

Robust Rank Aggregation and its Applications

Yuangang Pan

Faculty of Engineering and Information Technology
University of Technology Sydney

This dissertation is submitted for the degree of
Doctor of Philosophy

September 2019

Certificate of Original Authorship

I hereby declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. The contents of this dissertation are original and have not been submitted in whole or in part for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:

Signature: Signature removed prior to publication.

Date:

24 Dec 2019

I would like to dedicate this thesis to my family who, through thick and thin, has been there for me. Their support and drive are what have made me who I am.

Acknowledgements

First and foremost, my utmost gratitude goes to my supervisor Professor Ivor W. Tsang, for his cordiality and generousness in offering me this research opportunity in his group. Life has not been easy for me, especially at the beginning of this adventure. He has always been there for me, serving as a friend and a mentor. His illuminating instructions and insightful guidance helped me to avoid being unnecessarily stuck and kept me focused on the nature of science itself. His consistent support and encouragement helped me to survive the many struggles I experienced on my Ph.D. journey. His depth and breadth of knowledge, insightful vision, and wisdom have not only guided me through my Ph.D. research but also shaped what I have achieved in this dissertation. His dedication and professionalism also set a great example from which I can learn and benefit from in my future career.

During my time in Sydney, I have had the privilege of meeting many excellent people. Being more senior to me, I have appreciated how they have generously shared many insightful thoughts on life and research. Having them as companions in my daily life makes me feel life brighter. Their enthusiasm in research, diligent attitude, depth and breadth of knowledge guided me through my Ph.D. research and shaped what I have achieved. My warmest gratitude, therefore, goes to Dr. Qiong Wang, Dr. Yali Du, Dr. Jiangchao Yao, Yiming Jiang, Licheng Feng, Prof. Weijie Chen, Prof. Weiwei Liu, Dr. Muming Zhao, Dr. Donna Xu, Yueming Lyu, Xiaoqi Jiang and Xu Chen for many illuminating conversations on life and research.

Thankyou also to the University of Technology Sydney for offering me this rare opportunity. My appreciation also goes to those fine minds in our “Eating Group”. Your participation has made this journey for me so memorable. I would like to thank especially Dr. Yan Zhang, Yaxin Shi, Yinghua Yao, Jing Li, Xiaowei Zhou, Xingrui Yu. Thank you all!

Last but not least, I owe the deepest debt of gratitude to my family for their unbounded love and unselfish support. Those miseries my parents suffered for me will be buried in my heart forever.

Abstract

Rank aggregation (RA) refers to the task of recovering the total order over a set of items, given a collection of preferences over the items. The flexible collection of preferences enables successful application of RA in various fields, e.g., image rating and bioinformatics. A basic assumption underlying the vanilla RA is that all preferences are provided by homogeneous users. However, this assumption is rarely satisfied in real applications, due to the complex real situation. Therefore, RA usually suffers from model misspecification, namely the inconsistency between the collected preferences and the homogeneity assumption. Another challenge associated with RA is the scalability issue. In particular, RA usually involves ranking over tens of thousands of items, leading to an exponential volume of preferences for aggregation. Therefore, an inappropriate inference method would limit the application of the proposed model.

This thesis considered RA under model misspecification in the following three scenarios:

- In a crowdsourcing scenario, sufficient annotations from each user are available, which enables exploration of user heterogeneity to account for model misspecification. Therefore, I proposed a reliable CrowdsOURced Plackett-LucE (COUPLE) model, which introduces an uncertainty vector to make a fine-grained categorization of users. Meanwhile, a general Bayesian Moment Matching (OnlineGBMM) was proposed, to ensure an analytic Bayesian update with an almost twice differentiable likelihood function.
- In a general setting, typical model augmentation methods would cause overfitting, because insufficient annotations from each user are available. Inspired by the distributional robust literature, I proposed CoarsenRank, which performs regular RA over a neighborhood of preferences. The resultant inference would enjoy robustness against model misspecification. To this end, I first defined a neighborhood of the rank dataset using relative entropy. Then, I instantiated CoarsenRank with three popular probability ranking models and discussed the optimization strategies.

