

2015

Learning mixed membership models with a separable latent structure: theory, provably efficient algorithms, and applications

<https://hdl.handle.net/2144/13671>

"Downloaded from OpenBU. Boston University's institutional repository."

BOSTON UNIVERSITY
COLLEGE OF ENGINEERING

Dissertation

**LEARNING MIXED MEMBERSHIP MODELS WITH A
SEPARABLE LATENT STRUCTURE: THEORY,
PROVABLY EFFICIENT ALGORITHMS, AND
APPLICATIONS**

by

WEICONG DING

B.Sc., Tsinghua University, 2010

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2015

© 2015 by
WEICONG DING
All rights reserved

Approved by

First Reader

Prakash Ishwar, Ph.D.
Associate Professor of Electrical and Computer Engineering
Associate Professor of Systems Engineering

Second Reader

Venkatesh Saligrama, Ph.D.
Professor of Electrical and Computer Engineering
Professor of Systems Engineering

Third Reader

Luis E. Carvalho, Ph.D.
Assistant Professor of Mathematics and Statistics

Fourth Reader

Vivek Goyal, Ph.D.
Associate Professor of Electrical and Computer Engineering

Acknowledgments

I have been fortunate to work with my research advisor, Prof. Prakash Ishwar and my co-advisor Prof. Venkatesh Saligrama whose guidance was invaluable to my research achievements. I would like to express my deepest gratitude to them for shaping and refining my approach to research from boldly exploring uncharted areas, identifying and abstracting the most significant research questions, developing key insights and solution strategies, and archiving and presenting findings and conclusions. Their approach to research has greatly enriched and expanded my perspective of fundamental and applied research in data science and engineering. I would like to especially thank Prof. Ishwar for his patience and the enormous time and effort that he dedicated to every stage of my research. His enthusiasm, encouragement, and meticulous research attitude has inspired and enlightened me during throughout my doctoral studies. My thanks also go to Prof. Saligrama for his guidance and substantial knowledge of the literature. I am especially grateful for having the opportunity to work with both of my co-advisors and imbibe their diverse research styles. I wish to thank other members of my committee, Prof. Luis Carvalho and Prof. Vivek Goyal, for their valuable discussions and suggestions. I would also like to thank Dr. Mohammad H. Rohban and Prof. W. Clem Karl for insightful discussions on various projects we worked on together.

My life in Boston University was throughably enjoyable thanks to my great lab fellows: Cem Aksoylar, Tolga Bolukbasi, Feng Nan, Dr. Joe Wang, Jonathan Wu, and Dr. Ziming Zhang. I received generous help and encouragement from my best friends in Boston who have made me feel at home. I would like to sincerely thank Yuting Chen, Hao Chen, Dr. Wuyang Dai, Wenbo He, Wei Si, Dr. Jing Wang, Dr. Bowen Zhang, Yuting Zhang, and Qi Zhao.

Lastly, I would like to thank my family for all their love and support. Most of

all, I want to thank my dear wife, Kai Shen. Kai has been my pillar of strength and support throughout all the hard times. She has always believed in me and cheered me up no matter where she is across the world.

I would also like to acknowledge all the funding agencies that supported my research. The material in this dissertation is based upon work supported by the US Air Force Office of Scientific Research and the US National Science Foundation. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the agencies.

**LEARNING MIXED MEMBERSHIP MODELS WITH A
SEPARABLE LATENT STRUCTURE: THEORY,
PROVABLY EFFICIENT ALGORITHMS, AND
APPLICATIONS**

WEICONG DING

Boston University, College of Engineering, 2015

Major Professors: Prakash Ishwar, Ph.D.

Associate Professor of Electrical and Computer
Engineering

Associate Professor of Systems Engineering

Venkatesh Saligrama, Ph.D.

Professor of Electrical and Computer Engineering
Professor of Systems Engineering

ABSTRACT

In a wide spectrum of problems in science and engineering that includes hyperspectral imaging, gene expression analysis, and machine learning tasks such as topic modeling, the observed data is high-dimensional and can be modeled as arising from a data-specific probabilistic mixture of a small collection of latent factors. Being able to successfully learn the latent factors from the observed data is important for efficient data representation, inference, and prediction. Popular approaches such as variational Bayesian and MCMC methods exhibit good empirical performance on some real-world datasets, but make heavy use of approximations and heuristics for dealing with the highly non-convex and computationally intractable optimization objectives that

accompany them. As a consequence, consistency or efficiency guarantees for these algorithms are rather weak.

This thesis develops a suite of algorithms with provable polynomial statistical and computational efficiency guarantees for learning a wide class of high-dimensional Mixed Membership Latent Variable Models (MMLVMs). Our approach is based on a natural separability property of the shared latent factors that is known to be either exactly or approximately satisfied by the estimates produced by variational Bayesian and MCMC methods. Latent factors are called separable when each factor contains a novel part that is predominantly unique to that factor. For a broad class of problems, we establish that separability is not only an algorithmically convenient structural condition, but is in fact an inevitable consequence of having a relatively small number of latent factors in a high-dimensional observation space. The key insight underlying our algorithms is the identification of novel parts of each latent factor as extreme points of certain convex polytopes in a suitable representation space. We show that this can be done efficiently through appropriately defined random projections in the representation space. We establish statistical and computational efficiency bounds that are both polynomial in all the model parameters. Furthermore, the proposed random-projections-based algorithm turns out to be naturally amenable to a low-communication-cost distributed implementation which is attractive for modern web-scale distributed data mining applications.

We explore in detail two distinct classes of MMLVMs in this thesis: learning topic models for text documents based on their empirical word frequencies and learning mixed membership ranking models based on pairwise comparison data. For each problem, we demonstrate that separability is inevitable when the data dimension scales up and then establish consistency and efficiency guarantees for identifying all novel parts and estimating the latent factors. As a by-product of this analysis, we

obtain the first asymptotic consistency and polynomial sample and computational complexity results for learning permutation-mixture and Mallows-mixture models for rankings based on pairwise comparison data. We demonstrate empirically that the performance of our approach is competitive with the current state-of-the-art on a number of real-world datasets.

Contents

1	Introduction	1
1.1	Mixed Membership Latent Variable Models	3
1.1.1	General Modeling Framework for MMLVMs	3
1.1.2	Historical Development of MMLVMs	5
1.1.3	Approximation Approaches	6
1.2	Separability Property	7
1.3	Other Related Works	9
2	Topic Discovery through Random Projections	11
2.1	Generative Model and Our Main Results	11
2.2	Related Works	14
2.3	Topic Separability, Necessary and Sufficient Conditions, and the Geometric Intuitions	18
2.3.1	Key Structural Property: Topic Separability	19
2.3.2	Conditions on the Topic Mixing Weights	20
2.3.3	Geometric Implications and Random Projections Based Algorithm	23
2.4	Topic geometry with a finite sample size: word co-occurrence matrix representation, solid angle, and random projection based approach	26
2.4.1	Normalized Word Co-occurrence Matrix Representation	26
2.4.2	Solid Angle Extreme Point Robustness Measure	29
2.4.3	Efficient Estimation of Solid Angles using Random Projections	31

2.5	Algorithm and Analysis	34
2.6	Distributed Topic Discovery	37
2.6.1	Distributed Setting	38
2.6.2	Estimating Solid Angles from Distributed Servers	39
2.6.3	Analysis	41
2.6.4	Related Works in Distributed Topic Modeling	42
2.7	Empirical Results	42
2.7.1	Semi-Synthetic Dataset	43
2.7.2	Real World Text Corpus	48
3	Mixed Membership Ranking Models for Pairwise Comparisons	51
3.1	Motivating Example and Generative Framework	51
3.2	Related Works	55
3.3	Topical Ranking Model	57
3.3.1	Reduction to Topic Models	59
3.3.2	Separability Property	60
3.3.3	The Geometric Approach and Analysis	61
3.4	Mixed Membership Mallows Models	64
3.4.1	Mallows Distribution and Generative Model for M4	65
3.4.2	Reduction to Topic Model via Ranking Matrix	67
3.4.3	Overview of Algorithm, Key Insights, and Theoretical Results	69
3.5	Empirical Results	74
3.5.1	Semi-synthetic Simulation	75
3.5.2	Predicting pairwise comparisons	77
3.5.3	Predicting star ratings	80
4	Most Large MMLVMs are Separable	82
4.1	Separability in Topic Modeling	83

4.1.1	Discussion and Implications of Lemma 12	84
4.2	Separability in Topic Model for Rankings	86
4.2.1	Discussion and Implications of Lemma 13	87
4.3	Separability in Mixed Membership Mallows Model	88
4.4	Discussion and Implication of Lemma 14	89
5	Concluding Remarks and Outlook	91
5.1	Future Directions	92
A	Proofs of all Technical Results	93
A.1	Proof of Lemma 1	93
A.2	Proof of Lemma 2	94
A.3	Proof of Proposition 1 and Proposition 2	94
A.4	Proof of Lemma 3	96
A.5	Proof of Lemma 4	96
A.6	Proof of Lemma 5	97
A.7	Proof of Lemma 6	100
A.8	Proof of Lemma 7	100
A.9	Proof of Lemma 8	101
A.10	Proof of Theorem 2	101
A.11	Proof of Theorem 3	102
A.12	Proof of Theorem 4	107
A.13	Theorem 3 with Spherical Gaussian Directions	110
A.14	Proof of Theorem 5	113
A.15	Proof of Lemma 9	113
A.16	Proof of Lemma 10 and Theorem 6	114
A.17	Proof of Theorem 7	114
A.18	Proof of Proposition 4	115

A.19 Proof of Lemma 11	118
A.20 Proof of Theorem 8	118
A.21 Proof of Theorem 9	118
A.21.1 Consistency of Algorithm 6 in M4	118
A.21.2 Consistency of Algorithm 7 in M4	121
A.21.3 Consistency of the post-processing Algorithm 10 in M4	123
A.21.4 Overall sample complexity of the Algorithm 9	124
A.22 Proof of Lemma 12	125
A.23 Proof of Lemma 13	127
A.24 Proof of Lemma 14	128
A.25 Separability for Measures and Irreducibility	129
References	132
Curriculum Vitae	140

List of Tables

2.1	Normalized held-out log probability of RP, RecoverL2, and Gibbs Sampling on NYT test data. The Mean \pm STD's are calculated from 5 different random training-testing splits.	49
2.2	Examples of topics estimated by RP on NYT	50
3.1	Comparison of the proposed Mixed Membership Ranking Models to closely related works in mixed membership/mixture ranking models.	57
3.2	Testing RMSE on the Movielens dataset with $Q = 100$ most rated movies.	81
4.1	Validation of approximate separability of the topic matrices in Topic Modeling with Dirichlet prior using parameters from benchmark text corpus datasets using Monte Carlo simulation.	86
4.2	Validation of approximate separability of the ranking matrices in Mixed Membership Mallows Models using Monte Carlo simulations.	90

List of Figures

2.1	Generative process and the plate representation of standard topic models.	12
2.2	An example of separable topic matrix and the extreme point geometry.	19
2.3	Topic separability alone does not guarantee uniqueness.	20
2.4	Relationships between conditions on the normalized second- moment $\bar{\mathbf{R}}$	23
2.5	The extreme point geometry in the word co-occurrence matrix representation.	29
2.6	A distributed implementation of our proposed approach.	39
2.7	Reconstruction Error in ℓ_1 distance on semi-synthetic text corpus . .	45
2.8	Computation cost vs reconstruction error in semi-synthetic text corpus	46
2.9	Solid angle estimation on Semi-Syn NYT dataset and Semi-Syn+Sep NYT dataset.	48
3.1	Key idea of mixed membership ranking models.	52
3.2	Generative process and plate representation for Topical Ranking Model.	58
3.3	An example separable ranking matrix in Topic Ranking Model and the corresponding geometric structure.	60
3.4	Generative process and plate representation for M4.	66
3.5	An example of approximate separable ranking matrix in M4 and the corresponding geometric structure.	70
3.6	Reconstruction error in Kendall's tau distance on semi-synthetic comparison data.	76
3.7	Predictive likelihood for new comparisons on real-world data.	79

3.8	Predictive likelihood for new users on real-world data.	79
-----	---	----

List of Abbreviations

BPMF	Bayesian Probabilistic Matrix Factorization
BTL	Bradley-Terry-Luce Model
CTM	Correlated Topic Model
DDP	Data Dependent Projection
DTM	Dynamic Topic Model
HDP	Hierarchical Dirichlet Process
LDA	Latent Dirichlet Allocation
M4	Mixed Membership Mallows Model
MCMC	Monte Carlo Markov Chain
MMLVM	Mixed Membership Latent Variable Model
MMSB	Mixed Membership Stochastic Blockmodels
NMF	Nonnegative Matrix Factorization
PL	Prackett-Luce Model
pLSI	probabilistic Latent Semantic Index
PMF	Probabilistic Matrix Factorization
RP	Random Projection
VB	Variational Bayesian

Chapter 1

Introduction

Large amounts of data are now being generated at a web scale such as text messages on Twitter and user ratings on Yelp. To understand and make reliable prediction for these real world data, one has to deal with the high-dimensionality and variability in them. The key to overcome these challenges is to design models that are sufficiently rich to accommodate the key data characteristics yet are tractable for learning efficiently and reliably from the available observations.

This thesis focuses on the family of **Mixed Membership Latent Variable Models** (MMLVMs) that have been widely used in many important machine learning tasks including text analysis [Blei, 2012], preference prediction [Ding et al., 2015b, Gormley et al., 2009], community detection [Airoldi et al., 2008], etc. On a high level, the MMLVM views each observation as arising from a probabilistic mixture of a few latent factors that are shared among the dataset [Airoldi et al., 2014]. The primary learning problem in MMLVM is to estimate the shared latent factors from the observation. The standard MAP or ML estimator for latent factors in general is non-convex and typically NP-hard [e.g., Arora et al., 2012, Sontag and Roy, 2011]. The popular estimation and inference approaches include approximations like variational Bayesian (VB), Markov Chain Monte Carlo methods (MCMC) like Gibbs Sampling, and EM-type alternating optimization algorithms [Airoldi et al., 2014, Cichocki et al., 2009, Wainwright and Jordan, 2008]. These approaches have produced state-of-the-art empirical performances on many real-world machine learning tasks. Guarantees of

asymptotic consistency or efficiency for these approaches, however, are either weak or non-existent. This makes it difficult to evaluate the *model fidelity*: failure to produce satisfactory results could be due to the use of approximations and heuristics or due to model mis-specification which is more fundamental. Furthermore, these sub-optimal approaches are computationally intensive for large datasets [Arora et al., 2013, Ding et al., 2014b].

This thesis develops a novel approach for learning shared latent factors with provable statistical and computational efficiency guarantees. To overcome the hardness of learning problem of MMLVMs in its full generality, we propose to consider the set of MMLVMs with a natural separable property on the shared latent factors, wherein every latent factors contains a **novel** part, i.e., a part that is predominantly unique to that factor and is approximately absent from the other factors. The topic separability property is empirically motivated by the fact that for many real-world datasets, the empirical topic estimates produced by popular Variational Bayes and Gibbs Sampling approaches are approximately separable [Arora et al., 2013, Ding et al., 2014b]. Moreover, we show that the separability property turns out to be an inevitable outcome of having a relatively small number of latent factors in a high dimensional observation space for a wide range of MMLVMs [e.g., Ding et al., 2015a,c]. Therefore, separability is a natural approximation for most high-dimensional MMLVMs investigated in this thesis.

Our approach is based on the following geometric insight: if we associate each part of the observation with a co-occurrence vector in a suitable representation space, the novel parts unique to each factor will be extreme points of the convex polytope formed by co-occurrence vectors of all parts. We leverage this geometric insight and develop an approach based on robust extreme point detection using random projections. We establish both sample and computational complexity bounds that

are *polynomial* in all model parameters. Our approach is especially amenable to a distributed implementation suitable for large databased stored in network of serves [Ding et al., 2014b]. It can achieve the same statistical efficiency as the centralized version with an insignificant communication cost between the distributed serves.

We first apply our approach to the topic modeling problems [Ding et al., 2013a,b, 2014b]. We demonstrate that our approaches are empirically competitive with the popular approximation based methods on real-world text corpora. We then apply our approach to the problem of estimating and predicting preference behavior from pairwise comparisons [e.g., Ding et al., 2014a, 2015b,c]. We propose novel mixed membership ranking models that can capture a heterogeneous and inconsistent user-population in a natural way. As a by-product, we obtain the *first provable consistency and efficiency* results for permutation-mixture model [Farias et al., 2009] and Mallows-mixture model [Lu and Boutilier, 2014] which are special cases of our proposed models.

1.1 Mixed Membership Latent Variable Models

MMLVMs are widely used in modeling high-dimensional data such as document collections as combinations of different semantic topics, interactions in social networks driven by multiple user communities, etc. [Airoldi et al., 2014] In this section, we overview the general frameworks, history development, application, and the widely used approximation methods for MMLVMs.

1.1.1 General Modeling Framework for MMLVMs

We first sketch the general modeling framework of MMLVMs. To fix idea, we consider a collection of M observations denoted by $\mathbf{x}^1, \dots, \mathbf{x}^M$. On the population level, we posit K latent factors β^1, \dots, β^K that are shared among the dataset and they each defines a distinct probability on the observation. Each individual observation \mathbf{x}^m is associated with a data-specific mixing weights vector $\theta^m = [\theta_{m,1}, \dots, \theta_{m,K}]^\top \in \mathbb{R}^K$,

which is a realization from some prior distribution $\Pr(\theta; \alpha)$ on the K -dimensional simplex. α denotes the hyper-parameters of the prior distribution. The mixing weights for M observations are assumed to be sampled from $\Pr(\theta; \alpha)$ in an i.i.d fashion.¹ Conditioned on θ^m , the observation \mathbf{x}^m arises as a mixture of the shared K latent variables, i.e.,

$$p(\mathbf{x}^m | \theta^m) = f(\mathbf{x}^m; \sum_{k=1}^K \theta_{m,k} \beta^k) \quad (1.1)$$

where $f(\mathbf{x}^m, \sum_{k=1}^K \theta_{m,k} \beta^k)$ is the observation model. For example in the popular Latent Dirichlet Allocation (LDA) model [Blei et al., 2003], each latent factor β^k is a distribution over a vocabulary of size W . These latent factors are referred to as “topics” in this case. The prior distribution $\Pr(\theta; \alpha)$ is Dirichlet distribution. The observation model f is multinomial distribution and \mathbf{x}^m is empirical word counts by sampling N words i.i.d from the document-specific distribution $\sum_{k=1}^K \theta_{m,k} \beta^k$.

Smoothed MMLVMs: In a fully Bayesian setting, the latent factors β_k ’s are further modeled as being sampled from some prior on the spaces of latent factors. This smoothed setting makes it easier for fully Bayesian approaches such as MCMC. In literature, smoothed MMLVMs are often referred to the same terminology as MMLVMs. For instance in the smoothed LDA model, the topics $\beta^1, \dots, \beta^K \in \mathbb{R}^W$ are modeled as i.i.d samples from a symmetric Dirichlet prior on W dimensional simplex with a hyper-parameter β_0 [Blei et al., 2003].

Overall Goal: Our algorithmic goal is to discover the latent factors β^1, \dots, β^K from the observations $\mathbf{x}^1, \dots, \mathbf{x}^M$. Once the latent factors are estimated, we then adopt the standard procedure to inference θ^m and to make prediction for new observations [e.g., Blei, 2012, Lee and Seung, 1999, Wallach et al., 2009]. We note that our theoretical

¹The mixing weights θ^m ’s can also be modeled as following a dynamic structure such as in the Dynamic Topic Model [Blei and Lafferty, 2006]. For simplicity, we only consider the independent θ^m ’s in this thesis but our approach can be extend to handle dynamic models.

guarantees apply only to the latent factor estimation.

To be specific, the *asymptotic consistency* refers to the property that estimated latent factors converge to the ground truth in some metric as the number of observations $M \rightarrow \infty$. The *polynomial sample complexity* refers to a sufficient bound on the number of samples M required to achieve a reconstruction error of ϵ with a tolerance failure probability δ is at most a polynomial function of W , $1/\epsilon$, $\log(1/\delta)$ and other model parameters.

1.1.2 Historical Development of MMLVMs

The mixed membership modeling perspective can be traced back to 1970s under the name Grade of Membership (GoM) model that was designed for medical diagnosis observation [Woodbury et al., 1978]. However, it was not until early 2000s when MMLVMs achieved widespread success with the use of Bayesian approaches. There are several independent trends of researches that lead to the modern form of MMLVMs including GoM motivated by medical record data [Erosheva et al., 2007], admixture model motivated by genetic data [Pritchard et al., 2000], and Latent Dirichlet Allocation type of model in computer science [Blei et al., 2003].

Text Corpus and Topic Modeling: The seminal work of LDA [Blei et al., 2003] initiated extensive study of developing MMLVMs in modeling text-based observations including news articles, scientific papers, social media, and reviews and comments from web applications [e.g., Blei, 2012, Tang et al., 2014, and the reference therein]. In this context, each document is viewed as a bag of words and is modeled as a probabilistic mixture of a few shared latent semantic “topics”. Due to its popularity, MMLVMs are sometimes referred to as “topic modeling” [Airoldi et al., 2014].

Health Science Application: The mixed membership modeling perspective was first developed in the context of medical record analysis Woodbury et al. [1978] to discover sub-patterns of illness in particular diseases. MMLVMs have also been applied to

medical survey to discover distinct pathways and patterns of disabilities [Erosheva et al., 2007, Manrique-Vallier, 2014].

Ranking and Preference Data: MMLVMs have also been developed for rank data in political science to analyze latent factors in election data [Gormley et al., 2009]. Recently in [Ding et al., 2014a, 2015b,c, Kim et al., 2014], MMLVMs have been developed to learn the latent influencing factors for user preferences in a heterogeneous and inconsistent population. They are important in modern web-scale applications such as personalized recommendation, e-commerce, and information retrieval, etc.

Network Interactions and Overlapping Communities: Another popular application of MMLVMs is to estimate overlapped latent communities from network interactions Airoldi et al. [2008]. Chief among them is the Mixed Membership Stochastic Blockmodels (MMSB). MMSB has been extended in various directions to incorporate different types of network observations [e.g., Azizi et al., 2014, Gopalan and Blei, 2013, Huang et al., 2013].

1.1.3 Approximation Approaches

Exact parameter estimation and inference in MMLVMs are intractable in general [Arora et al., 2012, Sontag and Roy, 2011]. In practice, approximation techniques are used and falls into two major categories: sampling based approaches and structural approximation or optimization based approaches. Sampling based approaches are typically based on Monte Carlo Markov Chain (MCMC) and generate a sequence of approximately independent samples whose distributions converges to the true posterior [e.g., Airoldi et al., 2014, Griffiths and Steyvers, 2004, Wallach et al., 2009]. On the other hand, the most common structural approximation or optimization based approaches is the variational Bayesian methods (VB) [e.g., Awasthi and Risteski, 2015, Blei et al., 2003, Wainwright and Jordan, 2008]. In VB, the posterior distribution is approximated by a variational surrogate distribution which has a simpler form and

we minimize the distance between variational distribution and the true posterior in KL- divergence.

Variational Bayes and MCMC based algorithms in practice have similar empirical performances. MCMC approaches can be applied straightforwardly to a general family of MMLVMs but it requires many samples to approximate the true posterior. In contrast, VB method should be designed for each specific problem. However, the computation is typically relatively simple and in some cases, for instance in LDA, the update in each iteration has a closed form. Recently, [Awasthi and Risteski, 2015] show that VB approach for LDA model can be consistent if the topics are separable and the VB procedure is properly initialized.

1.2 Separability Property

We formally overview the key structural property, separability. For simplicity, we consider the observation in discrete space of size W and the corresponding latent factors β^k 's are $W \times 1$ dimensional pmfs. We then define a **latent factor matrix** $\beta = [\beta^1, \dots, \beta^K]$ which is a $W \times K$ dimensional non-negative matrix.

The separability is a structural property of β . Formally,

Definition 1. (λ -approximate separability) *A $W \times K$ nonnegative matrix β is λ - approximately separable for some constant $\lambda \geq 0$, if $\forall k = 1, \dots, K$, there exist at least one row i such that $\beta_{i,k} > 0$ and $\beta_{i,l} \leq \lambda\beta_{i,k}$, $\forall l \neq k$.*

The λ -approximate separability therefore requires the existence of rows of β that concentrate predominantly in one column and have relatively negligible occurrences in the other columns. We call these rows of β as λ -approximately novel rows. The smaller the value of λ , the sharper the concentration within a single latent factor and higher the novelty of the row and the separability of the latent factors. In the limiting case that $\lambda = 0$, we will say that latent factor matrix β is *exactly separable* [Arora et al., 2013, Ding et al., 2013b, 2014b].

Separability in Other Applications

Separability has been discovered and exploited in literature from different fields. The earliest concept we can identify in literature is the *Pure Pixel Index* condition in the Hyperspectral Image (HSI) unmixing problem [Boardman, 1993]. A number of algorithms have been proposed based on similar geometric property with ours [Bioucas-Dias et al., 2013, Gillis and Vavasis, 2014, and the references therein]. However, the guarantees exist only when there is no additive “noise”. Separability has also been studied in the context of Nonnegative Matrix Factorization (NMF). [Donoho and Stodden, 2004] first showed that exact separability can guarantee the uniqueness of NMF along with additional conditions. A subsequent work then develops algorithms for NMF by exploiting the separability in different aspects [Gillis and Vavasis, 2014, Recht et al., 2012].

Very recently, separability has been exploited in the context of topic discovery and other MMLVMs [e.g., Arora et al., 2012, 2013, Awasthi and Risteski, 2015, Bansal et al., 2014, Ding et al., 2013b, 2014b, 2015b,c, Kumar et al., 2013] with consistency or efficiency guarantees. They are closely related and will be discussed in later sections.

Separability is Inevitable in High Dimensional MMLVMs

Separability as in Definition 1 is a structural property on the latent factor matrix β . In this thesis we show it is satisfied with high probability if β is sampled from prior distributions that are typically used in the smoothed MMLVM setting. Concretely, we investigated the separability property for MMLVMs considered in this thesis:

- In the topic models to be discussed in chapter 2, β is a $W \times K$ column-stochastic matrix sampled from the following prior: the K columns vectors β^1, \dots, β^K are iid samples from a symmetric Dirichlet prior $\text{Dir}(\beta_0)$ for some $\beta_0 > 0$. Informally, we show for any small constant $\lambda \in (0, 1)$ and $\beta_0 \in (0, 1)$, the probability that β is λ -approximate separable is at least $1 - K \exp(-pW)$

where p is determined by K, λ and β_0 . The size of vocabulary W is very large in benchmark datasets.

- In the topic ranking model to be discussed in Section 3.3, β is a $W \times K$ binary non-negative matrix and $W = Q(Q - 1)$ is the number of ordered pairs of Q items. Each column of β correspond to a total ranking. Informally, we show if the K total rankings are i.i.d uniform samples from the set of all permutations, then, β is exact separable with probability at least $1 - K \exp(-2^{-K}Q)$.
- In the Mixed Membership Mallows Model to be discussed in Section 3.4, β is a $W \times K$ non-negative matrix and $W = Q(Q - 1)$ is the number of ordered pairs of Q items. Each column of β is determined by a Mallows component [Mallows, 1957]. We show if the reference rankings of K Mallows components are iid uniform samples from the set of all permutations, then, β is approximately separable with high probability.

We will discuss these results in detail in Chapter 4.

Separability for measures

We have defined the separability property for MMLVMs whose latent factors are a collection of distributions on a finite space (e.g., a fixed vocabulary of size W in topic models). It turns out that we can extend the separability property to the general setting in which we consider a collection of measures based on the same principle. Interestingly, while we make this generalization, we identify that the separability property is equivalent to the so-called **irreducibility property** that has been studied in the context of mixture models to establish their identifiability guarantees. [Blanchard and Scott, 2014, Scott, 2015]. We will discuss this connection in details in the appendix.

1.3 Other Related Works

Nonnegative Matrix Factorization: Our algorithm is closely related to the Nonnegative Matrix Factorization (NMF) [Lee and Seung, 1999]. The goal of NMF is to decompose a matrix as a product of two low-rank non-negative matrices and it is important in a number of application [Bioucas-Dias et al., 2013, Cichocki et al., 2009, Gillis and Vavasis, 2014]. The general NMF problem has been shown to be NP-hard [Vavasis, 2009]. The separability has been identified in the context of NMF as a necessary condition that can guarantee the uniqueness [Donoho and Stodden, 2004]. A subsequent of work developed efficient algorithms for NMF with separability properties [Cichocki et al., 2009, Gillis and Vavasis, 2014].

Bayesian Nonparametric Models: Bayesian nonparametric model such as Hierarchical Dirichlet Process [Teh et al., 2006] are important variations of MMLVMs. They assume that the number of latent factor are not fixed and can be potentially infinite. These methods provide alternative for adaptive model selection in MMLVMs. In this thesis, we always assume the true number of latent topics is known.

Organization:

The rest of this thesis is organized as follows. In Chapter 2, we discuss in detail the topic modeling problem for text document. We illustrate the key geometric properties of our approach. In Chapter 3, we discuss two novel mixed membership models for ranking preferences in pairwise comparisons. We show in Chapter 4 that the proposed separability property is inevitable in the models in Chapter 2 and 3. We conclude in Chapter 5 by showing that our method can be applied to a wide range of other MMLVMs.

Chapter 2

Topic Discovery through Random Projections

Topic modeling refers to a family of generative models and associated algorithms for discovering the topical structure common to a large corpus of documents. They are important for organizing, searching, and making sense of a large text corpus including news reports, scientific publications, web pages, and social media streaming [Blei, 2012]. They have also been applied to observations such as images, network logs, etc. [e.g., Carman et al., 2010, Li and Perona, 2005, Tang et al., 2014]

In this chapter, we use topic modeling as an example to develop the key geometric intuitions and sketch our approach that can be later generalized to other MMLVMs. In order to highlight the key ideas, in this chapter, we assume the latent factors to be *exact* separable. We will discuss the approximate separability in the next chapter.

2.1 Generative Model and Our Main Results

We consider a collection of M documents, and each document is composed of N words drawn from a fixed vocabulary of size W .¹ The documents are indexed by $m = 1, \dots, M$ and the distinct words in the vocabulary are labeled by $w = 1, \dots, W$. A “topic” is a $W \times 1$ distribution over the vocabulary. A topic model posits the existence of $K < \min\{W, M\}$ “topics”. They are collectively represented by the key columns β^1, \dots, β^K of a $W \times K$ “topic matrix” β . To generate each document m ,

¹The methods discussed in this thesis can handle $N \geq 2$. It remains an open question whether latent topics can be efficiently estimated when $N = 1$.

first, draw a $K \times 1$ distribution vector $\boldsymbol{\theta}^m$ from some prior $\Pr(\boldsymbol{\alpha})$ ² on K dimensional simplex as the topic mixing weights of document m ; then, draw N words as iid samples from a $W \times 1$ document distribution over the vocabulary $\mathbf{A}^m = \sum_{k=1}^K \boldsymbol{\beta}^k \theta_{k,m}$ which is a convex combination (probabilistic mixture) of the latent topics. A document m can then be represented as an empirical word-counts vector \mathbf{X}^m where $X_{w,m}$ is the number of times word w appears in document m [Arora et al., 2013, Blei, 2012, Blei et al., 2003, Ding et al., 2014b]. We represent the entire corpus by a $W \times M$ matrix $\mathbf{X} = [\mathbf{X}^1, \dots, \mathbf{X}^M]$.³ An graphical representation of topic models is depicted in Figure 2.1. In benchmark datasets such as a news article collection from NY Times [Bache and Lichman, 2013] to be used later in experiments, we get $W = 14,943$, $M = 300,000$, and on average $N = 298$. We observe that $N \ll W$, \mathbf{X} is very sparse, and M is very large. The number of topic in literature are typically set to be $K \approx 100$.

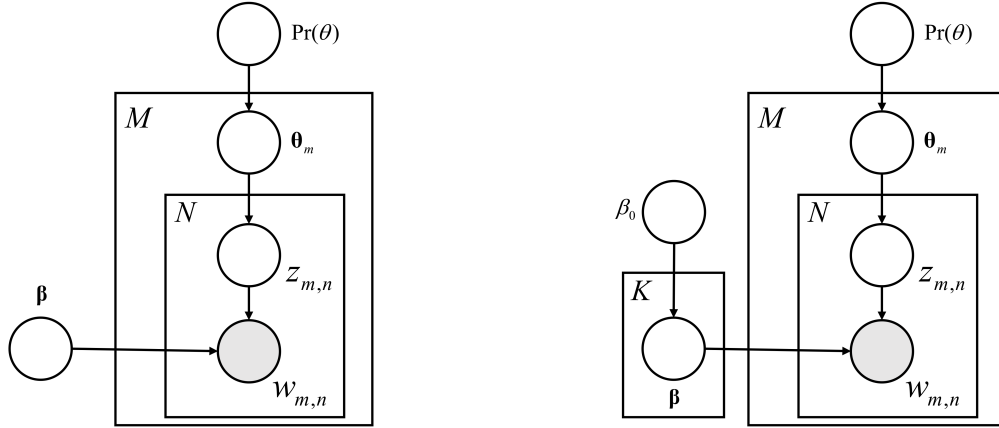
Learning Objective: We focus on the learning problem in topic model. Given empirical observation \mathbf{X} , our goal is to learn the latent topics $\boldsymbol{\beta}$. We do not solve the model selection problem and assume the number of topics K is known[Blei, 2012].

Technical Conditions: In this chapter, we assume that the topic matrix $\boldsymbol{\beta}$ is non-random and is exact separable ($\lambda = 0$ in Definitions 1). In addition to separability, we require some technical regularities on the prior distribution for topic weights $\boldsymbol{\theta}^m$'s. Specifically, let $\mathbf{a} = \mathbb{E}(\boldsymbol{\theta}^m)$ and $\mathbf{R} = \mathbb{E}(\boldsymbol{\theta}^m \boldsymbol{\theta}^{m\top})$, and define $\bar{\mathbf{R}} := \text{diag}(\mathbf{a})^{-1} \mathbf{R} \text{diag}(\mathbf{a})^{-1}$. We consider the following conditions,

Condition 1. (Simplicial) *A matrix \mathbf{B} is (row-wise) γ_s -simplicial if any row-vector of \mathbf{B} is at a distance of at least $\gamma_s > 0$ from the convex hull of the remaining row-vectors. A topic model is γ_s -simplicial if its normalized second-order moment $\bar{\mathbf{R}}$ is γ_s -simplicial.*

² $\boldsymbol{\alpha}$ denotes the hyper-parameters, e.g., the concentration parameter in Dirichlet.

³All the words in the same document are therefore independent drawings from the \mathbf{A}^m . This model ignores the sequential order and is the classic “bag-of-words” modeling paradigm [Blei, 2012]. When it is clear from the context, we will use $X_{m,n}$ to represent either the empirical word-count or, by suitable column-normalization of \mathbf{X} , the empirical word-frequency.



(Left, Deterministic Setting) Given β ; or (Right, Smoothed Setting) For $k = 1, \dots, K$, sample topics $\beta^k \in \mathbb{R}^W \sim \text{Dir}(\beta_0)$

For each document $m = 1, \dots, M$,

- 1) Sample a topic weight vector $\theta^m \in \mathbb{R}^K$ from some prior $\text{Pr}(\theta)$
- 2) For each word $n = 1, \dots, N$ in the document,
 - (a) Sample a word token $z_{m,n} \in \{1, \dots, W\}$ from $\text{Multinomial}(\theta^m)$
 - (b) Sample a word $w_{m,n}$ from $\beta^{z_{m,n}}$

Figure 2-1: Generative process and the graphical plate representation of a topic model. The boxes represent replicates. The outer plate represents (M) documents, and the inner plate represents (N) word topic-tokens and observed words of each document. Left figure represents the deterministic setting and right figure the smoothed settings of topic matrix β .

Condition 2. (Affine Independence) A matrix \mathbf{B} is (row-wise) γ_a -affine independent if $\min_{\lambda} \|\sum_{k=1}^K \lambda_k \mathbf{B}_k\|_2 / \|\lambda\|_2 \geq \gamma_a > 0$, where \mathbf{B}_k is the k -th row of \mathbf{B} and the minimum is over all $\lambda \in \mathbb{R}^K$ such that $\lambda \neq \mathbf{0}$ and $\sum_{k=1}^K \lambda_k = 0$. A topic model is γ_a -affine independent if its normalized second-order moment $\bar{\mathbf{R}}$ is γ_a -affine independent.

Here γ_s and γ_a are called the simplicial and affine-independence constants respectively.

They are condition numbers which measure the degree to which the conditions that they are respectively associated with hold. The larger that these condition numbers are, the easier it is to estimate the topic matrix. Going forward, we will say that a matrix is simplicial (resp. affine independent) if it is γ_s -simplicial (resp. γ_a -affine-independent) for some $\gamma_s > 0$ (resp. $\gamma_a > 0$).⁴ Common priors like Dirichlet (LDA)

⁴We use the Euclidean distance for Condition 1 and 2 in this thesis. We can, in principle, use the other distance metric in these definitions.

and log-Normal (CTM) satisfy the simplicial and affine independent condition.

Main Results: We develop a novel approach to learn β [Ding et al., 2013a,b, 2014b] with provable guarantees. Informally,

Theorem 1. (Informal) *Let topic matrix β be separable.*

- a If $\bar{\mathbf{R}}$ for topic weight prior is γ_s - simplicial, then, our proposed approach runs in **polynomial** time in terms of M, N, K, W , can **consistently** detect all the novel words for K distinct topics when $N \geq 2$ is fixed and $M \rightarrow \infty$, and fails with probability at most δ if $M \geq \mathbf{Poly}(W, \log(1/\delta), K, 1/N)$.*
- b If $\bar{\mathbf{R}}$ for topic weight prior is γ_a - affine independent, then, our proposed approach can further estimate β with ϵ element wise error with probability at least $1 - \delta$ if $M \geq \mathbf{Poly}(W, \log(1/\delta), K, 1/N, 1/\epsilon)$.⁵*

The asymptotic setting that N being fixed and $M \rightarrow \infty$ is motivated by the empirical text corpus in which the number of words in each document is small while the number of documents is large. Our algorithm can be applied to a general *family of topic priors* $\text{Pr}(\alpha)$ that satisfy Condition 1 and/or 2. In contrast, the standard VB or MCMC need to be designed for each specific topic prior.

Organization: We first overview the related works in topic modeling in Section 2.2. We depict the key geometric motivation in Section 2.3 as implications of separability combined with the necessary technical regularities on the mixing weights. We develop a word co-occurrence representation in Section 2.4 to consistently achieve this geometry. We also develop a extreme point measure, solid angle, to handle finite sampling perturbation in empirical observations. We summarize the centralized version of our algorithm in Section 2.5 and discuss a distributed implementation of our approach that can provably achieve the centralized guarantees in Section 2.6. We demonstrate a set of synthetic and real-world experiments in Section 2.7.

⁵The topic estimation is only consistent up to a column permutation.

2.2 Related Works

The idea of modeling text documents as mixtures of a few semantic topics was first proposed in Hofmann [1999] where the mixing weights were assumed to be deterministic. Latent Dirichlet Allocation (LDA) in the seminal work of Blei et al. [2003] extended this to a probabilistic setting by modeling topic mixing weights using Dirichlet priors. This setting has been further extended to include other topic priors such as the log-normal prior in the Correlated Topic Model Blei and Lafferty [2007]. LDA models and their derivatives have been successful on a wide range of problems in terms of achieving good empirical performance Airoldi et al. [2014], Blei [2012].

The prevailing approaches for estimation and inference problems in topic modeling are based on MAP or ML estimation Blei [2012]. However, the computation of posterior distributions conditioned on observations \mathbf{X} is intractable Blei et al. [2003]. Moreover, the MAP estimation objective is non-convex and has been shown to be \mathcal{NP} -hard Arora et al. [2012], Sontag and Roy [2011]. Therefore various approximation and heuristic strategies have been employed. These approaches fall into two major categories – sampling approaches and optimization approaches. Most sampling approaches are based on Markov Chain Monte Carlo (MCMC) algorithms that seek to generate (approximately) independent samples from a Markov Chain that is carefully designed to ensure that the sample distribution converges to the true posterior Griffiths and Steyvers [2004], Wallach et al. [2009]. Optimization approaches are typically based on the so-called Variational-Bayes methods. These methods optimize the parameters of a simpler parametric distribution so that it is close to the true posterior in terms of KL divergence Blei et al. [2003], Wainwright and Jordan [2008]. Expectation-Maximization-type algorithms are typically used in these methods. In practice, while both Variational-Bayes and MCMC algorithms have similar performance, Variational-Bayes is typically faster than MCMC Blei [2012], Hoffman et al.

[2010].

Nonnegative Matrix Factorization (NMF) is an alternative approach for topic estimation. NMF-based methods exploit the fact that both the topic matrix β and the mixing weights are nonnegative and attempt to decompose the empirical observation matrix \mathbf{X} into a product of a nonnegative topic matrix β and the matrix of mixing weights by minimizing a cost function of the form Cichocki et al. [2009], Hoffman et al. [2010], Lee and Seung [1999], Recht et al. [2012]

$$\sum_{m=1}^M d(\mathbf{X}^m, \beta \boldsymbol{\theta}^m) + \lambda \psi(\beta, \boldsymbol{\theta}^1, \dots, \boldsymbol{\theta}^M),$$

where $d(\cdot)$ is some measure of closeness and ψ is a regularization term which enforces desirable properties, e.g., sparsity, on β and the mixing weights. The NMF problem, however, is also known to be non-convex and \mathcal{NP} -hard Vavasis [2009] in general. Sub-optimal strategies such as alternating minimization, greedy gradient descent, and heuristics are used in practice Cichocki et al. [2009].

In contrast to the above approaches, a new approach has recently emerged which is based on imposing additional structure on the model parameters Anandkumar et al. [2014], Arora et al. [2012, 2013], Ding et al. [2013b, 2014b], Kumar et al. [2013]. These approaches show that the topic discovery problem lends itself to provably consistent and polynomial-time solutions by making assumptions about the *structure* of the topic matrix β and the distribution of the mixing weights. In this category of approaches are methods based on a tensor decomposition of the moments of \mathbf{X} Anandkumar et al. [2013, 2014]. The algorithm in Anandkumar et al. [2013] uses second order empirical moments and is shown to be asymptotically consistent when the topic matrix β has a special sparsity structure. The algorithm in Anandkumar et al. [2014] uses the third order tensor of observations. It is, however, strongly tied to the specific structure of the Dirichlet prior on the mixing weights and requires knowledge of the concentration

parameters of the Dirichlet distribution Anandkumar et al. [2014]. Furthermore, in practice these approaches are computationally intensive and require some initial coarse dimensionality reduction, gradient descent speedups, and GPU acceleration to process large-scale text corpora like the NYT dataset Anandkumar et al. [2014].

