Geometric Inference in Bayesian Hierarchical Models with Applications to Topic Modeling

by

Mikhail Yurochkin

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in the University of Michigan
2018

Doctoral Committee:

Associate Professor Long Nguyen, Chair
Associate Professor Qiaozhu Mei
Assistant Professor Yuekai Sun
Associate Professor Ambuj Tewari
Assistant Professor Shuheng Zhou

Mikhail Yurochkin

moonfolk@umich.edu

ORCID iD: 0000-0003-0153-6811

Dedicated to my family.

A C K N O W L E D G M E N T S

My PhD journey is coming to an end. These five years let me discover something that truly interests and motivates me — research in Statistics and Machine Learning. This, of course, would not have been possible without many people that I had a pleasure to work with and interact throughout my graduate school years.

First and foremost I want to say thank you to my adviser Long Nguyen. He was always there to inspire, teach and give fair criticism to help me think deeper. His level of understanding, broad knowledge and passion for research always fascinated me - let alone our email communication until 6AM. I am also very thankful for all the freedom Long gave me in pursuing different ideas and his thoughtful advice no matter what I was bringing to our meetings.

I would like to thank my internship mentors Nikolaos Vasiloglou and Hung Bui for the advice, interesting problems and freedom to pursue research ideas in the industrial setup. I am also grateful for their continued support after the internships and help in the job search process.

I am thankful to my thesis committee members, Professors Qiaozhu Mei, Ambuj Tewari, Shuheng Zhou and Yuekai Sun, for their helpful suggestions after my preliminary examination and in my final year, and to Professors Liza Levina and Ambuj Tewari for their help in navigating my first year of graduate school.

It was my pleasure to meet many wonderful and talented people throughout this journey. I have actively worked with and continue collaborating with Aritra Guha and Nhat Ho, my internship colleagues Zhiwei Fan and Dung Thai. Some of our joint work will appear in this thesis. I have had exciting research conversations and fun times with my friends and colleagues Elizabeth Hou, Jesus Arroyo, Liza Rebrova, Alexander Giessing, Abhinav Jain, Frank Cheng, Joanna Chiang, Kazem Shirani, Hossein Keshavarz, Jennifer Chu, Can Le, Naveen Narisetty, Ruofei Zhao, Tim NeCamp, Yumeng Li, Ai Rene Ong, Rui Shu, Tum Chaturapruek, Blake Wulfe and Uranzaya Batsaikhan. Thank you to Judy McDonald and David Clark for their great help with the paperwork matters and cheerful conversations we had.

Finally, this thesis is dedicated to my family, for their endless support, encouragement and love.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ABSTRACT

Unstructured data is available in abundance with the rapidly growing size of digital information. Labeling such data is expensive and impractical, making unsupervised learning an increasingly important field. Big data collections often have rich latent structure that statistical modeler is challenged to uncover. Bayesian hierarchical modeling is a particularly suitable approach for complex latent patterns. Graphical model formalism has been prominent in developing various procedures for inference in Bayesian models, however the corresponding computational limits often fall behind the demands of the modern data sizes. In this thesis we develop new approaches for scalable approximate Bayesian inference. In particular, our approaches are driven by the analysis of latent geometric structures induced by the models.

Our specific contributions include the following. We develop full geometric recipe of the Latent Dirichlet Allocation topic model. Next, we study several approaches for exploiting the latent geometry to first arrive at a fast weighted clustering procedure augmented with geometric corrections for topic inference, and then a nonparametric approach based on the analysis of the concentration of mass and angular geometry of the topic simplex, a convex polytope constructed by taking the convex hull of vertices representing the latent topics. Estimates produced by our methods are shown to be statistically consistent under some conditions. Finally, we develop a series of models for temporal dynamics of the latent geometric structures where inference can be performed in online and distributed fashion. All our algorithms are evaluated with extensive experiments on simulated and real datasets, culminating at a method several orders of magnitude faster than existing state-of-the-art topic modeling approaches, as demonstrated by experiments working with several million documents in a dozen minutes.

CHAPTER 1

Introduction

Bayesian hierarchical models are prominent tools for unsupervised learning of patterns in data. Unlabeled data is often available in vast amount - images, news articles, tweets, query logs, scientific articles, etc. Ghahramani (2004) provides an overview of unsupervised models such as Factor Analysis (Roweis & Ghahramani, 1999), probabilistic Principal Component Analysis (Roweis, 1998; Roweis & Ghahramani, 1999), Independent Component Analysis (Hyvärinen et al., 2004), mixture models (McLachlan & Peel, 2004), Hidden Markov Models (HMM) (Rabiner & Juang, 1986), State Space Models (SSM) (Kitagawa, 1996), which by now are over a decade old. Other important models from the same time period are admixtures (Pritchard et al., 2000), Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Finite Feature Model (FFM) (Ghahramani & Griffiths, 2005); nonparametric models such as Dirichlet Process (DP) (Ferguson, 1973), DP mixtures (Antoniak, 1974) and Beta Processes (Hjort, 1990). Larger data often has more structure like time stamps, location or publisher, hence requiring more sophisticated models to take into account all the information, and nonparametric approaches, as specifying number of patterns of interest apriori is very challenging and it is desirable to infer this quantity from the data instead. In the recent decades many complex Bayesian hierarchical and nonparametric models have been developed, where classical unsupervised models may play a role of building blocks. Hierarchical Dirichlet Process (HDP) (Teh et al., 2006) for nonparametric topic modeling, its extension for topics changing over time (Ahmed & Xing, 2012); document clustering and topic modeling with context (Nguyen et al., 2014b) utilizing HDP and Nested Dirichlet Process (Rodriguez et al., 2008); Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2011) and Hierarchical Beta-Bernoulli process (Thibaux & Jordan, 2007), its extensions combining IBP and HMM for time series data (Fox et al., 2009); nonparametric clustering of functional data (Nguyen & Gelfand, 2011) to name a few.

Unfortunately the data size and modeling are at crossroads — training sophisticated models on large data is often infeasible. Moreover, the problem remains acute for simpler models such as mixtures or Latent Dirichlet Allocation. Exact inference of the posterior

distribution or even Maximum Aposteriori estimation (MAP) is intractable and approximate inference techniques were developed. Markov Chain Monte Carlo (MCMC) (e.g., Gibbs sampler for LDA by Griffiths & Steyvers (2004)) and mean field variational inference (VI) (Jordan et al., 1999; Blei et al., 2017) are among the most common approaches. Variational inference historically was advocated as a faster alternative to sampling methods (e.g., comparison for the Boltzmann Machines by Anderson & Peterson (1987)). Nonetheless, classic versions of neither VI nor MCMC can efficiently handle modern data sizes of millions of observations.

Scalable approximate Bayesian inference is an active research area in the recent years and many approaches were proposed: divide-and-conquer and subsampling based MCMC (see Bardenet et al. (2017) for an overview), stochastic gradient MCMC (see Ma et al. (2015) for classification of some of the algorithms, Betancourt (2017) for introduction to Hamiltonian Monte Carlo, Mandt et al. (2017) for connection to stochastic gradient descent), stochastic variational inference (Hoffman et al., 2013).

It is evident that scalable inference even for simpler models remains an important problem up-to-date. Over the past couple decades MCMC and VI evolved a lot and became sophisticated approximate inference techniques applicable to many problems, however conceptually both approaches are rooted in the same fundamental concept of *graphical models* (Jordan et al., 2004). Graphical models formalism is based on graph theory and probability theory, where graph induced by the model is utilized to deduce various inference procedures. Such inference procedures, including Gibbs sampler and variational inference, belong to the family of *message-passing* algorithms (e.g., see review papers by Jordan et al. (1999); Ghahramani (2004); Wainwright et al. (2008)).

In this thesis we depart from the graphical models formalism and make first steps in developing a new way of approaching inference in Bayesian hierarchical models. Instead of a graph, our approach is based on the analysis of the *geometry* induced by the model. Specifically, we analyze geometric relationships among latent variables and data arising from the Latent Dirichlet Allocation model (Blei et al., 2003), which leads us to propose new inference algorithms for the latent geometry of the LDA, including nonparametric inference and inference for some of the LDA extensions. Key advantage of our algorithms is orders of magnitude speed ups comparative to graphical models based approaches.

In Section 1.1 we give an overview of the Latent Dirichlet Allocation, some of its extensions and estimation procedures. In Section 1.2 we review role of geometry in the LDA and some other Machine Learning problems. Key ingredients of our work are outlined in Section 1.3. Main problems and contributions are summarized in Section 1.4.

1.1 Latent Dirichlet Allocation

With the rapidly growing size of digital information it is important to be able to explore, organize and summarize it. Internet is full of news, blogs, Web pages, scientific articles and other text, image, audio or video types of data. We focus our attention on the class of algorithms called Topic Models. The goal of such algorithms is to summarize collections of data (e.g., text data) by finding topics in it, which are essentially distributions over words. One of the pioneering works in this field is a probabilistic topic model - Latent Dirichlet Allocation by Blei et al. (2003). Examples of topics learned from the New York Times articles are demonstrated in Table 1.1. In this section we review Latent Dirichlet Allocation, its graphical representation, some of its predecessors and extensions and common inference approaches.

Table 1.1: New York Times topics

Cooking	*Stem Cells*	*Antitrust*	*LGBT*	*Elections*
cup	cell	Microsoft	gay	ballot
minutes	stem	window	lesbian	Al Gore
tablespoon	research	company	right	election
add	human	software	sex	votes
teaspoon	scientist	case	marriage	recount
pepper	cloning	system	group	Florida
oil	patient	operating	couples	court
sugar	disease	computer	sexual	vote
butter	phones	antitrust	partner	voter
pan	researcher	court	issue	count

1.1.1 Model

First we summarize the data generative process of the Latent Dirichlet Allocation. Let $\alpha \in \mathbb{R}_+^K$ and $\eta \in \mathbb{R}_+^V$ be hyperparameters, where V denotes the number of words in a vocabulary, and K the number of topics. The K topics are represented as distributions on words:

$$\beta_k|\eta \sim \text{Dir}_V(\eta), \text{ for } k = 1, \ldots, K.$$

Each of the M documents can be generated as follows. First, draw the document topic proportions:

$$\theta_m|\alpha \sim \text{Dir}_K(\alpha), \text{ for } m = 1, \ldots, M.$$

Next, for each of the N_m words in document m, pick a topic label z and then sample a word d from the chosen topic:

$$z_{n_m}|\theta_m \sim \text{Categorical}(\theta_m);\ d_{n_m}|z_{n_m}, \beta_{1\ldots K} \sim \text{Categorical}(\beta_{z_{n_m}}).$$

Each of the resulting documents is a vector of length N_m with entries $d_{n_m} \in \{1, \ldots, V\}$, where $n_m = 1, \ldots, N_m$. Because these words are exchangeable by the modeling, they are equivalently represented as a vector of word counts $w_m \in \mathbb{N}^V$. Corresponding graphical model representation is shown in Figure 1.1. From the model construction it follows that each document can exhibit multiple topics to a different extent. This property, encoded in corresponding θ_m, distinguishes LDA from mixture models and makes it more suitable for modeling text data, where single topic may not be sufficient to describe a document.

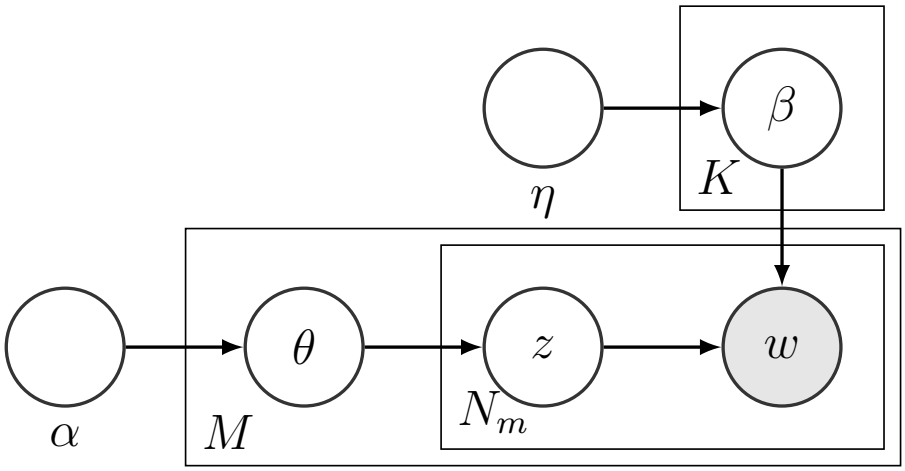


Figure 1.1: Latent Dirichlet Allocation graphical model representation.

Latent Semantic Analysis (Deerwester et al., 1990) is a precursor of the LDA model casting topics as basis vectors of low dimensional latent subspace. Hofmann (1999) added probabilistic component to it. Admixture (Pritchard et al., 2000) is an equivalent to LDA model widely used in the field of population genetics.

LDA is a fundamental building block for many more sophisticated topic models. It is a parametric model, meaning that we need to provide number of topics as an input, which could be challenging in practice. Nonparametric version, based on the Dirichlet process (Ferguson, 1973) is known as Hierarchical Dirichlet Process (Teh et al., 2006). Apriori topics are assumed independent, Correlated Topic model (Blei & Lafferty, 2006a) relaxes this assumption by using logistic-normal distribution instead of Dirichlet to model latent topic proportions of a document. Similar idea is used to allow time changing topics - Dynamic Topic Models (Blei & Lafferty, 2006b) and a nonparametric version by Ahmed & Xing

(2012). Nested Chinese Restaurant Process (Blei et al., 2010) allows for topic hierarchies. Supervised LDA (Mcauliffe & Blei, 2008) incorporates prediction in the model, allowing for documents with labels. Multilevel Clustering with Context (Nguyen et al., 2014b) allows to use context information (like authors, time and location stamps) and cluster documents by combining Hierarchical and Nested Dirichlet Processes.

1.1.2 Inference

Graphical model formalism leads to the two most common approaches for inference of the LDA topics — Gibbs sampling (Griffiths & Steyvers, 2004) based on the Multinomial-Dirichlet conjugacy and mean field variational inference (Blei et al., 2003) minimizing the Kullback-Leibler divergence between posterior and factorized posterior. Brief summary of these methods is given in the Supplement.

Neither Gibbs sampling nor variational inference in their respective original formulations scale well to large corpora of millions of documents. Training time gets off practical limits for more complex and nonparametric LDA extensions. In the past decade significant research efforts have been made to address this concern. Online (Hoffman et al., 2010), distributed (Newman et al., 2008) and both online and distributed (Broderick et al., 2013) algorithms have been developed. Online algorithms for Hierarchical Dirichlet Process (HDP) (Teh et al., 2006), a nonparametric counterpart of LDA, have also been proposed (Wang et al., 2011; Bryant & Sudderth, 2012). Both classes of inference algorithms (i.e. sampling and variational inference), their virtues notwithstanding, are known to exhibit certain deficiencies, which also propagate through their modern extensions. The problem can be traced back to the need for approximating or sampling from the posterior distributions of the latent variables representing the topic labels. Since these latent variables are not geometrically intrinsic — any permutation of the labels yields the same likelihood — the manipulation of these redundant quantities tend to slow down the computation, and compromise with the learning accuracy.

Another fruitful perspective on topic modeling can be obtained by partially stripping away the distributional properties of the probabilistic model and turning the estimation problem into a form of matrix factorization (Deerwester et al., 1990; Xu et al., 2003; Anandkumar et al., 2012; Arora et al., 2013). We call this the linear subspace viewpoint. For instance, the Latent Semantic Analysis approach (Deerwester et al., 1990) looks to find a latent subspace via singular-value decomposition, but has no topic structure. Notably, the RecoverKL by Arora et al. (2013) is one of the recent fast algorithms with provable guarantees coming from the linear subspace perspective.

1.2 Geometry in Machine Learning

In this section we summarize prior work concerning the geometry of the Latent Dirichlet Allocation. Then we give some examples of geometry arising in other Machine Learning problems.

1.2.1 Geometry of the Latent Dirichlet Allocation

The key idea underlining the geometric viewpoint of the Latent Dirichlet Allocation is that latent topics induce a convex set to be called a *topic polytope*. Observed documents then correspond to points randomly drawn inside this topic polytope. Probabilistic Latent Semantic Analysis (Hofmann, 1999) contains early hints of this viewpoint, however it does not explore the distribution of the documents inside the topic polytope. One of the key Latent Dirichlet Allocation (Blei et al., 2003) ideas is to place a Dirichlet prior on the inside of the topic polytope. Figure 4 of Blei et al. (2003) summarizes this idea. Geometric viewpoint had laid dormant for quite some time, until studied in depth in the work of Nguyen (2015), who demonstrated theoretically that posterior distribution of the LDA topics contracts towards the true latent topic polytope. The contraction behavior was also verified in practice (Tang et al., 2014). In this thesis we complete geometric interpretation of *all* of the LDA latent variables (i.e., topics and topic proportions) and use it to develop a series of efficient algorithms directly targeting estimation of the latent topic polytope.

1.2.2 Geometry in other problems

Geometry has always played an important role in Machine Learning. Support Vector Machines (Vapnik, 2006) is among famous classification algorithms with intuitive geometric motivation. Principal Component Analysis can be interpreted geometrically; k-means clustering is tightly related to a geometric problem of Centroidal Voronoi Tessellation (Du et al., 1999). Many problems go beyond Euclidean geometry — topological data analysis (Wasserman, 2016), manifold learning (Belkin et al., 2006), some of the deep learning (Bronstein et al., 2017).

Another important problem, related to Topic Modeling through the linear subspace perspective, is the Nonnegative Matrix Factorization (NMF) (Lee & Seung, 1999). Although NMF is a NP-hard problem in general (Vavasis, 2009), it may be solved provably when *separability* is assumed (Arora et al., 2012). Under this assumption NMF problem has an elegant geometric interpretation — some (unknown) rows of the data matrix form a polytope containing rest of the data. Damle & Sun (2017) utilized this geometric treatment of NMF

to propose scalable estimation algorithms. In topic modeling separability assumption is usually called an *anchor word* assumption — each topic is assumed to contain a word absent in all other topics. RecoverKL by Arora et al. (2013) is one of the recent fast algorithms utilizing the anchor word assumption. We will show that one of our geometric algorithms can be used to improve RecoverKL and make it nonparametric.

1.3 Key ideas in our framework

In this section we shall present the full geometric recipe of the Latent Dirichlet Allocation and outline some of the concepts we will use for the algorithms and models development.

In Section 1.1.2 we noted that the root of the inefficiency of sampling and variational inference methods can be traced to the need for approximating the posterior distributions of the latent variables representing the topic labels — these are not geometrically intrinsic as any permutation of the labels yields the same likelihood. By integrating out the latent variables that represent the topic labels, we obtain a geometric formulation of the LDA. Indeed, integrating z's out yields that, for $m = 1, \ldots, M$,

$$w_m|\theta_m, \beta_{1\ldots K}, N_m \sim \text{Multinomial}(p_{m1}, \ldots, p_{mV}, N_m),$$

where p_{mi} denotes probability of observing the i-th word from the vocabulary in the m-th document, and is given by

$$p_{mi} = \sum_{k=1}^{K} \theta_{mk}\beta_{ki} \text{ for } i = 1, \ldots, V;\ m = 1, \ldots, M.$$

Resulting graphical model representation is given in Fig. 1.2. The crucial geometric observation is that the exact same form of p_m would arise if we interpreted topic proportions θ_m as *barycentric coordinates* of document m with respect to the topic polytope $B = \text{Conv}(\beta_1, \ldots, \beta_K)$ and p_m as the corresponding Cartesian coordinates.

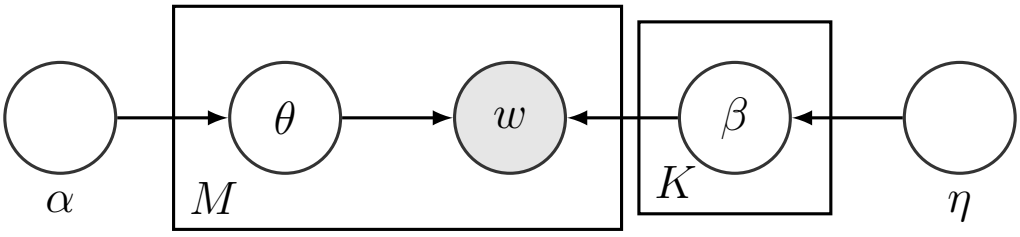
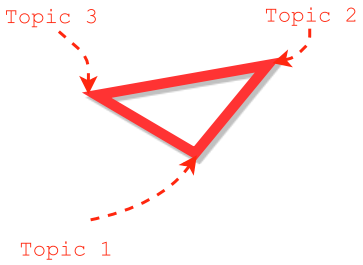
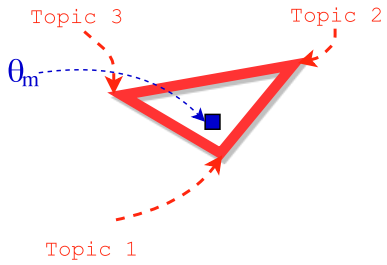


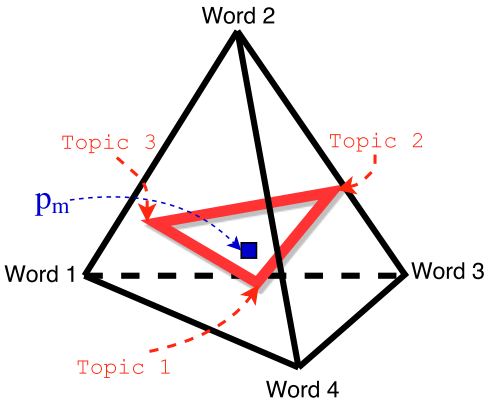
Figure 1.2: Latent Dirichlet Allocation with integrated topic labels.



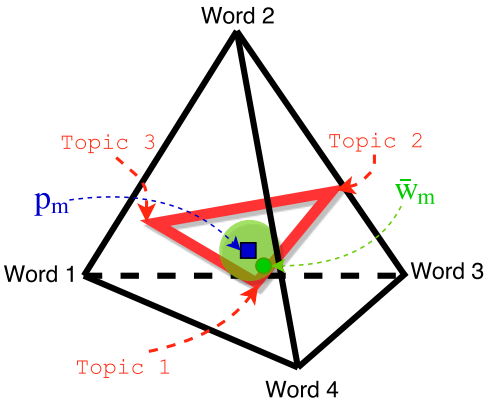
(a) Topic polytope



(b) Barycentric coordinates



(c) Cartesian coordinates



(d) Normalized document

Figure 1.3: Toy geometric illustration of the LDA document generation.

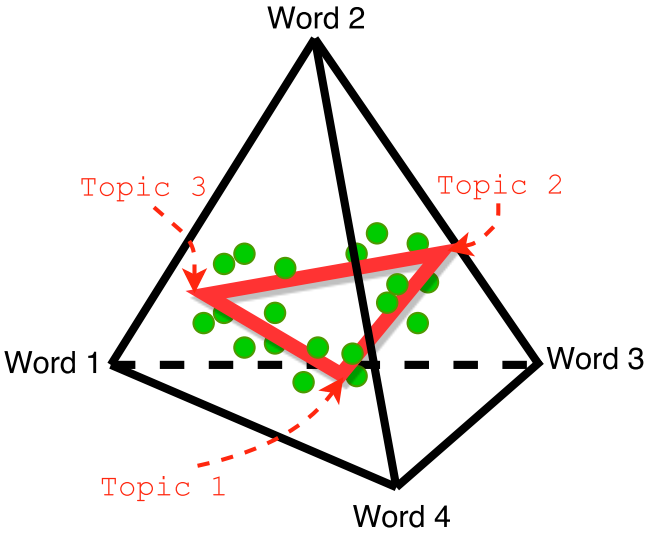


Figure 1.4: Toy geometric illustration of topic polytope and document collection.

Let us now restate the generative process of the LDA with each layer of the model accompanied by a toy graphical illustration from Fig. 1.3.

Generate K topics — extreme points of the topic polytope (Fig. 1.3a):

$$\beta_k|\eta \sim \text{Dir}_V(\eta), \text{ for } k = 1, \ldots, K.$$

For each of the M documents sample a vector of barycentric coordinates (Fig. 1.3b):

$$\theta_m|\alpha \sim \text{Dir}_K(\alpha).$$

Translate into Cartesian coordinate system (Fig. 1.3c):

$$p_m = \sum_k \theta_{mk}\beta_k.$$

Generate and normalize document (Fig. 1.3d):

$$w_m \sim \text{Multinomial}(p_m, N_m), \bar{w}_m = w_m/N_m.$$

Resulting normalized document $\bar{w}_m$ is a noisy estimate of p_m. The whole (normalized) corpora can be viewed as a cloud of points around the latent topic polytope (Fig. 1.4).

In this thesis we consider a variety of approaches targeting topic polytope estimation. We shall establish a geometric objective function, its connection to the LDA likelihood and subspace learning. To arrive at scalable geometric inference algorithms we analyze the concentration of mass inside the topic polytope. First, by investigating the Centroidal Voronoi Tessellation (Du et al., 1999) of the topic polytope, and then, by covering the topic polytope with suitable cones and spherical balls.

Given efficient nonparametric procedure for topic estimation, it is natural to ask if our geometric framework can be extended to more sophisticated topic modeling scenarios. To answer this question we build a series of models directly targeting the *latent geometry*. Specifically, we develop a nonparametric model for the temporal dynamics of the topic polytope. We shall utilize an isometric embedding of the unit sphere in the word simplex, so that the evolution of topic polytopes may be modeled by a collection of (random) trajectories of points residing on the unit sphere. Instead of worrying about mix-matching vertices in an ad hoc fashion, we appeal to the Bayesian nonparametric modeling framework that allow the number of topic vertices to be random and vary across time. The mix-matching between topic variables shall be guided by the assumption on the smoothness of the collection of global trajectories on the sphere using Von Mises-Fisher dynamics (Mardia & Jupp, 2009). The selection of active topics at each time point will be guided by a nonparametric prior on the random binary matrices via the (hierarchical) Beta-Bernoulli process (Thibaux & Jordan,

2007). For the inference part, we demonstrate connection between MAP estimation in the Beta-Bernoulli process and the Hungarian matching algorithm (Kuhn, 1955).

1.4 Main problems and contributions

In this section we provide an extended summary of the problems considered in this thesis and our contributions in regard to these problems.

- Geometric formulation of the Latent Dirichlet Allocation.
- Geometric procedure for estimating topic proportions given a topic simplex.
- Fast parametric geometric algorithm based on the analysis of Centroidal Voronoi Tessellation of a topic polytope.
- Fast nonparametric geometric algorithm based on the analysis of the concentration of mass inside the topic simplex and its coverage with suitable cones and spheres.
- New model for temporal dynamics of the topic polytope and corresponding online and distributed inference algorithm.
- Novel Bayesian hierarchical model for supervised adaptive selection of interactions of arbitrary order among predictor variables.

At a high level, we uncover and exploit geometry induced by the Latent Dirichlet Allocation model to develop several algorithms that are much faster and as accurate as their graphical models inspired counterparts. Being able to efficiently infer the latent geometry, we propose new models to allow for dynamically evolving latent geometry, where inference can be done in online and distributed fashion allowing for topic estimation from several millions of documents in a dozen of minutes.

1.4.1 Geometric analysis of the LDA and Geometric Dirichlet Means algorithm for topic inference

In Chapter 2 we further elaborate on the geometric viewpoint of the LDA outlined in Section 1.3. We develop and analyze a new class of algorithms for topic inference, which exploits both the convex geometry of topic models and the distributional properties they carry. The main contributions in Chapter 2 are the following: (i) we investigate a geometric loss function to be optimized, which can be viewed as a surrogate to the LDA's likelihood;

this leads to a novel estimation and inference algorithm — the Geometric Dirichlet Means (GDM) algorithm, which builds upon a weighted k-means clustering procedure and is augmented with a geometric correction for obtaining polytope estimates; (ii) we prove that the GDM algorithm is consistent, under conditions on the Dirichlet distribution and the geometry of the topic polytope; (iii) we propose a nonparametric extension of GDM and discuss geometric treatments for some of the LDA extensions; (v) finally we provide a thorough evaluation of our method against a Gibbs sampler, a variational algorithm, and the RecoverKL algorithm. Our method is shown to be comparable to a Gibbs sampler in terms of estimation accuracy, but much more efficient in runtime. It outperforms RecoverKL (Arora et al., 2013) algorithm in terms of accuracy, in some realistic settings of simulations and in real data.

1.4.2 Conic Scan-and-Cover algorithms for nonparametric topic modeling

In Chapter 3 we shall continue to amplify the geometric viewpoint to address *nonparametric topic modeling*, a setting in which the number of topics is unknown, as is the distribution inside the topic polytope (in some situations). We will propose algorithms for topic estimation by explicitly accounting for the concentration of mass and angular geometry of the topic polytope, typically a simplex in topic modeling applications. The geometric intuition is fairly clear: each vertex of the topic simplex can be identified by a ray emanating from its center (to be defined formally), while the concentration of mass can be quantified for the cones hinging on the apex positioned at the center. Such cones can be rotated around the center to scan for high density regions inside the topic simplex — under mild conditions such cones can be constructed efficiently to recover both the number of vertices and their estimates. We will also extend RecoverKL (Arora et al., 2013) to a nonparametric setting and show that the anchor word assumption appears to limit the number of topics one can efficiently learn.

1.4.3 Streaming dynamic and distributed inference of latent geometric structures

When data and the associated modeling are indexed by time dimension, it is of interest to study the temporal dynamics of the latent geometric structure that arises. In Chapter 4 we focus on the modeling and algorithm for analyzing the temporal dynamics of a *topic polytope*. A number of authors have extended the basic topic modeling framework to analyze

how topics evolve over time. The Dynamic Topic Models (DTM) (Blei & Lafferty, 2006b) demonstrated the importance of accounting for non-exchangeability between document groups, particularly when the time information is provided. Another approach is to fix the topics and only consider evolving topic popularity (Wang & McCallum, 2006). Hong et al. (2011) extended such an approach to multiple corpora. Ahmed & Xing (2012) proposed an interesting nonparametric construction extending DTM where topics can appear or eventually die out. The evolution of the latent geometric structure arising from the model (the topic polytope) is implicitly present in all these works, however it was not explicitly considered and analyzed.

Directly confronting the modeling and inference of the temporal dynamics of the topic polytope offers several opportunities and challenges. To start, what is the suitable space that we may consider the topic polytope to be represented? As topics evolve over time, so are the number of topics that may become active and dormant again, raising interesting choices about modeling of this evolution. Interesting issues arise in the inference, too. For instance, what is the principled way of mix-matching the vertices of a polytope to its reincarnation in the next time point? These challenges also open the door for directly modeling the geometric structure in a way that facilitates efficient inference.

To this end, we construct a sequence of Bayesian nonparametric models in increasing levels of complexity: the simpler model captures the temporal dynamics of a sequence of topic polytopes, while the full model describes the temporal dynamics of a collection of topic polytopes as they arise from multiple corpora. The semantic of topics that arise from our models can be interpreted as follows: there is a latent collection of global topics of unknown cardinality that evolve over time (e.g. topics in science or social topics in Twitter) and each year (or day) a subset of the global topics is elucidated by the community (i.e. some topics may be dormant at a given time point). The nature of each global topic may change smoothly (via varying word frequencies). Additionally, different subsets of global topics are associated with different groups (e.g. journals or Twitter location stamps), some become active and inactive over time.

It is interesting to note that our model construction leads to a suite of approximate inference algorithm that scales well in an online and distributed setting. In particular, the online MAP update of the latent topic polytope can be viewed as solving an optimal matching problem for which a fast Hungarian matching algorithm (Kuhn, 1955) can be applied. Our approach is able to perform dynamic nonparametric topic inference on 3 million documents in 12 minutes, which is significantly faster than prior static online and/or distributed topic modeling algorithms (Newman et al., 2008; Hoffman et al., 2010; Wang et al., 2011; Bryant & Sudderth, 2012; Broderick et al., 2013).

1.4.4 Multi-way Interacting Regression via Factorization Machines

A fundamental challenge in supervised learning, particularly in regression, is the need for learning functions which produce accurate prediction of the response, while retaining the explanatory power for the role of the predictor variables in the model. The standard linear regression method is favored for the latter requirement, but it fails the former when there are complex interactions among the predictor variables in determining the response. The challenge becomes even more pronounced in a high-dimensional setting – there are exponentially many potential interactions among the predictors, for which it is simply not computationally feasible to resort to standard variable selection techniques (cf. Fan & Lv (2010)).

There are numerous examples where accounting for the predictors' interactions is of interest, including problems of identifying epistasis (gene-gene) and gene-environment interactions in genetics (Cordell, 2009), modeling problems in political science (Brambor et al., 2006) and economics (Ai & Norton, 2003). In the business analytics of retail demand forecasting, a strong prediction model that also accurately accounts for the interactions of relevant predictors such as seasons, product types, geography, promotions, etc. plays a critical role in the decision making of marketing design.

A simple way to address the aforementioned issue in the regression problem is to simply restrict our attention to lower order interactions (i.e. 2- or 3-way) among predictor variables. This can be achieved, for instance, via a support vector machine (SVM) using polynomial kernels (Cristianini & Shawe-Taylor, 2000), which pre-determine the maximum order of predictor interactions. In practice, for computational reasons the degree of the polynomial kernel tends to be small. Factorization machines (Rendle, 2010) can be viewed as an extension of SVM to sparse settings where most interactions are observed only infrequently, subject to a constraint that the interaction order (a.k.a. interaction depth) is given. Neither SVM nor FM can perform any selection of predictor interactions, but several authors have extended the SVM by combining it with ℓ_1 penalty for the purpose of feature selection (Zhu et al., 2004) and gradient boosting for FM (Cheng et al., 2014) to select interacting features. It is also an option to perform linear regression on as many interactions as we can and combine it with regularization procedures for selection (e.g. LASSO (Tibshirani, 1996) or Elastic net (Zou & Hastie, 2005)). It is noted that such methods are still not computationally feasible for accounting for interactions that involve a large number of predictor variables.

In Chapter 5 we propose a regression method capable of adaptive selection of multi-way interactions of arbitrary order (MiFM for short), while avoiding the combinatorial complexity growth encountered by the methods described above. MiFM extends the basic factorization mechanism for representing the regression coefficients of interactions among

the predictors, while the interaction selection is guided by a prior distribution on random hypergraphs. The prior, which does not insist on the upper bound on the order of interactions among the predictor variables, is motivated from but also generalizes Finite Feature Model, a parametric form of the well-known Indian Buffet process (IBP) (Ghahramani & Griffiths, 2005). We introduce a notion of the hypergraph of interactions and show how a parametric distribution over binary matrices can be utilized to express interactions of unbounded order. In addition, our generalized construction allows us to exert extra control on the tail behavior of the interaction order. IBP was initially used for infinite latent feature modeling and later utilized in the modeling of a variety of domains (see a review paper by Griffiths & Ghahramani (2011)).

In developing MiFM, our contributions are the following: (i) we introduce a Bayesian multi-linear regression model, which aims to account for the multi-way interactions among predictor variables; part of our model construction includes a prior specification on the hypergraph of interactions — in particular we show how our prior can be used to model the incidence matrix of interactions in several ways; (ii) we propose a procedure to estimate coefficients of arbitrary interactions structure; (iii) we establish posterior consistency of the resulting MiFM model, i.e., the property that the posterior distribution on the true regression function represented by the MiFM model contracts toward the truth under some conditions, without requiring an upper bound on the order of the predictor interactions; and (iv) we present a comprehensive simulation study of our model and analyze its performance for retail demand forecasting and case-control genetics datasets with epistasis. The unique strength of the MiFM method is the ability to recover meaningful interactions among the predictors while maintaining a competitive prediction quality compared to existing methods that target prediction only.

1.5 Thesis organization

The reminder of this thesis proceeds as follows:

Chapter 2: Geometric Dirichlet Means algorithm for topic inference

This chapter develops geometric viewpoint of the LDA, defines corresponding geometric loss functions and proposes new geometrically inspired topic polytope estimation algorithm.

Chapter 3: Conic Scan-and-Cover algorithms for nonparametric topic modeling

This chapter continues with the geometric viewpoint of Chapter 2 to investigate concentration of mass inside the topic polytope and to propose a fast nonparametric topic estimation algorithm.

Chapter 4: Streaming dynamic and distributed inference of latent geometric struc-

tures

This chapter develops a model of temporal dynamics of the topic polytope and introduces an online and distributed estimation algorithm based on Chapter 3.

Chapter 5: Multi-way Interacting Regression via Factorization Machines

This chapter proposes a notion of hypergraph of interactions and develops a corresponding Bayesian hierarchical model for uncovering high-order interactions in the regression setting.

Chapter 6: Conclusions and suggestions

This chapter summarizes contributions of this thesis and discusses ideas for the future work.

Each chapter is sufficiently self-contained covering all relevant background knowledge and literature. Chapter 2 initiates the geometric journey taking off from a more classical likelihood perspective. Chapter 3 culminates the geometric exploration. Chapter 4 uses algorithmic results of Chapter 3, although it could be read independently. Chapter 5 addresses a problem in the supervised learning domain and can be read without the background in the previous chapters.

