \vec{\gamma}^*)] = f_G^{aa'}(\vec{\gamma}^*) f^{a'a}(\vec{\gamma}^*) - f_G^{a'a}(\vec{\gamma}^*) f^{aa'}(\vec{\gamma}^*) = 0$.

Assumption 2.5: The sequence $\mathcal{W}$ converges almost surely to a positive semi-definite matrix. This holds since $W_n = I$ for all t.

Premise (1): $g_G^{aa'}(d, \vec{\gamma})$ is first moment continuous. Since $|g_G^{aa'}(d, \vec{\gamma})| \leq 2$, by Lemma 2.1 of [58], we have that $g_G^{aa'}(d, \vec{\gamma})$ is first moment continuous.

Premise (2): Ω is compact, which is the assumption of our theorem.

Premise (3): $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d,\vec{\gamma})]$ has a unique zero at $\vec{\gamma}^*$. By Proposition 3 we have that $f^{aa'}(\gamma_a - \gamma_{a'})$ is monotonic increasing in $\gamma_a - \gamma_{a'}$ and $f^{a'a}(\gamma_{a'} - \gamma_a)$ is monotonic increasing in $\gamma_{a'} - \gamma_a$. Therefore, $\dfrac{f^{aa'}(\gamma_a - \gamma_{a'})}{f^{a'a}(\gamma_a - \gamma_{a'})}$ is monotonic increasing in $\gamma_a - \gamma_{a'}$. Hence if $\vec{\gamma}'$ is another zero point for $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d,\vec{\gamma})]$ with $\gamma'_m = 0$, then we must have that for all pairs (a,a'), $\gamma'_a - \gamma'_{a'} = \gamma^*_a - \gamma^*_{a'}$. Given that $\gamma'_m = \gamma^*_m = 0$, this means that $\vec{\gamma}' = \vec{\gamma}^*$, which is a contradiction. Therefore, $\vec{\gamma}^*$ is the only zero point of $E_{d|\vec{\gamma}^*}[g_G^{aa'}(d,\vec{\gamma})]$. □

A direct result of the above theorem is that, for any pairwise consistent breaking B_G for PL, RUM with flipped Gumbel distributions, and RUM with Normal distributions (e.g. the full breaking), Algorithm 4 is consistent for PL, RUM with flipped Gumbel distributions, and RUM with Normal distributions respectively.

Compared to the MC-EM algorithm [11], Algorithm 4 runs quicker since optimizing Equation (3.2) is much easier through e.g., gradient descent or Newton-Raphson. This is because $f^{aa'}(x)'$ is usually easy to compute, and sometimes has a concise analytic solution, as shown in the following example. Breaking is particularly helpful here since it enables an analytic expression for gradient.

Example 4 *Consider RUM with normal distributions whose variances are* 1. *For any* $a \neq a'$ *we have:*

$$f^{aa'}(x)' = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} e^{-\frac{(y-x)^2}{2}} dy = \frac{1}{2\sqrt{\pi}} e^{-\frac{x^2}{4}}$$

A similar formula exists for location families with normal distributions whose variances are not identical.

Why do we use the moment conditions in (3.4)? The following moment conditions seem to be more natural.

$$\begin{aligned} &g_G^{aa'}(d,\vec{\gamma}) = \\ &\quad X_G^{a\succ a'}(d) \times f_G^{a'a}(\vec{\gamma}) - X_G^{a\succ a'}(d) \times f_G^{aa'}(\vec{\gamma}) \end{aligned} \tag{3.6}$$

The only difference between (3.6) and (3.4) is that the former uses $f_G^{aa'}$ and $f_G^{a'a}$ while the latter uses $f^{aa'}$ and $f^{a'a}$. However, for models in the location family, optimizing (3.6) is often

hard due to the lack of analytical solutions to $f_G^{aa'}$ or $(f_G^{aa'})'$. As shown in Example 4, $(f^{aa'})'$ is easy to compute. This is the main reason we choose (3.4) over (3.6).

Why are we interested in breakings beyond the full breaking? The optimization problem (2) is $m-$dimensional, but requires as input the counts in equation 3.4 to be computed for every ordered pair of alternatives. Computing these counts scales a $O(m^2n)$ for full breaking but as $O(mn)$ for adjacent breaking or position-k breaking. For large n this can become the bottleneck with the difference between $O(m^2n)$ and $O(mn)$ making a meaningful difference and starting to become the bottleneck in computation [8]. In such cases we may would prefer to use a partial breaking and explore the tradeoff between computational efficiency and statistical efficiency. However, it is important to do this while maintaining consistency of the estimator.

3.7 Which Breakings are Consistent?

This section provides theoretical results on the consistency of partial breakings (breakings which take only part of the available ranks) for the location family. We will first present the theorems, then introduce four lemmas in Section 3.7.1, and finally in Section 3.7.2 use them as building blocks to provide proofs for the theorems. We start with the following positive results.

Theorem 12 *For PL, a breaking B_G is consistent if and only if G is the union of position-k breakings.*

In a similar way the following Theorem holds if we change PL to PL*.

Theorem 13 *For the RUM with flipped Gumbel distributions (PL*), B_G is consistent if and only if G is the union of position*-k breakings.*

Theorem 12 gives a complete characterization of pairwise consistent breakings for PL (thus answering an open question in [8]) and Theorem 13 gives a complete characterization of pairwise consistent breakings for the RUM with flipped Gumbel distributions.

Theorem 14 *Let $\mathcal{M}_r$ be a model in the (possibly) inhomogeneous location family where each utility distribution has support $(-\infty, \infty)$. If the PDF of each utility distribution in $\mathcal{M}_r$ is symmetric around its mean, then the only pairwise consistent breaking is the full breaking.*

Since the normal distribution is symmetric, we immediately have the following corollary of Theorem 14.

Corollary 5 *For the RUM with Normal distributions (the variances are not necessary identical), the only pairwise consistent breaking is the full breaking.*

Theorem 14 and Corollary 5 tell us that for certain natural models in the location family, the only pairwise consistent breaking is the full breaking. This will also be demonstrated by experimental results in the next section. The next theorem provides a quick check to see if the full breaking is the only pairwise consistent breaking by just checking the $m = 3$ case.

Theorem 15 *For any model in the homogeneous location family where each utility distribution has support $(-\infty, \infty)$, if the full breaking is the only pairwise consistent breaking for $m = 3$, then the full breaking is the only pairwise consistent breaking for any m.*

The last result of this section is a trichotomy theorem for single-edge breakings to be consistent for the homogeneous location family.

Theorem 16 *For any m and any model in the homogeneous location family (with support $(-\infty, \infty)$), exactly one of the following holds.*

1. *No single-edge breaking is consistent.*
2. *Among all single-edge breakings, only $\{1, 2\}$ is consistent.*
3. *Among all single-edge breakings, only $\{m-1, m\}$ is consistent.*

This theorem corresponds to a symmetry notion in the specific location family. Using this theorem and Theorem 14 we know that case (1) corresponds to the symmetric location families and we conjecture that the cases (2) and (3) correspond to negative and positive skewness in the location family distributions respectively.

The next example shows that each of the three cases in Theorem 16 (but not any two of them) holds for some natural location family.

Example 5 *By Corollary 5, the location family with normal distributions belongs to Case 1 in Theorem 16; by Theorem 12, PL belongs to Case 2 in Theorem 16; by Theorem 13, PL* belongs to Case 3 in Theorem 16.*

3.7.1 Four Core Lemmas

To prove the theorems we introduce some notation and four core lemmas in this subsection. For any model $\mathcal{M}_r$ in the location family, let $\mathcal{M}_r^*$ denote the model in the location family where the PDF of each distribution (conditioned on the mean parameter being 0) is flipped around the y-axis. That is, for any $i \leq m$ and any x, $\Pr_{\mathcal{M}_r,i}(x|0) = \Pr_{\mathcal{M}_r^*,i}(-x|0)$. For any breaking B_G, we let B_{G^*} denote the breaking such that $(i,j) \in G^*$ if and only if $(m+1-i, m+1-j) \in G$.

Example 6 *PL* is the RUM with flipped Gumbel distribution. Let $\mathcal{M}_N$ denote the RUM with normal distributions. We have $\mathcal{M}_N = \mathcal{M}_N^*$. For any $k \geq 2$, we have $(G_P^k)^* = G_{P^*}^{m-k}$.*

Lemma 3 *For any $\mathcal{M}_r$ in the location family, if B_G is consistent for $\mathcal{M}_r$, then B_{G^*} is consistent for $\mathcal{M}_r^*$.*

For any graph G and any $1 \leq k_1 < k_2 \leq m$, we let $G_{[k_1,k_2]}$ denote the subgraph of G where the vertices $1, \ldots, k_1 - 1$ and $k_2 + 1, \ldots, m$ are removed,and the vertices are renamed to $1, \ldots, k_2 + 1 - k_1$ by subtracting $k_1 - 1$ from all vertices.

Example 7 *For $m = 6$, a breaking B_G and its restriction to $[2,4]$ are shown in Figure 3.2.*

Lemma 4 *For any model $\mathcal{M}_r$ in the location family, if B_G is consistent then for any $1 \leq k_1 < k_2 \leq m$, either $G_{[k_1,k_2]} = \emptyset$, or $B_{G_{[k_1,k_2]}}$ is consistent for any location family for $k_2 - k_1 + 1$ alternatives where the utility distributions can be any combination of $k_2 - k_1 + 1$ utility distributions in $\mathcal{M}_r$.*

Lemma 5 *For any location family where each utility distribution has support $(-\infty, \infty)$, the single-edge breaking $B_{\{\{1,m\}\}}$ is not consistent.*

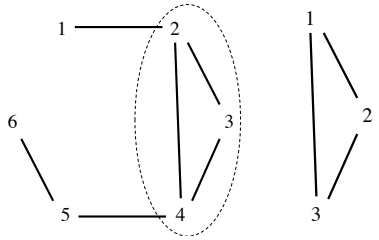


Figure 3.2: A breaking graph G and $G_{[2,4]}$ for $m = 6$.

The last lemma (specifically, part (3), (4), (5)) is a natural extension of Theorem 4 in [8].

Lemma 6 *Let B_{G_1}, B_{G_2} be a pair of breakings.*

- *Suppose both B_{G_1} and B_{G_2} are consistent,*

  (1) if $G_1 \cap G_2 = \emptyset$, then $B_{G_1 \cup G_2}$ is also consistent;

  (2) if $G_1 \subsetneq G_2$, then $B_{G_2 \setminus G_1}$ is also consistent.

- *Suppose B_{G_1} is consistent but B_{G_2} is not consistent,*

  (3) if $G_1 \cap G_2 = \emptyset$, then $B_{G_1 \cup G_2}$ is not consistent;

  (4) if $G_1 \subsetneq G_2$, then $B_{G_2 \setminus G_1}$ is not consistent.

  (5) if $G_2 \subsetneq G_1$, then $B_{G_1 \setminus G_2}$ is not consistent.

Proof: The proof is based on the following two observations. 1) If $G_1 \cap G_2 = \emptyset$, then $f^{aa'}_{G_1 \cup G_2}(d) = f^{aa'}_{G_1}(d) + f^{aa'}_{G_2}(d)$ and $X^{a \succ a'}_{G_1 \cup G_2}(d) = X^{a \succ a'}_{G_1}(d) + X^{a \succ a'}_{G_2}(d)$. 2) If $G_1 \subsetneq G_2$, then $f^{aa'}_{G_1 \setminus G_2}(d) = f^{aa'}_{G_1}(d) - f^{aa'}_{G_2}(d)$ and $X^{a \succ a'}_{G_1 \setminus G_2}(d) = X^{a \succ a'}_{G_1}(d) - X^{a \succ a'}_{G_2}(d)$. □

3.7.2 Proofs of the Theorems

Proof of Theorem 12. The "if" direction was proved in above ??. We now prove the "only if" part by induction on m. When $m = 3$, the theorem obviously holds. Suppose the theorem holds for l. When $m = l + 1$, we first apply Lemma 4 to $G_{[2,m]}$. By the induction hypothesis, $G_{[2,m]}$ must be the union of position-k breakings for some $k \geq 2$. Now apply Lemma 4 to $G_{[1,m-1]}$. There are two cases.

Case 1: for all $i \leq m - 1$, $\{1, i\} \in G$. We claim that $\{1, m\} \in G$. This is because $B_{\{1,m\} \cup G}$ is consistent, and $B_{\{1,m\}}$ is not consistent due to Lemma 5. Hence $B_{G \setminus \{1,m\}}$ is not consistent.

Case 2: for all $i \leq m-1$, $\{1,i\} \notin G$. In this case $\{1,m\} \notin G$ following a similar argument as in Case 1.

This means that the theorem holds for $m = l+1$, which proves the theorem. □

Proof of Theorem 13. The proof follows immediately after Theorem 12 and Lemma 3. □

Proof of Theorem 14. Let B_G denote a pairwise consistent breaking. We prove the theorem by induction on m. When $m = 3$, the full breaking is consistent and by Lemma 5, the single edge-breaking $B_{\{(1,3)\}}$ is not consistent. By Lemma 6 part (5), $B_{\{(1,2),(2,3)\}}$ is not consistent.

We now prove that the single-edge breaking $B_{\{(1,2)\}}$ is not consistent. For the sake of contradiction suppose it is. By Lemma 3, $B_{\{(1,2)\}^*} = B_{\{(2,3)\}}$ is consistent for $\mathcal{M}_r^*$. Since all utility distributions in $\mathcal{M}_r$ are symmetric, $\mathcal{M}_r^* = \mathcal{M}_r$. Therefore, $B_{\{(2,3)\}}$ is consistent for $\mathcal{M}_r$. By Lemma 6 part (1), $B_{\{(1,2),(1,3)\}}$ is consistent, which is a contradiction.

Similarly the single-edge breaking $B_{\{(2,3)\}}$ is not consistent. It follows from Lemma 6 part (5) that $B_{\{(1,2),(1,3)\}}$ and $B_{\{(1,3),(2,3)\}}$ are not consistent. Therefore, the only pairwise consistent breaking for $m = 3$ is the full breaking.

Suppose the theorem holds for $m = l$. When $m = l+1$, we first apply Lemma 4 to $G_{[2,m]}$ and $G_{[1,m-1]}$. By the induction hypothesis, $G_{[2,m]}$ ($G_{[1,m-1]}$) is either empty or the full graph. We have the following two cases.

Since $m > 3$, if $G_{[2,m]}$ is empty, then $G_{[1,m-1]}$ is empty as well. Since G is non-empty, $G = \{(1,m)\}$, which contradicts Lemma 5.

If $G_{[2,m]}$ is full, then $G_{[1,m-1]}$ is full as well. Hence G can be either the full graph G_F, or $G_F \setminus \{(1,m)\}$. By Lemma 5, $B_{\{(1,m)\}}$ is inconsistent, which means that $B_{G_F \setminus \{(1,m)\}}$ is not consistent (Lemma 6 part (5)).

Therefore, the only remaining case is that G is the full breaking, which means that the theorem holds for $m = l+1$, which proves the theorem. □

Proof of Theorem 15. The proof is similar to the proof of Theorem 14. We prove the theorem by induction on m. $m = 3$ is the assumption. Suppose the theorem holds for l. When

n	F−T	M−T	M−F
5	-10^{-4} (10^{-3})	**17 (.05)**	**17 (.05)**
50	.004 (.005)	**198 (1.3)**	**198 (1.3)**
100	**.008 (.0005)**	**359 (11)**	**359 (11)**
150	**.035 (.004)**	**970 (10)**	**970 (10)**
200	**.017 (.0015)**	**1021 (31)**	**1021 (31)**

(a) Run time (seconds).

F−T	M−T	M−F
.09 (.55)	.08 (.57)	**-.01 (.001)**
.27 (.4)	.26 (.37)	**-.01 (.001)**
.08 (.08)	.04 (.08)	**-.04 (.004)**
.34 (.1)	**.33 (.11)**	**-.01 (.001)**
.29 (.027)	**.27 (.022)**	**-.02 (.0057)**

(b) Kendall correlation.

Table 3.1: Paired t-tests for the three algorithms. F, T, M represents values for full breaking, top-3 breaking, and MC-EM, respectively. Mean (std) are shown. Significance results with 95% confidence are in bold.

$m = l + 1$, we first apply Lemma 4 to $G_{[2,m]}$. By the induction hypothesis, $G_{[2,m]}$ is either empty or full.

If $G_{[2,m]}$ is empty, then $G_{[1,m-1]}$ is empty as well. Hence if G is non-empty, then $G = \{(1, m)\}$, which contradicts Lemma 5.

If $G_{[2,m]}$ is full, then $G_{[1,m-1]}$ is full as well. Hence G can be either the full graph G_F, or $G_F \setminus \{(1, m)\}$. By Lemma 5, $B_{\{(1,m)\}}$ is inconsistent, which means that $B_{G_F \setminus \{(1,m)\}}$ is inconsistent (since G_F is always consistent by definition).

Therefore, the theorem holds for $m = l + 1$, which completes the proof. □

Proof of Theorem 16. For any $k_2 > k_1 + 1$, let us first consider $G_{[k_1,k_2]}$. By Lemma 5, $B_{\{(1,k_2-k_1+1)\}}$ is not consistent. Therefore by Lemma 4, any non-adjacent single-edge breaking is not consistent.

Now for an adjacent single-edge graph $\{(k_1, k_1 + 1)\}$ that is different from $\{(1, 2)\}$ and $\{(m-1, m)\}$, by applying Lemma 4 on $G_{[k_1-1,k_1+1]}$ and $G_{[k_1,k_1+2]}$, we have that both $B_{\{\{1,2\}\}}$ and $B_{\{(2,3)\}}$ are consistent for the model in the location family with $m = 3$ and any combination of 3 utility distributions in $\mathcal{M}_r$. By Lemma 6 part (1), $\{(1, 2), (2, 3)\}$ is consistent, which contradicts Lemma 6 part (5) applied to Lemma 5.

Now, we only need to prove that it is impossible for both $B_{\{(1,2)\}}$ and $B_{\{(m-1,m)\}}$ to be consistent. If on the contrary both are consistent, then we apply Lemma 4 on $G_{[1,3]}$ and $G_{[m-2,m]}$. Following a similar argument as in the previous paragraph, we can show a contradiction. This proves the theorem. □

We conjecture that the converse of Theorem 11 holds for natural models in the location

family.

3.8 Experiments

We implemented the MC-EM algorithm , Algorithm 4 with the full breaking, and Algorithm 4 with top-3 breaking for the Normal RUM with fixed variance. We evaluate the algorithms according to run-time and the following two representative criteria. For this, let $\vec{\gamma}^*$ denote the ground truth parameters, and $\vec{\gamma}$ denote the output of the algorithm.

- *Kendall Rank Correlation Coefficient*: Let $K(\vec{\gamma}, \vec{\gamma}^*)$ denote the *Kendall tau distance* between the ranking over components in $\vec{\gamma}$ and the ranking over components in $\vec{\gamma}^*$. The Kendall correlation is $1 - 2\frac{K(\vec{\gamma},\vec{\gamma}^*)}{m(m-1)/2}$.