Our work falls into the family of approaches that exploit the separability property of β and its geometric implications Arora et al. [2012, 2013], Awasthi and Risteski [2015], Bansal et al. [2014], Ding et al. [2013b, 2014b], Kumar et al. [2013]. An asymptotically consistent polynomial-time topic estimation algorithm was first proposed in Arora et al. [2012]. However, this method requires solving W linear programs, each with W variables and is computationally impractical. Subsequent work improved the computational efficiency Kumar et al. [2013], Recht et al. [2012], but theoretical guarantees of asymptotic consistency (when N fixed, and the number of documents $M \rightarrow \infty$) are unclear. Algorithms in Arora et al. [2013] and Ding et al. [2013b] are both practical and provably consistent. Each requires a stronger and slightly different technical condition on the topic mixing weights than Arora et al. [2012]. Specifically, Arora et al. [2013] imposes a full-rank condition on the second-order correlation matrix of the mixing weights and proposes a Gram-Schmidt procedure to identify the extreme points. Similarly, Ding et al. [2013b] imposes a diagonal-dominance condition on the same second-order correlation matrix and proposes a random projections based approach. These approaches are tied to the specific conditions imposed and they both fail to detect all the novel words and estimate topics when the imposed conditions (which are sufficient but not necessary for consistent novel word detection or topic estimation) fail to hold in some examples Ding et al. [2014b]. The random projections based algorithm proposed in Ding et al. [2014b] is both practical and provably consistent. Furthermore, it requires fewer constraints on the topic mixing weights.

We note that the separability property has been exploited in other recent work as well Awasthi and Risteski [2015], Bansal et al. [2014]. In Bansal et al. [2014], a singular value decomposition based approach is proposed for topic estimation. In Awasthi and Risteski [2015], it is shown that the standard Variational-Bayes approximation can be asymptotically consistent if β is separable. However, the additional constraints proposed essentially boil down to the requirement that each document contain predominantly only one topic. In addition to assuming the existence of such “pure” documents, Awasthi and Risteski [2015] also requires a strict initialization. It is thus unclear how this can be achieved using only the observations \mathbf{X} .

The separability property has been re-discovered and exploited in the literature across a number of different fields and has found application in several problems. To the best of our knowledge, this concept was first introduced as the *Pure Pixel Index* assumption in the Hyperspectral Image unmixing problem Boardman [1993]. This work assumes the existence of pixels in a hyper-spectral image containing predominantly one species. Separability has also been studied in the NMF literature in the context of ensuring the uniqueness of NMF Donoho and Stodden [2004]. Subsequent work has led to the development of NMF algorithms that exploit separability Gillis and Vavasis [2014], Recht et al. [2012]. The uniqueness and correctness results in this line of work has primarily focused on the noiseless case. We finally note that separability has also been recently exploited in the problem of learning multiple ranking preferences from pairwise comparisons for personal recommendation systems and information retrieval Ding et al. [2015b], Farias et al. [2009] and has led to provably consistent and efficient estimation algorithms.

2.3 Topic Separability, Necessary and Sufficient Conditions, and the Geometric Intuitions

In this section, we unravel the key ideas that motivate our algorithmic approach by focusing on the ideal case where there is no “sampling-noise”, i.e., each document is infinitely long ($N = \infty$). In the next section, we will turn to the finite N case. We recall that β and \mathbf{X} denote the $W \times K$ topic matrix and the $W \times M$ empirical word counts/frequency matrix respectively. Also, M, W , and K denote, respectively, the number of documents, the vocabulary size, and the number of topics. For convenience, we group the document-specific mixing weights, the θ^m 's, into a $K \times M$ weight matrix $\theta = [\theta^1, \dots, \theta^M]$ and the document-specific distributions, the \mathbf{A}^m 's, into a $W \times M$ document distribution matrix $\mathbf{A} = [\mathbf{A}^1, \dots, \mathbf{A}^M]$. The generative procedure that describes a topic model then implies that $\mathbf{A} = \beta\theta$. In the ideal case considered in this section ($N = \infty$), the empirical word *frequency* matrix $\mathbf{X} = \mathbf{A}$.

Notation: A vector \mathbf{a} without specification will denote a column-vector, $\mathbf{1}$ the all-ones column vector of suitable size, \mathbf{X}^i the i -th column vector and \mathbf{X}_j the j -th row vector of matrix \mathbf{X} , and $\bar{\mathbf{B}}$ a suitably row-normalized version (described later) of a non-negative matrix \mathbf{B} . Also, $[n] := \{1, \dots, n\}$.

2.3.1 Key Structural Property: Topic Separability

We first recall our key structural property, the exact separability,

Definition 2. (Exact Separability) *A topic matrix $\beta \in \mathbb{R}^{W \times K}$ is separable if for each topic k , there is some word i such that $\beta_{i,k} > 0$ and $\beta_{i,l} = 0, \forall l \neq k$.*

Topic separability implies that each topic contains word(s) which appear only in that topic. We refer to these words as the **novel words** of the K topics. Figure 2-2 shows an example of a separable β with $K = 3$ topics. Words 1 and 2 are novel to topic 1, words 3 and 4 to topic 2, and word 5 to topic 3. Other words that appear in multiple topics are called non-novel words (e.g., word 6). Identifying the novel words for K distinct topics is the key step of our proposed approach.

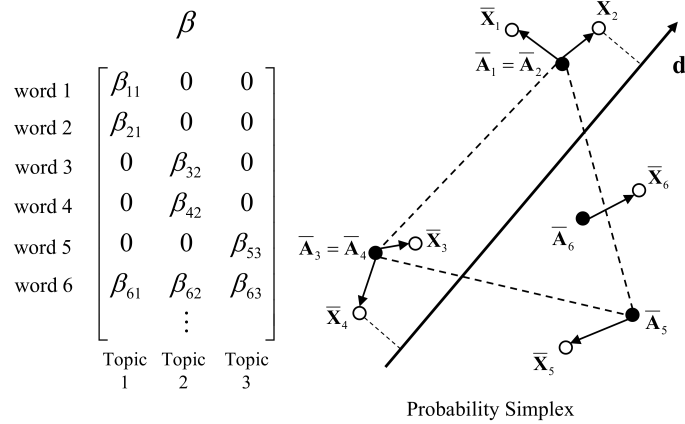


Figure 2.2: An example of separable topic matrix β (left) and the underlying geometric structure (right) of the row space of the normalized document distribution matrix $\bar{\mathbf{A}}$. Note: the word ordering is only for visualization and has no bearing on separability. Solid circles represent **rows** of $\bar{\mathbf{A}}$. Empty circles represent **rows** of $\bar{\mathbf{X}}$ when N is finite (in the ideal case, $\bar{\mathbf{A}} = \bar{\mathbf{X}}$). Projections of $\bar{\mathbf{A}}_w$'s (resp. $\bar{\mathbf{X}}_w$'s) along a random isotropic direction \mathbf{d} can be used to identify novel words.

Empirically, separability has been observed to be approximately satisfied by topic estimates produced by VB and MCMC algorithms [Arora et al., 2013, Awasthi and Risteski, 2015, Ding et al., 2014b]. More *fundamentally*, in very recent work [Ding et al., 2015a] to be Discussed in Chapter 4, it has been shown that topic separability is an inevitable consequence of having a relatively small number of topics in a very large vocabulary (high-dimensionality). To be more explicit, if we consider the standard smoothed setting of topic modeling where topics are sampled iid from a Dirichlet prior [Blei [2012], then, most of the topic matrices are approximately separable if $W \gg K$. Therefore, our geometric approach to be develop next indeed can be applied to most large topic models. As we will discuss next in Sec. 2.3.3, the topic separability property combined with additional conditions on the second-order statistics of the mixing weights leads to an intuitively appealing geometric property that can be exploited to develop a provably consistent and efficient topic estimation algorithm.

2.3.2 Conditions on the Topic Mixing Weights

$$\begin{array}{ccc}
 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ \dots & & \end{pmatrix} & \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ 0.5\boldsymbol{\theta}_1 + 0.5\boldsymbol{\theta}_2 \end{pmatrix} & = & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0.5 & 0.5 & 0 \\ \dots & & \end{pmatrix} & \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \\ 0.5\boldsymbol{\theta}_1 + 0.5\boldsymbol{\theta}_2 \end{pmatrix} \\
 \boldsymbol{\beta}^{(1)} & \boldsymbol{\theta} & & \boldsymbol{\beta}^{(2)} & \boldsymbol{\theta}
 \end{array}$$

Figure 2.3: Example showing that topic separability **alone** does not guarantee a unique solution to the problem of estimating $\boldsymbol{\beta}$ from \mathbf{X} . Here, $\boldsymbol{\beta}_1\boldsymbol{\theta} = \boldsymbol{\beta}_2\boldsymbol{\theta} = \mathbf{A}$ is a document distribution matrix that is consistent with two different topic matrices $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$ that are both separable.

Topic separability alone does not guarantee that there will be a unique $\boldsymbol{\beta}$ that is consistent with all the observations \mathbf{X} . This is illustrated in Fig. 2.3 Ding et al. [2013a]. Therefore, in an effort to develop provably consistent topic estimation algorithms, a number of different conditions have been imposed on the topic mixing weights $\boldsymbol{\theta}$ in the literature Arora et al. [2012, 2013], Ding et al. [2013b, 2014b], Kumar et al. [2013]. In this section, we identify the simplicial condition (Condition 1) as the necessary and sufficient conditions for consistent *detection* of novel words. We then show that the affine-independence condition (Condition 2) is necessary and sufficient for consistent *estimation* of a separable topic matrix. Our necessity results are *information-theoretic* and *algorithm-independent* in nature, meaning that they are independent of any specific statistics of the observations and the algorithms used. The novel words and the topics can only be identified up to a permutation and is accounted for in our results.

We first recall in Section 2.1, the simplicial and affine-independent conditions are defined based on a key quantity $\bar{\mathbf{R}} := \text{diag}(\mathbf{a})^{-1}\mathbf{R}\text{diag}(\mathbf{a})^{-1}$, where $\mathbf{a} := \mathbb{E}(\boldsymbol{\theta}^m)$ and $\mathbf{R} := \mathbb{E}(\boldsymbol{\theta}^m\boldsymbol{\theta}^{m\top})$. We implicitly assume that all the elements of \mathbf{a} to be strictly positive. Before we discuss the necessity and sufficiency, we point out that the affine independent condition is a stronger condition than simplicial:

Proposition 1. $\bar{\mathbf{R}}$ is γ_a -affine-independent $\Rightarrow \bar{\mathbf{R}}$ is at least γ_a -simplicial. The reverse implication is false in general.

The Simplicial Condition is both Necessary and Sufficient for Novel Word

Detection: We first focus on detecting all the novel words of the K distinct topics. For this task, the simplicial condition is an algorithm-independent, information-theoretic necessary condition. Formally,

Lemma 1. (*Simplicial Condition is Necessary for Novel Word Detection [Ding et al., 2013a, Lemma 1]*) Let β be separable and $W > K$. If there exists an algorithm that can consistently identify all novel words of all K topics from \mathbf{X} , then $\bar{\mathbf{R}}$ is simplicial.

The key insight behind this result is that when $\bar{\mathbf{R}}$ is non-simplicial, we can construct two distinct separable topic matrices with different sets of novel words which induce the same distribution on the empirical observations \mathbf{X} . Geometrically, the simplicial condition guarantees that the K rows of $\bar{\mathbf{R}}$ will be extreme points of the convex hull that they themselves form. Therefore, if $\bar{\mathbf{R}}$ is not simplicial, there will exist at least one redundant topic which is just a convex combination of the other topics.

It turns out that $\bar{\mathbf{R}}$ being simplicial is also sufficient for consistent novel word detection. This is a direct consequence of the consistency guarantees of our approach as outlined in Theorem 3.

Affine-Independence is Necessary and Sufficient for Separable Topic Es-

timation: We now focus on estimating a separable topic matrix β , which is a stronger requirement than detecting novel words. It naturally requires conditions that are stronger than the simplicial condition. Affine-independence turns out to be an algorithm-independent, information-theoretic necessary condition. Formally,

Lemma 2. (*Affine-Independence is Necessary for Separable Topic Estimation*) Let β be separable with $W \geq 2 + K$. If there exists an algorithm that can consistently estimate β from \mathbf{X} , then its normalized second-moment $\bar{\mathbf{R}}$ is affine-independent.

Similar to Lemma 1, if $\bar{\mathbf{R}}$ is not affine-independent, we can construct two distinct and separable topic matrices that induce the same distribution on the observation which makes consistent topic estimation impossible. Geometrically, every point in a convex set can be decomposed *uniquely* as a convex combination of its extreme points, if, and only if, the extreme points are affine-independent. Hence, if $\bar{\mathbf{R}}$ is not affine-independent, a non-novel word can be assigned to different subsets of topics.

The sufficiency of the affine-independence condition in separable topic estimation is again a direct consequence of the consistency guarantees of our approach as in Theorems 3 and 4. We note that since affine-independence implies the simplicial condition (Proposition 1), affine-independence is sufficient for novel word detection as well.

Connection to Other Conditions on the Mixing Weights: We briefly discuss other conditions on the mixing weights θ that have been exploited in the literature. In Arora et al. [2013], Kumar et al. [2013], \mathbf{R} (equivalently $\bar{\mathbf{R}}$) is assumed to have full-rank (with minimum eigenvalue $\gamma_r > 0$). In Ding et al. [2013b], $\bar{\mathbf{R}}$ is assumed to be diagonal dominant, i.e., $\forall i, j, i \neq j, \bar{\mathbf{R}}_{i,i} - \bar{\mathbf{R}}_{i,j} \geq \gamma_d > 0$. They are both sufficient conditions for detecting all the novel words of all distinct topics. The constants γ_r and γ_d are condition numbers which measure the degree to which the full-rank and diagonal dominance conditions hold respectively. They are counterparts of γ_s and γ_a and like them, the larger they are, the easier it is to consistently detect the novel words and estimate β . The relationships between these conditions are summarized in Proposition 2 and illustrated in Fig. 2.4.

Proposition 2. *Let $\bar{\mathbf{R}}$ be the normalized second-moment of the topic prior. Then,*

1. $\bar{\mathbf{R}}$ is full rank with minimum eigenvalue $\gamma_r \Rightarrow \bar{\mathbf{R}}$ is at least γ_r -affine-independent $\Rightarrow \bar{\mathbf{R}}$ is at least γ_r -simplicial.
2. $\bar{\mathbf{R}}$ is γ_d -diagonal dominant $\Rightarrow \bar{\mathbf{R}}$ is at least γ_d -simplicial.

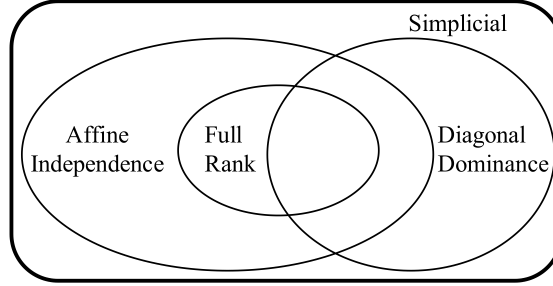


Figure 2-4: Relationships between Simplicial, Affine-Independence, Full Rank, and Diagonal Dominance conditions on the normalized second-moment $\bar{\mathbf{R}}$.

3. $\bar{\mathbf{R}}$ being diagonal dominant neither implies nor is implied by $\bar{\mathbf{R}}$ being affine-independent (or full-rank).

We note that in our earlier work Ding et al. [2014b], the provable guarantees for estimating the separable topic matrix require $\bar{\mathbf{R}}$ to have full rank. The analysis in this paper provably extends the guarantees to the affine-independence condition.

2.3.3 Geometric Implications and Random Projections Based Algorithm

We now demonstrate the geometric implications of topic separability combined with the simplicial/ affine-independence condition on the topic mixing weights. To highlight the key ideas we focus on the ideal case where $N = \infty$. Then, the empirical document word-frequency matrix $\mathbf{X} = \mathbf{A} = \boldsymbol{\beta}\boldsymbol{\theta}$.

Novel Words are Extreme Points: To expose the underlying geometry, we normalize the rows of \mathbf{A} and $\boldsymbol{\theta}$ to obtain row-stochastic matrices $\bar{\mathbf{A}} := \text{diag}(\mathbf{A}\mathbf{1})^{-1}\mathbf{A}$ and $\bar{\boldsymbol{\theta}} := \text{diag}(\boldsymbol{\beta}\mathbf{1})^{-1}\boldsymbol{\beta}$. Then since $\mathbf{A} = \boldsymbol{\beta}\boldsymbol{\theta}$, we have $\bar{\mathbf{A}} = \bar{\boldsymbol{\beta}}\bar{\boldsymbol{\theta}}$ where $\bar{\boldsymbol{\beta}} := \text{diag}(\mathbf{A}\mathbf{1})^{-1}\boldsymbol{\beta}\text{diag}(\boldsymbol{\theta}\mathbf{1})$ is a row-normalized ‘‘topic matrix’’ which is both row-stochastic and separable with the same sets of novel words as $\boldsymbol{\beta}$.

Now consider the row vectors of $\bar{\mathbf{A}}$ and $\bar{\boldsymbol{\theta}}$. First, it can be shown that if $\bar{\mathbf{R}}$ is simplicial (cf. Condition 1) then, with probability one, no row of $\bar{\boldsymbol{\theta}}$ will be in the convex hull of the others Ding et al. [2013a]. Next, the separability property ensures that if w is a novel word of topic k , then $\bar{\beta}_{wk} = 1$ and $\bar{\beta}_{wj} = 0 \forall j \neq k$ so that $\bar{\mathbf{A}}_w = \bar{\boldsymbol{\theta}}_k$.

Revisiting the example in Fig. 2·2, the rows of $\bar{\mathbf{A}}$ which correspond to novel words, e.g., words 1 through 5, are all row-vectors of $\bar{\boldsymbol{\theta}}$ and together form a convex hull of K extreme points. For example, $\bar{\mathbf{A}}_1 = \bar{\mathbf{A}}_2 = \bar{\boldsymbol{\theta}}_1$ and $\bar{\mathbf{A}}_3 = \bar{\mathbf{A}}_4 = \bar{\boldsymbol{\theta}}_2$. If, however, w is a non-novel word, then $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk} \bar{\boldsymbol{\theta}}_k$ lives inside the convex hull of the rows of $\bar{\boldsymbol{\theta}}$. In Fig. 2·2, row $\bar{\mathbf{A}}_6$ which corresponds to non-novel word 6, is inside the convex hull of $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2, \bar{\boldsymbol{\theta}}_3$. In summary, the novel words can be detected as extreme points of all the row-vectors of $\bar{\mathbf{A}}$. Also, multiple novel words of the same topic correspond to the same extreme point (e.g., $\bar{\mathbf{A}}_1 = \bar{\mathbf{A}}_2 = \bar{\boldsymbol{\theta}}_1$). Formally,

Lemma 3. (Novel Words are Extreme Points) *Let $\bar{\mathbf{R}}$ be simplicial and $\boldsymbol{\beta}$ be separable. Then, a word i is novel if, and only if, the i -th row of $\bar{\mathbf{A}}$ is an extreme point of the convex hull spanned by all the rows of $\bar{\mathbf{A}}$.*

To see how identifying novel words can help us estimate $\boldsymbol{\beta}$, recall that the row-vectors of $\bar{\mathbf{A}}$ corresponding to novel words coincide with the rows of $\bar{\boldsymbol{\theta}}$. Thus $\bar{\boldsymbol{\theta}}$ is known once one novel word for each topic is known. Also, for all words w , $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk} \bar{\boldsymbol{\theta}}_k$. Thus, if we can *uniquely* decompose $\bar{\mathbf{A}}_w$ as a convex combination of the extreme points, then the coefficients of the decomposition will give us the w -th row of $\bar{\boldsymbol{\beta}}$. A unique decomposition exists with certainty when $\bar{\mathbf{R}}$ is affine-independent and can be found by solving a constrained linear regression problem. This gives us $\bar{\boldsymbol{\beta}}$. Finally, noting that $\text{diag}(\mathbf{A}\mathbf{1})\bar{\boldsymbol{\beta}} = \boldsymbol{\beta} \text{diag}(\boldsymbol{\theta}\mathbf{1})$, $\boldsymbol{\beta}$ can be recovered by suitably renormalizing rows and then columns of $\bar{\boldsymbol{\beta}}$. To sum up,

Lemma 4. *Given \mathbf{A} and the novel words for K distinct topics. If $\bar{\mathbf{R}}$ is further affine independent, then, $\boldsymbol{\beta}$ can be estimated uniquely using constrained linear regressions.*

Lemmas 3 and 4 together provide a geometric approach for learning $\boldsymbol{\beta}$ from \mathbf{A} (equivalently $\bar{\mathbf{A}}$):

1. Find extreme points of rows of $\bar{\mathbf{A}}$. Cluster the rows of $\bar{\mathbf{A}}$ that correspond to the same extreme point into the same group.

2. Express the remaining rows of $\bar{\mathbf{A}}$ as convex combination of the K distinct extreme points.
3. Re-normalized $\bar{\beta}$ to obtain β .

Detecting Extreme Points using Random Projections: A key contribution of our approach is an efficient random projections based algorithm to detect novel words as extreme points. The idea is illustrated in Fig. 2.2: if we project every point of a convex body onto an isotropically distributed random direction \mathbf{d} , the maximum (or minimum) projection value must correspond to one of the extreme points with probability 1. On the other hand, the non-novel words will not have the maximum projection value along any random direction. Therefore, by repeatedly projecting all the points onto a few isotropically distributed random directions, we can detect all the extreme points with very high probability as the number of random directions increase. An explicit bound on the number of projections needed appears in Theorem 3.

Finite N in Practice: The geometric intuition discussed above was based on the row-vectors of $\bar{\mathbf{A}}$. When $N = \infty$, $\bar{\mathbf{A}} = \bar{\mathbf{X}}$ the matrix of row-normalized empirical word-frequencies of all documents. If N is finite but very large, $\bar{\mathbf{A}}$ can be well-approximated by $\bar{\mathbf{X}}$ thanks to the law of large numbers. However, in real-word text corpora, $N \ll W$ (e.g., $N = 298$ while $W = 14,943$ in the NYT dataset). Therefore, the row-vectors of $\bar{\mathbf{X}}$ are significantly perturbed away from the ideal rows of $\bar{\mathbf{A}}$ as illustrated in Fig. 2.2. We discuss the effect of small N and how we address the accompanying issues next.

2.4 Topic geometry with a finite sample size: word co-occurrence matrix representation, solid angle, and random projection based approach

The extreme point geometry sketched in Sec. 2.3.3 is perturbed when N is small as highlighted in Fig. 2.2. Specifically, the rows of the empirical word-frequency matrix \mathbf{X} deviate from the rows of \mathbf{A} . This creates several problems: (1) points in the convex hull corresponding to non-novel words may also become “outlier” extreme points (e.g., $\bar{\mathbf{X}}_6$ in Fig. 2.2); (2) some extreme points that correspond to novel words may no longer be extreme (e.g., $\bar{\mathbf{X}}_3$ in Fig. 2.2); (3) multiple novel words corresponding to the same extreme point may become multiple distinct extreme points (e.g., $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_2$ in Fig. 2.2). Unfortunately, these issues do not vanish as M increases with N fixed – a regime which captures the characteristics of typical benchmark datasets – because the dimensionality of the rows (equal to M) also increases. There is no “averaging” effect to smoothen-out the sampling noise.

Our solution is to seek a new representation, a statistic of \mathbf{X} , which can not only smoothen out the sampling noise of individual documents, but also preserve the same extreme point geometry induced by the separability and affine independence conditions. In addition, we also develop an extreme point robustness measure that naturally arises within our random projections based framework. This robustness measure can be used to detect and exclude the “outlier” extreme points.

2.4.1 Normalized Word Co-occurrence Matrix Representation

We construct a suitably normalized word co-occurrence matrix from \mathbf{X} as our new representation. The co-occurrence matrix converges almost surely to an ideal statistic as $M \rightarrow \infty$ for any fixed $N \geq 2$. Simultaneously, in the asymptotic limit, the original novel words continue to correspond to extreme points in the new representation and overall extreme point geometry is preserved.

The new representation is (conceptually) constructed as follows. First randomly divide all the words in each document into two equal-sized independent halves and obtain two $W \times K$ empirical word-frequency matrices \mathbf{X} and \mathbf{X}' each containing $N/2$ words. Then normalize their rows like in Sec. 2.3.3 to obtain $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}'}$ which are row-stochastic. The empirical word co-occurrence matrix of size $W \times W$ is then given by

$$\hat{\mathbf{E}} = M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top \quad (2.1)$$

We note that in our random projection based approach, $\hat{\mathbf{E}}$ is not *explicitly* constructed by multiplying $\bar{\mathbf{X}'}$ and $\bar{\mathbf{X}}$. Instead, we keep $\bar{\mathbf{X}'}$ and $\bar{\mathbf{X}}$ and exploit their sparsity properties to reduce the computational complexity of all subsequent processing.

Asymptotic Consistency: The first nice property of the word co-occurrence representation is its asymptotic consistency when N is fixed. As the number of documents $M \rightarrow \infty$, the empirical $\hat{\mathbf{E}}$ converges, almost surely, to an ideal word co-occurrence matrix \mathbf{E} of size $W \times W$. Formally,

Lemma 5. *Let $\hat{\mathbf{E}}$ be the empirical word co-occurrence matrix defined in Eq. (2.1). Then,*

$$\hat{\mathbf{E}} \xrightarrow[\text{almost surely}]{M \rightarrow \infty} \bar{\boldsymbol{\beta}}\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top =: \mathbf{E} \quad (2.2)$$

where $\bar{\boldsymbol{\beta}} := \text{diag}^{-1}(\boldsymbol{\beta}\mathbf{a})\boldsymbol{\beta} \text{diag}(\mathbf{a})$ and $\bar{\mathbf{R}} := \text{diag}^{-1}(\mathbf{a})\mathbf{R} \text{diag}^{-1}(\mathbf{a})$. Furthermore, if $\eta := \min_{1 \leq i \leq W} (\boldsymbol{\beta}\mathbf{a})_i > 0$, then $\Pr(\|\hat{\mathbf{E}} - \mathbf{E}\|_\infty \geq \epsilon) \leq 8W^2 \exp(-\epsilon^2 \eta^4 MN/20)$.

Here $\bar{\mathbf{R}}$ is the same normalized second-moment of the topic priors as defined in Sec. 2.3 and $\bar{\boldsymbol{\beta}}$ is a row-normalized version of $\boldsymbol{\beta}$. We make note of the abuse of notion for $\bar{\boldsymbol{\beta}}$ which was defined in Sec. 2.3.3. It can be shown that the $\bar{\boldsymbol{\beta}}$ defined in Lemma 5 is the limit of the one defined in Sec. 2.3.3 as $M \rightarrow \infty$. The convergence result in Lemma 5 shows that the word co-occurrence representation \mathbf{E} can be consistently estimated by $\hat{\mathbf{E}}$ as $M \rightarrow \infty$ and the deviation vanishes exponentially in M which is large in typical benchmark datasets.

Novel Words are Extreme Points: Another reason for using this word co-occurrence representation is that it preserves the extreme point geometry. Consider the ideal word co-occurrence matrix $\mathbf{E} = \bar{\beta}(\bar{\mathbf{R}}\bar{\beta}^\top)$. It is straightforward to show that if $\bar{\beta}$ is separable and $\bar{\mathbf{R}}$ is simplicial then $(\bar{\mathbf{R}}\bar{\beta}^\top)$ is also simplicial. Using these facts it is possible to establish the following counterpart of Lemma 3 for \mathbf{E} :

Lemma 6. (Novel Words are Extreme Points) *Let $\bar{\mathbf{R}}$ be simplicial and β be separable. Then, a word i is novel if, and only if, the i -th row of \mathbf{E} is an extreme point of the convex hull spanned by all the rows of \mathbf{E} .*

In another words, the novel words correspond to the extreme points of all the row-vectors of the ideal word co-occurrence matrix \mathbf{E} . Consider the example in Fig. 2.5 which is based on the same topic matrix β as in Fig. 2.2. Here, $\mathbf{E}_1 = \mathbf{E}_2, \mathbf{E}_3 = \mathbf{E}_4$, and \mathbf{E}_5 are $K = 3$ distinct extreme points of all row-vectors of \mathbf{E} and \mathbf{E}_6 , which corresponds to a non-novel word, is inside the convex hull.

Once the novel words are detected as extreme points, we can follow the same procedure as in Lemma 4 and express each row \mathbf{E}_w of \mathbf{E} as a unique convex combination of the K extreme rows of \mathbf{E} or equivalently the rows of $(\bar{\mathbf{R}}\bar{\beta}^\top)$. The weights of the convex combination are the $\bar{\beta}_{wk}$'s. We can then apply the same row and column renormalization to obtain β . The following result is the counterpart of Lemma 4 for \mathbf{E} :

Lemma 7. *Let \mathbf{E} and one novel word for each distinct topic be given. If $\bar{\mathbf{R}}$ is affine-independent, then β can be recovered uniquely via constrained linear regression.*

One can follow the same steps as in the proof of Lemma 4. The only additional step is to check that $\bar{\mathbf{R}}\bar{\beta}^\top = [\bar{\mathbf{R}}, \bar{\mathbf{R}}\mathbf{B}]$ is affine-independent if $\bar{\mathbf{R}}$ is affine-independent.

We note that the finite sampling noise perturbation $\hat{\mathbf{E}} - \mathbf{E}$ is still not 0 but vanishes as $M \rightarrow \infty$ (in contrast to the $\bar{\mathbf{X}}$ representation in Sec. 2.3.3). However, there is still a possibility of observing “outlier” extreme points if a non-novel word lies on the facet of the convex hull of the rows of \mathbf{E} . We next introduce an extreme point robustness

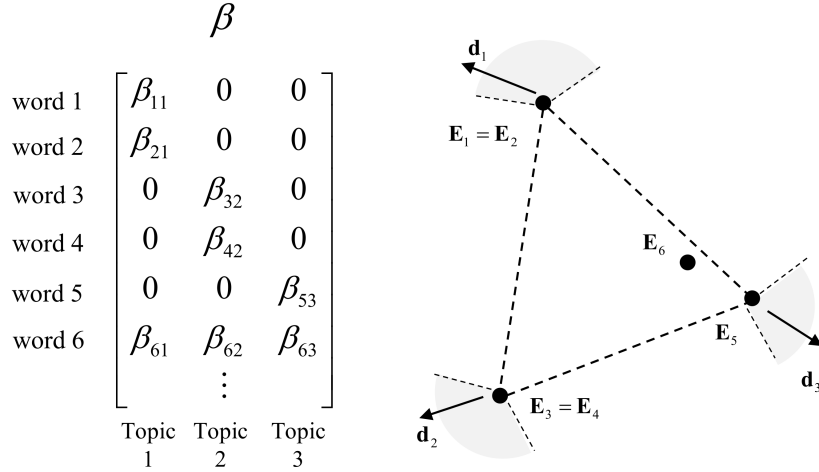


Figure 2-5: An example of separable topic matrix β (left) and the underlying geometric structure (right) in the word co-occurrence representation. Note: the word ordering is only for visualization and has no bearing on separability. The example topic matrix β is the same as in Fig. 2-2. Solid circles represent the **rows** of \mathbf{E} . The shaded regions depict the solid angles subtended by each extreme point. $\mathbf{d}^1, \mathbf{d}^2, \mathbf{d}^3$ are isotropic random directions along which each extreme point has maximum projection value. They can be used to estimate the solid angles.

measure based on a certain *solid angle* that naturally arises in our random projections based approach, and discuss how it can be used to detect and distinguish between “true” novel words and such “outlier” extreme points.

2.4.2 Solid Angle Extreme Point Robustness Measure

To handle the impact of a small but nonzero perturbation $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty$, we develop an extreme point “robustness” measure. This is necessary for not only applying our approach to real-world data but also to establish finite sample complexity bounds. Intuitively, a robustness measure should be able to distinguish between the “true” extreme points (row vectors that are novel words) and the “outlier” extreme points (row vectors of non-novel words that become extreme points due to the nonzero perturbation). Towards this goal, we leverage a key geometric quantity, namely,

the *Normalized Solid Angle* subtended by the convex hull of the rows of \mathbf{E} at an extreme point. To visualize this quantity, we revisit our running example in Fig. 2·5 and indicate the solid angles attached to each extreme point by the shaded regions. It turns out that this geometric quantity naturally arises in the context of random projections that was discussed earlier. To see this connection, in Fig. 2·5 observe that the shaded region attached to any extreme point coincides precisely with the set of directions along which its projection is larger (taking sign into account) than that of any other point (whether extreme or not). For example, in Fig. 2·5 the projection of $\mathbf{E}_1 = \mathbf{E}_2$ along \mathbf{d}_1 is larger than that of any other point. Thus, the solid angle attached to a point \mathbf{E}_i (whether extreme or not) can be formally defined as the set of directions $\{\mathbf{d} : \forall j : \mathbf{E}_j \neq \mathbf{E}_i, \langle \mathbf{E}_i, \mathbf{d} \rangle > \langle \mathbf{E}_j, \mathbf{d} \rangle\}$. This set is nonempty only for extreme points. The solid angle defined above is a set. To derive a scalar robustness measure from this set and tie it to the idea of random projections, we adopt a statistical perspective and define the normalized solid angle of a point as the *probability* that the point will have the maximum projection value along an isotropically distributed random direction. Concretely, for the i -th word (row vector), the normalized solid angle q_i is defined as

$$q_i := \Pr(\forall j : \mathbf{E}_j \neq \mathbf{E}_i, \langle \mathbf{E}_i, \mathbf{d} \rangle > \langle \mathbf{E}_j, \mathbf{d} \rangle) \quad (2.3)$$

where \mathbf{d} is drawn from an isotropic distribution in \mathbb{R}^W such as the spherical Gaussian. The condition $\mathbf{E}_i \neq \mathbf{E}_j$ in Eq. (2.3) is introduced to exclude the multiple novel words of the same topic that correspond to the same extreme point. For instance, in Fig. 2·5 $\mathbf{E}_1 = \mathbf{E}_2$, Hence, for q_1 , $j = 2$ is excluded. To make it practical to handle finite sample estimation noise we replace the condition $\mathbf{E}_j \neq \mathbf{E}_i$ by the condition $\|\mathbf{E}_i - \mathbf{E}_j\| \geq \zeta$ for some suitably defined ζ .

As illustrated in Fig. 2·5, the solid angle for all the extreme points are strictly positive given $\bar{\mathbf{R}}$ is γ_s -simplicial. On the other hand, for i that is non-novel, the

corresponding solid angle q_i is zero by definition. Hence the extreme point geometry in Lemma 6 can be re-expressed in term of solid angles as follows:

Lemma 8. (Novel Words have Positive Solid Angles) *Let $\bar{\mathbf{R}}$ be simplicial and β be separable. Then, word i is a novel word if, and only if, $q_i > 0$.*

We denote the smallest solid angle among the K distinct extreme points by $q_\wedge > 0$. This is a robust condition number of the convex hull formed by the rows of \mathbf{E} and is related to the simplicial constant γ_s of $\bar{\mathbf{R}}$.

In a real-world dataset we have access to only an empirical estimate $\hat{\mathbf{E}}$ of the ideal word co-occurrence matrix \mathbf{E} . If we replace \mathbf{E} with $\hat{\mathbf{E}}$, then the resulting empirical solid angle estimate \hat{q}_i will be very close to the ideal q_i if $\hat{\mathbf{E}}$ is close enough to \mathbf{E} . Then, the solid angles of “outlier” extreme points will be close to 0 while they will be bounded away from zero for the “true” extreme points. One can then hope to correctly identify all K extreme points by *rank-ordering* all empirical solid angle estimates and selecting the K distinct row-vectors that have the largest solid angles. This forms the basis of our proposed algorithm. The problem now boils down to efficiently estimating the solid angles and establishing the asymptotic convergence of the estimates as $M \rightarrow \infty$. We next discuss how random projections can be used to achieve these goals.

2.4.3 Efficient Estimation of Solid Angles using Random Projections

The definition of the normalized solid angle in Eq. (2.3) motivates an efficient algorithm based on *random projections* to estimate it. For convenience, we first rewrite Eq. (2.3) as

$$q_i = \mathbb{E} \left[\mathbb{I} \{ \forall j : \|\mathbf{E}_j - \mathbf{E}_i\| \geq \zeta, \mathbf{E}_i \mathbf{d} \geq \mathbf{E}_j \mathbf{d} \} \right] \quad (2.4)$$

and then propose to estimate it by

$$\hat{q}_i = \frac{1}{P} \sum_{r=1}^P \mathbb{I}(\forall j : \hat{E}_{i,i} + \hat{E}_{j,j} - 2\hat{E}_{i,j} \geq \zeta/2, \hat{\mathbf{E}}_i \mathbf{d}^r > \hat{\mathbf{E}}_j \mathbf{d}^r) \quad (2.5)$$

where $\mathbf{d}^1, \dots, \mathbf{d}^P \in \mathbf{R}^{W \times 1}$ are P iid directions drawn from an isotropic distribution in \mathbf{R}^W . Algorithmically, by Eq. (2.5), we approximate the solid angle q_i at the i -th word (row-vector) by first projecting all the row-vectors onto P iid isotropic random directions and then calculating the fraction of times each row-vector achieves the maximum projection value. It turns out that the condition $\hat{E}_{i,i} + \hat{E}_{j,j} - 2\hat{E}_{i,j} \geq \zeta/2$ is equivalent to $\|\mathbf{E}_i - \mathbf{E}_j\| \geq \zeta$ in terms of its ability to exclude multiple novel words from the same topic and is adopted for its simplicity.⁶

This procedure of taking random projections followed by calculating the number of times a word is a maximizer via Eq. (2.5) provides a consistent estimate of the solid angle in Eq. (2.3) as $M \rightarrow \infty$ and the number of projections P increases. The high-level idea is simple: as P increases, the empirical average in Eq. 2.5 converges to the corresponding expectation. Simultaneously, as M increases, $\hat{\mathbf{E}} \xrightarrow{a.s.} \mathbf{E}$. Overall, the approximation \hat{q}_i proposed in Eq (2.5) using random projections converges to q_i .

This random projections based approach is also computationally efficient for the following reasons. First, it enables us to avoid the explicit construction of the $W \times W$ dimensional matrix $\hat{\mathbf{E}}$: Recall that each column of \mathbf{X} and \mathbf{X}' has no more than $N \ll W$ non-zero entries. Hence \mathbf{X} and \mathbf{X}' are both sparse. Since $\hat{\mathbf{E}}\mathbf{d} = M\bar{\mathbf{X}}'(\bar{\mathbf{X}}^\top \mathbf{d})$, the projection can be calculated using two sparse matrix-vector multiplications. Second, it turns out that the number of projections P needed to guarantee consistency is small. In fact in Theorem 3 we provide a sufficient upper bound for P which is a polynomial function of $\log(W)$, $\log(1/\delta)$ and other model parameters, where δ is the probability that the algorithm fails to detect all the distinct novel words.

⁶We abuse the symbol ζ by using it to indicate different thresholds in these conditions.

Parallelization, Distributed and Online Settings: Another advantage of the proposed random projections based approach is that it can be *parallelized* and is naturally amenable to *online* or *distributed* settings. This is based on the following observation that each projection has an additive structure:

$$\widehat{\mathbf{E}}\mathbf{d}^r = M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top\mathbf{d}^r = M\sum_{m=1}^M\bar{\mathbf{X}}^{m'}\bar{\mathbf{X}}^{m\top}\mathbf{d}^r.$$

The P projections can also be computed independently. Therefore,

- In a *distributed* setting in which the documents are stored on distributed servers, we can first share the same random directions across servers and then aggregate the projection values. The communication cost is only the “partial” projection values and is therefore insignificant Ding et al. [2014b] and does not scale as the number of observations N, M increases.
- In an *online* setting in which the documents are streamed in an online fashion Hoffman et al. [2010], we only need to keep all the projection values and update the projection values (hence the empirical solid angle estimates) when new documents arrive.

The additive and independent structure guarantees that the statistical efficiency of these variations are the same as the centralized “batch” implementation. For the rest of this paper, we only focus on the centralized version.

Outline of Overall Approach: Our overall approach can be summarized as follows. (1) Estimate the empirical solid angles using P iid isotropic random directions as in Eq. 2.5. (2) Select the K words with distinct word co-occurrence patterns (rows) that have the largest empirical solid angles. (3) Estimate the topic matrix using constrained linear regression as in Lemma 4. We will discuss the details of our overall approach in the next section and establish guarantees for its computational and statistical efficiency.

2.5 Algorithm and Analysis

Algorithm 1 describes the main steps of the overall approach. The two main steps, novel word detection and topic matrix estimation are outlined in Algorithms 2 and 3 respectively. Algorithm 2 outlines the random projection and rank-ordering steps. Algorithm 3 describes the constrained linear regression and the renormalization steps in a combined way.

Algorithm 1 Overall-Approach

Input: Text documents $\bar{\mathbf{X}}, \bar{\mathbf{X}}'(W \times M)$; Number of topics K ; Number of iid random projections P ; Tolerance parameters $\zeta, \epsilon > 0$.

Output: Estimate of the topic matrix $\hat{\beta}(W \times K)$.

1: Set of Novel Words $\mathcal{I} \leftarrow \text{NovelWordDetect}(\bar{\mathbf{X}}, \bar{\mathbf{X}}', K, P, \zeta)$

2: $\hat{\beta} \leftarrow \text{EstimateTopics}(\mathcal{I}, \bar{\mathbf{X}}, \bar{\mathbf{X}}', \epsilon)$

Computational Efficiency: We first summarize the computational efficiency of Algorithm 1:

Theorem 2. *Let the number of novel words for each topic be a constant relative to M, W, N . Then, the running time of Algorithm 1 is $\mathcal{O}(MNP + WP + WK^3)$.*

This efficiency is achieved by exploiting the sparsity of \mathbf{X} and the property that there are only a small number of novel words in a typical vocabulary. A detailed analysis of the computational complexity is presented in the appendix. Here we point out that in order to upper bound the computation time of the linear regression in Algorithm 3 we used $\mathbf{O}(WK^3)$ for W matrix inversions, one for each of the words in the vocabulary. In practice, a gradient descent implementation can be used for the constrained linear regression which is much more efficient. We also note that these W optimization problems are decoupled given the set of detected novel words. Therefore, they can be parallelized in a straightforward manner Ding et al. [2014b].