Appendix

1.A Collapsed Gibbs Sampling

Here we briefly summarize Collapsed Gibbs sampler of Griffiths & Steyvers (2004). Due to the conjugacy of Categorical and Dirichlet distributions we can integrate out θ and β to be left only with latent variables z, for which closed form full conditional is available:

$$P(z_{mi} = k|z_{-mi}, W) \propto \frac{n_{-i,k}^{(w_{mi})} + \gamma}{n_{-i,k}^{(\cdot)} + V\gamma} \frac{n_{-i,k}^{(m)} + \alpha}{n_{-i,\cdot}^{(m)} + K\alpha}.$$

First ratio gives probability of word w_{mi} in topic k. $n_k^{(w_{mi})}$ is the number of times word w_{mi} has been assigned to topic k. $-i$ index means excluding current assignment of z_{mi}. Second ratio gives probability of topic k in document m. $n_k^{(m)}$ is the number of times a word from document m has been assigned to topic k. Markov Chain Monte Carlo algorithm is then to start with some initial state and sample topic assignments for each word in each document until chain approaches target posterior distribution. Estimates for topics and topic proportions are:

$$\beta_{kv} = \frac{n_k^{(v)} + \gamma}{n_k^{(\cdot)} + V\gamma}.$$

$$\theta_{mk} = \frac{n_k^{(m)} + \alpha}{n_\cdot^{(m)} + K\alpha}.$$

This algorithm converges asymptotically, but as in any MCMC it is hard to assess rate and it is quite slow. Also we have to do sampling for every word, which is one of the inefficiencies we are addressing in our work. Moreover it is problematic to estimate θ of unseen document, since we have to sample topic assignments to obtain posterior for the document.

1.B Variational Inference

Variational inference (Blei et al., 2003) is based on decoupling latent variables, and minimizing KL divergence between factorized posterior with unknown parameters and true posterior. Resulting objective is non-convex, therefore it is only guaranteed to converge to local optima. Next we give a short description of how VI works.

Let for a probabilistic graphical model z be latent variables and x - data. We are interested in the posterior $p(z|x)$, which is often intractable, as in the LDA case. We want to

approximate the true posterior with some easier-to-compute distribution $q(z)$ by minimizing

$$\begin{aligned}\text{KL}(q||p) &= \mathbb{E}_q\left[\log\frac{q(z)}{p(z|x)}\right] \\ &= \mathbb{E}_q\left[\log q(z)\right] - \mathbb{E}_q\left[\log p(z|x)\right] \\ &= -\left(\underbrace{\mathbb{E}_q\left[\log p(z,x)\right] - \mathbb{E}_q\left[\log q(z)\right]}_{\text{ELBO}}\right) + \mathbb{E}_q\left[\log p(x)\right].\end{aligned}$$

Evidence Lower Bound (ELBO) is a lower bound of the log-likelihood of the data and could be derived via Jensen's inequality. Last term $\mathbb{E}_q\left[\log p(x)\right]$ does not depend on $q(z)$, therefore minimizing KL is equivalent to maximizing ELBO. For LDA we take $q(z)$ as factorized (decoupling latent variables) posterior (Blei et al., 2003):

$$q(\beta,\theta,z|\lambda,\gamma,\phi) = \left[\prod_{k=1}^{K} q(\beta_k|\lambda_k)\right]\cdot\left[\prod_{m=1}^{M} q(\theta_m|\gamma_m)\right]\cdot\left[\prod_{m=1}^{M}\prod_{n=1}^{N_m} q(z_{mn}|\phi_{mw_{mn}})\right], \text{ where}$$

$$q(\beta_k|\lambda_k) = \text{Dir}_V(\beta_k|\lambda_k),\ q(\theta_m|\gamma_m) = \text{Dir}_K(\theta_m|\gamma_m), \text{and}$$

$$q(z_{mn}|\phi_{mw_{mn}}) = \text{Categorical}_K(z_{mn}|\phi_{mw_{mn}}).$$

We minimize ELBO over parameters (λ_{ki}), (γ_{mk}), (ϕ_{mik}) with $k \in \{1,...,K\}$, $i \in \{1,...,V\}$, $m \in \{1,...,M\}$. Common approach is to do coordinate ascent, but there are more modern techniques, that are faster.

CHAPTER 2

Geometric Dirichlet Means algorithm for topic inference

We propose a geometric algorithm for topic learning and inference that is built on the convex geometry of topics arising from the Latent Dirichlet Allocation (LDA) model and its nonparametric extensions. To this end we study the optimization of a geometric loss function, which is a surrogate to the LDA's likelihood. Our method involves a fast optimization based weighted clustering procedure augmented with geometric corrections, which overcomes the computational and statistical inefficiencies encountered by other techniques based on Gibbs sampling and variational inference, while achieving the accuracy comparable to that of a Gibbs sampler.[1] The topic estimates produced by our method are shown to be statistically consistent under some conditions. The algorithm is evaluated with extensive experiments on simulated and real data.[2]

2.1 Introduction

Most learning and inference algorithms in the probabilistic topic modeling literature can be delineated along two major lines: the variational approximation popularized in the seminal paper of Blei et al. (2003), and the sampling based approach studied by Pritchard et al. (2000) and other authors. Both classes of inference algorithms, their virtues notwithstanding, are known to exhibit certain deficiencies, which can be traced back to the need for approximating or sampling from the posterior distributions of the latent variables representing the topic labels. Since these latent variables are not geometrically intrinsic — any permutation of the labels yields the same likelihood — the manipulation of these redundant quantities tend to slow down the computation, and compromise with the learning accuracy.

[1] Code is available at https://github.com/moonfolk/Geometric-Topic-Modeling.

[2] This chapter has been published in Yurochkin & Nguyen (2016).

In this chapter we take a convex geometric perspective of the Latent Dirichlet Allocation, which may be obtained by integrating out the latent topic label variables. As a result, topic learning and inference may be formulated as a convex geometric problem: the observed documents correspond to points randomly drawn from a *topic polytope*, a convex set whose vertices represent the topics to be inferred. The original paper of Blei et al. (2003) (see also Hofmann (1999)) contains early hints about a convex geometric viewpoint, which is left unexplored. This viewpoint had laid dormant for quite some time, until studied in depth in the work of Nguyen and co-workers, who investigated posterior contraction behaviors for the LDA both theoretically and practically (Nguyen, 2015; Tang et al., 2014).

Another fruitful perspective on topic modeling can be obtained by partially stripping away the distributional properties of the probabilistic model and turning the estimation problem into a form of matrix factorization (Deerwester et al., 1990; Xu et al., 2003; Anandkumar et al., 2012; Arora et al., 2013). We call this the linear subspace viewpoint. For instance, the Latent Semantic Analysis approach (Deerwester et al., 1990), which can be viewed as a precursor of the LDA model, looks to find a latent subspace via singular-value decomposition, but has no topic structure. Notably, the RecoverKL by Arora et al. (2013) is one of the recent fast algorithms with provable guarantees coming from the linear subspace perspective.

The geometric perspective continues to be the main force driving this work. We develop and analyze a new class of algorithms for topic inference, which exploits both the convex geometry of topic models and the distributional properties they carry. The main contributions in this chapter are the following: (i) we investigate a geometric loss function to be optimized, which can be viewed as a surrogate to the LDA's likelihood; this leads to a novel estimation and inference algorithm — the Geometric Dirichlet Means algorithm, which builds upon a weighted k-means clustering procedure and is augmented with a geometric correction for obtaining polytope estimates; (ii) we prove that the GDM algorithm is consistent, under conditions on the Dirichlet distribution and the geometry of the topic polytope; (iii) we propose a nonparametric extension of GDM and discuss geometric treatments for some of the LDA extensions; (v) finally we provide a thorough evaluation of our method against a Gibbs sampler, a variational algorithm, and the RecoverKL algorithm. Our method is shown to be comparable to a Gibbs sampler in terms of estimation accuracy, but much more efficient in runtime. It outperforms RecoverKL algorithm in terms of accuracy, in some realistic settings of simulations and in real data.

The chapter proceeds as follows. Section 2.2 provides a brief background of the LDA and its convex geometric formulation. Section 2.3 carries out the contributions outlined above. Section 2.4 presents experiments results. We conclude with a discussion in Section

2.5.

2.2 Background on topic models

In this section we give an overview of the well-known Latent Dirichlet Allocation model for topic modeling (Blei et al., 2003), and the geometry it entails. Let $\alpha \in \mathbb{R}_+^K$ and $\eta \in \mathbb{R}_+^V$ be hyperparameters, where V denotes the number of words in a vocabulary, and K the number of topics. The K topics are represented as distributions on words: $\beta_k|\eta \sim \text{Dir}_V(\eta)$, for $k = 1, \ldots, K$. Each of the M documents can be generated as follows. First, draw the document topic proportions: $\theta_m|\alpha \sim \text{Dir}_K(\alpha)$, for $m = 1, \ldots, M$. Next, for each of the N_m words in document m, pick a topic label z and then sample a word d from the chosen topic:

$$z_{n_m}|\theta_m \sim \text{Categorical}(\theta_m); \; d_{n_m}|z_{n_m}, \beta_{1\ldots K} \sim \text{Categorical}(\beta_{z_{n_m}}). \tag{2.1}$$

Each of the resulting documents is a vector of length N_m with entries $d_{n_m} \in \{1, \ldots, V\}$, where $n_m = 1, \ldots, N_m$. Because these words are exchangeable by the modeling, they are equivalently represented as a vector of word counts $w_m \in \mathbb{N}^V$. In practice, the Dirichlet distributions are often simplified to be symmetric Dirichlet, in which case hyperparameters $\alpha, \eta \in \mathbb{R}_+$ and we will proceed with this setting.

Two most common approaches for inference with the LDA are Gibbs sampling (Griffiths & Steyvers, 2004), based on the Multinomial-Dirichlet conjugacy, and mean-field inference (Blei et al., 2003). The former approach produces more accurate estimates but is less computationally efficient than the latter. The inefficiency of both techniques can be traced to the need for sampling or estimating the (redundant) topic labels. These labels are not intrinsic — any permutation of the topic labels yield the same likelihood function.

Convex geometry of topics. By integrating out the latent variables that represent the topic labels, we obtain a geometric formulation of the LDA. Indeed, integrating z's out yields that, for $m = 1, \ldots, M$,

$$w_m|\theta_m, \beta_{1\ldots K}, N_m \sim \text{Multinomial}(p_{m1}, \ldots, p_{mV}, N_m),$$

where p_{mi} denotes probability of observing the i-th word from the vocabulary in the m-th document, and is given by

$$p_{mi} = \sum_{k=1}^{K} \theta_{mk}\beta_{ki} \text{ for } i = 1, \ldots, V;\ m = 1, \ldots, M. \tag{2.2}$$

The model's geometry becomes clear. Each topic is represented by a point β_k lying in the $V-1$ dimensional probability simplex Δ^{V-1}. Let $B := \mathrm{Conv}(\beta_1, \ldots, \beta_K)$ be the convex hull of the K topics β_k, then each document corresponds to a point $p_m := (p_{m1}, \ldots, p_{mV})$ lying inside the polytope B. This point of view has been proposed before (Hofmann, 1999), although topic proportions θ were not given any geometric meaning. The following treatment of θ lets us relate to the LDA's Dirichlet prior assumption and complete the geometric perspective of the problem. The Dirichlet distribution generates probability vectors θ_m, which can be viewed as the (random) *barycentric coordinates* of the document m with respect to the polytope B. Each $p_m = \sum_k \theta_{mk}\beta_k$ is a vector of Cartesian coordinates of the m-th document's multinomial probabilities. Given p_m, document m is generated by taking $w_m \sim \text{Multinomial}(p_m, N_m)$. In Section 2.4 we will show how this interpretation of topic proportions can be utilized by other topic modeling approaches, including for example the RecoverKL algorithm of Arora et al. (2013). In the following the model geometry is exploited to derive fast and effective geometric algorithm for inference and parameter estimation.

2.3 Geometric inference of topics

We shall introduce a geometric loss function that can be viewed as a surrogate to the LDA's likelihood. To begin, let $\boldsymbol{\beta}$ denote the $K \times V$ topic matrix with rows β_k, $\boldsymbol{\theta}$ be a $M \times K$ document topic proportions matrix with rows θ_m, and $\overline{W}$ be $M \times V$ normalized word counts matrix with rows $\bar{w}_m = w_m / N_m$.

2.3.1 Geometric surrogate loss to the likelihood

Unlike the original LDA formulation, here the Dirichlet distribution on $\boldsymbol{\theta}$ can be viewed as a prior on parameters $\boldsymbol{\theta}$. The log-likelihood of the observed corpora of M documents is

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_{m=1}^{M} \sum_{i=1}^{V} w_{mi} \log \left(\sum_{k=1}^{K} \theta_{mk}\beta_{ki} \right),$$

where the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are subject to constraints $\sum_i \beta_{ki} = 1$ for each $k = 1, \ldots, K$, and $\sum_k \theta_{mk} = 1$ for each $m = 1, \ldots, M$. Partially relaxing these constraints and keeping only the one that the sum of all entries for each row of the matrix product $\boldsymbol{\theta}\boldsymbol{\beta}$ is 1, yields the upper bound that $L(\boldsymbol{\theta}, \boldsymbol{\beta}) \leq L(\overline{W})$, where function $L(\overline{W})$ is given by

$$L(\overline{W}) = \sum_m \sum_i w_{mi} \log \bar{w}_{mi}.$$

We can establish a tighter bound, which will prove useful (the proof of this and other technical results are in the Supplement):

Proposition 2.1. Given a fixed topic polytope B and $\boldsymbol{\theta}$. Let U_m be the set of words present in document m, and assume that $p_{mi} > 0 \; \forall \; i \in U_m$, then

$$L(\overline{W}) - \frac{1}{2} \sum_{m=1}^{M} N_m \sum_{i \in U_m} (\bar{w}_{mi} - p_{mi})^2 \geq L(\boldsymbol{\theta}, \boldsymbol{\beta}) \geq L(\overline{W}) - \sum_{m=1}^{M} N_m \sum_{i \in U_m} \frac{1}{p_{mi}} (\bar{w}_{mi} - p_{mi})^2.$$

Since $L(\overline{W})$ is constant, the proposition above shows that maximizing the likelihood has the effect of minimizing the following quantity with respect to both $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$:

$$\sum_m N_m \sum_i (\bar{w}_{mi} - p_{mi})^2.$$

For each fixed $\boldsymbol{\beta}$ (and thus B), minimizing first with respect to $\boldsymbol{\theta}$ leads to the following

$$G(B) \;\; := \;\; \min_{\boldsymbol{\theta}} \sum_m N_m \sum_i (\bar{w}_{mi} - p_{mi})^2 = \sum_{m=1}^{M} N_m \min_{x : x \in B} \|x - \bar{w}_m\|_2^2, \qquad (2.3)$$

where the second equality in the above display is due $p_m = \sum_k \theta_{mk} \beta_k \in B$.

The proposition suggests a strategy for parameter estimation: $\boldsymbol{\beta}$ (and B) can be estimated by minimizing the geometric loss function G:

$$\min_B G(B) = \min_B \sum_{m=1}^{M} N_m \min_{x : x \in B} \|x - \bar{w}_m\|_2^2. \qquad (2.4)$$

In words, we aim to find a convex polytope $B \in \Delta^{V-1}$, which is closest to the normalized word counts $\bar{w}_m$ of the observed documents. It is interesting to note the presence of document length N_m, which provides the weight for the squared ℓ_2 error for each document. Thus, our loss function adapts to the varying length of documents in the collection. Without the weights, our objective is similar to the sum of squared errors of the Nonnegative

Matrix Factorization(NMF). Ding et al. (2006) studied the relation between the likelihood function of interest and NMF, but with a different objective of the NMF problem and without geometric considerations. Once $\hat{B}$ is solved, $\hat{\boldsymbol{\theta}}$ can be obtained as the barycentric coordinates of the projection of $\bar{w}_m$ onto $\hat{B}$ for each document $m = 1, \ldots, M$ (cf. Eq (2.3)). We note that if $K \leq V$, then B is a simplex and $\beta_1, \ldots, \beta_k$ in general positions are the extreme points of B, and the barycentric coordinates are unique. (If $K > V$, the uniqueness no longer holds).

Finally, $\hat{p}_m = \hat{\theta}_m^T \hat{\boldsymbol{\beta}}$ gives the Cartesian coordinates of a point in B that minimizes Euclidean distance to the maximum likelihood estimate: $\hat{p}_m = \underset{x \in B}{\operatorname{argmin}} \|x - \bar{w}_m\|_2$. This projection is not available in the closed form, but a fast algorithm is available (Golubitsky et al., 2012), which can easily be extended to find the corresponding distance and to evaluate our geometric objective.

2.3.2 Geometric Dirichlet Means algorithm

We proceed to devise a procedure for approximately solving the topic polytope B via Eq. (2.4): first, obtain an estimate of the underlying subspace based on weighted k-means clustering and then, estimate the vertices of the polytope that lie on the subspace just obtained via a geometric correction technique. Please refer to the Supplement for a clarification of the concrete connection between our geometric loss function and other objectives which arise in subspace learning and weighted k-means clustering literature, the connection that motivates the first step of our algorithm.

Geometric Dirichlet Means (GDM) algorithm estimates a topic polytope B based on the training documents (see Algorithm 2.1).

The algorithm is conceptually simple, and consists of two main steps: First, we perform a (weighted) k-means clustering on the M points $\bar{w}_1, \ldots, \bar{w}_M$ to obtain the K centroids $\mu_1, \ldots, \mu_K$, and second, construct a ray emanating from a (weighted) center of the polytope and extending through each of the centroids μ_k until it intersects with a sphere of radius R_k or with the simplex Δ^{V-1} (whichever comes first). The intersection point will be our estimate for vertices β_k, $k = 1, \ldots, K$ of the polytope B. The center C of the sphere is given in step 1 of the algorithm, while $R_k = \max_{1 \leq m \leq M} \|C - \bar{w}_m\|_2$, where the maximum is taken over those documents m that are clustered with label k.

To see the intuition behind the algorithm, let us consider a simple simulation experiment. We use the LDA data generative model with $\alpha = 0.1$, $\eta = 0.1$, $V = 5$, $K = 4$, $M = 5000$, $N_m = 100$.

Algorithm 2.1 Geometric Dirichlet Means (GDM)

Input: documents $w_1, \ldots, w_M$, K,
extension scalar parameters $m_1, \ldots, m_K$
Output: topics $\beta_1, \ldots, \beta_K$

$C = \frac{1}{M} \sum_m \bar{w}_m$ {find center of the data}
$\mu_1, \ldots, \mu_K$ = weighted k-means$(\bar{w}_1, \ldots, \bar{w}_M, K)$ {find centers of K clusters}.
for all $k = 1, \ldots, K$ **do**
  $\beta_k = C + m_k (\mu_k - C)$.
  if any $\beta_{ki} < 0$ **then** {threshold topic if it is outside vocabulary simplex Δ^{V-1}}
    for all $i = 1, \ldots, V$ **do**
      $\beta_{ki} = \frac{\beta_{ik} \mathbb{1}_{\beta_{ki}>0}}{\sum_i \beta_{ki} \mathbb{1}_{\beta_{ki}>0}}$.
    end for
  end if
end for
$\beta_1, \ldots, \beta_K$.

Multidimensional scaling is used for visualization (Fig. 2.1). We observe that the k-means centroids (pink) do not represent the topics very well, but our geometric modification finds extreme points of the tetrahedron: red and yellow spheres overlap, meaning we found the true topics. In this example, we have used a very small vocabulary size, but in practice V is much higher and the cluster centroids are often on the boundary of the vocabulary simplex, therefore we have to threshold the betas at 0. Extending length until R_k is our default choice for the extension parameters:

$$m_k = \frac{R_k}{\|C - \mu_k\|_2} \text{ for } k = 1, \ldots, K, \tag{2.5}$$

but we will see in our experiments that a careful tuning of the extension parameters based on optimizing the geometric objective (2.4) over a small range of m_k helps to improve the performance considerably. We call this **tGDM** algorithm (tuning details are presented in the Supplement). The connection between extension parameters and the thresholding is the following: if the cluster centroid assigns probability to a word smaller than the whole data does on average, this word will be excluded from topic k with large enough m_k. Therefore, the extension parameters can as well be used to control for the sparsity of the inferred topics.

2.3.3 Consistency of Geometric Dirichlet Means

We shall present a theorem which provides a theoretical justification for the Geometric Dirichlet Means algorithm. In particular, we will show that the algorithm can achieve consistent estimates of the topic polytope, under conditions on the parameters of the Dirichlet

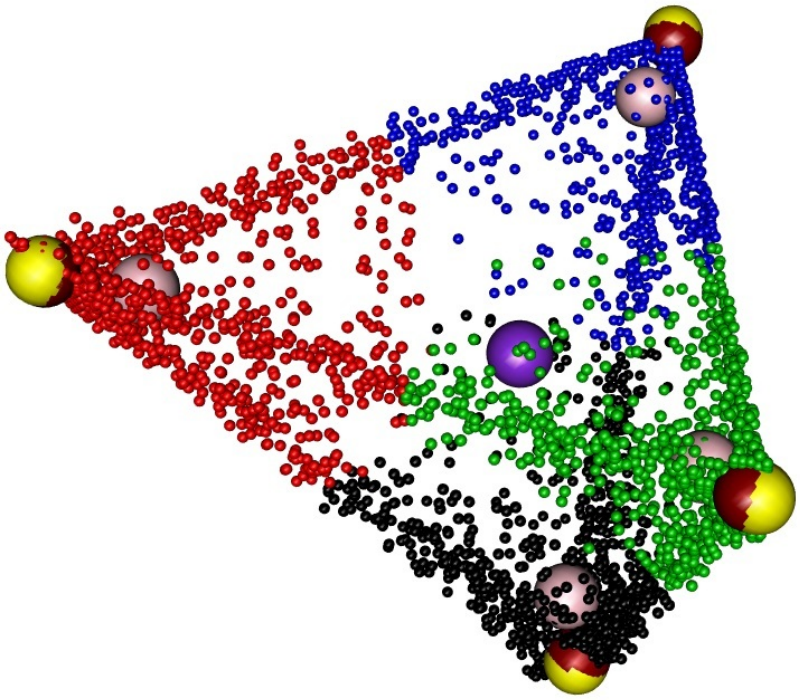


Figure 2.1: Visualization of GDM: Black, green, red and blue are cluster assignments; purple is the center, pink are cluster centroids, dark red are estimated topics and yellow are the true topics.

distribution of the topic proportion vector θ_m, along with conditions on the geometry of the convex polytope B. The problem of estimating vertices of a convex polytope given data drawn from the interior of the polytope has long been a subject of convex geometry — the usual setting in this literature is to assume the uniform distribution for the data sample. Our setting is somewhat more general — the distribution of the points inside the polytope will be driven by a symmetric Dirichlet distribution setting, i.e., $\theta_m \overset{iid}{\sim} \text{Dir}_K(\alpha)$. (If $\alpha = 1$ this results in the uniform distribution on B.) Let $n = K - 1$. Assume that the document multinomial parameters $p_1, \ldots, p_M$ (given in Eq. (2.2)) are the actual data. Now we formulate a geometric problem linking the population version of k-means and polytope estimation:

Problem 2.1. Given a convex polytope $A \in \mathbb{R}^n$, a continuous probability density function $f(x)$ supported by A, find a K-partition $A = \bigsqcup_{k=1}^{K} A_k$ that minimizes:

$$\sum_{k}^{K} \int_{A_k} \|\mu_k - x\|_2^2 f(x)\, dx,$$

where μ_k is the center of mass of A_k: $\mu_k := \frac{1}{\int_{A_k} f(x)\,dx} \int_{A_k} x f(x)\,dx$.

This problem is closely related to the Centroidal Voronoi Tessellations (Du et al., 1999). This connection can be exploited to show that

Lemma 2.1. Problem 2.1 has a unique global minimizer.

In the following lemma, a median of a simplex is a line segment joining a vertex of a simplex with the centroid of the opposite face.

Lemma 2.2. If $A \in \mathbb{R}^n$ is an equilateral simplex with symmetric Dirichlet density f parameterized by α, then the optimal centers of mass of the Problem 2.1 lie on the corresponding medians of A.

Based upon these two lemmas, consistency is established under two distinct asymptotic regimes.

Theorem 2.1. Let $B = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$ be the true convex polytope from which the M-sample $p_1, \ldots, p_M \in \Delta^{V-1}$ are drawn via Eq. (2.2), where $\theta_m \overset{iid}{\sim} \mathrm{Dir}_K(\alpha)$ for $m = 1, \ldots, M$.

(a) If B is also an equilateral simplex, then topic estimates obtained by the GDM algorithm using the extension parameters given in Eq. (2.5) converge to the vertices of B in probability, as α is fixed and $M \to \infty$.

(b) If M is fixed, while $\alpha \to 0$ then the topic estimates obtained by the GDM also converge to the vertices of B in probability.

2.3.4 nGDM: nonparametric geometric inference of topics

In practice, the number of topics K may be unknown, necessitating a nonparametric probabilistic approach such as the well-known Hierarchical Dirichlet Process (HDP) (Teh et al., 2006). Our geometric approach can be easily extended to this situation. The objective (2.4) is now given by

$$\min_B G(B) = \min_B \sum_{m=1}^{M} N_m \min_{x \in B} \|x - \bar{w}_m\|_2^2 + \lambda |B|, \tag{2.6}$$

where $|B|$ denotes the number of extreme points of convex polytope $B = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$. Accordingly, our nGDM algorithm now consists of two steps: (i) solve a penalized and weighted k-means clustering to obtain the cluster centroids (e.g. using DP-means (Kulis

& Jordan, 2012)); (ii) apply geometric correction for recovering the extreme points, which proceeds as before.

Our theoretical analysis can be also extended to this nonparametric framework. We note that the penalty term is reminiscent of the DP-means algorithm of Kulis & Jordan (2012), which was derived under a small-variance asymptotics regime. For the HDP this corresponds to $\alpha \to 0$ — the regime in part (b) of Theorem 2.1. This is an unrealistic assumption in practice. Our geometric correction arguably enables the accounting of the non-vanishing variance in data. We perform a simulation experiment for varying values of α and show that nGDM outperforms the KL version of DP-means (Jiang et al., 2012) in terms of perplexity. This result is reported in the Supplement.

2.4 Performance evaluation

Simulation experiments We use the LDA model to simulate data and focus our attention on the perplexity of held-out data and minimum-matching Euclidean distance between the true and estimated topics (Tang et al., 2014). We explore settings with varying document lengths (N_m increasing from 10 to 1400 - Fig. 2.2(a) and Fig. 2.3(a)), different number of documents (M increasing from 100 to 7000 - Fig. 2.2(b) and Fig. 2.3(b)) and when lengths of documents are small, while number of documents is large ($N_m = 50$, M ranging from 1000 to 15000 - Fig. 2.2(c) and Fig. 2.3(c)). This last setting is of particular interest, since it is the most challenging for our algorithm, which in theory works well given long documents, but this is not always the case in practice. We compare two versions of the Geometric Dirichlet Means algorithm: with tuned extension parameters (tGDM) and the default one (GDM) (cf. Eq. (2.5)) against the **variational EM** (VEM) algorithm (Blei et al., 2003) (with tuned hyperparameters), **collapsed Gibbs sampling** (Griffiths & Steyvers, 2004) (with true data generating hyperparameters), and **RecoverKL** (Arora et al., 2013) and verify the theoretical upper bounds for topic polytope estimation (i.e. either $(\log M/M)^{0.5}$ or $(\log N_m/N_m)^{0.5}$) - cf. Tang et al. (2014) and Nguyen (2015). We are also interested in estimating each document's topic proportion via the projection technique. RecoverKL produced only a topic matrix, which is combined with our projection based estimates to compute the perplexity (Fig. 2.3). Unless otherwise specified, we set $\eta = 0.1$, $\alpha = 0.1$, $V = 1200$, $M = 1000$, $K = 5$; $N_m = 1000$ for each m; the number of held-out documents is 100; results are averaged over 5 repetitions. Since finding exact solution to the k-means objective is NP hard, we use the algorithm of Hartigan & Wong (1979) with 10 restarts and the k-means++ initialization.

Our results show that (i) Gibbs sampling and tGDM have the best and almost identical

performance in terms of statistical estimation; (ii) RecoverKL and GDM are the fastest while sharing comparable statistical accuracy; (iii) VEM is the worst in most scenarios due to its instability (i.e. often producing poor topic estimates); (iv) short document lengths (Fig. 2.2(c) and Fig. 2.3(c)) do not degrade performance of GDM, (this appears to be an effect of the law of large numbers, as the algorithm relies on the cluster means, which are obtained by averaging over a large number of documents); (v) our procedure for estimating document topic proportions results in a good quality perplexity of the RecoverKL algorithm in all scenarios (Fig. 2.3) and could be potentially utilized by other algorithms. Additional simulation experiments are presented in the Supplement, which considers settings with varying N_m, α and the nonparametric extension.

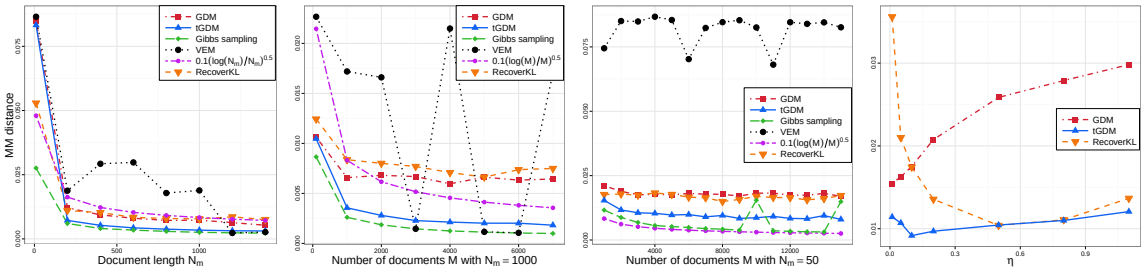


Figure 2.2: Minimum-matching Euclidean distance: increasing N_m, $M = 1000$ (a); increasing M, $N_m = 1000$ (b); increasing M, $N_m = 50$ (c); increasing η, $N_m = 50$, $M = 5000$ (d).

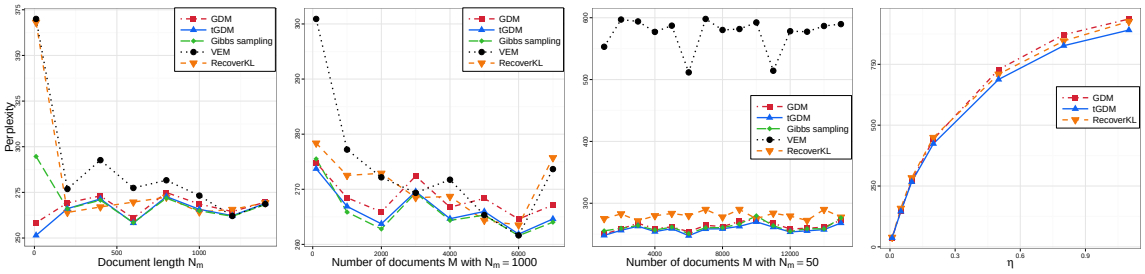


Figure 2.3: Perplexity of the held-out data: increasing N_m, $M = 1000$ (a); increasing M, $N_m = 1000$ (b); increasing M, $N_m = 50$ (c); increasing η, $N_m = 50$, $M = 5000$ (d).

Comparison to RecoverKL Both tGDM and RecoverKL exploit the geometry of the model, but they rely on very different assumptions: RecoverKL requires the presence of anchor words in the topics and exploits this in a crucial way (Arora et al., 2013); our method relies on long documents in theory, even though the violation of this does not

appear to degrade its performance in practice, as we have shown earlier. The comparisons are performed by varying the document length N_m, and varying the Dirichlet parameter η (recall that $\beta_k|\eta \sim \text{Dir}_V(\eta)$). In terms of perplexity, RecoverKL, GDM and tGDM perform similarly (see Fig.2.4(c,d)), with a slight edge to tGDM. Pronounced differences come in the quality of topic's word distribution estimates. To give RecoverKL the advantage, we considered manually inserting anchor words for each topic generated, while keeping the document length short, $N_m = 50$ (Fig. 2.4(a,c)). We found that tGDM outperforms RecoverKL when $\eta \leq 0.3$, an arguably more common setting, while RecoverKL is more accurate when $\eta \geq 0.5$. However, if the presence of anchor words is not explicitly enforced, tGDM always outperforms RecoverKL in terms of topic distribution estimation accuracy for all η (Fig. 2.2(d)). The superiority of tGDM persists even as N_m varies from 50 to 10000 (Fig. 2.4(b)), while GDM is comparable to RecoverKL in this setting.

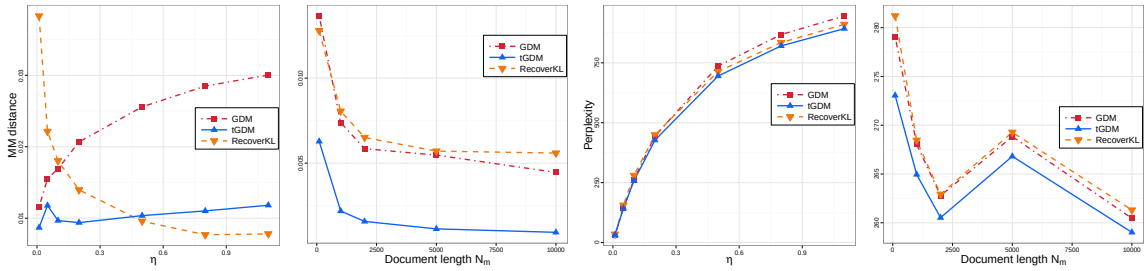


Figure 2.4: MM distance and Perplexity for varying η, $N_m = 50$ with anchors (a,c); varying N_m (b,d).

NIPS corpora analysis We proceed with the analysis of the NIPS corpus.[3] After preprocessing, there are 1738 documents and 4188 unique words. Length of documents ranges from 39 to 1403 with mean of 272. We consider $K = 5, 10, 15, 20$, $\alpha = \frac{5}{K}$, $\eta = 0.1$. For each value of K we set aside 300 documents chosen at random to compute the perplexity and average results over 3 repetitions. Our results are compared against Gibbs sampling, Variational EM and RecoverKL (Table 2.1). For $K = 10$, GDM with 1500 k-means iterations and 5 restarts in R took 50sec; Gibbs sampling with 5000 iterations took 10.5min; VEM with 750 variational, 1500 EM iterations and 3 restarts took 25.2min; RecoverKL coded in Python took 1.1min. We note that with recent developments (e.g., see Hoffman et al. (2013)) VEM could be made faster, but its statistical accuracy remains poor. Although RecoverKL is as fast as GDM, its perplexity performance is poor and is getting worse with more topics, which we believe could be due to lack of anchor words in the data. We present

[3] https://archive.ics.uci.edu/ml/datasets/Bag+of+Words

topics found by Gibbs sampling, GDM and RecoverKL for $K = 10$ in the Supplement.

Table 2.1: Perplexities of the 4 topic modeling algorithms trained on the NIPS dataset.

	GDM	RecoverKL	VEM	Gibbs sampling
$K = 5$	1269	1378	1980	1168
$K = 10$	1061	1235	1953	924
$K = 15$	957	1409	1545	802
$K = 20$	763	1586	1352	704

2.5 Discussion

We wish to highlight a conceptual aspect of GDM distinguishing it from moment-based methods such as RecoverKL. GDM operates on the document-to-document distance/similarity matrix, as opposed to the second-order word-to-word matrix. So, from an optimization viewpoint, our method can be viewed as the dual to RecoverKL method, which must require anchor-word assumption to be computationally feasible and theoretically justifiable. While the computational complexity of RecoverKL grows with the vocabulary size and not the corpora size, our convex geometric approach continues to be computationally feasible when number of documents is large: since only documents near the polytope boundary are relevant in the inference of the extreme points, we can discard most documents residing near the polytope's center.

We discuss some potential improvements and extensions next. The tGDM algorithm showed a superior performance when the extension parameters are optimized. This procedure, while computationally effective relative to methods such as Gibbs sampler, may still be not scalable to massive datasets. It seems possible to reformulate the geometric objective as a function of extension parameters, whose optimization can be performed more efficiently. In terms of theory, we would like to establish the error bounds by exploiting the connection of topic inference to the geometric problem of Centroidal Voronoi Tessellation of a convex polytope.

The geometric approach to topic modeling and inference may lend itself naturally to other LDA extensions, as we have demonstrated with nGDM algorithm for the HDP (Teh et al., 2006). Correlated topic models of Blei & Lafferty (2006a) also fit naturally into the geometric framework — we would need to adjust geometric modification to capture logistic normal distribution of topic proportions inside the topic polytope. Another interesting direction is to consider dynamic (Blei & Lafferty, 2006b) (extreme points of topic polytope

evolving over time) and supervised (Mcauliffe & Blei, 2008) settings. Such settings appear relatively more challenging, but they are worth pursuing further.