The synthetic data-sets are generated as follows. Let $m = 5$. The ground truth $\vec{\gamma}^*$ is generated from the Dirichlet distribution Dirichlet($\vec{1}$) which is a distribution on an $m-$dimensional unit simplex. Then, for any given $\vec{\gamma}^*$ we generate up to $n = 200$ full rankings from the location family with normal distributions. All experiments are run on a 2.4 Ghz, Intel Core 2 duo 32 bit laptop.

Table 3.1 (a) shows the paired t-test on running time for the three methods for $n = 5, 50, 100, 150, 200$, where F, T, M represents values for full breaking, top-3 breaking, and MC-EM, respectively. We clearly observe that the running time of Algorithm 4 with full breaking and Algorithm 4 with top-3 breaking are significantly lower than the running time of MC-EM.

Table 3.1 (b) show paired t-tests for the three methods, for Kendall correlation. We note that a higher Kendall correlation means that the estimation is more accurate. Surprisingly, for Kendall correlation, Algorithm 4 with full breaking outperforms MC-EM with 95% confidence for almost all n in our experiments despite that Algorithm 4 runs much quicker. Both algorithms are significantly better than Algorithm 4 with top-3 breaking with 95% confidence when n is not too small. The latter observation is because Algorithm 4 with top-3 breaking is not consistent for the location family with normal distributions.

In the sushi data-set [69], there are 10 kinds of sushi ($m = 10$) and the amount of data

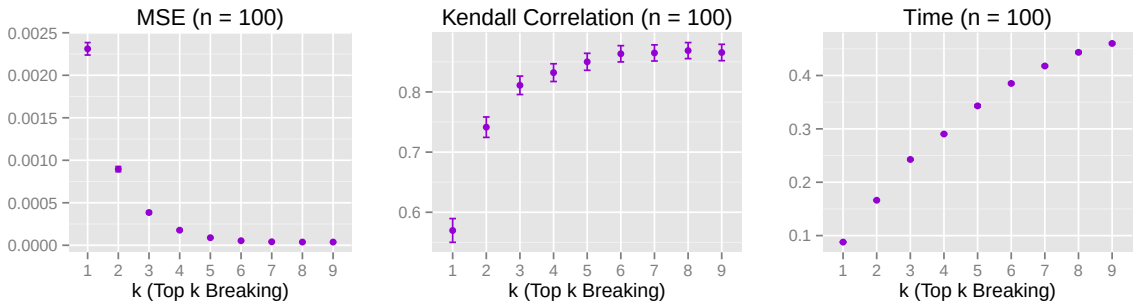


Figure 3.3: Comparison of GMM with top-k breakings as k is varied. The x-axis represents k in the top-k breaking. Error bars are 95% confidence intervals and $m = 10, n = 100$.

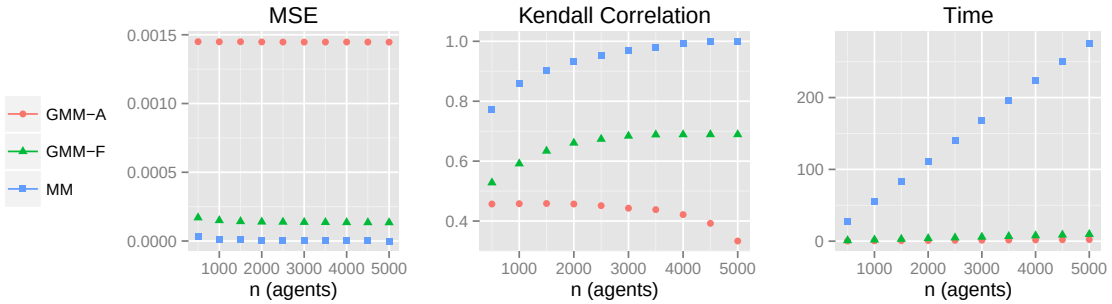


Figure 3.4: The MSE and Kendall correlation criteria and computation time for MM (10 iterations), GMM-F (full breaking), and GMM-A (adjacent breaking) on sushi data.

n is varied, randomly sampling with replacement. We set the ground truth to be the output of MM applied to all 5000 data points. This choice is motivated by providing a comparison of the out of the new algorithm with the MLE estimates. For the running time, we observe the same as for the synthetic data: GMM (adjacent breaking) runs quicker than GMM (full breaking), which runs quicker than MM.

Comparisons for MSE and Kendall correlation are shown in Figure 3.4. In both figures, 95% confidence intervals are plotted but too small to be seen. Statistics are calculated over 1970 trials. For MSE and Kendall correlation, we observe that MM converges quickest, followed by GMM (full breaking), which outperforms GMM (adjacent breaking) which does not converge. Differences between performances are all statistically significant with 95% confidence (with exception of Kendall correlation and both GMM methods for $n = 200$, where $p = 0.07$). This is different from comparisons for synthetic data (Figure 3.5). We

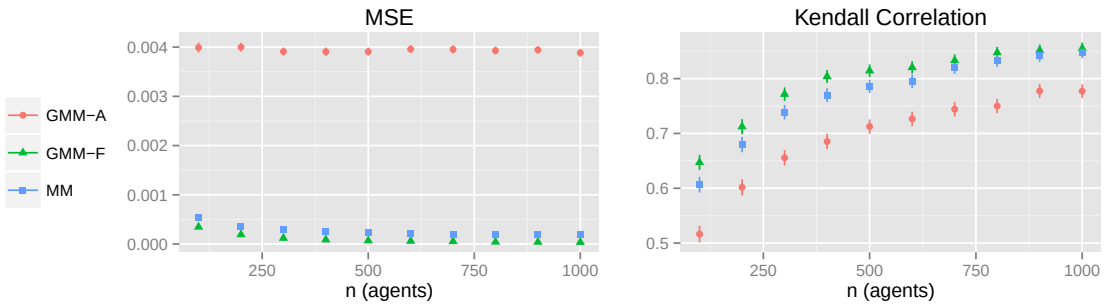


Figure 3.5: The MSE and Kendall correlation of MM (10 iterations), GMM-F (full breaking), and GMM-A (adjacent breaking). Error bars are 95% confidence intervals.

believe that the main reason is because PL does not fit sushi data well. Therefore, we cannot expect that GMM converges to the output of MM on the sushi dataset, since the consistency results (Corollary 3) assumes that the data is generated under PL.

3.9 Conclusions

We studied consistency of rank breaking for random utility models and provides a quick algorithm to compute parameters of these models. The method is based on generalized method of moments and uses a preprocess for turning complex forms of data as permutations to pairwise alternatives. The code is provided in the R package “StatRank" [7].

We plan to extend the algorithms and analysis to partial orders, non-location families such as RUMs parameterized by mean and variance, and to GRUMs [12] and GRUMs with multiple types [10]. The challenge with these settings is finding right conditions for the consistency when the breaking needs to be done on partial rankings.

Chapter 4

Random Utility for Personalized Rank Data and Elicitation

4.1 Introduction

In many situations, we need to know the preferences of agents over a set of alternatives in order to make decisions. For example, in recommender systems, we can compute recommendations of new products for a user based on his reported preferences over some products. In social choice, we need to know agent preferences over alternatives, to make a joint decision about which alternative is socially chosen. Predicting consumer behavior based on reported preferences is an important topic in econometrics [17, 19].

There are two closely related challenges in building a decision support system: preference elicitation and semi-autonomous decision making [26].

Given preferences, the decision making problem can typically be solved through optimization techniques (e.g., computing the choice that minimizes the maximum *regret*). However, there is often a *preference bottleneck*, where it is too costly or even impossible for users to report full information about their preferences. This happens, for example, in airline recommendation systems, where the number of possible itineraries is large [26]. Another instance is combinatorial voting, where agents vote on multiple related issues [70].

To overcome the preference bottleneck, a well accepted approach is *preference elicitation*. This aims to elicit as little as possible of the agents' preferences as is required, to make a good decision. Previous work focused on achieving one of the following two goals:

1. Social choice. We want to make a decision that will affect all agents. Applications include combinatorial auctions [107], voting [40, 75], and crowdsourcing [98].

2. Personalized choice. We want to "learn" an agent's preferences based on a part of her own preferences or preferences of other similar agents. Applications include product configuration [26], matching problems such as public school choice and recommender systems [66] . See [24, 66] for recent developments.

In this chapter, we focus on elicitation for *ordinal preferences*, which means that preferences are represented by rankings. We assume that preferences are generated by *general random utility models (GRUMs)*.

GRUMs are a significant extension of *random utility models (RUMs)* [111], where the effect of attributes of alternatives and agents are not considered. RUMs have been extensively studied and applied in prior work but generally in ways that are specialized to particular parametric forms; e.g., the Bradley-Terry model [29] and the Plackett-Luce model [77, 99].

In a GRUM, an agent's preferences are generated as follows: Each alternative is characterized by a *utility distribution*, and the agents rank the alternatives according to the *perceived utilities*, which are generated from the corresponding utility distributions. Parameters for each utility distribution are computed by a combination of attributes of the alternative and attributes of the agent. Parameters of the GRUM model the interrelationship between alternative attributes and agent attributes. See Section 4.2.1 for more details.

4.1.1 Contributions

We propose a general adaptive method (Algorithm 5) for preference elicitation within the *Bayesian experimental design* framework (see, e.g., [35]), guided by maximum expected information gain. In this chapter, we focus on a special case, where in each step a targeted agent reports her preferences in full.

We target an agent for elicitation who, based on agent attributes, will provide the greatest expected information gain. In addition to using classical criteria in Bayesian experimental design, we also propose two new criteria that are designed to best improve the quality of the inferred rank preferences, one for predicting social choice, and the other for predicting personalized choice.

Directly computing the optimal agent to target next can be challenging due to the lack of efficient algorithms for MAP inference and lack of efficient computation of observed Fisher information [45]. To overcome this, we extend the MC-EM algorithm and conditions for convergence developed for RUMs in Chapter 1 to handle GRUMs. We compute observed Fisher information within the E-step.

We test the proposed methods for MAP/MLE inference and preference elicitation for GRUMs on a synthetic data-set as well as the Sushi data-set [69].

We compare the performance under the new criteria and performance under the standard criteria from Bayesian experimental design literature. Results show that our elicitation framework can significantly improve the precision of estimation for a moderate number of samples in social choice, relative to random agent and some ordering elicitation criteria.

4.1.2 Related Work

GRUMs are a specific case of the generative model studied by Berry, Levinsohn and Pakes [17]. The BLP model explicitly considers unobserved attributes of alternatives and agents, whereas GRUMs only consider observed attributes. The focus of this chapter is to provide a platform for elicitation which has not been considered in the BLP setting.

However, most work on the BLP model has focused on calculating aggregate properties (for example, the demand curve) when a distribution of the values of unobserved attributes are given. Moreover, the methodologies developed in [17] and subsequent papers only work for the *logit model*. That is: the utility distributions are the standard Gumbel distribution, which is a special case. Even when there are no unobserved variables, BLP was not known to be computationally tractable, beyond the logit case.

The approximate method of maximum simulated likelihood has been proposed for GRUMs [118]. We focus on the use of MAP/MLE inference to drive preference elicitation for GRUMs. We developed an MC-EM algorithm for a large class of GRUMs. To the best of our knowledge, this is the first practical algorithm for MAP/MLE inference for general GRUMs, beyond the logit case. We note that RUMs are a special class of GRUMs. Therefore, the new algorithm naturally extends the algorithm developed in Chapter 1 for RUMs. [1]

For social choice, the elicitation scheme designed by Lu and Boutilier [75] aims at computing the outcomes of different commonly studied voting rules. In comparison, the proposed elicitation scheme aims at computing the MAP of GRUMs, which we believe to be different from any commonly studied voting rules.

Compared to the elicitation scheme designed by Pfeiffer et al. [98, 101] for the *Bradley-Terry model*, this chapter focuses on the more general family of GRUMs. Also, as we will see later in the chapter in Example 9, the elicitation scheme by Pfeiffer et al. is closely related to a well studied criterion under the Bayesian experimental design framework called *D-optimality*. The new elicitation framework presented here allows us to use many other classical criteria in Bayesian experimental design, including D-optimality. Experimental results on synthetic data show that D-optimality might not be a good choice for social choice for rankings.

The new elicitation framework considers the attributes of agents and alternatives, allowing for more options for elicitation (e.g. we can target an agent with specific attributes). The proposed method for elicitation is related to the general idea used for the same goal in [66, 34, 25]. However, the proposed method is more general, in the sense that we can handle orders with any length (e.g. Sushi dataset which includes full orders and not only pairwise data). It can also handle any partial order situation due to missing data or design of voting rule (e.g. k first voting or ranks for some missing parties).

[1] Inference and elicitation for GRUMs with unobserved attributes are two interesting directions for future research.

4.2 Preliminaries

In this section, we formally define GRUMs and their corresponding MAP mechanism. Further, we recall basic ideas in Bayesian experimental design.

4.2.1 General Random Utility Models

We consider a preference aggregation setting with a set of alternatives $\mathcal{C} = \{c_1, .., c_m\}$, and multiple agents indexed by $i \in \{1, \ldots, n\}$. In GRUMs, for every $j \leq m$, alternative j is characterized by a vector of $L \in \mathbb{M}$ real numbers, denoted by $\vec{z}_j$. And for every $i \leq n$, agent i is characterized by a vector of $K \in \mathbb{N}$ real numbers, denoted by $\vec{x}_i$.[2] Throughout the chapter, j denotes an alternative, i denotes an agent, l denotes the attribute of an alternative, and k denotes an agent attribute.

The agents' preferences are generated through the following process.Let u_{ij} be agent i's *perceived utility* for alternative j, and let B be a $K \times L$ real matrix that models the linear inter-relation between attributes of alternatives and attributes of agents.

$$u_{ij} = \delta_j + \vec{x}_i B(\vec{z}_j)^T + \epsilon_{ij}, \tag{4.1}$$

$$u_{ij} \sim \Pr(\cdot|\vec{x}_i, \vec{z}_j, \delta_j, B) \tag{4.2}$$

In words, agent i's utility for alternative j is composed of the following three parts:

1. δ_j: The *intrinsic utility* of alternative j, which is the same across all agents;

2. $\vec{x}_i B(\vec{z}_j)^T$: The *agent-specific utility*, where B is the same across all agents;

3. ϵ_{ij}: The random noise generated independently across agents and alternatives.

Given this, an agent ranks the alternatives according to her perceived utilities for the alternatives in the descending order. That is, for agent i, $c_{j_1} \succ_i c_{j_2}$ if and only if $u_{ij_1} > u_{ij_2}$.[3] The

[2] In this chapter we focus on the case where all $\vec{x}_i$ and $\vec{z}_j$ are numerical attributes rather than categorical attributes.

[3] For all reasonable GRUMs the situations with tied perceived utilities have zero probability measure.

parameters for a GRUM are denoted by $\Theta = (\vec{\delta}, B)$. When $K = L = 0$, the GRUM model degenerates to RUM.

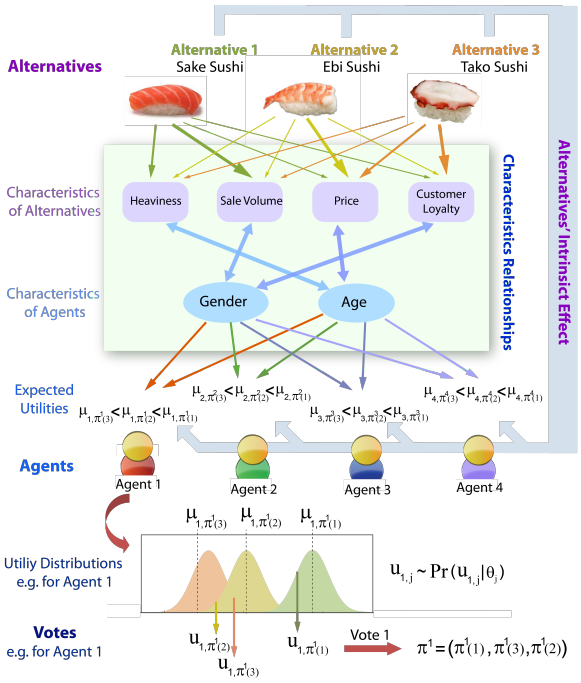


Figure 4.1: The generative process for GRUMs.

Example 8 Figure 4.1 illustrates a GRUM for three alternatives (different kinds of sushi) and n agents. Each alternative is characterized by its attributes including heaviness, price, and custom loyalty. Each agent is characterized by attributes including gender and age. Agent attributes have different relationships with alternative attributes. For instance, a person's salary can be related to a preference in regard to the sushi's price rather than heaviness. The outcome of this relationship is a vector of random utilities, assigned to the alternatives by each agent.

4.2.2 MAP Inference

Given a GRUM, the preference profile is viewed as *data*, $D = \{\pi^1, \ldots, \pi^n\}$, where each π^i is a permutation $(\pi^i(1), \ldots, \pi^i(m))$ of $\{1, \ldots, m\}$ that represents the full ranking $[c_{\pi^i(1)} \succ_i c_{\pi^i(2)} \succ_i \cdots \succ_i c_{\pi^i(m)}]$. We take the standard maximum a posteriori (MAP) approach to

estimate the parameters.

MAP is different from MLE which is used in former chapters. The purpose of MAP is to combine the prior in the estimation as well which is essential in the sequential estimation of parameters.

Recall that each agent's preferences are generated conditionally independently given the parameters Θ. Therefore, in GRUMs, the probability (likelihood) of the data given the ground truth Θ is: $\Pr(D \mid \Theta) = \prod_{i=1}^{n} \Pr(\pi^i \mid \Theta)$, where:

$$\Pr(\pi^i|\Theta) = \int_{u_{i\pi^i(1)}>\cdots>u_{i\pi^i(n)}} \prod_j \Pr(u_{i\pi^i(j)}|\vec{x}_i, \vec{z}_j, \Theta)\ du_{i\pi^i(j)}$$

Suppose we have a prior over the parameters, for MAP inference we aim at computing Θ to maximize the posterior function:

$$\Pr(\Theta|D) = \prod_{i=1}^{n} \Pr(\pi^i \mid \Theta) \Pr(\Theta)$$

After computing Θ^* that maximizes posterior, we can make joint decisions for the agents based on Θ^*.For example, we can choose the winner to be the alternative whose utility distribution has the highest mean, or choose a winning ranking over alternatives by ranking the means of the utility distributions.

4.2.3 One-Step Bayesian Experimental Design

Suppose we have a parametric probabilistic model. Let $\Pr(\Theta^*)$ denote the prior distribution over the parameters. A one-step Bayesian experimental design problem is composed of two parts: a set of *designs* $\mathscr{H}$ and a *quality function* $G(\cdot)$ defined on any *distribution* over the parametric space.