Asymptotic Consistency and Statistical Efficiency: We now summarize the asymptotic consistency and sample complexity bounds for Algorithm 1. The analysis

Algorithm 2 NovelWordDetect (via Random Projections)

Input: $\bar{\mathbf{X}}, \bar{\mathbf{X}}'$; Number of topics K ; Number of projections P ; Tolerance ζ ;

Output: The set of all novel words of K distinct topics \mathcal{I} .

```

1:  $\hat{q}_i \leftarrow 0, \forall i = 1, \dots, W, \hat{\mathbf{E}} \leftarrow M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top$ .
2: for all  $r = 1, \dots, P$  do
3:   Sample  $\mathbf{d}^r \in \mathbb{R}^W$  from an isotropic prior.
4:    $\mathbf{v} \leftarrow M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top \mathbf{d}^r$ 
5:    $i^* \leftarrow \arg \max_{1 \leq i \leq W} \mathbf{v}_i, \hat{q}_{i^*} \leftarrow \hat{q}_{i^*} + 1/P$ 
6:    $\hat{J}_{i^*} \leftarrow \{j : \hat{E}_{i^*,i^*} + \hat{E}_{j,j} - 2\hat{E}_{i^*,j} \geq \zeta/2\}$ 
7:   for all  $k \in \hat{J}_{i^*}^c$  do
8:      $\hat{J}_k \leftarrow \{j : \hat{E}_{k,k} + \hat{E}_{j,j} - 2\hat{E}_{k,j} \geq \zeta/2\}$ 
9:     if  $\{\forall j \in \hat{J}_k, v_k > v_j\}$  then
10:       $\hat{q}_k \leftarrow \hat{q}_k + 1/P$ 
11:     end if
12:   end for
13: end for
14:  $\mathcal{I} \leftarrow \emptyset, k \leftarrow 0, j \leftarrow 1$ 
15: while  $k < K$  do
16:    $i \leftarrow$  index of the  $j^{\text{th}}$  largest value of  $\{\hat{q}_1, \dots, \hat{q}_W\}$ .
17:   if  $\{\forall p \in \mathcal{I}, \hat{E}_{p,p} + \hat{E}_{i,i} - 2\hat{E}_{i,p} \geq \zeta/2\}$  then
18:      $\mathcal{I} \leftarrow \mathcal{I} \cup \{i\}, k \leftarrow k + 1$ 
19:   end if
20:    $j \leftarrow j + 1$ 
21: end while
22: Return  $\mathcal{I}$ .

```

is a combination of the consistency of the novel word detection step (Algorithm 2) and the topic estimation step (Algorithm 3). We state the results for both of these steps. First, for detecting all the novel words of the K distinct topics, we have the following result:

Theorem 3. *Let topic matrix β be separable and $\bar{\mathbf{R}}$ be γ -simplicial. If the projection directions are iid sampled from any isotropic distribution, then Algorithm 2 can identify all the novel words of the K distinct topics as $M, P \rightarrow \infty$. Furthermore, $\forall \delta \geq 0$, if*

$$M \geq 20 \frac{\log(2W/\delta)}{N\rho^2\eta^4} \text{ and } P \geq 8 \frac{\log(2W/\delta)}{q_\wedge^2} \quad (2.6)$$

then Algorithm 2 fails with probability at most δ . The model parameters are defined

Algorithm 3 EstimateTopics

Input: $\mathcal{I} = \{i_1, \dots, i_K\}$ set of novel words, one for each of the K topics; $\widehat{\mathbf{E}}$; precision parameter ϵ

Output: $\widehat{\boldsymbol{\beta}}$, which is the estimate of the $\boldsymbol{\beta}$ matrix

- 1: $\widehat{\mathbf{E}}_w^* = \left[\widehat{\mathbf{E}}_{w,i_1}, \dots, \widehat{\mathbf{E}}_{w,i_K} \right]$
 - 2: $\mathbf{Y} = (\widehat{\mathbf{E}}_{i_1}^{*\top}, \dots, \widehat{\mathbf{E}}_{i_K}^{*\top})^\top$
 - 3: **for all** $i = 1, \dots, W$ **do**
 - 4: Solve $\mathbf{b}^* := \arg \min_{\mathbf{b}} \|\widehat{\mathbf{E}}_i^* - \mathbf{b}\mathbf{Y}\|^2$
 - 5: subject to $b_j \geq 0, \sum_{j=1}^K b_j = 1$
 - 6: using precision ϵ for the stopping-criterion.
 - 7: $\widehat{\boldsymbol{\beta}}_i \leftarrow (\frac{1}{M}\mathbf{X}_i\mathbf{1})\mathbf{b}^*$
 - 8: **end for**
 - 9: $\widehat{\boldsymbol{\beta}} \leftarrow$ column normalize $\widehat{\boldsymbol{\beta}}$
-

as follows. $\rho = \min\{\frac{d}{8}, \frac{\pi d_2 q_\wedge}{4W^{1.5}}\}$ where $d = (1-b)^2\gamma^2/\lambda_{\max}$, $d_2 \triangleq (1-b)\gamma$, λ_{\max} is the maximum eigenvalue of $\bar{\mathbf{R}}$, $b = \max_{j \in \mathcal{C}_0, k} \bar{\beta}_{j,k}$, and \mathcal{C}_0 is the set of non-novel words. Finally, q_\wedge is the minimum solid angle of the extreme points of the convex hull of the rows of \mathbf{E} .

The detailed proof is presented in the appendix. The results in Eq. (2.6) provide a sufficient finite sample complexity bound for novel word detection. The bound is *polynomial* with respect to $M, W, K, N, \log(\delta)$ and other model parameters. The number of projections P that impacts the computational complexity scales as $\log(W)/q_\wedge^2$ in this sufficient bound where q_\wedge can be upper bounded by $1/K$. In practice, we have found that setting $P = \mathcal{O}(K)$ is a good choice Ding et al. [2014b].

We note that the result in Theorem 3 only requires the simplicial condition which is the *minimum* condition required for consistent novel word detection (Lemma 1). This theorem holds true if the topic prior $\bar{\mathbf{R}}$ satisfies stronger conditions such as affine-independence. We also point out that our proof in this paper holds for *any isotropic distribution* on the random projection directions $\mathbf{d}^1, \dots, \mathbf{d}^P$. The previous result in Ding et al. [2014b], however, only applies to some specific isotopic distributions such as the Spherical Gaussian or the uniform distribution in a unit ball. In practice, we

use Spherical Gaussian since sampling from such prior is simple and requires only $\mathcal{O}(W)$ time for generating each random direction.

Next, given the successful detection of the set of novel words for all topics, we have the following result for the accurate estimation of the separable topic matrix β :

Theorem 4. *Let topic matrix β be separable and $\bar{\mathbf{R}}$ be γ_a -affine-independent. Given the successful detection of novel words for all K distinct topics, the output of Algorithm 3 $\hat{\beta} \xrightarrow{p} \beta$ element-wise (up to a column permutation). Specifically, if*

$$M \geq \frac{2560W^2K \log(W^4K/\delta)}{N\gamma_a^2 a_{\min}^2 \eta^4 \epsilon^2} \quad (2.7)$$

then $\forall i, k, \hat{\beta}_{i,k}$ will be ϵ close to $\beta_{i,k}$ with probability at least $1 - \delta$, for any $0 < \epsilon < 1$. η is the same as in Theorem 3. a_{\min} is the minimum value in \mathbf{a} .

We note that the sufficient sample complexity bound in Eq. (2.7) is again polynomial in terms of all the model parameters. Here we only require $\bar{\mathbf{R}}$ to be affine-independent. Combining Theorem 3 and Theorem 4 gives the consistency and sample complexity bounds of our overall approach in Algorithm 1.

Alternatives for Algorithm 3: We note that due to Lemma 4, the topic estimation step in Algorithm 3 can be achieved using the empirical frequencies and is still consistent [Ding et al., 2013b]. We outline this alternative approach in Algorithm 4. It leads to the same computation complexity bounds. A detailed analysis on its sample complexity bounds can be find in [Ding et al., 2013b].

2.6 Distributed Topic Discovery

We now consider the application of our approach in the setting where documents are stored on a network of distributed servers. This is motivated by modern web-scale corpus such as Google online libraries, Twitter Streaming, etc. Due to the distributed nature of the data and limited communication bandwidth between servers, it is crucial to design a distributed topic modeling algorithm with small communication cost.

Algorithm 4 EstimateTopics (Using \mathbf{A})

Input: $\mathcal{I} = \{i_1, \dots, i_K\}$ the set of novel words for K topics; \mathbf{X}, \mathbf{X}' ; precision parameter ϵ

Output: $\hat{\beta}$, which is the estimation of β matrix

$\mathbf{Y} = (\tilde{\mathbf{X}}_{i_1}^\top, \dots, \tilde{\mathbf{X}}_{i_K}^\top)^\top, \mathbf{Y}' = (\tilde{\mathbf{X}}'_{i_1}{}^\top, \dots, \tilde{\mathbf{X}}'_{i_K}{}^\top)^\top$

for all $1 \leq i \leq W$ **do**

Solve $\hat{\beta}_i \leftarrow (\frac{1}{M}\mathbf{X}_i\mathbf{1}) \arg \min_{\mathbf{b}} M(\tilde{\mathbf{X}}_i - \mathbf{b}\mathbf{Y})(\tilde{\mathbf{X}}'_i - \mathbf{b}\mathbf{Y}')^\top$

Subject to $b_j \geq 0, \sum_{j=1}^K b_j = 1$

With precision ϵ

end for

$\hat{\beta} \leftarrow$ column normalize $\hat{\beta}$

This section shows that a distributed implementation of our random projection based approach can *provably* achieve the same statistical performance as the *centralize* counterpart while requiring insignificant communication cost.

2.6.1 Distributed Setting

We consider a collection of M documents that are archived among L servers. For simplicity we consider there are $H = M/L$ documents per server. The *servers* are all connected and there is a *fusion center* that outputs the result. We consider a simple scheme that each server is directly connected to and communicate with the fusion center. But our approach in principle can be applied to all the connected server network. The generative process is the same as in Figure 2-1. We assume that a common vocabulary of size W is shared across the system. An example structure is depicted in Figure 2-6.

For further reference, we denote by $\mathbf{X}^{(l)}$ ($W \times H$) the documents stored on the l -th server. Hence $\mathbf{X}^{(l)}$ is a slice of \mathbf{X} . Similarly, $\bar{\mathbf{X}}'^{(l)}$ and $\bar{\mathbf{X}}^{(l)}$ are obtained by first splitting each document stored on server l into two independent halves and followed by proper row-normalization. So simplicity and to highlight the connection to the centralized version, we require these row-normalization ensure the “global” matrix $\bar{\mathbf{X}}$

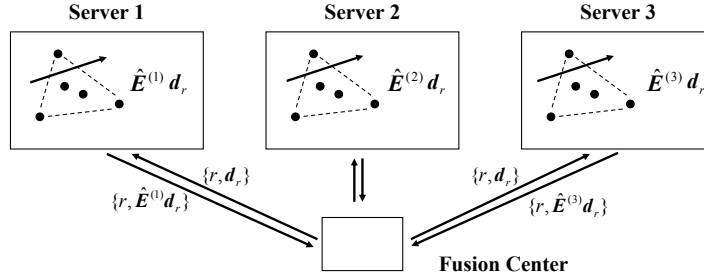


Figure 2.6: The structure of our proposed approach in estimating solid angles from distributed servers. Server 1 to 3 are example distributed servers connected to the Fusion Center. \mathbf{d}^r is an example projection directions that is synchronized across servers. $\hat{\mathbf{E}}^i \mathbf{d}^r$'s are the partial projection values calculated on each server.

and $\bar{\mathbf{X}}'$ are row-stochastic.⁷ We consider the asymptotic setting in which the total number of total documents $M \rightarrow \infty$ and N fixed. This can be achieved by either increase the number of servers L or the number of documents per server H .

2.6.2 Estimating Solid Angles from Distributed Servers

We use the same random projection based approach described in Section 2.4 and 2.5. The geometric property as well as the polynomial efficiency guarantees are exactly the same. Our objective here is to show this approach requires a insignificant communication cost if the collection of M documents are stored on L servers.

We first consider the key calculation step in Algorithm 2, i.e., to compute the projection values $\mathbf{v}_r = \hat{\mathbf{E}} \mathbf{d}^r = M \bar{\mathbf{X}}' \bar{\mathbf{X}}^\top \mathbf{d}^r$ for $r = 1, \dots, P$. We can rewrite it as,

$$\mathbf{v}_r = \hat{\mathbf{E}} \mathbf{d}^r = M \bar{\mathbf{X}}' \bar{\mathbf{X}}^\top \mathbf{d}^r = M \sum_{l=1}^L \bar{\mathbf{X}}'^{(l)} \bar{\mathbf{X}}^{(l)\top} \mathbf{d}^r := M \sum_{l=1}^L \mathbf{v}_{l,r} \quad (2.8)$$

based on the *additive structure* of second order co-occurrence and projections. Therefore, the projection values $\mathbf{v}_r \in \mathbb{R}^{W \times 1}$ along direction \mathbf{d}_r can be decomposed as the summation of “partial projection values” $\mathbf{v}_{l,r} = \bar{\mathbf{X}}'^{(l)} \bar{\mathbf{X}}^{(l)\top} \mathbf{d}^r$. These $W \times 1$ dimen-

⁷We can also normalize the rows of the empirical word-frequency matrices “locally” for each server and it does not effect the overall results.

sional partial projection values for L servers can be calculated locally, and the servers can communicate only these partial projection values $\mathbf{v}_{l,r}$. In sum, $P W \times 1$ projection values ($\mathcal{O}(WP)$ real numbers) need to be transmitted.

In addition, as we have discussed in Section 2.5, the condition $\hat{J}_i \leftarrow \{j : \hat{E}_{i,i} + \hat{E}_{j,j} - 2\hat{E}_{i,j} \geq \zeta/2\}$ in Eq. (2.5) can be calculated as if we project along $\mathbf{d} = \mathbf{e}_i$ which is the i -th standard base. Hence we can use the same partial decomposition trick as above and requires $\mathcal{O}(WP)$ real numbers to be transmitted. Third, we require all the L servers to use the same set of projection directions $\mathbf{d}^1, \dots, \mathbf{d}^P$. This requires $\mathcal{O}(WP)$ real numbers to be transmitted. In practice, one can use the same seed for a pseudo-random generator across the servers.

Overall, the communication cost is $\mathcal{O}(WP)$ real numbers for each server which does not scale with the sample size M, N and is relatively small. At the same time, the matrix-vector multiplications are parallelized on each individual server. Hence the computation time required is much smaller than the centralized counterpart. Abstractly, the overall procedure is as follows : (i) All the L servers are synchronized with the same set of projection directions $\mathbf{d}^r, r = 1, \dots, P$. (ii) Local projection values are calculated on each server and are transmitted to the fusion center. (iii) The fusion center executes the remaining steps as in Algorithm 2. The interaction structure is depicted in Figure 2.6.

We briefly discuss the remaining steps of our algorithm. We note that as shown in Algorithm 3, we only need the sub-collection of columns in \mathbf{E} corresponding to the extreme points for the constrained linear regression. There $\mathcal{O}(W \times K)$ values have been already computed and collected as projection onto specify \mathbf{e}_w 's as one steps in the random projections. Therefore, we do not need extra communication to conduct such regressions if the work is to be conducted in the fusion center. Alternatively, we can conduct these steps in parallel on the distributed servers since the W constrained

regressions are independent. Therefore, we can parallel the computation across all the servers with the same amount of communication cost.

Besides the scheme described above, another distributed implementation is proposed in [Ding et al., 2014b]. Instead of transmitting the partial projection values, one can transmit the index of word that achieves the maximum projection value based on the word co-occurrence representation estimated using the local documents on each servers. Provable guarantees can be established as $H \rightarrow \infty$.

Online Setting: In online algorithm setting, the entire set of documents are not available but are observed by algorithm piece-by-piece in a streaming fashion. This can be viewed as a variation of the distributed setting discussed above. Therefore, we (1) fix a collection of P projection directions $\mathbf{d}^1, \dots, \mathbf{d}^P$, (2) update the overall projection values after observing a new mini-batch of documents, and update the estimated solid angles, (3) perform the remaining steps. In contrast to the popular online VB or MCMC algorithms, our approach can provably achieve the same statistical guarantees as summarized in Theorem 3 and 4.

2.6.3 Analysis

We summarize the discussion on communication costs in the following Theorem:

Theorem 5. *Let topic matrix β be separable and topic prior be γ -simplicial. Let the M documents be stored evenly on L servers. The **distributed** implementation of Algorithm 2 using partial projection value decomposition has*

- a (Statistical Efficiency) the same asymptotic consistency and the same sample complexity bounds as in Theorem 3 and Theorem 4,*
- b (Low Communication Cost) a communication cost of $\mathcal{O}(WP)$ real numbers per server,*
- c (Parallelized Running Time) a $\mathcal{O}(MNP/L + W)$ running time per server, and a $\mathcal{O}(WPL + K^2)$ running time for the fusion center.*

2.6.4 Related Works in Distributed Topic Modeling

The distributed variations of the topic modeling algorithms have been studied based on different centralized counterpart. Base on the centralized MCMC or VB approaches [Blei, 2012, Griffiths and Steyvers, 2004], Collapsed Gibbs Sampling, online variational methods, and many other variations have been used to parallelize the estimation and inference [Asuncion et al., 2009, Newman et al., 2009, Smola and Narayanamurthy, 2010]. An alternative approach is to distribute the NMF with appropriate regularization [Gemulla et al., 2011, Liu et al., 2010]. These distributed approaches can empirically approximate the performance of their centralized counterparts.

Another possible direction is to attempt to parallelize the existing algorithms that come with computational and statistical guarantees as discussed in Section 2.2. However, it is unclear how to directly parallelize these approaches. [Arora et al., 2013] assumes the separability condition. The key step is based on Gram-Schmidt process over the rows of normalized empirical word co-occurrence matrix. The Gram-Schmidt process is inherently sequential and is hard to parallelize this key step with small communication cost. [Kumar et al., 2013] is based on a similar idea. [Ding et al., 2013b] proposed a data-dependent projection scheme and it requires to scan through all the entries of the word co-occurrence. Other than these separability based method, the distributed implementation of the tensor decomposition approach in [Anandkumar et al., 2012] is also unclear.

2.7 Empirical Results

In this section, we present experiment results over a wide range of datasets and measures. When the ground truth is available in Section 2.7.1, we compute the ℓ_1 *reconstruction error* between the ground truth topics and the estimates. For the real-

world text corpus, we report the *held-out probability* as a standard measure in topic modeling literature. We also *qualitatively* compare the semantic topics extracted by our approaches using the top probable words for each topics. We use the random projection based Algorithm 1 (denoted by RP) and its distribution variation in Section 2.6 (denoted by RP(distributed)). We note that we simulate the distributed variation only in software.

2.7.1 Semi-Synthetic Dataset

In order to validate our proposed algorithm, we generate “semi-synthetic” text corpora by sampling from a synthetic, yet realistic, ground truth topic model. To ensure that the semi-synthetic data is similar to real-world data, in terms of dimensionality, sparsity, and other characteristics, we use the following generative procedure adapted from Arora et al. [2013], Ding et al. [2014b].

We first train an LDA model (with $K = 100$) on a real-world dataset using a standard Gibbs Sampling method with default parameters (as described in Griffiths and Steyvers [2004], McCallum [2002]) to obtain a topic matrix β_0 of size $W \times K$. The real-world dataset that we use to generate our synthetic data is derived from a New York Times (NYT) articles dataset Bache and Lichman [2013]. The original vocabulary is first pruned based on document frequencies. Specifically, as is standard practice, only words that appear in more than 500 documents are retained. Thereafter, again as per standard practice, the words in the so-called stop-word list are deleted as recommended in Lewis et al. [2004]. After these steps, $M = 300,000$, $W = 14,943$, and the average document length $N = 298$. We then generate semi-synthetic datasets, for various values of M , by fixing $N = 300$ and using β_0 and a Dirichlet topic prior. As suggested in Griffiths and Steyvers [2004] and used in Arora et al. [2013], Ding et al. [2014b], we use symmetric hyper-parameters (0.03) for the Dirichlet topic prior.

The $W \times K$ topic matrix β_0 may not be separable. To enforce separability, we create a new *separable* $(W + K) \times K$ dimensional topic matrix β_{sep} by inserting K synthetic novel words (one per topic) having suitable probabilities in each topic. Specifically, β_{sep} is constructed by transforming β_0 as follows. First, for each synthetic novel word in β_{sep} , the value of the sole nonzero entry in its row is set to the probability of the most probable word in the topic (column) of β_0 for which it is a novel word. Then the resulting $(W + K) \times K$ dimensional nonnegative matrix is renormalized column-wise to make it column-stochastic. Finally, we generate semi-synthetic datasets, for various values of M , by fixing $N = 300$ and using β_{sep} and the same symmetric Dirichlet topic prior used for β_0 .

We use the name *Semi-Syn* to refer to datasets that are generated using β_0 and the name *Semi-Syn+Novel* for datasets generated using β_{sep} .

In our proposed random projections based algorithm, which we call RP, we set $P = 150 \times K$, $\zeta = 0.05$, and $\epsilon = 10^{-4}$. We compare RP against the provably efficient algorithm RecoverL2 in Arora et al. [2013] and the standard Gibbs Sampling based LDA algorithm (denoted by Gibbs) in Griffiths and Steyvers [2004], McCallum [2002]. In order to measure the performance of different algorithms in our experiments based on semi-synthetic data, we compute the ℓ_1 norm of the *reconstruction error* between $\hat{\beta}$ and β . Since all column permutations of a given topic matrix correspond to the same topic model (for a corresponding permutation of the topic mixing weights), we use a bipartite graph matching algorithm to optimally match the columns of $\hat{\beta}$ with those of β (based on minimizing the sum of ℓ_1 distances between all pairs of matching columns) before computing the ℓ_1 norm of the reconstruction error between $\hat{\beta}$ and β .

The results on both *Semi-Syn+Novel* NYT and *Semi-Syn* NYT are summarized in Fig. 2-7 for all three algorithms for various choices of the number of documents M . We note that in these figures the ℓ_1 norm of the error has been normalized by

the number of topics ($K = 100$).

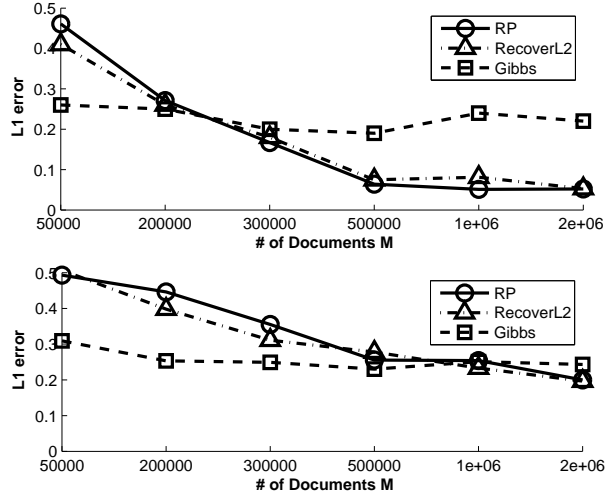


Figure 2.7: ℓ_1 norm of the error in estimating the topic matrix β for various M ($K = 100$): (Top) *Semi-Syn+Novel* NYT; (Bottom) *Semi-Syn* NYT. RP is the proposed algorithm, RecoverL2 is a provably efficient algorithm from Arora et al. [2013], and Gibbs is the Gibbs Sampling approximation algorithm in Griffiths and Steyvers [2004]. In RP, $P = 150K$, $\zeta = 0.05$, and $\epsilon = 10^{-4}$.

As Fig. 2.7 shows, when the separability condition is strictly satisfied (*Semi-Syn+Novel*), the reconstruction error of RP converges to 0 as M becomes large and outperforms the approximation-based Gibbs. When the separability condition is not strictly satisfied (*Semi-Syn*), the reconstruction error of RP is comparable to Gibbs (a practical benchmark).

We further consider the *distributed* setting discussed in Section 2.6 and the distributed implementation - RP(distributed). We simulate $L = 200$ distributed servers. The computation cost for the RP(distributed) is reported as the computation time per server+computation time for the fusion center. We plot the ℓ_1 error vs computation cost for RP(distrusted), DDP (algorithm proposed in [Ding et al., 2013b]) and RecoverL2 in Figure 2.8 (left) for different number of documents M . This plot fully depicts the merits of our approach, i.e., it can achieve the same level of statisti-

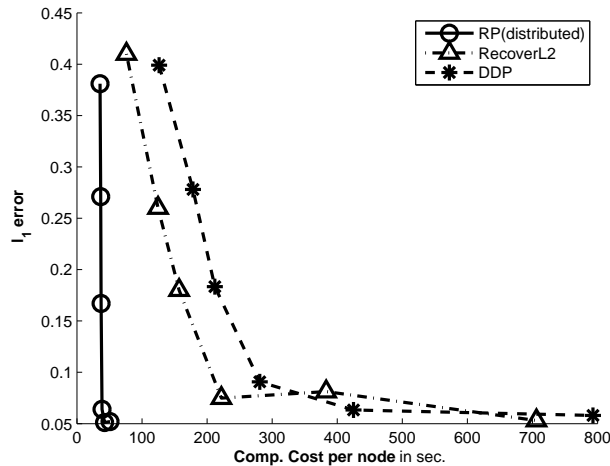


Figure 2-8: Computation cost vs. ℓ_1 reconstruction error on *Semi-Syn+Novel* NYT dataset for $M = 50k, 200k, 300k, 500k, 1m, 2m, L = 200$. RP(distributed) are compared against RecoverL2 in [Arora et al., 2013], DDP in [Ding et al., 2013b]. RecoverL2(distributed) and DDP(distributed) are naive distributed implementations that first estimate topics locally and then average across servers. $P = 150 \times K$ and L parallel threads are simulated for *centralized* RecoverL2 and DDP in Regression in both case. k =thousand. m =million.

cal accuracy with much lower computation time, when compared to the centralized state-of-the-art. We point out that the C-implementation we used for Gibbs Sampling requires 6918 sec. in estimation with 100 iterations for $M = 300,000$. It is much longer than the the range of computation time reported in Figure 2-8.

Solid Angle and Model Selection: In our proposed algorithm RP, the number of topics K (the model-order) needs to be specified. When K is unavailable, it needs to be estimated from the data. Although not the focus of this work, Algorithm 2, which identifies novel words by sorting and clustering the estimated solid angles of words, can be suitably modified to estimate K .

Indeed, in the ideal scenario where there is no sampling noise ($M = \infty, \hat{\mathbf{E}} = \mathbf{E}$, and $\forall i, \hat{q}_i = q_i$), only novel words have positive solid angles (\hat{q}_i 's) and the rows of $\hat{\mathbf{E}}$ corresponding to the novel words of the same topic are identical, i.e., the distance

between the rows is zero or, equivalently, they are within a neighborhood of size zero of each other. Thus, the number of distinct neighborhoods of size zero among the non-zero solid angle words equals K .

In the nonideal case M is finite. If M is sufficiently large, one can expect that the estimated solid angles of non-novel words will not all be zero. They are, however, likely to be much smaller than those of novel words. Thus to reliably estimate K one should not only exclude words with exactly zero solid angle estimates, but also those above some nonzero threshold. When M is finite, the the rows of $\hat{\mathbf{E}}$ corresponding to the novel words of the same topic are unlikely to be identical, but if M is sufficiently large they are likely to be close to each other. Thus, if the threshold ζ in Algorithm 2, which determines the size of the neighborhood for clustering all novel words belonging to the same topic, is made sufficiently small, then each neighborhood will have only novel words belonging to the same topic.

With the two modifications discussed above, the number of distinct neighborhoods of a suitably nonzero size (determined by $\zeta > 0$) among the words whose solid angle estimates are larger than some threshold $\tau > 0$ will provide an estimate of K . The values of τ and ζ should, in principle, decrease to zero as M increases to infinity. Leaving the task of unraveling the dependence of τ and ζ on M to future work, here we only provide a brief empirical validation on both the *Semi-Syn+Novel* and *Semi-Syn* NYT datasets. We set $M = 2,000,000$ so that the reconstruction error has essentially converged (see Fig. 2.7), and consider different choices of the threshold ζ .

We run Algorithm 2 with $K = 100$, $P = 150 \times K$, and a new line of code: 16': (**if** $\{\hat{q}_i = 0\}$, **break**); inserted between lines 16 and 17 (this corresponds to $\tau = 0$). The input hyperparameter $K = 100$ is not the actual number of estimated topics. It should be interpreted as specifying an upper bound on the number of topics. The value of (little) k when Algorithm 2 terminates (see lines 14–21) provides an estimate

of the number of topics.

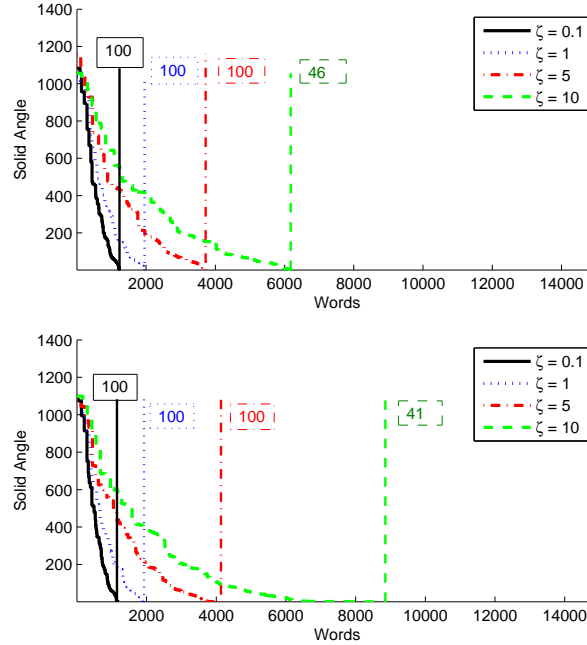


Figure 2.9: Solid-angles (in descending order) of all $14943+100$ words in the *Semi-Syn+Sep* NYT dataset (left) and all 14943 words in the *Semi-Syn* NYT dataset (right) estimated (for different values of ζ) by Algorithm 2 with $K = 100$, $P = 150 \times K$, $M = 2,000,000$, and a new line of code: 16': (**if** $\{\hat{q}_i = 0\}$, **break**); inserted between lines 16 and 17. The values of j and (little) k when Algorithm 2 terminates are indicated, respectively, by the position of the vertical dashed line and the rectangular box next to it for different ζ .

Figure 2.9 illustrates how the solid angles of all words, sorted in descending order, decay for different choices of ζ and how they can be used to detect the novel words and estimate the value of K . We note that in both the semi-synthetic datasets, for a wide range of values of ζ (0.1–5), the modified Algorithm 2 correctly estimates the value of K as 100. When ζ is large (e.g., $\zeta = 10$ in Fig. 2.9), many interior points would be declared as novel words and multiple ideal novel words would be grouped into one cluster resulting. This causes K to be underestimated (46 and 41 in Fig. 2.9).

2.7.2 Real World Text Corpus

We now describe results on the actual real-world NYT dataset that was used in Sec. 2.7.1 to construct the semi-synthetic datasets. Since ground truth topics are unavailable, we measure performance using the so-called *predictive held-out log-probability*. This is a standard measure which is typically used to evaluate how well a learned topic model fits real-world data. To calculate this for each of the three topic estimation methods (Gibbs Griffiths and Steyvers [2004], McCallum [2002], RecoverL2 Arora et al. [2013], and RP), we first randomly select 60,000 documents to test the goodness of fit and use the remaining 240,000 documents to produce an estimate $\hat{\beta}$ of the topic matrix. Next we assume a Dirichlet prior on the topics and estimate its concentration hyper-parameter α . In Gibbs, this estimate $\hat{\alpha}$ is a byproduct of the algorithm. In RecoverL2 and RP this can be estimated from $\hat{\beta}$ and \mathbf{X} . We then calculate the probability of observing the test documents given the learned topic model $\hat{\beta}$ and $\hat{\alpha}$:

$$\log \Pr(\mathbf{X}_{\text{test}} | \hat{\beta}, \hat{\alpha})$$

Since an exact evaluation of this predictive log-likelihood is intractable in general, we calculate it using the MCMC based approximation proposed in Wallach et al. [2009] which is now a standard approximation tool McCallum [2002]. For RP, we use $P = 150 \times K$, $\zeta = 0.05$, and $\epsilon = 10^{-4}$ as in Sec. 2.7.1. We report the held-out log probability, normalized by the total number of words in the test documents, averaged across 5 training/testing splits. The results are summarized in Table 2.1. As shown in Table 2.1, Gibbs has the best descriptive power for new documents. RP and RecoverL2 have similar, but somewhat lower values than Gibbs. This may be attributed to missing novel words that appear only in the test set and are crucial to the success of RecoverL2 and RP. Specifically, in real-world examples, there is a model-mismatch as a result of which the data likelihoods of RP and RecoverL2 suffer.

Table 2.1: Normalized held-out log probability of RP, RecoverL2, and Gibbs Sampling on NYT test data. The Mean \pm STD’s are calculated from 5 different random training-testing splits.

K	RecoverL2	Gibbs	RP
50	-8.22 \pm 0.56	-7.42 \pm 0.45	-8.54 \pm 0.52
100	-7.63 \pm 0.52	-7.50 \pm 0.47	-7.45 \pm 0.51
150	-8.03 \pm 0.38	-7.31 \pm 0.41	-7.84 \pm 0.48
200	-7.85 \pm 0.40	-7.34 \pm 0.44	-7.69 \pm 0.42

Finally, we *qualitatively* assess the topics produced by our RP algorithm. We show some example topics extracted by RP trained on the *entire* NYT dataset of $M = 300,000$ documents in Table 2.2 ⁸ For each topic, its most frequent words are

Table 2.2: Examples of topics estimated by RP on NYT

Topic label	Words in decreasing order of estimated probabilities
“weather”	weather wind air storm rain cold
“feeling”	feeling sense love character heart emotion
“election”	election zzz_florida ballot vote zzz_al_gore recount
“game”	yard game team season play zzz_nfl

listed. As can be seen, the estimated topics do form recognizable themes that can be assigned meaningful labels. The full list of all $K = 100$ topics estimated on the NYT dataset can be found in Ding et al. [2013b]. In [Ding et al., 2013b] we also provide a comparison of all the estimated topics produced by RP, RecoverL2, Gibbs on another dataset consists of all articles from NIPS conference.

⁸The zzz prefix in the NYT vocabulary is used to annotate certain special named entities. For example, zzz_nfl annotates NFL.

Chapter 3

Mixed Membership Ranking Models for Pairwise Comparisons

Partial rankings of items generated by a large user-population can now be observed and recorded over the web through transactions, reviews, check-ins and browsing history such as products from Amazon, businesses from Yelp, and movies from Netflix. The problem of predicting preference behavior for a diverse population is important in many applications including recommendation systems, e-commerce and information retrieval [e.g., Awasthi et al., 2014, Ding et al., 2015b,c, Gormley et al., 2009, Kim et al., 2014, Lu and Boutilier, 2014, Oh and Shah, 2014, Volkovs and Zemel, 2014]. This chapter demonstrate how MMLVMs can be used to model, learn, and ultimately predict the preference behavior of users in the form of pairwise comparisons.

The key contribution of this chapter is two-folded. *First*, from the modeling perspective, we propose a novel family of MMLVMs. To our best knowledge, it is the first work that systematically investigates MMLVMs in partial ranking observations. *Second*, we identify the corresponding separability property and propose provably consistent and efficient learning algorithms. As a by-product, we obtain the first provably consistent and efficient results for learning permutation-mixture [Ding et al., 2015b] and Mallows-mixture models [Ding et al., 2015c] from pairwise comparisons.

3.1 Motivating Example and Generative Framework

We propose two MMLVMs for pairwise comparisons that accounts for a *heterogeneous* population of *inconsistent and noisy* users. The key idea is to view the outcomes of comparisons of each user arises as a probabilistic mixture of a few latent ranking factors that are shared across the population [Ding et al., 2014a, 2015b,c]. This is especially appealing in the context of emerging web-scale applications where (i) there are multiple factors that influence individual preference behavior, (ii) each individual is influenced by multiple latent factors to different extents, (iii) the same latent factor can consistently result in different outcomes on different users, more so for similar items, and (iv) the number of comparisons available from each user is typically limited.

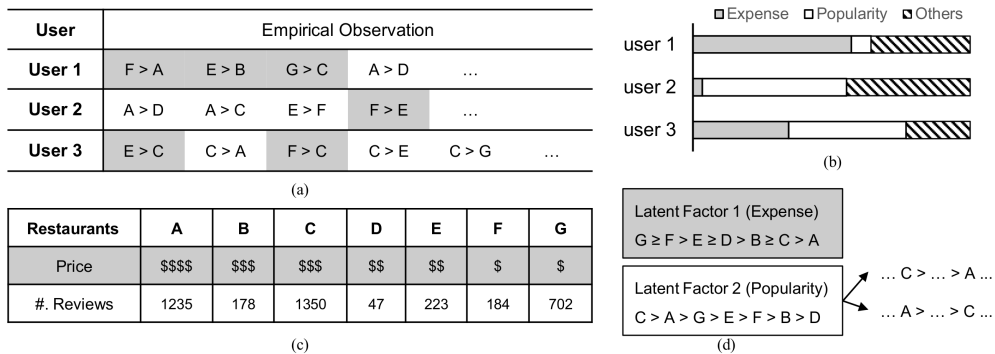


Figure 3.1: An illustration of how proposed MMLVMs can model noisy preferences of heterogeneous and inconsistent users. Say a set of ratings from Yelp for restaurants are obtained and anonymized from a local area (subplot (c)). Two example latent factors, “expense” and “popularity” (subplot (d)), influence the three users’ behavior (subplot (a)), with different weights (subplot (b)). This models heterogeneity. $A > B$ means A is preferred over B . The shading in subplot (a) indicates the most-likely influencing factor of each observation using the same color coding as in other subplots. This accounts for inconsistency. A and C are very close in “Popularity” and both $C > A$ and $A > C$ are possible when influenced by the same “Popularity” factor which accounts for noise. Comparisons can be observed or inferred from check-in, browsing history, GPS record, etc.

For instance consider an example when comparing the restaurants in a local area on Yelp as in Figure 3.1, “Expense” and “Popularity” are two example ranking factors

that can influence a typical user-population. Each of these factors imposes distinct preferences over the restaurants. Each user has her own importance weights over the expense, popularity, and other factors, (subplot (b)) and her comparisons are results of a mixture of these factors (subplot (a)). It is inadequate either to aggregate a global ranking for the population [Shah et al., 2015, Volkovs and Zemel, 2014] or to cluster the users into heterogeneous types and assume users within each cluster are similar [Awasthi et al., 2014, Oh and Shah, 2014]. We note that the contextual information such as prices in Figure 3.1 is not part of our input.

Problem Setup: We consider a universe of Q items $\mathcal{U} = \{1, \dots, Q\}$, and a population of M users that each compares N pairs of items. A comparison is denoted by an ordered pair $w_{m,n} = (i, j)$ if in the n -th comparison of user m , she considers items i, j and prefers i over j . $W = Q(Q - 1)$ is the number of all possible pairwise comparison results. We assume the pairs to be compared are sampled from some distribution μ on all pairs with $\mu_{i,j} = \mu_{j,i} > 0$ being the probability of comparing item i, j . We represent the empirical observations using a $W \times K$ dimensional matrix \mathbf{X} . $X_{(i,j),m}$ denotes the number of times that user m compares item i, j and prefers i over j .

Ranking Components and Ranking Matrix: We define a **ranking component** β^k to be a probabilistic model on partial rankings which plays the same role as the “topic” in topic modeling. In the context of pairwise comparisons it defines for each pair of item i, j the probability of i being preferred over j , if the two items are to be compared by some user. We denote this probability by $\beta_{(i,j),k} \geq 0$. The set of parameters $\{\beta_{(i,j),k}\}_{i \neq j}$ define a ranking component in pairwise comparisons. We denote by β a $W \times K$ dimensional **ranking matrix** whose W rows are indexed by all ordered pairs and $\beta_{(i,j),k}$ as defined above. The pairwise ranking matrix β can be defined using various probabilistic ranking models. For instance,

1. A ranking component β^k can be modeled as a total ranking over the Q items, denoted by a permutation σ^k . Conditioned on each component, the pairwise

comparisons are deterministic. $\beta_{(i,j),k} = \mathbb{I}(\sigma^k(i) < \sigma^k(j))$.¹ It is exploited in [Ding et al., 2014a, 2015b] and will be discussed in Section 3.3.

2. A ranking component β^k can be modeled as a Mallows distribution (parameterized by reference ranking σ^k and dispersion ϕ_k , see Eq. (3.6)). This is used in [Awasthi et al., 2014, Ding et al., 2015c] and will be discussed in Section 3.4.
3. A ranking component β^k can be modeled as the Bradley-Terry-Luce [Shah et al., 2015] with a set of score parameters $w_1^k, \dots, w_Q^k \geq 0$. For each pair of items i, j , $\beta_{(i,j),k} = \frac{w_i^k}{w_i^k + w_j^k}$. This is used in [Kim et al., 2014, Oh and Shah, 2014]

As we shall see next in Section 3.3 and 3.4, this ranking matrix allows us to associate the proposed mixed membership ranking models with a statistically equivalent topic model whose topic matrix provides an information-equivalent representation of the parameters of the ranking components.

Mixed Membership Ranking models: We posit K distinct latent ranking components β^1, \dots, β^K that are shared by the population of M users. The key idea is to model the comparisons made by each user as a *probabilistic mixture* of the K latent ranking components. For each user $m = 1, \dots, M$ is,

1. Sample $\theta^m \in \Delta^K$ from a prior distribution $\Pr(\theta)$
2. For each comparison $n = 1, \dots, N$:
 - (a) Sample a pair of items $\{i, j\}$ from μ
 - (b) Sample a ranking token $z_{m,n} \in \{1, \dots, K\} \sim \text{Multinomial}(\theta^m)$
 - (c) Sample $w_{m,n}$ from latent ranking component $\sigma^{z_{m,n}}$.

On a high-level, β^1, \dots, β^K capture the prevalent latent ranking factors in the population. The K dimensional probabilistic vector θ^m are the weights of each user

¹ $\sigma^k(i)$ is the position of the item i in σ^k and item i is preferred over j if $\sigma^k(i) < \sigma^k(j)$.

over the shared rankings. They capture the degree of influence of each ranking component on each user. Similar as in Chapter 2, our approach can be applied to a general family of prior distribution $\Pr(\theta)$ which satisfies some minimum technical conditions.

Overall Approach: Our approach is to view each comparison as “words”, the comparisons of a user as “document”, and the latent ranking factors as “topics”. This allows us to draw a formal statistical equivalency between the proposed mixed membership ranking model and a standard topic model. Therefore, any approach that is developed for topic modeling can be applied. We then identify the (approximate) separability in the proposed generative models. This allows us to apply the similar approach as we developed in Chapter 2, and establish asymptotic consistency and efficiency results. We also identify that when Q scales sufficiently faster than K , the (approximate) separability property is satisfied with high probability.

Learning Problem: We focus also on the estimation problem in this chapter. To be explicit, given \mathbf{X} and K , our goal is to learn the parameters of the shared latent ranking components as well as the parameters for the corresponding “topic priors”.