Appendix

2.A Proof of Proposition 2.1

Proof. Consider the KL divergence between two distributions parameterized by $\bar{w}_m$ and p_m, respectively:

$$\begin{aligned} D(P_{\bar{w}_m} \| P_{p_m}) &= \sum_{i \in U_m} \bar{w}_{mi} \log \frac{\bar{w}_{mi}}{p_{mi}} \\ &= \frac{1}{N_m} \left(\sum_{i \in U_m} w_{mi} \log \bar{w}_{mi} - \sum_{i \in U_m} w_{mi} \log p_{mi} \right). \end{aligned}$$

Then $L(\overline{W}) - L(\boldsymbol{\theta}, \boldsymbol{\beta}) = \sum_m N_m D(P_{\bar{w}_m} \| P_{p_m}) \geq 0$, due to the non-negativity of KL divergence. Now we shall appeal to a standard lower bound for the KL divergence (Cover & Thomas, 2006):

$$D(P_{\bar{w}_m} \| P_{p_m}) \geq \frac{1}{2} \sum_{i \in U_m} (\bar{w}_{mi} - p_{mi})^2,$$

and an upper bound via χ^2-distance (e.g. see Sayyareh (2011)):

$$D(P_{\bar{w}_m} \| P_{p_m}) \leq \sum_{i \in U_m} \frac{1}{p_{mi}} (\bar{w}_{mi} - p_{mi})^2.$$

Taking summation of both bounds over $m = 1, \ldots, M$ concludes the proof. □

2.B Connection between our geometric loss function and other objectives which arise in subspace learning and k-means clustering problems.

Recall that our geometric objective (Eq. (2.4)) is:

$$\min_B G(B) = \min_B \sum_{m=1}^{M} N_m \min_{x : x \in B} \|x - \bar{w}_m\|_2^2.$$

We note that this optimization problem can be reduced to two other well-known problems when the objective function and constraints are suitably relaxed/modified:

- A version of weighted low-rank matrix approximation is $\min_{\text{rank}(\hat{D})\leq r} \text{tr}((\hat{D}-D)^T Q(\hat{D}-D))$. If $Q = \text{diag}(N_1, \ldots, N_M)$, $D = \overline{W}$, $r = K$ and $\hat{D} = \boldsymbol{\theta}\boldsymbol{\beta}$, the problem looks similar to the geometric objective without constraints and has a closed form solution (Manton et al., 2003): $\hat{D} = Q^{-1/2} U \Sigma_K V^T$, where

$$Q^{1/2} D = U \Sigma_K V^T \tag{2.7}$$

  is the singular value decomposition and Σ_K is the truncation to K biggest singular values. Also note that here and further without loss of generality we assume $M \geq V$, if $M < V$ for the proofs to hold we replace $Q^{1/2}D$ with $(Q^{1/2}D)^T$.

- The k-means algorithm involves optimizing the objective (Hartigan & Wong, 1979; Lloyd, 1982; MacQueen, 1967): $\min_{x_1,\ldots,x_K} \sum_m \min_{i\in\{1,\ldots,K\}} \|\bar{w}_m - x_i\|_2^2$. Our geometric objective (2.4) is quite similar — it replaces the second minimization with minimizing over the convex hull of $\{x_1, \ldots, x_K\}$ and includes weight N_ms.

- The two problems described above are connected in the following way (Xu et al., 2003). Define the weighted k-means objective with respect to cluster assignments: $\sum_k \sum_{m\in C_k} N_m \|\bar{w}_m - \mu_k\|^2$, where μ_k is the centroid of the k-th cluster:

$$\mu_k = \frac{\sum_{m\in C_k} N_m \bar{w}_m}{\sum_{m\in C_k} N_m}. \tag{2.8}$$

  Let S_k be the optimal indicator vector of cluster k, i.e., m-th element is 1 if $m \in C_k$ and 0 otherwise. Define

$$Y_k = \frac{Q^{1/2} S_k}{\|Q^{1/2} S_k\|_F^2}. \tag{2.9}$$

  If we relax the constraint on S_k to allow any real values instead of only binary values, then Y can be solved via the following eigenproblem: $Q^{1/2}\overline{W}\overline{W}^T Q^{1/2} Y = \lambda Y$.

Let us summarize the above observations by the following:

Proposition 2.2. Given the $M \times V$ normalized word counts matrix $\overline{W}$. Let $\mu_1, \ldots, \mu_K$ be the optimal cluster centroids of the weighted k-means problem given by Eq. (2.8), and let v_ks be the columns of V in the SVD of Eq. (2.7). Then,

$$\text{span}(\mu_1, \ldots, \mu_K) = \text{span}(v_1, \ldots, v_K).$$

Proof. Following Ding & He (2004), let P_c be an operator projecting any vector onto $\mathrm{span}(\mu_1, \ldots, \mu_K)$: $P_c = \sum_k \mu_k \mu_k^T$. Recall that S_k is the indicator vector of cluster k and Y_k defined in Eq. (2.9). Then

$$\mu_k = \frac{\overline{W}^T Q S_k}{\|Q^{1/2} S_k\|_F^2} = \overline{W}^T Q^{1/2} Y_k, \text{ and}$$
$$P_c = \sum_k \overline{W}^T Q^{1/2} Y_k (\overline{W}^T Q^{1/2} Y_k)^T.$$

Now, note that Y_k's are the eigenvectors of $Q^{1/2}\overline{W}\overline{W}^T Q^{1/2}$, which are also left-singular vectors of $Q^{1/2}\overline{W} = U\Sigma V^T$, so

$$P_c = (Q^{1/2}\overline{W})^T Y_k ((Q^{1/2}\overline{W})^T Y_k)^T = \sum_k \lambda_k^2 v_k v_k^T,$$

which is the projection operator for $\mathrm{span}(v_1, \ldots, v_K)$. Hence, the two subspaces are equal. □

Prop. 2.2 and the preceding discussions motivate the GDM algorithm for estimating the topic polytope: first, obtain an estimate of the underlying subspace based on k-means clustering and then, estimate the vertices of the polytope that lie on the subspace just obtained.

2.C Proofs of technical lemmas

Recall Problem 2.1 from the main part:

Problem 2.1. *Given a convex polytope $A \in \mathbb{R}^n$, a continuous probability density function $f(x)$ supported by A, find a K-partition $A = \bigsqcup_{k=1}^{K} A_k$ that minimizes:*

$$\sum_k^K \int_{A_k} \|\mu_k - x\|_2^2 f(x)\, dx,$$

where μ_k is the center of mass of A_k: $\mu_k := \frac{1}{\int_{A_k} f(x)\, dx} \int_{A_k} x f(x)\, dx$.

Proof of Lemma 2.1

Proof. The proof follows from a sequence of results of Du et al. (1999), which we now summarize. First, if the K-partition $(A_1, \ldots, A_K)$ is a minimizer of Problem 2.1, then A_ks

are the Voronoi regions corresponding to the μ_ks. Second, Problem 2.1 can be restated in terms of the μ_ks to minimize $\mathcal{K}(\mu_1, \ldots, \mu_K) = \sum_k \int_{\hat{A}_k} \|\mu_k - x\|_2^2 f(x)\, dx$, where $\hat{A}_k$s are the Voronoi regions corresponding to their centers of mass μ_ks. Third, $\mathcal{K}(\mu_1, \ldots, \mu_K)$ is a continuous function and admits a global minimum. Fourth, the global minimum is unique if the distance function in $\mathcal{K}$ is strictly convex and the Voronoi regions are convex. Now, it can be verified that the squared Euclidean distance is strictly convex. Moreover, Voronoi regions are intersections of half-spaces with the convex polytope A, which can also be represented as an intersection of half-spaces. Therefore, the Voronoi regions of Problem 2.1 are convex polytopes, and it follows that the global minimizer is unique. □

Proof of Lemma 2.2

Proof. Since f is a symmetric Dirichlet density, the center of mass of A coincides with its centroid. Let $n = 3$. In an equilateral triangle, the centers of mass μ_1, μ_2, μ_3 form an equilateral triangle C. An intersection point of the Voronoi regions A_1, A_2, A_3 is the circumcenter and the centroid of C, which is also a circumcenter and centroid of A. Therefore, μ_1, μ_2, μ_3 are located on the medians of A with exact positions depending on the α. The symmetry and the property of circumcenter coinciding with centroid carry over to the general n-dimensional equilateral simplex (Westendorp, 2013). □

2.D Proof of Theorem 2.1

Proof. For part (a), let $(\hat{\mu}_1, \ldots, \hat{\mu}_K)$ be the minimizer of the k-means problem

$$\min_{\mu_1,\ldots,\mu_K} \sum_m \min_{i\in\{1,\ldots,K\}} \|p_m - \mu_i\|_2^2.$$

Let $\tilde{\mu}_1, \ldots, \tilde{\mu}_K$ be the centers of mass of the solution of Problem 2.1 applied to B and the Dirichlet density. By Lemma 1, these centers of mass are unique, as they correspond to the unique optimal K-partition. Accordingly, by the strong consistency of k-means clustering under the uniqueness condition (Pollard, 1981), as $M \to \infty$,

$$\text{Conv}(\hat{\mu}_1, \ldots, \hat{\mu}_K) \to \text{Conv}(\tilde{\mu}_1, \ldots, \tilde{\mu}_K) \text{ a.s.},$$

where the convergence is assessed in either Hausdorff or the minimum matching distance for convex sets (Nguyen, 2015). Note that $C = \frac{1}{M}\sum_m p_m$ is a strongly consistent estimate of the centroid C_0 of B, by the strong law of large numbers. Lemma 2 shows that $\tilde{\mu}_1, \ldots, \tilde{\mu}_K$

are located on the corresponding medians. To complete the proof, it remains to show that $\hat{R} := \max_{1\le m\le M} \|C - p_m\|_2$ is a weakly consistent estimate of the circumradius R_0 of B. Indeed, for a small $\epsilon > 0$ define the event $E_m^k = \{p_m \in B_\epsilon(\beta_k) \cap B\}$, where $B_\epsilon(\beta_k)$ is an ϵ-ball centering at vertex β_k. Since B is equilateral and the density over it is symmetric and positive everywhere in the domain, $\mathbb{P}(E_m^1) = \ldots = \mathbb{P}(E_m^K) =: b_\epsilon > 0$. Let $E_m = \bigcup_k E_m^k$, then $\mathbb{P}(E_m) = b_\epsilon K$. We have

$$\limsup_{M\to\infty} \mathbb{P}(|\hat{R} - R_0| > 2\epsilon) = \limsup_{M\to\infty} \mathbb{P}(\max_{1\le m\le M} \|C_0 - p_m\|_2 < R_0 - \epsilon) <$$
$$< \limsup_{M\to\infty} \mathbb{P}(\bigcap_{m=1}^{M} E_m^{\complement}) = \limsup_{M\to\infty} (1 - b_\epsilon K)^M = 0.$$

A similar argument allows us to establish that each R_k is also a weakly consistent estimate of R_0. This completes the proof of part (a). For a proof sketch of part (b), for each $\alpha > 0$, let $(\mu_1^\alpha, \ldots, \mu_K^\alpha)$ denote the K means obtained by the k-means clustering algorithm. It suffices to show that these estimates converge to the vertices of B. Suppose this is not the case, due to the compactness of B, there is a subsequence of the K means, as $\alpha \to 0$, that tends to K limit points, some of which are not the vertices of B. It is a standard fact of Dirichlet distributions that as $\alpha \to 0$, the distribution of the p_m converges weakly to the discrete probability measure $\sum_{k=1}^K \frac{1}{K}\delta_{\beta_k}$. So the k-means objective function tends to $\frac{M}{K}\sum_k \min_{i\in\{1,\ldots,K\}} \|\beta_k - \mu_i^\alpha\|_2^2$, which is strictly bounded away from 0, leading to a contradiction. This concludes the proof. □

2.E Tuned GDM

In this section we discuss details of the extension parameters tuning. Recall that GDM requires extension scalar parameters $m_1, \ldots, m_K$ as part of its input. Our default choice (Eq. (2.5)) is

$$m_k = \frac{R_k}{\|C - \mu_k\|_2} \text{ for } k = 1, \ldots, K,$$

where $R_k = \max_{m\in C_k} \|C - \bar{w}_m\|_2$ and C_k is the set of indices of documents belonging to cluster k. In some situations (e.g. outliers making extension parameters too big) tuning of the extension parameters can help to improve the performance, which we called **tGDM**

algorithm. Recall the geometric objective (2.4) and let

$$G_k(B) := \sum_{m \in C_k} N_m \min_{x: x \in B} \|x - \bar{w}_m\|_2^2, \tag{2.10}$$

which is simply the geometric objective evaluated at the documents of cluster k. For each $k = 1, \ldots, K$ we used line search procedure (Brent, 2013) optimization of $G_k(B)$ in an interval from 1 up to default m_k as in (2.5). Independent tuning for each k gives an approximate solution, but helps to reduce the running time.

2.F Additional experiments

Here we present some additional simulation results and NIPS topics.

2.F.1 Nonparametric analysis with nGDM

Based on simulations we show how nGDM can be used when number of topics is unknown and compare it against DP-means utilizing KL divergence (KL DP-means) by Jiang et al. (2012). We analyze settings with α ranging from 0.01 to 2. Recall that KL DP-means assumes $\alpha \to 0$. $V = 1200$, $M = 2000$, $N_m = 3000$, $\eta = 0.1$, true $K = 15$. For each value of α average over 5 repetitions is recorded and we plot the perplexity of 100 held-out documents. Fig. 2.F.1 supports our argument - for small values of α both methods perform equivalently well (KL DP-means due to variance assumption being satisfied and nGDM due to part (b) of Theorem 1), but as α gets bigger, we see how our geometric correction leads to improved performance.

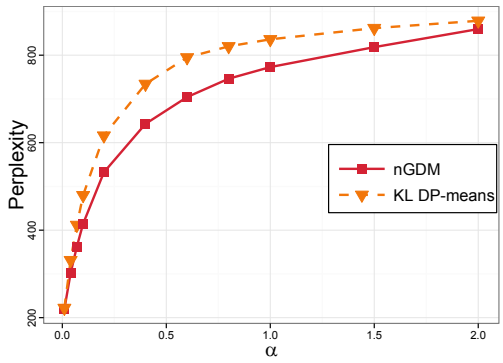


Figure 2.F.1: Perplexity for varying α

2.F.2 Documents of varying size

Until this point all documents are of the same length. Next, we evaluate the improvement of our method when document length varies. The lengths are randomly sampled from 50 to 1500 and the experiment is repeated 20 times. The weighted GDM uses document lengths as weights for computing the data center and training k-means. In both performance measures (Fig. 2.F.2 left and center) the weighted version consistently outperforms the unweighted one, while the tuned weighted version stays very close to Gibbs sampling results.

2.F.3 Effect of the document topic proportions prior

Recall that topic proportions are sampled from the Dirichlet distribution $\theta_m|\alpha \sim \text{Dir}_K(\alpha)$. We let α increase from 0.01 to 2. Smaller α implies that samples are close to the extreme points, and hence GDM estimates topics better. This also follows from Theorem 1(b) of the chapter. We see (Fig. 2.F.2 right) that our solution and Gibbs sampling are almost identical for small α, while VEM is unstable. With increased α Gibbs sampling remains the best, while our algorithm remains better than VEM. We also note that increasing α causes error of all methods to increase.

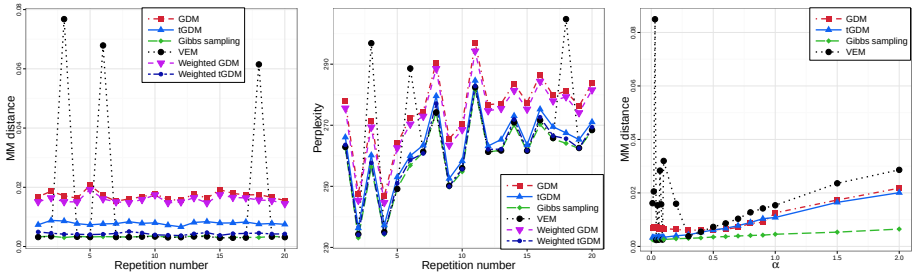


Figure 2.F.2: Minimum-matching Euclidean distance: varying N_m (left); increasing α (right). Perplexity for varying N_m (center).

2.F.4 Projection estimate analysis

Our objective function (2.4) motivates the estimation of document topic proportions by taking the barycentric coordinates of the projection of the normalized word counts of a document onto the topic polytope. To do this we utilized the projection algorithm of Golubitsky et al. (2012). Note that some algorithms (RecoverKL in particular) do not have a built in method for finding topic proportions of the unseen documents. Our projection based estimate can solve this issue, as it can find topic proportions of a document only based on

the topic polytope. Fig. 2.F.3 shows that perplexity with projection estimates closely follows corresponding results and outperforms VEM on the short documents (Fig. 2.F.3 (right)).

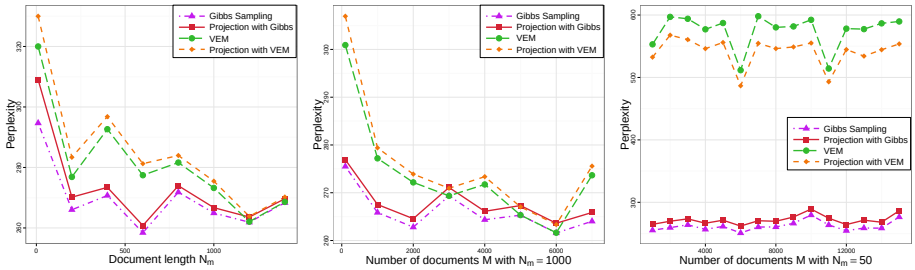


Figure 2.F.3: Projection method: increasing N_m, $M = 1000$ (left); increasing M, $N_m = 1000$ (center); increasing M, $N_m = 50$ (right).

Table 2.F.1: Top 10 words (columns) of each of the 10 learned topics of NIPS dataset

GDM topics									
analog	regress.	reinforc.	nodes	speech	image	mixture	neurons	energy	rules
circuit	kernel	policy	node	word	images	experts	neuron	characters	teacher
memory	bayesian	action	classifier	hmm	object	missing	cells	boltzmann	student
chip	loss	controller	classifiers	markov	visual	mixtures	cell	character	fuzzy
theorem	posterior	actions	tree	phonetic	objects	expert	synaptic	hopfield	symbolic
sources	theorem	qlearning	trees	speaker	face	gating	spike	temperature	saad
polynom.	hyperp.	reward	bayes	acoustic	pixel	posterior	activity	annealing	membership
separation	bounds	sutton	rbf	phoneme	pixels	tresp	firing	kanji	rulebased
recurrent	monte	robot	theorem	hmms	texture	loglikel.	visual	adjoint	overlaps
circuits	carlo	barto	boolean	hybrid	motion	ahmad	cortex	window	children

Gibbs sampler topics									
neurons	rules	mixture	reinforc.	memory	speech	image	analog	theorem	classifier
cells	language	bayesian	policy	energy	word	images	circuit	regress.	nodes
cell	recurrent	posterior	action	neurons	hmm	visual	chip	kernel	node
neuron	node	experts	robot	neuron	auditory	object	voltage	loss	classifiers
activity	tree	entropy	motor	capacity	sound	motion	neuron	bounds	tree
synaptic	memory	mixtures	actions	hopfield	phoneme	objects	vlsi	proof	clustering
firing	nodes	markov	controller	associative	acoustic	spatial	circuits	polynom.	character
spike	symbol	separation	trajectory	recurrent	hmms	face	digital	lemma	rbf
stimulus	symbols	sources	arm	attractor	mlp	pixel	synapse	teacher	cluster
cortex	grammar	principal	reward	boltzmann	segment.	pixels	gate	risk	characters

RecoverKL topics									
entropy	reinforc.	classifier	loss	ensemble	neurons	penalty	mixture	validation	image
image	controller	classifiers	theorem	energy	neuron	rules	missing	regress.	visual
kernel	policy	speech	bounds	posterior	spike	regress.	recurrent	bayesian	motion
energy	action	nodes	proof	bayesian	synaptic	bayesian	bayesian	crossvalid.	cells
ica	actions	word	lemma	speech	cells	energy	posterior	risk	neurons
images	memory	node	polynom.	boltzmann	firing	theorem	image	stopping	images
separation	robot	image	neurons	student	cell	analog	markov	tangent	receptive
clustering	trajectory	tree	regress.	face	activity	regulariz.	speech	image	circuit
sources	sutton	character	nodes	committee	synapses	recurrent	images	kernel	spatial
mixture	feedback	memory	neuron	momentum	stimulus	perturb.	object	regulariz.	object

CHAPTER 3

Conic Scan-and-Cover algorithms for nonparametric topic modeling

We propose new algorithms for topic modeling when the number of topics is unknown. Our approach relies on an analysis of the concentration of mass and angular geometry of the topic simplex, a convex polytope constructed by taking the convex hull of vertices representing the latent topics. Our algorithms are shown in practice to have accuracy comparable to a Gibbs sampler in terms of topic estimation, which requires the number of topics be given. Moreover, they are one of the fastest among several state of the art parametric techniques.[1] Statistical consistency of our estimator is established under some conditions.[2]

3.1 Introduction

A well-known challenge associated with topic modeling inference can be succinctly summed up by the statement that sampling based approaches may be accurate but computationally very slow, e.g., Pritchard et al. (2000); Griffiths & Steyvers (2004), while the variational inference approaches are faster but their estimates may be inaccurate, e.g., Blei et al. (2003); Hoffman et al. (2013). For nonparametric topic inference, i.e., when the number of topics is a priori unknown, the problem becomes more acute. The Hierarchical Dirichlet Process model (Teh et al., 2006) is an elegant Bayesian nonparametric approach which allows for the number of topics to grow with data size, but its sampling based inference is much more inefficient compared to the parametric counterpart. As pointed out in Chapter 2, the root of the inefficiency can be traced to the need for approximating the posterior distributions of the latent variables representing the topic labels — these are not geometrically intrinsic as any permutation of the labels yields the same likelihood.

[1] Code is available at https://github.com/moonfolk/Geometric-Topic-Modeling.

[2] This chapter has been published in Yurochkin et al. (2017a).

A promising approach in addressing the aforementioned challenges is to take a *convex geometric* perspective, where topic learning and inference may be formulated as a convex geometric problem: the observed documents correspond to points randomly drawn from a *topic polytope*, a convex set whose vertices represent the topics to be inferred. This perspective has been adopted to establish posterior contraction behavior of the topic polytope in both theory and practice (Nguyen, 2015; Tang et al., 2014). In Chapter 2 we extended this perspective and exploited the convex geometry to propose the Geometric Dirichlet Means (GDM) algorithm, which demonstrated attractive behaviors both in terms of running time and estimation accuracy. In this chapter we shall continue to amplify this viewpoint to address *nonparametric topic modeling*, a setting in which the number of topics is unknown, as is the distribution inside the topic polytope (in some situations).

We will propose algorithms for topic estimation by explicitly accounting for the concentration of mass and angular geometry of the topic polytope, typically a simplex in topic modeling applications. The geometric intuition is fairly clear: each vertex of the topic simplex can be identified by a ray emanating from its center (to be defined formally), while the concentration of mass can be quantified for the cones hinging on the apex positioned at the center. Such cones can be rotated around the center to scan for high density regions inside the topic simplex — under mild conditions such cones can be constructed efficiently to recover both the number of vertices and their estimates.

We also mention another fruitful approach, which casts topic estimation as a matrix factorization problem (Deerwester et al., 1990; Xu et al., 2003; Anandkumar et al., 2012; Arora et al., 2013). A notable recent algorithm coming from the matrix factorization perspective is RecoverKL (Arora et al., 2013), which solves non-negative matrix factorization (NMF) efficiently under assumptions on the existence of so-called anchor words. RecoverKL remains to be a parametric technique — we will extend it to a nonparametric setting and show that the anchor word assumption appears to limit the number of topics one can efficiently learn.

This chapter is organized as follows. In Section 3.2 we discuss recent developments in geometric topic modeling and introduce our approach; Sections 3.3 and 3.4 deliver the contributions outlined above; Section 3.5 demonstrates experimental performance; we conclude with a discussion in Section 3.6.

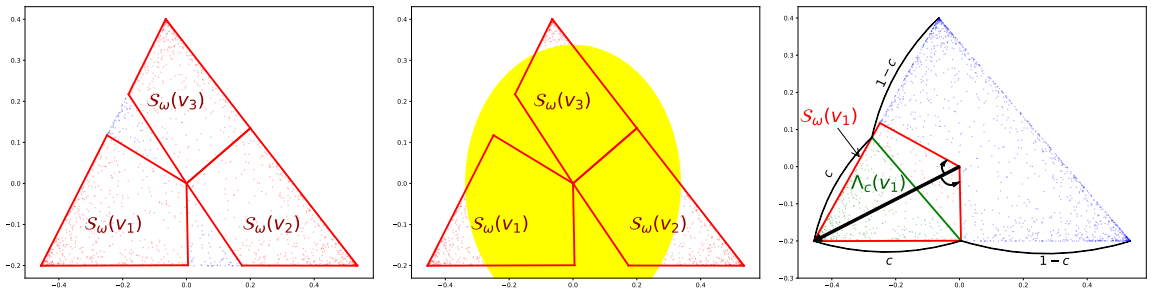
3.2 Geometric topic modeling

Background and related work In this section we present the convex geometry of the Latent Dirichlet Allocation (LDA) model of Blei et al. (2003), along with related theoretical and algorithmic results that motivate our work. Let V be vocabulary size and Δ^{V-1} be

the corresponding vocabulary probability simplex. Sample K topics (i.e., distributions on words) $\beta_k \sim \mathrm{Dir}_V(\eta)$, $k = 1, \ldots, K$, where $\eta \in \mathbb{R}_+^V$. Next, sample M document-word probabilities p_m residing in the *topic simplex* $B := \mathrm{Conv}(\beta_1, \ldots, \beta_K)$ (cf. Nguyen (2015)), by first generating their *barycentric coordinates* (i.e., topic proportions) $\theta_m \sim \mathrm{Dir}_K(\alpha)$ and then setting $p_m := \sum_k \beta_k \theta_{mk}$ for $m = 1, \ldots, M$ and $\alpha \in \mathbb{R}_+^K$. Finally, word counts of the m-th document can be sampled $w_m \sim \mathrm{Mult}(p_m, N_m)$, where $N_m \in \mathbb{N}$ is the number of words in document m. The above model is equivalent to the LDA when individual words to topic label assignments are marginalized out.

Nguyen (2015) established posterior contraction rates of the topic simplex, provided that $\alpha_k \leq 1 \, \forall k$ and *either* number of topics K is known *or* topics are sufficiently separated in terms of the Euclidean distance. In Chapter 2 we devised an estimate for B, taken to be a fixed unknown quantity, by formulating a geometric objective function, which is minimized when topic simplex B is close to the normalized documents $\bar{w}_m := w_m / N_m$. We showed that the estimation of topic proportions θ_m given B simply reduces to taking barycentric coordinates of the projection of $\bar{w}_m$ onto B. To estimate B given K, we proposed a Geometric Dirichlet Means (GDM) algorithm, which operated by performing a k-means clustering on the normalized documents, followed by a geometric correction for the cluster centroids. The resulting algorithm is remarkably fast and accurate, supporting the potential of the geometric approach. The GDM is not applicable when K is unknown, but it provides a motivation which our approach in this chapter is built on.

The Conic Scan-and-Cover approach To enable the inference of B when K is not known, we need to investigate the concentration of mass inside the topic simplex. It suffices to focus on two types of geometric objects: cones and spheres, which provide the basis for a complete coverage of the simplex. To gain intuition of our procedure, which we call Conic Scan-and-Cover (CoSAC) approach, imagine someone standing at a center point of a triangular dark room trying to figure out all corners with a portable flashlight, which can produce a *cone* of light. A room corner can be identified with the direction of the farthest visible data objects. Once a corner is found, one can turn the flashlight to another direction to scan for the next ones. See Fig. 3.1a, where red denotes the scanned area. To make sure that all corners are detected, the cones of light have to be open to an appropriate range of angles so that enough data objects can be captured and removed from the room. To make sure no false corners are declared, we also need a suitable stopping criterion, by relying only on data points that lie beyond a certain spherical radius, see Fig. 3.1b. Hence, we need to be able to gauge the concentration of mass for suitable cones and spherical balls in Δ^{V-1}. This is the subject of the next section.



(a) An incomplete coverage using 3 cones (containing red points).

(b) Complete coverage using 3 cones (red) and a ball (yellow).

(c) Cap $\Lambda_c(v_1)$ and cone $\mathcal{S}_\omega(v_1)$.

Figure 3.1: Complete coverage of topic simplex by cones and a spherical ball for $K = 3$, $V = 3$.

3.3 Geometric estimation of the topic simplex

We start by representing B in terms of its convex and angular geometry. First, B is centered at a point denoted by C_p. The centered probability simplex is denoted by $\Delta_0^{V-1} := \{x \in \mathbb{R}^V | x + C_p \in \Delta^{V-1}\}$. Then, write $b_k := \beta_k - C_p \in \Delta_0^{V-1}$ for $k = 1, \ldots, K$ and $\tilde{p}_m := p_m - C_p \in \Delta_0^{V-1}$ for $m = 1, \ldots, M$. Note that re-centering leaves corresponding barycentric coordinates $\theta_m \in \Delta^{K-1}$ unchanged. Moreover, the extreme points of centered topic simplex $\tilde{B} := \mathrm{Conv}\{b_1, \ldots, b_K\}$ can now be represented by their directions $v_k \in \mathbb{R}^V$ and corresponding radii $R_k \in \mathbb{R}_+$ such that $b_k = R_k v_k$ for any $k = 1, \ldots, K$.

3.3.1 Coverage of the topic simplex

The first step toward formulating a CoSAC approach is to show how $\tilde{B}$ can be *covered* with exactly K cones and one spherical ball positioned at C_p. A cone is defined as set $\mathcal{S}_\omega(v) := \{p \in \Delta_0^{V-1} | d_{\cos}(v, p) < \omega\}$, where we employ the angular distance (a.k.a. cosine distance) $d_{\cos}(v, p) := 1 - \cos(v, p)$ and $\cos(v, p)$ is the cosine of angle $\angle(v, p)$ formed by vectors v and p.

The Conical coverage It is possible to choose ω so that the topic simplex can be covered with exactly K cones, that is, $\bigcup_{k=1}^{K} \mathcal{S}_\omega(v_k) \supseteq \tilde{B}$. Moreover, each cone contains exactly one vertex. Suppose that C_p is the *incenter* of the topic simplex $\tilde{B}$, with r being the inradius. The incenter and inradius correspond to the maximum volume sphere contained in $\tilde{B}$. Let $a_{i,k}$ denote the distance between the i-th and k-th vertex of $\tilde{B}$, with $a_{min} \leq a_{i,k} \leq a_{max}$ for all i, k, and R_{max}, R_{min} such that $R_{min} \leq R_k := \|b_k\|_2 \leq R_{max}$ $\forall\, k = 1, \ldots, K$. Then we

can establish the following.

Proposition 3.1. For simplex $\tilde{B}$ and $\omega \in (\omega_1, \omega_2)$, where $\omega_1 = 1 - r/R_{max}$ and $\omega_2 = \max\{(a_{min}^2)/(2R_{max}^2), \max_{i,k=1,\dots,K}(1-\cos(b_i, b_k)\}$, the cone $\mathcal{S}_\omega(v)$ around any vertex direction v of $\tilde{B}$ contains exactly one vertex. Moreover, complete coverage holds: $\bigcup_{k=1}^{K} \mathcal{S}_\omega(v_k) \supseteq \tilde{B}$.

We say there is an *angular separation* if $\cos(b_i, b_k) \leq 0$ for any $i, k = 1, \dots, K$ (i.e., the angles for all pairs are at least $\pi/2$), then $\omega \in \left(1 - \frac{r}{R_{max}}, 1\right) \neq \emptyset$. Thus, under angular separation, the range ω that allows for full coverage is nonempty independently of K. Our result is in agreement with that of Nguyen (2015), whose result suggested that topic simplex B can be consistently estimated without knowing K, provided there is a minimum edge length $a_{min} > 0$. The notion of angular separation leads naturally to the Conic Scan-and-Cover algorithm. Before getting there, we show a series of results allowing us to further extend the range of admissible ω.

The inclusion of a spherical ball centered at C_p allows us to expand substantially the range of ω for which conical coverage continues to hold. In particular, we can reduce the lower bound on ω in Proposition 3.1, since we only need to cover the regions near the vertices of $\tilde{B}$ with cones using the following proposition. Fig. 3.1b provides an illustration.

Proposition 3.2. Let $\mathscr{B}(C_p, \mathcal{R}) = \{\tilde{p} \in \mathbb{R}^V | \|\tilde{p} - C_p\|_2 \leq \mathcal{R}\}$, $\mathcal{R} > 0$; ω_1, ω_2 given in Prop. 3.1, and

$$\omega_3 := 1 - \min\left\{\min_{i,k}\left(\frac{R_k \sin^2(b_i, b_k)}{\mathcal{R}} + \cos(b_i, b_k)\sqrt{1 - \frac{R_k^2 \sin^2(b_i, b_j)}{\mathcal{R}^2}}\right), 1\right\}, \quad (3.1)$$

then we have $\bigcup_{k=1}^{K} \mathcal{S}_\omega(v_k) \cup \mathscr{B}(C_p, \mathcal{R}) \supseteq \tilde{B}$ whenever $\omega \in (\min\{\omega_1, \omega_3\}, \omega_2)$.

Notice that as $\mathcal{R} \to R_{max}$, the value of $\omega_3 \to 0$. Hence if $\mathcal{R} \leq R_{min} \approx R_{max}$, the admissible range for ω in Prop. 3.2 results in a substantial strengthening from Prop. 3.1. It is worth noting that the above two geometric propositions do not require any distributional properties inside the simplex.

Coverage leftovers In practice complete coverage may fail if ω and $\mathcal{R}$ are chosen outside of corresponding ranges suggested by the previous two propositions. In that case, it is useful to note that leftover regions will have a very low mass. Next we quantify the mass inside a cone that *does* contain a vertex, which allows us to *reject* a cone that has low mass, therefore not containing a vertex in it.

Proposition 3.3. The cone $S_\omega(v_1)$ whose axis is a topic direction v_1 has mass

$$\mathbb{P}(\mathcal{S}_\omega(v_1)) > \mathbb{P}(\Lambda_c(b_1)) = \frac{\int_{1-c}^1 \theta_1^{\alpha_1-1}(1-\theta_1)^{\sum_{i\neq 1}\alpha_i-1}\mathrm{d}\theta_1}{\int_0^1 \theta_1^{\alpha_1-1}(1-\theta_1)^{\sum_{i\neq 1}\alpha_i-1}\mathrm{d}\theta_1} = \\ \frac{c^{\sum_{i\neq 1}\alpha_i}(1-c)^{\alpha_1}\Gamma(\sum_{i=1}^K \alpha_i)}{(\sum_{i\neq 1}\alpha_i)\Gamma(\alpha_1)\Gamma(\sum_{i\neq 1}\alpha_i)}\left[1+\frac{c\sum_{i=1}^K\alpha_i}{\sum_{i\neq 1}\alpha_i+1}+\frac{c^2(\sum_{i=1}^K\alpha_i)(\sum_{i=1}^K\alpha_i+1)}{(\sum_{i\neq 1}\alpha_i+1)(\sum_{i\neq 1}\alpha_i+2)}+\cdots\right], \tag{3.2}$$

where $\Lambda_c(b_1)$ is the simplicial cap of $\mathcal{S}_\omega(v_1)$ which is composed of vertex b_1 and a base parallel to the corresponding base of $\tilde{B}$ and cutting adjacent edges of $\tilde{B}$ in the ratio $c:(1-c)$.

See Fig. 3.1c for an illustration for the simplicial cap described in the proposition. Given the lower bound for the mass around a cone containing a vertex, we have arrived at the following guarantee.

Proposition 3.4. For $\lambda \in (0,1)$, let c_λ be such that $\lambda = \min_k \mathbb{P}(\Lambda_{c_\lambda}(b_k))$ and let ω_λ be such that

$$c_\lambda = \left(\left(2\sqrt{1-\frac{r^2}{R_{max}^2}}\right)\left(\sin(d)\cot(\arccos(1-\omega_\lambda))+\cos(d)\right)\right)^{-1}, \tag{3.3}$$

where angle $d \leq \min_{i,k} \angle(b_k, b_k - b_i)$. Then, as long as

$$\omega \in \left(\omega_\lambda, \max\left(\frac{a_{min}^2}{2R_{max}^2}, \max_{i,k=1,\ldots,K}(1-\cos(b_i,b_k)\right)\right), \tag{3.4}$$

the bound $\mathbb{P}(\mathcal{S}_\omega(v_k)) \geq \lambda$ holds for all $k = 1,\ldots,K$.

3.3.2 CoSAC: Conic Scan-and-Cover algorithm

Having laid out the geometric foundations, we are ready to present the Conic Scan-and-Cover (CoSAC) algorithm, which is a scanning procedure for detecting the presence of simplicial vertices based on data drawn randomly from the simplex. The idea is simple: iteratively pick the farthest point from the center estimate $\hat{C}_p := \frac{1}{M}\sum_m p_m$, say v, then construct a cone $\mathcal{S}_\omega(v)$ for some suitably chosen ω, and remove all the data residing in this cone. Repeat until there is no data point left.