A design $h \in \mathscr{H}$ is mathematically characterized by $\Pr(\cdot|\Theta^*, h)$, which controls the way the data D are generated for any ground truth parameter vector Θ^*. Therefore, for any given design h, we can compute the probability for data D as $\Pr(D|h)$. Given any data D and design h, we can compute the posterior distribution of parameters $\Pr(\cdot|D, h)$. One step refers to the

step by step procedure which chooses an optimal elicitation at every step. The objective of Bayesian experimental design is to choose the design h that maximizes the expected quality of the posterior of MAP parameters, where the randomness comes from the data that are generated given h. Formally, we aim at computing h^* as follows.

$$h^* = \arg\max_h \int G(\Pr(\cdot|D,h)) \times \Pr(D|h)\ dD \tag{4.3}$$

The quality function $G(\Pr(\cdot|D,h))$ represents the performance of the decision process for an observed data and a design. The purpose is to search for designs that provide good performance given all possible data-sets.

Often, directly computing (4.3) is hard. Even $G(\Pr(\cdot|D,h))$ is difficult to compute given D and h. Researchers have taken various approximations to (4.3). A common approach is to approximate $\Pr(\cdot|D,h)$ by a normal distribution $\mathcal{N}(\hat{\Theta}, [R(\hat{\Theta}) + I_h(\hat{\Theta})]^{-1})$, where:

- $\hat{\Theta}$ is the MAP of D,
- $R(\Theta)$ is the *precision matrix* of the prior over Θ, that is, $R = \nabla^2_\Theta \log \Pr(\Theta)$, and
- $I_h(\hat{\Theta})$ is the *Fisher information* matrix defined as follows. Let $X_\pi = \nabla_\Theta \log \Pr(\pi|\vec{\Theta}, h)$, we have

$$I_h(\hat{\Theta}) = E_\pi(X_\pi (X_\pi)^T|_{\Theta=\hat{\Theta}}).$$

Equivalently, if $\log \Pr(\pi|\Theta, h)$ is twice differentiable w.r.t. Θ for each ranking π, then

$$I_h(\hat{\Theta}) = -E_\pi(\nabla^2_\Theta \log \Pr(\pi|\Theta,h)|_{\Theta=\hat{\Theta}}).$$

If we approximate $\Pr(\cdot|D,h)$ by $\mathcal{N}(\hat{\Theta}, [R(\hat{\Theta}) + I_h(\hat{\Theta})]^{-1})$, then the most commonly studied quality functions are functions of $\hat{\Theta}$ and h. More precisely, they are functions of $\hat{\Theta}$ and $R(\hat{\Theta}) + I_h(\hat{\Theta})$. In such cases, we can rewrite $G(\mathcal{N}(\hat{\Theta}, I_h(\hat{\Theta}))) = G^*_R(\hat{\Theta}, h)$. Then, (4.3) becomes:

$$h^* = \arg\max_h \int G^*_R(\hat{\Theta}, h) \cdot \Pr(\hat{\Theta}|h) d\hat{\Theta} \tag{4.4}$$

Still the integration in (4.4) is often hard to compute, and is approximated by $G^*_R(\Theta^*, h)$, where Θ^* is the mode of $\Pr(\Theta)$. Some popular quality functions and corresponding approximations are summarized in Table 4.1.

Name	Quality function	Heuristics $G^*_R(\hat{\Theta}, h)$
D-optimality	Gain in Shannon information	$det(R + I_h(\hat{\Theta}))$
E-optimality	Minimum eigenvalue of the information matrix	$\lambda_{min}\{R + I_h(\hat{\Theta})\}$
social choice	Minimum inverse of pairwise coefficient of variation	Equation (4.5)
personalized choice	Minimum inverse of pairwise coefficient of variation	Equation (4.6)

Table 4.1: Different criteria for experimental design.

Example 9 *The adaptive elicitation approach by Pfeiffer et al. [98] is a special case of Bayesian D-optimality design, where $\mathscr{H}$ is the set of all pairwise questions between alternatives. Pfeiffer et al. derived formulas for $\Pr(\cdot|\Theta^*, h)$ for each $h \in \mathscr{H}$, and chose h^* according to (4.3). The quality function they use is the negative Shannon entropy, which is exactly D-optimality as shown in Table 4.1.*

4.3 A Preference Elicitation Scheme

In applications to preference elicitation, we adapt the one-step Bayesian experimental design to multiple iterations. For any iteration t, let D^t denote the preferences elicited in all previous iterations. The prior distribution $\Pr^t$ over parameters is the posterior of observing D^t, that is: for any Θ, $\Pr^t(\Theta) = \Pr(\Theta|D^t)$. Then we solve a standard one-step Bayesian experimental design problem w.r.t. the prior $\Pr^t$ to elicit a new agents' preferences, and then form D^{t+1} for the next iteration.

Our general elicitation framework for GRUMs is presented as Algorithm 5. To allow flexibility of using various criteria of Bayesian experimental design, we let the input consist of the heuristic $G^*_R(\hat{\Theta}, h)$, which is usually a function of $\hat{\Theta}$ and $R(\hat{\Theta}) + I_h(\hat{\Theta})$. To present the main idea, in this chapter the set of designs $\mathscr{H}$ is the multi-set of all agents attributes. That is, in each iteration (Steps 1∼3) we will compute an $h \in \mathscr{H}$ and query the preferences of a random agent whose attributes are h.[4] Steps 1∼3 are hard to compute. In this chapter,

[4]The elicitation scheme can be extended to other types of elicitation questions, for instance, pairwise

Algorithm 5 Preference Elicitation for GRUMs

Heuristic: $G^*_R(\hat{\Theta}, h)$.
Randomly choose an initial set of data D^1.
for $t = 1$ to T **do**
1: Compute $\Theta^t = \text{MAP}(D^t)$.
2: Compute the precision matrix R^t of $\Pr(\Theta|D^t)$ at Θ^t.
3: Compute $h^t \in \mathscr{H}$ that maximizes $G^*_{R^t}(\Theta^t, h^t)$.
4: Query an agent whose attributes are h^t. Let π^t denote her preferences. $D^{t+1} \leftarrow D^t \cup \{\pi^t\}$, $\mathscr{H} \leftarrow \mathscr{H} \setminus \{h^t\}$.
end for

we will use a multivariate normal distribution $\mathcal{N}(\hat{\Theta}, J_{D^t}(\hat{\Theta})^{-1})$ to approximate $\Pr(\Theta|D^t)$ in Step 2, where $J_{D^t}(\hat{\Theta})$ is the *observed Fisher information* matrix, and we immediately have $R^t = J_{D^t}(\hat{\Theta})$.[5] Given any data D, $J_D(\hat{\Theta}, h)$ is defined as follows. Again, let $\hat{\Theta} = \text{MAP}(D)$.

$$J_{D,h}(\hat{\Theta}) = \sum_{\pi \in D} (X_\pi \times (X_\pi)^T|_{\Theta=\hat{\Theta}}).$$

Equivalently, if $\log \Pr(\pi|\Theta, h)$ is twice differentiable w.r.t. Θ for each ranking π, then we have:

$$J_{D,h}(\hat{\Theta}) = -\sum_{\pi \in D} (\nabla^2_\Theta \log \Pr(\pi|\Theta, h)|_{\Theta=\hat{\Theta}}).$$

In Section 4.4 we propose an MC-EM algorithm to compute $\text{MAP}(D^t)$ in Step 1. In Section 4.4.3 we study how to compute the observed Fisher information matrix $R^t = J_{D^t}(\Theta^t)$, and use it for elicitation as well as accelerating MC-EC algorithm. Computation of the Fisher information matrix $I_h(\hat{\Theta})$ used in Step 3 will also be discussed in Section 4.4.3.

The choice of G^*_R is crucial for the performance of the elicitation algorithm. The two criteria summarized in Table 4.1 are generic criteria for making the posterior as certain as possible, which may not work well for eliciting the aggregated ranking or individual rankings.

4.3.1 A New Elicitation Criterion for Social Choice

The social choice ranking is the ranking over the components of $\vec{\delta}$. Therefore, if the objective is to elicit preferences for the aggregated ranking, it makes sense to make each pairwise compar-

comparisons and "top-k".

[5] See e.g. page 224 [15] for justification of this approximation.

ison as certain as possible. Following the idea in t-test, we propose to use $\frac{|\text{mean}(\delta_{j_1} - \delta_{j_2})|}{\text{std}(\delta_{j_1} - \delta_{j_2})}$ (which is the inverse of *coefficient of variation*) to evaluate the certainty in pairwise comparison between c_{j_1} and c_{j_2}. The larger the value is, the more certain we are about the comparison between c_{j_1} and c_{j_2}. Therefore, we propose to use the following quality function G distributions over Θ. We recall that $\Theta = (\vec{\delta}, B)$.

$$G(\Pr) = \min_{j_1 \neq j_2} \frac{|\text{mean}(\delta_{j_1} - \delta_{j_2})|}{\text{std}(\delta_{j_1} - \delta_{j_2})}.$$

In words, G is the minimum inverse of the coefficient of variation across all pairwise comparisons. The corresponding G_R^* is thus the following.

$$G_R^*(\Theta, h) = \min_{j_1 \neq j_2} \frac{|\text{mean}(\delta_{j_1} - \delta_{j_2})|}{\sqrt{\text{Var}(\delta_{j_1}) + \text{Var}(\delta_{j_2}) + 2\text{cov}(\delta_{j_1}, \delta_{j_2})}}, \tag{4.5}$$

Where $|\text{mean}(\delta_{j_1} - \delta_{j_2})|$ can be computed from Θ and $\sqrt{\text{Var}(\delta_{j_1}) + \text{Var}(\delta_{j_2}) + 2\text{cov}(\delta_{j_1}, \delta_{j_2})}$ can be computed from $R + I_h(\Theta)$.

4.3.2 Generalization to Personalized Choice

Following the idea proposed in the last subsection, we can define a similar quality function $G_{\vec{x}}(\Pr)$ for any agent with attributes $\vec{x}$. This makes the ranking of the alternatives w.r.t. the deterministic parts of the perceived utilities as certain as possible, as follows. For any $j \leq m$, let $\mu_j = \delta_j + \vec{x}B(\vec{z}_j)^T$. We note that μ_j is a linear combination of the parameters in Θ.

$$G_{\vec{x}}(\Pr) = \min_{j_1 \neq j_2} \frac{|\text{mean}(\mu_{j_1} - \mu_{j_2})|}{\text{std}(\mu_{j_1} - \mu_{j_2})} \tag{4.6}$$

$G_{\vec{x}}^*(\Theta, h)$ can be defined in a similar way. However, usually we want to predict the rankings for a population of agents, for which only a distribution over agent attributes is known. Mathematically, let Δ denote a probability distribution over $\mathbb{R}^L$. We can extend the criterion for personalized choice w.r.t. Δ as follows.

$$G_\Delta(\Pr) = \int_{\vec{x} \in \mathbb{R}^T} G_{\vec{x}}(\Pr) \cdot \Delta(\vec{x})\ d\vec{x}.$$

G_Δ is usually hard to compute since it involves integrating $G_{\vec{x}}$ over all $\vec{x}$ in support of Δ, which is often not analytically or computationally tractable. In the experiments, we will use the criterion defined in (4.5) for personalized ranking and surprisingly it works well.

4.4 An MC-EM Inference Algorithm

In this section, we extend MC-EM algorithm for RUMs to GRUMs. We focus on GRUMs where the conditional probability $\Pr(\cdot|\vec{x}_i, \vec{z}_j, \delta_j, B)$ belongs to the *exponential family*, which takes the following form: $\Pr(U = u|\vec{x}_i, \vec{z}_j, \delta_j, B) = e^{\eta_{ij} \cdot T(u) - A(\eta_{ij}) + H(u)}$, where η_{ij} is the vector of *natural parameters*, which is a function of $\vec{x}_i, \vec{z}_j, \Theta$. A is a function of η_{ij} and T and H are functions of u.

Let $U = (\vec{u_1}, \ldots, \vec{u_n})$ denote the latent space, where $\vec{u_i} = (u_{i1}, \ldots, u_{im})$ represent agent i's perceived utilities for the alternatives. The general framework of the proposed EM algorithm is illustrated in Algorithm 6. The algorithm has multiple iterations, and in each iteration there is an E-step and a general M-step with a regression due to the generalization of RUM. Therefore, the algorithm is a *general EM (GEM)* algorithm. We recall that $\Theta = (\vec{\delta}, B)$ represents the parameters.

Algorithm 6 Framework of the EM algorithm

In each iteration.

$$\begin{aligned} &\textbf{E-Step :} \quad Q(\Theta, \Theta^t) \\ &= E_{\vec{U}} \left\{ \log \prod_{i=1}^{n} \Pr(\vec{u}_i, \pi^i|\Theta) + \log(\Pr(\Theta))|D, \Theta^t \right\} \qquad (4.7) \\ &\textbf{M-step :} \quad \text{compute } \Theta^{t+1} \text{ s.t. } Q(\Theta^{t+1}, \Theta^t) > Q(\Theta^t, \Theta^t) \end{aligned}$$

The algorithm builds in the prior in both E-step and M-step and hence it is finding an MAP estimator. The algorithm is performed for a fixed number of iterations or until no Θ^{t+1} in the M-step can be found. However, the E-step cannot be done analytically in general, and we will use a Monte Carlo approximation for the E-step.

4.4.1 Monte Carlo E-Step: Gibbs Sampling

E-step is similar to the E-step in Chapter 2 with a modification that considers the prior. We recall that $\Pr(\cdot|\vec{x}_i, \vec{z}_j, \delta_j, B)$ belongs to the exponential family. We have the following calculation for iteration t, where $\mu_{ij} = \delta_{ij} + \vec{x}_i B(\vec{z}_j)^T$ for any given $\Theta = (\vec{\delta}, B)$, and $\mu_{ij}^t = \delta_{ij}^t + \vec{x}_j B^t(\vec{z}_i)^T$.

$$\begin{aligned} Q(\Theta, \Theta^t) &= E_{\vec{U}}\{\log \prod_{i=1}^{n} \Pr(\vec{U}_i, \pi^i \mid \Theta) + \log \Pr(\Theta) \mid D, \Theta^t\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} E_{u_{ij}}\{\log \Pr(u_{ij}|\Theta) \mid \pi^i, \Theta^t\} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} \eta_{ij} S_{ij}^t - A(\eta_{ij}) + W, \end{aligned}$$

$$\text{where} \quad S_{ij}^t = E_{u_{ij} \sim \Pr(u_{ij}|\eta_{ij}^t)}\{u_{ij}|\pi^i\}. \tag{4.8}$$

We use a Monte Carlo approximation similar to that used in Chapter 2, which involves sampling U from the distribution $\Pr(U \mid D, \Theta^t)$ using a Gibbs sampler, and then approximate S_{ij}^{t+1} by $\frac{1}{N}\sum_{k=1}^{N} u_{ij}^k$. Each step of the Gibbs sampler is sampling from a truncated exponential distribution, illustrated in Figure 2 in Chapter 2.

4.4.2 General M-Step

After we compute S_{ij}^{t+1}'s, the M-step aims at improving $Q(\Theta, \Theta^t)$:

$$Q(\Theta, \Theta^t) = \sum_{j=1}^{m}\sum_{i=1}^{n} \log \Pr_j(u_{ij} = S_{ij}^{t+1} \mid \Theta) + \log(\Pr(\Theta))$$

We use steps of Newton's method to improve $Q(\Theta, \Theta^t)$ in the M-step (we can use as many steps at each iteration to ensure the convergence for each M-step).

$$\Theta^{t+1} = \Theta^t - (\nabla_\Theta^2 Q(\Theta, \Theta^t)|_{\Theta^t})^{-1} \nabla_\Theta Q(\Theta, \Theta^t)|_{\Theta^t} \tag{4.9}$$

$\nabla^2_\Theta Q(\Theta, \Theta^t)$ and $\nabla_\Theta Q(\Theta, \Theta^t)$ can be computed immediately from S^t_{ij} as follows.

$$\nabla^2_\Theta Q(\Theta, \Theta^t) = \sum_{i=1}^{n}\sum_{j=1}^{m} \nabla^2_\Theta \eta_{ij} S^t_{ij} - \nabla^2_\Theta A(\eta_{ij})$$
$$\nabla_\Theta Q(\Theta, \Theta^t) = \sum_{i=1}^{n}\sum_{j=1}^{m} \nabla_\Theta \eta_{ij} S^t_{ij} - \nabla_\Theta A(\eta_{ij})$$

4.4.3 Computing Observed Fisher information

Computation of the observed Fisher information will not only be used in Step 2 of the new elicitation scheme Algorithm 5, but also will accelerate the GEM algorithm [73]. Fisher information can be computed by the following method proposed by Louis [73]. From the independence of agents we have: $J_D(\hat{\Theta}) = \sum_i J_{\pi^i}(\hat{\Theta})$, where,

$$J_{\pi^i}(\Theta) = E_{U_i}\{-\nabla^2_\Theta \log P(\pi^i, U_i|\Theta)|\Theta, \pi^i\}$$
$$- E_{U_i}\{\nabla_\Theta \log P(\pi^i, U_i|\Theta) \nabla_\Theta \log P(\pi^i, U_i|\Theta)^T|\Theta, \pi^i\}$$

$J_{\pi^i}(\hat{\Theta})$ is computed using the samples (u_{ij}'s) generated in MC step in every iteration of EM algorithm as follows.

$$\nabla^2_\Theta \log P(\pi^i, U_i|\Theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \nabla^2_\Theta \eta_{ij} U_{ij} - \nabla^2_\Theta A(\eta_{ij})$$
$$\nabla_\Theta \log P(\pi^i, U_i|\Theta) = \sum_{i=1}^{n}\sum_{j=1}^{m} \nabla_\Theta \eta_{ij} U_{ij} - \nabla_\Theta A(\eta_{ij})$$

The Fisher information matrix $I_h(\hat{\Theta})$ used in Step 3 of Algorithm 5 can be approximated by $\lim_{n\to\infty} \frac{J_{D_n}(\hat{\Theta})}{n}$, where D_n is the data-set of n rankings randomly generated according to $\Pr(\pi|\hat{\Theta})$. Therefore, we can use the techniques developed in this subsection to approximately compute $I_h(\hat{\Theta})$.

4.4.4 MC-EM Algorithm in Detail

The details of the proposed EM algorithm (with fixed number of iterations) are illustrated in Algorithm 7.

Algorithm 7 MAP for GRUM

Input: $D = (\pi^1, \ldots, \pi^n)$, Θ^{start}, $T \in \mathbb{N}$
Let $\Theta^0 = \Theta^{start}$
for $t = 1$ to T **do**
  for every $\pi^i \in D$ **do**
    Compute S_{ij}^{t+1} and $J(\Theta^{t+1})$ according to (4.8) for all $j \leq m$.
  end for
  Compute Θ^{t+1} according to (4.9).
end for

4.5 Global Optimality for Posterior Distribution

In this section, we generalize theorems on the global optimality of the likelihood function for RUMs in Chapter 2 to GRUMs. The EM algorithm tends to find local optimal of the posterior distribution, hence, proving global optimality of MAP helps to avoid issues due to EM. First, we present the concavity of the posterior distribution in GRUMs.