- RA for mental fatigue monitoring. Common practices for mental fatigue monitoring refer to predicting the reaction time (RT) by aggregating the EEG signal from multiple heterogeneous EEG channels. Let us consider the RT as the item score and view each EEG channel as a user. The mental fatigue monitoring task could be formulated as RA under model misspecification, particularly in a crowdsourcing scenario. To address this problem, a Self-Weight Ordinal REgression (SWORE) model with Brain Dynamics table (BDtable) is proposed. The SWORE model could give a reliable evaluation of brain dynamics preferences from multiple channels, while the BDtable is employed to calibrate the SWORE model by utilizing the proposed online generalized Bayesian moment matching (OGMM) algorithm.

Table of contents

List of figures	xi
List of tables	xiii
1 Introduction	1
1.1 Background	1
1.1.1 Ordinal Data Collection	1
1.1.2 Rank Aggregation Methods	3
1.2 The Challenges of Rank Aggregation	5
1.3 Thesis Contributions	7
1.3.1 Stagewise Learning for Crowdsourced Partial Preferences	7
1.3.2 Robust Rank Aggregation against Model Misspecification	8
1.3.3 Real-time Mental Fatigue Monitoring	9
1.4 Thesis Outline	10
1.5 Publications	11
2 Literature Review	13
2.1 Problem Statement	13
2.2 Probabilistic Ranking Models	14
2.2.1 Permutation-based Ranking Model	15
2.2.2 Score-based Ranking Models	15
2.3 Distance Measures	16
2.3.1 Kullback–Leibler divergence	16
2.3.2 Method of Moments	17
2.4 Efficient Inferences for Ranking Models	17
2.4.1 Online Stochastic Ranking	17
2.4.2 Online Bayesian Ranking	18
2.4.3 Data Augmentation	18

2.5	Robust Rank Aggregation	18
2.5.1	Robust Rank Aggregation using Order Statistics	18
2.5.2	Robust Rank Aggregation via Model Augmentation	19
2.6	Evaluation Measures	20
3	COUPLE: Stagewise Learning for Noisy Preferences	22
3.1	Towards the robust aggregation of noisy preferences	22
3.1.1	Intractability of classical models	22
3.1.2	The CrowdsOURced Plackett-LucE (COUPLE) model	25
3.1.3	Reliability of COUPLE model	29
3.1.4	Optimization Difficulty of COUPLE model	29
3.2	Connection to related models	31
3.3	Online Bayesian moment matching for COUPLE	32
3.3.1	Main routine of Bayesian Moment Matching(BMM)	33
3.3.2	Generalized Bayesian Moment Matching (GBMM)	34
3.3.3	Posterior update	35
3.3.4	The OnlineGBMM algorithm	36
3.4	Experiments	39
3.4.1	Experiment setup	39
3.4.2	Empirical results on large-scale synthetic datasets	40
3.4.3	Empirical results in ordinal peer grading	41
3.4.4	Empirical results in online image-rating	43
3.4.5	Computation Cost	45
3.5	Summary of This Chapter	46
4	CoarsenRank: Rank Aggregation against Model Misspecification	47
4.1	Rank aggregation under model misspecification	47
4.1.1	Problem statement	47
4.1.2	Previous attempts: convolving the ranking model with specific perturbation mechanisms	48
4.1.3	CoarsenRank: rank aggregation over the neighborhood of the ranking data	49
4.2	Coarsened rank aggregation	52
4.2.1	Inferring over the neighborhood brings distributional robustness	52
4.2.2	Coarsened probability ranking model	55
4.2.3	CoarsenRank VS. the Mallows model	58
4.3	Efficient Bayesian inference	58