Specifically, let $A = \{1,\ldots,M\}$ be the index set of the initially unseen data, then set $v := \underset{\tilde{p}_m : m\in A}{\operatorname{argmax}} \|\tilde{p}_m\|_2$ and update $A := A \setminus \mathcal{S}_\omega(v)$. The parameter ω needs to be sufficiently large to ensure that the farthest point is a good estimate of a true vertex, and that the scan will be completed in exactly K iterations; ω needs to be not too large, so that $\mathcal{S}_\omega(v)$ does not contain more than one vertex. The existence of such ω is guaranteed by Prop. 3.1.

In particular, for an equilateral $\tilde{B}$, the condition of the Prop. 3.1 is satisfied as long as $\omega \in (1 - 1/\sqrt{K-1}, 1 + 1/(K-1))$.

In our setting, K is unknown. A smaller ω would be a more robust choice, and accordingly the set A will likely remain non-empty after K iterations. See the illustration of Fig. 3.1a, where the blue regions correspond to A after $K = 3$ iterations of the scan. As a result, we proceed by adopting a stopping criteria based on Prop. 3.2: the procedure is stopped as soon as $\forall\, m \in A\ \|\tilde{p}_m\|_2 < \mathcal{R}$, which allows us to complete the scan in K iterations (as in Fig. 3.1b for $K = 3$).

The CoSAC algorithm is formally presented by Algorithm 3.1. Its running is illustrated in Fig. 3.2, where we show iterations 1, 26, 29, 30 of the algorithm by plotting norms of the centered documents in the active set A and cone $\mathcal{S}_\omega(v)$ against cosine distance to the chosen direction of a topic. Iteration 30 (right) satisfies stopping criteria and therefore CoSAC recovered correct $K = 30$. Note that this type of visual representation can be useful in practice to verify choices of ω and $\mathcal{R}$. The following theorem establishes the consistency of the CoSAC procedure.

Theorem 3.1. Suppose $\{\beta_1, \ldots, \beta_K\}$ are the true topics, incenter C_p is given, $\theta_m \sim \mathrm{Dir}_K(\alpha)$ and $p_m := \sum_k \beta_k \theta_{mk}$ for $m = 1, \ldots, M$ and $\alpha \in \mathbb{R}_+^K$. Let $\hat{K}$ be the estimated number of topics, $\{\hat{\beta}_1, \ldots, \hat{\beta}_{\hat{K}}\}$ be the output of Algorithm 3.1 trained with ω and $\mathcal{R}$ as in Prop. 3.2. Then $\forall\, \epsilon > 0$,

$$\mathbb{P}\left(\left\{\min_{j \in \{1,\ldots,\hat{K}\}} \|\beta_i - \hat{\beta}_j\| > \epsilon \text{ , for any } i \in \{1, \ldots, \hat{K}\}\right\} \cup \{K \neq \hat{K}\}\right) \to 0 \text{ as } M \to \infty.$$

Remark We found the choices $\omega = 0.6$ and $\mathcal{R}$ to be median of $\{\|\tilde{p}_1\|_2, \ldots, \|\tilde{p}_M\|_2\}$ to be robust in practice and agreeing with our theoretical results. From Prop. 3.3 it follows that choosing $\mathcal{R}$ as median length is equivalent to choosing ω resulting in an edge cut ratio c such that $1 - \frac{K}{K-1}(\frac{c}{1-c})^{1-1/K} \geq 1/2$, then $c \leq (\frac{K-1}{2K})^{K/(K-1)}$, which, for any equilateral topic simplex B, is satisfied by setting $\omega \in (0.3, 1)$, provided that $K \leq 2000$ based on the Eq. (3.3).

3.4 Document Conic Scan-and-Cover algorithm

In the topic modeling problem, p_m for $m = 1, \ldots, M$ are *not* given. Instead, under the bag-of-words assumption, we are given the frequencies of words in documents $w_1, \ldots, w_M$ which provide a point estimate $\bar{w}_m := w_m / N_m$ for the p_m. Clearly, if number of documents $M \to \infty$ and length of documents $N_m \to \infty\ \forall m$, we can use Algorithm 3.1 with the

plug-in estimates $\bar{w}_m$ in place of p_m, since $\bar{w}_m \to p_m$. Moreover, C_p will be estimated by $\hat{C}_p := \frac{1}{M}\sum \bar{w}_m$. In practice, M and N_m are finite, some of which may take relatively small values. Taking the topic direction to be the farthest point in the topic simplex, i.e., $v = \underset{\tilde{w}_m : m \in A}{\operatorname{argmax}} \|\tilde{w}_m\|_2$, where $\tilde{w}_m := \bar{w}_m - \hat{C}_p \in \Delta_0^{V-1}$, may no longer yield a robust estimate, because the variance of this topic direction estimator can be quite high (in Proposition 3.5 we show that it is upper bounded with $(1 - 1/V)/N_m$).

To obtain improved estimates, we propose a technique that we call "mean-shifting". Instead of taking the farthest point in the simplex, this technique is designed to shift the estimate of a topic to a high density region, where true topics are likely to be found. Precisely, given a (current) cone $\mathcal{S}_\omega(v)$, we re-position the cone by updating $v := \underset{v}{\operatorname{argmin}} \sum_{m \in \mathcal{S}_\omega(v)} \|\tilde{w}_m\|_2(1 - \cos(\tilde{w}_m, v))$. In other words, we re-position the cone by centering it around the *mean direction* of the cone weighted by the norms of the data points inside, which is simply given by $v \propto \sum_{m \in \mathcal{S}_\omega(v)} \tilde{w}_m / \operatorname{card}(\mathcal{S}_\omega(v))$. This results in reduced variance of the topic direction estimate, due to the averaging over data residing in the cone.

The mean-shifting technique may be slightly modified and taken as a local update for a subsequent optimization which cycles through the entire set of documents and iteratively updates the cones. The optimization is with respect to the following weighted spherical k-means objective:

$$\min_{\|v_k\|_2=1, k=1,\ldots K} \sum_{k=1}^{K} \sum_{m \in S^k(v_k)} \|\tilde{w}_m\|_2(1 - \cos(v_k, \tilde{w}_m)), \tag{3.5}$$

where cones $S^k(v_k) = \{m | d_{\cos}(v_k, \tilde{p}_m) < d_{\cos}(v_l, \tilde{p}_i)\ \forall l \neq k\}$ yield a disjoint data partition $\bigsqcup_{k=1}^{K} S^k(v_k) = \{1, \ldots, M\}$ (this is different from $\mathcal{S}_\omega(v_k)$). The rationale of spherical k-means optimization is to use full data for estimation of topic directions, hence further reducing the variance due to short documents. The connection between objective function (3.5) and topic simplex estimation is given in the Supplement. Finally, obtain topic norms R_k along the directions v_k using maximum projection: $R_k := \max_{m : m \in S^k(v_k)} \langle v_k, \tilde{w}_m \rangle$. Our entire procedure is summarized in Algorithm 3.2.

Remark In Step 9 of the algorithm, cone $\mathcal{S}_\omega(v)$ with a very low cardinality, i.e.,

$$\operatorname{card}(\mathcal{S}_\omega(v)) < \lambda M,$$

for some small constant λ, is discarded because this is likely an outlier region that does not actually contain a true vertex. The choice of λ is governed by results of Prop. 3.4.

For small $\alpha_k = 1/K,\ \forall k$, $\lambda \leq \mathbb{P}(\Lambda_c) \approx \frac{c^{(K-1)/K}}{(K-1)(1-c)}$ and for an equilateral $\tilde{B}$ we can choose d such that $\cos(d) = \sqrt{\frac{K+1}{2K}}$. Plugging these values into Eq. (3.3) leads to $c = \left(\left(2\sqrt{1-\frac{1}{K^2}}\right)\left(\sqrt{\frac{K-1}{2K}}(\frac{1-\omega}{\sqrt{1-(1-\omega)^2}}) + \sqrt{\frac{K+1}{2K}}\right)\right)^{-1}$. Now, plugging in $\omega = 0.6$ we obtain $\lambda \leq K^{-1}$ for large K. Our approximations were based on large K to get a sense of λ, we now make a conservative choice $\lambda = 0.001$, so that $(K)^{-1} > \lambda\ \forall K < 1000$. As a result, a topic is rejected if the corresponding cone contains less than 0.1% of the data.

Finding anchor words using Conic Scan-and-Cover Another approach to reduce the noise is to consider the problem from a different viewpoint, where Algorithm 3.1 will prove itself useful. RecoverKL by Arora et al. (2013) can identify topics with diminishing errors (in number of documents M), *provided* that topics contain anchor words. The problem of finding anchor words geometrically reduces to identifying rows of the word-to-word co-occurrence matrix that form a simplex containing other rows of the same matrix (cf. Arora et al. (2013) for details). An advantage of this approach is that noise in the word-to-word co-occurrence matrix goes to zero as $M \to \infty$ no matter the document lengths, hence we can use Algorithm 3.1 with "documents" being rows of the word-to-word co-occurrence matrix to learn anchor words nonparametrically and then run RecoverKL to obtain topic estimates. We will call this procedure cscRecoverKL.

Algorithm 3.1 Conic Scan-and-Cover (CoSAC)

Input: document generating distributions $p_1, \ldots, p_M$,
angle threshold ω, norm threshold $\mathcal{R}$
Output: topics $\beta_1, \ldots, \beta_k$

1: $\hat{C}_p = \frac{1}{M}\sum_m p_m$ {find center}; $\tilde{p}_m := p_m - \hat{C}_p$ for $m = 1, \ldots, M$ {center the data}
2: $A_1 = \{1, \ldots, M\}$ {initialize active set}; $k = 1$ {initialize topic count}
3: **while** $\exists m \in A_k : \|\tilde{p}_m\|_2 > \mathcal{R}$ **do**
4: $v_k = \underset{\tilde{p}_m : m \in A_k}{\operatorname{argmax}} \|\tilde{p}_m\|_2$ {find topic}
5: $\mathcal{S}_\omega(v_k) = \{m : d_{\cos}(\tilde{p}_m, v_k) < \omega\}$ {find cone of near documents}
6: $A_k = A_k \setminus \mathcal{S}_\omega(v_k)$ {update active set}
7: $\beta_k = v_k + \hat{C}_p$, $k = k + 1$ {compute topic}
8: **end while**

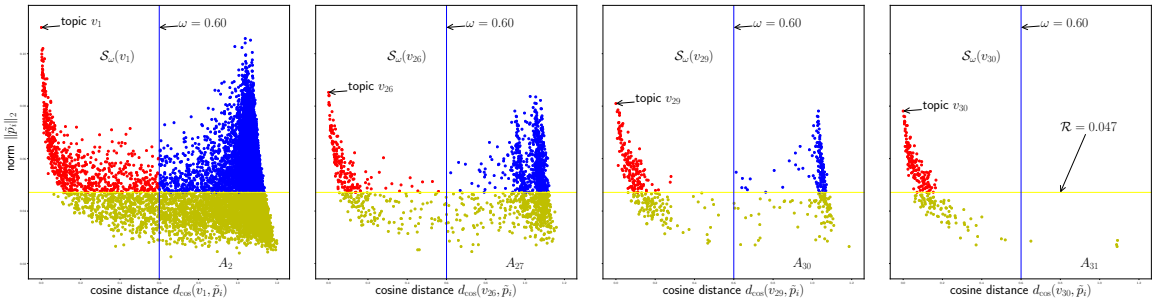


Figure 3.2: Iterations 1, 26, 29, 30 of the Algorithm 3.1. Red are the documents in the cone $\mathcal{S}_\omega(v_k)$; blue are the documents in the active set A_{k+1} for next iteration. Yellow are documents $\|\tilde{p}_m\|_2 < \mathcal{R}$.

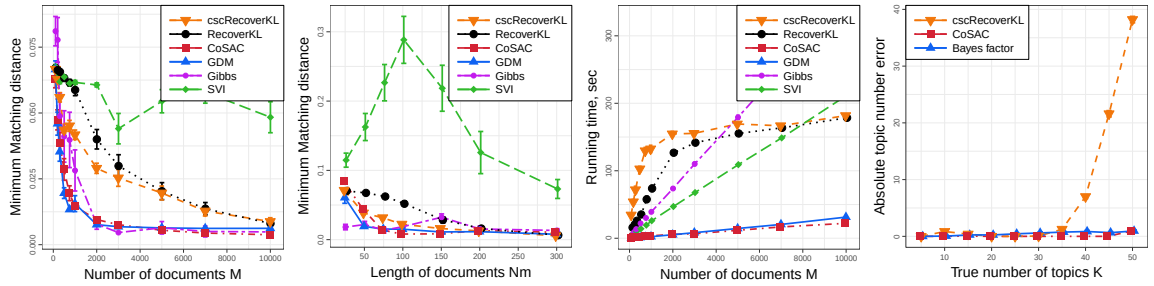


Figure 3.3: Minimum matching Euclidean distance for (a) varying corpora size, (b) varying length of documents; (c) Running times for varying corpora size; (d) Estimation of number of topics.

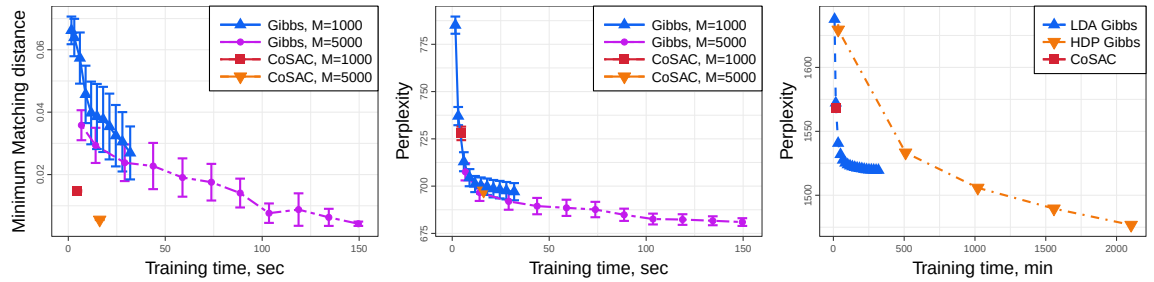


Figure 3.4: Gibbs sampler convergence analysis for (a) Minimum matching Euclidean distance for corpora sizes 1000 and 5000; (b) Perplexity for corpora sizes 1000 and 5000; (c) Perplexity for NYTimes data.

3.5 Experimental results

3.5.1 Simulation experiments

In the simulation studies we shall compare CoSAC (Algorithm 3.2) and cscRecoverKL based on Algorithm 3.1 both of which don't have access to the true K, versus popular parametric topic modeling approaches (trained with true K): Stochastic Variational Inference

Algorithm 3.2 CoSAC for documents

Input: normalized documents $\bar{w}_1, \ldots, \bar{w}_M$,
angle threshold ω, norm threshold $\mathcal{R}$, outlier threshold λ
Output: topics $\beta_1, \ldots, \beta_k$

$\hat{C}_p = \frac{1}{M}\sum_m \bar{w}_m$ {find center}; $\tilde{w}_m := \bar{w}_m - \hat{C}_p$ for $m = 1, \ldots, M$ {center the data}
$A_1 = \{1, \ldots, M\}$ {initialize active set}; $k = 1$ {initialize topic count}
while $\exists\, m \in A_k : \|\tilde{w}_m\|_2 > \mathcal{R}$ **do**
  $v_k = \underset{\tilde{w}_m : m \in A_k}{\operatorname{argmax}} \|\tilde{w}_m\|_2$ {initialize direction}
  while v_k not converged **do** {mean-shifting}
    $\mathcal{S}_\omega(v_k) = \{m : d_{\cos}(\tilde{w}_m, v_k) < \omega\}$ {find cone of near documents}
    $v_k = \sum_{m \in \mathcal{S}_\omega(v_k)} \tilde{w}_m / \operatorname{card}(\mathcal{S}_\omega(v_k))$ {update direction}
  end while
  $A_k = A_k \setminus \mathcal{S}_\omega(v_k)$ {update active set}
  if $\operatorname{card}(\mathcal{S}_\omega(v_k)) > \lambda M$ **then** $k = k + 1$ {record topic direction}
end while
$v_1, \ldots, v_k =$ weighted spherical k-means $(v_1, \ldots, v_k, \tilde{w}_1, \ldots, \tilde{w}_M)$
for l in $\{1, \ldots, k\}$ **do**
  $R_l := \max_{m : m \in S^l(v_l)} \langle v_l, \tilde{w}_m \rangle$ {find topic length along direction v_l}
  $\beta_l = R_l v_l + \hat{C}_p$ {compute topic}
end for

(SVI), Collapsed Gibbs sampler, RecoverKL and GDM (more details in the Supplement). The comparisons are done on the basis of minimum-matching Euclidean distance, which quantifies distance between topic simplices (Tang et al., 2014), and running times (perplexity scores comparison is given in Fig. 3.C.1). Lastly we will demonstrate the ability of CoSAC to recover correct number of topics for a varying K.

Estimation of the LDA topics First we evaluate the ability of CoSAC and cscRecoverKL to estimate topics $\beta_1, \ldots, \beta_K$, fixing $K = 15$. Fig. 3.3(a) shows performance for the case of fewer $M \in [100, 10000]$ but longer $N_m = 500$ documents (e.g. scientific articles, novels, legal documents). CoSAC demonstrates performance comparable in accuracy to Gibbs sampler and GDM.

Next we consider larger corpora $M = 30000$ of shorter $N_m \in [25, 300]$ documents (e.g. news articles, social media posts). Fig. 3.3(b) shows that this scenario is harder and CoSAC matches the performance of Gibbs sampler for $N_m \geq 75$. Indeed across both experiments CoSAC only made mistakes in terms of K for the case of $N_m = 25$, when it was underestimating on average by 4 topics and for $N_m = 50$ when it was off by around 1, which explains the earlier observation. Experiments with varying V and α are given in the

Supplement.

It is worth noting that cscRecoverKL appears to be strictly better than its predecessor. This suggests that our procedure for selection of anchor words is more accurate in addition to being nonparametric.

Running time A notable advantage of the CoSAC algorithm is its speed. In Fig. 3.3(c) we see that Gibbs, SVI, GDM and CoSAC all have linear complexity growth in M, but the slopes are very different and approximately are IN_m for SVI and Gibbs (where I is the number of iterations which has to be large enough for convergence), number of k-means iterations to converge for GDM and is of order K for the CoSAC procedure making it the fastest algorithm of all under consideration.

Next we compare CoSAC to per iteration quality of the Gibbs sampler trained with 500 iterations for $M = 1000$ and $M = 5000$. Fig. 3.4(b) shows that Gibbs sampler, when true K is given, can achieve good perplexity score as fast as CoSAC and outperforms it as training continues, although Fig. 3.4(a) suggests that much longer training time is needed for Gibbs sampler to achieve good topic estimates and small estimation variance.

Estimating number of topics Model selection in the LDA context is a quite challenging task and, to the best of our knowledge, there is no "go to" procedure. One of the possible approaches is based on refitting LDA with multiple choices of K and using Bayes Factor for model selection (Griffiths & Steyvers, 2004). Another option is to adopt the Hierarchical Dirichlet Process (HDP) model, but we should understand that it is not a procedure to estimate K of the LDA model, but rather a particular prior on the number of topics, that assumes K to grow with the data. A more recent suggestion is to slightly modify LDA and use Bayes moment matching (Hsu & Poupart, 2016), but, as can be seen from Figure 2 of their paper, estimation variance is high and the method is not very accurate (we tried it with true $K = 15$ and it took above 1 hour to fit and found 35 topics). Next we compare Bayes factor model selection versus CoSAC and cscRecoverKL for $K \in [5, 50]$. Fig. 3.3(d) shows that CoSAC consistently recovers *exact* number of topics in a wide range.

We also observe that cscRecoverKL does not estimate K well (underestimates) in the higher range. This is expected because cscRecoverKL finds the number of anchor words, *not* topics. The former is decreasing when later is increasing. Attempting to fit RecoverKL with more topics than there are anchor words might lead to deteriorating performance and our modification can address this limitation of the RecoverKL method.

Table 3.1: Modeling topics of NYTimes articles

	K	Perplexity	Coherence	Time
cscRecoverKL	27	2603	-238	37 min
HDP Gibbs	221 ± 5	1477 ± 1.6	-442 ± 1.7	35 hours
LDA Gibbs	80	1520 ± 1.5	-300 ± 0.7	5.3 hours
CoSAC	159	1568	-322	19 min

3.5.2 Real data analysis

In this section we demonstrate CoSAC algorithm for topic modeling on one of the standard bag of words datasets — NYTimes news articles. After preprocessing we obtained $M \approx 130,000$ documents over $V = 5320$ words. Bayes factor for the LDA selected the smallest model among $K \in [80, 195]$, while CoSAC selected 159 topics. We think that disagreement between the two procedures is attributed to the misspecification of the LDA model when real data is in play, which affects Bayes factor, while CoSAC is largely based on the geometry of the topic simplex.

The results are summarized in Table 3.1 — CoSAC found 159 topics in less than 20min; cscRecoverKL estimated the number of anchor words in the data to be 27 leading to fewer topics. Fig. 3.4(c) compares CoSAC perplexity score to per iteration test perplexity of the LDA (1000 iterations) and HDP (100 iterations) Gibbs samplers. Text files with top 20 words of all topics are available on GitHub. Sample of the CoSAC topics is demonstrated in Table 3.2. We note that CoSAC procedure recovered meaningful topics, contextually similar to LDA and HDP (e.g. elections, terrorist attacks, Enron scandal, etc.) and also recovered more specific topics about Mike Tyson, boxing and case of Timothy McVeigh which were present among HDP topics, but not LDA ones. We conclude that CoSAC is a practical procedure for topic modeling on large scale corpora able to find meaningful topics in a short amount of time.

3.6 Discussion

We have analyzed the problem of estimating topic simplex without assuming number of vertices (i.e., topics) to be known. We showed that it is possible to cover topic simplex using two types of geometric shapes, cones and a sphere, leading to a class of Conic Scan-and-Cover algorithms. We then proposed several geometric correction techniques to account for the noisy data. Our procedure is accurate in recovering the true number of topics, while remaining practical due to its computational speed. We think that angular geometric

Table 3.2: CoSAC NYTimes topics sample

Cooking	*Stem Cells*	*Antitrust*	*LGBT*	*Elections*
cup	cell	Microsoft	gay	ballot
minutes	stem	window	lesbian	Al Gore
tablespoon	research	company	right	election
add	human	software	sex	votes
teaspoon	scientist	case	marriage	recount
pepper	cloning	system	group	Florida
oil	patient	operating	couples	court
sugar	disease	computer	sexual	vote
butter	phones	antitrust	partner	voter
pan	researcher	court	issue	count

approach might allow for fast and elegant solutions to other clustering problems, although as of now it does not immediately offer a unifying problem solving framework like MCMC or variational inference. An interesting direction in a geometric framework is related to building models based on geometric quantities such as distances and angles.

Appendix

3.A Proofs of main theorems

We start by reminding the reader of our geometric setup. First, topic simplex $B := \mathrm{Conv}(\beta_1, \ldots, \beta_K)$ is centered at a point denoted by C_p. Let $\Delta_0^{V-1} := \{x \in \mathbb{R}^V : x + C_p \in \Delta^{V-1}\}$ — centered probability simplex. Then, write $b_k := \beta_k - C_p \in \Delta_0^{V-1}$ for $k = 1, \ldots, K$ and $\tilde{p}_m := p_m - C_p \in \Delta_0^{V-1}$ for $m = 1, \ldots, M$. Note that re-centering leaves corresponding barycentric coordinates $\theta_m \in \Delta^{K-1}$ unchanged. Moreover, the extreme points of centered topic simplex $\tilde{B} := \mathrm{Conv}\{b_1, \ldots, b_K\}$ can now be represented by their directions $v_k \in \mathbb{R}^V$ and corresponding radii $R_k \in \mathbb{R}_+$ such that $b_k = R_k v_k$ for any $k = 1, \ldots, K$.

3.A.1 Coverage of the topic simplex

Suppose that C_p is the incenter of the topic simplex $\tilde{B}$, with r being the inradius. Recall that the incenter and inradius correspond to the maximum volume sphere inside $\tilde{B}$. Let $a_{i,k}$ denote the distance between the i^{th} and k^{th} vertex of $\tilde{B}$, with $a_{min} \leq a_{i,k} \leq a_{max}$ for all i, k, and R_{max}, R_{min} such that $R_{min} \leq R_k := \|b_k\|_2 \leq R_{max}\ \forall\ k = 1, \ldots, K$

Proposition 3.1. *For simplex $\tilde{B}$ and $\omega \in (\omega_1, \omega_2)$, where $\omega_1 = 1 - r/R_{max}$ and $\omega_2 = \max\{(a_{min}^2)/(2R_{max}^2), \max_{i,k=1,\ldots,K}(1-\cos(b_i, b_k)\}$, the cone $\mathcal{S}_\omega(v)$ around any vertex direction v of $\tilde{B}$ contains exactly one vertex. Moreover, complete coverage holds: $\bigcup_{k=1}^{K} \mathcal{S}_\omega(v_k) \supseteq \tilde{B}$.*

Proof. Let $\omega_0 = \frac{a_{min}^2}{2R_{max}^2}$. Then, for any $k \in \{1, \ldots, K\}$, for any $\omega \leq \omega_0$, $\mathcal{S}_\omega(v_k)$ does not contain any other vertices. This can be explained as follows. Fix k, and choose $i \in \{1, \ldots, K\} \neq k$. Define $\phi_{i,k}$ as the angle at C_p made by the side connecting the vertex i and vertex k. Then from the cosine law for triangles, we have

$$\cos(\phi_{i,k}) = \frac{R_i^2 + R_k^2 - a_{i,k}^2}{2R_i R_k}.$$

Now, for any $\phi \leq \min_{i,k} \phi_{i,k}$, with $\omega_\phi = 1 - \cos(\phi)$, the cone $\mathcal{S}_{\omega_\phi}(v_k)$ does not cover any vertex other than vertex k, for any k. Now $\phi_1 = \min_{i,k} \phi_{i,k}$ satisfies

$$1 - \cos(\phi_1) \leq \frac{a_{min}^2}{2R_{max}^2} - \frac{(R_{max} - R_{min})^2}{2R_{max}R_{min}} \leq \frac{a_{min}^2}{2R_{max}^2}.$$

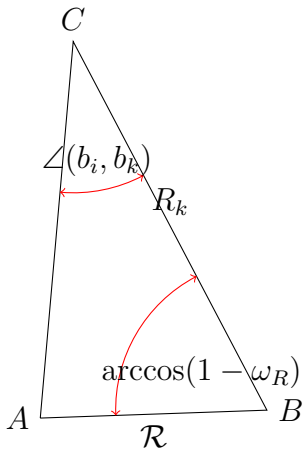


Figure 3.A.1: C : k^{th} vertex point, A : point where the adjacent side to the vertex has been cut off by the sphere, R_k: distance to k^{th} vertex from incenter, $\mathcal{R}$: radius of sphere, B : incenter

from which we obtain the upper bound for ω. For the lower bound, consider for vertex k, $\mathcal{S}(v_k)$ the cone connecting the incenter to facial incenters of facets containing vertex k. Then $\bigcup_{k=1}^{K} \mathcal{S}(v_k) \supseteq \tilde{B}$. Now for each k, $\mathcal{S}(v_k) \subseteq \mathcal{S}_{\omega_2}(v_k)$, where $\omega_2 = 1 - \cos(\phi_2)$, with ϕ_2 satisfying $\cos(\phi_2) \leq \min_{k\in\{1,\dots,K\}} \frac{r}{R_k}$. From this we get the lower bound. The restriction $2R_{max}^2 \leq a_{min}^2$ is needed to ensure that the set $\left\{\omega : 1 - (\frac{r}{R_{max}}) \leq \omega \leq (\frac{a_{min}^2}{2R_{max}^2})\right\}$ is non-empty. □

Proposition 3.2. *Let* $\mathscr{B}(C_p, \mathcal{R}) = \{\tilde{p} \in \mathbb{R}^V | \|\tilde{p} - C_p\|_2 \leq \mathcal{R}\}$, $\mathcal{R} > 0$; ω_1, ω_2 *given in Prop. 3.1, and*

$$\omega_3 := 1 - \min\left\{\min_{i,k}\left(\frac{R_k \sin^2(b_i, b_k)}{\mathcal{R}} + \cos(b_i, b_k)\sqrt{1 - \frac{R_k^2 \sin^2(b_i, b_j)}{\mathcal{R}^2}}\right), 1\right\}, \quad (3.6)$$

then we have $\bigcup_{k=1}^{K} \mathcal{S}_\omega(v_k) \cup \mathscr{B}(C_p, \mathcal{R}) \supseteq \tilde{B}$ *whenever* $\omega \in (\min\{\omega_1, \omega_3\}, \omega_2)$.

Proof. Let $\phi_{i,k} = \arccos(1 - \omega_{i,k})$ be the angle formed by the line joining the k^{th} vertex to the incenter C_p and the radial vector from incenter to the point where the sphere cuts the edge connecting i and k (segment AB on Fig. 3.A.1). From the sine law for a triangle we have

$$\cos(\phi_{i,k}) + \cot(b_i, b_k)\sin(\phi_{i,k}) - \frac{R_k}{\mathcal{R}} = 0. \quad (3.7)$$

Solving for $\phi_{i,k}$ we have $\cos(\phi_{i,k}) = \left(\frac{R_k \sin^2(b_i,b_k)}{\mathcal{R}} + \cos(b_i,b_k)\sqrt{1 - \frac{R_k^2 \sin^2(b_i,b_k)}{\mathcal{R}^2}} \right)$. Now, since we must choose the largest such ϕ over all i and k, the bound follows immediately. Notice that as $\mathcal{R} \to R_{max}$, the value of $\left(\frac{R_k \sin^2(b_i,b_k)}{\mathcal{R}} + \cos(b_i,b_k)\sqrt{1 - \frac{R_k^2 \sin^2(b_i,b_k)}{\mathcal{R}^2}} \right) \to 1$, whereas $\frac{r}{R_{max}} < 1$ strictly. Thus, as $\mathcal{R}$ increases the lower bound in this limiting scenario is dominated by $1 - \min_{i,k} \left(\frac{R_k \sin^2(b_i,b_k)}{\mathcal{R}} + \cos(b_i,b_k)\sqrt{1 - \frac{R_k^2 \sin^2(b_i,b_k)}{\mathcal{R}^2}} \right)$, thereby obtaining an improvement in the bound from Proposition 3.1. □

Proposition 3.3. *The cone $S_\omega(v_1)$ whose axis is a topic direction v_1 has mass*

$$\mathbb{P}(\mathcal{S}_\omega(v_1)) > \mathbb{P}(\Lambda_c(b_1)) = \frac{\int_{1-c}^1 \theta_1^{\alpha_1 - 1}(1-\theta_1)^{\sum_{i\neq 1}\alpha_i - 1}\mathrm{d}\theta_1}{\int_0^1 \theta_1^{\alpha_1 - 1}(1-\theta_1)^{\sum_{i\neq 1}\alpha_i - 1}\mathrm{d}\theta_1} = \frac{c^{\sum_{i\neq 1}\alpha_i}(1-c)^{\alpha_1}\Gamma(\sum_{i=1}^K \alpha_i)}{(\sum_{i\neq 1}\alpha_i)\Gamma(\alpha_1)\Gamma(\sum_{i\neq 1}\alpha_i)}\left[1 + \frac{c\sum_{i=1}^K \alpha_i}{\sum_{i\neq 1}\alpha_i + 1} + \frac{c^2(\sum_{i=1}^K \alpha_i)(\sum_{i=1}^K \alpha_i + 1)}{(\sum_{i\neq 1}\alpha_i + 1)(\sum_{i\neq 1}\alpha_i + 2)} + \cdots\right], \tag{3.8}$$

where $\Lambda_c(b_1)$ is the simplicial cap of $\mathcal{S}_\omega(v_1)$ which is composed of vertex b_1 and a base parallel to the corresponding base of $\tilde{B}$ and cutting adjacent edges of $\tilde{B}$ in the ratio $c : (1-c)$.

The truncated beta probability calculations in Proposition 3.3 can be found in Olver et al. (2010).

Proposition 3.4. *For $\lambda \in (0,1)$, let c_λ be such that $\lambda = \min_k \mathbb{P}(\Lambda_{c_\lambda}(b_k))$ and let ω_λ be such that*

$$c_\lambda = \left(\left(2\sqrt{1 - \frac{r^2}{R_{max}^2}} \right) (\sin(d)\cot(\arccos(1-\omega_\lambda)) + \cos(d)) \right)^{-1}, \tag{3.9}$$

where angle $d \leq \min_{i,k} \angle(b_k, b_k - b_i)$. Then, as long as

$$\omega \in \left(\omega_\lambda, \max\left(\frac{a_{min}^2}{2R_{max}^2}, \max_{i,k=1,\ldots,K}(1 - \cos(b_i,b_k)) \right) \right), \tag{3.10}$$

the bound $\mathbb{P}(\mathcal{S}_\omega(v_k)) \geq \lambda$ holds for all $k = 1,\ldots,K$.

Proof. Consider Figure 3.A.1, with length of $AC = a_{i,k}c$, where c is the proportion in which the cone cuts AC, the edge joining vertex i and vertex k. Now, from the sine law of a triangle,

$$\frac{R_k}{a_{i,k}c} = \sin(b_i,b_k)\cot\phi_{i,k} + \cos(b_i,b_k) \tag{3.11}$$

where $\phi_{i,k}$ is as defined in the proof of Proposition 3.2. Now $\frac{a_{i,k}}{R_k} \leq \frac{2(\sqrt{R_{max}^2 - r^2})}{R_{max}}$. The choice of $\phi_\lambda = \cos \omega_\lambda$ satisfies

$$c_\lambda \geq \frac{1}{2\sqrt{1 - \frac{r^2}{R_{max}^2}}} \min_{i,k} \frac{1}{\sin(b_i, b_k) \cot \phi_\lambda + \cos(b_i, b_k)} \tag{3.12}$$

therefore proves the theorem. Since, $\phi_\lambda \leq \frac{\pi}{2} - \angle(b_i, b_k)$, for all i, k, the function

$$\sin(b_i, b_k) \cot \phi_\lambda + \cos(b_i, b_k)$$

is increasing as the angle between b_i and b_k increases, as can be checked for maxima by the first derivative rule. Using the cosine law,

$$\cos(b_i, b_k) = \frac{-R_i^2 + R_k^2 + a_{i,k}^2}{2 a_{i,k} R_k}. \tag{3.13}$$

Minimizing this quantity with respect to i and k we get the result. □

3.A.2 Consistency of the Conic Scan-and-Cover algorithm

Under the LDA setup (as presented in Section 3.2), recall that $a_{i,k}$ is the length of the edge connecting the i^{th} and k^{th} vertex, i.e., $\|\beta_i - \beta_k\|_2 = a_{i,k}$, where $\|\cdot\|_2$ is the ℓ_2 norm. Let $\mathscr{B}(\cdot, \epsilon)$ denote an ϵ-ball in ℓ_2-norm. Then the following result states that with high probability there exists a document in a neighborhood of every vertex.

Lemma 3.1. Let $p_m := \sum_k \beta_k \theta_{mk}$ for $m = 1, \ldots, M$ as before. Then for any i and any $0 < \epsilon < \max_{k \neq i} a_{i,k}$,

$$\mathbb{P}(p_m \notin \mathscr{B}(\beta_i, \epsilon) \; \forall \; m \in \{1, \ldots, M\}) \leq \left(\frac{\int_0^{1 - (\epsilon / \max_{k \neq i} a_{i,k})} \theta_i^{\alpha_i - 1} (1 - \theta_i)^{\sum_{j \neq i} \alpha_j - 1} \mathrm{d}\theta_i}{\int_0^1 \theta_i^{\alpha_i - 1} (1 - \theta_i)^{\sum_{j \neq 1} \alpha_j - 1} \mathrm{d}\theta_i} \right)^M \tag{3.14}$$

Since $\left(\frac{\int_0^{1 - (\epsilon / \max_{k \neq i} a_{i,k})} \theta_i^{\alpha_i - 1} (1 - \theta_i)^{\sum_{j \neq i} \alpha_j - 1} \mathrm{d}\theta_i}{\int_0^1 \theta_i^{\alpha_i - 1} (1 - \theta_i)^{\sum_{j \neq 1} \alpha_j - 1} \mathrm{d}\theta_i} \right) < 1$, for all i because Beta distribution is absolutely continuous in $(0, 1)$, the bound on the right hand side goes to 0 as $M \to \infty$.

Let $\{\hat{\beta}_1, \ldots, \hat{\beta_K}\}$ be the topics identified by Conic Scan-and-Cover algorithm, with labels permuted according to the minimum matching distance criteria, with $\{\beta_1, \ldots, \beta_K\}$ being the true topics. Then the following result shows the consistency of the identified topics.