Theorem 17 *For the location family, if for every $j \leq m$ the joint probability density function for $\vec{\epsilon_i}$ and the prior $\Pr(\Theta)$ are log-concave, then $\Pr(\Theta|D)$ is concave up to a known transformation.*

For P-L, Ford, Jr. [47] proposed the following necessary and sufficient condition for the set of global maxima solutions to be bounded (more precisely, unique) when $\sum_{j=1}^{m} e^{\Theta_j} = 1$. The conditions are generalized to the case of RUMs in Chapter 2. We prove that this condition is also necessary and sufficient for global maxima solutions of the likelihood function of GRUMS to be bounded.

Condition 3 *Given the data D, in every partition of the alternatives $\mathcal{C}$ into two nonempty subsets $\mathcal{C}_1 \cup \mathcal{C}_2$, there exists $c_1 \in \mathcal{C}_1$ and $c_2 \in \mathcal{C}_2$ such that there is at least one ranking in D where $c_1 \succ c_2$.*

Theorem 18 *Suppose we fix $\mu_{11} = 0$. Then, the set S_D of global maxima solutions to $\Pr(D|\Theta)$ is bounded in Θ if and only if the data D satisfies Condition 3 and the linear model describing μ in terms of Θ is identifiable.*

4.6 Experimental Results

In this section, we report experimental results on synthetic data and a Sushi dataset from Kamishima [69] for three types of tests described below.

4.6.1 Social Choice and Synthetic Data

We first show the consistency of the model for social choice. We generate random data sets with $\delta_j \sim \text{Normal}(1,1), B_{ij} \sim \text{Normal}(0,1), X_i \sim \text{Normal}(0,1), Z_i \sim \text{Normal}(0,1)$, and then generate random utilities with the random noise ϵ_{ij} generated with mean zero and variance of 1. The results in Figure 4.2 are generated by varying the number of agents for which we have preference information. For each number of agents, we estimate the parameter set Θ, and evaluate the Kendall correlation between estimated and true ranks with respect to δ_j's. These results illustrate the improvement in estimated social choice order as the number of agents in the population increases.

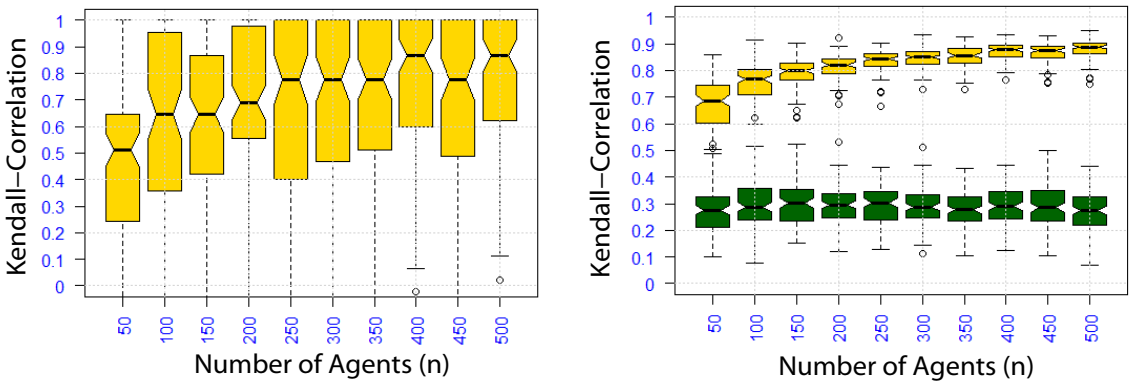
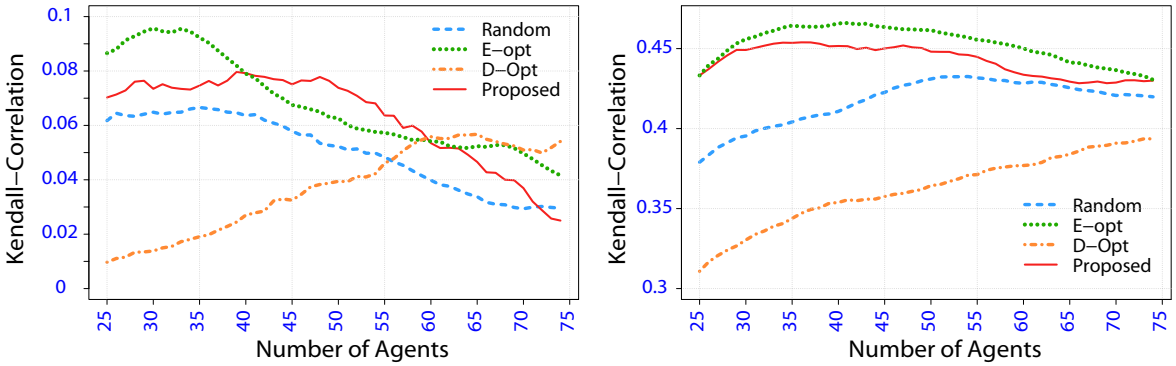


Figure 4.2: Asymptotic behavior for synthetic data and social choice in left panel. Asymptotic behavior for synthetic data and personalized choice in right panel. The y-axis is the average Kendal correlation between the estimated social choice and the ground truth order.

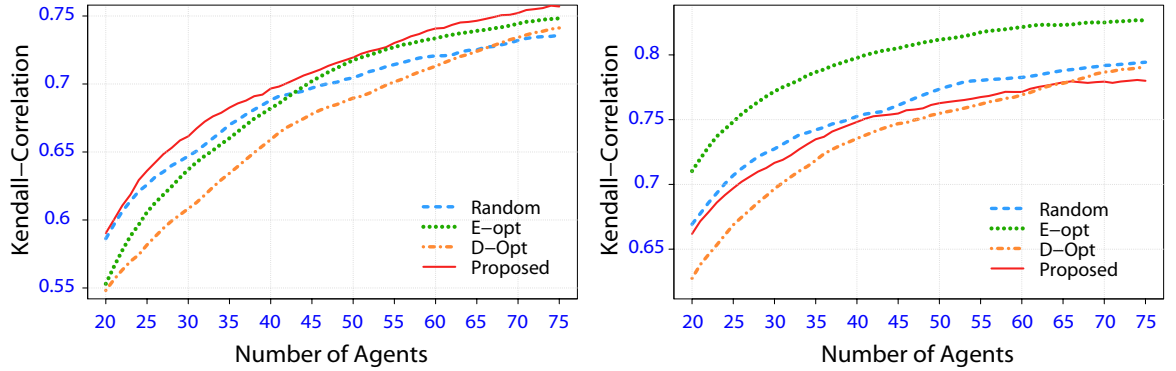
In studying elicitation for social choice, we test the performance of the elicitation schemes shown in Table 4.1, i.e. D-optimality, E-optimality, and the proposed criterion in (4.5), and compare the results to random elicitation. We adopt the following two synthetic datasets:



(a) Social choice: Dataset 1.

(b) Social choice: Dataset 2.

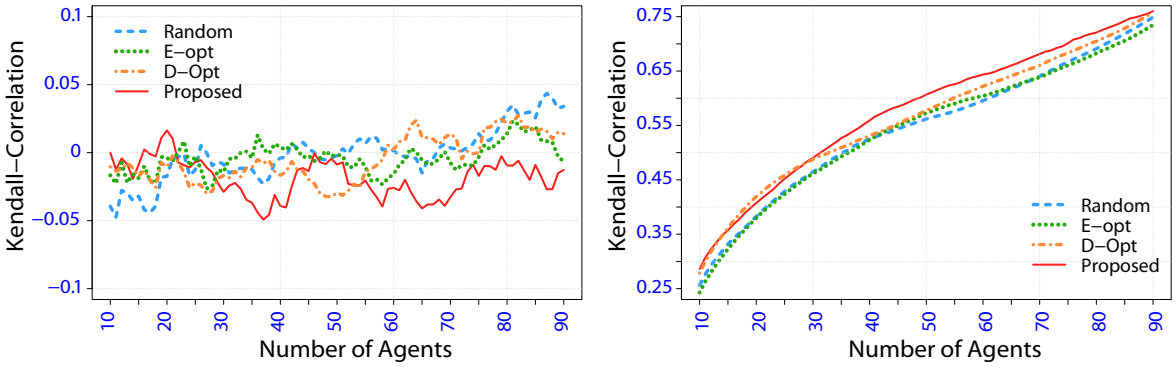
Figure 4.3: Comparison of elicitation criteria described in Table 4.1 for synthetic data and social choice.



(a) Personalized choice: Dataset 1.

(b) Personalized choice: Dataset 2.

Figure 4.4: Comparison of elicitation criteria described in Table 4.1 for synthetic data for personalized choice.



(a) Social choice: Sushi dataset.

(b) Personalized choice: Sushi dataset.

Figure 4.5: Comparison of elicitation criteria described in Table 4.1 for the Sushi dataset [69].

Data-set 1: $(B_{ij} \sim N(0,1), X_i \sim N(0,1), Z_i \sim N(0,1)), \delta_j \sim 0.1 * N(1,1)$ and the error term $\epsilon_{ij} \sim N(0,1)$.

Data-set 2: The same as Data-set 1, except that the $\delta_j \sim N(1,1)$ and the error term $\epsilon_{ij} \sim N(0,1/4)$.

Compared to the GRUM in data-set 1, the model adopted in data-set 2 has a stronger social component and less noise. For each data-set we generate 100 agents' preferences, and use the three criteria shown in Table 4.1 to elicit $n \in [1,100]$ rankings. For each n, we apply Algorithm 7 and compare the ranking over the learned δ_j's with the ground truth social choice ranking.

The results are shown in Figure 4.3 (graphs are smoothed with a moving window with length 25), where the x-axis is the number of agents whose preferences are elicited, and the y-axis is the Kendall correlation between the learned ranking and the ground truth ranking. We make the following observations.

• In Dataset 1 where the social component is small, it is not clear which criteria is better, as shown in Figure 4.3(a), and there are no statistically significant results.

• In data-set 2 where the social component is large, E-optimality generally works better than the proposed method, while both work better than random, which works surprisingly better than D-optimality, as shown in Figure 4.3(b). However, only a few of these observations are statistically significant with 90% confidence, for example, considering the interval of $[34,44]$ agents, E-optimality and the proposed method outperforms Random but the comparison between the other methods is not significant at 90%.

4.6.2 Personalized Choice and Synthetic Data

For personalized choice we first show the consistency results in Figure 4.2, where the bottom box-plot shows the Kendall correlation between noisy data (i.e., an individual agent's random utility and thus preference order) and the true preference order for each agent, and the top box-plot shows Kendall correlation between estimated agent preference orders and true preference orders, as obtained through the model.

Turning to preference elicitation, we compare the schemes in Table 4.1 with the random method for the same two datasets as were adopted for social choice. The results are shown in Figure 4.4 (graphs are smoothed with a moving window with length 20). For each group of 100 agents, and for any $n \in [1, 100]$ and each elicitation scheme, we compute the MAP of Θ, and use it to compute the Kendall correlation between the true preferences and the predicted preference for all 100 agents in this group. We make the following observations:

• In data-set 1, where the social component is small, when the number of agents used in elicitation is not too large (< 50), the proposed method works better than E-optimality, which is itself comparable to random. Both methods are better than D-optimality. See Figure 4.4(a). Some of these observations are statistically significant, for example, when $n = [24, 25]$, E-optimality works better than D-optimality with 90% significance, E-optimality works better than random with 75% significance, the proposed method works better than E-optimality with 75% significance, and the proposed method works better than D-optimality with 75% significance.

• In data-set 2, where the social component is large, E-optimality generally works better than the proposed method, both work better than random, and random is more effective than D-optimality, as shown in Figure 4.4(b). However, only a few of these observations are statistically significant with 90% confidence interval, for example E-optimality outperforms D-optimality when the number of agents is in the interval $[29, 42]$.

4.6.3 Sushi Data

In synthetic experiments, we have access to the ground truth. However, in the real world data (Sushi data) there are no data available as ground truth. In this experiment, we estimated parameters Θ using preferences from 1000 agents, randomly chosen from the 5000 agents in the data-set. And adopt those parameters as the ground truth for the experimental study. The categorical features are discarded from the data set.[6]

The results are shown in Figure 4.5 (graphs are smoothed with a moving window with

[6]We focus on non categorical features in this work. The method can be extended to categorical features.

length 10), where (a) shows comparisons for social choice (where we rank δ's), and (b) shows comparisons for personalized choice. We make the following observations:

• For social choice (a), none of the criteria work well (and note that the Kendall correlations are low for all criteria). We feel that this is reasonable since preferences over sushi is likely high personalized with a small social component to preferences.

• For personalized choice (b), we observe that the proposed method is generally the most effective, while the performance of E-optimality and D-optimality is very close to random. None of these results are statistically significant with 90% confidence.

4.7 Conclusions

We have proposed a method for preference elicitation of ordinal rank data, adopting the framework of Bayesian experimental design. This includes two new criteria, each optimal for social and personalized case respectively. The proposed criterion for social choice can significantly improve the precision of estimation, relative to random elicitation and some of the classical elicitation criteria. This work can also be seen as preference elicitation for learning to rank, since we focus on a learning to rank setting and design elicitation methods. In the future, we can adopt the methodology in other preference elicitation applications; for example recommendation systems, product prediction and so forth. Moreover, it is an interesting direction to use a similar technique to decide what alternatives to choose to elicit partial ranks on them is an interesting direction.

Chapter 5

Random Utility for Personalized Rank Data With Multiple Types

5.1 Introduction

Random utility models (RUM), which presume agent utility to be composed of a deterministic component and a stochastic unobserved (by the analyst) error component, are frequently used to model choices by individuals over alternatives. Examples from economics include the popular random coefficients logit model [17] where the data may involve a (partial) consumer ranking of products [19].

In a RUM, each agent receives an intrinsic utility that is common across all agents for a given choice of alternative, a pairwise-specific utility that varies with the interaction between agent characteristics and the characteristics of the agent's chosen alternative, as well as an agent-specific taste shock (noise) for his chosen alternative. These ingredients are used to construct a posterior/likelihood function of specific data moments, such as the fraction of agents of each type that choose each alternative.

To estimate preferences across heterogenous agents, one approach described in prior work [61, 68] is to assume a mixture models with a finite number of preference types. We build upon this work by developing an algorithm to learn the classification of agent types within

this mixture. Empirical researchers are increasingly being presented with rich data on the choices made by individuals, and asked to classify these agents into different types [90, 91] and to estimate the preferences of each type [20, 66]. Examples of individual-level data used in economics include household purchases from supermarket-scanner data [1, 63], and patients' hospital or treatment choices from health-care data [64].

The non probabilistic partitioning of agents into latent, discrete sets (or "types") can allow for the study of the underlying distribution of preferences across a population of heterogeneous agents. For example, preferences may be correlated with an agent characteristic, such as income, and the true classification of each agent's type, such as his income bracket, may be unobserved. In future work we can use a model of demand to estimate the elasticity in behavioral response of each type of agent and by aggregating these responses over the different types of agents, it can be possible to simulate the impact of a social or public policy [18], or simulate the counterfactual outcome of changing the options available to agents [60].

5.1.1 Our Contributions

This chapter focuses on estimating generalized random utility models (GRUM) when the observed data is partial orders of agents' rankings over alternatives and when latent types are present.

We build on Chapters 1 and 3 results on estimating GRUMs by allowing for an interaction between agent characteristics and the characteristics of the agent's chosen alternative. The interaction term helps us to avoid unrealistic substitution patterns due to the independence of irrelevant alternatives [83] by allowing agent utilities to be correlated across alternatives with similar characteristics. For example, this prevents a situation where removing the top choices of both a rich household and a poor household lead them to become equally likely to substitute to the same alternative choice. Our model also allows the marginal utilities associated with the characteristics of alternatives to vary across agent types.

To classify agents' types and estimate the parameters associated with each type, we propose an algorithm involving a novel application of reversible jump Markov Chain Monte Carlo

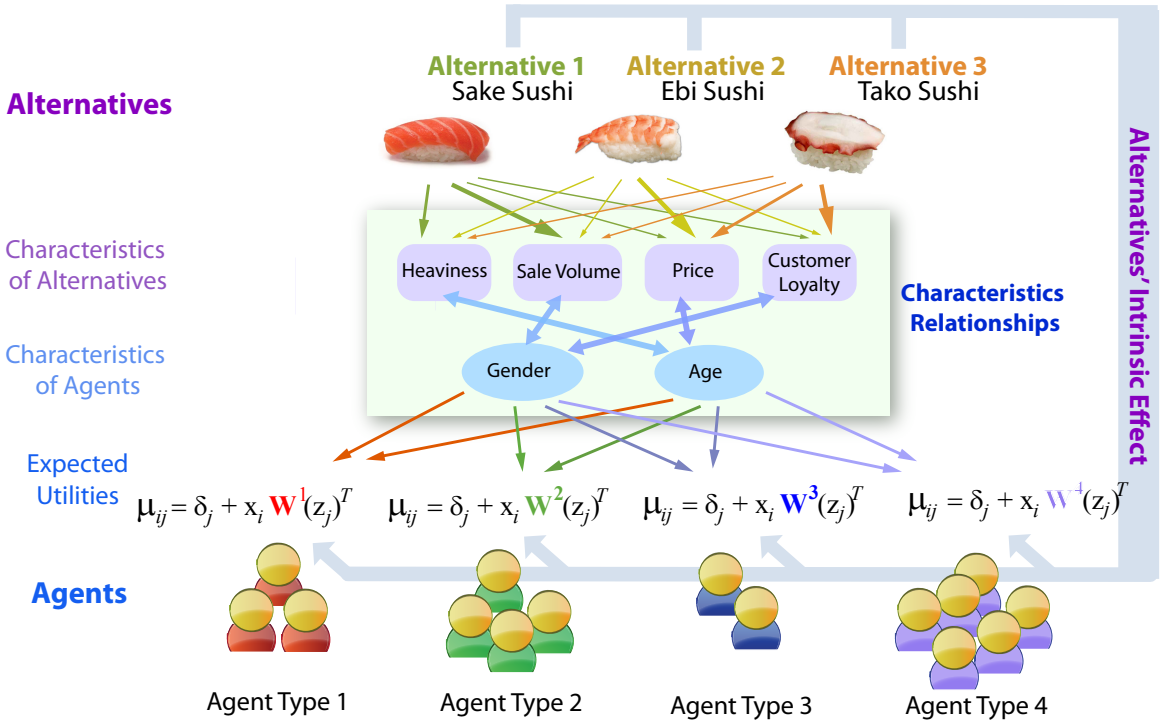


Figure 5.1: A GRUM with multiple types of agents

(RJMCMC) techniques. RJMCMC can be used for model selection and learning a posterior on the number of types in a mixture model [109]. Here, we use RJMCMC to cluster agents into different types, where each type exhibits demand for alternatives based on different preferences; i.e., different interaction terms between agent and alternative characteristics.

Allowing individuals to have characteristics and existence of types leads to the opportunity to understand how characteristics correlate with characteristics of the rank order distribution.