Organization: For the rest of this chapter, we first discuss the closely related works in section 3.2. We then discuss two specific models for the ranking components in section 3.3 and section 3.4. We analysis the separability structure, the corresponding ranking matrix, and the provable guarantees for each models. We demonstrate their empirical performance in predicting real-world movie comparisons in Ssction 3.5.

3.2 Related Works

First and for most, our proposed mixed membership models for pairwise comparisons take a *decomposition* perspective: to decompose each users’ comparisons as a mixture of a small number of common factors. This is fundamentally different from the *clustering* perspective in the popular mixture of ranking models in the literature.

Rank estimation from partial or full preference observations have been extensively

studied for several decades in various settings since the seminar works in [e.g., Mallows, 1957, Thurstone, 1927, Zermelo, 1929]. In the literature, we can identify two major categories of models. In the first category of models, the individual user rankings are modeled as independent drawings from a probability distribution which is centered around a single ground-truth global ranking. Efficient algorithms have been developed to estimate one global ranking that “optimally” agrees with the observations based on corresponding metric induced by the distribution [e.g., Ost, 2013, Gleich and Lim, 2011, Marden, 1995, Negahban et al., 2012, Qin et al., 2010, Rajkumar and Agarwal, 2014, Volkovs and Zemel, 2014]. Loosely speaking, this tacitly presupposes a fairly *homogeneous* population of users having very similar preferences. Chief among them are the permutation based Mallows model [Mallows, 1957], the random-utility theory based Plackett-Luce (PL) model [Plackett, 1975], and the score based Bradley-Terry-Luce (BTL) model [Rajkumar and Agarwal, 2014, Shah et al., 2015].

The second category is the family of mixture of ranking models: on the population level there are multiple distinct ranking components, and each user is associated with a *single* ranking scheme sampled from the mixture. Loosely speaking, this tacitly presupposes a *heterogeneous* population of users that can be clustered into different types by their preferences. Therefore each user is primarily influenced by only one factor. For example, in the popular *mixture of Mallows models* [Awasthi et al., 2014, Busse et al., 2007, Lebanon and Lafferty, 2002, Lu and Boutilier, 2014, Meila and Chen, 2010], each ranking component is Mallows [1957]. EM-based algorithms have been proposed from the observations in the form of pairwise comparisons [Lu and Boutilier, 2014], partial rankings [Lebanon and Lafferty, 2002], and full rankings Busse et al. [2007]. Recently, [Awasthi et al., 2014] proposed a provably correct algorithm based on tensor decomposition that can handle a mixture of 2 Mallows model using the top-3 ranked items as the observations which, in effect,

requires users to consider all items. This is impractical within the context of the target web-scale applications. Similarly, a subsequent work leverages the PL model into the mixture setting [Azari Soufiani et al., 2013], as well as the mixture of BTL models [Oh and Shah, 2014]. We note that in recent work of [Lu and Negahban, 2014, Oh and Shah, 2014], a provable algorithm have been developed to estimate shared BTL components using pairwise comparisons by a tensor decomposition method. [Farias et al., 2009, Jagabathula and Shah, 2008] considered using single permutations for each ranking component. They proposed combination algorithms that can learn the mixture model with consistency guarantees. Nevertheless, this consistence requires a property that is equivalent to the exact separability.

Table 3.1: Comparison to closely related works. DIS15c [Ding et al., 2015c], DIS15b [Ding et al., 2015b], GM09 [Gormley et al., 2009], KKS14 [Kim et al., 2014], FJS09[Farias et al., 2009], LB14[Lu and Boutilier, 2014], ABSV14[Awasthi et al., 2014], OS14 [Oh and Shah, 2014].

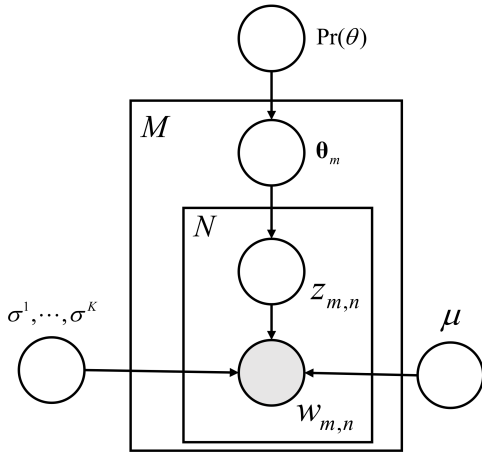
Method	Obs. type	Ranking component	Prior Distrib.	Consistency result	Computation complexity
DIS15c	pairwise	Mallows	general	provable	polynomial
DIS15b	pairwise	permutation	general	provable	polynomial
GM09	Full	Plackett-Luce	Dirichlet	not available	not available
KKS14	pairwise	BTL	Dirichlet	not available	not available
FJS09	pairwise	permutation	mixture	provable	combinatorial
LB14	pairwise	Mallows	mixture	not available	not available
ABSV14	top-3 rank	Mallows	mixture	provable	polynomial
OS14	pairwise	BTL	mixture	provable	polynomial

Our proposed models forms a third category of mixed membership ranking models. Similar to our models, [Gormley and Murphy, 2008, Gormley et al., 2009] proposed a mixed membership model by leveraging the PL model as latent ranking components. Their model was motivated by political science applications and considered full- rankings as observations. MCMC based algorithms are used in [Gormley et al., 2009] and it is not clear how the algorithm would scale for moderately large Q on the order of hundreds. [Kim et al., 2014] proposed recently a mixed membership model using BTL as latent component targeting at web-scale applications as our approaches.

They proposed to use variational methods and assume $\Pr(\theta)$ to be Dirichlet. Table 3.1 summarizes the closely related mixture of ranking models.

Rating based methods: Considerable work in preference prediction has focused on modeling numerical star ratings as is common in modern personalized recommendation and reviewing systems [e.g., Ricci et al., 2011, Salakhutdinov and Mnih, 2008a]. Although coming from a different feature space, our model shares the same mixed membership modeling perspective – the star ratings of each user is modeled as being influenced by a small number of latent factors shared by the population. We also note that an emerging trend in literature explores the idea of combining a topic model for text reviews simultaneously with a rating-based model for “star ratings” [Wang and Blei, 2011]. These approaches are, however, outside the scope of this thesis.

3.3 Topical Ranking Model



For each user $m = 1, \dots, M$,

- 1) Sample ranking weight $\theta^m \in \Delta^K$ from some prior distribution $\Pr(\theta)$
- 2) For each comparison $n = 1, \dots, N$,
 - a Sample a pair of items $\{i, j\}$ from μ ;
 - b Sample a ranking token $z \in \{1, \dots, K\} \sim \text{Multinomial}(\theta^m)$
 - d Output comparison $w_{m,n} = (i, j)$ if $\sigma^{z_{m,n}}(i) < \sigma^{z_{m,n}}(j)$; Otherwise $w_{m,n} = (j, i)$

Figure 3.2: Generative process of the Topical Ranking Model and its Graphical representation. The boxes represent replicates (the outer as users, and the inner as comparisons).

In this section, we model the K latent ranking components as permutations over the Q items. We denote these permutations as $\sigma^1, \dots, \sigma^K$, and define the $W \times K$ ranking matrix β as a binary matrix where $\beta_{(i,j),k} = \mathbb{I}(\sigma^k(i) < \sigma^k(j))$. The

corresponding generative procedure is summarize in Figure 3·2[Ding et al., 2014a, 2015b]. We refer to this generative model as Topical Ranking Model.

The generative procedure in Figure 3·2 induces the following conditional probabilities on observed comparisons $w_{m,n}$:

$$p(w_{m,n} = (i, j) | \boldsymbol{\theta}^m, \mu, \boldsymbol{\beta}) = \mu_{i,j} \sum_{k=1}^K \beta_{(i,j),k} \theta_{k,m} = \sum_{k=1}^K B_{(i,j),k} \theta_{k,m} \quad (3.1)$$

where $B_{(i,j),k} = \beta_{(i,j),k} \mu_{i,j}$. More than convenience, this $W \times K$ matrix \mathbf{B} is an information-equivalent representation of $\boldsymbol{\beta}$ and μ since,

$$\beta_{(i,j),k} = \frac{\beta_{(i,j),k} \mu_{i,j}}{(\beta_{(i,j),k} + \beta_{(j,i),k}) \mu_{i,j}} = \frac{B_{(i,j),k}}{B_{(i,j),k} + B_{(j,i),k}}, \quad \mu_{i,j} = B_{(i,j),k} + B_{(j,i),k} \quad (3.2)$$

Therefore, we can estimate $\boldsymbol{\beta}$ directly from \mathbf{B} .² In addition, \mathbf{B} is (1) column-stochastic, and (2) separable (Definition 1 or Definition 2) iff $\boldsymbol{\beta}$ is separable. For the rest of this section, when it is clear from the context, we will interchangeably refer to it also as the “ranking matrix”.

Learning Problem: Our objective is the estimation problem. Given empirical pairwise comparisons \mathbf{X} , we learn the parameters of the shared ranking components $\sigma^1, \dots, \sigma^K$, or equivalently the binary ranking matrix $\boldsymbol{\beta}$.

Overall Approach: As highlighted in Section 3.1, our approach is to formally reduce the proposed model to a statistical equivalent topic model whose topic matrix is related to the ranking matrix $\boldsymbol{\beta}$. We then identify the separability property for the ranking matrix. We prove that the ranking matrix is an inevitable consequence of a small number of latent factors in a universe of large number of items. We therefore apply the geometric approach outlined in Chapter 2 that can consistently learn the ranking matrix with polynomial sample and computational complexity.

²We can also estimate $\mu_{i,j}$ but it is not the main focus of our estimation problem.

3.3.1 Reduction to Topic Models

The key motivation of our approach is to establish a formal connection between the proposed Topical Ranking Model (in Fig. 3.2) and the topic models. On a high-level, we view each comparison of a user as a “word”, the comparisons made by a single user as a “document”, and each latent ranking component as a “topic”. Formally, we consider a topic model for a set of M documents, each composed of N words drawn from a vocabulary of size $W = Q(Q-1)$. We denote by β^{TM} the $W \times K$ topic matrix. We denote the topic weights as $\theta^{m,\text{TM}}$ and topic prior as $\text{Pr}^{\text{TM}}(\theta)$. We have,

Lemma 9. *The proposed Topic Ranking Model (Fig. 3.2) is statistically equivalent to a standard topic model (Fig. 2.1) whose topic matrix β^{TM} is set to be \mathbf{B} and the topic prior $\text{Pr}^{\text{TM}}(\theta)$ to be $\text{Pr}(\theta)$.*

We note that since \mathbf{B} has some additional structure induced by the property of a valid permutation³, given a general topic matrix β^{TM} , it is possible that there exist no equivalent Topic Ranking Model. The statistical equivalency in Lemma 9 shows that the our learning problem can be solved by any algorithm that can learn the topic matrix in a topic modeling. We next show how our provable approach developed in Chapter 2 can be applied to learn the ranking matrix β .

3.3.2 Separability Property

To apply our provable geometric approach in Chapter 2, we need to identify the key structural property: **Separability**. To be specific, we consider exact separability (Definition 1 and Definition 2) on the binary ranking matrix β .

To start with we consider an example ranking matrix β as in Figure 3.3. Here we have $Q = 3$ items and $K = 3$ distinct permutations $\sigma^1, \sigma^2, \sigma^3$. β has $W = Q(Q-1) = 6$ rows. In this example, β is exact separable where ordered pair $(1, 3)$ is novel to ranking σ^1 , the pair $(2, 1)$ to σ^2 , and the pair $(3, 2)$ to σ^3 . Formally, a novel “word”

³Namely, the totally and transitivity.

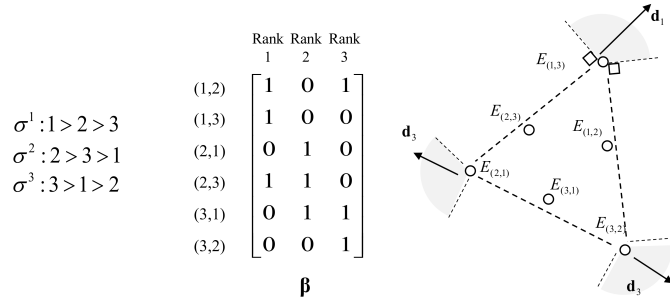


Figure 3.3: A separable ranking matrix β with $K = 3$ rankings over $Q = 3$ items, and the underlying geometry of the row vectors of \mathbf{E} . $(1,3), (2,1), (3,2)$ are novel pairs. Shaded regions depict the solid angles of the extreme points.

for ranking k is a pair of items i, j such that item i is uniquely preferred over j in σ^k while j is ranked higher than i in all the other rankings. We will refer to these pairs (rows of β) as novel pairs.

While the existence of novel pairs seems to be restrictive, we can show that it is in fact satisfied by most ranking matrix β when the number of items Q is large. To be explicit, we assume that the K rankings $\sigma^1, \dots, \sigma^K$ are sampled uniformly from the set of all permutations over the Q items. We show in Lemma 13 (Chapter 4) that β is separable with a probability that converges to 1 exponentially in Q .

Conditions on the Ranking Prior: We consider the same technical conditions on the prior for mixing weights θ^m as discussed in Section 2.3.2. Let $\mathbf{a} = \mathbb{E}(\theta^m)$ and $\mathbf{R} = \mathbb{E}(\theta^m \theta^{m\top})$ are, respectively, the expectation and correlation matrix of the mixing weights, and let $\bar{\mathbf{R}} = \text{diag}^{-1}(\mathbf{a})\mathbf{R}\text{diag}^{-1}(\mathbf{a})$ be the normalized second order moments. We assume the ranking prior to be γ_a -affine independent as in Condition 2 in this section. Therefore, the ranking prior is also at least γ_a -simplicial (Condition 1).

3.3.3 The Geometric Approach and Analysis

So far we have established the statistical connection between the proposed Topic Ranking Model and the topic model. We have also identify the (exact) separability

property in the context of Topic Ranking Model. As a consequence, we can directly apply the geometric approach developed in Chapter 2. In this section, we sketch our learning algorithm and summarize the theoretical analysis.

The Comparison Co-occurrence Matrix Representation: Following the discussion in Section 2.4.1, we adopt the second order moments of the empirical observation as our representation and construct the comparison co-occurrence matrix as follows. We first split all the comparisons of each user into two independent halves, and obtain two empirical comparison-frequency matrices \mathbf{X} and \mathbf{X}' of size $W \times K$. We then normalize their rows to obtain row-stochastic $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}'$. We then construct an empirical comparison co-occurrence matrix of size $W \times W$ as,

$$\hat{\mathbf{E}} = M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top \quad (3.3)$$

Due to the statistical equivalency in Lemma 9, the results in Lemma 5 can be directly applied hence we have,

$$\hat{\mathbf{E}} \xrightarrow[\text{almost surely}]{M \rightarrow \infty} \bar{\mathbf{B}}\bar{\mathbf{R}}\bar{\mathbf{B}}^\top =: \mathbf{E} \quad (3.4)$$

where $\bar{\mathbf{B}} = \text{diag}^{-1}(\mathbf{B}\mathbf{a})\mathbf{B} \text{diag}(\mathbf{a})$, $\bar{\mathbf{R}} = \text{diag}^{-1}(\mathbf{a})\mathbf{R} \text{diag}^{-1}(\mathbf{a})$.

Detecting Novel Pairs as Extreme Points: By Lemma 6 and Lemma 8, if the ranking matrix β (hence \mathbf{B}) is separable and the ranking prior is simplicial, then the novel pairs are the *extreme points* of the convex hull formed by all the row vectors of \mathbf{E} , and the solid angle subtended by the convex hull at these novel rows have non-zero values. Revisiting the example in Figure 3-3, (1, 3), (2, 1), (3, 2) are novel words and the row vectors $\mathbf{E}_{(1,3)}$, $\mathbf{E}_{(2,1)}$, $\mathbf{E}_{(3,2)}$ are extreme points. They have non-zero solid angles as indicated by the shaded regions in Figure 3-3. Formally, similar to Eq. (2.3), we define a solid angle for each pair $\mathbf{E}_{(i,j)}$ as,

$$q_{(i,j)} \triangleq p\{\forall(s,t) : \mathbf{E}_{(i,j)} \neq \mathbf{E}_{(s,t)}, \langle \mathbf{E}_{(i,j)}, \mathbf{d} \rangle > \langle \mathbf{E}_{(s,t)}, \mathbf{d} \rangle\} \quad (3.5)$$

where \mathbf{d} is isotropically distributed random direction. In parallel to Lemma 8,

Lemma 10. *Suppose the ranking matrix β is separable and topic prior, i.e., $\bar{\mathbf{R}}$, is γ_a -affine independent, then, $q_{(i,j)} > 0$ if and only if (i, j) is a novel pair.*

Once all the novel pairs are identified, the ranking matrix can be estimated using a constrained linear regression proposed in Lemma 7. In sum, our solution approach is to: (1) Estimate the solid angles $q_{(i,j)}$ and select K distinct pairs with largest solid angles, (2) Estimate equivalent topic matrix \mathbf{B} using constrained linear regression, and (3) Infer the ranking matrix β using Eq. 3.2. The main steps of our approach are outlined in Algorithm 5 and expanded in detail in Algorithms 6, 7 and 8. Algorithm 6 and 7 inherit the topic discovery algorithm in previous chapter.

Algorithm 8 further processes $\hat{\mathbf{B}}$ to obtain an estimate of the ranking matrix β that guarantees it to be binary and satisfies that $\hat{\beta}_{(i,j),k} + \hat{\beta}_{(j,i),k} = 1$ for all $i \neq j$ and all k . We should point out that we do not enforce the estimated pairwise preferences to be K valid total rankings. However, due to the asymptotic consistency, the estimation will be eventually a valid total ranking if we have access to more observations.

Algorithm 5 Ranking Recovery (Main Steps)

Input: Pairwise comparisons $\tilde{\mathbf{X}}, \tilde{\mathbf{X}}'(W \times M)$; Number of rankings K ; Number of projections P ; Tolerance parameters $\zeta, \epsilon > 0$.

Output: Ranking matrix estimate $\hat{\beta}$.

- 1: Novel Pairs $\mathcal{I} \leftarrow \text{NovelPairDetect}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}', K, Q, \zeta)$
 - 2: $\hat{\mathbf{B}} \leftarrow \text{EstimateRankings}(\mathcal{I}, \mathbf{X}, \epsilon)$
 - 3: $\hat{\beta} \leftarrow \text{PostProcess}(\hat{\mathbf{B}})$
-

Computation Complexity: The proposed ranking estimation Algorithm 5 has the similar computational efficiency as for topic modeling. Formally,

Theorem 6. *The running time of Algorithm 5 is $\mathcal{O}(MNP + Q^2P + Q^2K^3)$.*

We shall note that the last term Q^2K^3 is an upper bound on the linear regressions for Q^2 rows in β . K^3 is the complexity of each linear regression using matrix inversion

Algorithm 6 NovelPairDetect (via Random Projections)

Input: $\tilde{\mathbf{X}}, \tilde{\mathbf{X}}'$; number of rankings K ; number of projections P ; tolerance ζ ;

Output: \mathcal{I} : The set of all novel pairs of K distinct rankings.

$$\hat{\mathbf{E}} \leftarrow M\tilde{\mathbf{X}}'\tilde{\mathbf{X}}^\top$$

$$\forall(i, j), \mathcal{J}_{(i,j)} \leftarrow \{(s, t) : \hat{E}_{(i,j),(i,j)} - 2\hat{E}_{(i,j),(s,t)} + \hat{E}_{(s,t),(s,t)} \geq \zeta/2\},$$

for $r = 1, \dots, P$ **do**

Sample $\mathbf{d}_r \in \mathbb{R}^W$ from an isotropic prior

$$\hat{q}_{(i,j),r} \leftarrow \mathbb{I}\{\forall(s, t) \in \mathcal{J}_{(i,j)}, \hat{\mathbf{E}}_{(s,t)}\mathbf{d}_r \leq \hat{\mathbf{E}}_{(i,j)}\mathbf{d}_r\}, \forall(i, j)$$

end for

$$\hat{q}_{(i,j)} \leftarrow \frac{1}{P} \sum_{r=1}^P \hat{q}_{(i,j),r}, \forall(i, j)$$

$$k \leftarrow 0, l \leftarrow 1, \text{ and } \mathcal{I} \leftarrow \emptyset$$

while $k \leq K$ **do**

$(s, t) \leftarrow$ index of the l^{th} largest value among $\hat{q}_{(i,j)}$'s

if $(s, t) \in \bigcap_{(i,j) \in \mathcal{I}} \mathcal{J}_{(i,j)}$ **then**

$$\mathcal{I} \leftarrow \mathcal{I} \cup \{(s, t)\}, \quad k \leftarrow k + 1$$

end if

$$l \leftarrow l + 1$$

end while

Algorithm 7 Estimate Rankings

Input: $\mathcal{I} = \{(i_1, j_1), \dots, (i_K, j_K)\}$ the set of novel pairs of K rankings; \mathbf{X}, \mathbf{X}' ; precision ϵ

Output: $\hat{\mathbf{B}}$ as the estimate of \mathbf{B} .

$$\mathbf{Y} = (\tilde{\mathbf{X}}_{(i_1, j_1)}^\top, \dots, \tilde{\mathbf{X}}_{(i_K, j_K)}^\top)^\top, \quad \mathbf{Y}' = (\tilde{\mathbf{X}}'_{(i_1, j_1)}^\top, \dots, \tilde{\mathbf{X}}'_{(i_K, j_K)}^\top)^\top$$

for all (i, j) pairs **do**

$$\text{Solve } \hat{\boldsymbol{\beta}}_{(i,j)} \leftarrow \arg \min_{\mathbf{b}} M(\tilde{\mathbf{X}}_{(i,j)} - \mathbf{b}\mathbf{Y})(\tilde{\mathbf{X}}'_{(i,j)} - \mathbf{b}\mathbf{Y}')^\top$$

Subject to $b_k \geq 0, \sum_{k=1}^K b_k = 1$, With precision ϵ

$$\hat{\boldsymbol{\beta}}_{(i,j)} \leftarrow \left(\frac{1}{M}\mathbf{X}_{(i,j)}\mathbf{1}\right)\hat{\boldsymbol{\beta}}_{(i,j)}$$

end for

$$\hat{\mathbf{B}} \leftarrow \text{column normalize } \hat{\boldsymbol{\beta}}$$

but the iterative algorithms in practice have much lower computational complexity.

Also if we use the tricks as in [e.g., Wauthier et al., 2013], we can only compute the

regression for $Q \log(Q)K$ number of rows instead of Q^2 rows. However, these are not

the main focus of this thesis which is to establish the first provable and polynomial

computation complexity bound on the mixed membership ranking models.

Sample Complexity: For the sample complexity, we have,

Algorithm 8 Post Processing

Input: $\widehat{\mathbf{B}}$ as the estimate of \mathbf{B}

Output: $\widehat{\boldsymbol{\beta}}$ as the estimate of $\boldsymbol{\beta}$

- 1: $\widehat{\beta}_{(i,j),k} \leftarrow \frac{\widehat{B}_{(i,j),k}}{\widehat{B}_{(i,j),k} + \widehat{B}_{(j,i),k}}, \forall i, j \in \mathcal{U}, \forall k$
 - 2: $\widehat{\beta}_{(i,j),k} \leftarrow \text{Round}[\widehat{\beta}_{(i,j),k}], \forall i, j \in \mathcal{U}, \forall k$
-

Theorem 7. *Let the ranking matrix $\boldsymbol{\sigma}$ be separable and $\bar{\mathbf{R}}$ is γ_a -affine independent. Then the Algorithm 5 can consistently recover $\boldsymbol{\sigma}$ up to a column permutation as the number of users $M \rightarrow \infty$ and number of projections $P \rightarrow \infty$. Furthermore, for any isotropically drawn random direction \mathbf{d} , $\forall \delta > 0$, if*

$$M \geq \max \left\{ 40 \frac{\log(3W/\delta)}{N\rho^2\eta^4}, 320 \frac{W \log(3W/\delta)}{N\eta^6 \lambda_{\min}} \right\}$$

and $Q \geq 16 \frac{\log(3W/\delta)}{q_\wedge^2}$, then Algorithm 5 fails with probability at most δ . The other model parameters are: $\eta = \min_{1 \leq w \leq W} [\mathbf{B}\mathbf{a}]_w$, $\rho = \min\{\frac{d}{8}, \frac{\pi d_2 q_\wedge}{4W^{1.5}}\}$, $d_2 \triangleq (1-b)\gamma_a$, $d = (1-b)^2\gamma_a^2/\lambda_{\max}$, $b = \max_{j \in \mathcal{C}_0, k} \bar{B}_{j,k}$ and λ_{\max} is the maximum eigenvalues of $\bar{\mathbf{R}}$. q_\wedge is the minimum solid angle of the extreme points of the convex hull of the rows of \mathbf{E} .

3.4 Mixed Membership Mallows Models

We next discuss another MMLVMs for ranking, the Mixed Membership Mallows Model (M4) where the K shared latent ranking components are modeled as distinct Mallows distributions over the permutation space [Mallows, 1957, Marden, 1995]. This new M4 has a few major *conceptual advances*. *First*, from a modeling perspective, it is more general and subsumes the Topic Ranking Model in Section 3.3 as a special case. By adopting the Mallows distribution, it can capture the randomness within each ranking component and therefore subsumes the popular mixture of Mallows model in literature Awasthi et al. [2014], Lebanon and Lafferty [2002], Lu and Boutilier [2014], Meila and Chen [2010] as special cases. *Second*, to develop a provable learning algorithm for M4 we develop the notion of approximate separability and non-trivially extend the provable guarantees in the exact separable setting to

approximate separability. We also show that the ranking matrix induced by M4 is approximate separable with high probability.

3.4.1 Mallows Distribution and Generative Model for M4

Mallows Distribution: The building block of M4 is the classic Mallows model [Mallows, 1957, Marden, 1995] that defines a pmf over all the permutation of the Q items. A Mallows model is parameterized by a reference ranking σ^k and a dispersion parameter $\phi_k \in [0, 1)$.⁴ Specifically,

$$p_M(\sigma|\sigma^k, \phi_k) = \phi_k^{d(\sigma, \sigma^k)} / Z_k \quad (3.6)$$

where Z_k is the normalization constant. By definition, the Mallows distribution is centered around the reference permutation σ^k . The probability on a permutation σ decays exponentially with its Kendall's tau distance $d(\sigma, \sigma^k)$ to σ^k at a rate governed by ϕ_k . This captures the fact that for two items similar in one latent ranking factor, the outcome of their pairwise comparison is more random. We also note that the closer ϕ_k is to 1, the more spread the Mallows pmf is.

M4 then views the ordered pairs produced by each user as a probabilistic mixture of K latent component Mallows pmfs which capture *heterogeneous* influencing ranking factors. We summarize the generative process of M4 in Figure 3-4. It imposes the following distribution on the observed pairwise comparisons $w_{m,n}$,

$$p(w_{m,n} = (i, j) | \theta^m) = \mu_{i,j} \sum_{k=1}^K \sum_{\sigma(i) < \sigma(j)} p_M(\sigma|\sigma^k, \phi_k) \theta_{k,m} \quad (3.7)$$

As we have seen in the Topic Ranking Model, in M4, each user can be influenced by multiple latent factors to different extents, the comparisons produced by each users can potentially be *inconsistent*. In addition, the use of a Mallows model for each

⁴A Mallows model with $\phi_k = 1$ is a uniform distribution on all permutations. It is unidentifiable since any reference permutation σ^k would induce the same distribution.

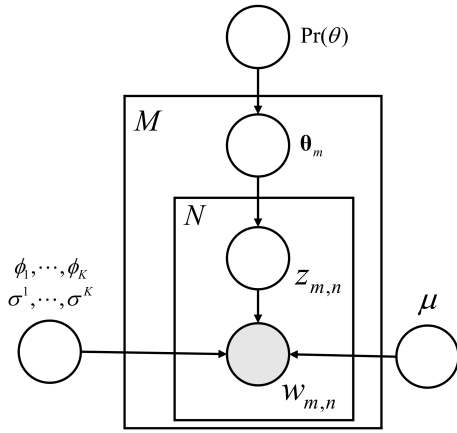


Figure 3-4: Generative process of the Mixed Membership Mallows Model and its Graphical representation. The boxes represent replicates (the outer as users, and the inner as comparisons).

For each user $m = 1, \dots, M$,

- 1) Sample ranking weight $\theta^m \in \Delta^K$ from some prior distribution $\Pr(\theta)$
- 2) For each comparison $n = 1, \dots, N$,
 - a Sample a pair of items i, j from μ
 - b Sample ranking token $z \in \{1, \dots, K\} \sim \text{Multinomial}(\theta^m)$
 - c Sample a ranking $\sigma_{m,n} \sim z$ -th Mallows component (σ^z, ϕ_z)
 - d Output ordered pair $w_{m,n} = (i, j)$ if $\sigma_{m,n}(i) < \sigma_{m,n}(j)$; Otherwise output $w_{m,n} = (j, i)$

latent factor allows one to capture potential *randomness* in the outcomes of item comparisons for items that are very similar.

Learning Problem: Given empirical comparisons \mathbf{X} and K , our primary objective is to learn the parameters of the shared latent Mallows components, i.e., the reference rankings σ^k 's and the dispersion parameters ϕ_k 's. For the problem of inferring θ^m [Blei et al., 2003] and predicting preferences for new observations, we use standard tools [Wallach et al., 2009]. Establishing guarantees for the inference and prediction problems is not addressed in this work and remains an open question.

Connection to other ranking models: Before we discuss the M4 model in detail, we point out that the proposed M4 is a much more general family and it subsumes the following models in literature as special cases.

Proposition 3. *In Mixed Membership Mallows Model,*

- a) *If $\phi_k \rightarrow 0, k = 1, \dots, K$, then, each Mallows component reduces to a single permutation σ^k , and the M_4 reduces to the model in Ding et al. [2015b] (hence subsumes [Farias et al., 2009, Jagabathula and Shah, 2008] as special cases).*
- b) *If the topic prior $\Pr(\theta)$ has non-zero probability only on the vertices of K -dimension simplex, then, the M_4 reduces to the mixture of Mallows model*

Awasthi et al. [2014], Lu and Boutilier [2014]

Therefore, all our theoretical guarantees apply to the mixture of Mallows model and the mixed membership models in [Ding et al., 2015b]. While our results coincide with those in [Ding et al., 2015b], they are the *first provably consistency and efficiency guarantees* for mixture of Mallows model from pairwise comparisons as a by-product [Awasthi et al., 2014, Lu and Boutilier, 2014, Meila and Chen, 2010].

3.4.2 Reduction to Topic Model via Ranking Matrix

Recall that our solution strategy is to formally associate M4 with a topic model whose topic matrix provides an *information-equivalent representation* of the parameters of M4. To do this, it is convenient to define a $W \times K$ ranking matrix β as,

$$\beta_{(i,j),k} := \sum_{\sigma: \sigma(i) < \sigma(j)} p_M(\sigma | \sigma^k, \phi_k) \quad (3.8)$$

We note that β is completely determined by σ^k 's and ϕ_k 's. Statistically, $\beta_{(i,j),k}$ is the probability that item i is preferred over item j in a ranking sampled from the k -th Mallows component. This matrix β is defined in analogous to the topic matrix in a topic model [Blei, 2012]. Similarly as in Section 3.3, we also define a $W \times K$ matrix \mathbf{B} as $B_{(i,j),k} = \mu_{i,j} \beta_{(i,j),k}$ and refer to it also as ranking matrix. Given these definitions, the observation probability in Eq. (3.7) can be simplified as,

$$p(w_{m,n} = (i, j) | \theta^m) = \mu_{i,j} \sum_{k=1}^K \sum_{\sigma(i) < \sigma(j)} p_M(\sigma | \sigma^k, \phi_k) \theta_{k,m} \quad (3.9)$$

$$= \mu_{i,j} \sum_{k=1}^K \beta_{(i,j),k} \theta_{k,m} = \sum_{k=1}^K B_{(i,j),k} \theta_{k,m} \quad (3.10)$$

We first show that the underlying parameters σ^k 's and ϕ_k 's can be recovered directly from β although it is not straightforward in Eq. 3.8,

Proposition 4. *Let the ranking matrix β be defined as in Eq. (3.8). Then, $\forall(i, j)$ and $\forall k$, we have,*

- a. *If $\sigma^k(i) < \sigma^k(j)$, then $\beta_{(i,j),k} > 0.5 > \beta_{(j,i),k}$; $\beta_{(i,j),k} + \beta_{(j,i),k} = 1$*
- b. *If $\sigma^k(j) = \sigma^k(i) + 1$ and $\phi_k < 1$, $1/\beta_{(i,j),k} = 1 + \phi_k$;*
- c. *If $\sigma^k(i) > \sigma^k(j)$ and $L = \sigma^k(i) - \sigma^k(j) + 1$, then, $\beta_{(i,j),k} \leq \frac{L\phi_k^{L-1}}{1+L\phi_k^{L-1}}$*

Prop. 4 a) shows that σ_k 's can be recovered from β by rounding its entries to the nearest integer. Prop. 4 b) shows that the dispersion parameter ϕ_k can be recovered from β . Thus, β does indeed provide an information-equivalent representation of M4. For the rest we focus on the estimation of \mathbf{B} . We note that Prop. 4 c. is a more general property of b. and motivates us to investigate the approximate separability property on β . Before we move on, we note that by the definition of separability in Definition 1, β is λ -approximate separable iff \mathbf{B} is λ -approximate separable.

Reduction to Topic Model: The proposed M4 shares the same structure as in the Topical Ranking Model in section 3.3. Therefore, we can establish a statistical reduction of M4 to a topic model. Noting that \mathbf{B} is also column-stochastic, and similarly as Lemma 9, we have

Lemma 11. *The Mixed Membership Mallows Model (Fig. 3.4) is statistically equivalent to a topic model whose topic matrix β^{TM} is set to be \mathbf{B} and the topic prior $\Pr^{TM}(\theta)$ to be $\Pr(\theta)$.*

3.4.3 Overview of Algorithm, Key Insights, and Theoretical Results

We have just reduced the learning problem of M4 to the estimation the ranking matrix β (Eq. (3.8)), which can be solved by any estimation algorithm in topic modeling. However, here are major technical difficulties in directly applying our previous developed geometric approach with polynomial sample and computational complexity guarantees. Specifically, the exact separability condition used in previous approach can not be satisfied by M4 since all the entries of β is strictly positive (see Eq. (3.8)).

As highlighted in the beginning of this section, two nontrivial technical innovations are needed to overcome this difficulty. (1) we consider the general “approximate separability” property (Definition 1) and prove that most instances of the ranking matrix in M4, when appropriately sampled, are approximately separable if $Q \gg K$. (2) we generalize the results extreme point geometry property measured by the solid angle (Eq. (3.5)) for learning β from exact to approximate separability. We introduce these key technical advances in this section.

A. Approximate Separability in M4

We first identify the separability property in M4. We consider the approximate separability (Definition 1) on the ranking matrix β . Intuitively, λ -approximate separability requires the existence of ordered pairs that have negligible probability in all-but-one of the Mallows components, i.e., the row entries concentrate predominantly in one column. We call such pairs (rows of β) as λ -approximately novel pairs (rows) for each latent Mallows component.

Figure 3.5 shows an example approximate separable ranking matrix β . In this example, the pairs 1, 2, 3 are, respectively, novel for the first, second, and third Mallows components. Since $\beta_{(i,j),k}$ is a pairwise comparison probability, row (i, j) being approximately novel means that i is preferred over j in only one factor and i is mostly likely to be preferred below j in the remaining. To show that this is reasonable and achievable for some small constant λ in ranking matrix of M4, we recall the property c. in Prop. 4 where

$$\beta_{(i,j),k} \leq \frac{L\phi_k^{L-1}}{1 + L\phi_k^{L-1}}$$

for $\sigma^k(i) > \sigma^k(j)$ and $L = \sigma^k(i) - \sigma^k(j) + 1$. Since $\phi_k < 1$, if L increases, the corresponding $\beta_{(i,j),k}$ is very close to 0. Therefore, to achieve the approximately separability in M4, for a novel pair of item i and j , i is uniquely preferred over j in

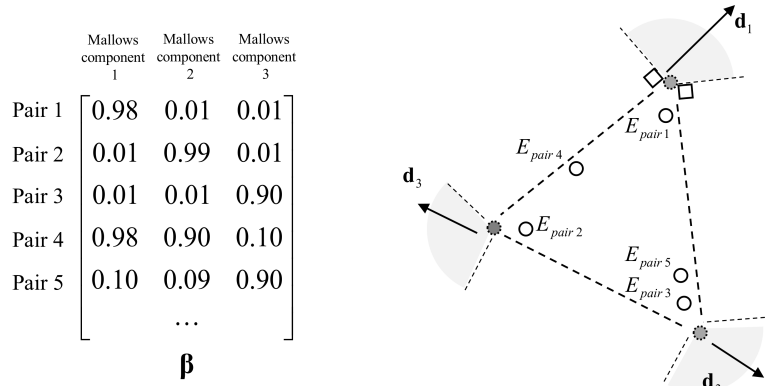


Figure 3-5: An example of approximate separable β with $K = 3$, and the underlying geometry of the row vectors of \mathbf{E} . Pair 1, 2, 3 are approximate novel pairs for Mallows component 1, 2, and 3. The shaded dash circles represent the ideal extreme points with exact separable β and the shaded regions depict their solid angles. The numbers in β are from $\phi_k = 0.1$. $\beta_{(i,j),k} \approx 0.01$ when $L = 3$, $\beta_{(i,j),k} \approx 0.1$ when $L = 2$. $L = \sigma_k(i) - \sigma_k(j) + 1$.

one reference ranking (hence $\beta_{(i,j),k} > 1/2$ by Prop. 4 a.), while j is ranked higher than i by a large margin L in all other reference rankings (hence $\beta_{(i,j),l} < L\phi_l^{L-1}$ for $l \neq k$).

We will prove in Chapter 4 that most ranking matrices β in M4 are approximately separable with high probability. Our approach can therefore be applied to most large M4. For the rest of this section, we will always assume that the ranking matrix β is λ -approximate separable for some small constant $\lambda > 0$.

B. Robust Novel Word Detection via Random Projection

We now demonstrate how the extreme point geometry considered in the previous problems can be extended in the approximate separability setting in M4. We adopt the same comparison co-occurrence matrix representation and construct it as in section 3.3. We first split all the comparisons of each user into two independent halves, and obtain two empirical comparison-frequency matrices \mathbf{X} and \mathbf{X}' of size $W \times K$. We then normalize their rows to obtain row-stochastic $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}'$. We then construct

an empirical comparison co-occurrence matrix of size $W \times W$ as,

$$\widehat{\mathbf{E}} = M\bar{\mathbf{X}}'\bar{\mathbf{X}}^\top \quad (3.11)$$

Due to the statistical equivalency in Lemma 11, the results in Lemma 5 can be directly applied hence we have $\widehat{\mathbf{E}} \xrightarrow[\text{almost surely}]{M \rightarrow \infty} \bar{\mathbf{B}}\bar{\mathbf{R}}\bar{\mathbf{B}}^\top =: \mathbf{E}$ where $\bar{\mathbf{B}} = \text{diag}^{-1}(\mathbf{B}\mathbf{a})\mathbf{B}\text{diag}(\mathbf{a})$, $\bar{\mathbf{R}} = \text{diag}^{-1}(\mathbf{a})\mathbf{R}\text{diag}^{-1}(\mathbf{a})$. \mathbf{a} and \mathbf{R} are the expectation vector and correlation matrix of the mixing weights respectively. For simplicity, we assume $\bar{\mathbf{R}}$ to be full rank hence it is simplicial and affine independent.

Robust Extreme Points are approximate Novel pairs: Recall in Lemma 10 when β is exactly separable, the novel pairs (rows) in \mathbf{E} are exactly the set of extreme points. These are indicated as the shaded dash circles in Figure 3-5.

We still focus on the rows of \mathbf{E} . We also use the solid angle defined as the probability that a row vector $\mathbf{E}_{(i,j)}$ has the maximum projection value along an isotropically distributed direction $\mathbf{d} \in \mathbb{R}^{W \times 1}$:

$$q_{(i,j)} \triangleq p\{\forall(s,t) : \|\mathbf{E}_{(i,j)} - \mathbf{E}_{(s,t)}\| \geq \zeta, \mathbf{E}_{(i,j)}\mathbf{d} > \mathbf{E}_{(s,t)}\mathbf{d}\} \quad (3.12)$$

Consider now that β is λ -approximate separable with small enough $\lambda > 0$. The rows of \mathbf{E} (empty circles in Figure 3-5) can be viewed as a small perturbation from the ideal case. As a consequence, (a) The rows of approximately novel pairs – $\mathbf{E}_{\text{pair1}}$, $\mathbf{E}_{\text{pair2}}$, and $\mathbf{E}_{\text{pair3}}$ in empty circle – are inside the ideal convex hull and are close to the ideal extreme points. The corresponding solid angles subtended will be close to that of the ideal extreme points which are lower bounded away from 0. (b) The non-novel rows could become extreme points but would be close to the convex hull formed by the approximate novel rows (e.g., $\mathbf{E}_{\text{pair4}}$ in Figure 3-5). But in this case the associated solid angles will be very close to 0.

To sum up, when the deviation introduced by λ -approximate separability is small,

the solid angle can measure the “robustness” of an extreme point. If we sort the non-zero solid angles for all the rows in \mathbf{E} , the distinct K rows with largest solid angles must correspond to $c\lambda$ -approximate novel pairs for some constant c and a properly defined ζ in Eq. (3.12).

Overall Algorithm: In sum, we adopt the same algorithmic procedure as in section 3.3.3. We first detect approximately novel pairs for K distinct Mallows components by sorting the solid angles of all pairs using a few i.i.d isotropic random projections (Algorithm 6). Once the approximate novel pairs for K distinct Mallows components are identified, \mathbf{B} hence β can be estimated using constrained linear regression (Algorithm 7). We then post-process β to get σ_k, ϕ_k 's of the shared Mallows components by Prop. 4 (Algorithm 10). These steps are outlined in Algorithm 5. In Algorithm 10: step 1 estimates all the pairwise relations $\sigma_{(i,j),k} = \mathbb{I}(\sigma_k(i) < \sigma_k(j))$ in σ_k . Step 2 aggregates them to the positions of each item in σ_k . Step 3 estimates ϕ_k .