Theorem 3.1. *Suppose* $\{\beta_1, \ldots, \beta_K\}$ *are the true topics, incenter* C_p *is given,* $\theta_m \sim \mathrm{Dir}_K(\alpha)$ *and* $p_m := \sum_k \beta_k \theta_{mk}$ *for* $m = 1, \ldots, M$ *and* $\alpha \in \mathbb{R}_+^K$. *Let* $\{\hat{\beta}_1, \ldots, \hat{\beta}_{\hat{K}}\}$ *be the output of the Conic Scan-and-Cover algorithm trained with* ω *and* $\mathcal{R}$ *as in Proposition 3.2. Then* $\forall\, \epsilon > 0$,

$$\mathbb{P}\left(\left\{\min_{j\in\{1,\ldots,\hat{K}\}} \|\beta_i - \hat{\beta}_j\| > \epsilon \text{ , for any } i \in \{1, \ldots, \hat{K}\}\right\} \cup \{K \neq \hat{K}\}\right) \to 0 \text{ as } M \to \infty.$$

Proof. From the description of the Conic Scan-and-Cover algorithm it suffices to prove that for the suitable choice of $\omega, \mathcal{R}$ as in Proposition 3.2 there holds

$$\mathbb{P}(\exists\, x_i \in \{p_1, \ldots, p_m\} \text{ such that } \|\beta_i - x_i\| < \epsilon\, \forall\, i \in \{1, \ldots, K\}) \to 1 \text{ as } M \to \infty.$$

But this probability expression is bounded from below by $1 - \sum_{i=1}^K \mathbb{P}(p_m \notin \mathscr{B}(\beta_i, \epsilon)\ \forall\ m \in \{1, \ldots, M\})$. The conclusion now follows from Lemma 3.1. □

3.A.3 Variance argument for multinomial setup

In the topic modeling problem we are not given p_m for $m = 1, \ldots, M$. Under the bag-of-words assumption we have access to the frequencies of words in documents $w_1, \ldots, w_M$ which provide a point estimate $\bar{w}_m := w_m/N_m$ for the p_m. The following proposition establishes a bound on the variation of $\bar{w}_m$ from p_m.

Proposition 3.5.

$$\mathbb{E}[\|\bar{w}_m - p_m\|_2^2] \leq \frac{1 - (1/V)}{N_m}. \tag{3.15}$$

Proof. By iterated expectation identity,

$$\begin{aligned}
\mathbb{E}[\|\bar{w}_m - p_m\|_2^2] &= \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^V \|\bar{w}_{mi} - p_{mi}\|_2^2 \middle| p_m\right]\right] \\
&= \mathbb{E}\left[\sum_{i=1}^V \frac{p_{mi}(1 - p_{mi})}{N_m}\right] \\
&= \frac{1 - \mathbb{E}[\sum_{i=1}^V p_{mi}^2]}{N_m} \leq \frac{1 - (1/V)}{N_m}.
\end{aligned}$$

The second equality follows because conditioned on p_m, each $w_{mi} \sim \mathrm{Bin}(N_m, p_{mi})$. The last inequality follows from Cauchy-Schwartz Inequality. □

3.B Spherical k-means for topic modeling

We aim to clarify the role of Step 11 of the document Conic Scan-and-Cover algorithm, a geometric correction technique based on weighted spherical k-means optimization.

3.B.1 Topic directions as solutions to weighted spherical k-means

The weighted spherical k-means objective takes the form

$$\min_{\|v_k\|_2=1,k=1,\dots K} \sum_{k=1}^{K} \sum_{m\in S^k(v_k)} r_m(1-\cos(v_k,\tilde{p}_m)), \tag{3.16}$$

where $S^k(v_k) := \{m|\cos(v_k,\tilde{p}_m)) > \cos(v_l,\tilde{p}_m)\ \forall l \neq k)\}$. Next observe that:

$$r_m \cos(v_k,\tilde{p}_m) = \langle v_k,\tilde{p}_m\rangle = \sum_{i=1}^{K} \theta_{mi}\langle v_k, b_i\rangle = \sum_{i=1}^{K} \theta_{mi} R_i \alpha_i(v_k), \tag{3.17}$$

so our objective 3.16 becomes:

$$\max_{\|v_k\|_2=1,k=1,\dots K} \sum_{k=1}^{K} \sum_{m\in S^k(v_k)} \sum_{i=1}^{K} \theta_{mi} R_i \alpha_i(v_k). \tag{3.18}$$

Now, if $R_1 = \ldots = R_K$ and $\alpha_i(b_k) = \alpha_i(b_l)\ \forall\ k, l \neq i$, which implies that topic simplex is equilateral, we see that cluster boundaries of topic directions are given by $m \in S^k(b_k)$ iff $\theta_{mk} > \theta_{ml}\ \forall\ l \neq k$. Observe that the corresponding partition is defined by the geometric *medians* of topic simplex, which in turn partitions it into equal volume parts. Then, assuming that the topic simplex B is symmetric, combined with the symmetricity of the Dirichlet distribution of θ_m-s, it follows that b_k is the centroid of $S^k(b_k)$ for $k = 1, \dots, K$.

3.B.2 Role of the spherical k-means in CoSAC algorithm for documents

The result of Section 3.B.1 shows that weighted spherical k-means with Lloyd type updates (Lloyd, 1982) will converge to the directions of the true topics if it is initialized in their respective neighborhoods and equilaterality of B and symmetricity of Dirichlet for document topic proportions is satisfied.

Recall that goal of the Conic Scan-and-Cover is to find the number of topics and their directions, while Mean Shifting was used to address the noise in the data. We proceed to

compare weighted spherical k-means by itself (with 500 iterations, which makes it slower than CoSAC) versus document Conic Scan-and-Cover with only Mean Shifting and the full document Conic Scan-and-Cover algorithm to see the effect of the spherical k-means post-processing step. Results in Fig. 3.B.1 are for the same scenarios as in Section 3.5 – that is when either documents are short $N_m \in [25, 300]$ but corpora is large $M = 30000$ or when documents are longer $N_m = 500$ and corpora is smaller $M \in [100, 10000]$. We see that spherical k-means by itself does not succeed, whereas when used as a postprocessing step for CoSAC it allows for a slight improvement when documents are short. This is because it operates on the full data partition when taking averages for direction estimation, while Mean Shifting only has access to the data in its respective cone $\mathcal{S}_\omega(v)$. Using more data is important for noise reduction when documents are short as suggested by our analysis.

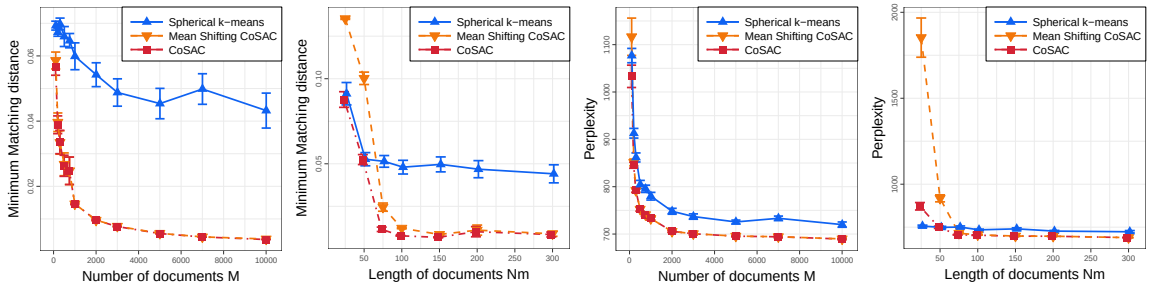


Figure 3.B.1: Minimum matching Euclidean distance for (a) varying corpora size, (b) varying length of documents. Perplexity for (c) varying corpora sizes, (d) varying length of documents.

3.C Additional experiments

3.C.1 Perplexity comparison

In this section we present perplexity scores comparison for the experiments of Section 3.5. For simulation experiments we used $V = 2000$, symmetric $\alpha = \eta = 0.1$. To compute held-out perplexity for the CoSAC we employed projection based estimates for topic proportions θ_m from Chapter 2, which led to a slightly worse perplexity scores for CoSAC and GDM in comparison to Gibbs sampler. However, CoSAC (except for $N_m = 25$, when it slightly underestimates K) shows competitive performance without requiring K as an input. We note that as before cscRecoverKL outperforms RecoverKL in all cases.

3.C.2 Varying vocabulary size V

Our next experiment investigates the influence of vocabulary size V. We set $N_m = 500$, $M = 5000$, $K = 15$, symmetric $\alpha = \eta = 0.1$ and varied V from 2000 to 15000. We discovered that $\omega = 0.6$ is too small for $V > 10000$, meaning that CoSAC algorithm does not find enough documents in the corresponding cones and keeps discarding without recording topics (per Step 9 of Algorithm 3.2). This can be explained by the fact that vectors tending to be far apart in high dimensions and relatively (to $V > 10000$) small values of corpora size M and document lengths N_m. On the other hand, setting $\omega = 0.75$ worked well for all values of V in this experiment. Results are reported in Fig. 3.C.1(c), (d) and Fig. 3.C.2(d). Document CoSAC with $\omega = 0.75$ recovered true $K = 15$ for all values of V and showed better recovery than GDM and Gibbs sampler in terms of minimum matching distance, while Gibbs sampler had slightly better perplexity for higher values of V. It is worth reminding that unlike CoSAC, both GDM and the Gibbs sampling based method requires the number of topics K be given.

3.C.3 Impact of α

Recall that, per the LDA model, topic proportions $\theta \sim \text{Dir}_K(\alpha)$. Cases with $\alpha > 1$ were previously shown (Nguyen, 2015) to exhibit slower convergence rates of the LDA's posterior estimation (via Gibbs sampler, for instance). Geometrically, large α implies that documents are more likely concentrated near the center of the topic simplex, leaving fewer documents near the vertices; this entails that geometric inference is more challenging. In our choices for parameters $\omega, \mathcal{R}, \lambda$ we relied on small values of α as a more practical scenario. Specifically, we considered $\omega = 0.8$ for this experiment to achieve full coverage of the topic simplex. In our previous experiments we set $\alpha = 0.1$. Now, we consider a larger range, $\alpha \in [0.01, 1.5]$, to gauge its impact more fully. Results are reported in Fig. 3.C.2(a), (b) and (c). For smaller values of α CoSAC is demonstrated to be the best algorithm of all under consideration. As α increases, CoSAC can still recover correct K with high accuracy, although the quality of topic estimates deteriorates faster than for Gibbs sampler and GDM. We think that further work on estimation procedures for topic radii R_ks (recall that topics are estimated as direction and length along this direction $b_k = R_k v_k$) might address this issue. In this work we considered maximum projection (Step 13 of Algorithm 3.2) to estimate R_ks, which might not be as accurate when documents are mostly near the center of the topic simplex (i.e., for higher α).

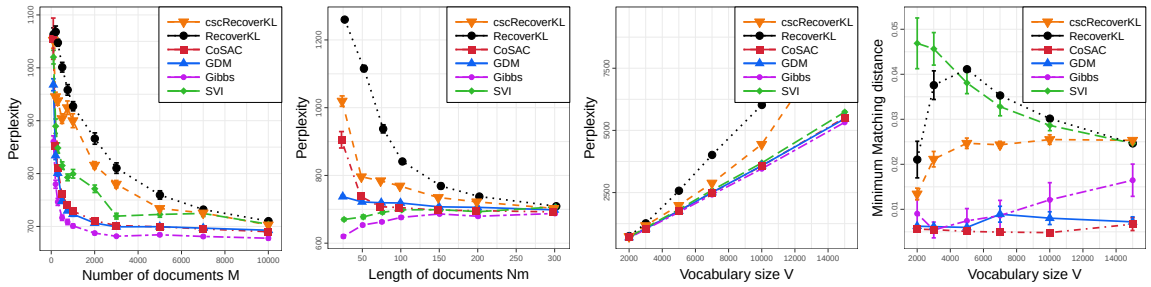


Figure 3.C.1: Perplexity for (a) varying corpora size, (b) varying length of documents, (c) varying vocabulary size; (d) Minimum matching Euclidean distance for varying vocabulary size.

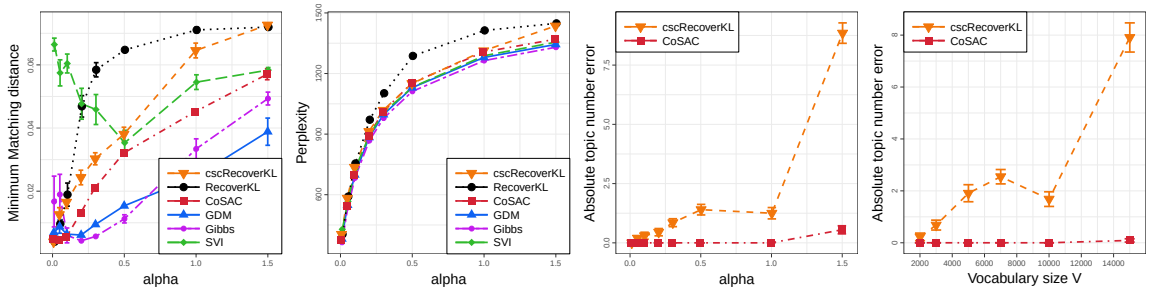


Figure 3.C.2: Varying α (a) Minimum matching Euclidean distance, (b) Perplexity, (c) Estimation of number of topics; (d) Estimation of number of topics for varying vocabulary size.

3.D Implementation details

In this section we give details about the implementations of the algorithms used in simulation studies and real data. We implemented Conic Scan-and-Cover (CoSAC) algorithm in Python with the help of Scipy (Jones et al., 2001–) sparse matrix modules. Geometric Dirichlet Means (GDM) was implemented with the help of Scikit-learn (Pedregosa et al., 2011) k-means implementation (with 10 restarts to avoid local minima of k-means) combined with a geometric correction technique. Codes for CoSAC and GDM are available at `https://github.com/moonfolk/Geometric-Topic-Modeling`. For RecoverKL (Arora et al., 2013) we applied code from one of the coauthors. To implement cscRecoverKL we used our CoSAC implementation (Algorithm 3.1 with outlier threshold λ as in Algorithm 3.2) to find anchor words and then recovery routine from the aforementioned code. For the Gibbs sampling (Griffiths & Steyvers, 2004) we used an lda package in Python that utilizes Cython to achieve C speed. Gibbs sampler was trained with $\alpha = 0.1$, $\eta = 0.01$ and 500 iterations in simulations studies and $\alpha = 0.1$, $\eta = 0.1$ and 1000 iterations in the NYTimes

articles[3] analysis. For the SVI (Hoffman et al., 2013) we used Gensim implementation (Řehůřek & Sojka, 2010) with automatic hyperparameters estimation, 50 iterations and 10 passes. Finally for HDP (Teh et al., 2006) we used C++ implementation with default hyperparameter settings and 100 iterations. For all experiments (except large vocabulary sizes and bigger α), per discussions in Sections 3.3.2 and 3.4, parameters of the CoSAC were set to $\omega = 0.6, n = 0.001M$ and $\mathcal{R}$ as median of the centered and normalized document norms. Spherical k-means post-processing step was run for 30 iterations. For cscRecoverKL we set $\omega = 0.4, \lambda = 0.015$ ($\lambda = 0.005$ for real data) and $\mathcal{R}$ as corresponding median of the norms. Note that cscRecoverKL takes word-to-word co-occurrence matrix as input, therefore sample size is V and "documents" are rows of this matrix. Exploring distributional properties of the simplex spanned by the anchor words is outside of the scope of this work, therefore parameter choices were made empirically based on the visual analysis illustrated by Fig. 3.2. All simulated results are reported after 20 repetitions of the data generation for each scenario and NYTimes results for LDA and HDP are reported over 10 refits of the corresponding Gibbs samplers.

[3] https://archive.ics.uci.edu/ml/datasets/bag+of+words

CHAPTER 4

Streaming dynamic and distributed inference of latent geometric structures

We develop new models and algorithms for learning about the temporal dynamics of the topic polytopes and related geometric objects that arise in topic model based inference. Our model is nonparametric and the corresponding inference algorithm is able to discover new topics as the time progresses. We establish the connection between the modeling of topic polytope evolution, Beta-Bernoulli process and the Hungarian matching algorithm. Our method is shown to be several orders of magnitude faster than existing scalable topic modeling approaches, as demonstrated by experiments working with several million documents in a dozen minutes.[1]

4.1 Introduction

A large number of unsupervised learning tasks associated with latent variable models can be fruitfully cast as the inference of a latent geometric structure arising from the model. Such geometric viewpoints often yield improved understanding about the inferential and computational behavior of the model. They also lead to more scalable inference and learning algorithms. When data and the associated modeling are indexed by time dimension, it is of interest to study the temporal dynamics of the latent geometric structure that arises. In this chapter, we focus on the modeling and algorithm for analyzing the temporal dynamics of *topic polytope*, a fundamental geometric object arising from the topic modeling literature.

The convex geometry of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and admixtures (Pritchard et al., 2000) was exploited to analyze the posterior contraction behavior of latent topics in both theory and practice (Nguyen, 2015; Tang et al., 2014). Chapter 2 studied a convex geometric interpretation of topic modeling that allowed for faster inference

[1] This work is submitted to ICML 2018

than variational inference and Gibbs sampling (Griffiths & Steyvers, 2004). Chapter 3 followed this line of work to arrive at a fast geometric algorithm for learning nonparametric topic models.

A number of authors have extended the basic topic modeling framework to analyze how topics evolve over time. The Dynamic Topic Models (DTM) (Blei & Lafferty, 2006b) demonstrated the importance of accounting for non-exchangeability between document groups, particularly when the time information is provided. Another approach is to fix the topics and only consider evolving topic popularity (Wang & McCallum, 2006). Hong et al. (2011) extended such an approach to multiple corpora. Ahmed & Xing (2012) proposed an interesting nonparametric construction extending DTM where topics can appear or eventually die out. The evolution of the latent geometric structure arising from the model (the topic polytope) is implicitly present in all these works, however it was not explicitly considered and analyzed.

Directly confronting the modeling and inference of the temporal dynamics of the topic polytope offers several opportunities and challenges. To start, what is the suitable space that we may consider the topic polytope to be represented? As topics evolve over time, so are the number of topics that may become active and dormant again, raising interesting choices about modeling of this evolution. Interesting issues arise in the inference, too. For instance, what is the principled way of mix-matching the vertices of a polytope to its reincarnation in the next time point? These challenges also open the door for directly modeling the geometric structure in a way that facilitates efficient inference.

In particular, in this work we shall utilize an isometric embedding of the unit sphere in the word simplex, so that the evolution of topic polytopes may be modeled by a collection of (random) trajectories of points residing on the unit sphere. Instead of worrying about mix-matching vertices in an ad hoc fashion, we appeal to the Bayesian nonparametric modeling framework that allow the number of topic vertices to be random and vary across time. The mix-matching between topic variables shall be guided by the assumption on the smoothness of the collection of global trajectories on the sphere using Von Mises-Fisher dynamics (Mardia & Jupp, 2009). The selection of active topics at each time point will be guided by a nonparametric prior on the random binary matrices via the (hierarchical) Beta-Bernoulli process (Thibaux & Jordan, 2007).

To this end, we construct a sequence of Bayesian nonparametric models in increasing levels of complexity: the simpler model captures the temporal dynamics of a sequence of topic polytopes, while the full model describes the temporal dynamics of a collection of topic polytopes as they arise from multiple corpora. The semantic of topics that arise from our models can be interpreted as follows: there is a latent collection of global topics of

unknown cardinality that evolve over time (e.g. topics in science or social topics in Twitter) and each year (or day) a subset of the global topics is elucidated by the community (i.e. some topics may be dormant at a given time point). The nature of each global topic may change smoothly (via varying word frequencies). Additionally, different subsets of global topics are associated with different groups (e.g. journals or Twitter location stamps), some become active and inactive over time.

It is interesting to note that our model construction leads to a suite of approximate inference algorithm that scales well in an online and distributed setting. In particular, the online MAP update of the latent topic polytope can be viewed as solving an optimal matching problem for which a fast Hungarian matching algorithm can be applied. Our approach is able to perform dynamic nonparametric topic inference on 3 million documents in 12 minutes, which is significantly faster than prior static online and/or distributed topic modeling algorithms (Newman et al., 2008; Hoffman et al., 2010; Wang et al., 2011; Bryant & Sudderth, 2012; Broderick et al., 2013).

The remainder of the chapter is organized as follows. In Section 4.2 we propose an approach for defining a Markovian process over the space of topic polytopes (simplices). In Section 4.3 we present a series of models for polytope dynamics across time and over a collection of corpora. In Section 4.4 we describe our algorithms for online dynamic and/or distributed inference. Section 4.5 demonstrates experimental results. We conclude with a discussion in Section 4.6.

4.2 Temporal dynamics of a topic polytope

A fundamental object of inference in this work is the topic polytope arising in topic modeling (Blei et al., 2003; Nguyen, 2015). Given a vocabulary of V words, a topic is defined as a probability distribution on the vocabulary. Thus a topic is taken to be a point in the vocabulary simplex, namely, Δ^{V-1}, and a topic polytope for a corpus of documents is defined as a convex hull of topics associated with the documents. Geometrically, the topics correspond to the vertices (extreme points) of the (latent) topic polytope to be inferred from data.

In order to infer about the temporal dynamics of a topic polytope, one might consider the evolution of each topic variable, say $\theta^{(t)}$, which represents a vertex of the polytope at time t. A standard approach is due to Blei & Lafferty (2006b), who proposed to use a Gaussian Markov chain $\theta^{(t)}|\theta^{(t-1)} \sim \mathcal{N}(\theta^{(t-1)}, \sigma I)$ in $\mathbb{R}^V$ for modeling temporal dynamics and a logistic normal transformation $\pi(\theta^{(t)})_i := \frac{\exp(\theta_i^{(t)})}{\sum_i \exp(\theta_i^{(t)})}$ for mapping into Δ^{V-1}. This popular construction does carry some disadvantages: the mapping into Δ^{V-1} is many-to-one, and

while this can be remedied by a suitable reparameterization, the domain of the preimage is noncompact ($\mathbb{R}^V$), resulting in inefficiency of inference.

Motivated by the directional representation of topic polytope studied in Chapter 3, we shall represent each topic variable as a point in a unit sphere $\mathbb{S}^{V-2}$, which possesses a natural isometric embedding in the vocabulary simplex Δ^{V-1}, so that the temporal dynamics of a topic variable can be visualized as a (random) trajectory on $\mathbb{S}^{V-2}$. This trajectory shall be modeled as a Markovian process on $\mathbb{S}^{V-2}$: $\theta^{(t)}|\theta^{(t-1)} \sim \mathrm{vMF}(\theta^{(t-1)}, \tau_0)$. Von Mises-Fisher (vMF) distribution is commonly used in the field of directional statistics (Mardia & Jupp, 2009) to model points on a unit sphere. Its likelihood is proportional to the cosine similarity, a similarity measure popular in text mining applications (Feldman & Sanger, 2007).

Isometric embedding of $\mathbb{S}^{V-2}$ in vocabulary simplex We recall the directional representation of topic polytope from Chapter 3: let $B = \{\beta_1, \ldots, \beta_K\}$ be a collection of vertices of a topic polytope. Each vertex is represented as $\beta_k := C + R_k v_k$, where $C \in \mathrm{Conv}(B)$ is a reference point in a convex hull of B, $v_k \in \mathbb{R}^V$ is a topic direction and $R_k \in \mathbb{R}_+$. Moreover, $R_k \in [0, 1]$ is determined so that the tip of direction vector v_k resides on the boundary of Δ^{V-1}. Since the effective dimensionality of v_k is $V - 2$, we can now define a one-to-one and isometric map taking v_k onto $\mathbb{S}^{V-2}$ in the following way: map of the vocabulary simplex $\Delta^{V-1} \in \mathbb{R}^V$ where it is first translated so that C becomes an origin and then rotated into $\mathbb{R}^{V-1}$, where resulting topics, say $\theta_1, \ldots, \theta_K \in \mathbb{S}^{V-2}$, are normalized to the unit length. Observe that this geometric map is an isometry and hence invertible. It also preserves angles between vectors, therefore we can evaluate vMF density without performing the map explicitly, by simply setting $\theta_k := \frac{\beta_k - C}{\|\beta_k - C\|}$.

Figure 4.1 provides a geometric illustration. Let $V = 3$ and vocabulary simplex be a red triangle. Two topics on the boundary (face) of the vocabulary simplex are $\beta_1 = C + v_1$ and $\beta_2 = C + v_2$. Green dot C is the reference point and $\alpha = \angle(v_1, v_2)$. In Fig. 4.1 (left) we translate C into the origin and rotate vocabulary simplex from xyz to xy plane. In Fig. 4.1 (center) we show resulting position of the vocabulary simplex and add a unit sphere (blue) in $\mathbb{R}^2$. Corresponding to β_1, β_2 topics are the points θ_1, θ_2 on the sphere with $\angle(\theta_1, \theta_2) = \alpha$. Now we apply the inverse translation and rotation to *both* vocabulary simplex and unit sphere. Result is shown in Fig. 4.1 (right) — we are back to $\mathbb{R}^3$ and $\angle(\theta_1, \theta_2) = \angle(v_1, v_2) = \alpha$, where $\theta_k = \frac{\beta_k - C}{\|\beta_k - C\|_2}$. In Fig. 4.2 we give an example of the dynamics.

In the preceeding paragraphs, the evolution of each topic variable is modeled as a random trajectory residing in a unit sphere, and so the evolution of a collection of topics can be modeled by a collection of corresponding trajectories on the sphere. A modeling challenge is how to account for the fact that the number of "active" topics may be unknown and

vary over time. Moreover, a topic may be activated, become dormant, and then resurface after some time. A natural model-based solution for such situations is to employ Bayesian nonparametric modeling elements and adapt them to handle the geometric structures of our inferential interest. This is the focus of the next section.

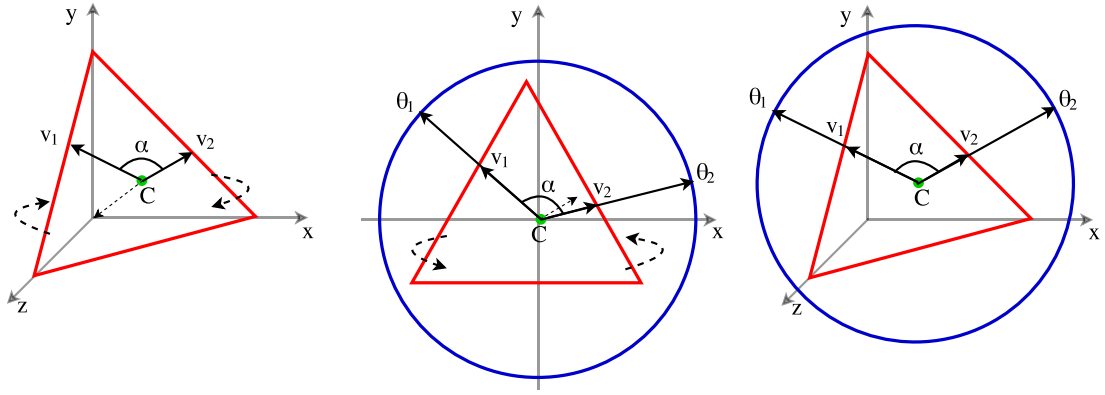


Figure 4.1: Invertible transformation between unit sphere and a standard simplex

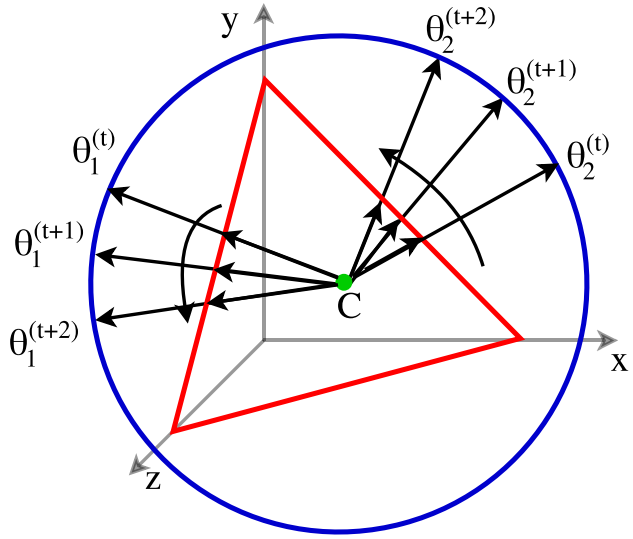


Figure 4.2: Topic dynamics

4.3 Hierarchical Bayesian modeling for one or multiple topic polytopes

In this section we present a sequence of models with increasing levels of complexity for capturing various aspects of topic modeling in online and distributed settings. We start by proposing a hierarchical model for online learning of the temporal dynamics of a single topic polytope, allowing for varying number of vertices over time. Then we show how to borrow strength from *multiple* topic polytopes learned on different corpora drawing on a common pool of global topics. We conclude with a construction of a "full" model for modeling evolution of global topics across time and across groups of corpora.

4.3.1 Dynamic model for single topic polytope

At a high level, our model contains a collection of global trajectories taking values on a unit sphere. Each trajectory shall be endowed with a Von Mises-Fisher dynamic described in the previous section. At each time point, a random topic polytope is constructed by selecting a (random) subset of points on the trajectory evaluated at time t. The random selection is guided by a Beta-Bernoulli process prior (Thibaux & Jordan, 2007). This construction is motivated by a modeling technique of Nguyen (2010), who studied a hierarchical Dirichlet process type model for inference of smooth trajectories on an Euclidean domain. Our model using Beta-Bernoulli process, as opposed to the Dirichlet-type priors, as a building block is more appropriate for the purpose of topic discovery. Due to the isometric embedding of $\mathbb{S}^{V-2}$ in Δ^{V-1} described in the previous section, from here on we shall refer to topics as points on $\mathbb{S}^{V-2}$.

First, generate an unbounded collection of global topic trajectories using Beta process prior (cf. Thibaux & Jordan (2007)):

$$Q|\gamma_0, H \sim \text{BP}(\gamma_0, H), \tag{4.1}$$

where H is a base measure on the space of trajectories on $\mathbb{S}^{V-2}$, γ_0 the concentration parameter of the Beta process. It follows that

$$Q = \sum_i q_i \delta_{\theta_i},$$

where each $\theta_i \sim H$ is a sequence of T random elements on the unit sphere $\theta_i := \{\theta_i^{(t)}\}_{t=1}^T$,

which are generated as follows:

$$\begin{aligned} \theta_i^{(t)}|\theta_i^{(t-1)} &\sim \text{vMF}(\theta_i^{(t-1)}, \tau_0) \text{ for } t = 1, \ldots, T \text{ and} \\ \theta_i^{(0)} &\sim \text{vMF}(\cdot, 0) \text{ – uniform on } \mathbb{S}^{V-2}. \end{aligned} \tag{4.2}$$

At any given time $t = 1, \ldots, T$, the process Q induces a marginal measure Q_t, whose support is given by the atoms of Q as they are evaluated at time t. Now, select a subset of the global topics that are active at t via the Bernoulli process

$$T^{(t)}|Q_t \sim \text{BeP}(Q_t). \tag{4.3}$$

This means that $T^{(t)} := \sum_i b_i^{(t)} \delta_{\theta_i^{(t)}}$, where

$$b_i^{(t)}|q_i \sim \text{Bern}(q_i), \ \forall i. \tag{4.4}$$

We can also view $T^{(t)}$ as the collection of supporting atoms $\{\theta_i^{(t)} : b_i^{(t)} = 1, i = 1, 2, \ldots\}$, which represent the topics that are active at time t. Finally, assume that noisy measurements of each of these topic variables are generated via:

$$v_k^{(t)}|T^{(t)} \sim \text{vMF}(T_k^{(t)}, \tau_1) \text{ for } k = 1, \ldots, K^{(t)}, \tag{4.5}$$

where $K^{(t)} := \text{card}(T^{(t)})$, number of topics active at t.

The noisy estimates for the topics at any particular time point may come from either the global topics observed until the previous time point or a so far unexplored topic. The collection of random variables that link up the observed noisy estimates at any time point to the global topics observed thus far is of interest. In that regard, let $B^{(t)} = ((B_{ik}^{(t)}))$ denote the binary matrix representing the presence of connection between observed topic estimates and global topics at time point t, i.e, $B_{ik}^{(t)} = 1$ if the vector $v_k^{(t)}$ is a noisy estimate for $\theta_i^{(t)}$.

The joint posterior probability for the hidden topics $\theta^{(t)}$ given observed noisy topics $v^{(t)}$ is:

$$\begin{aligned} &\mathbb{P}\left(\theta^{(0)}, \{\theta^{(t)}, B^{(t)}\}_{t=1}^T | \{v^{(t)}\}_{t=1}^T\right) \propto \\ &\mathbb{P}(\theta^{(0)}) \prod_{t=1}^T \mathbb{P}(\theta^{(t)}, B^{(t)}|\theta^{(t-1)}, \{B^{(a)}\}_{a=1}^{t-1}) \mathbb{P}(v^{(t)}|\theta^{(t)}, B^{(t)}). \end{aligned}$$

Now, using probabilistic calculations, we get

$$\begin{aligned}
&\mathbb{P}(\theta^{(t)}, B^{(t)}|\theta^{(t-1)}, \{B^{(a)}\}_{a=1}^{t-1})\mathbb{P}(v^{(t)}|\theta^{(t)}, B^{(t)}) \propto \\
&\mathbb{P}(\theta^{(t)}, B^{(t)}|\theta^{(t-1)}, v^{(t)}, \{B^{(a)}\}_{a=1}^{t-1}) \propto \\
&\prod_{i=1}^{L_{t-1}} \left(\frac{m_i^{(t-1)}}{t - m_i^{(t-1)}} \right)^{\sum_{k=1}^{K^{(t)}} B_{ik}^{(t)}} c(\tau_0) \exp(\tau_0 \langle \theta_i^{(t-1)}, \theta_i^{(t)} \rangle) \\
&\frac{\exp(-\frac{\gamma_0}{t})(\gamma_0/t)^{L_t - L_{t-1}}}{(L_t - L_{t-1})!} \exp(\tau_1 \sum_{i=1}^{L_t} \sum_{k=1}^{K^{(t)}} B_{ik}^{(t)} \langle \theta_i^{(t)}, v_k^{(t)} \rangle),
\end{aligned} \tag{4.6}$$

where $m_i^{(t)}$ is the number of occurrences of topic i up to time point t — cf. popularity of a dish in the Indian Buffet Process (IBP) metaphor (Ghahramani & Griffiths, 2005) and L_t is the number of global topics at time t. Inner product comes from the vMF prior in Eq. (4.2) and $c(\tau_0)$ is the corresponding normalizing constant.

Consider next a scenario for online inference for which we are interested in a sequential posterior formulation. To this end, we express the logarithm of this posterior distribution in a form of a *matching* problem.

$$\begin{aligned}
&\log(\mathbb{P}(\theta^{(t)}, L_t, B^{(t)}|\theta^{(t-1)}, v^{(t)}, \{B^{(a)}\}_{a=1}^{t-1})) = \\
&\sum_{i=1}^{L_{t-1}} \sum_k B_{ik}^{(t)} [(\log \frac{m_i^{(t-1)}}{t - m_i^{(t-1)}}) + \langle \tau_1 v_k^{(t)} + \tau_0 \theta_i^{(t-1)}, \theta_i^{(t)} \rangle] \\
&+ \sum_{i=L_{t-1}+2}^{L_t} \sum_k B_{ik}^{(t)} \{\tau_1 + c_0 + \log(\gamma_0/t) - \log(i - L_{t-1})! \\
&+ \log(i - L_{t-1} - 1)!\} \\
&+ \sum_k B_{(L_{t-1}+1)k}^{(t)} \{\tau_1 + c_0 + \log(\gamma_0/t)\} + C,
\end{aligned} \tag{4.7}$$

where c_0 is the log inverse volume of a unit sphere - a term coming from the uniform prior on $\theta^{(0)}$ in Eq. (4.2), and C is a constant. We highlight the connection of the above expression to the objective of an optimal *matching* problem: given a cost matrix workers should be assigned to tasks, at most one worker per task and one task per worker. We see from Eq. (4.7) that $B(t)$ can be viewed as an assignment matrix and each new topic $v_k^{(t)}$ should be assigned to either one of the global topics $\theta_i^{(t)}$ or start a new topic with cost for the later option guided by the prior. We interpret such assignment as topic discovery; details of the inference algorithm are deferred to Section 4.4.

4.3.2 Hierarchical Beta process for multiple topic polytopes

We take our modeling in a different direction, by allowing the presence of multiple corpora, each of which maintains its own topic polytope. Large text corpora often can be partitioned based on some grouping criteria, e.g. scientific papers by journals, news by different media agencies or tweets by location stamp. In this subsection we model the collection of topic polytopes observed at a single time point by employing the Hierarchical Beta Process prior (HBP) (Thibaux & Jordan, 2007). The modeling of temporal dynamics of a collection of polytopes will be described in the following subsection.

First, generate distribution of global topics Q as in Eq. (4.1). Since we are interested only in a single time point in this subsection, the base measure H is simply a $\mathrm{vMF}(\cdot, 0)$, the uniform distribution over $\mathbb{S}^{V-2}$. Next, for each group $j = 1, \ldots, J$, generate a group specific distribution over topics:

$$\begin{aligned} G_j|Q &\sim \mathrm{BP}(\gamma_j, Q) \\ G_j &:= \sum_i p_{ji}\delta_{\theta_i}, \end{aligned} \tag{4.8}$$

where p_{ji} vary around corresponding q_i. The distributional properties of p_{ij} are described in Thibaux & Jordan (2007). We now proceed similarly to Eq. (4.4):

$$\begin{aligned} T_j|G_j &\sim \mathrm{BeP}(G_j) \\ T_j &:= \sum_i b_{ji}\delta_{\theta_i}, \text{ where} \\ b_{ji}|p_{ji} &\sim \mathrm{Bern}(p_{ji}),\ \forall i. \end{aligned} \tag{4.9}$$

Notice that each group $T_j := \{\theta_i : b_{ji} = 1, i = 1, 2, \ldots\}$ selects only a subset from the collection of global topics, which is consistent with the idea of partitioning by journals: some topics of ICML are not represented in KDD and vice a versa. The last step is analogous to Eq. (4.5):

$$v_{jk}|T_j \sim \mathrm{vMF}(T_{jk}, \tau_1) \text{ for } k = 1, \ldots, K_j, \tag{4.10}$$

where $K_j := \mathrm{card}(T_j)$ – number of topics present in group j.