We apply the approach to a real-world data-set involving consumers' preference rankings and also conduct experiments on synthetic data to perform coverage analysis of RJMCMC. The results show that our method is scalable, and that the clustering of types provides a better fit to real world data. The proposed learning algorithm is based on Bayesian methods to find posteriors on the parameters. This differentiates us from previous estimation approaches in econometrics that rely on techniques based on the generalized method of moments.[1]

[1]There are alternative methods to RJMCMC, such as the saturation method [31]. However, the memory required to keep track of former sampled memberships in the saturation method quickly becomes infeasible

The main theoretical contribution establishes identifiability of mixture models over data consisting of partial orders. Previous theoretical results have established identifiability for data consisting of vectors of real numbers [4, 56], but not for data consisting of partial orders. We also establish conditions under which the GRUM likelihood function is uni-modal for the case of observable types. We do not provide results on the log concavity of the general likelihood problem with latent type, leaving this for future study.[2]

5.1.2 Related work

Prior work in econometrics has focused on developing models that use data aggregated across types of agents, such as geographical location of a market, and that allow heterogeneity by using random coefficients on either agents' preference parameters [17, 19] or on a set of dummy variables that define types of agents [16, 84], or by imposing additional structure on the covariance matrix of idiosyncratic taste shocks [50]. In practice, this approach typically relies on restrictive functional assumptions about the distribution of consumer taste shocks that enter the RUM in order to reduce computational burden. For example, the logit model [83] assumes i.i.d. draws from a Type I extreme value distribution. This may lead to biased estimates, in particular when the number of alternatives grow large [14].

Previous work on clustering ranking data for variations of the Placket-Luce (PL) model [90, 91] has been restricted to settings without agent and alternative characteristics. Moreover, Gormley et al. [90] and Chu et al. [36] performed clustering for RUMs with normal distributions, but this was limited to pairwise comparison data. Inference of GRUMs for partial ranks involved similar computational challenges addressed in Chapter 1. Moreover, in mixture models, assuming an arbitrary number of types can lead to biased results, and reduces the statistical efficiency of the estimators [48].

The multiple type model in the second chapter does not consider observed user character-

---

given the combinatorial nature of our problem.

[2] In the chapter the notation z is used for observed characteristics as opposed to latent characteristics in Chapter 2.

istics and tries to capture the heterogeneity using latent user characteristics.

To the best of our knowledge, we are the first to study the identifiability and inference of GRUMs with multiple types. Inference for GRUMs has been generalized in Chapter 3, However, the methods in Chapter 3 do not consider existence of multiple types. The proposed method here also applies to partial orders. The inference method establishes a posterior on the number of types, resolving the common issue of how the researcher should select the number of types. Moreover, the use of RJMCMC in order to compute a posterior on the number of parameters is a novel approach which allows us to have a full posterior on the model.

5.2 Model

Suppose we have N agents and M alternatives $\{c_1, .., c_M\}$, and there are S types (subgroups) of agents and $s(n)$ is agent n's type. The types are latent in this model.

Agent characteristics are observed and defined as an $N \times K$ matrix X, and alternative characteristics are observed and defined as an $L \times M$ matrix Z, where K and L are the number of agent and alternative characteristics respectively.

Let u_{nm} be agent n's *perceived utility* for alternative m, and let $W^{s(n)}$ be a $K \times L$ real matrix that models the linear relation between the attributes of alternatives and the attributes of agents. We have,

$$u_{nm} = \delta_m + \vec{x}_n W^{s(n)} (\vec{z}_m)^T + \epsilon_{nm}, \tag{5.1}$$

where $\vec{x}_n$ is the nth row of the matrix X and $\vec{z}_m$ is the mth column of the matrix Z. In words, agent n's utility for alternative m consists of the following three parts:

1. δ_m: The *intrinsic utility* of alternative m, which is the same across all agents;

2. $\vec{x}_n W^{s(n)} (\vec{z}_m)^T$: The *agent-specific utility*, which is unique to all agents of type $s(n)$, and where $W^{s(n)}$ has at least one nonzero element;

3. ϵ_{nm}: The *random noise* (agent-specific taste shock), which is generated independently across agents and alternatives.

The number of parameters for each type is $P = KL + M$.

See Figure 5.2 for an illustration of the model. In order to write the model as a linear regression, we define matrix $A^{(n)}_{M\times P}$, such that $A^{(n)}_{KL+m,m} = 1$ for $1 \le m \le M$ and $A^{(n)}_{KL+m,m'} = 0$ for $m \neq m'$ and $A^{(n)}_{(k-1)L+l,m} = \vec{x}_n(k)\vec{z}_m(l)$ for $1 \le l \le L$ and $1 \le k \le K$. We also need to shuffle the parameters for all types into a $P \times S$ matrix Ψ, such that $\Psi_{KL+m,s} = \delta$ and $\Psi_{(k-1)L+l,s} = W^s_{kl}$ for $1 \le k \le K$ and $1 \le l \le L$. We adopt $B^{(n)}_{S\times 1}$ to indicate the type of agent n, with $B^{(n)}_{s(n),1} = 1$ and $B^{(n)}_{s,1} = 0$ for all $s \neq s(n)$. We also define an $M \times 1$ matrix, $U^{(n)}$, as $U^{(n)}_{m,1} = u_{nm}$. We can now rewrite (5.1) as:

$$U^{(n)} = A^{(n)}\Psi B^{(n)} + \epsilon \tag{5.2}$$

Suppose that an agent has type s with probability γ_s. Given this, the random utility model can be written as, $\Pr(U^{(n)}|X^{(n)}, Z, \Psi, \Gamma) = \sum_{s=1}^{S} \gamma_s \Pr(U^{(n)}|X^{(n)}, Z, \Psi^s)$, where Ψ^s is the sth column of the matrix Ψ. An agent ranks the alternatives according to her perceived utilities for the alternatives. Define rank order π^n as a permutation $(\pi^n(1), \ldots, \pi^n(m))$ of $\{1, \ldots, M\}$. π^n represents the full ranking $[c_{\pi^i(1)} \succ_i c_{\pi^i(2)} \succ_i \cdots \succ_i c_{\pi^i(m)}]$ of the alternatives $\{c_1, .., c_M\}$. That is, for agent n, $c_{m_1} \succ_n c_{m_2}$ if and only if $u_{nm_1} > u_{nm_2}$ (In this model, situations with tied perceived utilities have zero probability measure).

The model for observed data $\pi^{(n)}$, can be written as:

$$\Pr(\pi^{(n)}|X^{(n)}, Z, \Gamma, \Psi) = \int_{\pi^{(n)}=order(U^{(n)})} \Pr(U^{(n)}|X^{(n)}, Z, \Psi, \Gamma) = \sum_{s=1}^{S} \gamma_s \Pr(\pi^{(n)}|X^{(n)}, Z, \Psi^s)$$

Note that $X^{(n)}$ and Z are observed characteristics, while Γ and Ψ are unknown parameters. $\pi = order(U)$ is the ranking implied by U, and $\pi(i)$ is the ith largest utility in U. $D = \{\pi^1, .., \pi^N\}$ denotes the collection of all data for different agents. We have that

$$\Pr(D|X, Z, \Psi, \Gamma) = \prod_{n=1}^{N} \Pr(\pi^{(n)}|X^{(n)}, Z, \Psi, \Gamma)$$

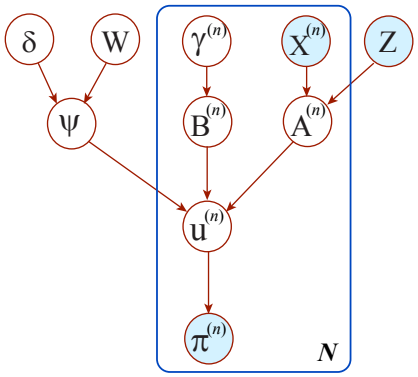


Figure 5.2: Graphical representation of the multiple type GRUM generative process.

5.3 Strict Log-concavity and Identifiability

In this section, we establish conditions for identifiability of the types and parameters for the model. Identifiability is a necessary property in order for researchers to be able to infer economically-relevant parameters from an econometric model. Establishing identifiability in a model with multiple types and ranking data requires a different approach from classical identifiability results for mixture models [4, 56, e.g.]. Moreover, we establish conditions for uni-modality of the likelihood for the parameters Γ and Ψ, when the types are observed. Although our main focus is on data with unobservable types, establishing the conditions for uni-modality conditioned on known types remains an essential step because of the sampling and optimization aspects of RJMCMC. We sample from the parameters conditional on the algorithm's specification of types.

The uni-modality result establishes that the sampling approach is exploring a uni-modal distribution conditional on its specified types. Despite adopting a Bayesian point of view in presenting the model, we adopt a uniform prior on the parameter set, and only impose nontrivial priors on the number of types in order to obtain some regularization. Given this, we present the theory with regards to the likelihood function from the data rather than the posterior on parameters.

5.3.1 Strict Log-concavity of the Likelihood Function

For agent n, we define a set G^n of function g^n's whose positivity is equivalent to giving an order π^n. More precisely, we define $g^n_m(\vec{\psi}, \vec{\epsilon}) = [\mu_{n\pi^n(m)} + \epsilon_{n\pi^n(m)}] - [\mu_{n\pi^n(m+1)} + \epsilon_{n\pi^n(m+1)}]$ for $m = 1, .., M-1$ where $\mu_{nj} = \delta_j + \sum_{k,l} x_n(k) W^{s(n)}_{kl} z_j(l)$ for $1 \leq j \leq M$. Here, $\vec{\psi}$ is a vector of $KL + M$ variables consisting of all δ_j's and W_{kl}'s. We have, $L(\vec{\psi}, \pi^n) = L(\vec{\psi}, G^n) = \Pr(g^n_1(\vec{\psi}, \vec{\epsilon}) \geq 0, ..., g^n_{M-1}(\vec{\psi}, \vec{\epsilon}) \geq 0)$. This is because $g^n_m(\vec{\psi}, \vec{\epsilon}) \geq 0$ is equivalent to saying alternative $\pi^n(m)$ is preferred to alternative $\pi^n(m+1)$ in the RUM sense.

Then using the result in Chapter 1 and [102], $L(\vec{\psi}) = L(\vec{\psi}, \pi)$ is **logarithmic concave** in the sense that $L(\lambda\vec{\psi} + (1-\lambda)\vec{\psi'}) \geq L(\psi)^{\lambda} L(\psi')^{1-\lambda}$ for any $0 < \lambda < 1$ and any two vectors $\vec{\psi}$, $\vec{\psi'} \in \mathbb{R}^{LK+M}$. Let's consider all n agents together. We study the function, $l(\Psi, D) = \sum_{n=1}^{N} \log Pr(\pi^n | \vec{\psi}^{s(n)})$. By log-concavity of $L(\vec{\psi}, \pi)$ and using the fact that sum of concave functions is concave, we know that $l(\Psi, D)$ is concave in Ψ, viewed as a vector in $\mathbb{R}^{SKL+M}$. To show uni-modality, we need to prove that this concave function has a unique maximum. Namely, we need to be able to establish the conditions for when the equality holds. If our data is subject to some mild condition, which implies boundedness of the parameter set that maximizes $l(\Psi, D)$, Theorem 19 bellow tells us when the equality holds. This condition has been explained in Chapter 1 as condition (1).

Before stating the main result, we define the following auxiliary $(M-1)N' \times (SKL+M-1)$ matrix $\widetilde{A} = \widetilde{A}^{N'}$ (Here, let $N' \leq N$ be a positive number that we will specify later.) such that, $\widetilde{A}_{(M-1)(n-1)+m,(s-1)KL+(K-1)l+k}$ is equal to $x_n(k)(z_m(l) - z_M(l))$if $s = s(n)$ and is equal to 0 if $s \neq s(n)$, for all $1 \leq n \leq N'$, $1 \leq m \leq M-1$, $1 \leq s \leq S$, $1 \leq k \leq K$, and $1 \leq l \leq L$. Also, $\widetilde{A}_{(M-1)(n-1)+m,SKL+m'}$ is equal to 1 if $m = m'$ and is equal to 0 if $m \neq m'$, for all $1 \leq m, m' \leq M-1$ and $1 \leq n \leq N'$.

Theorem 19 *Suppose there is an $N' \leq N$ such that rank $\widetilde{A}^{N'} = SKL + M - 1$. Then $l(\Psi) = l(\Psi, D)$ is strictly concave up to δ-shift, in the sense that,*

$$l(\lambda\Psi + (1-\lambda)\Psi') \geq \lambda l(\Psi) + (1-\lambda) l(\Psi'), \tag{5.3}$$

for any $0 < \lambda < 1$ and any $\Psi, \Psi' \in \mathbb{R}^{SKL+M}$, and the equality holds if and only if there exists

$c \in \mathbb{R}$, such that:

$$\begin{cases} \delta_m = \delta'_m + c & \text{for all } 1 \le m \le M \\ W^s_{kl} = W'^s_{kl} & \text{for all } s, k, l \end{cases}$$

Remark 1 *We remark that the strictness "up to δ-shift" is natural. A δ-shift results in a shift in the intrinsic utilities of all the products, which does not change the utility difference between products. So such a shift does not affect our outcome. In practice, we may set one of the δ's to be 0 and then our algorithm will converge to a single maximum.*

Remark 2 *It's easy to see that N' must be larger than or equal to $1 + \frac{SKL}{M-1}$. The reason we introduce N' is to avoid cumbersome calculations involving N.*

5.3.2 Identifiability of the Model

In this section, we show the case of unobserved types our model is identifiable for a certain class of CDFs for the noise in random utility models. Let's first specify this class of "nice" CDFs:

Definition 5 *Let $\phi(x)$ be a smooth pdf defined on $\mathbb{R}$ or $[0, \infty)$, and let $\Phi(x)$ be the associated CDF. For each $i \ge 1$, we write $\phi^{(i)}(x)$ for the i-th derivative of $\phi(x)$. Let $g_i(x) = \frac{\phi^{(i+1)}(x)}{\phi^{(i)}(x)}$. The function Φ is called **nice** if it satisfies one of the following two mutually exclusive conditions:*

(a) $\phi(x)$ is defined on $\mathbb{R}$. For any $x_1, x_2 \in \mathbb{R}$, the sequence $\frac{g_i(x_1)}{g_i(x_2)}$ converges to some value in $\mathbb{R}$ (as $i \to \infty$) only if either $x_1 = x_2$; or $x_1 = -x_2$ and $\frac{g_i(x_1)}{g_i(x_2)} \to -1$ as $i \to \infty$.

(b) $\phi(x)$ is defined on $[0, \infty)$. For any $x_1, x_2 \ge 0$, the ratio $\frac{\phi^{(i)}(x_1)}{\phi^{(i)}(x_2)}$ is independent of i for i sufficiently large. Moreover, we require that $\phi(x_1) = \phi(x_2)$ if and only if $x_1 = x_2$.

This class of nice functions contains Normal distributions and exponential distributions.

Identifiability is formalized as follows: Let $\mathcal{C} = \{\{\gamma_s\}_{s=1}^S \mid S \in \mathbb{Z}_{>0}, \gamma_i \in \mathbb{R}_{>0}, \sum_{s=1}^S \gamma_s = 1\}$. Suppose, for two sequences $\{\gamma_s\}_{s=1}^S$ and $\{\gamma'_s\}_{s=1}^{S'}$, we have:

$$\sum_{s=1}^{S} \gamma_s \Pr(\pi | X^{(n)}, Z, \Psi) = \sum_{s=1}^{S'} \gamma'_s \Pr(\pi | X^{(n)}, Z, \Psi') \tag{5.4}$$

for all possible orders π of M products, and for all agents n. Then, we must have $S = S'$ and (up to a permutation of indices $\{1, \cdots, S\}$) $\gamma_s = \gamma'_s$ and $\Psi = \Psi'$ (up to δ-shift).

For now, let's fix the number of agent characteristics, K. One observation is that the number $x_n(k)$, for any characteristic k, reflects certain characteristics of agent n. Varying the agent n, this amount $x_n(k)$ is in a bounded interval in $\mathbb{R}$. Suppose the collection of data D is sufficiently large. Based on this, assuming that N can be be arbitrarily large, we can assume that the $x_n(k)$'s form a dense subset in a closed interval $I_k \subset \mathbb{R}$. Hence, (5.4) should hold for any $X \in I_k$, leading to the following theorem:

Theorem 20 *Define an $(M-1) \times L$ matrix $\widetilde{Z}$ by setting $\widetilde{Z}_{m,l} = z_m(l) - z_M(l)$. Suppose the matrix $\widetilde{Z}$ has rank L, and suppose,*

$$\sum_{s=1}^{S} \gamma_s \Pr(\pi | X, Z, \Psi) = \sum_{s=1}^{S'} \gamma'_s \Pr(\pi | X, Z, \Psi'), \tag{5.5}$$

for all $x(k) \in I_k$ and all possible orders π of M products. Here, the probability measure is associated with a nice CDF. Then we must have $S = S'$ and (up to a permutation of indices $\{1, \cdots, S\}$), $\gamma_s = \gamma'_s$ and $\Psi = \Psi'$ (up to δ-shift).

Here, we illustrate the idea for the simple case, with two alternatives ($m = 2$) and no agent or alternative characteristics ($K = L = 1$). Equation (5.5) is merely a single identity. Unwrapping the definition, we obtain:

$$\sum_{s=1}^{S} \gamma_s \Pr(\epsilon_1 - \epsilon_2 > \delta_1 - \delta_2 + xW^s(z_1 - z_2)) = \sum_{s=1}^{S'} \gamma'_s \Pr(\epsilon_1 - \epsilon_2 > \delta'_1 - \delta'_2 + xW'^s(z_1 - z_2)). \tag{5.6}$$

Without loss of generality, we may assume $z_1 = 1$, $z_2 = 0$, and $\delta_2 = 0$. We may further assume that the interval $I = I_1$ contains 0. (Otherwise, we just need to shift I and δ accordingly.) Given this, the problem reduces to the following lemma:

Lemma 7 *Let $\Phi(x)$ be a nice CDF. Suppose,*

$$\sum_{s=1}^{S} \gamma_s \Phi(\delta + xW^s) = \sum_{s=1}^{S'} \gamma'_s \Phi(\delta' + xW'^s), \tag{5.7}$$

for all x in a closed interval I containing 0. Then we must have $S = S'$, $\delta = \delta'$ and (up to a permutation of $\{1, \cdots, S\}$) $\gamma_s = \gamma_s$, $W^s = W'^s$.

By applying this to (5.6), we can show identifiablity for the simple case of $m = 2$ and $K = L = 1$.

Theorem 20 guarantees identifiability in the limit case that we observe agents with characteristics that are dense in an interval. Beyond the theoretical guarantee, we would in practice expect (5.6) to have a unique solution with a enough agents with different characteristics. Lemma 7 is a new identifiability result for scalar observations from a set of truncated distributions.

5.4 RJMCMC for Parameter Estimation

We are using a uniform prior for the parameter space and regularize the number of types with a geometric prior. We use a Gibbs sampler to sample from the posterior. In each of T iterations, we sample utilities u^n for each agent, matrix ψ_s for each type, and set of assignments of agents to alternatives $\mathbf{S}^n$. The utility of each agent for each alternative conditioned on the data and other parameters is sampled from a truncated exponential family (e.g. Normal) distribution. In order to sample agent i's utility for alternative j (u_{ij}), we set thresholds for lower and upper truncation based on agent i's former samples of utility for the two alternatives that are ranked one below and one above alternative j, respectively.