Algorithm 9 M4 Estimation (Main Steps)

Input: Pairwise comparisons $\tilde{\mathbf{X}}, \tilde{\mathbf{X}}'(W \times M)$; Number of latent components K ; Number of projections P ; Tolerance parameters $\zeta, \epsilon > 0$
Output: Reference ranking $\hat{\sigma}_k$ and dispersion $\hat{\phi}_k, k = 1, \dots, K$
 1: Novel Pairs $\mathcal{I} \leftarrow \text{NovelPairDetect}(\tilde{\mathbf{X}}, \tilde{\mathbf{X}}', K, P, \zeta)$ (Alg. 6)
 2: $\hat{\mathbf{B}} \leftarrow \text{EstimateRankingMatrix}(\mathcal{I}, \mathbf{X}, \epsilon)$ (Alg. 7)
 3: $\hat{\sigma}_1, \dots, \hat{\sigma}_K, \hat{\phi}_1, \dots, \hat{\phi}_K \leftarrow \text{PostProcess}(\hat{\mathbf{B}})$ (Alg. 10)

Algorithm 10 Post Processing (Mixed Membership Mallows Model)

Input: $\hat{\mathbf{B}}$ as the estimate of \mathbf{B}

Output: $\hat{\sigma}_k, \hat{\phi}_k, k = 1, \dots, K$

- 1: $\hat{\beta}_{(i,j),k} \leftarrow \frac{\hat{B}_{(i,j),k}}{\hat{B}_{(i,j),k} + \hat{B}_{(j,i),k}}, \forall i, j \in \mathcal{U}, \forall k$
 - 2: $\hat{\sigma}_{(i,j),k} \leftarrow \text{Round}[\hat{\beta}_{(i,j),k}], \forall i, j \in \mathcal{U}, \forall k$
 - 3: $\hat{\sigma}_k \leftarrow \text{GlobalRank}(\hat{\sigma}_{(i,j),k}, \forall i, j) \forall k$ (First count the number of times each item wins in all pairwise comparison and then sort.)
 - 4: $\hat{\phi}_k \leftarrow \frac{1}{Q-1} \sum_{i=1}^{Q-1} \frac{1}{\hat{\beta}_{(\sigma_k^{-1}(i), \sigma_k^{-1}(i+1)), k}} - 1, \forall k$ ($\sigma_k^{-1}(i)$ is the item in the i -th position in ranking σ_k .)
-

C. Complexity Analysis

Our approach has similar polynomial computation complexity as in previous sections. Formally, we have,

Theorem 8. *The running time of Algorithm 5 is $\mathcal{O}(MNP + Q^2P + Q^2K^3)$.*

Theorem 9. *Let the ranking matrix β be λ -approximate separable and the second order moments \mathbf{R} of ranking prior to be full rank. If*

$$\lambda \leq \frac{a_{\min}\kappa(1-\phi)q_{\wedge}}{8K^2a_0\sqrt{\log(W/q_{\wedge})}} \quad (3.13)$$

and $M, P \rightarrow \infty$, then, Algorithm 5 can consistently recover all the reference rankings of the latent Mallows distributions. Moreover, $\forall \delta > 0$, if

$$M \geq \max \left\{ \frac{640W^2 \log(3W/\delta)}{N\eta^4 d^2 q_{\wedge}^2}, \frac{320W \log(3W/\delta)}{N\eta^4 \lambda_{\min}^2 a_{\min}^2 (1-\phi)^2} \right\}$$

and for

$$P \geq 32 \frac{\log(3W/\delta)}{q_{\wedge}^2}$$

the proposed algorithm fails with probability at most δ . The other model parameters are defined as follows: $\eta = \min_{1 \leq w \leq W} [\mathbf{B}\mathbf{a}]_w$; a_{\max} , a_{\min} are the max/min of entries of \mathbf{a} ; $a_0 = \max_{i,j} a_i/a_j$; $\mathbf{Y} = \bar{\mathbf{R}}\bar{\mathbf{B}}$; $\kappa = \lambda_{\min}/\lambda_{\max}$ is the condition number of $\bar{\mathbf{R}}$; q_{\wedge} be the minimum normalized solid angle formed by row vectors of \mathbf{Y} ; $d = 6\kappa/K$; $\phi_k \leq \phi < 1$. N is the number of comparisons of each user.

The detailed proofs are summarized in [Ding et al., 2015c]. Eq. (3.13) provides an explicit sufficient upper bound on the required λ -approximate separable degree. It is roughly inverse polynomial in K . By Prop. 4.d, the margin L required to satisfy λ in Eq. (3.13) should scale as $O(\log(K))$ which is small. We note that in the complexity bounds, the term $1 - \phi$ represents the spread of the Mallows components and determines the hardness of estimation: for smaller ϕ , λ can be larger and the required M is smaller. When $\phi \rightarrow 1$, Eq. (3.13) reduces to $\lambda = 0$ and $M \geq \infty$ which is not achievable and the corresponding Mallows distribution is un-identifiable.

3.5 Empirical Results

We present the empirical performances proposed Mixed Membership Ranking models first on semi-synthetic dataset in order to for validation purpose, and then on real-world datasets in order to demonstrate that the proposed model can indeed effectively capture the variability that one encounters in the real world. We focus on the collaborative filtering applications where population heterogeneity and user inconsistency are the well-known characteristics [e.g., Ricci et al., 2011, Salakhutdinov and Mnih, 2008a].

We use Movielens, a benchmark movie-rating dataset widely used in the literature.⁵ The rating-based data is selected due to its public availability and widespread use, but we convert it to pairwise comparisons data and focus on modeling from a ranking viewpoint. This procedure has been suggested and widely used in the rank--aggregation literature [e.g., Lu and Boutilier, 2014, Volkovs and Zemel, 2014]. For the semi-synthetic datasets, we evaluate the *reconstruction error* between the learned rankings $\hat{\beta}$ and the ground truth. In M4, this mean the reference rankings of the shared Mallows components. We adopt the standard *Kendall's tau distance* between two rankings. It proportional to the number of pairs in which two rankings differ. For the real-world datasets where true parameters are not available, we use the *held-out likelihood*, a standard metric in ranking prediction [Lu and Boutilier, 2014] and in topic modeling Wallach et al. [2009].

In addition, we consider the standard task of rating prediction via our proposed ranking model. Our aim here is to illustrate that our model is suitable for real-word data. We do not optimize tuning parameters in order to achieve the best result. We measure the performance by *root-mean-square-error* (RMSE) which is the standard in literature[e.g., Salakhutdinov and Mnih, 2008a, Toscher et al.].

⁵Another large benchmark, Netflix dataset, has been removed from public domain due to privacy issues. Movielens is currently available at <http://grouplens.org/datasets/movielens/>

3.5.1 Semi-synthetic Simulation

We first use *semi-synthetic* dataset to validate the performance of our algorithm. In order to match the dimensionality and other characteristics that are representative of real-world examples, we generate the semi-synthetic pairwise comparisons dataset using a benchmark movie star-ratings dataset, Movielens. The original dataset has approximately 1 million ratings for 3952 movies from $M = 6040$ users. The ratings range from 1 star to 5 stars. We follow the procedure in [Lu and Boutilier, 2014, Volkovs and Zemel, 2014] to generate the semi-synthetic dataset as follows. We consider the $Q = 100$ most frequently rated movies and train a latent factor model on the star-ratings data using a state-of-the-art matrix factorization based algorithm [Salakhutdinov and Mnih, 2008a]. This approach is selected for its state-of-the-art performance on many real-world collaborative filtering tasks. This procedure learns a $Q \times K$ movie-factor matrix whose columns are interpreted as scores of the Q movies over the K latent factors [Salakhutdinov and Mnih, 2008a, Volkovs and Zemel, 2014]. **(For Topic Ranking Model)** By sorting the scores of each column of the movie-factor matrix, we obtain K rankings for generating the semi-synthetic dataset. We set $K = 10$ as suggested by [Lu and Boutilier, 2014, Salakhutdinov and Mnih, 2008a]. We note that the resulting ranking matrix σ satisfies the separability condition. **(For M4)** We set the reference rankings are above. We then set the same dispersion parameters for all Mallows components as $\phi_k = \phi$ for $\phi = 0, 0.1, 0.2, 0.5$.

The other model parameters are set as follows. $\mu_{i,j} = 1/\binom{Q}{2}$, $\forall i, j \in \mathcal{U}$. The prior distribution for θ^m is set to be Dirichlet $\Pr(\theta^m | \alpha) = \frac{1}{C} \prod_{k=1}^K \theta_k^{\alpha_k - 1}$ as suggested by [Lu and Boutilier, 2014]. The parameters α_k 's are determined by $\alpha_k = \alpha_0 a_k$, where the concentration parameter $\alpha_0 = 0.1$. **For Topic Ranking Models**, we set the expectation $\mathbf{a} = [a_1, \dots, a_K]^\top$ to be sampled uniformly from the $K = 10$ dimensional

simplex for each random realization.⁶ **For M4 models**, we only consider the symmetric Dirichlet prior. We note that the correlation matrix \mathbf{R} of the Dirichlet distribution has full rank [Arora et al., 2013, Ding et al., 2014b]. We fix $N = 300$ comparisons per user to approximate the observed average pairwise comparisons in the Movielens dataset and vary M .

Since the estimation of β is determined only up to a column permutation, we align the columns of β and $\hat{\beta}$ using bipartite matching based on the Kendall’s tau distance of the reference rankings. In the case of Topic Mallows Model whose output is a binary ranking matrix, it is exactly the ℓ_1 distance between two binary columns of the ranking matrix β and $\hat{\beta}$. We further normalize the ℓ_1 error by $W = Q \times (Q - 1)$ and average across the K ranking components and the error measure is a number between $[0, 1]$.

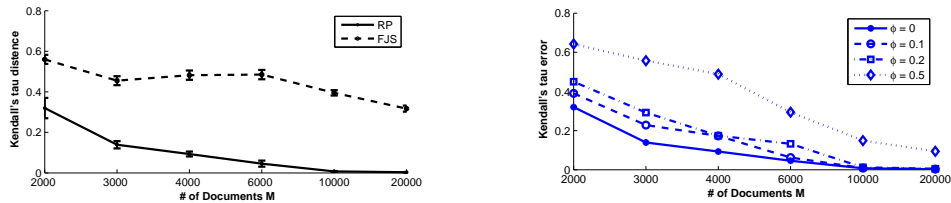


Figure 3-6: The normalized Kendall’s tau distance error of the estimated rankings, as functions of M , estimated by RP and FJS from the semi-synthetic dataset with $Q = 100, N = 300, K = 10$.

For the Topic Ranking Model simulation, we compare our proposed algorithm (denoted by RP) against the algorithm proposed in [Farias et al., 2009, Jagabathula and Shah, 2008] (denoted by FJS). To the best of our knowledge, this is the most recent algorithm with consistency guarantees for $K > 1$.⁷ We compared how the

⁶This is designed to be able to compare against the baseline method, FJS[Farias et al., 2009] since their methods requires the K entries in the topic expectations to be distinct.

⁷Although FJS was developed for a mixture ranking model setting, we can show that can be used in our mixed membership ranking model settings as in Figure 3-2. The key reason is that FJS only exploits the first order statistics. Our model induces the same distribution and the asymptotically for the first order statistics as the mixture models. We also verify that all the other technical conditions in [Farias et al., 2009] are satisfied.

estimation error varies with the number of users M , and the results are depicted in Figure 3-6. For each setting, we average over 10 Monte Carlo runs.

Evidently, our algorithm shows superior performance over FJS as in left of Figure 3-6. More specifically, since our ground truth ranking matrix is separable, as M increases, the estimation error of RP converges to zero, and the convergence is much faster than FJS. We note that only for $M \geq 100,000$ does the error of the FJS algorithm eventually start approaching 0. The right of Figure 3-6 depicts how the estimation error varies with the number of users M with different values of dispersion. We can see that the reconstruction error in reference rankings for $\phi = 0, 0.1, 0.2$ converges to zero at different rates as a function of M . For M4 with $\phi = 0.5$, it converges to a small but non-zero number when $M \rightarrow \infty$. We note that for the ground-truth ranking matrix β , it is $\lambda = 0, 0.01, 0.05, 0.20$ approximate separable for $\phi = 0, 0.1, 0.2, 0.5$ respectively. Our approach therefore can correctly detect the reference rankings when λ is small. When λ is mild, it can still detect most of the reference rankings correctly.

3.5.2 Predicting pairwise comparisons

We apply our mixed membership models to the real-world Movielens dataset and consider the task of predicting pairwise comparisons. We train and evaluate our model using the comparisons obtained from the star-ratings of the Movielens dataset. This procedure of generating comparisons from star-ratings is motivated by [Lu and Boutilier, 2014, Volkovs and Zemel, 2014]. We consider two settings: (1) new comparison prediction, and (2) new user prediction. We focus on the $Q = 100$ most frequently rated movies and obtain a subset of 183,000 star-ratings from $M = 5940$ users. The pairwise comparisons are generated from the star ratings following [Lu and Boutilier, 2014, Volkovs and Zemel, 2014]: for each user m , we **select** pairs of movies i, j that user m rated, and **compare** the stars of the two movies to generate

comparisons.

To **select** pairs of items to compare, we consider: (a) (Full) all pairs of movies that a user has rated, or (b) (Partial) randomly select $5N_{star,m}$ pairs where $N_{star,m}$ is the number of movies user m has rated.

To **compare** a pair of movies i, j rated by a user, $w_{m,n} = (i, j)$ if the star rating of i is higher than j . For ties, we consider: (i)(Both) generate $w_{m,1} = (i, j)$ and $w_{m,2} = (j, i)$, (ii) (Ignore) do nothing, and (iii) (Random) select one of $w_{m,1}, w_{m,2}$ with equal probability.

New comparison prediction: In this setting, for each user, a subset of her ratings are used to generate the training comparisons while the remaining are for testing comparisons. We follow the training/testing split as in [Salakhutdinov and Mnih, 2008a].⁸ We convert both the training ratings and testing ratings into training comparisons and testing comparisons independently.

We evaluate the performance by the predictive log-likelihood of the testing data, i.e., $\Pr(\mathbf{w}_{test} | \mathbf{w}_{train}, \hat{\beta})$. Given the estimate $\hat{\beta}$, we first fit a Dirichlet prior model as in [Arora et al., 2013, Ding et al., 2014b]. We then calculate the prediction log-likelihood using the approximation in [Wallach et al., 2009]. We first compare binary ranking matrix generated by RP algorithm for Topic Ranking Model (TRM) against the FJS algorithm. Figure 3·7(left) summarizes the results for different strategies in generating the pairwise comparisons with $K = 10$ held fixed, and Figure 3·7(right) summarizes the results as function of K . The log-likelihood is normalized by the total number of pairwise comparisons tested. As depicted in Figure 3·7, the log-likelihood produced by the proposed algorithm RP is higher, by a large margin, compared to FJS. The improvement in predictive accuracy is robust to how the comparison data is constructed and the number of latent factors K .

New user prediction: In this setting, we split the first 4000 users (in the original

⁸The training/testing split is available at <http://www.cs.toronto.edu/~rsalakhu/BPMF.html>

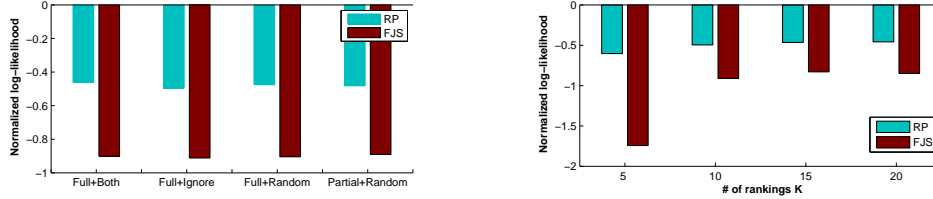


Figure 3-7: The normalized log-likelihood for new comparison predictions, (left) under different comparison generating strategies with $K = 10$, and (right) for various number of latent factors K with Full+Ignore strategy, on $Q = 100$ most rated movies in Movielens dataset. RP denotes Alg. 5 for Topic Ranking Model. FJS denotes the algorithm in [Farias et al., 2009]. The log-likelihood are normalized with the number of test pairs.

dataset) in the Movielens dataset for training comparisons while the remaining users’ comparisons are used for testing. We use the held-out log-likelihood, i.e., $\Pr(\mathbf{w}_{test}|\hat{\sigma})$ to measure the performance. The log-likelihoods are again calculated using the standard Gibbs Sampling approximation [Wallach et al., 2009].

We first use the $Q = 100$ most rated movies compare our algorithm RP for TRM against the FJS algorithm. The log-likelihoods are then normalized by the total number of comparisons in the testing phase. The number of latent ranking components is held fixed at $K = 10$. We summarize the results results in Figure 3-8 (left). The results agrees with the previous experiments.

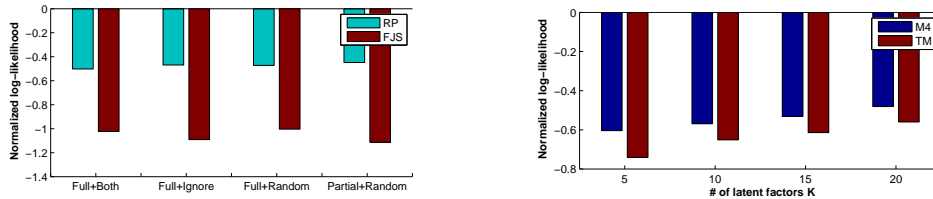


Figure 3-8: The normalized log-likelihood for new user prediction (left) different strategies for constructing comparisons with $K = 10$ for the $Q = 100$ most rated movies by RP for TRM and FJS algorithm [Farias et al., 2009], and (right) Full+Ignore strategy for various K for the $Q = 200$ most rated movies by RP algorithm for TRM and M4 models. The log-likelihood are normalized with the number of test pairs.

We then use the $Q = 200$ most frequently rated movies and compare the performance using TRM and M4 models for different settings of K in Figure 3.8 (right). One can see that M4 can further improve the prediction accuracy of TRM for different choices of K .

3.5.3 Predicting star ratings

To further demonstrate that our model can capture real-world user behavior, we consider the standard rating prediction task in recommendation system Toscher et al.. We first train the proposed mixed membership models using the training comparisons, and then predict ratings by aggregating the prediction of properly defined test comparisons. The purpose of this experiment is not to optimize to achieve the best empirical result in the rich literature on rating prediction.

We use the same training/testing rating split from Salakhutdinov and Mnih [2008a], and focus on the $Q = 100$ most rated movies in MovieLens following Ding et al. [2015b]. We convert the training ratings into training comparisons (for each user, all pairs of movies she rated in the training set are converted into comparisons based on the stars and the ties are ignored) and train a M4 model. The ranking prior is set to be Dirichlet. To predict stars rating $r_{i,m}$ of user m for movie i , we consider the following method: for $s = 1, \dots, 5$, we set $r_{i,m} = s$, and compare it against the movies user m has rated in the training set. This generates a set of pairwise comparisons $\mathbf{w}_{i,m}(s)$. For example, if user m has rated movies A, B, C with 4, 2, 5 stars respectively in the training set and we are predicting her rating for movie D . Then for $s = 3$, $\mathbf{w}_{D,m}(3) = \{(A, D), (D, B), (C, D)\}$. We choose s such that,

$$\hat{r}_{i,m} = \arg \max_s p(\mathbf{w}_{i,m}(s) | \mathbf{w}_{train}, \hat{\beta}).$$

We evaluate the performance using the standard root-mean-square-error (**RMSE**)

metric Toscher et al..⁹ We compared our proposed mixed membership ranking models, Topic Ranking Model (TRM), the Mixed Membership Mallows Model (M4), against two benchmark rating-based algorithms, Probability Matrix Factorization (PMF) in Salakhutdinov and Mnih [2008b], and Bayesian probability matrix factorization (BPMF) in Salakhutdinov and Mnih [2008a] that have robust empirical performance in literature¹⁰. Both PMF and BPMF are latent factor models and the number of latent factors K has the similar interpretation as in M4. Note that the ratings predicted by our algorithm are integers from 1 to 5, we also round the output of BPMF to the nearest integers from 1 to 5 (BPMF-int).

We report the RMSE for different choices of K in Table 3.2. It is clear that M4 improves upon TRM in which the latent factors are restricted to single permutations. On the other hand, when compared to the rating based algorithms, the RMSE of M4 approach are comparable to BPMF and outperforms BPMF-int and PMF although they are coming from a different feature space. We note that the BPMF typically provides robust and benchmark results on real-world problems. This demonstrates that our approaches is suitable for real-world noisy user behavior.

Table 3.2: Testing RMSE on the Movielens dataset with $Q = 100$ most rated movies.

K	PMF	BPMF	BPMF-int	TM	M4
10	1.0491	0.8254	0.8723	0.8840	0.8509
15	0.9127	0.8236	0.8734	0.8780	0.8296
20	0.9250	0.8213	0.8678	0.8721	0.8241

⁹Normalized Discounted Cumulative Gain (nDCG) is another standard metric. It requires, however, to predict a total ranking and is inapplicable in our test setting.

¹⁰We use the suggested settings to optimize the hyper-parameters and use the implementation and data split from <http://www.cs.toronto.edu/~rsalakhu/BPMF.html>

Chapter 4

Most Large MMLVMs are Separable

In previous chapters we have seen that the *separability* has a intuitive appeal and there are some empirical evidences to support it. Yet it might appear to be somewhat restrictive. In this chapter, we demonstrate that separability is not only a natural and convenient structural property, but is, in fact, an *inevitable consequence of high-dimensionality* [Ding et al., 2015a,c]. In particular, if we consider a smoothed setting in which the K latent factors (e.g., topics in topic modeling or ranking components in ranking models) are randomly sampled from some prior, then, the resulting MMLVMs will be (approximately) separable with probability tending to one as W , the dimension of observation, scales to infinity sufficiently faster than K , the number of latent factors. We explicitly show that the topic matrix studied in Chapter 2, the binary ranking matrix for the Topic Ranking Model in Section 3.3, and the ranking matrix for the Mixed Membership Mallows Model in Section 3.4 are (approximately) separable with high probability with suitable priors.

Although the three models we will discuss in this Chapter are distinct, the proofs of the inevitability of separability property share the same framework: we reduce the analysis of the separability properties of latent variable matrix β to that of a related $W \times K$ dimensional random matrix whose *rows are independent*. Then computing the probability that β is approximately separable can be reduced to examining of the probability that each independent row vector in the related matrix is approximately novel to one of the K topics. We then apply the union bound argument to derive a

lower-bound on the probability of β being λ -separable.

4.1 Separability in Topic Modeling

We first consider the topic models whose generative procedures are summarized in Figure 2.1. Following typical settings in literature [Blei, 2012, Blei and Lafferty, 2007, Blei et al., 2003, Tang et al., 2014, Wallach et al., 2009], we assume that the topic matrix β is a realization of the following prior on the $W \times K$ column-stochastic matrix: the K column vectors of β are i.i.d. samples from a symmetric Dirichlet prior $\text{Dir}(\beta_0)$ with concentration parameter $\beta_0 > 0$. Noting that each entry in β is non-zero with probability 1, we consider the approximate separability with small $\lambda > 0$.

Main Results: We calculate a lower bounds on the probability that β is λ -approximately separable. This lower bound depends on W (the size of vocabulary), K (the number of latent topics), and β_0 (the concentration parameter of the Dirichlet prior on the columns of β). It converges to 1 as W increases. Formally,

Lemma 12. *Let the K columns of the topic matrix β be generated i.i.d from $\text{Dir}(\beta_0)$ for $\beta_0 \in (0, 1)$. Then, the probability that β is λ -approximately separable is at least*

$$1 - Kc_1 \exp(-c_2 W \beta_0) - K \exp(-W p_1(\beta_0, \lambda/4, K)) \quad (4.1)$$

where c_1, c_2 are some absolute constants and $p_1(\beta_0, \lambda/4, K)$ is the probability that a $1 \times K$ row vector with independent $\text{gamma}(\beta_0, 1)$ -distributed entries is a $\lambda/4$ -approximately novel row for the first topic. This can be lower bounded as follows:

$$p_1(\beta_0, c, K) \geq \frac{c_3}{K} \left(\frac{c}{cK + 1 - c} \right)^{\beta_0 K} \quad (4.2)$$

where c_3 is some absolute constant.

We first reduce the analysis of aseparability properties of β to a $W \times K$ dimensional random matrix whose entries are i.i.d gamma distributed. This is a special property of Dirichlet prior. Then computing the probability that β is approximately separable

reduces to examining of the probability that each independent row vector in the related matrix is approximately novel to one of the K topics.

4.1.1 Discussion and Implications of Lemma 12

We discuss some insights and implications that follow from Lemma 12. A direct observation from the lower bounds in Eq. (4.1) is that the probability of β not being λ -approximately separable vanishes exponentially in W , the size of vocabulary, which is typically very large. If we require that the probability that β is not λ -approximately separable should decay at a polynomial rate with respect to W (with K held fixed), i.e., $\frac{2}{W^a}$ for some positive degree $a > 0$, then by Eq. (4.1), it suffices to require that,

$$\frac{W}{\log(W)} \geq (a + 1) / \min\{c_2\beta_0, p_1\} \quad (4.3)$$

If the number of latent topics K also scales, noting that p_1 is a function of K , we need to require that W scale as

$$\frac{W}{\log(W)} \geq (a + 1) \max \left\{ \frac{1}{c_2\beta_0}, \frac{K}{c_3} \left(K - 1 + \frac{1}{\lambda} \right)^{\beta_0 K} \right\} \quad (4.4)$$

Role of hyper-parameter β_0 : Equation (4.4) indicates that if β_0 is moderately small,¹ the topic matrix is more likely to be separable and can be estimated using algorithms, such as those in [Arora et al., 2013, Ding et al., 2013b, 2014b], that come with provable guarantees.

In fact, this implication of our analysis agrees with the practical guidelines adopted in literature to set the hyper-parameters. First, it has been empirically observed that topic models with a smaller β_0 can be more efficiently learned using approximation methods compared to those with a larger β_0 [e.g., Tang et al., 2014]. In the literature,

¹Due to the term $1/c_2\beta_0$ in Eq. (4.4) (or $\exp(-c_2\beta_0W)$ in Eq. (4.1)), extremely small β_0 would makes it difficult for β to be separable. In principle the two terms in Eq. (4.4) lead to an optimal β_0 which is moderately small. In simulation, however, we do not observe such phenomenon. We conjecture this is the artifact of our proof scheme but we do not have a fully treatment now.

the hyper-parameter β_0 is often set to a small positive number [e.g. Blei, 2012, Blei et al., 2003, Newman et al., 2009, Steyvers and Griffiths, Tang et al., 2014]. This is in accordance with our alternative explanation using the separability condition.

Further, a small β_0 can indeed compensate for the exponential dependency of W on K in Eq. (4.4). As reported, empirically satisfactory results are often obtained with $\beta_0 \approx 0.01$ and the number of latent topics ranging from $K = 50 \sim 200$ [Newman et al., 2009, Steyvers and Griffiths, Tang et al., 2014, Wallach et al., 2009]. For these values, the exponent $\beta_0 K$ in Eq. (4.4) would range from 0.5 to 2. Hence the requirement in Eq. (4.4) can be satisfied for moderate values of W .

Finally, we note that in popular topic modeling packages such as McCallum [2002], the default hyper-parameter setting is $\beta_0 = 0.01$. In other packages such as [top, Griffiths and Steyvers, 2004], it is even suggested that the hyper-parameter be set according to the rule $\beta = c/W$, for some constant $c \approx 200$.

Role of approximate separability degree λ : In terms of the degree of approximate separability, i.e., the small constant λ , a scenario of special interest is when the weight (entry in the topic matrix) of each novel word in its corresponding topic is much larger than its cumulative weight in all the remaining topics, e.g., $\sum_{k=2}^K \beta_{i,k} \ll \beta_{i,1}$ if word i is a λ -approximately novel word for topic 1. This translates to $\lambda(K-1) \ll 1$ or $\lambda \ll 1/K$. In this scenario, Eq. (4.4) can be further simplified as,

$$\frac{W}{\log(W)} \geq (a+1) \max \left\{ \frac{1}{c_2 \beta_0}, \frac{K}{\lambda^{\beta_0 K}} \right\}.$$

Validation using Parameters in Benchmark Dataset: We conduct the following simulation so demonstrate that the size of problems in real-world text corpus favors separability. We first obtained the parameters of some benchmark datasets in literature, specifically, the size of the vocabulary W and the number of latent topics specified K . We then generated random realizations of the topic matrix β and

checked if the λ -approximate separability condition is satisfied.

Dataset	Vocab. size W	# Topics K	Prob. 0.01-separable
NIPS	12,419	50	$100 \pm 0\%$
Wikipedia	109,611	50	$99.9 \pm 0.3\%$
Twitter	122,035	50	$100 \pm 0.0\%$
NYT	102,660	100	$99.6 \pm 0.6\%$
PubMed	141,043	100	$99.9 \pm 0.3\%$

Table 4.1: Probability of generating ($\beta_0 = 0.01$) a 0.01-approximately separable β matrix for different W, K values taken from some real-world benchmark topic-modeling datasets. The statistics of Wikipedia and Twitter are from [Tang et al., 2014], NIPS and NYT are from [Arora et al., 2013, Ding et al., 2014b], PubMed are used from [Wallach et al., 2009]. NIPS, NYT, PubMed can also be obtained from [Bache and Lichman, 2013]. The probability is estimated using 1000 Monte Carlo runs. The 3σ -confidence intervals are provided.

As discussed in previous sections, we set $\beta_0 = 0.01$ and consider $\lambda = 0.01$ -approximate separability. For each setting, we generated 1000 Monte Carlo runs to estimate the probability of generating a 0.01-approximately separable matrix. The results are summarized in Table 4.1. We can observe that in most examples, the topic matrix is 0.01-approximately separable with very high probability.

4.2 Separability in Topic Model for Rankings

We consider the ranking matrix β as in the Topic Ranking Model in Figure 3.2. Here $\beta_{(i,j),k} = \mathbb{I}(\sigma^k(i) < \sigma^k(j))$ and σ^k 's are the permutations defining the latent ranking factors. Noting that β is binary in this case, we only consider the exact separability with $\lambda = 0$. We adopt a uniform prior on the columns of β : $\sigma^1, \dots, \sigma^K$ are sampled i.i.d uniformly from the all set of permutations over the Q items. With uniform prior we aim to show most of the ranking matrices in this generative model are exact separable. Formally, we have the following result,

Lemma 13. *Let the K permutations $\sigma^1, \dots, \sigma^K$ be sampled i.i.d uniformly from the set of all permutations over the Q items. Then, the probability that the $W \times K$ binary*

ranking matrix β being exact separable is at least

$$1 - K \exp(-Q \frac{1}{2^K}) \quad (4.5)$$

The key idea of our proof is to consider a subset of rows in β that are independent under the uniform prior. To be specific, two pairs (rows) with distinct items are independent if σ^k 's are uniformly sampled from the set of all permutation. We can then consider a subset of rows such as $(1, 2), (3, 4), \dots, (Q - 1, Q)$ and it suffices for this sub-matrix to be separable.

4.2.1 Discussion and Implications of Lemma 13

We discuss some the insights and implications that follow from Lemma 13. First, following the lower-bound in Eq. 4.5, the probability of β being not exact separable vanishes exponentially in Q , the number of items. In our application scenario Q is considered to be moderately large. Further, if we require that the probability of β being not separable to decay at a polynomial rate with respect to Q with K held fixed, (i.e., $\frac{1}{Q^a}$ for some positive polynomial degree $a > 0$), then, by Eq. 4.5, it suffices to require that,

$$\frac{Q}{\log(Q)} \geq (a + 1)2^K \quad (4.6)$$

We note that [Farias et al., 2009] also considered the setting of K uniformly i.i.d permutations and considered a property which is equivalent to the exact separability. They proposed a different method and also obtained similar results. Specifically, they show if $K = o(\log(Q))$, then the probability of β being non-separable is at most $o(1)$ as $Q \rightarrow \infty$. While it is of the same order as in Eq. 4.6, it is not clear how to guarantee a polynomial or exponential vanishing rate of non-separable probability in Q . In addition, our analysis framework can be applied to other settings such as that in Section 4.3.

A Loose Lower Bound: We note that the result in Lemma 13 is a very loose bound since it only consider a sufficient case of a sub-matrix of β being separable. To validate this we consider the empirical settings in the semi- synthetic data in section 3.5. In this setting, we have $K = 10$ and $Q = 100$. Using 1000 Monte Carlo simulation, the probability of β being separable is $92.9\% \pm 1.6\%$ (95% confidence interval). On the other hand, the probability lower bound in Eq. (4.5) is negative. It remains in further work to improve this lower bound.

4.3 Separability in Mixed Membership Mallows Model

We consider the ranking matrix β of the Mixed Membership Mallows Model (M4) as defined in Figure 3-4. By its definition (see Eq. (3.8)), the k -th column of β is determined by the parameters of the k -th Mallows components, i.e., the reference permutation σ^k and the dispersion parameter ϕ_k . Noting that each entry of β is strictly positive, we consider the separability with some small constant $\lambda > 0$. We considered the following prior on the ranking matrix β . First, the K reference permutations $\sigma^1, \dots, \sigma^K$ are sampled uniformly from all permutations over the Q items. Second, we assume the dispersion parameters to be strictly less than one, i.e., $\phi_k \leq \phi < 1$. By imposing a uniform distribution and minimum assumptions on the dispersion parameters, we aim to show that most M4 models are approximately separable.

Although by definition the entries of β are strictly positive, if the position of i is above j with a large margin in the reference ranking, $\beta_{(j,i),k}$ will be very close to zero. In fact, as in the Proposition 4 c), $\beta_{(j,i),k}$ is close to 0 exponentially in terms of the distance between the position of i, j in the k -th reference ranking. Formally,

Lemma 14. *Let the reference rankings $\sigma_1, \dots, \sigma_K$ be sampled i.i.d uniformly from the set of all permutations, and the dispersion parameters $\phi_k \leq \phi < 1, k = 1, \dots, K$. Then, the probability that the corresponding ranking matrix β being λ -approximately*

separable for any $\lambda \in (0, 1)$ is at least

$$1 - K \exp\left(-\frac{Q}{L(\phi, \lambda)^{2K-1}}\right) \quad (4.7)$$

where $L(\phi, \lambda) = \text{ceil}\left(2\frac{\log(\lambda)}{\log(\phi)}\right)$, and $\text{ceil}(x)$ is the minimum integer no smaller than x .

The key intuition in proving Lemma 14 is again to consider an independent subset of rows from the β matrix. We exploit the fact that for two disjoint subset of items, their relative positions in a reference permutation are independent if the reference permutation is uniformly sampled from all permutations set. We then split the items into disjoint groups of size $L(\phi, \lambda)$ so that two items in the group can be at least L distant away in the reference ranking and the corresponding $\beta_{(i,j),l} < \lambda\beta_{(i,j),k}$, $l \neq k$. We defer the details in the supplementary.

4.4 Discussion and Implication of Lemma 14

We also discuss the implications that follow from Lemma 14. First, by the lower-bound in Eq. (4.7) the probability of β not being λ separable vanishes exponentially in the number of items Q . If we require that the probability of β being not separable to decay at a polynomial rate with respect to Q with K and ϕ held fixed, (i.e., $1/Q^a$ for some positive polynomial degree $a > 0$), then, by Eq. (4.7) it suffices to require,

$$\frac{Q}{\log(Q)} \geq (a + 1)L(\phi, \lambda)^{2K-1} \quad (4.8)$$

We next note that the dependence of the position difference L on ϕ and λ is $\log(\lambda)/\log(\phi)$. $\phi < 1$ and $\lambda < 1$. Note that we are interested in the case when λ is sufficiently small, the logarithm dependency makes L remain relatively small.

A Loose Lower Bound: We again note that Eq. (4.7) is a loose lower bound on the separability probability. To validate this we consider the empirical setting in the semi-synthetic data in section 3.5 for M4 where we have $Q = 100$ and $K = 10$. We set $\phi_1 = \dots \phi_k = \phi$. Following the discussion in section 4.1.1 we set $\lambda = 0.05$.

We use 1000 Monte Carlo simulation to check the the empirical probability of β being λ -approximate separable. Some simulation results are summarized in Table 4.2. However, in each setting, the lower bound in Eq. (4.7) is a negative value. It also remains in further work to improve this lower bound.

ϕ	0.01	0.1	0.2	0.5	0.8
$Q = 100$	$93.3 \pm 1.6\%$	$87.0 \pm 2.1\%$	$79.3 \pm 2.5\%$	$42.6 \pm 3.1\%$	$0 \pm 0\%$
$Q = 200$	$100 \pm 0.0\%$	$100 \pm 0.0\%$	$100 \pm 0.0\%$	$99.8 \pm 0.3\%$	$80.3 \pm 2.5\%$

Table 4.2: Probability of a random β being 0.05-approximate separable with $Q = 100$, $K = 10$ for different values of ϕ . We also consider the case when $Q = 200$. The results are calculated using 1000 Monte Carlo runs. The 3σ -confidence intervals are provided.

Chapter 5

Concluding Remarks and Outlook

This thesis proposed a novel approach for estimating the shared latent factors in a family of Mixed Membership Latent Variable Models. We exploited a natural structural property, separability, which is an inevitable consequence of the high-dimensionality of the observation space. We leveraged a key geometric property that the novel parts of the latent factors correspond to extreme points in a second order co-occurrence representation space. We proposed a random-projection based approach that can learn the latent factors consistently with polynomial computation and sample complexity guarantees. Our approach can be applied to a wide family of MMLVMs whose membership weights prior satisfies some information-theoretical necessary and sufficient conditions.

We applied our approach to two distinct problems, topic modeling for text analysis and user-preference prediction in personalized recommendation systems, which are typically solved using different approaches in the literature. We demonstrated that empirically the performance of our proposed approach can match the state-of-the-arts in the two problems. As a by-product, we gave the first provably consistent and polynomial complexity algorithm for learning mixture of permutation and mixture of Mallows models for which theoretical guarantees were unavailable.

Our projection based approach is especially amenable to the setting where a large number of observations are stored on a network of *distributed* servers. Through theoretical analysis and simulation we show that our distributed implementation can

provably achieve the centralized statistical performance with insignificant communication cost between servers.

5.1 Future Directions

This thesis is a first step towards leveraging the separability property as a key structural property in a wide range of MMLVMs. There are several aspects of this property that are worth a deeper investigation.

Other problems in MMLVMs: In this thesis we have focused on the learning problem, i.e., the problem of estimating the latent factors such as the topic matrix, and established theoretical guarantees. There are, however, other problems such as inference, prediction, and model selection that have not been deeply addressed in this thesis. While these problems are also very important, they are known to be intractable and *NP*-hard. In this thesis, we adopted the standard MCMC based approximations whenever prediction was required. An interesting future direction would be to study whether the separability property can be leveraged to develop consistent and efficient algorithms for inference, prediction, and model selection..

Other Mixed Membership Models: In this thesis, we applied our approach to two types of MMLVMs, namely topic models and the mixed membership ranking models. These two families of models share a number of common features. Most importantly, the probability model of the observation conditioned on the mixing weights are both “bag-of-words” models in a discrete observation space. Extending our approach to other MMLVMs would be another interesting research direction..

Appendix A

Proofs of all Technical Results

A.1 Proof of Lemma 1

Proof. The proof is by contradiction. We will show that if $\bar{\mathbf{R}}$ is non-simplicial, we can construct two topic matrices $\beta^{(1)}$ and $\beta^{(2)}$ whose sets of novel words are not identical and yet \mathbf{X} has the same distribution under both models. The difference between constructed $\beta^{(1)}$ and $\beta^{(2)}$ is not a result of column permutation. This will imply the impossibility of consistent novel word detection.

Suppose $\bar{\mathbf{R}}$ is non-simplicial. Then we can assume, without loss of generality, that its first row is within the convex hull of the remaining rows, i.e., $\bar{\mathbf{R}}_1 = \sum_{j=2}^K c_j \bar{\mathbf{R}}_j$, where $\bar{\mathbf{R}}_j$ denotes the j -th row of $\bar{\mathbf{R}}$, and $c_2, \dots, c_K \geq 0$, $\sum_{j=2}^K c_j = 1$ are convex combination weights. Compactly, $\mathbf{e}^\top \bar{\mathbf{R}} \mathbf{e} = 0$ where $\mathbf{e} := [-1, c_2, \dots, c_K]^\top$. Recalling that $\bar{\mathbf{R}} = \text{diag}(\mathbf{a})^{-1} \mathbf{R} \text{diag}(\mathbf{a})^{-1}$, where \mathbf{a} is a positive vector and $\mathbf{R} = \mathbb{E}(\boldsymbol{\theta}^m \boldsymbol{\theta}^{m\top})$ by definition, we have

$$0 = \mathbf{e}^\top \bar{\mathbf{R}} \mathbf{e} = (\text{diag}(\mathbf{a})^{-1} \mathbf{e})^\top \mathbb{E}(\boldsymbol{\theta}^m \boldsymbol{\theta}^{m\top}) (\text{diag}(\mathbf{a})^{-1} \mathbf{e}) = \mathbb{E}(\|\boldsymbol{\theta}^{m\top} \text{diag}(\mathbf{a})^{-1} \mathbf{e}\|_2^2),$$

which implies that $\boldsymbol{\theta}^{m\top} \text{diag}(\mathbf{a})^{-1} \mathbf{e} \stackrel{a.s.}{=} 0$. From this it follows that if we define two non-negative rows $\mathbf{b}_1 := b [a_1^{-1}, 0, \dots, 0]$ and $\mathbf{b}_2 = b [(1 - \alpha)a_1^{-1}, \alpha c_2 a_2^{-1}, \dots, \alpha c_K a_K^{-1}]$, where $b > 0, 0 < \alpha < 1$ are constants, then $\mathbf{b}_1 \boldsymbol{\theta}^{m\top} \stackrel{a.s.}{=} \mathbf{b}_2 \boldsymbol{\theta}^{m\top}$ for any distribution on $\boldsymbol{\theta}^m$.

Now we construct two separable topic matrices $\beta^{(1)}$ and $\beta^{(2)}$ as follows. Let \mathbf{b}_1 be the first row and \mathbf{b}_2 be the second in $\beta^{(1)}$. Let \mathbf{b}_2 be the first row and \mathbf{b}_1 the second in $\beta^{(2)}$. Let $\mathbf{B} \in \mathbb{R}^{W-2 \times K}$ be a valid separable topic matrix. Set the remaining $(W - 2)$ rows of both $\beta^{(1)}$ and $\beta^{(2)}$ to be $\mathbf{B}(I_K - \text{diag}(\mathbf{b}_1 + \mathbf{b}_2))$. We can choose b to be small enough to ensure that each element of $(\mathbf{b}_1 + \mathbf{b}_2)$ is strictly less than 1. This will ensure that $\beta^{(1)}$ and $\beta^{(2)}$ are column-stochastic and therefore valid separable topic matrices. Observe that \mathbf{b}_2 has at least two non-zero components. Thus, word 1 is novel for $\beta^{(1)}$ but non-novel for $\beta^{(2)}$.