We again define matrix B as the binary matrix representing the assignment of global topics to the noisy topic estimates, i.e., $B_{ijk} = 1$ if the k^{th} topic estimate for group j arises as a noisy estimate of global topic θ_i. The *matching* problem is now different: we don't have any information about the global topics as there is no history, however instead we should match a *collection* of topic polytopes to a global topic polytope.

It can be seen from Thibaux & Jordan (2007) that the matrix of topic assignments is

distributed a priori by an Indian Buffet Process (IBP) with parameter γ_0. The posterior probability for global topics θ_i and assignment matrix B given the topic estimates v_{jk} has the following form:

$$\mathbb{P}(B, \theta|v) \propto \exp(\tau_1 \sum_{i,j,k} B_{jik}\langle\theta_i, v_{jk}\rangle) \tag{4.11}$$
$$\frac{(\gamma_0 c_0)^L}{f(L)} \exp(-\gamma_0 \sum_{j=1}^{J} 1/j) \prod_{i=1}^{L} \frac{(J-m_i)!(m_i-1)!}{J!},$$

where L is any upper bound on the number of global topics, i.e. $L = \sum_{j=1}^{J} K_j$ and the value of $f(L)$ depends on the partitions of L. Similarly to previous case, m_i is the popularity of global topic i after the assignment ((-1)! = 0 when $m_i = 0$).

4.3.3 Dynamic hierarchical Beta process

We are now ready to present Dynamic Hierarchical Beta Process model (dHBP) which combines the constructions described in subsections 4.3.1 and 4.3.2 to enable inference of the temporal dynamics of collections of topic polytopes. We start by specifying the upper level Beta process given by Eq. (4.1) and base measure H given by Eq. (4.2). Next, for each group $j = 1, \ldots, J$, generate a group specific distribution over topics:

$$\begin{aligned} Q_j|Q &\sim \text{BP}(\gamma_j, Q) \\ Q_j &:= \sum_i p_{ji}\delta_{\theta_i}. \end{aligned} \tag{4.12}$$

At any given time t, each group j selects a subset from the common pool of global topics:

$$\begin{aligned} T_j^{(t)}|Q_{jt} &\sim \text{BeP}(Q_{jt}) \\ T_j^{(t)} &:= \sum_i b_{ji}^{(t)}\delta_{\theta_i^{(t)}}, \text{ where} \\ b_{ji}^{(t)}|p_{ji} &\sim \text{Bern}(p_{ji}),\ \forall i. \end{aligned} \tag{4.13}$$

Let $T_j^{(t)} := \{\theta_i^{(t)} : b_{ji}^{(t)} = 1, i = 1, 2, \ldots\}$ be the corresponding collection of atoms – topics active at time t in group j. Noisy measurements of these topics are generated by:

$$v_{jk}^{(t)}|T_j^{(t)} \sim \text{vMF}(T_{jk}^{(t)}, \tau_1) \text{ for } k = 1, \ldots, K_j^{(t)}, \tag{4.14}$$

where $K_j^{(t)} := \text{card}(T_j^{(t)})$, the number of topics active at time t in group j.

As in Section 4.3.1 we are interested in the posterior distribution of the global topics at time t given the state of the global topics at time $t-1$:

$$\begin{aligned}&\mathbb{P}(\theta^{(t)}, L_t, B^{(t)}|\theta^{(t-1)}, v^{(t)}) \propto \\ &\exp(\tau_0 \sum_i \langle \theta_i^{(t)}, \theta_i^{(t-1)} \rangle) \\ &\cdot \exp\left(\sum_{i,j,k} \tau_1 B_{jik}^{(t)} \langle \theta_i^{(t)}, v_{jk}^{(t)} \rangle\right) F(L_t, L_{t-1}, \{m_{ji}^{(t)}\}),\end{aligned} \tag{4.15}$$

where $F(L_t, L_{t-1}, \{m_{ji}^{(t)}\})$ is a function dependent on the count $m_{ji}^{(t)}$ for topic i, groups j and up to time point t. Analogous to the Chinese Restaurant Franchise (Teh et al., 2006), one can think of an Indian Buffet Franchise in the case of HBP. A headquarter buffet provides some dishes each day and the local branches serve a subset of those dishes. Although this analogy seems intuitive, we are not aware of a corresponding Gibbs sampler and it remains to be a question of future studies. Therefore we are not able to handle the term $F(L_t, L_{t-1}, \{m_{ij}^{(t)}\})$ directly and propose a heuristic replacement — stripping away popularity of topics across groups and only considering group specific topic popularity (groups still remain dependent through the atom locations). The resulting approximate inference algorithm is given in the next section.

4.4 Streaming dynamic distributed inference

Here we present our algorithm for Streaming Dynamic Distributed (SDD) topic estimation. In this work we focus on modeling topic polytopes and the first step is to obtain such polytopes. To do so we partition data into batches according to available group and/or time information and utilize the very fast Conic Scan-and-Cover algorithm (see Chapter 3) to obtain polytope estimates for each of the batches in a trivially parallel fashion. To perform polytope-to-sphere transformation we use our construction of Section 4.2 using estimates of word marginal probabilities as the reference point C. Such estimate is simply a data mean which can be updated online. After the transformation we obtain the collections $v_j^{(t)}$ for $t = 1, \ldots, T$ and $j = 1, \ldots, J$.

Our primary practical interests are the questions of topic discovery and evolution over time. To address the topic discovery question we want to identify a point in time where topic appeared first and to interpret the dynamics we want to estimate probabilities for each of the words in each topic at every time step. To this end we focus our attention on Maximum Aposteriori (MAP) estimation based on our posterior distribution calculations of Section

4.3.

Streaming Dynamic Matching (SDM) We have already laid out the foundation for the inference algorithm through the cost representation of the log posterior probability in Eq. (4.7). The remaining part is the estimation of $\theta^{(t)}$. We can readily obtain a MAP estimate of $\theta^{(t)}$ given the assignment matrix $B^{(t)}$, thanks to the self-conjugacy of vMF distribution:

$$\hat{\theta}_i^{(t)} | B_{ik}^{(t)} = 1, \theta_i^{(t-1)} = \frac{\tau_0 \theta_i^{(t-1)} + \tau_1 v_k^{(t)}}{\|\tau_0 \theta_i^{(t-1)} + \tau_1 v_k^{(t)}\|_2} \tag{4.16}$$

We rewrite parametric part of Eq. (4.7) as

$$\sum_{i,k} B_{ik}^{(t)} \mathrm{R}(i,k), \text{ where} \tag{4.17}$$
$$\mathrm{R}(i,k) = \log \frac{m_i^{(t-1)}}{t - m_i^{(t-1)}} + \|\tau_0 \theta_i^{(t-1)} + \tau_1 v_k^{(t)}\|_2.$$

Now it is evident that we can maximize Eq. (4.17) using Hungarian algorithm (Kuhn, 1955). Recall that we only considered parametric part of the problem. It is possible to extend to nonparametric estimation using the Hungarian algorithm based on Eq. (4.7). However, due to large values of c_0 (volume of sphere surface goes to 0 in high dimensions), the discovery of new topics is over-encouraged. We instead adopt a geometrically intuitive thresholding — if for $B_{ik}^{(t)} = 1$, $\cos(v_k^{(t)}, \theta_i^{(t-1)}) < c$, then let $v_k^{(t)}$ be a new topic. We also note that reward term $\mathrm{R}(i,k)$ is closely related to the cosine distance between corresponding topics, supporting the geometric interpretation of our approach.

Distributed Matching (DM) To obtain the MAP estimate via Eq. (4.11) we again analyze the parametric subproblem in the form admissible to the Hungarian algorithm:

$$\begin{aligned} \sum_j \sum_{i,k} B_{jik} (\tau_1 (\|v_{jk} + \sum_{-j,i,k} B_{-jik} v_{-jk}\|_2 \\ - \|\sum_{-j,i,k} B_{-jik} v_{-jk}\|_2) + \log m_{-ji}), \end{aligned} \tag{4.18}$$

where m_{-ji} is the popularity of global topic i outside of group j. The problem of matching is across groups and topics and is not amendable to closed form Hungarian algorithm. However based on Eq. (4.18) we see that for a fixed group the assignment optimization reduces to a case for the Hungarian algorithm. We utilize this to perform iterative updates for different j, which guarantees convergence to a local optimum. Topics are assigned as

new based on same argument as for the SDM case.

Streaming Dynamic Distributed Matching (SDDM) Finally we combine our results to perform approximate inference of the model in Section 4.3.3. The assignment for a single group j is done based on the following objective:

$$\sum_{i,k} B^{(t)}_{jik} \Bigg((\|\tau_1 v^{(t)}_{jk} + \tau_1 \sum_{-j,i,k} B^{(t)}_{-jik} v^{(t)}_{-jk} + \tau_0 \theta^{(t-1)}_i \|_2 \\ - \| \sum_{-j,i,k} B^{(t)}_{-jik} v^{(t)}_{-jk} + \tau_0 \theta^{(t-1)}_i \|_2) + \log m^{(t)}_{ji} \Bigg), \tag{4.19}$$

where $m^{(t)}_{ji}$ denotes the popularity of topic i in group j at time t – a heuristic term replacing the Indian Buffet Franchise term.

4.5 Experiments

The objective of the experiments is to demonstrate the learning of latent temporal dynamics and topic discovery, the ability to perform learning in a distributed and streaming fashion, and to demonstrate scalability of our approaches. We analyze two datasets: Early Journal Content (EJC) available from JSTOR[2] and collection of Wikipedia articles partitioned by categories and in time according to their popularity.

4.5.1 Early Journal Content

Early Journal Content dataset spans years from 1665 up to 1922. Years before 1882 contain very few articles and we aggregated them into a single timepoint. After preprocessing, dataset has 400k scientific articles from over 400 unique journals. Vocabulary was truncated to 4516 words. We set all articles from the last available year (1922) aside for the testing purposes. We compare three variations of our model with CoSAC and parametric models such as Stochastic Variational Bayes (SVB) (Broderick et al., 2013) and Dynamic Topic Models (DTM) (Blei & Lafferty, 2006b) trained with 100 topics. Perplexity scores on the held out data, training times, computing resources and number of topics are reported in Table 4.1. SDM achieves the best perplexity, while SDDM outperforms DM, which suggests that modeling time is beneficial. For this dataset modeling groups negatively effects perplexity which may be due to majority of the groups having very few articles (i.e. less than 100) - a

[2] http://www.jstor.org/dfr/about/sample-datasets

setup challenging for many topic modeling algorithms. We report details about parameter settings in the Supplement. Next we present a case study of a topic based on the SDM results.

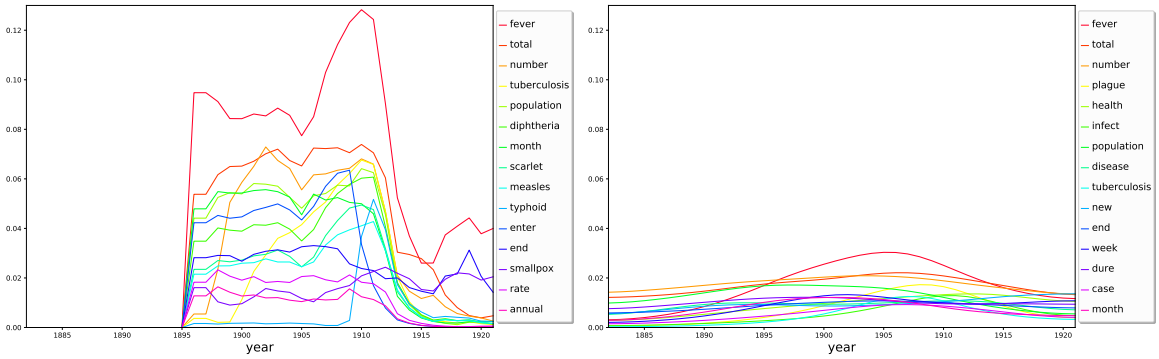
Case study: epidemics. The beginning of the 20th century is known to have a vast history of disease epidemics of various kinds. Vaccines or effective treatments against the majority of them were developed shortly after. Examples of such disease are smallpox, typhoid, yellow fever and scarlet fever. One of the journals represented in the EJC dataset is the "Public Health Report", however publications from it are only available starting 1896. One of the primary objectives of the journal was to reflect epidemic disease infections. As one of the objectives of our modeling approach is to do topic discovery, it is interesting to see if the model can discover an epidemics related topic around 1896.

Figure 4.3a shows that our model correctly discovered a new topic is 1896 semantically related to epidemics. We plot the evolution of probabilities of the top 15 words in this topic across time. We observe that word "typhoid" increases in probability towards 1910 in the "epidemics" topic, which aligns with historical events such as Typhoid Mary in 1907 and chlorination of public drinking water in the US in 1908 for controlling the typhoid fever. The probability of "tuberculosis" increases too aligning with foundation of the National Association for the Study and Prevention of Tuberculosis in 1904.

Another interesting aspect is comparison to the analogous topics of the DTM (shown in Fig. 4.3b). DTM curves are smoother because DTM is based on Gaussian Kalman filtering with forward-backward updates, which induces smoother curves. In our model we consider an online learning setup and hence we are only doing a single forward path. It is unclear which approach is better — smoothing (i.e. backward updates) allows to better account for uncertainty, which can be beneficial for smaller data. However it tempers with the notion of new topic discovery as it spreads high probability words over time and results in almost flat curves.

Table 4.1: Modeling topics of EJC

	Perplexity	Time	Topics	Cores used
SDM	**1181**	24min	124	1
DM	1323	8min	128	20
SDDM	1289	**3.2min**	88	20
DTM	1194	56hours	100	1
SVB	1840	3hours	100	20
CoSAC	1191	51min	132	1



(a) SDM (b) DTM

Figure 4.3: Epidemics: evolution of top 15 words

4.5.2 Streaming Wiki corpus

We have collected articles from Wikipedia together with their page view counts for the 12 months of 2017 and category information (e.g., Arts, History, Health, etc.). We used categories as groups and partitioned the data across time according to the page view counts. We provide detailed description of the dataset construction in the Supplement. Total number of documents is 3 million and we reduced vocabulary to 7359 words (similar to Hoffman et al. (2010)). In Table 4.2 we report training times and perplexity on held out documents from category Art from December 2017. SDDM shows the best performance both in terms of time and perplexity, which supports the idea of modeling both group and time information. DM ignores the time component and produces worst results as expected.

Table 4.2: Modeling Wikipedia articles

	Perplexity	Time	Topics	Cores used
SDM	1236	35min	182	20
DM	1260	14min	183	20
SDDM	**1228**	**12min**	184	20

4.6 Discussion

Now that we presented our models and algorithms, we will discuss some additional related literature and then possible applications of our modeling approach outside of topic modeling.

4.6.1 Related literature

The idea of using time series as base measure in the Bayesian nonparametric settings was previously explored by Nguyen (2010) in the context of Hierarchical Dirichlet Process (HDP) (Teh et al., 2006). His approach is suitable for modeling evolving clustering structure of time data, however mixture based models are not appropriate for the modeling of polytopes. Polytopes (i.e. topics) are collections of distinct points and therefore should originate from distinct vertices of the global polytope. Mixture model based inference violates this constraint as multiple vertices of a polytope from a single batch may be assigned to same vertex of the global polytope. We showed that Beta Process based constructions are a better approach for modeling geometric object as they account for the desired "distinctiveness" and lead to a natural matching based inference.

Von Mises-Fisher distribution was previously used as a prior for topics *and* documents represented as normalized tf-idf scores (Rasiwasia & Vasconcelos, 2013). Such representation of documents strips away important geometric information (i.e. norms of documents) and in our work we aimed to maintain a more classical perspective, where topics represent word probabilities and documents are generated using multinomial distribution, whereas vMF distribution is used to model the latent dynamics.

Fox et al. (2009) utilized Beta-Bernoulli process in the time series modeling to incorporate notion of switching regimes of an autoregressive process and the corresponding Indian Buffet Process was used to select subsets of the latent states of the Hidden Markov Model.

4.6.2 Geometric inference beyond topic modeling

In this work we have demonstrated several modeling approaches of polytopes in nonparametric, online, dynamic, distributed fashion. Sophisticated Bayesian nonparametric models are crucial for complicated large-scale data. Unfortunately the data size and modeling are at crossroads — training sophisticated models on large data is often infeasible. Our work suggests a new approach for nonparametric learning at large scale. We note that the latent geometric structure of interest is solely encoded in the base measure H (recall Eq. (4.1)). It is of interest to continue exploring choices of H for other geometric structures such as collections of k-means centroids, principal components, etc. Once an appropriate base measure is constructed, our Beta process based models can be utilized to do dynamic and distributed learning amendable to large data sizes.

Appendix

4.A Experiments details

Here we provide all hyperparameter settings for the experiments. The Dynamic Topic Model (Blei & Lafferty, 2006b) was trained using code from https://github.com/blei-lab/dtm with default parameter settings and $K = 100$. In 56 hours we were able to complete LDA initialization and two EM iterations of the dynamic updates. Streaming Variational Bayes (Broderick et al., 2013) was trained based on the code from https://github.com/tbroderick/streaming_vb with default parameters, $K = 100$, batch size of 1024 and 4 asynchronous batches per evaluation. Algorithm ran noticeably slower on our computing cluster using 20 cores than reported in the paper and we only could complete 2 EM iterations per batch in a reasonable amount of time, which perhaps explains the poor perplexity score. For CoSAC, we used code from https://github.com/moonfolk/Geometric-Topic-Modeling with default parameters and $\omega = 0.7$. Same setting of CoSAC was used for learning topics on batches in this framework.

For our algorithms we chose cosine threshold $c = -0.1$, which corresponds to a new topic being discovered when its best matching global topic is slightly above 90° apart from the batch topic, which is an intuitive choice geometrically. The role of τ_0 and τ_1 is two fold — their total magnitude controls the effect of the prior. Larger values mean less influence of the topic popularity counts. For SDM and SDDM, τ_0 influences the rate of the dynamics — the larger values imply slower dynamics. For SDM and SDDM we set $\tau_0 = 10$ and $\tau_1 = 2$, and for DM we set $\tau_1 = 10$.

4.B Datasets

4.B.1 Early Journal Content

For the EJC dataset we get from http://www.jstor.org, the data is well-structured and the preprocessed 1-gram format is available along with the corresponding meta data in xml format for each document. We perform stemming on every word shown in the dataset and remove all stop words given in ENGLISH_STOP_WORDS from sklearn.feature_extraction.stop_words and nltk.corpus.stopwords.words. Words the length of which are shorter than 3 are removed. We also remove those words that appear in more than 99% of the documents and those appearing in less than 1% of the documents. For documents, we remove those *outliers*, in which a same word appears

more than 200 times. Those documents which contain less than 100 words (considering only words in the preprocessed vocabulary) are also excluded. After preprocessing, there are 4516 words in the vocabulary and approximately 400k documents left. We batch the documents based on both of their publication years and the journals they are published in.

4.B.2 Wiki

For Wikipedia data, we need four main components: *the vocabulary*, *the original Wikipedia page texts*, *the page view counts* (for each Wikipedia page being considered), and the *category-title mapping*. The data acquisition and processing for each component is described separately as follows.

The vocabulary We take the vocabulary from Wiktionary:Frequency_lists. Originally there are 10000 words. We remove words shorter than 3 characters and remove all stop words given in ENGLISH_STOP_WORDS from `sklearn.feature_extraction.stop_words` and `nltk.corpus.stopwords.words` After preprocessing, there are 7359 words left in the vocabulary.

The original Wikipedia page texts We download the Wiki data dumps (2017/08/20) from `https://meta.wikimedia.org/wiki/Data_dump_torrents` containing about 9 million Wikipedia pages after decompression. We split the whole file into multiple text files in which each individual file contains the content of a single Wikipedia page. We then use *MeTA*, a modern C++ data sciences toolkit to transform all of these raw texts files into 1-gram format using the preprocessed vocabulary.

The page view counts We use the Pageview API provided by AQS (Analytics Query Service) to get the page view count information of each Wikipedia page by specifying the time period of interest (year 2017) and the granularity (monthly). The API will return the page view count information as a json file in which the page view count is given for every month in the year 2017.

The category-title mapping Wikipedia pages are categorized in a structure called *category tree*. There are 22 *top-level* categories (e.g., Arts, Culture, Events, etc). Under each category, there could be relevant Wikipedia pages and subcategories, and each subcategory could also contain another set of subcategories and corresponding pages. Unlike the name suggests, *category tree* is *cyclic* and is in fact not tree-structured: each category could be under multiple categories and each page could belong to multiple categories/subcategories

and there could be *cycles*. Thus trying to traverse the whole tree to get articles under a certain category is infeasible. Considering all the categories/subcategories in the category tree could also be distractive. Thus, we only focus on the *top-level* categories and exclude the category *Reference works* since it is of little interest. We use *wptools* to traverse the category tree for each of the top-level categories of interest *individually* by specifying the category name during the traversal. We store the mapping information between each category and the corresponding Wikipage titles for later processing.

Components aggregation We first perform *intersection* over the titles of Wikipedia pages extracted from the original Wikipedia page texts, page view counts and category-title mapping, and drop those Wikipedia pages the texts in which contain less than 10 words (the total word count) from the vocabulary. Since there is a severe inconsistency in titles extracted from all the three components, performing intersection results in loss of a large portion of Wikipedia pages. After dropping all unqualified Wikipedia pages, we have approximately 500k remaining Wikipedia pages. We assign each Wikipedia page a timestamp based on its page view counts across the 12 months in 2017, the month in which the page gets most view counts is assigned to the page as its timestamp. Since our model naturally considers data in *two-level* batches (for example, we batch the Wikipedia pages based on time and category), and there are overlapping Wikipedia pages among different batches (one Wikipedia page could belong to multiple categories), we have about three million Wikipedia pages in the final dataset incarnation across all the batches.

CHAPTER 5

Multi-way Interacting Regression via Factorization Machines

We propose a Bayesian regression method that accounts for multi-way interactions of arbitrary orders among the predictor variables. Our model makes use of a factorization mechanism for representing the regression coefficients of interactions among the predictors, while the interaction selection is guided by a prior distribution on random hypergraphs, a construction which generalizes the Finite Feature Model. We present a posterior inference algorithm based on Gibbs sampling, and establish posterior consistency of our regression model.[1] Our method is evaluated with extensive experiments on simulated data and demonstrated to be able to identify meaningful interactions in applications in genetics and retail demand forecasting.[2]

5.1 Introduction

A fundamental challenge in supervised learning, particularly in regression, is the need for learning functions which produce accurate prediction of the response, while retaining the explanatory power for the role of the predictor variables in the model. The standard linear regression method is favored for the latter requirement, but it fails the former when there are complex interactions among the predictor variables in determining the response. The challenge becomes even more pronounced in a high-dimensional setting – there are exponentially many potential interactions among the predictors, for which it is simply not computationally feasible to resort to standard variable selection techniques (cf. Fan & Lv (2010)).

There are numerous examples where accounting for the predictors' interactions is of interest, including problems of identifying epistasis (gene-gene) and gene-environment

[1] Code is available at https://github.com/moonfolk/MiFM.

[2] This chapter has been published in Yurochkin et al. (2017b).

interactions in genetics (Cordell, 2009), modeling problems in political science (Brambor et al., 2006) and economics (Ai & Norton, 2003). In the business analytics of retail demand forecasting, a strong prediction model that also accurately accounts for the interactions of relevant predictors such as seasons, product types, geography, promotions, etc. plays a critical role in the decision making of marketing design.

A simple way to address the aforementioned issue in the regression problem is to simply restrict our attention to lower order interactions (i.e. 2- or 3-way) among predictor variables. This can be achieved, for instance, via a support vector machine (SVM) using polynomial kernels (Cristianini & Shawe-Taylor, 2000), which pre-determine the maximum order of predictor interactions. In practice, for computational reasons the degree of the polynomial kernel tends to be small. Factorization machines (Rendle, 2010) can be viewed as an extension of SVM to sparse settings where most interactions are observed only infrequently, subject to a constraint that the interaction order (a.k.a. interaction depth) is given. Neither SVM nor FM can perform any selection of predictor interactions, but several authors have extended the SVM by combining it with ℓ_1 penalty for the purpose of feature selection (Zhu et al., 2004) and gradient boosting for FM (Cheng et al., 2014) to select interacting features. It is also an option to perform linear regression on as many interactions as we can and combine it with regularization procedures for selection (e.g. LASSO (Tibshirani, 1996) or Elastic net (Zou & Hastie, 2005)). It is noted that such methods are still not computationally feasible for accounting for interactions that involve a large number of predictor variables.

In this work we propose a regression method capable of adaptive selection of multi-way interactions of arbitrary order (MiFM for short), while avoiding the combinatorial complexity growth encountered by the methods described above. MiFM extends the basic factorization mechanism for representing the regression coefficients of interactions among the predictors, while the interaction selection is guided by a prior distribution on random hypergraphs. The prior, which does not insist on the upper bound on the order of interactions among the predictor variables, is motivated from but also generalizes Finite Feature Model, a parametric form of the well-known Indian Buffet process (IBP) (Ghahramani & Griffiths, 2005). We introduce a notion of the hypergraph of interactions and show how a parametric distribution over binary matrices can be utilized to express interactions of unbounded order. In addition, our generalized construction allows us to exert extra control on the tail behavior of the interaction order. IBP was initially used for infinite latent feature modeling and later utilized in the modeling of a variety of domains (see a review paper by Griffiths & Ghahramani (2011)).

In developing MiFM, our contributions are the following: (i) we introduce a Bayesian multi-linear regression model, which aims to account for the multi-way interactions among

predictor variables; part of our model construction includes a prior specification on the hypergraph of interactions — in particular we show how our prior can be used to model the incidence matrix of interactions in several ways; (ii) we propose a procedure to estimate coefficients of arbitrary interactions structure; (iii) we establish posterior consistency of the resulting MiFM model, i.e., the property that the posterior distribution on the true regression function represented by the MiFM model contracts toward the truth under some conditions, without requiring an upper bound on the order of the predictor interactions; and (iv) we present a comprehensive simulation study of our model and analyze its performance for retail demand forecasting and case-control genetics datasets with epistasis. The unique strength of the MiFM method is the ability to recover meaningful interactions among the predictors while maintaining a competitive prediction quality compared to existing methods that target prediction only.

The chapter proceeds as follows. Section 5.2 introduces the problem of modeling interactions in regression, and gives a brief background on the Factorization Machines. Sections 5.3 and 5.4 carry out the contributions outlined above. Section 5.5 presents results of the experiments. We conclude with a discussion in Section 5.6.

5.2 Background and related work

Our starting point is a model which regresses a response variable $y \in \mathbb{R}$ to observed covariates (predictor variables) $x \in \mathbb{R}^D$ by a non-linear functional relationship. In particular, we consider a multi-linear structure to account for the interactions among the covariates in the model:

$$\mathbb{E}(Y|x) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{j=1}^{J} \beta_j \prod_{i \in Z_j} x_i. \tag{5.1}$$

Here, w_i for $i = 0, \ldots, D$ are bias and linear weights as in the standard linear regression model, J is the number of multi-way interactions where Z_j, β_j for $j = 1, \ldots, J$ represent the interactions, i.e., sets of indices of interacting covariates and the corresponding interaction weights, respectively. Fitting such a model is very challenging even if dimension D is of magnitude of a dozen, since there are $2^D - 1$ possible interactions to choose from in addition to other parameters. The goal of our work is to perform interaction selection and estimate corresponding weights. Before doing so, let us first discuss a model that puts a priori assumptions on the number and the structure of interactions.

5.2.1 Factorization Machines

Factorization Machines (FM) (Rendle, 2010) is a special case of the general interactions model defined in Eq. (5.1). Let $J = \sum_{l=2}^{d} \binom{D}{l}$ and $Z := \bigcup_{j=1}^{J} Z_j = \bigcup_{l=2}^{d} \{(i_1, \ldots, i_l) | i_1 < \ldots < i_l;\ i_1, \ldots, i_l \in \{1, \ldots, D\}\}$. I.e., restricting the set of interactions to $2, \ldots, d$-way, so (5.1) becomes:

$$\mathbb{E}(Y|x) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{l=2}^{d} \sum_{i_1=1}^{D} \ldots \sum_{i_l=i_{l-1}+1}^{D} \beta_{i_1,\ldots,i_l} \prod_{t=1}^{l} x_{i_t}, \tag{5.2}$$

where coefficients $\beta_j := \beta_{i_1,\ldots,i_l}$ quantify the interactions. In order to reduce model complexity and handle sparse data more effectively, Rendle (2010) suggested to factorize interaction weights using PARAFAC (Harshman, 1970): $\beta_{i_1,\ldots,i_l} := \sum_{f=1}^{k_l} \prod_{t=1}^{l} v_{i_t,f}^{(l)}$, where $V^{(l)} \in \mathbb{R}^{D \times k_l}$, $k_l \in \mathbb{N}$ and $k_l \ll D$ for $l = 2, \ldots, d$. Advantages of the FM over SVM are discussed in details by Rendle (2010). FMs turn out to be successful in the recommendation systems setups, since they utilize various context information (Rendle et al., 2011; Nguyen et al., 2014a). Parameter estimation is typically achieved via stochastic gradient descent technique, or in the case of Bayesian FM (Freudenthaler et al., 2011) via MCMC. In practice only $d = 2$ or $d = 3$ are typically used, since the number of interactions and hence the computational complexity grow exponentially. We are interested in methods that can adapt to fewer interactions but of arbitrarily varying orders.

5.3 MiFM: Multi-way Factorization Machine

We start by defining a mathematical object that can encode sets of interacting variables $Z_1, \ldots, Z_J$ of Eq. (5.1) and selecting an appropriate prior to model it.

5.3.1 Modeling hypergraph of interactions

Multi-way interactions are naturally represented by hypergraphs, which are defined as follows.

Definition 5.1. Given D vertices indexed by $S = \{1, \ldots, D\}$, let $Z = \{Z_1, \ldots, Z_J\}$ be the set of J subsets of S. Then we say that $G = (S, Z)$ is a hypergraph with D vertices and J hyperedges.

A hypergraph can be equivalently represented as an incidence binary matrix. Therefore, with a bit abuse of notation, we recast Z as the matrix of interactions, i.e., $Z \in \{0, 1\}^{D \times J}$, where $Z_{i_1 j} = Z_{i_2 j} = 1$ iff i_1 and i_2 are part of a hyperedge indexed by column/interaction j.

Placing a prior on multi-way interactions is the same as specifying the prior distribution on the space of binary matrices. We will at first adopt the Finite Feature Model (FFM) prior (Ghahramani & Griffiths, 2005), which is based on the Beta-Bernoulli construction: $\pi_j|\gamma_1,\gamma_2 \overset{iid}{\sim} \text{Beta}(\gamma_1,\gamma_2)$ and $Z_{ij}|\pi_j \overset{iid}{\sim} \text{Bernoulli}(\pi_j)$. This simple prior has the attractive feature of treating the variables involved in each interaction (hyperedge) in an symmetric fashion and admits exchangeabilility among the variables inside interactions. In Section 5.4 we will present an extension of FFM which allows to incorporate extra information about the distribution of the interaction degrees and explain the choice of the parametric construction.

5.3.2 Modeling regression with multi-way interactions

Now that we know how to model unknown interactions of arbitrary order, we combine it with the Bayesian FM to arrive at a complete specification of MiFM, the Multi-way interacting Factorization Machine. Starting with the specification for hyperparameters:

$$\begin{aligned}
&\sigma \sim \Gamma(\alpha_1/2,\beta_1/2), \qquad \lambda \sim \Gamma(\alpha_0/2,\beta_0/2), \qquad \mu \sim \mathcal{N}(\mu_0,1/\gamma_0),\\
&\lambda_k \sim \Gamma(\alpha_0/2,\beta_0/2), \qquad \mu_k \sim \mathcal{N}(\mu_0,1/\gamma_0) \text{ for } k=1,\dots,K.
\end{aligned}$$

Interactions and their weights:

$$\begin{aligned}
&w_i|\mu,\lambda \sim \mathcal{N}(\mu,1/\lambda) \text{ for } i=0,\dots,D, \qquad Z \sim \text{FFM}(\gamma_1,\gamma_2),\\
&v_{ik}|\mu_k,\lambda_k \sim \mathcal{N}(\mu_k,1/\lambda_k) \text{ for } i=1,\dots,D;\ k=1,\dots,K.
\end{aligned}$$

Likelihood specification given data pairs $(y_n, x_n = (x_{n1},\dots,x_{nD}))_{n=1}^N$:

$$y_n|\Theta \sim \mathcal{N}(y(x_n,\Theta),\sigma), \text{ where } y(x,\Theta) := w_0 + \sum_{i=1}^D w_i x_i + \sum_{j=1}^J \sum_{k=1}^K \prod_{i\in Z_j} x_i v_{ik} \tag{5.3}$$

for $n = 1,\dots,N$, and $\Theta = \{Z, V, \sigma, w_{0,\dots,D}\}$. Note that while the specification above utilizes Gaussian distributions, the main innovation of MiFM is the idea to utilize incidence matrix of the hypergraph of interactions Z with a low rank matrix V to model the mean response as in Eq. 5.1. Therefore, within the MiFM framework, different distributional choices can be made according to the problem at hand — e.g. Poisson likelihood and Gamma priors for count data or logistic regression for classification. Additionally, if selection of linear terms is desired, $\sum_{i=1}^D w_i x_i$ can be removed from the model since FFM can select linear interactions besides higher order ones.

5.3.3 MiFM for Categorical Variables

In numerous real world scenarios such as retail demand forecasting, recommender systems, genotype structures, most predictor variables may be categorical (e.g. color, season). Categorical variables with multiple attributes are often handled by so-called "one-hot encoding", via vectors of binary variables (e.g., IS_blue; IS_red), which must be mutually exclusive. The FFM cannot immediately be applied to such structures since it assigns positive probability to interactions between attributes of the same category. To this end, we model interactions between categories in Z, while with V we model coefficients of interactions between attributes. For example, for an interaction between "product type" and "season" in Z, V will have individual coefficients for "jacket-summer" and "jacket-winter" leading to a more refined predictive model of jackets sales (see examples in Section 5.5.2).

We proceed to describe MiFM for the case of categorical variables as follows. Let U be the number of categories and d_u be the set of attributes for the category u, for $u = 1, \ldots, U$. Then $D = \sum_{u=1}^{U} \mathrm{card}(d_u)$ is the number of binary variables in the one-hot encoding and $\bigsqcup_{u=1}^{U} d_u = \{1, \ldots, D\}$. In this representation the input data of predictors is X, a $N \times U$ matrix, where x_{nu} is an active attribute of category u of observation n. Coefficients matrix $V \in \mathbb{R}^{D \times K}$ and interactions $Z \in \{0, 1\}^{U \times J}$. All priors and hyperpriors are as before, while the mean response (5.3) is replaced by:

$$y(x, \Theta) := w_0 + \sum_{u=1}^{U} w_{x_u} + \sum_{k=1}^{K} \sum_{j=1}^{J} \prod_{u \in Z_j} v_{x_u k}. \tag{5.4}$$

Note that this model specification is easy to combine with continuous variables, allowing MiFM to handle data with different variable types.

5.3.4 Posterior Consistency of the MiFM

In this section we shall establish posterior consistency of MiFM model, namely: the posterior distribution Π of the conditional distribution $P(Y|X)$, given the training N-data pairs, contracts in a weak sense toward the truth as sample size N increases.

Suppose that the data pairs $(x_n, y_n)_{n=1}^{N} \in \mathbb{R}^D \times \mathbb{R}$ are i.i.d. samples from the joint distribution $P^*(X, Y)$, according to which the marginal distribution for X and the conditional distribution of Y given X admit density functions $f^*(x)$ and $f^*(y|x)$, respectively, with

respect to Lebesgue measure. In particular, $f^*(y|x)$ is defined by

$$\begin{aligned} &Y = y_n|X = x_n, \Theta^* \sim \mathcal{N}(y(x_n, \Theta^*), \sigma), \text{ where } \Theta^* = \{\beta_1^*, \ldots, \beta_J^*, Z_1^*, \ldots, Z_J^*\}, \\ &y(x, \Theta^*) := \sum_{j=1}^{J} \beta_j^* \prod_{i \in Z_j^*} x_i, \text{ and } x_n \in \mathbb{R}^D, y_n \in \mathbb{R}, \beta_j^* \in \mathbb{R}, Z_j^* \subset \{1, \ldots, D\} \end{aligned} \tag{5.5}$$

for $n = 1, \ldots, N, j = 1, \ldots, J$. In the above Θ^* represents the *true* parameter for the conditional density $f^*(y|x)$ that generates data sample y_n given x_n, for $n = 1, \ldots, N$. A key step in establishing posterior consistency for the MiFM (here we omit linear terms since, as mentioned earlier, they can be absorbed into the interaction structure) is to show that our PARAFAC type structure can approximate arbitrarily well the true coefficients $\beta_1^*, \ldots, \beta_J^*$ for the model given by (5.1).