We use reversible-jump MCMC [55] for sampling from conditional distributions of the assignment function (see Algorithm 8). We consider three possible moves for sampling from the assignment function $\mathbf{S}(n)$:

(1) Increase the number of types by one, through moving a random agent to a new type of its own. The acceptance ratio for this move is:

$$\Pr_{split} = \min\{1, \frac{\Pr(S+1)\Pr(\mathcal{M}^{(t+1)}|D)}{\Pr(S)\Pr(\mathcal{M}^{(t)}|D)} . \frac{\frac{1}{S+1}}{\frac{1}{S}} . \frac{p_{+1}}{p_{-1}} . \frac{1}{p(\alpha)} . \mathcal{J}_{(t)\to(t+1)}\},$$

where $\mathcal{M}^{(t)} = \{u, \psi, B, \mathbf{S}, \pi\}^{(t)}$, and $\mathcal{J}_{(t)\to(t+1)} = 2^P$ is the Jacobian of the transformation

from the previous state to the proposed state and $\Pr(S)$ is the prior (regularizer) for the number of types.

(2) Decrease the number of types by one, through merging two random types. The acceptance ratio for the merge move is:

$$\Pr_{merge} = \min\{1, \frac{\Pr(S-1)\Pr(\mathcal{M}^{(t+1)}|D)}{\Pr(S)\Pr(\mathcal{M}^{(t)}|D)} . \frac{\frac{1}{S-1}}{\frac{1}{S}} . \frac{p_{-1}}{p_{+1}} . \mathcal{J}_{(t)\to(t+1)}\}$$

(3) Leave the number of types unchanged and consider moving one random agent from one type to another. This case reduces to a standard Metropolis-Hastings, where because of the normal symmetric proposal distribution, the proposal is accepted with probability:

$$\Pr_{mh} = \min\{1, \frac{\Pr(\mathcal{M}^{(t+1)}|D)}{\Pr(\mathcal{M}^{(t)}|D)}\}$$

Algorithm 8 RJMCMC to update $\mathbf{S}^{(t+1)}(n)$ from $\mathbf{S}^{(t)}(n)$

Set p_{-1}, p_0, p_{+1}, Find S: number of distinct types in $\mathbf{S}^{(t)}(n)$

Propose move ν from $\{-1, 0, +1\}$ with probabilities p_{-1}, p_0, p_{+1}, respectively.

case $\nu = +1$**:**

Select random type M_s and agent $n \in M_s$ uniformly and Assign n to module M_{s_1} and remainder to M_{s_2} and Draw vector $\alpha \sim \mathcal{N}(0, 1)$ and Propose $\psi_{s_1} = \psi_s - \alpha$ and $\psi_{s_2} = \psi_s + \alpha$ and Compute proposal $\{u^n, \pi^n\}^{(t+1)}$

Accept $\mathbf{S}^{(t+1)}(M_{s_1}) = S + 1$, $\mathbf{S}^{(t+1)}(M_{s_2}) = s$ with $\Pr_{split}$ from update $S = S + 1$

case $\nu = -1$**:**

Select two random types M_{s_1} and M_{s_2} and Merge into one type M_s and Propose $\psi_s = (\psi_{s_1} + \psi_{s_1})/2$ and Compute proposed $\{u^n, \pi^n\}^{(i+1)}$

Accept $\mathbf{S}^{(t+1)}(n) = s_1$ for $\forall n \; s.t. \; \mathbf{S}^{(t)}(n) = s_2$ with $\Pr_{merge}$ update $S = S - 1$

case $\nu = 0$**:**

Select two random types M_{s_1} and M_{s_2} and Move a random agent n from M_{s_1} to M_{s_2} and Compute proposed $\{u^{(n)}, \pi^{(n)}\}^{(t+1)}$

Accept $\mathbf{S}^{(t+1)}(n) = s_2$ with probability $\Pr_{mh}$

end switch

5.5 Experimental Study

We evaluate the performance of the algorithm on both synthetic data and a real world data set in which we observe agents' characteristics and their orderings on alternatives. For the synthetic data, we generate data with different numbers of types and perform RJMCMC in order to estimate the parameters and number of types. The algorithm is implemented in MATLAB and scales linearly in the number of samples and agents. It takes on average 60 ± 5 seconds to generate 50 samples for $N = 200$, $M = 10$, $K = 4$ and $L = 3$ on an i5 2.70GHz Intel(R).

Coverage Analysis for the number of types S for Synthetic Data: In this experiment, the data is generated from a randomly chosen number of clusters S for $N = 200$, $K = 3$, $L = 3$ and $M = 10$ and the posterior on S is estimated using RJMCMC. The prior is chosen to be $\Pr(S) \propto \exp(-3SKL)$. We consider a noisy regime by generating data from noise level of $\sigma = 1$, where all the characteristics (X,Z) are generated from $\mathcal{N}(0, 1)$. We repeat the experiment 100 times. Given this, we estimate 60%, 90% and 95% confidence intervals for the number of types from the posterior samples. We also estimate the *coverage* percentage, which is defined to be the percentage of samples which include the true number of types in the interval. The simulations show 61%, 73%, 88%, 93% for the intervals 60%, 75%, 90%, 95% respectively, which indicates that the method is providing reliable intervals for the number of types.

Performance for Synthetic Data: We generate data randomly from a model with between 1 and 4 types. N is set to 200, and M is set to 10 for $K = 4$ and $L = 3$. We draw $10,000$ samples from the stationary posterior distribution. The prior for S has chosen to be $\exp(-\alpha SKL)$ where α is uniformly chosen in $(0, 10)$. We repeat the experiment 5 times. Table 5.1 shows that the algorithm successfully provides larger log posterior when the number of types is the number of true types.

Clustering Performance for Real World Data: We have tested our algorithm on a sushi data-set, where $5,000$ users provide rankings on $M = 10$ different kinds of sushi [69]. We fit the multi-type GRUM for different number of types, on 100 randomly chosen subsets of the sushi data with size $N = 200$, and using the same prior we used in synthetic case. We provide the performance on the Sushi data in Table 5.1. It can be seen that GRUM with 3 types has significantly better performance in terms of log posterior (with the prior that we choose, log posterior can be seen as log likelihood penalized for number of parameters) than GRUM with one, two or four types. We have taken non-categorical agent features age, time for filling the questionnaire, region ID and prefecture ID) and sushi features as price,heaviness and sales volume.

5.6 Extended proofs

5.6.1 On Strict Logarithmic Concavity

The main purpose of this section is to establish a "strict" version of the logarithmic concavity results in Prékopa [102]. As an application, we shall prove Theorem 19.

Let us first prove following Lemma.

Lemma 8 *Suppose $\vec{\epsilon}$ is a vector of M real numbers that are generated according to a distribution whose pdf is strictly logarithmic concave in $\mathbb{R}^M$. Consider the function*

$$L(\vec{\psi}, \pi) = L(\vec{\psi}, G) = \Pr(g_1(\vec{\psi}, \vec{\epsilon}) \geq 0, ..., g_{M-1}(\vec{\psi}, \vec{\epsilon}) \geq 0) \tag{5.8}$$

*Then using the result in [11] and [102], $L(\vec{\psi}) = L(\vec{\psi}, \pi)$ is **logarithmic concave** in the sense that $L(\lambda\vec{\psi} + (1-\lambda)\vec{\psi'}) \geq L(\psi)^{\lambda} L(\psi')^{1-\lambda}$ for any $0 < \lambda < 1$ and any two vectors $\vec{\psi}$, $\vec{\psi'} \in \mathbb{R}^{LK+M}$.*

This is a direct consequence of Theorem 9 in [102]. Since its proof inspires our work on strict log-concavity, it is worth presenting here.

Proof: Similar to approach in [102], we consider sets $H(\vec{\psi}) = \{\vec{\epsilon}\,|\, g_m(\psi, \vec{\epsilon}) \geq 0,\ m = 1, \cdots, M-1\}$. Then $L(\vec{\psi}) = Pr(\vec{\epsilon} \in H(\vec{\psi}))$. We also have $H(\lambda\vec{\psi} + (1-\lambda)\vec{\psi'}) = \lambda H(\vec{\psi}) + (1-\lambda)H(\vec{\psi'})$ because our g_m's are linear functions. By Theorem 2 in [102], the probability measure Pr is strictly log-concave. So we have

$$\begin{aligned} L(\lambda\vec{\psi} + (1-\lambda)\vec{\psi'}) &= Pr(\vec{\epsilon} \in H(\lambda\vec{\psi} + (1-\lambda)\vec{\psi'})) \\ &= Pr(\vec{\epsilon} \in \lambda H(\vec{\psi}) + (1-\lambda)H(\vec{\psi'})) \\ &\geq (Pr(\vec{\epsilon} \in H(\vec{\psi})))^{\lambda}(Pr(\vec{\epsilon} \in H(\vec{\psi'})))^{1-\lambda} \\ &= L(\vec{\psi})^{\lambda} L(\vec{\psi'})^{1-\lambda} \end{aligned} \tag{5.9}$$

as desired. □

However, in practice, it is important to know when the equality in 5.9 holds. To answer this question, we need a "strict" version of log-concavity theory.

Strictly Logarithmic Concave Measure

Mimicing the major ideas from [102], we define strictly log-concave measures and strictly log-concave functions. Roughly speaking, they are the same as log-concave measures and log-concave functions, but subject to a uniqueness condition on when the equality holds.

Definition 6 *A measure P defined on the Borel measurable subsets of $\mathbb{R}^m$ is said to be* ***strictly logarithmic concave*** *if*

$$\Pr(\lambda A + (1-\lambda)B) \geq \Pr(A)^{\lambda} \Pr(B)^{1-\lambda}$$

for every $0 < \lambda < 1$ and for all convex subsets $A, B \subset \mathbb{R}^m$, and the equality holds if and only if $\mu(A \triangle B) = 0$. (Here μ stands for Lebesgue measure and $\triangle$ is the symmetric difference.)

Definition 7 *A positive continuous function $h(x)$ on $\mathbb{R}^m$ (resp., on a convex subset X of $\mathbb{R}^m$) is said to be* ***strictly logarithmic concave*** *if for every pair $x_1, x_2 \in \mathbb{R}^m$ (resp., $x_1, x_2 \in X$) and every $0 < \lambda < 1$, we have*

$$h(\lambda x_1 + (1-\lambda)x_2) \geq h(x_1)^{\lambda} h(x_2)^{1-\lambda},$$

and the equality holds if and only if $x_1 = x_2$.

The following technical lemma is needed later.

Lemma 9 *(a) Let* h *be a logarithmic concave function on* $\mathbb{R}^m$. *Suppose four points* x_1, x_2, y_1, y_2 *lie on the same line, with* x_1, y_1 *lie inside the line segment connecting* x_2, y_2. *Moreover assume that* $\lambda x_1 + (1-\lambda)y_1 = \lambda x_2 + (1-\lambda)y_2$ *for some* $0 < \lambda < 1$. *Then*

$$h(x_1)^\lambda h(y_1)^{1-\lambda} \geq h(x_2)^\lambda h(y_2)^{1-\lambda}$$

(b) Let h *be a strictly logarithmic concave function on* $\mathbb{R}^m$. *Let* $x \in \mathbb{R}^m$ *and* $a > 0$ *be a real number. Then there exists* $\epsilon > 0$ *such that*

$$h(x) \geq h(y)^\lambda h(z)^{1-\lambda} + \epsilon$$

whenever $\lambda y + (1-\lambda)z = x$ *and* $d(x, z) \geq a$. *Moreover, this* ϵ *is uniform in* x *and* a *if they vary in compact neighborhoods.*

Proof: (a) Let $\lambda_1 = \frac{y_2 - x_1}{y_2 - x_2}$ and $\lambda_2 = \frac{y_2 - y_1}{y_2 - x_2}$. Then $0 < \lambda_1, \lambda_2 < 1$ and

$$x_1 = \lambda_1 x_2 + (1-\lambda_1)y_2,$$

$$y_1 = \lambda_2 x_2 + (1-\lambda_2)y_2.$$

By log-concavity, we have

$$h(x_1) \geq h(x_2)^{\lambda_1} h(y_2)^{1-\lambda_1}$$

and

$$h(y_1) \geq h(x_2)^{\lambda_2} h(y_2)^{1-\lambda_2}$$

So

$$h(x_1)^\lambda h(y_1)^{1-\lambda} \geq h(x_2)^{\lambda\lambda_1 + (1-\lambda)\lambda_2} h(y_2)^{\lambda(1-\lambda_1) + (1-\lambda)(1-\lambda_2)}$$

Part (a) follows from the fact that $\lambda\lambda_1 + (1-\lambda)\lambda_2 = \lambda$ and $\lambda(1-\lambda_1) + (1-\lambda)(1-\lambda_2) = 1-\lambda$.

(b) If $d(x,z) \geq a$, then $h(x) > h(y)^{\lambda} h(z)^{1-\lambda}$ due to strict log-concavity. By part (a), $h(x) - h(y)^{\lambda} h(z)^{1-\lambda}$ is the smallest when $d(x,z) = a$. Define a function

$$g(y,z) := h(x) - h(y)^{\lambda} h(z)^{1-\lambda}$$

It is a continuous function on $\mathbb{R}^m \times \mathbb{R}^m$ and it is positive on the compact set

$$U := \{(y,z) \in \mathbb{R}^{2m} \,|\, d(x,z) = a,\ \lambda y + (1-\lambda) z = x\}$$

So it achieves a minimum $\epsilon > 0$ on U. This ϵ is exactly the one we desired.

Finally, the uniformity of ϵ follows from the continuity of g and the fact that U is contained in a ball of radius $\max\{a, (1-\lambda)a/\lambda\}$ centered at (x,x).

□

Finally, we present the following generalization of Theorem 2 in [102].

Theorem 21 *Let P be a probability measure on $\mathbb{R}^m$ generated by a probability density of the form $f(x) = e^{-Q(x)}$ where $Q(x)$ is a strictly convex function. (Namely, f is a strictly logarithmic concave function.) Then P is a strictly logarithmic concave probability measure.*

Proof: First, we recall the following result used in the proof of Theorem 2 in [102]. This is the inequality (2.4) in [102].

Lemma 10 *Let f, g be nonnegative Borel measurable functions on $\mathbb{R}^m$ and $0 < \lambda < 1$ be a real number. Let*

$$r(t) := \sup_{\lambda x + (1-\lambda) y = t} f(x) g(y).$$

Then we have inequality

$$\int_{\mathbb{R}^m} r(t)\, dt \geq (\int_{\mathbb{R}^m} f^{1/\lambda}(x)\, dx)^{\lambda} (\int_{\mathbb{R}^m} g^{1/(1-\lambda)}(y)\, dy)^{1-\lambda}.$$

Come back to the proof of the Theorem. We need to show that

$$\Pr(\lambda A + (1-\lambda) B) > \Pr(A)^{\lambda} \Pr(B)^{1-\lambda}$$

if $\mu(A \triangle B) > 0$.

Let $f_1(x) = f(x)$ if $x \in A$ and $f_1(x) = 0$ otherwise;

Let $f_2(x) = f(x)$ if $x \in B$ and $f_2(x) = 0$ otherwise;

Let $f_3(x) = f(x)$ if $x \in \lambda A + (1-\lambda)B$ and $f_3(x) = 0$ otherwise.

Without loss of generality, let's assume that $\mu(A \backslash B) > 0$. Notice that the set $V := (\lambda A + (1-\lambda)B \backslash B$ has positive Lebesgue measure. Pick a closed m-dimensional ball $B_a(x_0)$ inside V of small enough radius $a > 0$. We claim that there exist $\epsilon > 0$ such that

$$f_3(t) \geq \epsilon + \sup_{\lambda x + (1-\lambda) y = t} f_1(x)^\lambda f_2(x)^{1-\lambda}$$

for all $t \in B_{a/2}(x_0)$.

Indeed, by Lemma 9 (b), we know for each $t \in B_{a/2}(x_0)$,

$$f_3(t) > \epsilon_t + \sup_{\lambda x + (1-\lambda) y = t,\ d(t,y) > a/2} f_1(x)^\lambda f_2(y)^\lambda$$

for some $\epsilon_t > 0$. Moreover, this ϵ_t varies uniformly in the ball $B_{a/2}(x_0)$. So we can simply take $\epsilon = \inf_{t \in B_{a/2}(x_0)} \epsilon_t > 0$.

Finally, the following inequality concludes the proof:

$$\begin{aligned}
&\int_{\lambda A + (1-\lambda)B} f(x)\, dx = \int_{\mathbb{R}^m} f_3(t)\, dt \\
&= \int_{\mathbb{R}^m} (f_3(t) - \sup_{\lambda x + (1-\lambda) y = t} f_1(x)^\lambda f_2(y)^{1-\lambda})\, dt \\
&+ \int_{\mathbb{R}^m} \sup_{\lambda x + (1-\lambda) y = t} f_1(x)^\lambda f_2(y)^{1-\lambda}\, dt \\
&\geq \epsilon \mu(B_{a/2}(x_0)) + (\int_{\mathbb{R}^m} f_1(x)\, dx)^\lambda (\int_{\mathbb{R}^m} f_2(y)\, dy)^{1-\lambda} \\
&> (\int_A f(x)\, dx)^\lambda (\int_B f(y)\, dy)^{1-\lambda}
\end{aligned}$$

□

Proof of Theorem 19

Proof:[Proof of Theorem 19] Based on the proof of Lemma 8, the equality holds if and only if inequality (5.9) is equality. By Theorem 21, we must have $\mu(H(\vec{\psi}^{(n)}) \triangle H(\vec{\psi}'^{(n)})) = 0$. But $H(\vec{\psi})$ are closed convex sets cut out by hyperplanes of the form

$$\epsilon_{n\pi(m)} - \epsilon_{n\pi(m+1)} \geq \delta_{\pi(m+1)} - \delta_{\pi(m)} + \sum_{k,l} x_n(k)(z_{\pi(m+1)}(l) - z_{\pi(m)}(l))W_{kl}^{s(n)}.$$

So $\mu(H(\vec{\psi}^{(n)}) \triangle H(\vec{\psi}'^{(n)})) = 0$ if and only if $H(\vec{\psi}^{(n)}) = H(\vec{\psi}'^{(n)})$, which happens if and only if

$$\delta_m - \delta_M + \sum_{k,l} x_n(k)(z_m(l) - z_M(l))W_{kl}^{s(n)} = (\delta_m)' - (\delta_M)' + \sum_{k,l} x_n(k)(z_m(l) - z_M(l))(W_{kl}^{s(n)})'$$

for all n, k, l and $m = 1, \cdots, M-1$. Namely, the vector

$$\vec{\tau} = \left((W_{kl}^s - (W_{kl}^s)')_{s,k,l}, (\delta_m - \delta_M - (\delta_m)' + (\delta_M)')_m\right) \in \mathbb{R}^{SKL+M}$$

is a solution of $\widetilde{A}\vec{\tau}^T = 0$. By our assumption, $\widetilde{A}$ has full rank. So $\vec{\tau} = 0$, which says

$$\begin{cases} \delta_m = (\delta_m)' + c \quad \text{where } c = \delta_M - (\delta_M)' \\ W_{kl}^s = (W_{kl}^s)' \end{cases}$$

This concludes the proof of Theorem 19.