By construction, $\beta^{(1)}\theta \stackrel{a.s.}{=} \beta^{(2)}\theta$, i.e., the distribution of \mathbf{X} conditioned on θ is the same for both models. Marginalizing over θ , the distribution of \mathbf{X} under each topic matrix is the same. Thus no algorithm can consistently distinguish between $\beta^{(1)}$ and $\beta^{(2)}$ based on \mathbf{X} . \square

A.2 Proof of Lemma 2

Proof. The proof is by contradiction. Suppose that $\bar{\mathbf{R}}$ is not affine-independent. Then there exists a $\lambda \neq \mathbf{0}$ with $\mathbf{1}^\top \lambda = 0$ such that $\lambda^\top \bar{\mathbf{R}} = \mathbf{0}$ so that $\lambda^\top \bar{\mathbf{R}} \lambda = 0$. Recalling that $\bar{\mathbf{R}} = \text{diag}(\mathbf{a})^{-1} \mathbf{R} \text{diag}(\mathbf{a})^{-1}$, we have,

$$0 = \lambda^\top \bar{\mathbf{R}} \lambda = (\text{diag}(\mathbf{a})^{-1} \lambda)^\top \mathbb{E}(\theta^m \theta^{m\top}) (\text{diag}(\mathbf{a})^{-1} \lambda) = \mathbb{E}(\|\theta^{m\top} \text{diag}(\mathbf{a})^{-1} \lambda\|^2),$$

which implies that $\theta^{m\top} \text{diag}(\mathbf{a})^{-1} \lambda \stackrel{a.s.}{=} 0$. Since $\lambda \neq \mathbf{0}$, we can assume, without loss of generality, that the first t elements of λ , $\lambda_1, \dots, \lambda_t > 0$, the next s elements of λ , $\lambda_{t+1}, \dots, \lambda_{t+s} < 0$, and the remaining elements are 0 for some $s, t : s > 0, t > 0, s+t \leq K$. Therefore, if we define two non-negative and non-zero row vectors $\mathbf{b}_1 := b[\lambda_1 a_1^{-1}, \dots, \lambda_t a_t^{-1}, 0, \dots, 0]$ and $\mathbf{b}_2 := -b[0, \dots, 0, \lambda_{t+1} a_{t+1}^{-1}, \dots, \lambda_s a_s^{-1}, 0, \dots, 0]$, where $b > 0$ is a constant, then $\mathbf{b}_1 \theta^m \stackrel{a.s.}{=} \mathbf{b}_2 \theta^m$.

Now we construct two topic matrices $\beta^{(1)}$ and $\beta^{(2)}$ as follows. Let \mathbf{b}_1 be the first row and \mathbf{b}_2 the second in β_1 . Let \mathbf{b}_2 be the first row and \mathbf{b}_1 the second in β_2 . Let $\mathbf{B} \in \mathbb{R}^{W-2 \times K}$ be a valid topic matrix and assume that it is **separable**. Set the remaining $(W-2)$ rows of both β_1 and β_2 to be $\mathbf{B}(I_K - \text{diag}(\mathbf{b}_1 + \mathbf{b}_2))$. We can choose b to be small enough to ensure that each element of $(\mathbf{b}_1 + \mathbf{b}_2)$ is strictly less than 1. This will ensure that $\beta^{(1)}$ and $\beta^{(2)}$ are column-stochastic and therefore valid topic matrices. We note that the supports of \mathbf{b}_1 and \mathbf{b}_2 are disjoint and both are non-empty. They appear in distinct topics.

By construction, $\beta^{(1)}\theta \stackrel{a.s.}{=} \beta^{(2)}\theta \Rightarrow$ the distribution of the observation \mathbf{X} conditioned on θ is the same for both models. Marginalizing over θ , the distributions of \mathbf{X} under the topic matrices are the same. Thus no algorithm can distinguish between β_1 and β_2 based on \mathbf{X} . \square

A.3 Proof of Proposition 1 and Proposition 2

(1) $\bar{\mathbf{R}}$ is γ_a -affine-independent $\Rightarrow \bar{\mathbf{R}}$ is at least γ_a -simplicial.

Proof. By definition of affine independence, $\|\sum_{k=1}^K \lambda_k \bar{\mathbf{R}}_k\|_2 \geq \gamma_a \|\boldsymbol{\lambda}\|_2 > 0$ for all $\boldsymbol{\lambda} \in \mathbb{R}^K$ such that $\sum_{k=1}^K \lambda_k = 0$ and $\boldsymbol{\lambda} \neq \mathbf{0}$. If for each $i \in [K]$ we set $\lambda_k = 1$ for $k = i$ and choose $\lambda_k \leq 0, \forall k \neq i$ then (i) $\|\boldsymbol{\lambda}\|_2 \geq 1$, (ii) $\{-\lambda_k, k \neq i\}$ are convex weights, i.e., they are nonnegative and sum to 1, and (iii) $\sum_{k=1}^K \lambda_k \bar{\mathbf{R}}_k = \bar{\mathbf{R}}_i - \sum_{k \neq i} (-\lambda_k) \bar{\mathbf{R}}_k$. Therefore, for all $i \in [K]$, $\|\bar{\mathbf{R}}_i - \sum_{k \neq i} (-\lambda_k) \bar{\mathbf{R}}_k\|_2 \geq \gamma_a > 0$ which proves that $\bar{\mathbf{R}}$ is at least γ_a -simplicial. For the reverse implication, consider

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0 \\ 0.5 & 0.5 & 0 & 1 \end{bmatrix}.$$

It is simplicial but is not affine independent (the $1, 1, -1, -1$ combination of the 4 rows would be $\mathbf{0}$). \square

(2) $\bar{\mathbf{R}}$ is full rank with minimum eigenvalue $\gamma_r \Rightarrow \bar{\mathbf{R}}$ is at least γ_r -affine-independent.

Proof. The Rayleigh-quotient characterization of the minimum eigenvalue of a symmetric, positive-definite matrix $\bar{\mathbf{R}}$ gives $\min_{\boldsymbol{\lambda} \neq \mathbf{0}} \|\boldsymbol{\lambda}^\top \bar{\mathbf{R}}\|_2 / \|\boldsymbol{\lambda}\|_2 = \gamma_r > 0$. Therefore, $\min_{\boldsymbol{\lambda} \neq \mathbf{0}, \mathbf{1}^\top \boldsymbol{\lambda} = 0} \|\boldsymbol{\lambda}^\top \bar{\mathbf{R}}\|_2 / \|\boldsymbol{\lambda}\|_2 \geq \gamma_r > 0$. One can construct examples that contradict the reverse implication:

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

which is affine independent, but not linear independent. \square

(3) $\bar{\mathbf{R}}$ is γ_d -diagonal dominant $\Rightarrow \bar{\mathbf{R}}$ is at least γ_d -simplicial.

Proof. Noting that $\bar{\mathbf{R}}_{i,i} - \bar{\mathbf{R}}_{i,j} \geq \gamma_d > 0$ for all i, j , then the distance of the first row of $\bar{\mathbf{R}}$, $\bar{\mathbf{R}}_1$, to any convex combination of the remaining rows, $\sum_{j=2}^K c_j \bar{\mathbf{R}}_j$, where c_2, \dots, c_K are convex combination weights, can be lower bounded by, $\|\bar{\mathbf{R}}_1 - \sum_{j=2}^K c_j \bar{\mathbf{R}}_j\|_2 \geq |\bar{\mathbf{R}}_{1,1} - \sum_{j=2}^K c_j \bar{\mathbf{R}}_{j,1}| = |\sum_{j=2}^K c_j (\bar{\mathbf{R}}_{1,1} - \bar{\mathbf{R}}_{j,1})| \geq \gamma_d > 0$. Therefore, $\bar{\mathbf{R}}$ is at least γ_d -simplicial. It is straightforward to construct examples that contradict the reverse implication:

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

which is affine independent, hence simplicial, but not diagonal dominant. \square

(4) $\bar{\mathbf{R}}$ being diagonal dominant neither implies nor is implied by $\bar{\mathbf{R}}$ being affine-independent.

Proof. Consider the following two examples:

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

and

$$\bar{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 0.5 & 0.5 \\ 0 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 & 0 \\ 0.5 & 0.5 & 0 & 1 \end{bmatrix}.$$

They are the examples for the two sides of this assertion. \square

A.4 Proof of Lemma 3

Proof. Recall that $\bar{\mathbf{A}} = \bar{\boldsymbol{\beta}}\bar{\boldsymbol{\theta}}$ where $\bar{\mathbf{A}}$ and $\bar{\boldsymbol{\theta}}$ are row-normalized version of \mathbf{A} and $\boldsymbol{\theta}$, $\bar{\boldsymbol{\beta}} := \text{diag}(\mathbf{A}\mathbf{1})^{-1}\boldsymbol{\beta}\text{diag}(\boldsymbol{\theta}\mathbf{1})$. $\bar{\boldsymbol{\beta}}$ is row-stochastic and is separable if $\boldsymbol{\beta}$ is separable. If w is a novel word of topic k , $\bar{\beta}_{wk} = 1$ and $\bar{\beta}_{wj} = 0$, $\forall j \neq k$. We have then $\bar{\mathbf{A}}_w = \bar{\boldsymbol{\theta}}_k$. If w is a non-novel word, $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk}\bar{\boldsymbol{\theta}}_k$ is a convex combination of the rows of $\bar{\boldsymbol{\theta}}$.

It follows directly from the proof of Lemma 1 that $\bar{\boldsymbol{\theta}}$ is simplicial with probability one if $\bar{\mathbf{R}}$ is simplicial. Therefore all the row-vectors of $\bar{\boldsymbol{\theta}}$ are extreme points of the convex hull they formed and this concludes our proof. \square

A.5 Proof of Lemma 4

Proof. By Lemma 3, detecting K distinct novel words for K topics is equivalent to knowing $\bar{\boldsymbol{\theta}}$ up to a row permutation. Noting that $\bar{\mathbf{A}}_w = \sum_k \bar{\beta}_{wk}\bar{\boldsymbol{\theta}}_k$. it follows that $\bar{\beta}_{wk}, k = 1, \dots, K$ is one optimal solution to the following constrained optimization problem:

$$\min \left\| \bar{\mathbf{A}}_w - \sum_{k=1}^K b_k \bar{\boldsymbol{\theta}}_k \right\|^2 \text{ s.t } b_k \geq 0, \sum_{k=1}^K b_k = 1$$

By the proof of Lemma 2, if $\bar{\mathbf{R}}$ is affine-independent, $\bar{\boldsymbol{\theta}}$ is also affine-independent. Therefore, this optimal solution is unique. If this is not true, then there exist two distinct solutions b_1^1, \dots, b_K^1 and b_1^2, \dots, b_K^2 such that $\bar{\mathbf{A}}_w = \sum_{k=1}^K b_k^1 \bar{\boldsymbol{\theta}}_k = \sum_{k=1}^K b_k^2 \bar{\boldsymbol{\theta}}_k$.

$\sum b_k^1 = \sum b_k^2 = 1$. We then obtain $\sum_{k=1}^K (b_k^1 - b_k^2) \bar{\boldsymbol{\theta}}_k = \mathbf{0}$ where the coefficients $b_k^1 - b_k^2$ are not all zero and $\sum_k b_k^1 - b_k^2 = 0$. This contradicts the affine-independence definition.

Finally, we check the renormalization steps. Recall that $\text{diag}(\mathbf{A}\mathbf{1})\bar{\boldsymbol{\beta}} = \boldsymbol{\beta} \text{diag}(\boldsymbol{\theta}\mathbf{1})$. $\text{diag}(\mathbf{A}\mathbf{1})$ can be directly obtained from the observations. So we can first renormalize the rows of $\bar{\boldsymbol{\beta}}$. Removing $\text{diag}(\boldsymbol{\theta}\mathbf{1})$ is then simply a column renormalization operation. Recall that $\boldsymbol{\beta}$ is column-stochastic. It is not necessary to know the exact the value of $\text{diag}(\boldsymbol{\theta}\mathbf{1})$. To sum up, by solving a constrained linear regression followed by suitable row renormalization, we can obtain a unique solution which is the ground truth topic matrix. This concludes the proof of Lemma 4. \square

A.6 Proof of Lemma 5

Lemma 5 establishes the second order co-occurrence estimator in Eq. (2.1). We first provide a generic method to establish the explicit convergence bound for a function $\psi(\mathbf{X})$ of d random variables X_1, \dots, X_d , then apply it to establish Lemma 5

Proposition 5. *Let $\mathbf{X} = [X_1, \dots, X_d]$ be d random variables and $\mathbf{a} = [a_1, \dots, a_d]$ be positive constants. Let $\mathcal{E} := \bigcup_{i \in \mathcal{I}} \{|X_i - a_i| \geq \delta_i\}$ for some constants $\delta_i > 0$, and $\psi(\mathbf{X})$ be a continuously differentiable function in $\mathcal{C} := \mathcal{E}^c$. If for $i = 1, \dots, d$, $\Pr(|X_i - a_i| \geq \epsilon) \leq f_i(\epsilon)$ are the individual convergence rates and $\max_{\mathbf{X} \in \mathcal{C}} |\partial_i \psi(\mathbf{X})| \leq C_i$, then,*

$$\Pr(|\psi(\mathbf{X}) - \psi(\mathbf{a})| \geq \epsilon) \leq \sum_i f_i(\gamma) + \sum_{i=1}^d f_i\left(\frac{\epsilon}{dC_i}\right)$$

Proof. Since $\psi(\mathbf{X})$ is continuously differentiable in \mathcal{C} , $\forall \mathbf{X} \in \mathbf{C}, \exists \lambda \in (0, 1)$ such that

$$\psi(\mathbf{X}) - \psi(\mathbf{a}) = \nabla^\top \psi((1 - \lambda)\mathbf{a} + \lambda\mathbf{X}) \cdot (\mathbf{X} - \mathbf{a})$$

Therefore,

$$\begin{aligned} & \Pr(|\psi(\mathbf{X}) - \psi(\mathbf{a})| \geq \epsilon) \\ & \leq \Pr(\mathbf{X} \in \mathcal{E}) + \Pr\left(\sum_{i=1}^d |\partial_i \psi((1 - \lambda)\mathbf{a} + \lambda\mathbf{X})| |X_i - a_i| \geq \epsilon \mid \mathbf{X} \in \mathcal{C}\right) \end{aligned}$$

$$\begin{aligned}
&\leq \sum_{i \in \mathcal{I}} \Pr(|X_i - a_i| \geq \delta_i) + \sum_{i=1}^d \Pr(\max_{\mathbf{x} \in \mathcal{C}} |\partial_i \psi(\mathbf{x})| |X_i - a_i| \geq \epsilon/d) \\
&= \sum_{i \in \mathcal{I}} f_i(\delta_i) + \sum_{i=1}^d f_i\left(\frac{\epsilon}{dC_i}\right)
\end{aligned}$$

□

Now we turn to prove Lemma 5. Recall that $\bar{\mathbf{X}}$ and $\bar{\mathbf{X}}'$ are obtained from \mathbf{X} by first splitting each user's comparisons into two independent halves and then re-scaling the rows to make them row-stochastic hence $\bar{\mathbf{X}} = \text{diag}^{-1}(\mathbf{X}\mathbf{1})\mathbf{X}$. Also recall that $\bar{\boldsymbol{\beta}} = \text{diag}^{-1}(\boldsymbol{\beta}\mathbf{a})\boldsymbol{\beta} \text{diag}(\mathbf{a})$, $\bar{\mathbf{R}} = \text{diag}^{-1}(\mathbf{a})\mathbf{R} \text{diag}^{-1}(\mathbf{a})$, and $\bar{\boldsymbol{\beta}}$ is row stochastic. For any $1 \leq i, j \leq W$,

$$\begin{aligned}
\hat{E}_{i,j} &= M \frac{1}{\sum_{m=1}^M X'_{i,m}} \left(\sum_{m=1}^M X'_{i,m} X_{j,m} \right) \frac{1}{\sum_{m=1}^M X_{j,m}} = \frac{1/M \sum_{m=1}^M (X'_{i,m} X_{j,m})}{(1/M \sum_{m=1}^M X'_{i,m})(1/M \sum_{m=1}^M X_{j,m})} \\
&= \frac{\frac{1}{MN^2} \sum_{m=1, n=1, n'=1}^{M, N, N} \mathbb{I}(w_{m,n} = i) \mathbb{I}(w'_{m,n'} = j)}{\frac{1}{MN} \sum_{m=1, n=1}^{M, N} \mathbb{I}(w_{m,n} = i) \frac{1}{MN} \sum_{m=1, n=1}^{M, N} \mathbb{I}(w'_{m,n} = i)} \\
&:= \frac{F_{i,j}(M, N)}{G_i(M, N) H_j(M, N)}
\end{aligned}$$

From the Strong Law of Large Numbers and the generative topic modeling procedure,

$$\begin{aligned}
F_{i,j}(M, N) &\xrightarrow{a.s.} \mathbb{E}(\mathbb{I}(w_{m,n} = i) \mathbb{I}(w'_{m,n'} = j)) = (\boldsymbol{\beta}\mathbf{R}\boldsymbol{\beta}^\top)_{i,j} := p_{i,j} \\
G_i(M, N) &\xrightarrow{a.s.} \mathbb{E}(\mathbb{I}(w'_{m,n} = i)) = (\boldsymbol{\beta}\mathbf{a})_i := p_i \\
H_j(M, N) &\xrightarrow{a.s.} \mathbb{E}(\mathbb{I}(w_{m,n} = j)) = (\boldsymbol{\beta}\mathbf{a})_j := p_j
\end{aligned}$$

and $\frac{(\boldsymbol{\beta}\mathbf{R}\boldsymbol{\beta}^\top)_{i,j}}{(\boldsymbol{\beta}\mathbf{a})_i(\boldsymbol{\beta}\mathbf{a})_j} = \mathbf{E}_{i,j}$ by definition. Using McDiarmid's inequality, we obtain

$$\Pr(|F_{i,j} - p_{i,j}| \geq \epsilon) \leq 2 \exp(-\epsilon^2 MN)$$

$$\Pr(|G_i - p_i| \geq \epsilon) \leq 2 \exp(-2\epsilon^2 MN)$$

$$\Pr(|H_j - p_j| \geq \epsilon) \leq 2 \exp(-2\epsilon^2 MN)$$

In order to calculate $\Pr\{|\frac{F_{i,j}}{G_i H_j} - \frac{p_{i,j}}{p_i p_j}| \geq \epsilon\}$, we apply the results from Proposition 5.

Let $\psi(x_1, x_2, x_3) = \frac{x_1}{x_2 x_3}$ with $x_1, x_2, x_3 > 0$, and $a_1 = p_{i,j}$, $a_2 = p_i$, $a_3 = p_j$. Let

$\mathcal{I} = \{2, 3\}$, $\delta_2 = \gamma p_i$, and $\delta_3 = \gamma p_j$. Then $|\partial_1 \psi| = \frac{1}{x_2 x_3}$, $|\partial_2 \psi| = \frac{x_1}{x_2^2 x_3}$, and $|\partial_3 \psi| = \frac{x_1}{x_2 x_3^2}$.

If $F_{i,j} = x_1$, $G_i = x_2$, and $H_j = x_3$, then $F_{i,j} \leq G_i$, $F_{i,j} \leq H_j$. Then note that

$$\begin{aligned} C_1 &= \max_c |\partial_1 \psi| = \max_c \frac{1}{G_i H_j} \leq \frac{1}{(1-\gamma)^2 p_i p_j} \\ C_2 &= \max_c |\partial_2 \psi| = \max_c \frac{F_{i,j}}{G_i^2 H_j} \leq \max_c \frac{1}{G_i H_j} \leq \frac{1}{(1-\gamma)^2 p_i p_j} \\ C_3 &= \max_c |\partial_3 \psi| = \max_c \frac{F_{i,j}}{G_i H_j^2} \leq \max_c \frac{1}{G_i H_j} \leq \frac{1}{(1-\gamma)^2 p_i p_j} \end{aligned}$$

By applying Proposition 5, we get

$$\begin{aligned} &\Pr\left\{\left|\frac{F_{i,j}}{G_i H_j} - \frac{p_{i,j}}{p_i p_j}\right| \geq \epsilon\right\} \\ &\leq \exp(-2\gamma^2 p_i^2 MN) + \exp(-2\gamma^2 p_j^2 MN) + 2 \exp(-\epsilon^2 (1-\gamma)^4 (p_i p_j)^2 MN/9) \\ &\quad + 4 \exp(-2\epsilon^2 (1-\gamma)^4 (p_i p_j)^2 MN/9) \\ &\leq 2 \exp(-2\gamma^2 \eta^2 MN) + 6 \exp(-\epsilon^2 (1-\gamma)^4 \eta^4 MN/9) \end{aligned}$$

where $\eta = \min_{1 \leq i \leq W} p_i$. There are many strategies for optimizing the free parameter γ . We set $2\gamma^2 = \frac{(1-\gamma)^4}{9}$ and solve for γ to obtain

$$\Pr\left\{\left|\frac{F_{i,j}}{G_i H_j} - \frac{p_{i,j}}{p_i p_j}\right| \geq \epsilon\right\} \leq 8 \exp(-\epsilon^2 \eta^4 MN/20)$$

Finally, by applying the union bound to the W^2 entries in $\widehat{\mathbf{E}}$, we obtain the results.

A.7 Proof of Lemma 6

Proof. We first show that when $\bar{\mathbf{R}}$ is γ_s simplicial and $\boldsymbol{\beta}$ is separable, then $\mathbf{Y} = \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top$ is at least γ_s -simplicial. Without loss of generality we assume that word $1, \dots, K$ are the novel words for topic 1 to K . By definition, $\bar{\boldsymbol{\beta}}^\top = [\mathbf{I}_K, \mathbf{B}]$ hence $\mathbf{Y} = \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top = [\bar{\mathbf{R}}, \bar{\mathbf{R}}\mathbf{B}]$. Therefore, for convex combination weights $c_2, \dots, c_K \geq 0$ such that $\sum_{j=2}^K c_j = 1$,

$$\|\mathbf{Y}_1 - \sum_{j=2}^K c_j \mathbf{Y}_j\| \geq \|\bar{\mathbf{R}}_1 - \sum_{j=2}^K c_j \bar{\mathbf{R}}_j\| \geq \gamma_s > 0$$

Therefore the first row vector \mathbf{Y}_1 is at least γ_s distant away from the convex hull of the remaining rows. Similarly, any row of \mathbf{Y} is at least γ_s distant away from the convex hull of the remaining rows hence \mathbf{Y} is at least γ_s simplicial. The rest of the proof will be exactly the same as for Lemma 6. \square

A.8 Proof of Lemma 7

Proof. We first show that when $\bar{\mathbf{R}}$ is γ_a affine independent and $\boldsymbol{\beta}$ is separable, then $\mathbf{Y} = \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top$ is at least γ_a affine independent. Similarly as in the proof of Lemma 6, we assume that word $1, \dots, K$ are the novel words for topic 1 to K . By definition, $\bar{\boldsymbol{\beta}}^\top = [\mathbf{I}_K, \mathbf{B}]$ hence $\mathbf{Y} = \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top = [\bar{\mathbf{R}}, \bar{\mathbf{R}}\mathbf{B}]$. $\forall \boldsymbol{\lambda} \in \mathbb{R}^K$ such that $\boldsymbol{\lambda} \neq \mathbf{0}$, $\sum_{k=1}^K \lambda_k = 0$, then,

$$\left\| \sum_{k=1}^K \mathbf{Y}_k \right\|_2 / \|\boldsymbol{\lambda}\|_2 \geq \left\| \sum_{k=1}^K \bar{\mathbf{R}}_k \right\|_2 / \|\boldsymbol{\lambda}\|_2 \geq \gamma_a$$

Hence \mathbf{Y} is affine independent. The The rest of the proof will be exactly the same as that for Lemma 4.

We note that once the novel words for K topics are detection, we can use only the corresponding columns of \mathbf{E} for linear regression. Formally, let \mathbf{E}^* be the $W \times K$ matrix formed by the columns of the \mathbf{E} that correspond to K distinct novel words. Then, $\mathbf{E}^* = \bar{\boldsymbol{\beta}}\bar{\mathbf{R}}$. The rest of the proof is again the same as that for Lemma 4. \square

A.9 Proof of Lemma 8

Proof. We first check that if $q_w > 0$, w must be a novel word. Without loss of generality let word $1, \dots, K$ be novel words for K distinct topics. $\forall w, \mathbf{E}_w = \sum \bar{\beta}_{wk} \mathbf{E}_k$. $\forall \mathbf{d} \in \mathbb{R}^W$,

$$\langle \mathbf{E}_w, \mathbf{d} \rangle = \sum \bar{\beta}_{wk} \langle \mathbf{E}_k, \mathbf{d} \rangle \leq \max_k \langle \mathbf{E}_k, \mathbf{d} \rangle$$

and the last equality holds if, and only if, there exist some k such that $\bar{\beta}_{wk} = 1$ which implies w is a novel words.

We then show that for a novel word w , $q_w > 0$. We need to show for each topic k , when \mathbf{d} is sampled from an isotropic distribution in \mathbf{R}^W , there exist a set of directions \mathbf{d} with non-zero probability such that $\langle \mathbf{E}_k, \mathbf{d} \rangle > \langle \mathbf{E}_l, \mathbf{d} \rangle$ for $l = 1, \dots, K, l \neq k$. First, one can check by definition that $\mathbf{Y} = (\mathbf{E}_1^\top, \dots, \mathbf{E}_K^\top)^\top = \bar{\mathbf{R}} \bar{\boldsymbol{\beta}}^\top$ is at least γ_s -simplicial if $\bar{\mathbf{R}}$ is γ_s -simplicial. Let \mathbf{E}_1^* be the projection of \mathbf{E}_1 onto the simplex formed by the remaining row vectors $\mathbf{E}_2, \dots, \mathbf{E}_K$. By the orthogonality principle, $\langle \mathbf{E}_1 - \mathbf{E}_1^*, \mathbf{E}_k - \mathbf{E}_1^* \rangle \leq 0$ for $k = 2, \dots, K$. Therefore, for $\mathbf{d}^1 = \mathbf{E}_1^\top - \mathbf{E}_1^{*\top}$,

$$\mathbf{E}_1 \mathbf{d}^1 - \mathbf{E}_k \mathbf{d}^1 = \|\mathbf{d}^1\|^2 - (\mathbf{E}_k - \mathbf{E}_1^*) \mathbf{d}^1 \geq \gamma_s^2 > 0$$

Due to the continuity of the inner product, there exist a neighbor on the unite sphere around $\mathbf{d}^1 / \|\mathbf{d}^1\|_2$ that \mathbf{E}_1 has maximum projection value. This conclude our proof. \square

A.10 Proof of Theorem 2

Proof. We first consider the random projection steps (step 3 to 12 in Alg. 2). For projection along direction \mathbf{d}^r , we first calculate projection values $\mathbf{r} = \bar{\mathbf{X}}' \bar{\mathbf{X}}^\top \mathbf{d}^r$, find the maximizer index i^* and the corresponding set \hat{J}_{i^*} , and then evaluate $\mathbb{I}(\forall j \in \hat{J}_w, v_w > v_j)$ for all the words w in $\hat{J}_{i^*}^c = \{1, \dots, W\} \setminus \hat{J}_{i^*}$. (I) The set $\hat{J}_{i^*}^c$ have up to $|\mathcal{C}_k|$ elements asymptotically, where k is the topic associated with word i^* . This is considered a small constant $\mathcal{O}(1)$; (II) Note that $\hat{\mathbf{E}} \mathbf{d}_r = M \bar{\mathbf{X}}' (\bar{\mathbf{X}}^\top \mathbf{d}_r)$ and each column of $\bar{\mathbf{X}}$ has at most $N \ll W$ non-zero entries. Calculating the $W \times 1$ projection value vector \mathbf{v} requires two sparse matrix-vector multiplications and takes $\mathcal{O}(MN)$ time. Finding the maximum requires \mathbf{W} running time; (III) To evaluate one set $\hat{J}_i \leftarrow \{j : \hat{E}_{i,i} + \hat{E}_{j,j} - 2\hat{E}_{i,j} \geq \zeta/2\}$ we need to calculate $\hat{E}_{i,j}, j = 1, \dots, W$. This can be viewed as projecting $\hat{\mathbf{E}}$ along $\mathbf{d} = \mathbf{e}_i$ and takes $\mathcal{O}(MN)$. We also note that

the diagonal entries $\mathbf{E}_{w,w}, w = 1, \dots, W$ can be calculated once using $\mathcal{O}(W)$ time. To sum up, these steps takes $\mathcal{O}(MNP + WP)$ running time.

We then consider the detecting and clustering steps (step 14 to 21 in Alg. 2). We note that all the conditions in Step 17 have been calculated in the previous steps, and recall that the number of novel words are small constant per topic, then, this step will require a running time of $\mathcal{O}(K^2)$.

We last consider the topic estimation steps in Algorithm 3. Here all the corresponding inputs for the linear regression have already been computed in the projection step. Each linear regression has K variables and we upper bound its running time by $\mathcal{O}(K^3)$. Calculating the row-normalization factors $\frac{1}{M}\mathbf{X}\mathbf{1}$ requires $\mathcal{O}(MN)$ time. The row and column re-normalization each requires at most $\mathcal{O}(WK)$ running time. Overall, we need a $\mathcal{O}(WK^3 + MN)$ running time.

Other steps are also efficient. Splitting each document into two independent halves takes linear time in N for each document since we can achieve it using random permutation over N items. To generate each random direction \mathbf{d}_r requires $\mathcal{O}(W)$ complexity if we use the spherical Gaussian prior. While we can directly sort the empirical estimated solid angles (in $\mathcal{O}(W \log(W))$ time), we only search for the words with largest solid angles whose number is a constant w.r.t W , therefore it would take only $\mathcal{O}(W)$ time. \square

A.11 Proof of Theorem 3

We focus on the case when the random projection directions are sampled from **any** isotropic distribution. Our proof is not tied to the special form of the distribution; just its isotropic nature. We first provide some useful propositions. We denote by \mathcal{C}_k the set of all novel word of topic k , for $k \in [K]$, and denote by \mathcal{C}_0 the set of all non-novel words. We first show,

Proposition 6. *Let \mathbf{E}_i be the i -th row of \mathbf{E} . Suppose β is separable and $\bar{\mathbf{R}}$ is γ_s -simplicial, then the following is true: For all $k = 1, \dots, K$,*

	$\ \mathbf{E}_i - \mathbf{E}_j\ $	$E_{i,i} - 2E_{i,j} + E_{j,j}$
$i \in \mathcal{C}_1, j \in \mathcal{C}_1$	0	0
$i \in \mathcal{C}_1, j \notin \mathcal{C}_1$	$\geq (1-b)\gamma_s$	$\geq (1-b)^2\gamma_s^2/\lambda_{\max}$

where $b = \max_{j \in \mathcal{C}_{0,k}} \bar{\beta}_{j,k}$ and $\lambda_{\max} > 0$ is the maximum eigenvalue of $\bar{\mathbf{R}}$

Proof. We focus on the case $k = 1$ since the proofs for other values of k are analogous. Let $\bar{\boldsymbol{\beta}}_i$ be the i -th row vector of matrix $\bar{\boldsymbol{\beta}}$. To show the above results, recall that $\mathbf{E} = \bar{\boldsymbol{\beta}}\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top$. Then

$$\begin{aligned}\|\mathbf{E}_i - \mathbf{E}_j\| &= \|(\bar{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}_j)\bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top\| \\ E_{i,i} - 2E_{i,j} + E_{j,j} &= (\bar{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}_j)\mathbf{R}'(\bar{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}_j)^\top.\end{aligned}$$

It is clear that when $i, j \in \mathcal{C}_1$, i.e., they are both novel word for the same topic, $\bar{\boldsymbol{\beta}}_i = \bar{\boldsymbol{\beta}}_j = \mathbf{e}_1$. Hence, $\|\mathbf{E}_i - \mathbf{E}_j\| = 0$ and $E_{i,i} - 2E_{i,j} + E_{j,j} = 0$. When $i \in \mathcal{C}_1, j \notin \mathcal{C}_1$, we have $\bar{\boldsymbol{\beta}}_i = [1, 0, \dots, 0]$, $\bar{\boldsymbol{\beta}}_j = [\bar{\beta}_{j,1}, \bar{\beta}_{j,2}, \dots, \bar{\beta}_{j,K}]$ with $\bar{\beta}_{j,1} < 1$. Then,

$$\bar{\boldsymbol{\beta}}_i - \bar{\boldsymbol{\beta}}_j = [1 - \bar{\beta}_{j,1}, -\bar{\beta}_{j,2}, \dots, -\bar{\beta}_{j,K}] = (1 - \bar{\beta}_{j,1})[1, -c_2, \dots, -c_K] := (1 - \bar{\beta}_{j,1})\mathbf{e}^\top$$

and $\sum_{l=2}^K c_l = 1$. Therefore, defining $\mathbf{Y} := \bar{\mathbf{R}}\bar{\boldsymbol{\beta}}^\top$, we get

$$\|\mathbf{E}_i - \mathbf{E}_j\|_2 = (1 - \bar{\beta}_{j,1})\|\mathbf{Y}_1 - \sum_{l=2}^K c_l \mathbf{Y}_l\|_2$$

Noting that \mathbf{Y} is at least γ_s -simplicial, we have $\|\mathbf{E}_i - \mathbf{E}_j\|_2 \geq (1 - b)\gamma_s$ where $b = \max_{j \in \mathcal{C}_0, k} \bar{\beta}_{j,k} < 1$.

Similarly, note that $\|\mathbf{e}^\top \bar{\mathbf{R}}\| \geq \gamma$ and let $\bar{\mathbf{R}} = \mathbf{U}\Sigma\mathbf{U}^\top$ be its singular value decomposition. If λ_{\max} is the maximum eigenvalue of $\bar{\mathbf{R}}$, then we have

$$E_{i,i} - 2E_{i,j} + E_{j,j} = (1 - \bar{\beta}_{j,1})^2(\mathbf{e}^\top \bar{\mathbf{R}})\mathbf{U}\Sigma^{-1}\mathbf{U}^\top(\mathbf{e}^\top \bar{\mathbf{R}})^\top \geq (1 - b)^2\gamma_s^2/\lambda_{\max}.$$

The inequality in the last step follows from the observation that $\mathbf{e}^\top \bar{\mathbf{R}}$ is within the column space spanned by \mathbf{U} . \square

The results in Proposition 6 provide two statistics for identifying novel words of the same topic, $\|\mathbf{E}_i - \mathbf{E}_j\|$ and $E_{i,i} - 2E_{i,j} + E_{j,j}$. While the first is straightforward, the latter is efficient to calculate in practice with better computational complexity. Specifically, its empirical version, the set \mathcal{J}_i in Algorithm 2

$$\mathcal{J}_i = \{j : \hat{E}_{i,i} - \hat{E}_{i,j} - \hat{E}_{j,i} + \hat{E}_{j,j} \geq d/2\}$$

can be used to discover the set of novel words of the same topics asymptotically. Formally,

Proposition 7. *If $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq (1-b)^2\gamma_s^2/8\lambda_{\max}$, then,*

1. *For a novel word $i \in \mathcal{C}_k$, $\mathcal{J}_i = \mathcal{C}_k^c$*
2. *For a non-novel word $j \in \mathcal{C}_0$, $\mathcal{J}_i \supset \mathcal{C}_k^c$*

Now we start to show that Algorithm 2 can detect all the novel words of the K distinct rankings consistently. As illustrated in Lemma 8, we detect the novel words by ranking ordering the solid angles q_i . We denote the minimum solid angle of the K extreme points by q_\wedge . Our proof is to show that the estimated solid angle in Eq (2.5),

$$\hat{p}_i = \frac{1}{P} \sum_{r=1}^P \mathbb{I}\{\forall j \in \mathcal{J}_i, \widehat{\mathbf{E}}_j \mathbf{d}^r \leq \widehat{\mathbf{E}}_i \mathbf{d}^r\} \quad (\text{A.1})$$

converges to the ideal solid angle

$$q_i = \Pr\{\forall j \in \mathcal{S}(i), (\mathbf{E}_i - \mathbf{E}_j) \mathbf{d} \geq 0\} \quad (\text{A.2})$$

as $M, P \rightarrow \infty$. $\mathbf{d}^1, \dots, \mathbf{d}^P$ are iid directions drawn from a isotropic distribution. For a novel word $i \in \mathcal{C}_k, k = 1, \dots, K$, let $\mathcal{S}(i) = \mathcal{C}_k^c$, and for a non-novel word $i \in \mathcal{C}_0$, let $\mathcal{S}(i) = \mathcal{C}_0^c$.

To show the convergence of \hat{p}_i to p_i , we consider an intermediate quantity,

$$p_i(\widehat{\mathbf{E}}) = \Pr\{\forall j \in \mathcal{J}_i, (\widehat{\mathbf{E}}_i - \widehat{\mathbf{E}}_j) \mathbf{d} \geq 0\}$$

First, by Hoeffding's lemma, we have the following result.

Proposition 8. $\forall t \geq 0, \forall i$,

$$\Pr\{|\hat{p}_i - p_i(\widehat{\mathbf{E}})| \leq t\} \geq 2 \exp(-2Pt^2) \quad (\text{A.3})$$

Next we show the convergence of $p_i(\widehat{\mathbf{E}})$ to solid angle q_i :

Proposition 9. Consider the case when $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \frac{d}{8}$ and $\bar{\mathbf{R}}$ is γ_s -simplicial. If i is a novel word, then,

$$q_i - p_i(\widehat{\mathbf{E}}) \leq \frac{W\sqrt{W}}{\pi d_2} \|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty$$

Similarly, if j is a non-novel word, we have,

$$p_j(\widehat{\mathbf{E}}) - q_j \leq \frac{W\sqrt{W}}{\pi d_2} \|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty$$

where $d_2 \triangleq (1-b)\gamma_s$, $d = (1-b)^2\gamma_s^2/\lambda_{\max}$.

Proof. First note that, by the definition of \mathcal{J}_i and Proposition 6, if $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \frac{d}{8}$, then, for a novel word $i \in \mathcal{C}_k$, $\mathcal{J}_i = \mathcal{S}(i)$. And for a non-novel word $i \in \mathcal{C}_0$, $\mathcal{J}_i \supseteq \mathcal{S}(i)$. For convenience, let

$$\begin{aligned} A_j &= \{\mathbf{d} : (\widehat{\mathbf{E}}_i - \widehat{\mathbf{E}}_j)\mathbf{d} \geq 0\} & A &= \bigcap_{j \in \mathcal{J}_i} A_j \\ B_j &= \{\mathbf{d} : (\mathbf{E}_i - \mathbf{E}_j)\mathbf{d} \geq 0\} & B &= \bigcap_{j \in \mathcal{S}(i)} B_j \end{aligned}$$

For i being a novel word, we consider

$$q_i - p_i(\widehat{\mathbf{E}}) = \Pr\{B\} - \Pr\{A\} \leq \Pr\{B \cap A^c\}$$

Note that $\mathcal{J}_i = \mathcal{S}(i)$ when $\|\widehat{\mathbf{E}} - \mathbf{E}\| \leq d/8$,

$$\begin{aligned} \Pr\{B \cap A^c\} &= \Pr\{B \cap (\bigcup_{j \in \mathcal{S}(i)} A_j^c)\} \leq \sum_{j \in \mathcal{S}(i)} \Pr\{(\bigcap_{l \in \mathcal{S}(i)} B_l) \cap A_j^c\} \leq \sum_{j \in \mathcal{S}(i)} \Pr\{B_j \cap A_j^c\} \\ &= \sum_{j \in \mathcal{S}(i)} \Pr\{(\widehat{\mathbf{E}}_i - \widehat{\mathbf{E}}_j)\mathbf{d} < 0, \text{ and } (\mathbf{E}_i - \mathbf{E}_j)\mathbf{d} \geq 0\} = \sum_{j \in \mathcal{S}(i)} \frac{\phi_j}{2\pi} \end{aligned}$$

where ϕ_j is the angle between $\mathbf{e}_j = \mathbf{E}_i - \mathbf{E}_j$ and $\widehat{\mathbf{e}}_j = \widehat{\mathbf{E}}_i - \widehat{\mathbf{E}}_j$ for any isotropic distribution on \mathbf{d} . Noting that $\phi \leq \tan(\phi)$,

$$\Pr\{B \cap A^c\} \leq \sum_{j \in \mathcal{S}(i)} \frac{\tan(\phi_j)}{2\pi} \leq \sum_{j \in \mathcal{S}(i)} \frac{1}{2\pi} \frac{\|\widehat{\mathbf{e}}_j - \mathbf{e}_j\|_2}{\|\mathbf{e}_j\|_2} \leq \frac{W\sqrt{W}}{\pi d_2} \|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty$$

where the last inequality is obtained by the relationship between the ℓ_∞ norm and

the ℓ_2 norm, and the fact that for $j \in \mathcal{S}(i)$, $\|\mathbf{e}_j\|_2 = \|\mathbf{E}_i - \mathbf{E}_j\|_2 \geq d_2 \triangleq (1-b)\gamma_s$. Therefore for a novel word i , we have,

$$q_i - p_i(\hat{\mathbf{E}}) \leq \frac{W\sqrt{W}}{\pi d_2} \|\hat{\mathbf{E}} - \mathbf{E}\|_\infty$$

Similarly for a non-novel word $i \in \mathcal{C}_0$, $\mathcal{J}_i \supseteq \mathcal{S}(i)$,

$$\begin{aligned} p_i(\hat{\mathbf{E}}) - q_i &= \Pr\{A\} - \Pr\{B\} = \Pr\{A \cap B^c\} \leq \sum_{j \in \mathcal{S}(i)} \Pr\left\{\left(\bigcap_{l \in \hat{\mathcal{S}}(i)} A_l\right) \cap B_j^c\right\} \\ &\leq \sum_{j \in \mathcal{S}(i)} \Pr\{A_j \cap B_j^c\} \leq \frac{W\sqrt{W}}{\pi d_2} \|\hat{\mathbf{E}} - \mathbf{E}\|_\infty \end{aligned}$$

□

A direct implication of Proposition 9 is,

Proposition 10. $\forall \epsilon > 0$, let $\rho = \min\{\frac{d}{8}, \frac{\pi d_2 \epsilon}{W^{1.5}}\}$. If $\|\hat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho$, then, $q_i - p_i(\hat{\mathbf{E}}) \leq \epsilon$ for a novel word i and $p_j(\hat{\mathbf{E}}) - q_j \leq \epsilon$ for a non-novel word j .

We now prove Theorem 3. In order to correctly detect all the novel words of K distinct topics, we decompose the error event to be the union of the following two types,

1. *Sorting error*, i.e., $\exists i \in \bigcup_{k=1}^K \mathcal{C}_k, \exists j \in \mathcal{C}_0$ such that $\hat{p}_i < \hat{p}_j$. This event is denoted as $A_{i,j}$ and let $A = \bigcup A_{i,j}$.
2. *Clustering error*, i.e., $\exists k, \exists i, j \in \mathcal{C}_k$ such that $i \notin \mathcal{J}_j$. This event is denoted as $B_{i,j}$ and let $B = \bigcup B_{i,j}$

We point out that the event A, B are different from the notations we used in Proposition 9. According to Proposition 10, we also define $\rho = \min\{\frac{d}{8}, \frac{\pi d_2 q_\Delta}{4W^{1.5}}\}$ and the event that $C = \{\|\mathbf{E} - \hat{\mathbf{E}}\|_\infty \geq \rho\}$. We note that $B \subsetneq C$.