4.3.1	Data augmentation	59
4.3.2	EM algorithm with closed-formed updating rules	60
4.3.3	Gibbs sampling for Coarsened rank aggregation	61
4.3.4	A data-driven strategy for choosing α	62
4.4	Noisy assumption and rank model assumption	63
4.5	Experimental evaluation	64
4.5.1	Experimental setting	64
4.5.2	Detailed descriptions of datasets	65
4.5.3	Exploring the efficacy of the calibration step	67
4.5.4	Deviance Information Criterion (DIC) for choosing the hyperparameter α	67
4.5.5	The Kendall tau correlation of CoarsenRank in four real applications	68
4.5.6	The computational cost comparison of all methods	69
4.6	Summary of This Chapter	71
5	SWORE: Online Bayesian Ranking for Real-time Mental Fatigue Monitoring	72
5.1	Setup and Problem Statement	74
5.1.1	Mental Fatigue Monitoring	74
5.1.2	Real-time Mental Fatigue Evaluation	75
5.2	Proposed Approach	75
5.2.1	Brain Dynamic Preferences	76
5.2.2	Heterogeneous Brain Dynamic Preferences	77
5.2.3	Self-Weighted Ordinal Regression Model	79
5.3	Efficient Online Updating Strategy	80
5.3.1	Online Reservoir Sampling for BDtable	80
5.3.2	Online Generalized Bayesian Moment Matching for SWORE	81
5.4	Real-time Mental Fatigue Evaluation	86
5.5	Numerical Experiments	87
5.5.1	Experiment Setup	87
5.5.2	Online Mental Fatigue Evaluation	88
5.5.3	Parameter Sensitivity and Model Uncertainty	90
5.6	Summary of This Chapter	94
6	Conclusion and Future Work	95
6.1	Conclusion	95
6.2	Future Work	98

Appendix A	Appendix	100
A.1	Proof for Proposition 1	100
A.2	Detailed derivations for Equation 3.8	101
A.3	Proof for Theorem 1	101
A.4	Proof for Corollary 2	102
A.5	Proof for Theorem 2	103
A.6	Detailed derivations for Equation 5.12	106
References		107

List of figures

1.1	Crowdsourced Partial Preferences Collection. The tasks (subsets) with “✓” are assigned to the worker w . The notation W in the corner denotes that W crowd workers complete the annotation independently.	2
1.2	Challenges of Rank Aggregation	6
1.3	The organization of this thesis.	11
2.1	The overall structure of literature review.	14
3.1	Stagewise annotation process	25
3.2	An intuitive example for the first stage of the robust stagewise learning strategy.	27
3.3	Robust stagewise learning strategy. For brevity, I omit the subscripts i which indicates the stage in which the item has been selected.	28
3.4	Bayesian Moment Matching: (1) define $q(\theta)q(\eta_w)$ in the same form with the prior; (2) match the moments between $q(\eta_w)$ and $P(\eta_w \tilde{X} = \rho^i)$; (3) match the moments between $q(\theta)$ and $P(\theta \tilde{X} = \rho^i)$	33
3.5	Online Generalized Bayesian Moment Matching (OnlineGBMM) for COUPLE: Step (a) estimate $q(\theta)q(\eta_w)$ with generalized Bayesian moment matching; Step (b) replace prior $P(\theta)P(\eta_w)$ with approximate posterior $q(\theta)q(\eta_w)$	37
3.6	To verify the <i>reliability</i> of COUPLE preliminarily, the accuracy (%) with varying percentage of samples on large-scale synthetic datasets is provided.	41
3.7	To verify the <i>complexity analysis</i> in Table 3.1, I collected the time cost of the four models when I conducted the experiment on the <i>BabyFace</i> dataset.	45
4.1	The logic stream of CoarsenRank. Condition 1: perform rank aggregation over a neighborhood of the collected preferences (See Equation 4.7 and 4.8). Condition 2: adopt relative entropy as the divergence measure and assign an exponential prior for the size of the neighborhood (See Corollaries 1 and 2).	49