Lemma 5.1. Given natural number $J \geq 1$, $\beta_j \in \mathbb{R} \setminus \{0\}$ and $Z_j \subset \{1, \ldots, D\}$ for $j = 1, \ldots J$, exists $K_0 < J$ such that for all $K \geq K_0$ system of polynomial equations $\beta_j = \sum_{k=1}^{K} \prod_{i \in Z_j} v_{ik}$, $j = 1, \ldots, m$ has at least one solution in terms of $v_{11}, \ldots, v_{DK}$.

The upper bound $K_0 = J - 1$ is only required when *all* interactions are of the depth $D - 1$. This is typically not expected to be the case in practice, therefore smaller values of K are often sufficient.

By conditioning on the training data pairs (x_n, y_n) to account for the likelihood induced by the PARAFAC representation, the statistician obtains the posterior distribution on the parameters of interest, namely, $\Theta := (Z, V)$, which in turn induces the posterior distribution on the conditional density, to be denoted by $f(y|x)$, according to the MiFM model (5.3) without linear terms. The main result of this section is to show that under some conditions this posterior distribution Π will place most of its mass on the true conditional density $f^*(y|x)$ as $N \to \infty$. To state the theorem precisely, we need to adopt a suitable notion of weak topology on the space of conditional densities, namely the set of $f(y|x)$, which is induced by the weak topology on the space of joint densities on X, Y, that is the set of $f(x, y) = f^*(x)f(y|x)$, where $f^*(x)$ is the true (but unknown) marginal density on X (see Ghosal et al. (1999), Sec. 2 for a formal definition).

Theorem 5.1. Given any true conditional density $f^*(y|x)$ given by (5.5), and assuming that the support of $f^*(x)$ is bounded, there is a constant $K_0 < J$ such that by setting $K \geq K_0$, the following statement holds: for any weak neighborhood U of $f^*(y|x)$, under the MiFM model, the posterior probability $\Pi(U|(X_n, Y_n)_{n=1}^N) \to 1$ with P^*-probability one, as $N \to \infty$.

The proof's sketch for this theorem is given in the Supplement.

5.4 Prior constructions for interactions: FFM revisited and extended

The adoption of the FFM prior on the hypergraph of interactions carries a distinct behavior in contrast to the typical Latent Feature modeling setting. In a standard Latent Feature modeling setting (Griffiths & Ghahramani, 2011), each row of Z describes one of the data points in terms of its feature representation; controlling row sums is desired to induce sparsity of the features. By contrast, for us a column of Z is identified with an interaction; its sum represents the interaction depth, which we want to control a priori.

Interaction selection using MCMC sampler One interesting issue of practical consequence arises in the aggregation of the MCMC samples (details of the sampler are in the Supplement). When aggregating MCMC samples in the context of *latent feature modeling* one would always obtain exactly J latent features. However, in *interaction modeling*, different samples might have no interactions in common (i.e. no exactly matching columns), meaning that support of the resulting posterior estimate can have up to $\min\{2^D - 1, IJ\}$ unique interactions, where I is the number of MCMC samples. In practice, we can obtain marginal distributions of all interactions across MCMC samples and use those marginals for selection. One approach is to pick J interactions with highest marginals and another is to consider interactions with marginal above some threshold (e.g. 0.5). We will resort to the second approach in our experiments in Section 5.5 as it seems to be in more agreement with the concept of "selection". Lastly, we note that while a data instance may a priori possess unbounded number of features, the number of possible interactions in the data is bounded by $2^D - 1$, therefore taking $J \to \infty$ might not be appropriate. In any case, we do not want to encourage the number of interactions to be too high for regression modeling, which would lead to overfitting. The above considerations led us to opt for a parametric prior such as the FFM for interactions structure Z, as opposed to going fully nonparametric. J can then be chosen using model selection procedures (e.g. cross validation), or simply taken as the model input parameter.

Generalized construction and induced distribution of interactions depths We now proceed to introduce a richer family of prior distributions on hypergraphs of which the FFM is one instance. Our construction is motivated by the induced distribution on the column sums and the conditional probability updates that arise in the original FFM. Recall that under the FFM prior, interactions are a priori independent. Fix an interaction j, for the remainder of this section let Z_i denote the indicator of whether variable i is present in interaction j

or not (subscript j is dropped from Z_{ij} to simplify notation). Let $M_i = Z_1 + \ldots + Z_i$ denote the number of variables among the first i present in the corresponding interaction. By the Beta-Bernoulli conjugacy, one obtains $\mathbb{P}(Z_i = 1|Z_1, \ldots, Z_{i-1}) = \frac{M_{i-1}+\gamma_1}{i-1+\gamma_1+\gamma_2}$. This highlights the "rich-gets-richer" effect of the FFM prior, which encourages the existence of very deep interactions while most other interactions have very small depths. In some situations we may prefer a relatively larger number of interactions of depths in the medium range.

An intuitive but somewhat naive alternative sampling process is to allow a variable to be included into an interaction according to its present "shallowness" quantified by $(i - 1 - M_{i-1})$ (instead of M_{i-1} in the FFM). It can be verified that this construction will lead to a distribution of interactions which concentrates most its mass around $D/2$; moreover, exchangeability among Z_i would be lost. To maintain exchangeability, we define the sampling process for the sequence $Z = (Z_1, \ldots, Z_D) \in \{0, 1\}^D$ as follows: let $\sigma(\cdot)$ be a random uniform permutation of $\{1, \ldots, D\}$ and let $\sigma_1 = \sigma^{-1}(1), \ldots, \sigma_D = \sigma^{-1}(D)$. Note that $\sigma_1, \ldots, \sigma_D$ are discrete random variables and $\mathbb{P}(\sigma_k = i) = 1/D$ for any $i, k = 1, \ldots, D$. For $i = 1, \ldots, D$, set

$$\begin{aligned}\mathbb{P}(Z_{\sigma_i} = 1|Z_{\sigma_1}, \ldots, Z_{\sigma_{i-1}}) &= \tfrac{\alpha M_{i-1}+(1-\alpha)(i-1-M_{i-1})+\gamma_1}{i-1+\gamma_1+\gamma_2},\\ \mathbb{P}(Z_{\sigma_i} = 0|Z_{\sigma_1}, \ldots, Z_{\sigma_{i-1}}) &= \tfrac{(1-\alpha)M_{i-1}+\alpha(i-1-M_{i-1})+\gamma_2}{i-1+\gamma_1+\gamma_2},\end{aligned} \tag{5.6}$$

where $\gamma_1 > 0, \gamma_2 > 0, \alpha \in [0, 1]$ are given parameters and $M_i = Z_{\sigma_1} + \ldots + Z_{\sigma_i}$. The collection of Z generated by this process shall be called to follow FFM$_\alpha$. When $\alpha = 1$ we recover the original FFM prior. When $\alpha = 0$, we get the other extremal behavior mentioned at the beginning of the paragraph. Allowing $\alpha \in [0, 1]$ yields a richer spectrum spanning the two distinct extremal behaviors.

Details of the process and some of its properties are given in the Supplement. Here we briefly describe how FFM$_\alpha$ *a priori* ensures "poor gets richer" behavior and offers extra flexibility in modeling interaction depths compared to the original FFM. The depth of an interaction of D variables is described by the distribution of M_D. Consider the conditionals obtained for a Gibbs sampler where index of a variable to be updated is random and based on $\mathbb{P}(\sigma_D = i|Z)$ (it is simply $1/D$ for FFM$_1$). Suppose we want to assess how likely it is to *add* a variable into an existing interaction via the expression $\sum_{i:Z_i^{(k)}=0} \mathbb{P}(Z_i^{(k+1)} = 1, \sigma_D = i|Z^{(k)})$, where $k + 1$ is the next iteration of the Gibbs sampler's conditional update. This probability is a function of $M_D^{(k)}$; for small values of $M_D^{(k)}$ it quantifies the tendency for the "poor gets richer" behavior. For the FFM$_1$ it is given by $\frac{D-M_D^{(k)}}{D}\frac{M_D^{(k)}+\gamma_1}{D-1+\gamma_1+\gamma_2}$. In Fig. 5.1(a) we show that FFM$_1$'s behavior is opposite of "poor

gets richer", while $\alpha \leq 0.7$ appears to ensure the desired property. Next, in Fig.5.1 (b-f) we show the distribution of M_D for various α, which exhibits a broader spectrum of behavior.

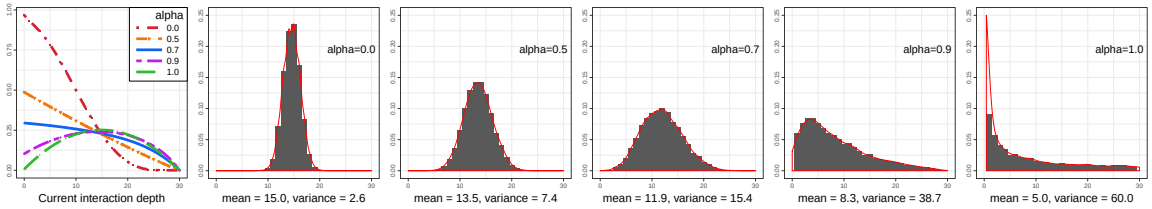


Figure 5.1: $D = 30$, $\gamma_1 = 0.2$, $\gamma_2 = 1$ (a) Probability of increasing interaction depth; (b-f) FFM$_\alpha$ M_D distributions with different α.

5.5 Experimental Results

5.5.1 Simulation Studies

We shall compare MiFM methods against a variety of other regression techniques in the literature, including Bayesian Factorization Machines (FM), lasso-type regression, Support Vector Regression (SVR), multilayer perceptron neural network (MLP).[3] The comparisons are done on the basis of prediction accuracy of responses (Root Mean Squared Error on the held out data), quality of regression coefficient estimates and the interactions recovered.

5.5.1.1 Predictive Performance

In this set of experiments we demonstrate that MiFMs with either $\alpha = 0.7$ or $\alpha = 1$ have dominant predictive performance when high order interactions are in play.

In Fig. 5.2(a) we analyzed 70 random interactions of varying orders. We see that MiFM can handle arbitrary complexity of the interactions, while other methods are comparative only when interaction structure is simple (i.e. linear or 2-way on the right of the Fig. 5.2(a)).

Next, to assess the effectiveness of MiFM in handling categorical variables (cf. Section 5.3.3) we vary the number of continuous variables from 1 (and 29 attributes across categories) to 30 (no categorical variables). Results in Fig. 5.2(b) demonstrate that our models can handle both variable types in the data (including continuous-categorical interactions), and still exhibit competitive RMSE performance.

[3]Random Forest Regression and optimization based FM showed worse results than other methods.

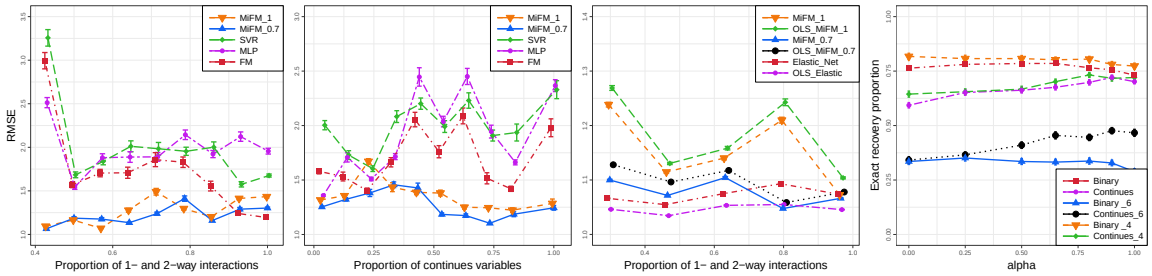


Figure 5.2: RMSE for experiments: (a) interactions depths; (b) data with different ratio of continuous to categorical variables; (c) quality of the MiFM_1 and $\text{MiFM}_{0.7}$ coefficients; (d) MiFM_α exact recovery of the interactions with different α and data scenarios

5.5.1.2 Interactions Quality

Coefficients of the interactions This experiment verifies the posterior consistency result of Theorem 5.1 and validates our factorization model for coefficients approximation. In Fig. 5.2(c) we compare MiFMs versus OLS fitted with the corresponding sets of chosen interactions. Additionally we benchmark against Elastic net (Zou & Hastie, 2005) based on the expanded data matrix with interactions of all depths included, that is $2^D - 1$ columns, and a corresponding OLS with only selected interactions.

Selection of the interactions In this experiments we assess how well MiFM can recover true interactions. We consider three interaction structures: a realistic one with five linear, five 2-way, three 3-way and one of each $4, \ldots, 8$-way interactions, and two artificial ones with 15 either only 4- or only 6-way interactions to challenge our model. Both binary and continuous variables are explored. Fig. 5.2(d) shows that MiFM can *exactly* recover up to 83% of the interactions and with $\alpha = 0.8$ it recovers 75% of the interaction in 4 out of 6 scenarios. Situation with 6-way interactions is more challenging, where 36% for binary data is recovered and almost half for continuous. It is interesting to note that lower values of α handle binary data better, while higher values are more appropriate for continuous, which is especially noticeable on the "only 6-way" case. We think it might be related to the fact that high order interactions between binary variables are very rare in the data (i.e. product of 6 binary variables is equal to 0 most of the times) and we need a prior eager to explore ($\alpha = 0$) to find them.

5.5.2 Real world applications

5.5.2.1 Finding epistasis

Identifying epistasis (i.e. interactions between genes) is one of the major questions in the field of human genetics. Interactions between multiple genes and environmental factors can often tell a lot more about the presence of a certain disease than any of the genes individually (Templeton, 2000). Our analysis of the epistasis is based on the data from Himmelstein et al. (2011). These authors show that interactions between single nucleotide polymorphisms (SNPs) are often powerful predictors of various diseases, while individually SNPs might not contain important information at all. They developed a model free approach to simulate data mimicking relationships between complex gene interactions and the presence of a disease. We used datasets with five SNPs and either 3-,4- and 5-way interactions or only 5-way interactions. For this experiment we compared MiFM_1, MiFM_0; refitted logistic regression for each of our models based on the selected interactions (LMiFM_1 and LMiFM_0), Multilayer Perceptron with 3 layers and Random Forest.[4] Results in Table 5.1 demonstrate that MiFM produces competitive performance compared to the very best black-box techniques on this data set, while it also selects interacting genes (i.e. finds epistasis). We don't know which of the 3- and 4-way interactions are present in the data, but since there is only one possible 5-way interaction we can check if it was identified or not — both MiFM_1 and MiFM_0 had a 5-way interaction in at least *95%* of the posterior samples.

Table 5.1: Prediction Accuracy on the Held-out Samples for the Gene Data

	MiFM_1	MiFM_0	LMiFM_1	LMiFM_0	MLP	RF
3-, 4-, 5-way	0.775	0.771	0.883	0.860	0.870	0.887
only 5-way	0.649	0.645	0.628	0.623	0.625	0.628

5.5.2.2 Understanding retail demand

We finally report the analysis of data obtained from a major retailer with stores in multiple locations all over the world. This dataset has 430k observations and 26 variables spanning over 1100 binary variables after the one-hot encoding. Sales of a variety of products on different days and in different stores are provided as response. We will compare MiFM_1 and MiFM_0, both fitted with $K = 12$ and $J = 150$, versus Factorization Machines in terms of adjusted mean absolute percent error $\text{AMAPE} = 100\frac{\sum_n |\hat{y}_n - y_n|}{\sum_n y_n}$, a common metric for evaluating sales forecasts. FM is currently a method of choice by the company for this data

[4] FM, SVM and logistic regression had low accuracy of around 50% and are not reported.

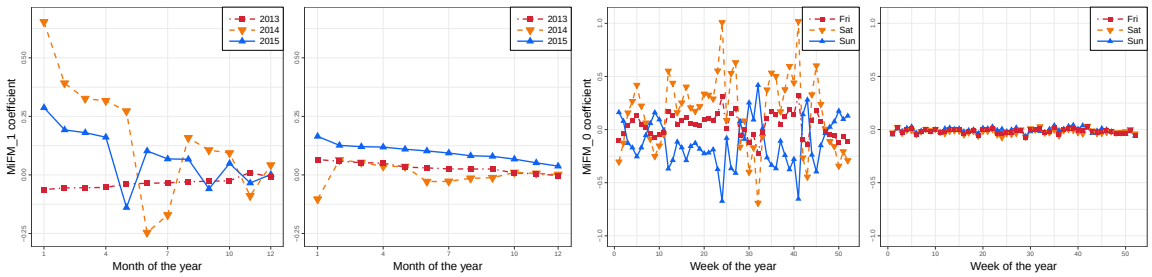


Figure 5.3: MiFM_1 store - month - year interaction: (a) store in Merignac; (b) store in Perols; MiFM_0 city - store - day of week - week of year interaction: (c) store in Merignac; (d) store in Perols.

set, partly because the data is sparse and is similar in nature to the recommender systems. AMAPE for MiFM_1 is 92.4; for MiFM_0 - 92.45; for FM - 92.0.

Posterior analysis of predictor interactions The unique strength of MiFM is the ability to provide valuable insights about the data through its posterior analysis. MiFM_1 recovered 62 non-linear interactions among which there are five 3-way and three 4-way. MiFM_0 selected 63 non-linear interactions including nine 3-way and four 4-way. We note that choice $\alpha = 0$ was made to explore deeper interactions and as we see MiFM_0 has more deeper interactions than MiFM_1. Coefficients for a 3-way interaction of MiFM_1 for two stores in France across years and months are shown in Fig. 5.3(a,b). We observe different behavior, which would not be captured by a low order interaction. In Fig. 5.3(c,d) we plot coefficients of a 4-way MiFM_0 interaction for the same two stores in France. It is interesting to note negative correlation between Saturday and Sunday coefficients for the store in Merignac, while the store in Perols is not affected by this interaction - this is an example of how MiFM can select interactions between attributes across categories.

5.6 Discussion

We have proposed a novel regression method which is capable of learning interactions of arbitrary orders among the regression predictors. Our model extends Finite Feature Model and utilizes the extension to specify a hypergraph of interactions, while adopting a factorization mechanism for representing the corresponding coefficients. We found that MiFM performs very well when there are some important interactions among a relatively high number (higher than two) of predictor variables. This is the situation where existing modeling techniques may be ill-equipped at describing and recovering. There are several future directions that we would like to pursue. A thorough understanding of the fully nonparametric version of the FFM_α is of interest, that is, when the number of columns is

taken to infinity. Such understanding may lead to an extension of the IBP and new modeling approaches in various domains.

Appendix

In the Supplementary material we will start by proving consistency of the MiFM theorem, then we will show several important results related to FFM$_\alpha$: how exchangeability is achieved using uniform permutation prior on the order in which variables enter the process, how it leads to a Gibbs sampler using distribution of the index of the variable entering FFM$_\alpha$ last and how to obtain distribution of the interaction depths M_D and compute its expectation. Lastly we will present a Gibbs sampling algorithm for the MiFM under the FFM$_\alpha$ prior on interactions structure Z.

5.A Proof of the Consistency Theorem 5.1

First let us remind the reader of the problem setup. Suppose that the data pairs $(x_n, y_n)_{n=1}^N \in \mathbb{R}^D \times \mathbb{R}$ are i.i.d. samples from the joint distribution $P^*(X, Y)$, according to which marginal distribution for X and the conditional distribution of Y given X admit density functions $f^*(x)$ and $f^*(y|x)$, respectively, with respect to Lebesgue measure. In particular, $f^*(y|x)$ is defined as in Eq. (5.5):

$$
\begin{aligned}
&Y = y_n | X = x_n, \Theta^* \sim \mathcal{N}(y(x_n, \Theta^*), \sigma), \text{ where } \Theta^* = \{\beta_1^*, \ldots, \beta_J^*, Z_1^*, \ldots, Z_J^*\}, \\
&y(x, \Theta^*) := \sum_{j=1}^{J} \beta_j^* \prod_{i \in Z_j^*} x_i, \text{ and } x_n \in \mathbb{R}^D, y_n \in \mathbb{R}, \beta_j^* \in \mathbb{R}, Z_j^* \subset \{1, \ldots, D\}, \\
&\text{for } n = 1, \ldots, N, j = 1, \ldots, J.
\end{aligned}
$$

In the above Θ^* represents the *true* parameter for the conditional density $f^*(y|x)$ that generates data sample y_n given x_n, for $n = 1, \ldots, N$. On the other hand, the statistical modeler has access only to the MiFM:

$$
\begin{aligned}
&Z \sim \text{FFM}_\alpha(\gamma_1, \gamma_2), \; v_{ik} | \mu_k, \lambda_k \sim \mathcal{N}(\mu_k, \frac{1}{\lambda_k}) \text{ for } i = 1, \ldots, D; \; k = 1, \ldots, K, \\
&y_n | \Theta \sim \mathcal{N}(y(x_n, \Theta), \sigma), \text{ where } y(x, \Theta) := \sum_{j=1}^{J} \sum_{k=1}^{K} \prod_{i \in Z_j} x_i v_{ik}, \\
&\text{for } n = 1, \ldots, N, \text{ and } \Theta = (Z, V).
\end{aligned}
\tag{5.7}
$$

We omitted linear terms in the MiFM since they can naturally be parts of the interaction structure Z and discarded hyperpriors for the ease of representation. Now we show that under some conditions posterior distribution Π will place most of its mass on the true conditional density $f^*(y|x)$ as $N \to \infty$.

Theorem 5.1. *Given any true conditional density $f^*(y|x)$ given by (5.5), and assuming that the support of $f^*(x)$ is bounded, there is a constant $K_0 < J$ such that by setting $K \geq K_0$, the following statement holds: for any weak neighborhood U of $f^*(y|x)$, under the MiFM model (5.7), the posterior probability $\Pi(U|(X_n, Y_n)_{n=1}^N) \to 1$ with P^*-probability one, as $N \to \infty$.*

A key part in the proof of this theorem is to clarify the role of parameter K, and the fact that under model (5.7), the regression coefficient β_j associated with interaction j is parameterized by $\beta_j := \sum_{k=1}^K \prod_{i \in Z_j} v_{ik}$, for $j = 1, \ldots, J$, which for some suitable choice of $\Theta = (Z, V)$ can represent exactly the true parameters $\beta_1^*, \ldots, \beta_J^*$, provided that K is sufficiently large. The following basic lemma is informative.

Lemma 5.2. Let $m \in [1, J]$ be a natural number, $\beta_j \in \mathbb{R} \setminus \{0\}$ for $j = 1, \ldots, m$. Suppose that the m subsets $Z_j \subset \{1, \ldots, D\}$ for $j = 1, \ldots m$ have non-empty intersection, then as long as $K \geq m$, the system of polynomial equations

$$\sum_{k=1}^{K} \prod_{i \in Z_j} v_{ik} = \beta_j, j = 1, \ldots, m \tag{5.8}$$

has at least one solution in terms of $v_{11}, \ldots, v_{DK}$ such that the following collection of K vectors in $\mathbb{R}^m$, namely, $\{(\prod_{i \in Z_1} v_{ik}, \ldots, \prod_{i \in Z_m} v_{ik}), k = 1, \ldots, K\}$ contains m linearly independent vectors.

Proof. Let i_0 be an element of the intersection of all Z_j, for $j = 1, \ldots, m$. We consider system (5.8) as linear with respect to $\{v_{i_0 1}, \ldots, v_{i_0 K}\}$, where corresponding coefficients are given by $\prod_{i \in Z_j \setminus \{i_0\}} v_{i,k}$, which we can pick to form a matrix of nonzero determinant. Hence by Rouché–Capelli theorem the system has at least one solution if $K \geq m$ and, since $\beta_j \neq 0$ for $\forall j$, the resulting $\{(\prod_{i \in Z_1} v_{ik}, \ldots, \prod_{i \in Z_m} v_{ik}), k = 1, \ldots, K\}$ contains at least m linearly independent vectors. □

Lemma 5.1. *Given natural number $J \geq 1$, $\beta_j \in \mathbb{R} \setminus \{0\}$ and $Z_j \subset \{1, \ldots, D\}$ for $j = 1, \ldots J$, exists $K_0 < J : \forall K \geq K_0$ system of polynomial equations (5.8) has at least one solution in terms of $v_{11}, \ldots, v_{DK}$.*

Proof. The proof proceeds by performing an elimination process on the collection of variables v_{ik} according to an ordering that we now define. Let $J_i = \text{card}(\{Z_j | i \in Z_j\})$ for $i = 1, \ldots, D$. Define $J^0 = \min_i J_i$ and $i_0 = \operatorname{argmin}_i J_i$. If $K \geq J^0$ by Lemma 5.2 we can find a solution of the reduced system of equations

$$\sum_{k=1}^{K} \prod_{i \in Z_j} v_{i,k} = \beta_j, \, j \in \{j | i_0 \in Z_j\},$$

while maintaining the linear independence needed to apply Lemma 5.2 again further. Now we know that we can find a solution for equations indexed by $\{j | i_0 \in Z_j\}$. We remove them from system (5.8) and recompute $J^1 = \min_{i \neq i_0} J_i$ and $i_1 = \operatorname{argmin}_{i \neq i_0} J_i$ to apply Lemma 5.2 again. Iteratively we will remove all the equations, meaning that there is at least one solution. Note that J_i are decreasing since whenever we remove equations, number of Z_js containing certain i can only decrease. Therefore, we will need $K \geq K_0 := \max(J^0, J^1, \ldots, 0)$ in order to apply Lemma 5.2 on every elimination step. □

From the proof of Lemma 5.1, it can be observed that $K_0 = \max(J^0, J^1, \ldots) \ll J$ when we anticipate only few interactions per variable, whereas the upper bound $K_0 = J - 1$ is attained when there are only $(D-1)$-way interactions. Now we are ready to present a proof of the main theorem.

Proof. (of main theorem). By Lemma 5.1 and the fact that the probability of a finite number of independent continuous random vectors being linearly dependent is 0 it follows that under the MiFM prior on V as in (5.7) and $\forall \beta_1, \ldots, \beta_J \in \mathbb{R} \setminus \{0\}$, distinct $Z_1, \ldots, Z_J$ and $\epsilon > 0$

$$\Pi \left(\sum_{j=1}^{J} (\beta_j - \sum_{k} \prod_{i \in Z_j} v_{ik})^2 < \epsilon \, | Z_1, \ldots, Z_J \right) > 0. \tag{5.9}$$

From Eq. (5.6) it follows that for any $Z_1, \ldots, Z_J$, the prior probability of the corresponding incidence matrix is bounded away from 0. Combining this with (5.9), we now establish that the probability of the true model parameters to be arbitrary close to the MiFM parameters under the MiFM prior as in (5.7):

$$\Pi \left((\sum_{j=1}^{J} \beta_j - \sum_{j=1}^{J} \sum_{k} \prod_{i \in Z_j} v_{ik})^2 < \epsilon \right) > 0, \, \forall \epsilon > 0. \tag{5.10}$$

We shall appeal to Schwartz's theorem (cf. Ghosal et al. (1999)), which asserts that the desired posterior consistency holds as soon as we can establish that the true joint distribution

$P^*(X,Y)$ lies in the Kullback-Leibler support of the prior Π on the joint distribution $P(X,Y)$. That is,

$$\Pi\left(\mathrm{KL}(P^*||P) < \epsilon\right) > 0, \text{ for } \forall \epsilon > 0. \tag{5.11}$$

Since the KL divergence of the two Gaussian distributions is proportional to the mean difference, we have ($\mathbb{E}^*_X$ denotes expectation with respect to the true marginal distribution of X)

$$\begin{aligned}&\mathrm{KL}(P^*||P) \propto \mathbb{E}^*_X \frac{1}{2}(y(X,\Theta) - y(X,\Theta^*))^2 \propto \\ &\mathbb{E}^*_X(\sum_{j=1}^{J}\beta_j \prod_{i\in Z_j} x_i - \sum_{j=1}^{J}\sum_{k}\prod_{i\in Z_j} v_{ik}x_i)^2 \lesssim (\sum_{j=1}^{J}\beta_j - \sum_{j=1}^{J}\sum_{k}\prod_{i\in Z_j} v_{ik})^2.\end{aligned} \tag{5.12}$$

Due to (5.10) this quantity can be made arbitrarily close to 0 with positive probability. Therefore (5.11) and then Schwartz theorem hold, which concludes the proof. □

5.B Analyzing FFM$_\alpha$

5.B.1 Model definition and exchangeability

Here we remind the reader the construction of FFM$_\alpha$ — the distribution over finite collection of binary random variables that we used to model interactions. Let D be the number of variables in the data and $Z \in \{0,1\}^D$ is j-th interaction (subscript j is dropped to simplify notation). Let $\sigma(\cdot)$ be a random uniform permutation of $\{1,\dots,D\}$ and let $\sigma_1 = \sigma^{-1}(1),\dots,\sigma_D = \sigma^{-1}(D)$. Note that $\sigma_1,\dots,\sigma_D$ are discrete random variables and $\mathbb{P}(\sigma_k = i) = 1/D$ for any $i,k = 1,\dots,D$. Next recall FFM$_\alpha$ from Eq. (5.6):

$$\begin{aligned}\mathbb{P}(Z_{\sigma_i} = 1|Z_{\sigma_1},\dots,Z_{\sigma_{i-1}}) &= \tfrac{\alpha M_{i-1} + (1-\alpha)(i-1-M_{i-1}) + \gamma_1}{i-1+\gamma_1+\gamma_2},\\ \mathbb{P}(Z_{\sigma_i} = 0|Z_{\sigma_1},\dots,Z_{\sigma_{i-1}}) &= \tfrac{(1-\alpha) M_{i-1} + \alpha(i-1-M_{i-1}) + \gamma_2}{i-1+\gamma_1+\gamma_2},\end{aligned}$$

where $\gamma_1 > 0, \gamma_2 > 0, \alpha \in [0,1]$ are given parameters and $M_i = Z_{\sigma_1} + \ldots + Z_{\sigma_i}$. Due to the random permutation of indices, distribution of $Z_1,\dots,Z_D$ is exchangeable because any ordering of variables entering the process has same probability. Next, we need to integrate the permutation part out to obtain a tractable full conditional representation.

5.B.2 Gibbs sampling for FFM$_\alpha$ and distribution of interaction depths M_D

To construct a Gibbs sampler for the the FFM$_\alpha$ we will use an additional latent variable - index of the variable entering the process last, σ_D. Additionally observe that when permutation is integrated out $\mathbb{P}(Z_1, \ldots, Z_D) = \mathbb{P}(M_D = Z_1 + \ldots + Z_D)$ since $\mathbb{P}(M_D = m)$ is precisely the summation over all possible orderings of $Z_1, \ldots, Z_D$ such that $Z_1 + \ldots + Z_D = m$.

$$\begin{aligned}
&\mathbb{P}(\sigma_D = i | Z_1, \ldots, Z_D) \propto \\
&Z_i \mathbb{P}(\sigma_D = i | Z_{\sigma_D} = 1, Z) \mathbb{P}(Z_{\sigma_D} = 1 | M_{D-1} = \sum_{k=1}^{D} Z_k - 1) \mathbb{P}(M_{D-1} = \sum_{k=1}^{D} Z_k - 1) + \\
&+ (1 - Z_i) \mathbb{P}(\sigma_D = i | Z_{\sigma_D} = 0, Z) \mathbb{P}(Z_{\sigma_D} = 0 | M_{D-1} = \sum_{k=1}^{D} Z_k) \mathbb{P}(M_{D-1} = \sum_{k=1}^{D} Z_k),
\end{aligned} \tag{5.13}$$

then if $Z_i = 1$ and $\sum_{k=1}^{D} Z_k = m$ we obtain

$$\begin{aligned}
\mathbb{P}(\sigma_D = i | Z_{-i}, Z_i = 1) = \mathbb{P}(\sigma_D = i | M_D = m, Z_i = 1) = \\
= \frac{\mathbb{P}(M_{D-1} = m - 1) \mathbb{P}(Z_{\sigma_D} = 1 | M_{D-1} = m - 1)}{m \mathbb{P}(M_D = m)},
\end{aligned} \tag{5.14}$$

where $\mathbb{P}(Z_{\sigma_D} = 1 | M_{D-1} = m - 1)$ and $\mathbb{P}(Z_{\sigma_D} = 0 | M_{D-1} = m)$ can be computed as in Eq. 5.6. Our next step is to analyze probability $\mathbb{P}(M_D = m)$. Indeed it is easy to obtain this distribution recursively:

$$\begin{aligned}
\mathbb{P}(M_D = m) = \mathbb{P}(M_{D-1} = m) \mathbb{P}(Z_{\sigma_D} = 0 | M_{D-1} = m) + \\
+ \mathbb{P}(M_{D-1} = m - 1) \mathbb{P}(Z_{\sigma_D} = 1 | M_{D-1} = m - 1).
\end{aligned} \tag{5.15}$$

The base of recursion is given by the following identities:

$$\begin{aligned}
&\mathbb{P}(M_0 = 0) = 1, \\
&\mathbb{P}(M_i = 0) = \prod_{k=0}^{i-1} \frac{\alpha(i - 1 - k) + \gamma_2}{k + \gamma_1 + \gamma_2} = \prod_{k=0}^{i-1} \frac{\alpha k + \gamma_2}{k + \gamma_1 + \gamma_2}, \\
&\mathbb{P}(M_i = i) = \prod_{k=0}^{i-1} \frac{\alpha k + \gamma_1}{k + \gamma_1 + \gamma_2}.
\end{aligned} \tag{5.16}$$

The above formulation allows us compute $\mathbb{P}(M_i = k), D \geq i \geq k$ dynamically (com-

putations are very fast since we only need to perform $\frac{(D+1)(D+2)}{2} - 1$ calculations) *before* running MiFM inference and utilize the table of probabilities during it. The last step of the Gibbs sampler is clearly the update of the $Z_i|\sigma_D = i, Z_{-i}$ which is done simply using the FFM$_\alpha$ definition 5.6. Recall Figure 1 (a) of the main text which illustrates the behavior of

$$\sum_{i:Z_i^{(k)}=0} \mathbb{P}(Z_i^{(k+1)} = 1, \sigma_D = i|Z^{(k)}) = \mathbb{P}(Z_{\sigma_D} = 0|Z)\mathbb{P}(Z_i = 1|\sigma_D = i, Z_{-i}),$$

and since we choose index of a variable to update based on the probability of it being last, the expression above reads as the probability that we choose to update a variable not present in the interaction and then add it to the interaction, therefore increasing the depth of the interaction.

5.B.3 Mean Behavior of the FFM$_\alpha$

From Eq. (5.15) it follows that

$$\begin{aligned}
&\mathbb{E}M_D = \sum_{m=0}^{D} m\mathbb{P}(M_D = m) = \\
&= \frac{1}{D-1+\gamma_1+\gamma_2}\Big\{(1-2\alpha)\mathbb{E}M_{D-1}^2 + (\alpha(D-1)+\gamma_2)\mathbb{E}M_{D-1} + \\
&+ (2\alpha-1)\mathbb{E}(M_{D-1}+1)^2 + ((1-\alpha)D - \alpha + \gamma_1)\mathbb{E}(M_{D-1}+1)\Big\} \\
&= \frac{1}{D-1+\gamma_1+\gamma_2}\Big\{\mathbb{E}M_{D-1}(D+2\alpha+\gamma_1+\gamma_2-2) + D(1-\alpha) + \alpha + \gamma_1 - 1\Big\}.
\end{aligned} \tag{5.17}$$

For $\alpha = 0$, this relation is simplified to be

$$\begin{aligned}
(D-1+\gamma_1+\gamma_2)\mathbb{E}M_D &= \mathbb{E}M_{D-1}(D+\gamma_1+\gamma_2-2) + (D+\gamma_1-1) = \\
&= (D+\gamma_1-1) + \ldots + \gamma_1 = \frac{1}{2}D(D+2\gamma_1-1).
\end{aligned} \tag{5.18}$$

5.C Gibbs Sampler for the MiFM

Our Gibbs sampling algorithm consists of two parts — updating factorization coefficients V (based on the results from Freudenthaler et al. (2011)) and then updating interactions Z based on the analysis of Section 5.B.2. Recall the MiFM model construction. First we have

a layer of hyperpriors:

$$\sigma \sim \Gamma(\frac{\alpha_1}{2}, \frac{\beta_1}{2}), \qquad \lambda \sim \Gamma(\frac{\alpha_0}{2}, \frac{\beta_0}{2}), \qquad \mu \sim \mathcal{N}(\mu_0, \frac{1}{\gamma_0}),$$
$$\lambda_k \sim \Gamma(\frac{\alpha_0}{2}, \frac{\beta_0}{2}),\ \mu_k \sim \mathcal{N}(\mu_0, \frac{1}{\gamma_0}) \text{ for } k = 1, \dots, K,$$

Then interactions and their weights:

$$w_i|\mu, \lambda \sim \mathcal{N}(\mu, \frac{1}{\lambda}) \text{ for } i = 0, \dots, D, \qquad Z \sim \text{FFM}_\alpha(\gamma_1, \gamma_2),$$
$$v_{ik}|\mu_k, \lambda_k \sim \mathcal{N}(\mu_k, \frac{1}{\lambda_k}) \text{ for } i = 1, \dots, D;\ k = 1, \dots, K,$$

And finally the model's likelihood from Eq. (5.3)

$$y_n|\Theta \sim \mathcal{N}(y(x_n, \Theta), \frac{1}{\sigma}), \text{ where}$$
$$y(x, \Theta) := w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{j=1}^{J} \sum_{k=1}^{K} \prod_{i \in Z_j} x_i v_{ik},$$
$$\text{for } n = 1, \dots, N, \text{ and } \Theta = \{Z, V, \sigma, w_{0,\dots,D}\}.$$

Inference in the context of Bayesian modeling is often related to learning the posterior distribution $\mathbb{P}(\Theta|X, Y)$. Then, if one wants point estimates, certain statistics of the posterior can be used, i.e. mean or median. In most situations (including MiFM) analytical form of the posterior is intractable, but with the help of Bayes rule it is often possible to compute it up to a proportionality constant:

$$\begin{aligned}\mathbb{P}(\Theta, \mu, \gamma, \mu_1, \dots \mu_K, \lambda_1, \dots, \lambda_K|Y) \propto \prod_{n=1}^{N} \mathbb{P}(y_n|Z, V, \sigma, w_{0,\dots,D}) \cdot \\ \cdot \mathbb{P}(Z)\mathbb{P}(V|\mu_1, \dots, \mu_K, \lambda_1, \dots, \lambda_K)\mathbb{P}(\sigma, \mu, \gamma, \mu_1, \dots \mu_K, \lambda_1, \dots, \lambda_K).\end{aligned} \tag{5.19}$$

One can maximize this quantity to obtain MAP estimate, but this is very complicated due to the combinatorial complexity of interactions in Z and, additionally, often leads to overfitting. We use Gibbs sampling procedure for learning the posterior of our model. Due to normal-normal conjugacy and a priori independence of Z and other latent variables, we can derive closed form full conditional (i.e. variable given all the rest and the data) distributions for each of the latent variables in the model.