□

5.6.2 On Identifiability

The main purpose of this section is to prove Theorem 20. We first recall the definition of *nice functions.*

Definition 8 *Let $\phi(x)$ be a smooth pdf defined on $\mathbb{R}$ or $[0, \infty)$ and let $\Phi(x)$ be the associated cdf. For each $i > 0$, we write $\phi^{(i)}(x)$ for the i-th derivative of $\phi(x)$. Let $g_i(x) = \frac{\phi^{(i+1)}(x)}{\phi^{(i)}(x)}$. The function Φ is called* ***nice*** *if it satisfies one of the following two mutually exclusive conditions:*

(a) ***(Type 1)*** *For any two x_1, x_2, the sequence $\frac{g_i(x_1)}{g_i(x_2)}$ converges to some value in $\mathbb{R}$ (as $i \to \infty$) only if either*

- $x_1 = x_2$; *or*
- $x_1 = -x_2$ *and* $\frac{g_i(x_1)}{g_i(x_2)} \to -1$ *as* $i \to \infty$.

(b) ***(Type 2)*** *For all* x_1, x_2, *the ratio* $\frac{g_i(x_1)}{g_i(x_2)}$ *converges to* 1, *as* $i \to \infty$. *Moreover, for any* $x_1 \neq x_2$, *there exists an odd positive number* m *such that* $\phi^{(m)}(x_1) \neq \phi^{(m)}(x_2)$.

Proof: [Proof of Lemma 7] Let $\phi(x)$ be the pdf associated to the cdf $\Phi(x)$. By assumption, ϕ is nice, which means $\phi(x)$ is of Type 1 or Type 2 as in the above definition.

Consider the Taylor expansion at 0. Note that the $(i+1)$-th derivatives of $\Phi(\delta + xW^s)$ is just $(W^s)^{i+1}\phi^{(i)}(\delta + xW^s)$. So, the induced identity on the $(i+1)$-th Taylor coefficient is

$$\sum_{s=1}^{S} \gamma_s (W^s)^{i+1} \phi^{(i)}(\delta) = \sum_{s=1}^{S'} \gamma'_s (W'^s)^{i+1} \phi^{(i)}(\delta') \tag{5.10}$$

Let us assume

$$|W^1| > |W^2| > \cdots > |W^S|,$$

$$|W'^1| > |W'^2| > \cdots > |W'^{S'}|,$$

and $|W^1| \geq |W'^1|$.

Dividing the $(i+2)$-th coefficient by the $(i+1)$-th coefficient, we obtain

$$\frac{\phi^{(i+1)}(\delta)}{\phi^{(i)}(\delta)} \cdot \frac{\sum_{s=1}^{S} \gamma_s (W^s)^{i+2}}{\sum_{s=1}^{S} \gamma_s (W^s)^{i+1}} = \frac{\phi^{(i+1)}(\delta')}{\phi^{(n)}(\delta')} \cdot \frac{\sum_{s=1}^{S'} \gamma'_s (W'^s)^{i+2}}{\sum_{s=1}^{S'} \gamma'_s (W'^s)^{i+1}}$$

Let $g_n(\delta) = \frac{\phi^{(i+1)}(\delta)}{\phi^{(i)}(\delta)}$. Then $\frac{g_i(\delta)}{g_i(\delta')} \to \frac{W'^1}{W^1} \in \mathbb{R}$ as $i \to \infty$. Now let's discuss Type 1 and Type 2 separately.

(i) (**Type 1**)

In this case, we must have $\delta = \delta'$, $W'^1 = W^1$ or, $\delta = -\delta'$, $W'^1 = -W^1$. However, if i is odd, the equation (5.10) tells us that $\phi^{(i)}(\delta)$ and $\phi^{(i)}(\delta')$ must have the same sign. This rules out the possibility of $\delta = -\delta'$. Thus $\delta = \delta'$ and $W^1 = W'^1$. Now equation (5.10) becomes

$$\sum_{s=1}^{S} \gamma_s (W^s)^{i+1} = \sum_{s=1}^{S'} \gamma'_s (W'^s)^{i+1}.$$

A classical identifiability result concludes that $S = S'$, $\gamma_s = \gamma'_s$, and $W^s = W'^s$ for all s (after a permutation).

(ii) (**Type 2**)

In this case, $\frac{W'^1}{W^1}$ must equal 1. Namely, $W^1 = W'^1$. Now look at equation (5.10). Since

$$\frac{g_i(\delta)}{g_i(\delta')} = \frac{\phi^{(i+1)}(\delta)/\phi^{(i+1)}(\delta')}{\phi^{(i)}(\delta)/\phi^{(i)}(\delta')} \to 1$$

as $i \to \infty$, we know that $\frac{\phi^{(i)}(\delta)}{\phi^{(i)}(\delta')}$ does not grow as fast as exponentially. So, again by the classical identifiability result, we know that $\gamma_1 = \gamma'_1$. Repeating this process, we know that $W^2 = W'^2$, $\gamma_2 = \gamma'_2$, and so on. Therefore, we also know $\phi^{(i)}(\delta) = \phi^{(i)}(\delta')$ for all odd i. However, by assumption, we must have $\delta = \delta'$.

□

Proof:[Proof of Theorem 20] Consider all possible permutations in which product 2 is more preferred to product 1. Define $\mathfrak{S}(1;2) := \{\pi \mid 1 \text{ shows after } 2 \text{ in the order } \pi\}$. Then

$$\sum_{s=1}^{S} \gamma_s \Pr(u_1 > u_2 | X, Z, \Psi) = \sum_{\pi \in \mathfrak{S}(1;2)} \sum_{s=1}^{S} \gamma_s \Pr(\pi | X, Z, \Psi)$$

So

$$\sum_{s=1}^{S} \gamma_s \Pr(u_1 > u_2 | X, Z, \Psi) = \sum_{s=1}^{S'} \gamma'_s \Pr(u_1 > u_2 | X, Z, \Psi')$$

Unwinding the definition, this means

$$\begin{aligned} &\sum_{s=1}^{S} \gamma_s \Pr(\epsilon_2 > \epsilon_1 | \delta_1 - \delta_2 + \sum_{k,l} x(k) W^s_{kl}(z_1(l) - z_2(l))) \\ =&\sum_{s=1}^{S'} \gamma'_s \Pr(\epsilon_2 > \epsilon_1 | \delta'_1 - \delta'_2 + \sum_{k,l} x(k) W'^s_{kl}(z_1(l) - z_2(l))) \end{aligned}$$

Namely,

$$\begin{aligned} &\sum_{s=1}^{S} \gamma_s \Phi(\delta_1 - \delta_2 + \sum_{k,l} x(k) W^s_{kl}(z_1(l) - z_2(l))) \\ =&\sum_{s=1}^{S'} \gamma'_s \Phi(\delta'_1 - \delta'_2 + \sum_{k,l} x(k) W'^s_{kl}(z_1(l) - z_2(l))) \end{aligned}$$

Again, we may assume all of the intervals I_k contain 0. If we fix $x(2), \cdots, x(K)$, we can think of $x(1)$ as a variable. By the previous Lemma, we must have

- $S = S'$
- $\delta_1 - \delta_2 + \sum_{k \geq 2} W^s_{kl}(z_1(l) - z_2(l)) = \delta'_1 - \delta'_2 + \sum_{k \geq 2} W'^s_{kl}(z_1(l) - z_2(l))$
- after a permutation of $\{1, \cdots, S\}$, $\gamma_s = \gamma'_s$, and $\sum_l W^s_{1l}(z_1(l) - z_2(l)) = \sum_l W'^s_{1l}(z_1(l) - z_2(l))$.

Since $x(k)$'s can be arbitrary in the intervals I_k's, we must have $\delta_1 - \delta_2 = \delta'_1 - \delta'_2$ and

$$\sum_l W^s_{kl}(z_1(l) - z_2(l)) = \sum_l W'^s_{kl}(z_1(l) - z_2(l))$$

for all $1 \leq k \leq K$. Now we can repeat the above for any two products. In particular, we know that $\delta = \delta'$ (up to a shift), and

$$\sum_l (W^s_{kl} - W'^s_{kl})(z_m(l) - z_M(l)) = 0$$

for all $1 \leq m \leq M - 1$. By assumption, the $M - 1$ by L matrix $Z' = (z_m(l) - z_M(l))$ had rank L. So the above systems of equation has a unique solution. Namely, $W^s_{kl} = W'^s_{kl}$ for all k, l, s. □

5.6.3 Examples of Nice CDFs

Normal Distributions

Let $\phi(x) = e^{-\frac{x^2}{2}}$. Write $\phi^{(i)}(x) = f_i(x)e^{-\frac{x^2}{2}}$. For example, $f_0(x) = 1$, $f_1(x) = -x$, and so on. We have the recursive relation $f_{i+1}(x) = -xf_i(x) + f'_{i-1}(x)$. In particular, we know that $f_i(x)$ is a polynomial in $\mathbb{R}[x]$ of degree i.

Lemma 11 *We have the following recursive relations.*

(a) $f_{i+1}(x) = -xf_i(x) - (i-1)f_{i-1}(x)$

(b) $f'_{i+1}(x) = -if_i(x)$

Proof: Assume the result holds for stage i. For stage $i+1$, we have

$$f_{i+2}(x) = -xf_{i+1}(x) + f'_{i+1}(x) = -xf_{i+1}(x) - if_i(x)$$

and

$$\begin{aligned} f'_{i+2}(x) &= (-xf_{i+1}(x) - if_i(x))' \\ &= -f_{i+1}(x) - xf'_{i+1}(x) - if'_i(x) \\ &= -f_{i+1}(x) - ixf_i(x) - if'_i(x) \\ &= -f_{i+1}(x) - i(xf_i(x) + f'_i(x)) \\ &= -f_{i+1}(x) - if_{i+1}(x) \\ &= -(i+1)f_{i+1}(x) \end{aligned}$$

□

Define $g_i(x) = \frac{f_{i+1}(x)}{f_i(x)}$, which is, *a priori*, a rational function with real coefficients. Dividing $f_i(x)$ on both side of the relation (a) in the previous lemma, we obtain

$$g_i(x) = -x - \frac{i-1}{g_{i-1}(x)}$$

Lemma 12 *Given any $\delta \in \mathbb{R}$, the sequence $\{g_i(\delta)\}$ does not converge to any number in $\mathbb{R} \cup \{\pm\infty\}$, as $i \to \infty$.*

Proof: If $\{g_i(\delta)\}$ does converge to some $a \in \mathbb{R}$, then

$$a = \lim_{i\to\infty} g_i(\delta) = \lim_{i\to\infty} (-\delta - \frac{i-1}{g_{i-1}(\delta)}) = -\delta - \lim_{i\to\infty} \frac{i-1}{g_{i-1}(\delta)} \to \infty,$$

a contradiction.

On the other hand, if $g_i(x) \to +\infty$, then $-\delta - \frac{i-1}{g_{i-1}(\delta)} \to +\infty$. But it's less than $|\delta|$, a contradiction. Similarly, $g_i(\delta)$ cannot converge to $-\infty$. □

Lemma 13 *Let δ, δ' be two real numbers. Then $\frac{g_i(\delta)}{g_i(\delta')} \to c \in \mathbb{R}$ (as $i \to \infty$) if and only if either $c = 1$, $\delta = \delta'$ or $c = -1$, $\delta = -\delta'$.*

Proof: We have $g_i(\delta) + \delta = -\frac{i-1}{g_{i-1}(\delta)}$ and $g_i(\delta') + \delta' = -\frac{i-1}{g_{i-1}(\delta')}$. Let $c_i = \frac{g_i(\delta)}{g_i(\delta')}$. Then

$$\frac{g_i(\delta) + \delta}{g_i(\delta') + \delta'} = \frac{g_{i-1}(\delta')}{g_{i-1}(\delta)}.$$

Thus

$$c_i + \frac{\delta - c_i\delta'}{g_i(\delta') + \delta'} = \frac{1}{c_{i-1}}.$$

Taking limit, we get

$$\lim_{i\to\infty} \frac{\delta - c\delta'}{g_i(\delta') + \delta'} = \frac{1}{c} - c.$$

However, according to the lemma, $\frac{1}{g_i(\delta')+\delta'}$ does not converge to any real number. So we must have $\delta - c\delta' = 0$. This implies $\frac{1}{c} - c = 0$. Namely, $c = \pm 1$. If $c = 1$, we must have $\delta = \delta'$ and if $c = -1$, we get $\delta = -\delta'$.

On the other hand, it's easy to see that $\frac{g_i(\delta)}{g_i(\delta')} \equiv 1$ if $\delta = \delta'$, while $\frac{g_i(\delta)}{g_i(\delta')} = -1$ if $\delta = -\delta'$. This completes the proof. □

Exponential Distributions

Let $\phi(x) = \lambda e^{-\lambda x}$ $(x \geq 0)$. Then $\phi^{(i)}(x) = (-1)^i\lambda^{i+1}e^{-\lambda x}$ and $g_i(x) = \frac{\phi^{(i+1)}(x)}{\phi^{(i)}(x)} = -\lambda$, a constant! In particular, for any x_1, x_2, the ratio $\frac{g_i(x_1)}{g_i(x_2)}$ is always 1. Moreover, if $x_1 \neq x_2$, then $\frac{\phi^{(1)}(x_1)}{\phi^{(1)}(x_2)} = e^{\lambda(x_2-x_1)} \neq 1$. Namely, $\phi^{(1)}(x_1) \neq \phi^{(1)}(x_2)$. Therefore, $\phi(x)$ is a nice pdf of type 2.

Gamma Distributions

Let $\phi(x) = \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$ where $\alpha, \beta > 0$, $\alpha \neq 1$, and $x > 0$. Write $\phi^{(i)}(x) = f_i(x) \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-i-1}e^{-\beta x}$. For example, $f_0(x) = 1, f_1(x) = \alpha - 1 - \beta x$, and so on.

We have the recursion relation

$$f_i(x) = (\alpha - i - \beta x)f_{i-1}(x) + xf'_{i-1}(x).$$

In particular, we know that $f_i(x)$ is a polynomial in $\mathbb{R}[x]$ of degree i.

Lemma 14 *We have the following recursive relations:*

(a) $f_i(x) = (\alpha - i - \beta x)f_{i-1}(x) - (i-1)\beta x f_{i-2}(x)$.

(b) $f_i'(x) = -n\beta f_{i-1}(x)$.

Proof: Assume the result holds for stage i. For stage $i+1$, we have

$$\begin{aligned} f_{i+1}(x) &= (\alpha - i - 1 - \beta x) f_i(x) + x f_i'(x) \\ &= (\alpha - i - 1 - \beta x) f_i(x) - x(i\beta x f_{i-1}(x)) \end{aligned}$$

and

$$\begin{aligned} f_{i+1}'(x) &= ((\alpha - i - 1 - \beta x) f_i(x) - i\beta x f_{i-1}(x))' \\ &= -\beta f_i(x) + (\alpha - i - 1 - \beta x) f_i'(x) - i\beta f_{i-1}(x) - n\beta x f_{i-1}'(x) \\ &= -\beta f_i(x) - (\alpha - i - 1 - \beta x) i\beta f_{i-1}(x) - i\beta f_{i-1}(x) - i\beta x f_{i-1}'(x) \\ &= -\beta f_i(x) - i\beta((\alpha - i - \beta x) f_{i-1}(x) + x f_{i-1}'(x)) \\ &= -\beta f_i(x) - i\beta f_i(x) \\ &= -(i+1)\beta f_i(x) \end{aligned}$$

□

Notice that $g_i(x) = \frac{\phi^{(i+1)}(x)}{\phi^{(i)}(x)} = \frac{1}{x} \cdot \frac{f_{i+1}(x)}{f_i(x)}$. Replacing i by $i+1$, the recursion in Lemma 14 gives

$$f_{i+1}(x) = (\alpha - i - 1 - \beta x) f_i(x) - x(i\beta x f_{i-1}(x)).$$

Diving by $x f_i(x)$ on both sides, we obtain

$$g_i(x) = \frac{\alpha - 1 - i - \beta x}{x} - \frac{i\beta}{x g_{i-1}(x)}.$$

Lemma 15 *For any given $x > 0$, we have $g_i(x) \sim -\frac{i}{x} + o(i)$ for i sufficiently large.*

Consequently, for any x_1, x_2, we must have $\frac{g_i(x_1)}{g_i(x_2)} \to \frac{x_2}{x_1}$ as $i \to \infty$.

5.7 Conclusions

In this chapter, we have proposed an extension of GRUMs in which we allow agents to adopt heterogeneous types. We develop a theory establishing the identifiability of the mixture

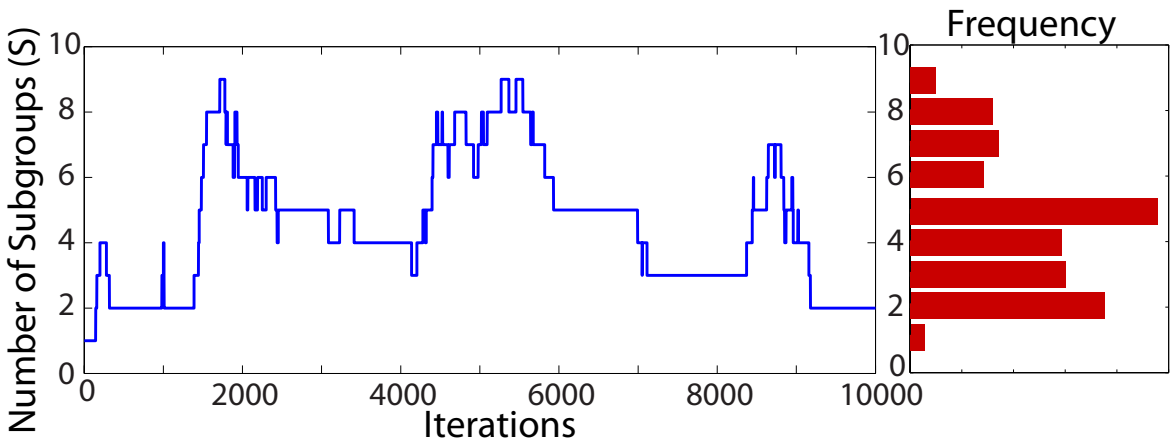


Figure 5.3: Left Panel: 10000 samples for S in Synthetic data, where the true S is 5. Right Panel: Histogram of the samples for S with max at 5 and mean at 4.56.

	Synthetic True types				Sushi
Type	One	two	Three	Four	sushi
one type	**-2069**	-2631	-2780	-2907	-2880
two types	-2755	**-2522**	**-2545**	-2692	-2849
three types	-2796	-2642	**-2582**	-2790	**-2819**
four types	-2778	-2807	-2803	**-2593**	-2850

Table 5.1: Performance of the method for different number of true types and number of types in algorithm in terms of log posterior. All the standard deviations are between 15 and 20. Bold numbers indicate the best performance in their column with statistical significance of 95%.

model when we observe ranking data. Our theoretical results for identifiability show that the number of types and the parameters associated with them can be identified. Moreover, we prove uni-modality of the likelihood (or posterior) function when types are observable.