4.2	The performance comparison of CoarsenRank and PL-EM algorithm under the case of with or without calibration step. “W/” denotes “with” while “W/O” denotes “without”	66
4.3	(a)-(d) The diagnostic plot of DIC VS. α on four datasets, respectively. The α used in the experiment are marked as “*” in each figure.	68
4.4	Performance improvement of various methods over PL-EM on four datasets, following $\frac{S_*-S_0}{S_0}$. S_0 is the correlation of PL-EM in Kendall tau correlation.	69
4.5	The computation cost of all baselines on four datasets, respectively.	70
5.1	Real-Time Mental Fatigue Evaluation. First, I maintain a fixed size Brain Dynamics table (BDtable) as the reference, which consists of the representative EEG signals and the corresponding reaction times (RTs). For a new collected EEG signal x_t , the Self-Weight Ordinal REgression (SWORE) model could give a coarse estimation of the reaction time RT_t by interpolating it among the bunch of maintained RTs using the brain dynamics preferences.	73
5.2	Event-related lane-departure driving paradigm	74
5.3	Gradient flattening w.r.t. sigmoid function. The dash line represents the original sigmoid function.	77
5.4	OGMM with Data Augmentation. Note that: (1) sample the corrupted EEG signal $\Delta\tilde{x}$ from the predefined corrupting distribution $P(\Delta\tilde{x} \Delta x)$; (2) define $q(w)q(\pi)$ in the same form as the prior (product of a Normal with Betas); (3) estimate $q(w)q(\pi)$ with generalized Bayesian moment matching; (4) replace prior $P(w)P(\pi)$ with approximate posterior $q(w)q(\pi)$	85
5.5	The PDF of prediction accuracy.	88
5.6	Real-Time Mental Fatigue Evaluation.	89
5.7	The negative Log-likelihood of brain dynamic preferences on training and test dataset w.r.t different level of data augmentation size. Note that “Aug-n” denotes the data augmentation size T is set to n	90
5.8	Box plot of the prediction accuracy on the test dataset. The symbol “+” denotes the outliers.	94

List of tables

1.1	The applications of rank aggregation	3
3.1	computation cost of COUPLE and other models. Assume t_1 is the cost of extracting a pairwise preference from a k -ary preference and t_2 is the cost of completing an update in Algorithm 1.	38
3.2	To verify the <i>reliability</i> of COUPLE in ordinal peer grading, the accuracy (%) on two ordinal peer grading datasets, i.e., PO and FR datasets, is provided. Accuracy is represented by a mean with standard deviation. As PeerGrader is SGD-based algorithms, I iterated PeerGrader until convergence and only measured the accuracy only once.	42
3.3	Six workers with lowest work quality identified by COUPLE, CrowdBT, or PeerGrader on PO and FR datasets, respectively.	42
3.4	To verify the efficacy of COUPLE and CrowdBT on <i>noisy worker detection</i> , the accuracy (%) on cleaned PO and FR datasets is provided. The accuracy is represented by a mean with standard deviation. “PO(COUPLE)” denotes the cleaned PO dataset processed by COUPLE. This definition applies to other similar notations.	43
3.5	To verify the <i>reliability</i> of COUPLE in the real-world challenges, the accuracy (%) on the <i>BabyFace</i> dataset is provided. To verify the efficacy of COUPLE and other methods on <i>noisy worker detection</i> , the accuracy (%) on the Cleaned <i>BabyFace</i> dataset is provided. The accuracy is represented by the mean with the standard deviation. As PeerGrader is SGD-based algorithms, I iterated PeerGrader until convergence and collect the accuracy only once.	44
3.6	Six workers with lowest work quality identified by COUPLE, CrowdBT, or PeerGrader on <i>BabyFace</i> datasets.	44

4.1	Comparison between various ranking models in terms of their noisy assumption and ranking model assumption. “—” denotes the assumption does not apply to a particular ranking model.	63
4.2	The statistics of four real ranking datasets	66
4.3	Experiment results of various rank aggregation methods on four real datasets. Best results are marked in bold.	69
4.4	The computation cost (s) of all baselines on four datasets, respectively. Best results are marked in bold.	70
5.1	Test accuracy (in %, the larger the better) w.r.t hyperparameter (μ, Σ) with hyperparameter (α, β) fixed to $(5, 5)$, dropout rate $\theta = 0.5$, data augmentation number $T = 1$. The best parameter settings are marked in gray. Some parameter settings do not consistently perform very well and may fail on some participants (marked in bold).	92
5.2	Test accuracy (in %, the larger the better) w.r.t hyperparameter (α, β) with hyperparameter (μ, Σ) fixed to $(10^{-2}, 10^{-4})$, dropout rate $\theta = 0.5$, data augmentation number $T = 1$. The best parameter settings are marked in gray. Some parameter settings do not consistently perform very well and may fail on some participants (marked in bold).	93