Inference and Interpretability in Latent Variable Modeling

by

Aritra Guha

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2020

Doctoral Committee:

Associate Professor Long Nguyen, Chair
Professor Moulinath Banerjee
Assistant Professor Zhenke Wu
Assistant Professor Gongjun Xu

Aritra Guha

aritra@umich.edu

ORCID iD: 0000-0003-3866-2918

To Maa, Baba, Chhottu, Swagata and Jethu

ACKNOWLEDGEMENTS

I write this with great grief as devastating pictures of Cyclone Amphan having wreaked havoc on my beloved city crop up on my news feed. To Kolkata, the city that shall always be "home".

My PhD journey is drawing to an end. As I reflect on the past 6 years, I realise how fortunate I have been to meet and interact with so many great people each of whom have contributed significantly to my life.

First and foremost, I would like to thank my adviser, Long Nguyen, for his amazing guidance and support throughout this journey. Long has been a constant source of inspiration and positive energy, which has helped me immensely to navigate the academic challenges of PhD life. I have been constantly fascinated by his knowledgeability and depth of understanding. The innumerable interactions with him have taught me to appreciate every problem no matter how insignificant it may seem. Long is extremely caring towards his students and his door has always remained open for any query that his students may have, academic or otherwise. Time and again there have been scenarios where a brief 5-minute query has led to hours of discussions with Long, which have taught me so much about Statistics. I am also thankful to Long for introducing me to a wonderful group of peers. Long is a complete mentor. He not only ensures the appropriate academic growth of his students but also cares a lot for their mental well-being. The occasional Thanksgiving and Christmas parties have been extremely cherishable moments and I would like to thank Long and his family for graciously hosting us so many times. Long has been the best possible adviser and an absolute role model for me, not only for his academic achievements but also for his kindness, humility and outlook towards life.

I would also like to thank Moulinath Banerjee, Gongjun Xu and Zhenke Wu for agreeing to be on my thesis defense committee, and Stilian Stoev for his invaluable advice in my first year of graduate life. I am grateful to Bodhida (Bodhisattva Sen) for his guidance over the summer of 2013 and invaluable advice over the years. Yuekai Sun, Yves Atchade and Long have also painstakingly written so many letters of recommendation for me and I cannot thank them enough for their efforts. I also take this opportunity to thank all the faculty members for their graduate courses that have shaped up my knowledge of Statistics, and all the administrative staff for their help with anything and everything I did not have a solution to. I believe that a PhD is a culmination of not just 5-6 years of work but a life time of effort from so many others who contribute in numerous ways. For that I am especially grateful to all my teachers at school (Aditya Academy and Bharatiya Vidya Bhavan) and at ISI, especially Paromita Mam, Sabita Mam and Mira Mam for shaping up my academic career.

I have had the great fortune to meet and make so many new friends in the Department. I have been extremely fortunate to have worked and interacted with Nhat, Mikhail, Rayleigh, Vincenzo, Bach, Hossein, Giuseppe, Tim, Jiahui, Jawad and so many others from whom I have learnt a lot.

Over the years at school and ISI, I have made great many friends who I have shared innumerable life experiences with: Soumya Subhra, Arkopal, Debanjan, Joydeep, Abhishek (Tewari), Sagnikda, Rupak, Paromita, Pragya, Surya, Rohit, Saketh, Avi, Soumendu, Sutanoy, Purvasha, Snehashis, Arnab (Hazra), Arkaprava, Sayan, Rudrendu, Abir, Deepanjan, Suprajo and so many more. I am forever indebted to each and every one of them for their support throughout my life. I am honored to have found an amazing family away from home in Ann Arbor: Anwesha, Diptavo, Rounakda, Shrijitadi, Avijitda, Debarghya, Subha, Rupam, Debraj, Soumik, Moulida and Bharamardi, thank you for the countless adda sessions and late night hangouts and coffee chats. I am also grateful to all the members of Ann Arbor and Troy cricket fraternities for the camraderie and all the cricket outings. They definitely made the summers all that more exciting.