Updating hyperprior parameters

$$\sigma \sim \Gamma\left(\frac{\alpha_1+N}{2};\ \frac{\sum_{n=1}^{N}(y_n-y(x_n,\Theta))^2+\beta_1)}{2}\right), \tag{5.20}$$

$$\lambda \sim \Gamma\left(\frac{\alpha_0+D+1}{2};\ \frac{\sum_{i=0}^{D}(w_i-\mu)^2+\beta_0}{2}\right), \tag{5.21}$$

$$\mu \sim \mathcal{N}\left(\frac{\sum_{i=0}^{D} w_i+\gamma_0\mu_0}{D+1+\gamma_0};\ \frac{1}{\lambda(D+1+\gamma_0)}\right), \tag{5.22}$$

$$\lambda_k \sim \Gamma\left(\frac{\alpha_0+D}{2};\ \frac{\sum_{i=1}^{D}(v_{ik}-\mu_k)^2+\beta_0}{2}\right), \tag{5.23}$$

$$\mu_k \sim \mathcal{N}\left(\frac{\sum_{i=1}^{D} v_{ik}+\gamma_0\mu_0}{D+\gamma_0};\ \frac{1}{\lambda_k(D+\gamma_0)}\right), \tag{5.24}$$

$$\text{for } k = 1, \ldots, K.$$

Updating factorization coefficients V For updating coefficients of the model we can utilize the multi-linear property also used for the Factorization Machines MCMC updates (Freudenthaler et al., 2011). Note that for any $\theta \in \{w_0, \ldots, w_D, v_{11}, \ldots, v_{DK}\}$ we can write $y(x, \Theta) = l_\theta(x) + \theta m_\theta(x)$, where $l_\theta(\cdot)$ are all the terms independent of θ and $m_\theta(\cdot)$ are the terms multiplied by θ. For example, if $\theta = w_0$, then $m_\theta(x) = 1$ and $l_\theta(x) = \sum_{i=1}^{D} w_i x_i + \sum_{j=1}^{J}\sum_{k=1}^{K}\prod_{i\in Z_j} x_i v_{ik}$. Next we give updating distribution that can be used for any $\theta \in \{w_0, \ldots, w_D, v_{11}, \ldots, v_{DK}\}$.

$$\begin{aligned}
&\theta \sim \mathcal{N}(\mu_\theta^*, \sigma_\theta^2), \text{where } \sigma_\theta^2 = \left(\sigma\sum_{n=1}^{N} m_\theta(x_n)^2 + \lambda_\theta\right)^{-1}, \\
&\mu_\theta^* = \sigma_\theta^2\left(\sigma\sum_{n=1}^{N}(y_n - l_\theta(x_n))m_\theta(x_n) + \mu_\theta\lambda_\theta\right),
\end{aligned} \tag{5.25}$$

and $\mu_\theta, \lambda_\theta$ are the corresponding hyperprior parameters.

Updating interactions Z Posterior updates of Z can be decomposed into prior times the likelihood:

$$\mathbb{P}(Z_i|Z_{-i}, V, Y) \propto \mathbb{P}(Z_i|Z_{-i})\mathbb{P}(Y|V, Z), \tag{5.26}$$

where second part is the Gaussian likelihood as in Eq. (5.3). To sample $Z_i|Z_{-i}$ we use the construction from Section 5.B, where we first sample the value of Z_{σ_D} for fixed j:

$$\begin{aligned}
\mathbb{P}(Z_{\sigma_D} = 1|Z) &= \mathbb{P}(\sigma_D = i|M_D = m, Z_i = 1) = \\
&= \frac{\mathbb{P}(M_{D-1} = m-1)\mathbb{P}(Z_{\sigma_D} = 1|M_{D-1} = m-1)}{\mathbb{P}(M_D = m)},
\end{aligned} \tag{5.27}$$

and then uniformly choose and index i to update among $\{i : Z_i = Z_{\sigma_D}\}$. Next Z_i can simply be updated using the process construction 5.6 assuming it to be last. Recall that $\mathbb{P}(M_D = m)$ should be computed beforehand using Eq. (5.15).

CHAPTER 6

Conclusions and suggestions

In this thesis we have investigated inference and modeling of latent geometric structures arising in topic modeling. We have also proposed a Bayesian hierarchical approach for interaction selection in the supervised setup. Our contributions can be summarized as follows:

- Geometric formulation of the Latent Dirichlet Allocation.
- Geometric procedure for estimating topic proportions given a topic simplex.
- Fast parametric geometric algorithm based on the analysis of Centroidal Voronoi Tessellation of a topic polytope.
- Fast nonparametric geometric algorithm based on the analysis of the concentration of mass inside the topic simplex and its coverage with suitable cones and spheres.
- New model for temporal dynamics of the topic polytope and corresponding online and distributed inference algorithm.
- Novel Bayesian hierarchical model for supervised adaptive selection of interactions of arbitrary order among predictor variables.

Our results are majorly related to Latent Dirichlet Allocation and topic modeling, however we believe that geometric viewpoint may lend itself naturally for improving inference in other models and applications. Next we summarize some important directions and open questions of future and ongoing work.

6.1 Going beyond discrete data

Latent Dirichlet Allocation is a model of discrete data, e.g., word frequencies in text documents. On the other hand, algorithms proposed in Chapters 2 and 3 are not restricted

to discrete domain. Moreover, only requirement for performing algorithmic steps is the ability to evaluate inner product between data points. Therefore it may be possible to extend our algorithms to domain of reals and even functions (Ramsay, 2006). Furthermore it is of interest to consider distributions as data points, e.g., short documents equipped with uncertainty "balls" to better account for the noise.

6.1.1 Simplex Nest: geometric admixture modeling

In the ongoing work we explore extension of LDA to the data in the domain of reals, which we call the Simplex Nest model. Below are model, objective function and algorithm sketches.

6.1.1.1 Model Description

Let $\beta_1, \ldots, \beta_K \in \mathbb{R}^V$ be the vertices of the nest and $B := \mathrm{Conv}(\beta_1, \ldots, \beta_K)$ the corresponding nest. Note that unlike LDA, vertices are not constrained to be in unit simplex and V is simply a notation for data dimension. An observation w_m, $m = 1, \ldots, M$ is generated as follows:

$$\theta_m \sim \mathrm{Dir}_K(\alpha), \tag{6.1}$$

$$p_m := \textstyle\sum_{k=1}^K \beta_k \theta_{mk} \text{ an offspring is born}, \tag{6.2}$$

$$w_m | p_m \sim F(\cdot | p_m) \text{ offspring leaves the parent nest.} \tag{6.3}$$

When $F(\cdot)$ is Multinomial and $\beta_1, \ldots, \beta_K$ are constrained to unit simplex, we recover the classic LDA model. Gaussian likelihood kernel could be a natural choice for many applications.

6.1.1.2 Objective function

Our goal is to simultaneously study the topics of each document and cluster the documents (in topic modeling terms), which we assume to have at most $\overline{k}$ clusters. To do such, we utilize the idea of Geometric Dirichlet Mean (GDM) of Chapter 2. Our objective is as follows:

$$\sum_{m=1}^{M} N_m d^2(w_m, B) + W_2^2(\frac{1}{M}\sum_{i=1}^{M} \delta_{\widetilde{w}_m}, H) \tag{6.4}$$

where $\widetilde{w}_m = \underset{x \in B}{\arg\min} \|x - w_m\|$ for $1 \leq m \leq M$, i.e., $\widetilde{w}_m$ is the projection of w_m on B; $H \in \mathcal{O}_{\overline{k}}(B)$ is the set of probability measures with at most $\overline{k}$ atoms, and $d^2(w_m, B) =$

$\inf_{x \in B} \|x - w_m\|^2$ for $1 \leq m \leq M$. We want to minimize w.r.t. B and H. Algorithm 6.1 is an intuitive geometric approach for solving the problem when $\overline{k} = K$.

Some intuition First term in the objective is the Geometric objective of GDM (Eq. (2.3)). It evaluates how close the simplex B is to the data. Second term can be thought of as k-means objective of the projected points, hence it penalizes the volume (and dimension if we try to go nonparametric) of B. Since there are no constraints on $\beta_1, \ldots, \beta_K$, penalizing volume is crucial. Without it, the problem is simply selecting a sufficiently large B in the SVD subspace.

Algorithm 6.1 Geometric Admixture Nesting

Input: observations $w_1, \ldots, w_M$; K; extension parameters $m_1, \ldots, m_K$
Output: simplex vertices $\beta_1, \ldots, \beta_K$

1: $C = \frac{1}{M} \sum_m w_m$ {find data center}
2: $\mu_1, \ldots, \mu_K$ = kmeans$(w_1, \ldots, w_M)$ {initialize atoms of H}
3: **while** objective 6.4 not converged **do**
4: $S_k = \{m : \operatorname{argmin}_{k=1,\ldots,K} \|w_m - \mu_k\|_2^2 = k\}$ for $k = 1, \ldots, K$ {get data partitioning according to H}
5: $C_k := \frac{1}{|S_k|} \sum_{m \in S_k} w_m$ for $k = 1, \ldots, K$
6: $\beta_k := C + m_k(C_k - C)$ for $k = 1, \ldots, K$ {update simplex as in GDM}
7: $\mu_k := \frac{1}{|S_k|} \sum_{m \in S_k} \widetilde{w}_m$ {update atoms of H}
8: **end while**

6.2 Geometric Inference as a general approach

Graphical model formalism and correspondingly variational inference and MCMC offer unifying inference approach for many Bayesian hierarchical models. We have shown that taking geometric viewpoint may lead to more efficient inference, however developing formal geometric viewpoint of a general class of Bayesian hierarchical models remains to be an open and very challenging question. Zoubin Ghahramani in his 2004 review paper of unsupervised learning (Ghahramani, 2004) referred to creating MCMC methods as a form of *art*. Nowadays probabilistic programming languages can automatically perform inference for large class of models, e.g. STAN (Carpenter et al., 2016) based on a variant of Hamiltonian Monte Carlo (Hoffman & Gelman, 2014) and automatic differentiation variational inference (Kucukelbir et al., 2017). However, before sufficiently deep level of understanding was achieved and automatic approaches developed, the art of MCMC and variational inference had to be carried out many times for numerous models. We think that

before considering geometric inference on a higher level of modeling abstraction, it is crucial to solve the geometric puzzle locally, for a variety of models, to gain understanding. Cases that appear particularly interesting are mixture models, when the latent geometric structure is no longer a polytope, demanding for different approaches for exploring concentrations of points in space; Correlated Topic Models (Blei & Lafferty, 2006a), when distribution inside the polytope offers new challenges; supervised and semi-supervised problems.

We think that angular geometric approach proposed in Chapter 3 may be extended to the case of mixture models. Particularly it is fascinating how complex high dimensional structure can be visually observed on a series of 2D plots as in Figure 3.2. Understanding how corresponding plots would look like for mixture models may allow for efficient geometric algorithms. Another interesting direction along the same lines is to investigate distributions of cosine distances and norms. Recall the directional representation from Chapter 3, where $\Delta_0^{V-1} := \{x \in \mathbb{R}^V | x + C_p \in \Delta^{V-1}\}$, $b_k := \beta_k - C_p \in \Delta_0^{V-1}$ for $k = 1, \ldots, K$, $\tilde{p}_m := p_m - C_p \in \Delta_0^{V-1}$ for $m = 1, \ldots, M$ and C_p is a point inside the topic polytope B. The random variables of interest are $Z_{jm} := 1 - \cos(b_j, \tilde{p}_m)$ and $r_m := ||\tilde{p}_m||_2$ for some topic j and document m. Analyzing joint distribution of (Z_{jm}, r_m) may provide insights for mixtures similar to Figure 3.2. For Latent Dirichlet Allocation, suppose C_p is a circumcenter, then circumradius $R := ||b_k||_2$ for any $k = 1, \ldots, K$. Hereafter we omit document index m.

Distribution of $Z_j|r$

$$\begin{aligned} Z_j &= 1 - \cos(b_j, \tilde{p}) = 1 - \frac{<b_j, \sum_k \theta_k b_k>}{Rr} = \\ & 1 - \frac{\sum_{k \neq j} \theta_k <b_k, b_j> + \theta_j R^2}{Rr} = \\ & 1 - \frac{R}{r}\left(\sum_{k \neq j} \theta_k \cos(b_k, b_j) + \theta_j\right). \end{aligned} \tag{6.5}$$

If all topics are orthogonal we can simplify $Z_j = 1 - \frac{R}{r}\theta_j$. Marginal distribution of $\theta_j \sim \text{Beta}(\alpha_j, \alpha_0 - \alpha_j)$, where $\alpha_0 := \sum_k \alpha_k$. Using change of variables we can obtain

$$f_{Z_j|r}(x) = \left(\frac{r}{R}\right)^{\alpha_j} \frac{1}{B(\alpha_0 - \alpha_j, \alpha_j)} (1-x)^{\alpha_j - 1} \left(x\frac{r}{R} + (1 - \frac{r}{R})\right)^{\alpha_0 - \alpha_j - 1} \tag{6.6}$$

If $r = R$, then $Z_j|r = R \sim \text{Beta}(\alpha_0 - \alpha_j, \alpha_j)$. Recall that small values of Z_j correspond to document being close in angle to topic j. Hence, when $r = R$ and $\alpha_j > \alpha_0 - \alpha_j$, document is close to the topic j, which is also true from the Dirichlet distribution perspective of topic proportions θ.

Distribution of r Distribution of r is a sum of dependent random variables $r^2 = \sum_{i=1}^{V} p_i^2$. Burghouts et al. (2008) analyzed such distributions based on the results from extreme value theory. Each p_i is bounded, therefore, based on Theorem 1 of Burghouts et al. (2008), r is a Weibull-distributed random variable:

$$f_r(y) = \frac{\gamma}{\sigma}\left(\frac{y}{\sigma}\right)^{\gamma-1} \exp\left(-\left(\frac{y}{\sigma}\right)^{\gamma}\right) \tag{6.7}$$

We can obtain the joint distribution of (Z_j, r) by combining the two results.

6.3 Modeling geometry

There are two main directions that we would like to outline in this section.

Modeling latent geometric structures. Ability to do computations in parallel is crucial for design of many scalable approaches. E.g., divide-and-conquer is concerned with learning posterior estimates of data batches and aggregating them to obtain overall posterior estimate. Broderick et al. (2013) have proposed similar framework for variational inference. Large datasets often have vast amount of metadata associated to them, which may allow for more principled data partitioning. In Chapter 4 we utilized that scientific articles may be partitioned by journals and year, similarly news articles may be partitioned by publisher and day, and tweets by location and minute to form data batches of manageable size. Then for each of the batches we used CoSAC to extract topics. More generally, at a granular scale, for a fixed time point and location, it is reasonable to believe that a simple approach is sufficient to learn the patterns in a batch of data. Depending on the application, for explaining batch pattern descriptors, one can consider mixture models of various kinds (McLachlan & Peel, 2004), component analysis (Jolliffe, 1986; Hyvärinen et al., 2004), etc. Correspondingly, pattern descriptors could be mixture components means, components of Principal Component Analysis or latent factors of Factor Analysis.

Modeling pattern descriptors may even be a necessity when data privacy is a concern. Merugu & Ghosh (2005) defined a model integration problem for distributed setting where only pattern descriptors for each of the data sources are available due to privacy issues. They proposed using mixture of Gaussians for parametric model integration. We think that mixture is not an appropriate model for pattern descriptors in many cases. Pattern descriptors are often ought to be distinct (e.g., topics, principal components, etc.) and therefore should originate in distinct global patterns. Mixture model based inference violates this constraint as multiple pattern descriptors of a single batch may be assigned to the same

global descriptor. In Chapter 4 we proposed a series of models based on the Beta-Bernoulli processes for modeling dynamic topics and accounting for their distinctiveness. We noted that the latent geometric structure of interest is solely encoded in the base measure H (recall Eq. (4.1)). It is of interest to continue exploring choices of H for other geometric structures such as collections of mixture components, principal components, etc. Once an appropriate base measure is constructed, our models can be utilized to do dynamic and distributed learning amendable to large data sizes.

Developing geometric priors. The Dirichlet prior on topics of the Latent Dirichlet Allocation plays a regularization role (Blei et al., 2003) and does not contain any semantic meaning. In Chapter 3 we introduced notion of the angular separation, which is in agreement with theoretical results of Nguyen (2015), who showed that topic polytope can be consistently estimated without knowing true number of topics, provided that minimum distance between topics is bounded away from zero. These results suggest that it is of interest to construct a geometric prior on topic polytope enforcing well separated vertices. It would also make sense semantically, as topics in practice are desired to be distinct. Tang et al. (2014) mentioned that imposing certain geometric constraints through the prior specification could improve posterior contraction behavior.

One of the directions to achieve separation in the prior is to consider repulsive prior, studied by Petralia et al. (2012) for the case of mixtures. Another approach is "diversity-promoting" modeling (Xie et al., 2016). A geometrically appealing idea is to develop priors on the pairwise distance matrix of the topic simplex vertices, which would allow to directly control the separation. Such matrix is a key object of interest in the field of Distance Geometry (Blumenthal, 1970; Liberti et al., 2014; Liberti & Lavor, 2016). Additionally, this distance matrix can be used to infer almost any geometric property of the simplex: its determinant, known as Cayley-Menger determinant, can be used to compute volume of the simplex, its minors can be used to find the incenter barycentric coordinates, etc.

6.4 Inference of dynamic latent geometric structures

Models that we proposed in Chapter 4 open up a series of inference challenges. We explored an online MAP estimation procedure without revisiting the past. In some cases it could be of interest to solve the problem when all the data is available at once, i.e. to find the global MAP estimate or even the posterior distribution. For Gaussian distributions the problem of global MAP estimation can be addressed via Kalman filtering (Grewal, 2011), but the problem becomes more challenging in the case of the Von Mises-Fisher dynamics used

in our work. Next, even for a case of Gaussian dynamics, the mix-matching introduces extra layer of complexity in the context of global MAP estimation. Potential approach is to combine ideas of Kalman filtering and Viterbi algorithm (Viterbi, 1967) to solve the joint problem, starting from a simpler Gaussian dynamics case. Different approach, additionally allowing for posterior inference, is particle filtering (Doucet & Johansen, 2009). Lastly, it is of interest to consider repulsive priors idea (Petralia et al., 2012; Xie et al., 2016) in the base measure initial state of the dynamics, such that the topic discovery may be done in a more principled way, replacing the geometric heuristic we employed.

6.5 Visualization, inference and modeling of the interactions

Visualization. The question of identifying interactions arises often in applied sciences. While working as a consultant in Consulting for Statistics, Computing and Analytics Research (CSCAR), I had met multiple clients who could benefit from the capabilities of the MiFM (Chapter 5). Applied researchers often lack sufficient programming experience to dive into the Python code on their own. Posterior estimate produced by the MiFM is rich in structure, and even with enough Python experience it may be challenging to navigate through it. We think that a visualization engine for the MiFM posterior could be of broad interest in the applied sciences and in business analytics to drive marketing and sales insights.

Inference. We have investigated some of the density estimation properties of the MiFM posterior (i.e., Theorem 5.1). However, some of the important questions concerning the properties of the posterior distribution remain open. Suppose there is a true interaction structure, then we want to know if this structure is identifiable and how well corresponding posterior mode approximates it. We have shown in Figure 5.2 (d) that proportion of correctly identified interactions empirically is related to the nature of the predictor variables. We think that analyzing the case of binary predictor variables is a promising direction to continue this line of thinking.

Modeling. In Section 5.3.3 we have discussed how MiFM can be applied to categorical variables. We chose to use categories in the hypergraph of interactions, but allowed for different coefficients for the attributes. It may be of interest to consider extending MiFM to capture interactions among attributes in the hypergraph of interactions. For instance, this can be done by adding an extra layer, where given that category is active in the interaction, one

of the attributes is chosen according to Categorical distribution parametrized by a probability vector with Dirichlet prior.

In the MiFM, an interaction was mapped into its numerical realization via cumulative product of the feature values. This is a common way of representing an interaction in the regression setting. Meanwhile, idea of modeling hypergraph of interactions via our FFM$_\alpha$ construction may be used broader, for example in the factor graph context (Ghahramani, 2004). Additionally, function class for interactions may be extended to monomials by allowing for integer values in the hypergraph of interactions matrix representation (e.g., by adopting Beta-Negative Binomial Process construction (Zhou et al., 2012)). Finally, a thorough understanding of the fully nonparametric version of the FFM$_\alpha$ is of interest, that is, when the number of columns is taken to infinity. Such understanding may lead to an extension of the IBP and new modeling approaches in various domains.

BIBLIOGRAPHY

Ahmed, Amr and Xing, Eric P. Timeline: A dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. *arXiv preprint arXiv:1203.3463*, 2012.

Ai, Chunrong and Norton, Edward C. Interaction terms in logit and probit models. *Economics letters*, 80(1):123–129, 2003.

Anandkumar, Anima, Foster, Dean P, Hsu, Daniel J, Kakade, Sham M, and Liu, Yi-Kai. A spectral algorithm for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, pp. 917–925, 2012.

Anderson, James R and Peterson, Carsten. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

Antoniak, Charles E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pp. 1152–1174, 1974.

Arora, Sanjeev, Ge, Rong, Kannan, Ravindran, and Moitra, Ankur. Computing a nonnegative matrix factorization–provably. In *Proceedings of the 44th annual ACM symposium on Theory of computing*, pp. 145–162. ACM, 2012.

Arora, Sanjeev, Ge, Rong, Halpern, Yonatan, Mimno, David, Moitra, Ankur, Sontag, David, Wu, Yichen, and Zhu, Michael. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 280–288, 2013.

Bardenet, Rémi, Doucet, Arnaud, and Holmes, Chris. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(1):1515–1557, 2017.

Belkin, Mikhail, Niyogi, Partha, and Sindhwani, Vikas. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7(Nov):2399–2434, 2006.

Betancourt, Michael. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, 2017.

Blei, D. M. and Lafferty, J. Correlated topic models. In *Advances in Neural Information Processing Systems*, pp. 147–154, 2006a.

Blei, D. M. and Lafferty, J. D. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006b.

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

Blei, David M, Griffiths, Thomas L, and Jordan, Michael I. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7, 2010.

Blei, David M, Kucukelbir, Alp, and McAuliffe, Jon D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Blumenthal, Leonard Mascot. *Theory and applications of distance geometry*. Chelsea New York, 1970.

Brambor, Thomas, Clark, William Roberts, and Golder, Matt. Understanding interaction models: Improving empirical analyses. *Political Analysis*, 14(1):63–82, 2006.

Brent, R. P. *Algorithms for minimization without derivatives*. Courier Corporation, 2013.

Broderick, Tamara, Boyd, Nicholas, Wibisono, Andre, Wilson, Ashia C, and Jordan, Michael I. Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, pp. 1727–1735, 2013.

Bronstein, Michael M, Bruna, Joan, LeCun, Yann, Szlam, Arthur, and Vandergheynst, Pierre. Geometric deep learning: going beyond Euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, 2017.

Bryant, Michael and Sudderth, Erik B. Truly nonparametric online variational inference for hierarchical Dirichlet processes. In *Advances in Neural Information Processing Systems*, pp. 2699–2707, 2012.

Burghouts, Gertjan, Smeulders, Arnold, and Geusebroek, Jan-Mark. The distribution family of similarity distances. In *Advances in Neural Information Processing Systems*, pp. 201–208, 2008.

Carpenter, Bob, Gelman, Andrew, Hoffman, Matt, Lee, Daniel, Goodrich, Ben, Betancourt, Michael, Brubaker, Michael A, Guo, Jiqiang, Li, Peter, Riddell, Allen, et al. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20(2):1–37, 2016.

Cheng, Chen, Xia, Fen, Zhang, Tong, King, Irwin, and Lyu, Michael R. Gradient boosting factorization machines. In *Proceedings of the 8th ACM Conference on Recommender systems*, pp. 265–272. ACM, 2014.

Cordell, Heather J. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10(6):392–404, 2009.

Cover, Thomas M. and Thomas, Joy A. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

Cristianini, Nello and Shawe-Taylor, John. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

Damle, Anil and Sun, Yuekai. A geometric approach to archetypal analysis and nonnegative matrix factorization. *Technometrics*, 59(3):361–370, 2017.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41 (6):391, Sep 01 1990.

Ding, Chris and He, Xiaofeng. K-means clustering via principal component analysis. In *Proceedings of the 21st International Conference on Machine Learning*, pp. 29, 2004.

Ding, Chris, Li, Tao, and Peng, Wei. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. In *AAAI*, volume 42, pp. 137–143, 2006.

Doucet, Arnaud and Johansen, Adam M. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12(656-704):3, 2009.

Du, Qiang, Faber, Vance, and Gunzburger, Max. Centroidal Voronoi Tessellations: Applications and algorithms. *SIAM Review*, 41(4):637–676, 1999.

Fan, Jianqing and Lv, Jinchi. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.

Feldman, Ronen and Sanger, James. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

Ferguson, Thomas S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pp. 209–230, 1973.

Fox, Emily, Jordan, Michael I, Sudderth, Erik B, and Willsky, Alan S. Sharing features among dynamical systems with Beta processes. In *Advances in Neural Information Processing Systems*, pp. 549–557, 2009.

Freudenthaler, Christoph, Schmidt-Thieme, Lars, and Rendle, Steffen. Bayesian factorization machines. 2011.

Ghahramani, Zoubin. Unsupervised learning. In *Advanced lectures on machine learning*, pp. 72–112. Springer, 2004.

Ghahramani, Zoubin and Griffiths, Thomas L. Infinite latent feature models and the Indian buffet process. In *Advances in Neural Information Processing Systems*, pp. 475–482, 2005.

Ghosal, Subhashis, Ghosh, Jayanta K, Ramamoorthi, RV, et al. Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158, 1999.

Golubitsky, Oleg, Mazalov, Vadim, and Watt, Stephen M. An algorithm to compute the distance from a point to a simplex. *ACM Commun. Comput. Algebra*, 46:57–57, 2012.

Grewal, Mohinder S. Kalman filtering. In *International Encyclopedia of Statistical Science*, pp. 705–708. Springer, 2011.

Griffiths, Thomas L and Ghahramani, Zoubin. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

Griffiths, Thomas L and Steyvers, Mark. Finding scientific topics. *PNAS*, 101(suppl. 1): 5228–5235, 2004.

Harshman, Richard A. Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis. 1970.

Hartigan, J. A. and Wong, M. A. Algorithm as 136: A K-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (applied Statistics)*, 28(1):100–108, 1979.

Himmelstein, Daniel S, Greene, Casey S, and Moore, Jason H. Evolving hard problems: generating human genetics datasets with a complex etiology. *BioData mining*, 4(1):1, 2011.

Hjort, Nils Lid. Nonparametric Bayes estimators based on Beta processes in models for life history data. *The Annals of Statistics*, pp. 1259–1294, 1990.

Ho, N., Nguyen, X., Yurochkin, M., Bui, H., Huynh, V., and Phung, D. Multilevel clustering via Wasserstein means. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Hoffman, Ma. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, May 2013.

Hoffman, Matthew, Bach, Francis R, and Blei, David M. Online learning for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, pp. 856–864, 2010.

Hoffman, Matthew D and Gelman, Andrew. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1): 1593–1623, 2014.

Hofmann, Thomas. Probabilistic latent semantic analysis. In *Proceedings of the 15th conference on Uncertainty in Artificial Intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc., 1999.

Hong, Liangjie, Dom, Byron, Gurumurthy, Siva, and Tsioutsiouliklis, Kostas. A time-dependent topic model for multiple text streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 832–840. ACM, 2011.

Hsu, Wei-Shou and Poupart, Pascal. Online Bayesian moment matching for topic modeling with unknown number of topics. In *Advances in Neural Information Processing Systems*, pp. 4529–4537, 2016.

Hyvärinen, Aapo, Karhunen, Juha, and Oja, Erkki. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

Jiang, Ke, Kulis, Brian, and Jordan, Michael I. Small-variance asymptotics for exponential family Dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pp. 3158–3166, 2012.

Jolliffe, Ian T. Principal component analysis and factor analysis. In *Principal component analysis*, pp. 115–128. Springer, 1986.

Jones, Eric, Oliphant, Travis, Peterson, Pearu, et al. SciPy: Open source scientific tools for Python, 2001–. URL http://www.scipy.org/.

Jordan, Michael I, Ghahramani, Zoubin, Jaakkola, Tommi S, and Saul, Lawrence K. An introduction to variational methods for graphical models. *Machine Learning*, 37(2): 183–233, 1999.

Jordan, Michael I et al. Graphical models. *Statistical Science*, 19(1):140–155, 2004.

Kitagawa, Genshiro. Monte Carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

Kucukelbir, Alp, Tran, Dustin, Ranganath, Rajesh, Gelman, Andrew, and Blei, David M. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18(1):430–474, 2017.

Kuhn, Harold W. The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 2(1-2):83–97, 1955.

Kulis, Brian and Jordan, Michael I. Revisiting k-means: New algorithms via Bayesian nonparametrics. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1131–1138, 2012.

Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788, 1999.

Liberti, Leo and Lavor, Carlile. Six mathematical gems from the history of distance geometry. *International Transactions in Operational Research*, 23(5):897–920, 2016.

Liberti, Leo, Lavor, Carlile, Maculan, Nelson, and Mucherino, Antonio. Euclidean distance geometry and applications. *Siam Review*, 56(1), 2014.

Lloyd, S. Least squares quantization in PCM. *Information Theory, IEEE Transactions on*, 28(2):129–137, Mar 1982.

Ma, Yi-An, Chen, Tianqi, and Fox, Emily. A complete recipe for stochastic gradient MCMC. In *Advances in Neural Information Processing Systems*, pp. 2917–2925, 2015.

MacQueen, J. B. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297, 1967.

Mandt, Stephan, Hoffman, Matthew D, and Blei, David M. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

Manton, Jonathan H, Mahony, Robert, and Hua, Yingbo. The geometry of weighted low-rank approximations. *Signal Processing, IEEE Transactions on*, 51(2):500–514, 2003.

Mardia, Kanti V and Jupp, Peter E. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

Mcauliffe, J. D. and Blei, D. M. Supervised topic models. In *Advances in Neural Information Processing Systems*, pp. 121–128, 2008.

McLachlan, Geoffrey and Peel, David. *Finite mixture models*. John Wiley & Sons, 2004.

Merugu, Srujana and Ghosh, Joydeep. A distributed learning framework for heterogeneous data sources. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 208–217. ACM, 2005.

Newman, David, Smyth, Padhraic, Welling, Max, and Asuncion, Arthur U. Distributed inference for Latent Dirichlet Allocation. In *Advances in Neural Information Processing Systems*, pp. 1081–1088, 2008.

Nguyen, Trung V, Karatzoglou, Alexandros, and Baltrunas, Linas. Gaussian process factorization machines for context-aware recommendations. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 63–72. ACM, 2014a.

Nguyen, Vu, Phung, Dinh, Nguyen, XuanLong, Venkatesh, Svetha, and Bui, Hung Hai. Bayesian nonparametric multilevel clustering with group-level contexts. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 288–296, 2014b.

Nguyen, XuanLong. Inference of global clusters from locally distributed data. *Bayesian Analysis*, 5(4):817–845, 2010.

Nguyen, XuanLong. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21(1):618–646, 02 2015.

Nguyen, XuanLong and Gelfand, Alan E. The Dirichlet labeling process for clustering functional data. *Statistica Sinica*, 21(3):1249, 2011.

Olver, Frank W. J., Lozier, Daniel M., Boisvert, Ronald F., and Clark, Charles W. NIST handbook of mathematical functions, cambridge university press, 2010. URL http://dlmf.nist.gov/8.17.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

Petralia, Francesca, Rao, Vinayak, and Dunson, David B. Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pp. 1889–1897, 2012.

Pollard, David. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1): 135–140, 01 1981.

Pritchard, Jonathan K, Stephens, Matthew, and Donnelly, Peter. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Rabiner, Lawrence and Juang, B. An introduction to hidden Markov models. *IEEE ASSP Magazine*, 3(1):4–16, 1986.

Ramsay, James O. *Functional data analysis*. Wiley Online Library, 2006.

Rasiwasia, Nikhil and Vasconcelos, Nuno. Latent Dirichlet Allocation models for image classification. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11): 2665–2679, 2013.

Řehůřek, Radim and Sojka, Petr. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Rendle, Steffen. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 995–1000. IEEE, 2010.

Rendle, Steffen, Gantner, Zeno, Freudenthaler, Christoph, and Schmidt-Thieme, Lars. Fast context-aware recommendations with factorization machines. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 635–644. ACM, 2011.

Rodriguez, Abel, Dunson, David B, and Gelfand, Alan E. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

Roweis, Sam and Ghahramani, Zoubin. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999.

Roweis, Sam T. EM algorithms for PCA and SPCA. In *Advances in Neural Information Processing Systems*, pp. 626–632, 1998.

Sayyareh, Abdolreza. A new upper bound for Kullback-Leibler divergence. *Appl. Math. Sci*, 67:3303–3317, 2011.

Tang, Jian, Meng, Zhaoshi, Nguyen, Xuanlong, Mei, Qiaozhu, and Zhang, Ming. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 190–198, 2014.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

Templeton, Alan R. Epistasis and complex traits. *Epistasis and the evolutionary process*, pp. 41–57, 2000.

Thibaux, Romain and Jordan, Michael I. Hierarchical Beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics*, pp. 564–571, 2007.

Tibshirani, Robert. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

Vapnik, Vladimir. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Vavasis, Stephen A. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 20(3):1364–1377, 2009.

Viterbi, Andrew. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE transactions on*, 13(2):260–269, 1967.

Wainwright, Martin J, Jordan, Michael I, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

Wang, Chong, Paisley, John, and Blei, David. Online variational inference for the hierarchical Dirichlet process. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, pp. 752–760, 2011.

Wang, Xuerui and McCallum, Andrew. Topics over time: a non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 424–433. ACM, 2006.

Wasserman, Larry. Topological data analysis. *Annual Review of Statistics and Its Application*, 2016.

Westendorp, Gerard. A formula for the n-circumsphere of an n-simplex, April 2013. Retrieved from http://westy31.home.xs4all.nl/.

Xie, Pengtao, Zhu, Jun, and Xing, Eric. Diversity-promoting Bayesian learning of latent variable models. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 59–68, 2016.

Xu, Wei, Liu, Xin, and Gong, Yihong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pp. 267–273. ACM, 2003.

Yurochkin, M., Guha, A., and Nguyen, X. Conic Scan-and-Cover algorithms for non-parametric topic modeling. In *Advances in Neural Information Processing Systems*, pp. 3881–3890, 2017a.

Yurochkin, M., Nguyen, X., and Vasiloglou, N. Multi-way interacting regression via factorization machines. In *Advances in Neural Information Processing Systems*, pp. 2595–2603, 2017b.

Yurochkin, Mikhail and Nguyen, XuanLong. Geometric Dirichlet Means Algorithm for topic inference. In *Advances in Neural Information Processing Systems*, pp. 2505–2513, 2016.

Zhou, Mingyuan, Hannah, Lauren, Dunson, David B, and Carin, Lawrence. Beta-negative binomial process and Poisson factor analysis. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pp. 1462–1471, 2012.

Zhu, Ji, Rosset, Saharon, Hastie, Trevor, and Tibshirani, Rob. 1-norm support vector machines. In *Advances in Neural Information Processing Systems*, pp. 49–56, 2004.

Zou, Hui and Hastie, Trevor. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.