Therefore,

$$Pe = \Pr\{A \cup B\} \leq \Pr\{A \cap C^c\} + \Pr\{C\}$$

$$\begin{aligned}
&\leq \sum_{i \text{ novel}, j \text{ non-novel}} \Pr\{A_{i,j} \cap B^c\} + \Pr\{C\} \\
&\leq \sum_{i,j} \Pr(\hat{p}_i - \hat{p}_j < 0 \cap \|\hat{\mathbf{E}} - \mathbf{E}\|_\infty \geq \rho) + \Pr(\|\hat{\mathbf{E}} - \mathbf{E}\|_\infty > \rho)
\end{aligned}$$

The second term can be bound by Lemma 5. Now we focus on the first term. Note that

$$\begin{aligned}
\hat{p}_i - \hat{p}_j &= \hat{p}_i - \hat{p}_j - p_i(\hat{\mathbf{E}}) + p_i(\hat{\mathbf{E}}) - q_i + q_i - p_j(\hat{\mathbf{E}}) + p_j(\hat{\mathbf{E}}) - q_j + q_j \\
&= \{\hat{p}_i - p_i(\hat{\mathbf{E}})\} + \{p_i(\hat{\mathbf{E}}) - q_i\} + \{p_j(\hat{\mathbf{E}}) - \hat{p}_j\} + \{q_j - p_j(\hat{\mathbf{E}})\} + q_i - q_j
\end{aligned}$$

and the fact that $q_i - q_j \geq q_\wedge$, then,,

$$\begin{aligned}
&\Pr(\hat{p}_i < \hat{p}_j \cap \|\hat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho) \\
&\leq \Pr(p_i(\hat{\mathbf{E}}) - \hat{p}_i \geq q_\wedge/4) + \Pr(\hat{p}_j - p_j(\hat{\mathbf{E}}) \geq q_\wedge/4) \\
&\quad + \Pr(q_i - p_i(\hat{\mathbf{E}}) \geq q_\wedge/4) \cap \|\hat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho) \\
&\quad + \Pr(p_j(\hat{\mathbf{E}}) - q_j \geq q_\wedge/4) \cap \|\hat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho) \\
&\leq 2 \exp(-Pq_\wedge^2/8) + \Pr(q_i - p_i(\hat{\mathbf{E}}) \geq q_\wedge/4) \cap \|\hat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho) \\
&\quad + \Pr(p_j(\hat{\mathbf{E}}) - q_j \geq q_\wedge/4) \cap \|\hat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho)
\end{aligned}$$

The last equality is by Proposition 8. For the last two terms, by Proposition 10 is 0.

Therefore, applying Lemma 5 we obtain,

$$Pe \leq 2W^2 \exp(-Pq_\wedge^2/8) + 8W^2 \exp(-\rho^2 \eta^4 MN/20)$$

And this concludes Theorem 3.

A.12 Proof of Theorem 4

Without loss of generality, let $1, \dots, K$ be the novel words of topic 1 to K . We first consider the solution of the constrained linear regression. To simplify the notation,

we denote $\mathbf{E}_i = [E_{i,1}, \dots, E_{i,K}]$ are the first K entries of a row vector without the super-scripts as in Algorithm 3.

Proposition 11. *Let $\bar{\mathbf{R}}$ be γ_a -affine-independent. The solution to the following optimization problem*

$$\hat{\mathbf{b}}^* = \arg \min_{b_j \geq 0, \sum b_j = 1} \|\hat{\mathbf{E}}_i - \sum_{j=1}^K b_j \hat{\mathbf{E}}_j\|$$

converges to the i -th row of $\bar{\boldsymbol{\beta}}$, $\bar{\boldsymbol{\beta}}_i$, as $M \rightarrow \infty$. Moreover,

$$\Pr(\|\hat{\mathbf{b}}^* - \bar{\boldsymbol{\beta}}_i\|_\infty \geq \epsilon) \leq 8W^2 \exp\left(-\frac{\epsilon^2 MN \gamma_a^2 \eta^4}{320K}\right)$$

where η is define the same as in Lemma 5.

Proof. We note that $\bar{\boldsymbol{\beta}}_i$ is the optimal solution to the following problem with ideal word co-occurrence statistics

$$\mathbf{b}^* = \arg \min_{b_j \geq 0, \sum b_j = 1} \|\mathbf{E}_i - \sum_{j=1}^K b_j \mathbf{E}_j\|$$

Define $f(\mathbf{E}, \mathbf{b}) = \|\mathbf{E}_i - \sum_{j=1}^K b_j \mathbf{E}_j\|$ and note the fact that $f(\mathbf{E}, \mathbf{b}^*) = 0$. Let $\mathbf{Y} = [\mathbf{E}_1^\top, \dots, \mathbf{E}_K^\top]^\top$. Then,

$$\begin{aligned} f(\mathbf{E}, \mathbf{b}) - f(\mathbf{E}, \mathbf{b}^*) &= \|\mathbf{E}_i - \sum_{j=1}^K b_j \mathbf{E}_j\| - 0 = \left\| \sum_{j=1}^K (b_j - b_j^*) \mathbf{E}_j \right\| \\ &= \sqrt{(\mathbf{b} - \mathbf{b}^*)^\top \mathbf{Y} \mathbf{Y}^\top (\mathbf{b} - \mathbf{b}^*)} \geq \|\mathbf{b} - \mathbf{b}^*\| \gamma_a \end{aligned}$$

The last equality is true by the definition of affine-independence. Next, note that,

$$\begin{aligned} |f(\mathbf{E}, \mathbf{b}) - f(\hat{\mathbf{E}}, \mathbf{b})| &\leq \|\mathbf{E}_i - \hat{\mathbf{E}}_i + \sum_{j=1}^K b_j (\hat{\mathbf{E}}_j - \mathbf{E}_j)\| \\ &\leq \|\mathbf{E}_i - \hat{\mathbf{E}}_i\| + \sum_{j=1}^K b_j \|\hat{\mathbf{E}}_j - \mathbf{E}_j\| \\ &\leq 2 \max_w \|\hat{\mathbf{E}}_w - \mathbf{E}_w\| \end{aligned}$$

Combining the above inequalities, we obtain,

$$\|\hat{\mathbf{b}}^* - \mathbf{b}^*\| \leq \frac{1}{\gamma_a} \{f(\mathbf{E}, \hat{\mathbf{b}}^*) - f(\mathbf{E}, \mathbf{b}^*)\}$$

$$\begin{aligned}
&= \frac{1}{\gamma_a} \{f(\mathbf{E}, \hat{\mathbf{b}}^*) - f(\hat{\mathbf{E}}, \hat{\mathbf{b}}^*) + f(\hat{\mathbf{E}}, \hat{\mathbf{b}}^*) - f(\hat{\mathbf{E}}, \mathbf{b}^*) + f(\hat{\mathbf{E}}, \mathbf{b}^*) - f(\mathbf{E}, \mathbf{b}^*)\} \\
&\leq \frac{1}{\gamma_a} \{f(\mathbf{E}, \hat{\mathbf{b}}^*) - f(\hat{\mathbf{E}}, \hat{\mathbf{b}}^*) + f(\hat{\mathbf{E}}, \mathbf{b}^*) - f(\mathbf{E}, \mathbf{b}^*)\} \\
&\leq \frac{4K^{0.5}}{\gamma_a} \|\hat{\mathbf{E}} - \mathbf{E}\|_\infty
\end{aligned}$$

where the last term converges to 0 almost surely. The convergence rate follows directly from Lemma 5. \square

We next consider the row renormalization. Let $\hat{\mathbf{b}}^*(i)$ be the optimal solution in Proposition 11 for the i -th word, and consider

$$\hat{\mathbf{B}}_i := \hat{\mathbf{b}}^*(i)^\top \left(\frac{1}{M} \mathbf{X} \mathbf{1}_{M \times 1} \right) \rightarrow \beta_i \text{diag}(\mathbf{a}) \quad (\text{A.4})$$

To show the convergence rate of the above step, it is straightforward to apply the result in Lemma 5

Proposition 12. *For the row-scaled estimation $\hat{\mathbf{B}}_i$ as in Eq. (A.4), we have,*

$$\Pr(|\hat{\mathbf{B}}_{i,k} - \beta_{i,k} a_k| \geq \epsilon) \leq 8W^2 \exp\left(-\frac{\epsilon^2 MN \gamma_a^2 \eta^4}{1280K}\right)$$

Proof. By Proposition 11, we have,

$$\Pr(|\hat{\mathbf{b}}^*(i)_k - \bar{\beta}_{i,k}| \geq \epsilon/2) \leq 8W^2 \exp\left(-\frac{\epsilon^2 MN \gamma_a^2 \eta^4}{1280K}\right)$$

Recall that in Lemma 5 by McDiarmid's inequality, we have

$$\Pr\left(\left|\frac{1}{M} \mathbf{X} \mathbf{1}_{M \times 1} - \mathbf{B}_i \mathbf{a}\right| \geq \epsilon/2\right) \leq \exp(-\epsilon^2 MN/2)$$

Therefore,

$$\Pr(|\hat{\mathbf{B}}_{i,k} - \beta_{i,k} a_k| \geq \epsilon) \leq 8W^2 \exp\left(-\frac{\epsilon^2 MN \gamma_a^2 \eta^4}{1280K}\right) + \exp(-\epsilon^2 MN/2)$$

where the second term is dominated by the first term. \square

Finally, we consider the column normalization step to remove the effect of $\text{diag}(\mathbf{a})$:

$$\widehat{\boldsymbol{\beta}}_{i,k} := \widehat{\mathbf{B}}_{i,k} / \sum_{w=1}^W \widehat{\mathbf{B}}_{w,k} \quad (\text{A.5})$$

And $\sum_{w=1}^W \widehat{\mathbf{B}}_{w,k} \rightarrow \mathbf{a}_k$ for $k = 1, \dots, K$. A worst case analysis on its convergence is,

$$\Pr\left(\left|\sum_{w=1}^W \widehat{\mathbf{B}}_{w,k} - \mathbf{a}_k\right| > \epsilon\right) \leq W \Pr\left(\left|\widehat{\mathbf{B}}_{i,k} - \boldsymbol{\beta}_{i,k} a_k\right| \geq \epsilon/W\right) \leq 8W^3 \exp\left(-\frac{\epsilon^2 MN \gamma_a^2 \eta^4}{1280W^2 K}\right)$$

Combining all the result above, we can show $\forall i = 1, \dots, W, \forall k = 1, \dots, K$,

$$\Pr\left(\left|\widehat{\boldsymbol{\beta}}_{i,k} - \boldsymbol{\beta}_{i,k}\right| > \epsilon\right) \leq 8W^4 K \exp\left(-\frac{\epsilon^2 MN \gamma_a^2 a_{\min}^2 \eta^4}{2560W^2 K}\right)$$

where $a_{\min} > 0$ is the minimum value of entries of \mathbf{a} . This concludes the result of Theorem 4.

A.13 Theorem 3 with Spherical Gaussian Directions

The proof in Section A.11 holds for any isotropic distribution on \mathbf{d} . If we assume \mathbf{d} to be the standard spherical Gaussian distribution, we can have better sample complexity bounds. First note that,

Proposition 13. *Let $\mathbf{X}^n, \mathbf{X} \in \mathbb{R}^m$ be two random vectors, $\mathbf{a}, \boldsymbol{\epsilon} \in \mathbb{R}^m$ be two vectors and $\boldsymbol{\epsilon} > \mathbf{0}$.*

$$\left|\Pr\{\mathbf{X}^n \leq \mathbf{a}\} - \Pr\{\mathbf{X} \leq \mathbf{a}\}\right| \leq \Pr(\exists i : |X_i^n - X_i| \geq \epsilon_i) + \Pr(\mathbf{a} - \boldsymbol{\epsilon} \leq \mathbf{X} \leq \mathbf{a} + \boldsymbol{\epsilon})$$

The inequality is element-wise.

Proof. Note that

$$\begin{aligned} \Pr\{\mathbf{X}^n \leq \mathbf{a}\} &\leq \Pr\{\mathbf{X}^n \leq \mathbf{a}, \forall i : |X_i^n - X_i| \leq \epsilon_i\} + \Pr\{\mathbf{X}^n \leq \mathbf{a}, \exists i : |X_i^n - X_i| \geq \epsilon_i\} \\ &\leq \Pr\{\mathbf{X} \leq \mathbf{a} + \boldsymbol{\epsilon}\} + \Pr\{\exists i : |X_i^n - X_i| \geq \epsilon_i\} \end{aligned}$$

Similarly, by swapping \mathbf{X}^n and \mathbf{X} , we have,

$$\Pr\{\mathbf{X} \leq \mathbf{a} - \boldsymbol{\epsilon}\} \leq \Pr\{\mathbf{X}^n \leq \mathbf{a}\} + \Pr\{\exists i : |X_i^n - X_i| \geq \epsilon_i\}$$

Combining them concludes the proof. \square

Proposition 14. *Let the random projection directions be $\mathbf{d} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_W)$ in Algorithm 2 of the main paper. Then, $\forall \epsilon > 0$, let $\rho = \min\{\frac{d}{8}, \frac{\sqrt{\pi}\epsilon d_2}{4K\sqrt{W \log(2W/\epsilon)}}\}$. If $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho$, then, $q_i - p_i(\widehat{\mathbf{E}}) \leq \epsilon$ for a novel pair i and $p_j(\widehat{\mathbf{E}}) - q_j \leq \epsilon$ for a non-novel pair j .*

Proof. Recall the definition of q_i and $p_i(\widehat{\mathbf{E}})$,

$$q_i = \Pr\{\forall j \in \mathcal{S}(i), \mathbf{E}_i \mathbf{d} \geq \mathbf{E}_j \mathbf{d}\}, \quad p_i(\widehat{\mathbf{E}}) = \Pr\{\forall j \in \mathcal{J}_i, \widehat{\mathbf{E}}_i \mathbf{d} \geq \widehat{\mathbf{E}}_j \mathbf{d}\}$$

When i is a novel word, $\mathcal{S}(i) = \mathcal{J}_i$ for $\|\widehat{\mathbf{E}} - \mathbf{E}\|_\infty \leq \rho \leq d/8$, therefore, by Proposition 13, we have,

$$|q_i - p_i(\widehat{\mathbf{E}})| \leq \Pr(\exists j \in \mathcal{J}_i : |\mathbf{e}_{i,j} \mathbf{d}| \geq \delta) + \Pr(\forall j \in \mathcal{J}_i : |\mathbf{z}_{i,j} \mathbf{d}| \leq \delta) \quad (\text{A.6})$$

where $\mathbf{e}_{i,j} = \mathbf{E}_i - \widehat{\mathbf{E}}_i + \widehat{\mathbf{E}}_j - \mathbf{E}_j$ and $\mathbf{z}_{i,j} = \mathbf{E}_i - \mathbf{E}_j$. To apply the union bound to the second term in Eq. (A.6), it suffice to consider only $j \in \bigcup_{k=1}^K \mathcal{C}_k$. Therefore, by union bounding both the first and second terms, we obtain,

$$|q_i - p_i(\widehat{\mathbf{E}})| \leq \sum_j \Pr(|\mathbf{e}_{i,j} \mathbf{d}| \geq \delta) + \sum_j \Pr(|\mathbf{z}_{i,j} \mathbf{d}| \leq \delta)$$

Note that $\mathbf{e}_{i,j} \mathbf{d} \sim \mathcal{N}(0, \|\mathbf{z}_{i,j}\|_2^2)$ and $\mathbf{z}_{i,j} \mathbf{d} \sim \mathcal{N}(0, \|\mathbf{a}_{i,j}\|_2^2)$ conditioned on $\widehat{\mathbf{E}}$. Using the properties of the Gaussian distribution we have,

$$\Pr(|\mathbf{z}_{i,j} \mathbf{d}| \leq \delta) = \int_{-\delta}^{\delta} \frac{1}{\sqrt{2\pi} \|\mathbf{z}_{i,j}\|} e^{-t^2/2\|\mathbf{z}_{i,j}\|^2} dt \leq \frac{\sqrt{2/\pi}}{\|\mathbf{z}_{i,j}\|} \delta$$

By Proposition 6, $\|\mathbf{z}_{i,j}\| \geq d_2$ for $j \in \mathcal{J}_i$, therefore, $\Pr(|\mathbf{z}_{i,j} \mathbf{d}| \leq \delta) \leq \frac{\sqrt{2/\pi}}{d_2} \delta$. Similarly, note that

$$\Pr(|\mathbf{e}_{i,j} \mathbf{d}| \geq \delta | \widehat{\mathbf{E}}) = 2Q(\delta / \|\mathbf{e}_{i,j}\|) \leq \exp(-\delta^2/2\|\mathbf{e}_{i,j}\|^2)$$

by the property of the Q -function. Note that

$$\|\mathbf{e}_{i,j}\| \leq \|\mathbf{E}_i - \widehat{\mathbf{E}}_i\| + \|\widehat{\mathbf{E}}_j - \mathbf{E}_j\| \leq 2W^{0.5}\|\mathbf{E} - \widehat{\mathbf{E}}\|_\infty$$

Then, by marginalizing over $\widehat{\mathbf{E}}$ we obtain, $\Pr(|\mathbf{e}_{i,j}\mathbf{d}| \geq \delta) \leq \exp(-\delta^2/8W\|\mathbf{E} - \widehat{\mathbf{E}}\|_\infty^2)$. Combining these results, we obtain,

$$|q_i - p_i(\widehat{\mathbf{E}})| \leq K \frac{\sqrt{2/\pi}}{d_2} \delta + W \exp(-\delta^2/8W\|\mathbf{E} - \widehat{\mathbf{E}}\|_\infty^2)$$

hold true for any $\delta > 0$. Therefore, if we set $\delta = \frac{\epsilon_0 \rho}{2K\sqrt{2/\pi}}$, and require

$$\|\mathbf{E} - \widehat{\mathbf{E}}\|_\infty \leq \frac{\sqrt{\pi\epsilon}d_2}{4K\sqrt{W\log(2W/\epsilon)}}$$

then $|q_i - p_i(\widehat{\mathbf{E}})| \leq \epsilon$. In summary, we require $\|\mathbf{E} - \widehat{\mathbf{E}}\|_\infty \leq \min\{\frac{\sqrt{\pi\epsilon}d_2}{4K\sqrt{W\log(2W/\epsilon)}}, d/8\}$. We note that the argument above holds true for a non-novel pair as well. \square

In Proposition 14, the bound on $\|\mathbf{E} - \widehat{\mathbf{E}}\|_\infty$ is, $\min\{\frac{d}{8}, \frac{\sqrt{\pi\epsilon}d_2}{4K\sqrt{W\log(2W/\epsilon)}}\}$ which is an improvement over the result in Proposition 10, $\min\{\frac{d}{8}, \frac{\pi d_2 \epsilon}{W^{1.5}}\}$ where we could reduce the dependence on W from $W\sqrt{W}$ to $K\sqrt{W}$. Since $K \ll W$, we obtain a gain over the general isotropic distribution. This leads to lightly improved results for the overall sample complexity bounds:

Theorem 3(with Spherical Gaussian Random Projections) Let β be separable and $\bar{\mathbf{R}}$ be γ_s simplicial. Then Algorithm 2 can consistently identify all the novel words of K distinct topics as $M, P \rightarrow \infty$. Furthermore, if the random directions $\mathbf{d}^1, \dots, \mathbf{d}^P$ are drawn iid from a spherical Gaussian distribution, then $\forall \delta > 0$, if

$$M \geq \max 20 \frac{\log(2W/\delta)}{N\rho^2\eta^4}, \text{ and } , P \geq 16 \frac{\log(3W/\delta)}{q_\wedge^2}$$

then Algorithm 2 fails with probability at most δ . The other model parameters are defined as $\eta = \min_{1 \leq w \leq W} [\beta \mathbf{a}]_w$, $\rho = \min\{\frac{d}{8}, \frac{\sqrt{\pi}d_2q_\wedge}{4K\sqrt{W\log(2W/q_\wedge)}}\}$, $d_2 \triangleq (1-b)\gamma_s$, $d = (1-b)^2\gamma_s^2/\lambda_{\max}$, $b = \max_{j \in \mathcal{C}_0, k} \bar{\beta}_{j,k}$ and λ_{\max} is the maximum eigenvalues of $\bar{\mathbf{R}}$.

q_\wedge is the minimum normalized solid angle of the extreme points of the convex hull of the rows of \mathbf{E} .

A.14 Proof of Theorem 5

Proof. The proof of (a) is the same as the proofs of Theorem 3 and 4.

We then summarize the communication costs for part (b). The P directions of size $W \times 1$ requires WP real-numbers. Noting that the size of \mathcal{J}_r^* for the maximizers of P projections are constant we need to calculate the related statistics $E_{i,i} - E_{i,j} - E_{i,j} + E_{j,j}$ for $\mathcal{O}(P)$ words and in sum it requires $\mathcal{O}(WP)$ real-numbers on partial estimations of $E_{i,j}$'s to be transmitted. The partial projection values requires again $\mathcal{O}(WP)$ real-numbers to be transmitted. In sum, the novel word detection step requires a $\mathcal{O}(WP)$ communication cost.

The estimation step requires no further cost if it is conducted on the fusion center. On the other hand, if one would distribute it to different servers, only WK/L real-numbers need to be transmitted to each of them. Row normalizing $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{X}}'$ requires $2W$ real-numbers. In sum, we obtain the communication cost in Theorem 5, (b).

The proof of (c) is the same as the proofs of Theorem 2. \square

A.15 Proof of Lemma 9

Proof. First we check that \mathbf{B} is column stochastic:

$$\sum_{(i,j)} B_{(i,j),k} = \sum_{(i,j) : i < j} (\beta_{(i,j),k} + \beta_{(j,i),k}) \mu_{i,j} = \sum_{(i,j) : i < j} \mu_{i,j} = 1 \quad (\text{A.7})$$

Hence \mathbf{B} is a valid topic matrix. We then need to show that the distribution on the comparisons $\mathbf{w} = \{w_{m,n}\}$ and on the words in topic model $\mathbf{w}^{\text{TM}} = \{w_{m,n}^{\text{TM}}\}$ are the same. From Eq. (3.1),

$$\begin{aligned} p(\mathbf{w}|\mathbf{B}) &= \prod_{m=1}^M \int p(w_{m,1}, \dots, w_{m,N} | \boldsymbol{\theta}_m, \mathbf{B}) \text{Pr}(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \\ &= \prod_{m=1}^M \int \left(\prod_{n=1}^N \sum_{k=1}^K B_{w_{m,n},k} \theta_{k,m} \right) \text{Pr}(\boldsymbol{\theta}_m) d\boldsymbol{\theta}_m \end{aligned}$$

□

A.16 Proof of Lemma 10 and Theorem 6

The proofs are exactly the same as those in Lemma 8 and Theorem 2.

A.17 Proof of Theorem 7

First, if $\bar{\mathbf{R}}$ is γ_a -affine independent, by Proposition 1 it is at least γ_a simplicial. The consistency as well as final sample complexity for novel word detection (Algorithm 6) and the linear regression estimator (Algorithm 7) can be proved in the same way as that of Theorem 3 and Theorem 4 (up to Proposition 12). We only need to show the consistency of the post-processing steps in Algorithm 8.

First, we note that by the definition of the ranking matrix,

$$\hat{\beta}_{(i,j),k} \leftarrow \frac{\hat{\mathbf{B}}_{(i,j),k}}{\hat{\mathbf{B}}_{(i,j),k} + \hat{\mathbf{B}}_{(j,i),k}} \doteq \frac{\beta_{(i,j),k} \mu_{i,j} a_k}{\beta_{(i,j),k} \mu_{i,j} a_k + \beta_{(j,i),k} \mu_{i,j} a_k}$$

Noting that in the last step of Algorithm 8 we compared the estimates to 1/2, the estimate of the binary matrix β would be consistent if $|\hat{\mathbf{B}}_{(i,j),k} - \mathbf{B}_{(i,j),k} a_k| \leq 0.5 \mu_{i,j} a_k$. In another word, we can only require a finite estimation error in β in order to guarantee the consistent estimation. In addition, we note that $\eta := \min_{w=1}^W (\mathbf{B}\mathbf{a})_w$ is a lower bound of $\mu_{i,j} a_k$. Putting the above results together, the error probability of Algorithm 5 can be upper bounded by

$$Pe \leq 2W^2 \exp(-Pq_{\wedge}^2/8) + 8W^2 \exp(-\rho^2 \eta^4 MN/20) + 8W^2 \exp\left(-\frac{MN \lambda_{\min} \eta^6}{160W}\right)$$

with the model parameters defined in the same way as in the Theorem. This leads to the sample complexity results in the theorem.

A.18 Proof of Proposition 4

Proof. First, by the definition that $\beta_{(i,j),k}$ is the probability that item i being preferred over j , we have $\beta_{(i,j),k}, \beta_{(i,j),k} + \beta_{(j,i),k} = 1$. For the rest, we represent the k -th reference ranking σ_k (where $\sigma^k(i) < \sigma^k(j)$) as

$$\{\mathcal{I}\}, i, \{\mathcal{II}\}, j, \{\mathcal{III}\}$$

All the calligraphic symbols represent a ordering of a subset of items. The leftmost is the 1st item. Let σ be a permutation $\sigma(i) < \sigma(j)$, there exist exactly one ‘‘complementary’’ ranking σ^c by swapping only the position of i and j :

$$\begin{aligned} \sigma &: \{\mathcal{A}\}, i, \{\mathcal{B}\}, j, \{\mathcal{C}\} \\ \sigma^c &: \{\mathcal{A}\}, j, \{\mathcal{B}\}, i, \{\mathcal{C}\} \end{aligned}$$

The set of σ with $\sigma(i) < \sigma(j)$ is then exactly half of all the permutations. By the form of σ and σ^c , $d(\sigma, \sigma^k)$ and $d(\sigma^c, \sigma^k)$ only differ by the pairwise relations between i, j , and items in \mathcal{B} . We further set $n_{\mathcal{I}} = |\mathcal{B} \cap \mathcal{I}|$, $n_{\mathcal{II}} = |\mathcal{B} \cap \mathcal{II}|$, $n_{\mathcal{III}} = |\mathcal{B} \cap \mathcal{III}|$. The number of disagreeing pairs (due to i, j , and \mathcal{B}) between σ and σ_k is $n_{\mathcal{I}} + n_{\mathcal{III}}$, the disagreeing pairs between σ^c and σ_k is $n_{\mathcal{I}} + n_{\mathcal{III}} + 2n_{\mathcal{II}} + 1$. The term 1 is induced by i, j . $n_{\mathcal{II}} \geq 0$. In sum, we have $p_M(\sigma^c | \sigma^k, \phi_k) = \phi_k^{1+2n_{\mathcal{II}}} p_M(\sigma | \sigma^k, \phi_k)$.

Therefore, by definition,

$$\begin{aligned} \beta_{(i,j),k} &= \sum_{\sigma : \sigma(i) < \sigma(j)} p_M(\sigma | \sigma^k, \phi_k) = \sum_{\sigma^c : \sigma^c(i) > \sigma^c(j)} \frac{1}{\phi_k^{1+2n_{\mathcal{II}}}} p_M(\sigma^c | \sigma^k, \phi_k) \\ &\geq \frac{1}{\phi_k} \sum_{\sigma^c : \sigma^c(i) > \sigma^c(j)} p_M(\sigma^c | \sigma^k, \phi_k) = \frac{1}{\phi_k} \beta_{(j,i),k} \end{aligned}$$

Noting that $\beta_{(i,j),k}, \beta_{(i,j),k} + \beta_{(j,i),k} = 1$, we have $\beta_{(i,j),k} \geq \frac{1}{1+\phi_k} > 0.5$. Specifically, let $\mathcal{II} = \emptyset$, i.e., $\sigma^k(j) = \sigma^k(i) + 1$, we have $n_{\mathcal{II}} = 0$ and the equality above holds. Therefore, we obtain the results in *b*) that $1/\beta_{(i,j),k} = 1 + \phi_k$.

We now prove part *c*). Our proof here is based on the so-called *Repeated Insertion Model* (RIM) [Doignon et al., 2004, Lu and Boutilier, 2014]. RIM is a generative procedure for sampling a ranking from a given reference ranking. Given a reference ranking σ_0 , the process of a realization of RIM is as follows: one sequentially place the i -th item in the reference permutation into the j_i -th position (of the current partial

sequence of length i), $1 \leq j_i \leq i$, in a probabilistic fashion:

$$p_i(j_i = l) = \frac{\phi^{i-l}}{1 + \phi^l + \dots + \phi^{i-1}} \quad (\text{A.8})$$

for all $1 \leq l \leq i$, $1 \leq i \leq Q$. [Doignon et al., 2004, Lu and Boutilier, 2014] showed that when the inserting probabilities are defined as above, RIM induces a pmf on all the permutations that is **identical** to that of a Mallows model with reference ranking σ_0 and a dispersion parameter ϕ . In this context, $\beta_{(i,j),k}$ is the probability that item j is inserted after i in the RIM procedure.

Now, consider $\sigma^k(i) < \sigma^k(j)$ so that from part *a*) we have $0 < \beta_{(j,i),k} < 0.5 < \beta_{(i,j),k} < 1$. Without loss of generality, we consider $\sigma^k(i) = i$ hence $\sigma^k : 1 \succ 2 \succ \dots \succ Q$ where \succ indicates “prefer over”. Noting that $\beta_{(i,j),k}$ is the probability that item j is inserted after item i in the sequential procedure of RIM, it does not depend on all the items ranked after item j . By symmetry if we reverse “prefer over” to “prefer below”, this probability also does not depend on the items before i . Therefore, without loss of generality, we set $i = 1$ and consider $j = 2, \dots, Q$. To simplify the notation, we drop the subscript and let $\phi_k = \phi$.

We decompose the calculation of $\beta_{(j,1),k}$ into the following problem of determining the probability that item 1 being on the r -th position in the current partial sequence after inserting the s -th item. Here $r = 1, \dots, s$ and $s = 1, \dots, j-1$. We want to show by induction that $q_{r,s} = \frac{\phi^{r-1}}{1 + \phi^1 + \dots + \phi^{s-1}}$, and then calculate $\beta_{(1,j),k}$. As a initial point, after inserting the second item $s = 2$, by the definition in Eq. A.8, we have $q_{1,2} = \frac{1}{1+\phi}$ and $q_{2,2} = \frac{\phi}{1+\phi}$. Now, assume for all $s = 1, \dots, s$, the assumption hold true, then after inserting item $s + 1$, for all $1 < r \leq s + 1$, (we consider $r = 1$ separately) by the definition of RIM,

$$q_{r,s+1} = q_{r,s} p_{s+1}(j_{s+1} > r) + q_{r-1,s} p_{s+1}(j_{s+1} < r)$$

here j_{s+1} is the position of item $s + 1$ after inserting it into the partial sequence as in the definition of RIM. By the induction assumption,

$$q_{r,s} = \frac{\phi^{r-1}}{1 + \phi^1 + \dots + \phi^{s-1}} \quad q_{r-1,s} = \frac{\phi^{r-2}}{1 + \phi^1 + \dots + \phi^{s-1}}$$

And the probability of inserting item $s + 1$ follows the rule defined in Eq. (A.8).

Therefore,

$$\begin{aligned}
q_{r,s+1} &= \frac{\phi^{r-1}}{1 + \phi^1 + \dots + \phi^{s-1}} \Pr(j_{s+1} > r) + \frac{\phi^{r-2}}{1 + \phi^1 + \dots + \phi^{s-1}} \Pr(j_{s+1} < r) \\
&= \frac{\phi^{r-1}}{1 + \phi^1 + \dots + \phi^{s-1}} \frac{1 + \phi + \dots + \phi^{s-r-1}}{1 + \dots + \phi^s} + \frac{\phi^{r-2}}{1 + \phi^1 + \dots + \phi^{s-1}} \frac{\phi^{s-r+1} + \dots + \phi^s}{1 + \dots + \phi^s} \\
&= \frac{\phi^{r-1}}{1 + \dots + \phi^{s+1-1}}
\end{aligned}$$

One can separately check the case when $r = 1$. In sum, we conclude our induction hypothesis that

$$q_{r,s} = \frac{\phi^{r-1}}{1 + \phi^1 + \dots + \phi^{s-1}}$$

Now we can calculate $\beta_{(1,j),k}$ by its definition,

$$\begin{aligned}
\beta_{(1,j),k} &= \sum_{r=1}^{j-1} q_{r,j-1} \Pr(j_j > r) = \sum_{r=1}^{j-1} \frac{\phi^{r-1}(1 + \dots + \phi^{j-r-1})}{(1 + \dots + \phi^{j-2})(1 + \dots + \phi^{j-1})} \\
&= \frac{\sum_{r=1}^{j-1} \sum_{l=r-1}^{n-2} \phi^l}{(1 + \dots + \phi^{j-2})(1 + \dots + \phi^{j-1})} \\
&= \frac{1 - j\phi^{j-1} + (j-1)\phi^j}{(1 - \phi)^2(1 + \dots + \phi^{j-2})(1 + \dots + \phi^{j-1})} \\
&= \frac{1 - j\phi^{j-1} + (j-1)\phi^j}{(1 - \phi^{j-1})(1 - \phi^j)}
\end{aligned}$$

By taking $\beta_{(j,1),k} = 1 - \beta_{(1,j),k}$, we have,

$$\beta_{(j,1),k} = \frac{\phi^{j-1}(j-1 - j\phi + \phi^j)}{(1 - \phi^{j-1})(1 - \phi^j)}$$

To simplify the above probabilities, we take their ratio and simply the expression as,

$$\frac{\beta_{(1,j),k}}{\beta_{(j,1),k}} = \frac{1 - j\phi^{j-1} + (j-1)\phi^j}{\phi^{j-1}(j-1 - j\phi + \phi^j)} \geq \frac{1}{(j-1)\phi^{j-1}}$$

The last inequality is obtained by:

$$\frac{1 - j\phi^{j-1} + (j-1)\phi^j}{(j-1 - j\phi + \phi^j)} \geq \frac{1}{c}$$

$$\begin{aligned}
&\Leftrightarrow c(1 - \phi^j) - cj\phi^{j-1}(1 - \phi) \geq j(1 - \phi) - (1 - \phi^j) \\
&\Leftrightarrow (c + 1)(1 - \phi^j) \geq j(1 - \phi)(1 + c\phi^{j-1}) \\
&\Leftrightarrow (c + 1)(1 + \dots + \phi^{j-1}) \geq j(1 + c\phi^{j-1})
\end{aligned}$$

setting $c = j - 1$ achieves the last inequality since $\phi^r \geq \phi^{j-1}$ for $r \leq j - 1$. Therefore, one conclude that $\beta_{(j,1),k} \leq (j - 1)\phi^{j-1}/1 + (j - 1)\phi^{j-1}$. For future reference, we denote by $L = \sigma_k(j) - \sigma_k(i)$ as the difference in distance between the two items in the reference ranking (when $\sigma_k(j) > \sigma_k(i)$). \square

A.19 Proof of Lemma 11

Note that $B_{(i,j),k} = \mu_{i,j} \sum_{\sigma: \sigma(i) < \sigma(j)} p_M(\sigma | \sigma^k, \phi_k)$, we can check that,

$$\sum_{(i,j)} B_{(i,j),k} = \sum_{(i,j)} \mu_{i,j} \sum_{\sigma: \sigma(i) < \sigma(j)} p_M(\sigma | \sigma^k, \phi_k) = \sum_{(i,j) : i < j} \mu_{i,j} \sum_{\sigma} p_M(\sigma | \sigma^k, \phi_k) = 1$$

The rest of the proof is the same as that of Lemma 9.

A.20 Proof of Theorem 8

The proof is the same as that of Theorem 2 and Theorem 6.

A.21 Proof of Theorem 9

Indexing convention: For convenience, for the rest of this section we will index the $W = Q(Q - 1)$ rows of \mathbf{B} and \mathbf{E} by just a single index i instead of an ordered pair (i, j) as in the main paper.

A.21.1 Consistency of Algorithm 6 in M4

Recall that $\mathbf{E} = \bar{\mathbf{B}}\mathbf{Y}$ where $\mathbf{Y} = \bar{\mathbf{R}}\bar{\mathbf{B}}$. We decouple the effect of λ -separability from the error in estimating \mathbf{E} . Note that the estimation error converges to 0 as $M, N \rightarrow \infty$ as shown in Lemma 5, we shall only focus on the perturbation on solid angle as a result of the λ -approximate separability in this proof.

For i being a λ -approximate novel row, let $\mathbf{E}_i^0 = \mathbf{Y}_k$ as the corresponding row of \mathbf{Y} . Otherwise, let $\mathbf{E}_i^0 = \mathbf{E}_i$ be the rows of \mathbf{E} . For each approximate novel row i , define the original solid angle as,

$$q_i^0 = \Pr(\forall j : \|\mathbf{E}_j^0 - \mathbf{E}_i^0\| \geq d : \mathbf{E}_i^0 \mathbf{u} - \mathbf{E}_j^0 \mathbf{u} > 0) \quad (\text{A.9})$$

and define the λ -approximate solid angle as,

$$q_i = \Pr(\forall j : \|\mathbf{E}_j - \mathbf{E}_i\| \geq d : \mathbf{E}_i \mathbf{u} - \mathbf{E}_j \mathbf{u} > 0) \quad (\text{A.10})$$

for i being a λ approximately novel row. Therefore, for any constant $c > 0$,

$$|q_i^0 - q_i| \leq \Pr(\exists j, *, |\mathbf{E}_i^0 \mathbf{u} - \mathbf{E}_j^0 \mathbf{u} - \mathbf{E}_i \mathbf{u} + \mathbf{E}_j \mathbf{u}| \geq c) + \Pr(\forall j, *, |\mathbf{E}_i^0 \mathbf{u} - \mathbf{E}_j^0 \mathbf{u}| \leq c) \quad (\text{A.11})$$

where we have replace the distance constraints with $*$ for convenience. We note that $\mathbf{E}_j^0 = \sum_{k=1}^K \bar{B}_{jk} \mathbf{Y}_k$. Without loss of generality, assume that i is a λ -approximate novel row for \mathbf{Y}_1 , then, $\mathbf{E}_i^0 = \mathbf{Y}_1$. Taking a closer look at the second term in the above equation, we have,

$$|\mathbf{E}_i^0 \mathbf{u} - \mathbf{E}_j^0 \mathbf{u}| = \left| \sum_{k=2}^K \bar{B}_{jk} (\mathbf{Y}_k - \mathbf{Y}_1) \mathbf{u} \right| \leq \sum_{k=2}^K \bar{B}_{jk} |(\mathbf{Y}_k - \mathbf{Y}_1) \mathbf{u}|$$

And note that $\mathbf{Y}_k, k = 2, \dots, K$ are among the \mathbf{E}_j^0 's, therefore, the second term in (A.11) is equivalent to $\Pr(j = k, \dots, K, |(\mathbf{Y}_k - \mathbf{Y}_1) \mathbf{u}| \leq c)$ hence by union bounding, we have,

$$\Pr(\forall j, *, |\mathbf{E}_i^0 \mathbf{u} - \mathbf{E}_j^0 \mathbf{u}| \leq c) \leq \sum_{k=2}^K \Pr(|(\mathbf{Y}_k - \mathbf{Y}_1) \mathbf{u}| \leq c)$$

Note that $(\mathbf{Y}_k - \mathbf{Y}_1) \mathbf{u} \sim \mathcal{N}(0, \|\mathbf{Y}_k - \mathbf{Y}_1\|_2^2)$, by the property of Gaussian distri-

bution, we have,

$$\Pr(|(\mathbf{Y}_k - \mathbf{Y}_1)\mathbf{u}| \leq c) = \int_{-c}^c \frac{1}{\sqrt{2\sigma}\|\mathbf{Y}_k - \mathbf{Y}_1\|} e^{-t^2/2\|\mathbf{Y}_k - \mathbf{Y}_1\|^2} dt \leq \frac{c}{\|\mathbf{Y}_k - \mathbf{Y}_1\|}$$

For now we denote by ρ_{\min} the minimum of $\|\mathbf{Y}_k - \mathbf{Y}_l\|$, therefore, the second term in (A.11) can be upper-bound by $\frac{c(K-1)}{\rho_{\min}}$. For the first term in (A.11), let $\mathbf{e}_{i,j} = \mathbf{E}_i^0 - \mathbf{E}_j^0 - \mathbf{E}_i + \mathbf{E}_j$ and note that $\mathbf{e}_{i,j}\mathbf{u} \sim \mathcal{N}(0, \|\mathbf{e}_{i,j}\|_2^2)$, then,

$$\Pr(|\mathbf{e}_{i,j}\mathbf{u}| \geq c) = 2Q(c/\|\mathbf{e}_{i,j}\|) \leq \exp(-c^2/2\|\mathbf{e}_{i,j}\|_2^2)$$

Further, $\|\mathbf{e}_{i,j}\| \leq \|\mathbf{E}_i^0 - \mathbf{E}_i\| + \|\mathbf{E}_j^0 - \mathbf{E}_j\|$. For j which is not a λ -approximate novel row and is one of the j 's in (A.10), $\|\mathbf{E}_j^0 - \mathbf{E}_j\| = 0$. For j being a λ -approximate novel row and is one of the j 's in (A.10), hence j correspond to another topic. Therefore, by the same argument,

$$\|\mathbf{E}_i^0 - \mathbf{E}_i\| = \|\mathbf{Y}_1 - \sum_{k=1}^K \bar{B}_{ik}\mathbf{Y}_k\| \leq \sum_{k=2}^M \bar{B}_{ik}\|\mathbf{Y}_1 - \mathbf{Y}_k\| \leq \lambda \sum_{k=2}^M \|\mathbf{Y}_1 - \mathbf{Y}_k\|$$

Combining the steps together, for Eq. (A.11), we require,

$$|q_i^0 - q_i| \leq \frac{c(K-1)}{\rho_{\min}} + W \exp(-[\frac{c}{\lambda K \rho_{\max}}]^2) \leq q_{\wedge}/3$$

where q_{\wedge} is the minimum solid angle of \mathbf{Y} . This is require so that the estimated solid angle for the λ -approximate novel rows is well-separated from the solid angle of the remaining non-novel rows. Recall that ρ_{\min} and ρ_{\max} is defined as the minimum and maximum values of $\|\mathbf{Y}_i - \mathbf{Y}_j\|, 1 \leq i \neq j \leq K$. To parse the above equation, we set $c = \frac{q_{\wedge}\rho_{\min}}{3K}$ and therefore, we require

$$\lambda \leq \frac{q_{\wedge}\rho_{\min}}{3K^2\rho_{\max}\sqrt{\log(W/q_{\wedge})}} \leq \frac{q_{\wedge}\kappa}{3K^2\sqrt{\log(W/q_{\wedge})}}$$

We can now apply the same argument to the other rows i whose d -neighbor does

not enclose a novel word. We thus require $d \geq 12\lambda K \sqrt{\log(W/q_\wedge)}/q_\wedge$. To combine the two results, we can set

$$d = \mathcal{O}(\kappa/K) \tag{A.12}$$

To summarize the discussion, we have,

Proposition 15. *If λ is small enough such that,*

$$\lambda \leq \frac{q_\wedge^\kappa}{3K^2 \sqrt{\log(W/q_\wedge)}} \tag{A.13}$$

with d set as in (A.12). Then, for $M, N \rightarrow \infty$ and the number of projections $P \rightarrow \infty$, the proposed algorithm can find $\mathcal{O}\left(2K \sqrt{\log(W/q_\wedge)}/q_\wedge\right)$ λ -approximately novel rows for K distinct topics.