Finally, this thesis is dedicated to my family. I would like to express my deep gratitude to my parents, Aruna and Abhijit Guha for their selfless love, unconditional sacrifice, support and belief in me since the day I was born. I would like to thank all my family members for their love, and especially my grandmother Binapani Dattagupta, and my uncles Arup and Barun Dattagupta for their guidance in life. I consider myself extremely lucky to have met Swagata. I am grateful to her for all the love and support over the past few years and for all the discussions on realities and perspectives. Her presence made the dark times a lot brighter. Special thanks also go to my brother, Arkajit for bearing with all my tantrums and leg-pullings over the past 23 years of life.

Last but not the least, I am especially thankful to Jethu, Nandadulal Datta for instilling in me the love for Mathematics and for teaching me about circles. I would never have made it this far if not for his guidance during my school years. To end, I would like to iterate one of Jethu's numerous one-liners, "God is a circle whose center is everywhere, circumference nowhere".

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

In this age of technology, more and more data is generated as an outcome of complex processes through heterogeneous mechanisms. Statistical models therefore need to invoke the complexities for appropriate inference. One such situation is when data is generated from heterogeneous sub-populations. Hierarchical models form the state-of-the-art methods for such scenarios. However, Markov Chain Monte Carlo methods, which form the traditional inference method can be quite cumbersome to implement for such large scale models, resulting in high time complexities. Sometimes, they may also suffer from inconsistency issues. On the other hand, mean-field variational inference methods even though fast, can suffer from inaccuracy in estimation. This dissertation focuses on understanding such complex models and drawing inference from each of those sub-populations and develops alternative techniques for inference, with statistical guarantees.

Our specific contributions are as follows. We provide an in-depth analysis of two specific types of latent variable models, namely, mixture and admixture models and also develop an unsupervised learning scheme with applications to autonomous vehicles.

In the mixture model context, firstly, we develop an understanding of the posterior contraction behavior of parameter estimation in the case of infinite mixture models, corresponding to two choices of priors, one parametric and the other nonparametric. Next, we provide an in-depth analysis of Bayesian mixtures under various misspecification settings and provide an asymptotic characterization pertaining to such scenarios. Our study reveals a deep perception of the role, the kernel plays on the statistical decisions of a practitioner. Next, in the context of admixture models, we develop a geometric estimation mechanism to the well-known Latent Dirichlet Allocation model. Finally, we provide a model-free inference scheme to robustly estimate and evaluate parameters of various sub-populations, in the applied setting when the heterogeneous sub-populations for data generation are derived from car driving scenarios, via the use of unsupervised learning.