We propose a scalable algorithm for inference, which can be parallelized for use on very large data sets. Our experimental results show that models with multiple types provide a significantly better fit in real-world data. By clustering agents into multiple types, our estimation algorithm allows choices to be correlated across agents of the same type, without making any *a priori* assumptions on how types of agents are to be partitioned.

This use of machine learning techniques complements various approaches in economics [21, 17, 18] by allowing the researcher to have additional flexibility in dealing with missing data or unobserved agent characteristics. We expect the development of these techniques to grow in importance as large, individual-level data-sets become increasingly available. In future research we intend to pursue applications of this method to problems of economic interest e.g. demand estimation and econometrics.

Chapter 6

Conclusions

There are two important components that drive research presented in this thesis. The first component is provided by the explosion in data on human choice behavior. Every day there are billions of clicks on Google search results, millions of purchases on Amazon, and billions of likes on Facebook. This data can benefit from richer models, better able to capture the underlying complexity and heterogeneity of choice behavior.

The second component is the increase in computation power that allows us to compute more efficiently. This provides the capability for estimation and inference with models that would not be tractable with just twentieth century computational powers.

This parallel advance in measurement and computation is leading to a new era in the digital revolution. However, the research in understanding choice has considerable inertia, prompting the need for new approaches to model building and inference.

This thesis provides a step forward in extending choice models and the algorithms that are available for estimation and inference. There remains a lot more to explore from the direction of econometrics, social choice and recommender systems. Next steps should be to apply these new models to large data sets and look for a deeper understanding of human choice. This will require a better appreciation and synthesis across the existing literatures in psychophysics, sociology, economics, computer science and statistics.

I wish to finish with the words of R.A.Fisher:

More attention to the History of Science is needed, as much by scientists as by historians, and especially by biologists, and this should mean a deliberate attempt to understand the thoughts of the great masters of the past, to see in what circumstances or intellectual milieu their ideas were formed, where they took the wrong turning or stopped short on the right track.

Bibliography

[1] Daniel A. Ackerberg. Advertising, learning, and consumer choice in experience goods: an empirical examination. *International Economic Review*, 44(3):1007–1040, 2003.

[2] Ernest Adams and Samuel Messick. An axiomatic formulation and generalization of successive intervals scaling. *Psychometrika*, 23(4):355–368, 1958.

[3] Nir Ailon. Aggregation of partial rankings, p-ratings and top-m lists. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2007.

[4] N. Atienza, J. Garcia-Heras, and J.M. Muñoz-Pichardo. A new condition for identifiability of finite mixture distributions. *Metrika*, 63(2):215–221, 2006.

[5] Bruce Aune. *Philosophy and Phenomenological Research*, 50(4):pp. 845–848, 1990.

[6] Hossein Azari Soufiani and Edoardo M Airoldi. Graphlet decomposition of a weighted network. *arXiv preprint arXiv:1203.2821*, 2012.

[7] Hossein Azari Soufiani and William Chen. *StatRank: Statistical Rank Aggregation: Inference, Evaluation, and Visualization*, 2013. R package version 0.0.2.

[8] Hossein Azari Soufiani, William Z. Chen, David C. Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2706–2714, Lake Tahoe, NV, USA, 2013.

[9] Hossein Azari Soufiani, David M. Chickering, Denis X. Charles, and David C. Parkes. Approximating the shapley value via multi-issue decompositions. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '14, pages 1209–1216, Richland, SC, 2014. International Foundation for Autonomous Agents and Multiagent Systems.

[10] Hossein Azari Soufiani, Hansheng Diao, Zhenyu Lai, and David C. Parkes. Generalized random utility models with multiple types. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, USA, 2013.

[11] Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Random utility theory for social choice. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 126–134, Lake Tahoe, NV, USA, 2012.

[12] Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Preference Elicitation For General Random Utility Models. In *Proceedings of Uncertainty in Artificial Intelligence (UAI)*, Bellevue, Washington, USA, 2013.

[13] Hossein Azari Soufiani, David C. Parkes, and Lirong Xia. Computing parametric ranking models via rank-breaking. In *Proceedings of The 31st International Conference on Machine Learning*, pages 360–368, 2014.

[14] Patrick Bajari and C. Lanier Benkard. Discrete choice models as structural models of demand: Some economic implications of common approaches. Technical report, Working Paper, 2003.

[15] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. James O. Berger, 2nd edition, 1985.

[16] James Berkovec and John Rust. A nested logit model of automobile holdings for one vehicle households. *Transportation Research Part B: Methodological*, 19(4):275–285, 1985.

[17] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890, 1995.

[18] Steven Berry, James Levinsohn, and Ariel Pakes. Voluntary export restraints on automobiles: evaluating a trade policy. *The American Economic Review*, 89(3):400–430, 1999.

[19] Steven Berry, James Levinsohn, and Ariel Pakes. Differentiated products demand systems from a combination of micro and macro data: The new car market. *Journal of Political Economy*, 112(1):68–105, 2004.

[20] Steven Berry and Ariel Pakes. Some applications and limitations of recent advances in empirical industrial organization: Merger analysis. *The American Economic Review*, 83(2):247–252, 1993.

[21] Steven T Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262, 1994.

[22] Henry David Block and Jacob Marschak. Random orderings and stochastic theories of responses. In *Contributions to Probability and Statistics*, pages 97–132, 1960.

[23] Henry David Block, Jacob Marschak, et al. Random orderings and stochastic theories of responses. *Contributions to probability and statistics*, 2:97–132, 1960.

[24] Edwin Bonilla, Shengbo Guo, and Scott Sanner. Gaussian process preference elicitation. In *Advances in Neural Information Processing Systems 23*, pages 262–270. 2010.

[25] Craig Boutilier. On the foundations of expected expected utility. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 285–290, Acapulco, Mexico, 2003.

[26] Craig Boutilier. Computational Decision Support: Regret-based Models for Optimization and Preference Elicitation. In P. H. Crowley and T. R. Zentall, editors, *Com-*

parative Decision Making: Analysis and Support Across Disciplines and Applications. Oxford University Press, 2013.

[27] Ralph A Bradley. Another interpretation of a model for paired comparisons. *Psychometrika*, 30(3):315–318, 1965.

[28] Ralph Allan Bradley. Some statistical methods in taste testing and quality evaluation. *Biometrics*, (9):22–38, 1953.

[29] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[30] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[31] Stephen P Brooks, PAULO Giudici, and Gareth O Roberts. Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):3–39, 2003.

[32] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng, editors. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, 2011.

[33] Francois Caron and Arnaud Doucet. Efficient Bayesian Inference for Generalized Bradley-Terry Models. *Journal of Computational and Graphical Statistics*, 21(1):174–196, 2012.

[34] Urszula Chajewska, Daphne Koller, and Ron Parr. Making rational decisions using adaptive utility elicitation. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 363–369, Austin, TX, USA, 2000.

[35] Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273—304, 1995.

[36] Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, pages 1019–1041, 2005.

[37] Marquis de Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris: L'Imprimerie Royale, 1785.

[38] Vincent Conitzer. *Computational aspects of preference aggregation*. PhD thesis, Carnegie Mellon University, 2006.

[39] Vincent Conitzer, Matthew Rognlie, and Lirong Xia. Preference functions that score rankings and maximum likelihood estimation. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 109–115, Pasadena, CA, USA, 2009.

[40] Vincent Conitzer and Tuomas Sandholm. Vote elicitation: Complexity and strategy-proofness. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 392–397, Edmonton, AB, Canada, 2002.

[41] Vincent Conitzer and Tuomas Sandholm. Common voting rules as maximum likelihood estimators. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 145–152, Edinburgh, UK, 2005.

[42] Gerard Debreu. The mathematization of economic theory. *American Economic Review*, 81(1):1–7, 1991.

[43] A. Donagan. *Choice: the Essential Element in Human Action*. Studies in philosophical psychology. Routledge & Kegan Paul, 1987.

[44] Cynthia Dwork, Ravi Kumar, Moni Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th World Wide Web Conference*, pages 613–622, 2001.

[45] F. Y. Edgeworth. On the probable errors of frequency-constants (contd.). *Journal of the Royal Statistical Society*, 71(3):pp. 499–512, 1908.

[46] Patricia Everaere, Sébastien Konieczny, and Pierre Marquis. The strategy-proofness landscape of merging. *Journal of Artificial Intelligence Research*, 28:49–105, 2007.

[47] Lester R. Ford, Jr. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8):28–33, 1957.

[48] Chris Fraley and Adrian E. Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *THE COMPUTER JOURNAL*, 41(8):578–588, 1998.

[49] Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer, 2006.

[50] John Geweke, Michael Keane, and David Runkle. Alternative computational approaches to inference in the multinomial probit model. *Review of Economics and Statistics*, pages 609–632, 1994.

[51] Isobel Claire Gormley and Thomas Brendan Murphy. Analysis of Irish third-level college applications data. *Journal of the Royal Statistical Society Series A*, 169(2):361–379, 2006.

[52] Isobel Claire Gormley and Thomas Brendan Murphy. A latent space model for rank data. In *Statistical Statistical Network Analysis: Models, Issues and New Directions. LNCS*, volume 4503, pages 90–107, 2007.

[53] Isobel Claire Gormley and Thomas Brendan Murphy. Exploring voting blocs within the irish exploring voting blocs within the irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103(483):1014–1027, 2008.

[54] Isobel Claire Gormley and Thomas Brendan Murphy. A grade of membership model for rank data. *Bayesian Analysis*, 4(2):265–296, 2009.

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

[56] Bettina Grün and Friedrich Leisch. Identifiability of finite mixtures of multinomial logit models with varying and fixed effects. *Journal of Classification*, 25(2):225–247, 2008.

[57] John Guiver and Edward Snelson. Bayesian inference for Plackett-Luce ranking models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML-09, pages 377–384, Montreal, Quebec, Canada, 2009.

[58] Lars Peter Hansen. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029–1054, 1982.

[59] David A Harville. Assigning probabilities to the outcomes of multi-entry competitions. *Journal of the American Statistical Association*, 68(342):312–316, 1973.

[60] Jerry A. Hausman. Valuation of new goods under perfect and imperfect competition. In *The economies of new goods*, pages 207–248. University of Chicago Press, 1996.

[61] James J. Heckman and Burton Singer. Econometric duration analysis. *Journal of Econometrics*, 24(1-2):63–132, 1984.

[62] Edith Hemaspaandra, Holger Spakowski, and Jörg Vogel. The complexity of Kemeny elections. *Theoretical Computer Science*, 349(3):382–391, December 2005.

[63] Igal Hendel and Aviv Nevo. Measuring the implications of sales and consumer inventory behavior. *Econometrica*, 74(6):1637–1673, 2006.

[64] Katherine Ho. The welfare effects of restricted hospital choice in the us medical care market. *Journal of Applied Econometrics*, 21(7):1039–1079, 2006.

[65] Fangxin Hong and Rainer Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, 24(3):374–382, 2008.

[66] Neil Houlsby, Jose Miguel Hernandez-Lobato, Ferenc Huszar, and Zoubin Ghahramani. Collaborative Gaussian processes for preference learning. In *Proceedings of the Annual*

Conference on Neural Information Processing Systems (NIPS), pages 2105–2113. Lake Tahoe, NV, USA, 2012.

[67] David R. Hunter. MM algorithms for generalized Bradley-Terry models. In *The Annals of Statistics*, volume 32, pages 384–406, 2004.

[68] Kamel Jedidi, Harsharanjeet S. Jagpal, and Wayne S. DeSarbo. Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, 16(1):39–59, 1997.

[69] Toshihiro Kamishima. Nantonac collaborative filtering: Recommendation based on order responses. In *Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 583–588, Washington, DC, USA, 2003.

[70] Jérôme Lang and Lirong Xia. Sequential composition of voting rules in multi-issue domains. *Mathematical Social Sciences*, 57(3):304–324, 2009.

[71] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[72] Tie-Yan Liu. *Learning to Rank for Information Retrieval*. Springer, 2011.

[73] Thomas A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 44:226–233, 1982.

[74] Tyler Lu and Craig Boutilier. Learning Mallows models with pairwise preferences. In *Proceedings of the Twenty-Eighth International Conference on Machine Learning (ICML 2011)*, pages 145–152, Bellevue, WA, USA, 2011.

[75] Tyler Lu and Craig Boutilier. Robust approximation and incremental elicitation in voting protocols. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 287–293, Barcelona, Catalonia, Spain, 2011.

[76] R. Duncan Luce and Howard Raiffa. *Games and Decisions.* John Wiley and Sons, New York, 1957. Dover republication 1989.

[77] Robert Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis.* Wiley, 1959.

[78] Colin L. Mallows. Non-null ranking model. *Biometrika*, 44(1/2):114–130, 1957.

[79] Andrew Mao, Ariel D. Procaccia, and Yiling Chen. Better human computation through principled voting. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, Bellevue, WA, USA, 2013.

[80] John I. Marden. *Analyzing and modeling rank data.* Chapman & Hall, 1995.

[81] John I Marden. *Analyzing and modeling rank data*, volume 64. CRC Press, 1996.

[82] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In *Frontiers of Econometrics*, pages 105–142, New York, NY, 1974. Academic Press.

[83] Daniel McFadden. The measurement of urban travel demand. *Journal of Public Economics*, 3(4):303–328, 1974.

[84] Daniel McFadden. Modelling the choice of residential location. In Daniel McFadden, A Karlqvist, L Lundqvist, F Snickars, and J Weibull, editors, *Spatial Interaction Theory and Planing Models*, pages 75–96. New York: Academic Press, 1978.

[85] Daniel McFadden. Econometric models for probabilistic choice among products. *Journal of Business*, pages S13–S29, 1980.

[86] Daniel McFadden. Economic choices. *American Economic Review*, pages 351–378, 2001.

[87] Daniel McFadden, Antti Talvitie, Stephen Cosslett, Ibrahim Hasan, Michael Johnson, Fred Reid, and Kenneth Train. *Demand model estimation and validation*, volume 5. Institute of Transportation Studies, 1977.

[88] Daniel McFadden and Kenneth Train. Mixed MNL models for discrete response. *Journal of applied Econometrics*, 15(5):447–470, 2000.

[89] Geoffrey McLachlan and David Peel. *Finite mixture models*. Wiley. com, 2004.

[90] Gormley-Claire McParland, Damien. Clustering ordinal data via latent variable models. IFCS 2013 Conference of the International Federation of Classification Societies, Tilburg University, The Netherlands, 2013.

[91] Marina Meila and Harr Chen. Dirichlet process mixtures of generalized mallows models. *arXiv preprint arXiv:1203.3496*, 2012.

[92] Carl N. Morris. Natural Exponential Families with Quadratic Variance Functions. *Annals of Statistics*, 10(1):65–80, 1982.

[93] Frederick Mosteller. Remarks on the method of paired comparisons: III. a test of significance for paired comparisons when equal standard deviations and equal correlations are assumed. *Psychometrika*, 16(2):207–218, 1951.

[94] R. M. Neal. Software for slice sampling, March 2008.

[95] Radford M. Neal. Slice sampling. *Annals of Statistics*, 31:705–767, 2003.

[96] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2483–2491, Lake Tahoe, NV, USA, 2012.

[97] Emanuel Parzen. On estimation of a probability density function and mode. In *The Annals of Mathematical Statistics*, pages 847–1226, 1962.

[98] Thomas Pfeiffer, Xi Alice Gao, Andrew Mao, Yiling Chen, and David G. Rand. Adaptive Polling and Information Aggregation. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, pages 122–128, Toronto, Canada, 2012.

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

[100] Daryl Pregibon and William D Heavlin. Performance tournaments with crowdsourced judges. 2013.

[101] Daryl Pregibon and Williiam D Heavlin. Performance tournaments with crowdsourced judges. In *Proceedings of the American Statistical Association, section on marketing statistics*, 732 North Washtington Street, Alexandria, VA 22314-1943, 2013.

[102] András Prékopa. Logarithmic concave measures and related topics. In *Stochastic Programming*, pages 63–82. Academic Press, 1980.

[103] Ariel D. Procaccia, Sashank J. Reddi, and Nisarg Shah. A maximum likelihood approach for selecting sets of alternatives. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, 2012.

[104] Frank Proschan and Yung L. Tong. Chapter 29. log-concavity property of probability measures. *FSU techinical report Number M-805*, pages 57–68, 1989.

[105] Magnus Roos, Jörg Rothe, and Björn Scheuermann. How to calibrate the scores of biased reviewers by quadratic programming. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 255–260, 2011.

[106] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function. In *The Annals of Mathematical Statistics*, pages 569–878, 1956.

[107] Tuomas Sandholm and Craig Boutilier. Preference elicitation in combinatorial auctions. In Peter Cramton, Yoav Shoham, and Richard Steinberg, editors, *Combinatorial Auctions*, chapter 10, pages 233–263. MIT Press, 2006.

[108] Hal Stern. Models for distributions on permutations. *Journal of the American Statistical Association*, 85(410):pp. 558–564, 1990.

[109] Mahlet G. Tadesse, Naijun Sha, and Marina Vannucci. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617, 2005.

[110] Daniel Tarlow, Ryan Prescott Adams, and Richard S. Zemel. Randomized optimum models for structured prediction. In *Proceedings of the Fifthteenth International Con-*

ference on Artificial Intelligence and Statistics (AISTATS-11), pages 1221–1229, La Palma, Canary Islands, 2012.

[111] Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.

[112] Nicolaus Tideman. *Collective Decisions and Voting: The Potential for Public Choice*. Ashgate Publishing, 2006.

[113] Kenneth Train. A recursive estimator for random coefficient models. *University of California, Berkeley*, 2007.

[114] Kenneth Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[115] Kenneth E Train. Em algorithms for nonparametric estimation of mixing distributions. *Journal of Choice Modelling*, 1(1):40–69, 2008.

[116] Michel Truchon. Borda and the maximum likelihood approach to vote aggregation. *Mathematical Social Sciences*, 55(1):96–102, 2008.

[117] John von Neumann and Oskar Morgenstern. *Theory of games and economic behavior*. Princeton University Press, 1947.

[118] Joan Walker and Moshe Ben-Akiva. Generalized random utility model. *Mathematical Social Sciences*, 43(3):303–343, 2002.

[119] Greg C. G. Wei and Martin A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

[120] Fen Xia, Tie-Yan Liu, Jue Wang, Wensheng Zhang, and Hang Li. Listwise Approach to Learning to Rank: Theorem and Algorithm. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning (ICML-08)*, Helsinki, Finland, 2008.

[121] Lirong Xia and Vincent Conitzer. A maximum likelihood approach towards aggregating partial orders. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 446–451, Barcelona, Catalonia, Spain, 2011.

[122] Lirong Xia, Vincent Conitzer, and Jérôme Lang. Aggregating preferences in multi-issue domains by using maximum likelihood estimators. In *Proceedings of the Ninth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, pages 399–406, Toronto, Canada, 2010.

[123] John I Yellott Jr. The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144, 1977.

[124] H. Peyton Young. Optimal voting rules. *Journal of Economic Perspectives*, 9(1):51–64, 1995.