A.21.2 Consistency of Algorithm 7 in M4

We now consider the error accumulated in steps in Algorithm 1 in the appendix. Assume the Algorithm 2 (in the main paper) is correct, we obtain K row vectors, $\mathbf{E}_j, j = 1, \dots, K$, as λ -approximate novel pairs for the K distinct Mallows components. Without loss of generality, \mathbf{E}_j approximately novel to the j -th Mallows component (j -th column). We further denote by \mathbf{E}_j^0 the ideal extreme points, i.e., $\mathbf{E}_j^0 = \mathbf{Y}_j$ for $j = 1, \dots, K$. Note that by definition, $\mathbf{E}_i = \sum_{k=1}^K \bar{B}_{ik} \mathbf{E}_k^0$ for $i = 1, \dots, K, k \neq i$, we have $\bar{B}_{ik} \leq \lambda \bar{B}_{ii}$. $\bar{\mathbf{B}}$ is a row-stochastic matrix. For $i = 1, \dots, W$, the corresponding row vector $\bar{\mathbf{B}}_i$ is the optimal solution of the following constrained linear regression,

$$\mathbf{b}^* = \arg \min_{b_j \geq 0, \sum b_j = 1} \left\| \mathbf{E}_i - \sum_{j=1}^K b_j \mathbf{E}_j^0 \right\|$$

Now consider the empirical version we have access to which is,

$$\hat{\mathbf{b}}^* = \arg \min_{b_j \geq 0, \sum b_j = 1} \left\| \hat{\mathbf{E}}_i - \sum_{j=1}^K b_j \hat{\mathbf{E}}_j \right\|$$

To bound the error between $\widehat{\mathbf{b}}^*$ and \mathbf{b}^* due to approximate separability, we can establish the following property:

Proposition 16. *Suppose that for $j = 1, \dots, K$, $\|\widehat{\mathbf{E}}_j - \mathbf{E}_j^0\|_2 \leq \delta_1$ and $\|\widehat{\mathbf{E}}_i - \mathbf{E}_i\|_2 \leq \delta_2$ a fixed i . Assume also that $\widehat{\mathbf{E}}_j, j = 1, \dots, K$ are at most λ -approximately separable and $(K-1)\lambda \leq 1$, then,*

$$\|\widehat{\mathbf{b}}^* - \mathbf{b}^*\|_2 \leq 4 \frac{\delta_1 + \delta_2}{(1 - (K-1)\lambda)\lambda_{\min}}$$

where λ_{\min} denotes the minimum eigenvalue of $\bar{\mathbf{R}}$.

Proof. Let $f(\mathbf{E}^0, \mathbf{b}) = \|\mathbf{E}_i - \sum_{j=1}^K b_j \mathbf{E}_j^0\|$ for any \mathbf{b} and note that for the optimal solution \mathbf{b}^* , $f(\mathbf{E}^0, \mathbf{b}^*) = 0$. Let $\mathbf{Y} = [\mathbf{E}_1^{0\top}, \dots, \mathbf{E}_K^{0\top}]^\top$, we have,

$$\begin{aligned} f(\mathbf{E}, \mathbf{b}) - f(\mathbf{E}, \mathbf{b}^*) &= \|\mathbf{E}_i - \sum_{j=1}^K b_j \mathbf{E}_j^0\| - 0 = \left\| \sum_{j=1}^K (b_j - b_j^*) \mathbf{E}_j^0 \right\| \\ &= \sqrt{(\mathbf{b} - \mathbf{b}^*) \mathbf{Y} \mathbf{Y}^\top (\mathbf{b} - \mathbf{b}^*)^\top} \geq \|\mathbf{b} - \mathbf{b}^*\| \lambda_{\min, Y} \end{aligned}$$

Recall that $\mathbf{Y} = \bar{\mathbf{R}} \bar{\mathbf{B}}^\top$ and let $\bar{\mathbf{B}}^\top = [B_K, B_r]^\top$ where the $K \times K$ B_K are approximately separable. Note that $B_{K,(i,j)}/B_{K,(i,i)} \leq \lambda$ and $\lambda(K-1) \leq 1$, then, by the Gershgorin circle theorem, the minimum eigenvalue of B_K is lower-bounded by $\frac{1-(K-1)\lambda}{1+(K-1)\lambda} > \frac{1-(K-1)\lambda}{2}$. Therefore, $\lambda_{\min, Y} \geq \lambda_{\min} \frac{1-(K-1)\lambda}{2}$ where λ_{\min} is the minimum eigenvalue of $\bar{\mathbf{R}}$. Next, note that for any probability vector \mathbf{b} ,

$$\begin{aligned} |f(\mathbf{E}, \mathbf{b}) - f(\widehat{\mathbf{E}}, \mathbf{b})| &\leq \|\mathbf{E}_i - \widehat{\mathbf{E}}_i + \sum_{j=1}^K b_j (\widehat{\mathbf{E}}_j - \mathbf{E}_j^0)\| \\ &\leq \|\mathbf{E}_i - \widehat{\mathbf{E}}_i\| + \sum_{j=1}^K b_j \|\widehat{\mathbf{E}}_j - \mathbf{E}_j^0\| \leq \delta_2 + \delta_1 \end{aligned}$$

Combining the above inequalities, we obtain,

$$\begin{aligned} \|\widehat{\mathbf{b}}^* - \mathbf{b}^*\| &\leq \frac{1}{\lambda_{\min, Y}} \{f(\mathbf{E}, \widehat{\mathbf{b}}^*) - f(\mathbf{E}, \mathbf{b}^*)\} \\ &= \frac{1}{\lambda_{\min, Y}} \{f(\mathbf{E}, \widehat{\mathbf{b}}^*) - f(\widehat{\mathbf{E}}, \widehat{\mathbf{b}}^*) + f(\widehat{\mathbf{E}}, \widehat{\mathbf{b}}^*) - f(\widehat{\mathbf{E}}, \mathbf{b}^*) + f(\widehat{\mathbf{E}}, \mathbf{b}^*) - f(\mathbf{E}, \mathbf{b}^*)\} \\ &\leq \frac{1}{\lambda_{\min, Y}} \{f(\mathbf{E}, \widehat{\mathbf{b}}^*) - f(\widehat{\mathbf{E}}, \widehat{\mathbf{b}}^*) + f(\widehat{\mathbf{E}}, \mathbf{b}^*) - f(\mathbf{E}, \mathbf{b}^*)\} \\ &\leq \frac{4}{\lambda_{\min}(1 - \lambda(K-1))} (\delta_1 + \delta_2) \end{aligned}$$

□

A.21.3 Consistency of the post-processing Algorithm 10 in M4

We first consider the row normalization step in Algorithm 10. Note that, $b^*(i, j)_k = \bar{B}_{(i,j),k} = \frac{\mu_{i,j}\beta_{(i,j),k}a_k}{\sum \mu_{i,j}\beta_{(i,j),l}a_l}$. We define the row-scaling factor,

$$p_{i,j} = \sum_m X_{(i,j),m} / (\sum_m X_{(i,j),m} + \sum_m X_{(j,i),m})$$

and by definition $p_{i,j} \rightarrow \sum \beta_{(i,j),l}a_l \leq 1$ as $M \rightarrow \infty$. If we define $c_{(i,j),k} \leftarrow p_{i,j}b^*(i, j)_k$ as intermediate step, and then compute $c_{(i,j),k} / (c_{(i,j),k} + c_{(j,i),k})$. Note that $c_{(i,j),k} = \beta_{(i,j),k}a_k$ in the ideal case, in order to learn the hidden ranking correctly, we only need $c_{(i,j),k} / (c_{(i,j),k} + c_{(j,i),k}) = \beta_{(i,j),k}$ to remain in the correct interval of either $[0, 0.5]$ or $[0.5, 1]$. Therefore, the error in estimating $c_{(i,j),k}$ should satisfy,

$$|c_{(i,j),k} - \hat{c}_{(i,j),k}| \leq a_k |0.5 - \beta_{(i,j),k}|$$

Recall that $p_{i,j}$ can be estimated much accurate than b^* , Therefore, we can consider the error in c as the result of error in \hat{b}^* . Note that the minimum of the $|0.5 - \beta_{(i,j),k}|$ is achieved if the position of item i, j in the reference ranking are next to each other and $|0.5 - \sigma_{(i,j),k}| \geq \frac{1-\phi}{2(1+\phi)} \geq (1-\phi)/4$. Therefore, we require,

$$|\hat{b}^*(i, j)_k - b^*(i, j)_k| p_{i,j} \leq a_k (1-\phi)/4$$

Let $a_{\min} = \min a_k$ and note that $p_{i,j} < 1$, using result in Prop. 16, we require,

$$\delta_1 + \delta_2 \leq a_{\min} \lambda_{\min} (1 - (K-1)\lambda)(1-\phi)/8 \quad (\text{A.14})$$

Now, we express δ_1 and δ_2 in terms of λ . Note that $\delta_2 = \|\hat{\mathbf{E}}_i - \mathbf{E}_i\|$ and $\delta_1 = \|\hat{\mathbf{E}}_j - \mathbf{E}_j^0\| \leq \|\hat{\mathbf{E}}_j - \mathbf{E}_j\| + \|\mathbf{E}_j^0 - \mathbf{E}_j\|$. δ_2 and the first term in δ_1 converges to 0 exponentially in M, N and does not depend on λ . Hence we focus on the term $\|\mathbf{E}_j^0 - \mathbf{E}_j\|$. Note

that $\|\mathbf{E}_j^0 - \mathbf{E}_j\| = \|\sum_{k \neq j} \bar{B}_{jk}(\mathbf{E}_k^0) - (1 - \bar{B}_{jj})\mathbf{E}_j^0\|$. Let $v = [-(1 - \bar{B}_{11}), \bar{B}_{12}, \dots, \bar{B}_{1K}]$ (wlog, consider $j = 1$), then, $\|\mathbf{E}_j^0 - \mathbf{E}_j\| \leq \|v\|\lambda_{\max, Y}$. Following the same steps in Prop. 16 and denoting λ_{\max} to be the maximum eigenvalue of $\bar{\mathbf{R}}$, we have, $\lambda_{\max, Y} \leq (1 + (K - 1)\lambda)\lambda_{\max}$, and $\|v\| \leq \lambda(K - 1)/(1 + (K - 1)\lambda)$. Combining the results, we have,

$$\|\mathbf{E}_j^0 - \mathbf{E}_j\| \leq \lambda(K - 1)\lambda_{\max}$$

Let's consider $K\lambda \ll 1$ and using all the results above, we need,

$$\lambda \leq \frac{a_{\min}\lambda_{\min}(1 - \phi)}{8K\lambda_{\max}}$$

Formally, to combine the above two sections, we have,

Proposition 17. *Assume K rows that λ -approximately novel pairs for K distinct Mallows components are selected. The remaining steps, i.e., constrained linear regression, row-scaling, and post-processing can recover the true reference rankings of all Mallows component when $M \rightarrow \infty$ and $\lambda \leq \frac{a_{\min}\kappa(1 - \phi)}{8(K - 1)}$ where $a_{\min} = \min_k a_k$, $\kappa = \lambda_{\min}/\lambda_{\max} > 0$ is the condition number of $\bar{\mathbf{R}}$, and $\phi = \max_k \phi_k < 1$.*

A.21.4 Overall sample complexity of the Algorithm 9

We can directly combine the results from Prop. 15, 16 and 17 to obtain the consistency results for the overall algorithm.

Proof. First note that $\bar{B}_{i,k} = \mu_i\beta_{i,k}a_k$. Therefore, if β is λ -approximately separable, then, $\bar{\mathbf{B}}$ is at most $a_0\lambda$ -approximately separable. Now, assuming that $\lambda a_0 \leq \frac{q_\wedge^\kappa}{3K^2\sqrt{\log(W/q_\wedge)}}$, by proposition 15, the novel word step via random projection can select roughly $c_1K\lambda a_0/q_\wedge$ -approximately separable novel words if $M, N \rightarrow \infty$ and $P \rightarrow \infty$.

Now apply proposition 17, we require $c_1K\lambda a_0/q_\wedge \leq \frac{a_{\min}\kappa(1 - \phi)}{8K}$, therefore,

$$\lambda \leq \frac{a_{\min}\kappa(1 - \phi)q_\wedge}{8c_1K^2a_0} = \frac{a_{\min}\kappa(1 - \phi)q_\wedge}{8K^2a_0\sqrt{\log(W/q_\wedge)}}$$

Note that this is stronger than previous constraints. In sum, given these constraints, and let $M, P \rightarrow \infty$, the estimation on the center rankings are consistent. The sample complexity follows directly from results in Lemma 5 and Theorem 3. \square

A.22 Proof of Lemma 12

Proof. Each $\text{Dir}(\beta_0)$ -distributed column β_k can be generated by first sampling each of its W entries *independently* from a gamma distribution with parameter β_0 , and then dividing all the column entries by their sum in order to make the column-sum equal to one (column-normalization). We will refer to the un-normalized $W \times K$ random matrix with independent $\text{gamma}(\beta_0, 1)$ -distributed entries as the “gamma random matrix”. Our overall analysis approach is to (a) first calculate the probability that a row of the gamma random matrix is $\lambda/4$ -approximately novel for a topic, i.e., $p_1(\beta_0, \lambda/4, K)$ as defined in Lemma 12, and (b) then show that all the column-normalization factors will concentrate around their means when W is large and will therefore not impact the approximate-separability property of the gamma random matrix.

To formalize the above ideas, let $\mu_{w,k}, w = 1, \dots, W, k = 1, \dots, K$ be i.i.d samples from the $\text{gamma}(\beta_0, 1)$ distribution. We denote by $b_k = \sum_{w=1}^W \mu_{w,k}$ the column-normalization factor for the k -th column. Let \mathcal{A} denote the event that all the normalization factors $b_k, k = 1, \dots, K$, are within a $W\beta_0/2$ radius of their means $W\beta_0$. Let \mathcal{B} denote the event that the gamma random matrix has at least one $\lambda/4$ -approximately novel word for each topic. When event \mathcal{A} occurs, then $\forall i, j, i \neq j, b_i/b_j \in (1/4, 4)$. Then the $\lambda/4$ -approximate novel words of the gamma random matrix will become at most λ -approximate novel words after column-normalization. Thus, for the event that β is λ -approximately separable to occur it is sufficient that the intersection of events $\mathcal{A} \cap \mathcal{B}$ occurs.

For event \mathcal{B} , we define $p_1(\beta_0, \lambda/4, K)$ to be the probability that the first row of the gamma random matrix is $\lambda/4$ approximately novel for the first column (topic 1). Since all entries in the gamma random matrix are i.i.d., the probability that any row of the gamma random matrix is approximately novel for any column would be exactly the same for all rows and columns (by symmetry). Next, note that $\mathcal{B}^c = \bigcup_{k=1}^K \mathcal{B}_k$ where \mathcal{B}_k is the event that *none* of the W rows in the gamma random matrix is $\lambda/4$ approximately novel for the k -th topic. Since the rows of the gamma random matrix are independent, we have

$$\Pr(\mathcal{B}_k) = (1 - p_1(\beta_0, \lambda/4, K))^W \leq \exp(-Wp_1)$$

Therefore, using the union bound, we get $\Pr(\mathcal{B}^c) \leq K \exp(-Wp_1)$.

We then consider $\mathcal{A} = \{\forall k, |b_k - W\beta_0| \leq W\beta_0/2\}$. Note that by law of large

numbers for sub-Gaussian random variables, we have $\Pr(|b_k - W\beta_0| > \frac{1}{2}W\beta_0) \leq c_1 \exp(-c_2W\beta_0)$ for some absolute constants c_1 and c_2 . Therefore,

$$\Pr(\mathcal{A}^c) \leq Kc_1 \exp(-c_2W\beta_0)$$

. Putting it all together, the probability that β is λ -approximately separable is lower bounded by the probability of the intersection of \mathcal{A} and \mathcal{B} , which is lower bounded by $c_1K \exp(-c_2W\beta_0) + K \exp(-Wp_1)$. It remains to derive an explicit formula or bound for p_1 . This is summarized in Lemma 15. \square

Lemma 15. *Let $\mu = [\mu_1, \dots, \mu_K]$ be a $1 \times K$ row vector where the μ_k 's are i.i.d samples from the $\text{gamma}(\beta_0, 1)$ distribution. Then, the probability that μ is a c -approximately novel row for topic 1, $p_1(\beta_0, c, K)$, can be lower bounded as follows:*

$$p_1(\beta_0, c, K) \geq \frac{c_3}{K} \left(\frac{c}{cK + 1 - c} \right)^{\beta_0 K} \quad (\text{A.15})$$

Proof. By the definition of separability,

$$\begin{aligned} p_1(\beta_0, c, K) &= \Pr(\mu_2 \leq c\mu_1, \dots, \mu_K \leq c\mu_1) \\ &= \int_0^\infty \Pr(\mu_2 \leq c\mu_1, \dots, \mu_K \leq c\mu_1 | \mu_1) p(\mu_1) d\mu_1 = \int_0^\infty \gamma(\beta_0, c\mu_1)^{K-1} p(\mu_1) d\mu_1 \end{aligned}$$

where $\gamma(\beta_0, c\mu_1) = \int_0^{c\mu_1} \frac{x^{\beta_0-1} \exp(-x)}{\Gamma(\beta_0)} dx$ is the incomplete gamma function (i.e., the CDF of the gamma distribution). We first consider a lower bound for the incomplete gamma function in closed-form,

$$\gamma(\beta_0, c\mu_1) = \int_0^{c\mu_1} \frac{x^{\beta_0-1} \exp(-x)}{\Gamma(\beta_0)} dx \geq \frac{\exp(-c\mu_1)}{\Gamma(\beta_0)} \int_0^{c\mu_1} x^{\beta_0-1} dx = \frac{\exp(-c\mu_1)}{\Gamma(\beta_0)} \frac{(c\mu_1)^{\beta_0}}{\beta_0}.$$

Putting it all together we have

$$\begin{aligned} p_1(\beta_0, c, K) &= \int_0^\infty \gamma(\beta_0, c\mu_1)^{K-1} p(\mu_1) d\mu_1 \geq \int_0^\infty \left(\frac{\exp(-c\mu_1)}{\Gamma(\beta_0)} \frac{(c\mu_1)^{\beta_0}}{\beta_0} \right)^{K-1} \frac{\mu_1^{\beta_0-1} \exp(-\mu_1)}{\Gamma(\beta_0)} d\mu_1 \end{aligned}$$

$$\begin{aligned}
&= \frac{c^{\beta_0(K-1)}}{\Gamma(\beta_0)^K \beta_0^{K-1}} \int_0^\infty \mu_1^{\beta_0 K - 1} \exp(-\mu_1(cK - c + 1)) d\mu_1 = \frac{c^{\beta_0(K-1)}}{\Gamma(\beta_0)^K \beta_0^{K-1}} \frac{\Gamma(K\beta_0)}{(cK + 1 - c)^{\beta_0 K}} \\
&= \frac{\Gamma(K\beta_0)}{\Gamma(\beta_0)} \frac{1}{(\Gamma(\beta_0)\beta_0)^{K-1}} \frac{1}{c^{\beta_0}} \frac{c^{\beta_0 K}}{(cK + 1 - c)^{\beta_0 K}}
\end{aligned}$$

To proceed further, first note that $\beta_0\Gamma(\beta_0) = \Gamma(\beta_0 + 1)$ and we consider $\beta_0 \in (0, 1)$. Using the fact that $\Gamma(1) = \Gamma(2) = 1$ and $\Gamma(x) < \Gamma(1) = \Gamma(2)$ for all $x \in (1, 2)$, we get $\beta_0\Gamma(\beta_0) = \Gamma(\beta_0 + 1) < 1$. Hence the term $\frac{1}{(\Gamma(\beta_0)\beta_0)^{K-1}} > 1$. Next note that for $\beta_0 K > 2$, the gamma function is increasing. Therefore, for large K , $\Gamma(\beta_0 K) > \Gamma(\beta_0)$. In the region where $\beta_0 K < 1$, one can show that $\Gamma(K\beta_0)/\Gamma(\beta_0) = O(1/K)$. We also note that $c < 1$ and $\beta_0 < 1$ so that $c^{\beta_0} < 1$. Hence for $p_1(\beta_0, c, K)$, we have,

$$p_1(\beta_0, c, K) \geq \frac{c_3}{K} \frac{c^{\beta_0 K}}{(cK + 1 - c)^{\beta_0 K}}.$$

□

A.23 Proof of Lemma 13

Proof. If a ranking σ^k is sampled uniformly from the set of all permutations, then,

- (a) The preferences of $\{i, j\}$ and $\{s, t\}$ is independent if i, j, s, t are distinct items.
- (b) $\beta_{(i,j),k} = \mathbb{I}(\sigma_k(i) < \sigma_k(j))$ is Bernoulli random variable with $p = 0.5$ for any pair of items i, j

Therefore, it suffice to consider the separability of a subset of the $W = Q(Q - 1)$ rows that corresponding to disjoint pairs of items. Given Q items, we can construct $Q/2$ distinct pairs and the we consider the separability condition of the corresponding $Q \times K$ sub-matrix (for each pair i, j , both row (i, j) and (j, i) are considered). For each corresponding row vector in β , i.e., $[\beta_{(i,j),1}, \dots, \beta_{(i,j),K}]$, the probability it being a novel row for topic 1, denoted by $p_1(K)$, can be straightforwardly calculated as,

$$\begin{aligned}
p_1(K) &= \Pr\{\beta_{(i,j),1} = 1, \beta_{(i,j),2} = \dots = \beta_{(i,j),K} = 0 \\
&\quad \text{or } \beta_{(i,j),1} = 0, \beta_{(i,j),2} = \dots = \beta_{(i,j),K} = 1\} = 2^{-(K-1)}
\end{aligned}$$

Now we can follow the same approach as in Lemma 12. Let $\mathcal{B} = \bigcup_{k=1}^K \mathcal{B}_k$ where \mathcal{B}_k indicates the event that none of the $Q/2$ disjoint pairs are novel for topic k . By

definition, $\Pr(\mathcal{B}_k) = (1 - p_k(K))^{Q/2} \leq \exp(-Qp_k/2)$. Then, by union bound, we have,

$$\Pr(\mathcal{B}) \leq K \exp(-Qp_k/2) \leq K \exp(-\frac{Q}{2K})$$

□

A.24 Proof of Lemma 14

Proof. By Lemma 4 c), if i is preferred over j in reference ranking σ_1 and under j in other reference rankings and the distance of their positions are at least L (in the rankings $2, \dots, K$), then, the corresponding row is at most $2L\phi^{L-1}$ approximately novel row for the first Mallows component since $\beta_{(i,j),1} > 1/2$. This holds true for novel rows for any other Mallows components. Here, we set $\phi = \max_k \phi_k < 1$.

Let σ^k be sampled uniformly at random from the set of all the permutations. Then, for two groups of disjoint items, the relative rankings within one group is independent to the other group. Therefore, we divide the Q items into Q/L groups of disjoint items, each containing L items, denoted by $\{i_{t,1}, \dots, i_{t,L}\}$, for $t = 1, \dots, Q/L$. For simplicity we assume Q is a multiple of L but we can always take $\text{ceil}(Q/L)$. Then, all the partial rankings within each group t are independent to that of another group s .

We now consider for each of these L -tuples, the probability that there exist two items i, j such that i is above j in the group for first reference permutation σ_1 , and that i ranked last and j ranked first in all the other permutations. We denote this probability by $p_1(\phi; \lambda, k)$. By definition, we have,

$$\begin{aligned} p_1(\phi; \lambda, k) &\geq \Pr\{\exists i, j \in \{i_{t,1}, \dots, i_{t,L}\}, s.t., \sigma^1(i) < \sigma^1(j), \\ &\quad \sigma^2(i) > \dots > \sigma^2(j), \dots, \sigma^K(i) > \dots > \sigma^K(j)\} \\ &= L(L-1) \left(\frac{1}{L(L-1)} \right)^{K-1} \frac{1}{2} = (L(L-1))^{-(K-2)} \frac{1}{2} \end{aligned}$$

and this pair would constitute an novel row for the first Mallows component with at most $2L\phi^L$ approximate separable degree. This is true that i, j 's distance is at least L is $\sigma_2, \dots, \sigma^K$ regardless of the effect of the other groups. λ can be arbitrarily small if L is large.

Now, let $\mathcal{B}_k, k = 1, \dots, K$ denote the event that none of the Q/L groups has a λ -approximately separable row. Then, by union bound, the even of not being separable

can be upper bounded by,

$$\begin{aligned} \Pr(\bigcup \mathcal{B}_k) &\leq K \exp(-p_1 Q/L) \leq K \exp\left(-\frac{2Q}{(L^{K-1}(L-1)^{K-2})}\right) \\ &< K \exp(-p_1 Q/L) \leq K \exp\left(-\frac{2Q}{L^{2K-3}}\right) \end{aligned}$$

as a upper bound for the probability of β note being separable. For a given λ , we can choose $L = L(\phi, \lambda)$ such that $2L\phi^L \leq \lambda$, and this translate to the results in Lemma 14. \square

A.25 Separability for Measures and Irreducibility

We defined and studied the notion of separability for a $W \times K$ topic matrix β which is a finite collection of K probability distributions over a finite set (of size W). It turns out that we can extend the notion separability to a finite collection of measures over a measurable space. This necessitates making a small technical modification to the definition of separability to accommodate the possibility of only having “novel subsets” that have zero measure. We also show that our generalized definition of separability is equivalent to the so-called **irreducibility** property of a finite collection of measures that has recently been studied in the context of mixture models to establish conditions for the identifiability of the mixing components Blanchard and Scott [2014], Scott [2015].

Consider a collection of K measures ν_1, \dots, ν_K over a measurable space $(\mathcal{X}, \mathcal{F})$, where \mathcal{X} is a set and \mathcal{F} is a σ -algebra over \mathcal{X} . We define the generalized notion of separability for measures as follows.

Definition 3. (Separability) *A collection of K measures ν_1, \dots, ν_K over a measurable space $(\mathcal{X}, \mathcal{F})$ is separable if for all $k = 1, \dots, K$,*

$$\inf_{A \in \mathcal{F}: \nu_k > 0} \max_{j: j \neq k} \frac{\nu_j(A)}{\nu_k(A)} = 0. \quad (\text{A.16})$$

Separability requires that for each measure ν_k , there exists a sequence of mea-

surable sets $A_n^{(k)}$, of nonzero measure with respect to ν_k , such that, for all $j \neq k$, the ratios $\nu_j(A_n^{(k)})/\nu_k(A_n^{(k)})$ vanish asymptotically. Intuitively, this means that for each measure there exists a sequence of nonzero-measure measurable subsets that are asymptotically “novel” for that measure. When \mathcal{X} is a finite set as in topic modeling, this reduces to the existence of novel words as in Definition 2 and $A_n^{(k)}$ are simply the sets of novel words for topic k .

The separability property just defined is equivalent to the so-called irreducibility property. Informally, a collection of measures is irreducible if *only nonnegative linear combinations of them can produce a measure*. Formally,

Definition 4. (Irreducibility) *A collection of K measures ν_1, \dots, ν_K over a measurable space $(\mathcal{X}, \mathcal{F})$ is irreducible if the following condition holds: If $\forall A \in \mathcal{F}$, $\sum_{k=1}^K c_k \nu_k(A) \geq 0$, then for all $k = 1, \dots, K$, $c_k \geq 0$.*

For a collection of nonzero measures,¹ these two properties are equivalent. Formally,

Lemma 16. *A collection of nonzero measures ν_1, \dots, ν_K over a measurable space $(\mathcal{X}, \mathcal{F})$ is irreducible if and only if it is separable. In particular, a topic matrix β is irreducible if and only if it is separable.*

Proof. We first show that irreducibility implies separability, or equivalently, if the collection is not separable, then it is not irreducible. Suppose that $\{\nu_1, \dots, \nu_K\}$ is not separable. Then there exists some $k \in [K]$ and a $\delta > 0$ such that,

$$\inf_{A: \nu_k(A) > 0} \max_{j: j \neq k} \frac{\nu_j(A)}{\nu_k(A)} = \delta > 0.$$

Then $\forall A \in \mathcal{F} : \nu_k(A) > 0$, $\max_{j: j \neq k} \frac{\nu_j(A)}{\nu_k(A)} \geq \delta$. This implies that $\forall A \in \mathcal{F} : \nu_k(A) > 0$,

$$\sum_{j: j \neq k} \nu_j(A) - \delta \nu_k(A) \geq 0.$$

¹A measure ν is nonzero if there exists at least one measurable set A for which $\nu(A) > 0$.

On the other hand, $\forall A \in \mathcal{F} : \nu_k(A) = 0$, we have

$$\sum_{j: j \neq k} \nu_j(A) - \delta \nu_k(A) = \sum_{j: j \neq k} \nu_j(A) \geq 0.$$

Thus the linear combination $\sum_{j \neq k} \nu_j - \delta \nu_k$ with one strictly negative coefficient $-\delta$ is nonnegative over all measurable A . This implies that the collection of measures $\{\nu_1, \dots, \nu_K\}$ is not irreducible.

We next show that separability implies irreducibility. If the collection of measures $\{\nu_1, \dots, \nu_K\}$ is separable, then by the definition of separability, $\forall k, \exists A_n^{(k)} \in \mathcal{F}, n = 1, 2, \dots$, such that $\nu_k(A_n^{(k)}) > 0$ and $\forall j \neq k, \frac{\nu_j(A_n^{(k)})}{\nu_k(A_n^{(k)})} \rightarrow 0$ as $n \rightarrow \infty$. Now consider any linear combination of measures $\sum_{i=1}^K c_i \nu_i$ which is nonnegative over all measurable sets, i.e., for all $A \in \mathcal{F}, \sum_{i=1}^K c_i \nu_i(A) \geq 0$. Then $\forall k = 1, \dots, K$ and all $n \geq 1$ we have,

$$\begin{aligned} & \sum_{i=1}^K c_i \nu_i(A_n^{(k)}) \geq 0 \\ \Rightarrow & \nu_k(A_n^{(k)}) \left(c_k + \sum_{j \neq k} c_j \frac{\nu_j(A_n^{(k)})}{\nu_k(A_n^{(k)})} \right) \geq 0 \\ \Rightarrow & c_k \geq - \sum_{j \neq k} c_j \frac{\nu_j(A_n^{(k)})}{\nu_k(A_n^{(k)})} \rightarrow 0 \text{ as } n \rightarrow \infty. \end{aligned}$$

Therefore, $c_k \geq 0$ for all k and the collection of measures is irreducible. \square

References

- http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.
- Enhanced statistical rankings via targeted data collection. In *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, GA, USA, 2013.
- E. Airoldi, D. Blei, S. Fienberg, and E. Xing. Mixed membership stochastic block-models. *Journal of Machine Learning Research*, 9:1981–2014, 2008.
- E. Airoldi, D. Blei, E. Erosheva, and S. Fienberg. *Handbook of Mixed Membership Models and Their Applications*. Chapman and Hall/CRC, 2014.
- A. Anandkumar, D. Foster, D. Hsu, S. Kakade, and Y. K. Liu. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, Dec. 2012.
- A. Anandkumar, D. Hsu, A. Javanmard, and S. Kakade. Learning linear bayesian networks with latent variables. In *Proceedings the 30th International Conference on Machine Learning*, Atlanta, GA, Jun. 2013.
- A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research*,, 15:2239–2312, 2014.
- S. Arora, R. Ge, and A. Moitra. Learning topic models – going beyond SVD. In *Proceedings of the IEEE 53rd Annual Symposium on Foundations of Computer Science*, New Brunswick, NJ, USA, Oct. 2012.
- S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *Proc. of the 30th International Conference on Machine Learning*, Atlanta, GA, USA, Jun. 2013.
- A. Asuncion, P. Smyth, and M. Welling. Asynchronous distributed learning of topic models. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pages 81–88, 2009.
- P. Awasthi and A. Risteski. On some provably correct cases of variational inference for topic models. *arXiv:1503.06567 [cs.LG]*, 2015.

- P. Awasthi, A. Blum, O. Sheffet, and A. .Vijayaraghavan. Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. Montreal, Canada, Dec. 2014.
- H. Azari Soufiani, H. Diao, Z. Lai, and D. C. Parkes. Generalized random utility models with multiple types. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 73–81. Lake Tahoe, NV, USA, Dec. 2013.
- E. Azizi, E. Airoidi, and J. Galagan. Learning modular structures from network data and node variables. In *Proceedings of The 31st International Conference on Machine Learning*, Beijing, China, Jul. 2014.
- K. Bache and M. Lichman. UCI machine learning repository, 2013.
- T. Bansal, C. Bhattacharyya, and R. Kannan. A provable SVD-based algorithm for learning topics in dominant admixture corpus. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Monteral, Canada, Dec. 2014.
- J. Bioucas-Dias, A. Plaza, G. Camps-Valls, P. Scheunders, N. Nasrabadi, and J. Chanussot. Hyperspectral remote sensing data analysis and future challenges. *IEEE Geoscience and Remote Sensing Magazine*, 1(2):6–36, 2013.
- G. Blanchard and C. Scott. Decontamination of mutually contaminated models. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, pages 1–9, 2014.
- D. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.
- D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.
- D. Blei and J. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.
- D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Mar. 2003.
- J. Boardman. Automating spectral unmixing of aviris data using convex geometry concepts. In *Proceedings of the Annual JPL Airborne Geoscience Workshop*, page 1114, 1993.
- L. Busse, P. Orbanz, and J. Buhmann. Cluster analysis of heterogeneous rank data. In *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, USA, Jul. 2007.

- M. J. Carman, F. Crestani, M. Harvey, and M. Baillie. Towards query log based personalization using topic models. In *CIKM'10*, Toronto, Canada, Oct. 2010.
- A. Cichocki, R. Zdunek, A. H. Phan, and S. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. Wiley, 2009.
- W. Ding, P. Ishwar, M. H. Rohban, and V. Saligrama. Necessary and Sufficient Conditions for Novel Word Detection in Separable Topic Models. In *Advances in on Neural Information Processing Systems 26 (NIPS 2013), Workshop on Topic Models: Computation, Application*, Lake Tahoe, NV, USA, Dec. 2013a.
- W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama. Topic discovery through data dependent and random projections. In *Proceedings of the 30th International Conference on Machine Learning*, Atlanta, GA, USA, Jun. 2013b.
- W. Ding, P. Ishwar, and V. Saligrama. A Topic Modeling approach to Rank Aggregation. In *Advances in on Neural Information Processing Systems 27 (NIPS 2014), workshop on Analysis of Rank data*, Montreal, Canada, Dec. 2014a.
- W. Ding, M. H. Rohban, P. Ishwar, and V. Saligrama. Efficient Distributed Topic Modeling with Provable Guarantees. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, Reykjavik, Iceland, Apr. 2014b.
- W. Ding, P. Ishwar, and V. Saligrama. Most large Topic Models are approximately separable. In *2015 Information Theory and Applications Workshop (ITA 2015)*, San Diego, CA, 2015a.
- W. Ding, P. Ishwar, and V. Saligrama. A Topic Modeling approach to Ranking. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, San Diego, CA, May 2015b.
- W. Ding, P. Ishwar, and V. Saligrama. Learning mixed membership mallows model from pairwise comparisons. In *arXiv: 1504.00757 [cs.LG]*, 2015c.
- J. Doignon, A. Pekeč, and M. Regenwetter. The repeated insertion model for rankings: Missing link between two subset choice models. *Psychometrika*, 69(1):33–54, 2004.
- D. Donoho and V. Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, Lake Tahoe, NV, USA, Dec. 2004.
- E. Erosheva, S. Fienberg, and C. Joutard. Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, 1(2):346, 2007.

- V. Farias, S. Jagabathula, and D. Shah. A data-driven approach to modeling choice. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*. Vancouver, Canada, Dec. 2009.
- R. Gemulla, E. Nijkamp, P. J. Haas, and y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD*, 2011.
- N. Gillis and S. A. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714, 2014.
- D. Gleich and L. Lim. Rank Aggregation via Nuclear Norm Minimization. In *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining*, San Diego, CA, USA, 2011.
- P. Gopalan and D. Blei. Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences of the United States of America*, 110(36):14534–14539, 2013.
- I. Gormley and T. Murphy. A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, pages 1452–1477, 2008.
- I. C. Gormley, T. B. Murphy, et al. A grade of membership model for rank data. *Bayesian Analysis*, 4(2):265–295, 2009.
- T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235, 2004.
- M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, Lake Tahoe, NV, USA, Dec. 2010.
- T. Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd ACM SIGIR*, pages 50–57, 1999.
- F. Huang, U. N. Niranjan, M. U. Hakeem, P. Verma, and A. Anandkumar. Fast detection of overlapping communities via online tensor methods on gpus. *arXiv: 1309.0787 [cs.LG]*, 2013.
- S. Jagabathula and D. Shah. Inferring rankings under constrained sensing. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*. Vancouver, Canada, Dec. 2008.

- Y. Kim, W. Kim, and K. Shim. Latent ranking analysis using pairwise comparisons. In *2014 IEEE International Conference on Data Mining (ICDM)*, pages 869–874, Shenzhen, China, Dec. 2014.
- A. Kumar, V. Sindhwani, and P. Kambadur. Fast conical hull algorithms for near-separable non-negative matrix factorization. In *the 30th International Conference on Machine Learning*, Atlanta, GA, USA, Jun. 2013.
- G. Lebanon and J. D. Lafferty. Cranking: Combining rankings using conditional probability models on permutations. In *Proceedings of the 19th International Conference on Machine Learning*, Sydney, Australia, Jul. 2002.
- D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct. 1999.
- D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, Dec. 2004.
- F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Proc. of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, pages 524–531, San Diego, CA, Jun. 2005.
- C. Liu, H. Yang, J. Fan, L. He, and Y.-M. Y. Wang. Distributed nonnegative matrix factorization for web-scale dyadic data analysis on mapreduce. In *Proceedings of the 19th International Conference on World Wide Web*, Raleigh, NC, USA, 2010.
- T. Lu and C. Boutilier. Effective Sampling and Learning for Mallows Models with Pairwise-Preference Data. *Journal of Machine Learning Research*, 2014.
- Y. Lu and S. N. Negahban. Individualized rank aggregation using nuclear norm regularization. *arXiv preprint arXiv:1410.0860*, 2014.
- C. Mallows. Non-null ranking models. I. *Biometrika*, pages 114–130, 1957.
- D. Manrique-Vallier. Longitudinal mixed membership trajectory models for disability survey data. *The Annals of Applied Statistics*, 8(4):2268–2291, 2014.
- J. Marden. *Analyzing and modeling rank data*. Chapman and Hall, 1995.
- A. McCallum. Mallet: A machine learning for language toolkit. 2002.
- M. Meila and H. Chen. Dirichlet process mixtures of generalized mallows models. *Proceedings of the 26th Conference in Uncertainty in Artificial Intelligence*, 2010.
- S. Negahban, S. Oh, and D. Shah. Iterative Ranking from Pairwise Comparisons. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. Lake Tahoe, NV, USA, Dec. 2012.

- D. Newman, A. Asuncion, P. Smyth, and M. Welling. Distributed algorithms for topic models. *The Journal of Machine Learning Research*, 10:1801–1828, 2009.
- S. Oh and D. Shah. Learning mixed multinomial logit model from ordinal data. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, Montreal, Canada, Dec. 2014.
- R. Plackett. The analysis of permutations. *Applied Statistics*, pages 193–202, 1975.
- J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.
- T. Qin, X. Geng, and T.-Y. Liu. A new probabilistic model for rank aggregation. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*. Vancouver, Canada, Dec. 2010.
- A. Rajkumar and S. Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *Proceedings of the 31st International Conference on Machine Learning*, Beijing, China, Jun. 2014.
- B. Recht, C. Re, J. Tropp, and V. Bittorf. Factoring nonnegative matrices with linear programs. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, Lake Tahoe, NV, Dec. 2012.
- F. Ricci, L. Rokach, and B. Shapira. *Introduction to recommender systems handbook*. Springer, 2011.
- R. Salakhutdinov and A. Mnih. Bayesian probabilistic matrix factorization using markov chain monte carlo. In *Proceedings of the 25th International Conference on Machine Learning*, pages 880–887, Helsinki, Finland, Jun. 2008a.
- R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in neural information processing systems 21 (NIPS 2008)*, pages 1257–1264, 2008b.
- C. Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 838–846, 2015.
- N. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, San Diego, CA, May 2015.
- A. Smola and S. Narayanamurthy. An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, 3(1-2):703–710, 2010.

- D. Sontag and D. Roy. Complexity of inference in latent dirichlet allocation. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 1008–1016. 2011.
- M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.
- J. Tang, Z. Meng, X. Nguyen, Q. Mei, and M. Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proc. of the 31st International Conference on Machine Learning*, Beijing, China, Jul. 2014.
- Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.
- L. L. Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.
- A. Toscher, M. Jahrer, and R. M. Bell. The bigchaos solution to the netflix grand prize. Available from: http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf.
- S. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, Oct. 2009.
- M. Volkovs and R. Zemel. New Learning Methods for Supervised and Unsupervised Preference Aggregation. *Journal of Machine Learning Research*, 15:1135–1176, 2014.
- M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.
- H. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proc. of the 26th International Conference on Machine Learning*, Montreal, Canada, Jun. 2009.
- C. Wang and D. Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 448–456, 2011.
- F. L. Wauthier, M. I. Jordan, and N. Jojic. Efficient Ranking from pairwise comparisons. In *Proceedings of the 30th International Conference on Machine Learning*, pages 109–117, Atlanta, GA, USA, Jun. 2013.
- M. Woodbury, J. Clive, and A. Garson. Mathematical typology: a grade of membership technique for obtaining disease definition. *Computers and biomedical research*, 11(3):277–298, 1978.

E. Zermelo. Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460, 1929.

CURRICULUM VITAE