CHAPTER I

Introduction

With the advent of technology, a large amount of today's data is generated by means of complex mechanisms. Data may be available in various forms - for example, unlabelled data as in images, tweets, articles or time series data as in daily weather reports, traffic scenarios including but not limited to inter-vehicular interactions. Moreover, data available is often high-dimensional in nature as vast amount of information is generated at low costs. For example, large-scale biological datasets obtained through next-generation sequencing, proteomics or brain-imaging are often high-dimensional. Other kinds of data may be more personal in nature, such as mobile app based monitoring of driving, health or other individual-specific activities. Any suitable Statistical method therefore needs to be considerate of the appropriate complexities of the data involved, for efficient inference. The accommodations of appropriate Statistical models are two-fold. Firstly, the model should be mindful of the heterogeneity of the processes and be suitable for inferring from each of the subpopulations in the data-generating scheme. Secondly, the inference scheme should also be scalable for large dimensional data. Unsupervised learning schemes such as probabilistic Principal component analysis (cf. *Tipping and Bishop* (1999); *Roweis and Ghahramani* (1999)), Factor analysis (cf. *Anderson and Rubin* (1956)), Independent component analysis (cf. *Hyvärinen et al.* (2004)) are well-known scalable techniques for large datasets. *Ghahramani* (2004) provides a succint account of unsupervised learning techniques. However, they often fail to capture the heterogeneity of the data because of their model-free constitution. On the other hand, Bayesian hierarchical models such as Mixture models (cf. *McLachlan and Basford* (1988)), Admixtures (cf. *Pritchard et al.* (2000)), Hierarchical Dirichlet processes (HDP) (cf. *Teh et al.* (2006)), Hidden Markov models (cf. *Baum and Petrie* (1966); *Baum et al.* (1970)) form the state-of the-art methods for modelling of heterogeneous subpopulations. However, they may suffer from inconsistency issues (cf. *Miller and Harrison* (2014)). Markov Chain Monte Carlo algorithms (cf. *Griffiths and Steyvers* (2004); *Escobar and West* (1995); *MacEachern and Mueller* (1998); *Neal* (2000); *Teh et al.* (2006); *Fox et al.* (2009)) form the state-of-the-art methods for inference with hierarchical models. However, a large number of latent variables in combination with complex modeling structures often makes it difficult for MCMC algorithms to be scalable. A useful alternative is provided by Variational Inference algorithms (cf. *Jordan et al.* (1999); *Blei et al.* (2003); *Blei and Jordan* (2006); *Hoffman et al.* (2013); *Mandt et al.* (2017)). Several recent papers (cf. *Wang and Blei* (2018, 2019)) explore the asymptotic consistency for Variational Inference algorithms. However, finite sample outcomes for VI algorithms do not produce the accurate Posterior distribution. Moreover, asymptotic consistency cannot be guaranteed for graphical models with complex structures. As a result, the need for scalable and statistically efficient algorithms is ever-present. This thesis focuses on understanding inference-related questions for such complex models and develops appropriate techniques for scalable inference with statistical guarantees. For this work, our focus is only on mixture and admixture models.

The remainder of this chapter elaborates on the key contributions of this thesis.

Relevant background material is included in each section to make the sections self-contained.

1.1 Model complexity, estimation and interpretability in Bayesian mixture modeling

Mixture models form the basic building blocks in latent variable modeling and provide an interpretable way for a statistician to analyze data from heterogeneous population sources (cf. *McLachlan and Basford* (1988); *Lindsay* (1995); *Mengersen et al.* (2011)). With practical applicability in model-based clustering techniques and modeling complex distributions(cf. *McLachlan and McGriffin* (1994)), mixture models provide a wide range of scopes for statistical modeling. Mixture models view data as samples from an assembly of latent sub-populations, each assuming its own distribution with corresponding parameters.

More concretely, we work with the following specific formulation of mixture models. Consider discrete mixing measures $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$. Here, $\boldsymbol{p} = (p_1, \ldots, p_k)$ is a vector of mixing weights, while atoms $\{\theta_i\}_{i=1}^{k}$ are elements in a given compact space Θ. Mixing measure G is combined with a kernel function $f(\cdot|\theta)$ with respect to Lebesgue measure μ to yield a mixture density:

$$p_G(\cdot) = \int f(\cdot|\theta) \mathrm{d}G(\theta) = \sum_{i=1}^{k} p_i f(\cdot|\theta_i).$$

When $k < \infty$, we call this a *finite mixture model* with k components. For an *infinite mixture model*, k is allowed to take the value ∞.. The atoms θ_i's are representatives of the underlying subpopulations.

The choice of the kernel f, and the prior on the unknown distribution G, affect the outcome of inference drastically for a practitioner of Bayesian mixture models. For a Bayesian, the choice of priors for mixture models is essentially restricted to two primary options:

(i) a nonparametric prior via use of Dirichlet process distribution on the mixing measures.

(ii) a parametric counterpart of Dirichlet process mixtures via use of suitable priors (eg. Poisson) on number of components.

Estimation of the true but unknown number of mixture components is an important inference question relative to mixture models. The choice of the prior drastically influences the efficiency of estimation corresponding to the parameters representing these components. A common misconception that may have initially contributed to the enthusiasm for Bayesian nonparametric modeling is that the use of such nonparametric models eliminates altogether the need for determining the number of mixture components, because the learning of such a quantity is "automatic" from the posterior samples of the mixing measure. However, *Miller and Harrison* (2014) explicitly demonstrated that the common practice of drawing inference about the number of mixture components via the DP mixture by counting the number of support points in the sample of the Dirichlet posterior leads to an asymptotically inconsistent estimate.

With this in context, we propose *Merge-Truncate-Merge* in Chapter II of this thesis. It is a post-processing algorithm that resolves the inconsistency issue by allowing to consistently estimate the true number of components with Dirichlet process mixtures. Moreover, the algorithm outputs posterior samples that retain the original parameter contraction rates pertaining to samples from the Dirichlet process posterior. Additionally, we also show that the parametric choice of prior given in (ii) yields optimal rates of convergence of the mixing measure (up to a logarithmic term), in addition to correctly recovering the number of mixture components, under considerably weak conditions.

Following George Box's famous quote, "all models are wrong, but some are useful", a natural question that may arise in the context of inference related to mixtures is: what happens to a mixture model based statistical procedure when the model is actually misspecified?

Misspecification may be of various different categories. Misspecifications of the kernel may skew the performance of the methods in question and lead to an incorrect conclusion because of non-robustness properties of the model. With regard to interpretability of parameter estimation under the misspecified model regime, our results reveal several new insights. Given the model is misspecified, the statistician might choose to indulge in heavy-tailed kernels which allow for fast contraction of parameter estimates, thereby implying that a given data set probably has relatively faster influence on the movement of mass from the prior to the posterior distribution. In that regard, Laplace location mixtures may be preferred to Gaussian location mixtures, provided that the bias due to misspecification is not too large. On the other hand, when this is not the case, it is advisable to have a more "conservative" approach by adopting Gaussian kernels instead, despite the latter's lagging posterior contraction behavior. Overall, the ultimate model choice under misspecification will reside on resolving the tension between the aforementioned bias and contracting variance.

Other kinds of misspecifications may arise out of a biased choice for the support of the prior. Theoretical results often rely on critical assumptions which may create such scenarios in practice. For example, a vast array of works that deal with asymptotically optimal estimation procedures for the population density (cf. *Ghosal et al.* (1999);

Ghosal and van der Vaart (2007); *Shen et al.* (2013)) rely on the critical assumption that the space of parameters is bounded. It is also a common assumption for works that deal with the theoretical understanding of the parameter estimation regime (cf. *Nguyen* (2013); *Gao and van der Vaart* (2016); *Scricciolo* (2017)). However, such an assumption may be unfavorable in situations where this bounded support in incorrectly specified. While a small support allows for misspecification of the parameter space thereby leading to a bias in estimation, a large or unbounded support leads to a slow contraction rate thereby resulting in higher variability in estimation. In practice, it is common to allow the base prior for the parameters to have an unbounded support to overcome the additional step of estimating the support of the prior. As a result there is a glaring mismatch between the practical application and theoretical understanding for parameter estimation problems with Bayesian nonparametric priors. In Chapter III of this thesis, we provide a solution to this problem via the use of sieve estimates.

Sieve methods have been implemented in the density estimation context by many authors such as *Ghosal and van der Vaart* (2001); *Shen and Wong* (1994); *Wong and Shen* (1995); *Van de Geer* (1993); *Birge and Massart* (1998). However, to the best of our knowledge, there has been no such treatment in the context of parameter estimation. As part of the sieve estimation procedure we allow the support of the prior to change gradually with the sample-size. This enables us to overcome the bias in parameter estimation while appropriately increasing the variance at a suitably chosen rate, thereby providing a solution to the eternal bias-variance trade-off problem in this context. Our theory reveals that this rate of change of support is much faster when the chosen kernel is light-tailed as compared to heavy-tailed ordinary smooth kernels. This might be counterintuitive to the results in Chapter II since it implies that for supersmooth kernels a possible large change in bias of estimation results in only a relatively negligible change in variance. To address this puzzling issue, we develop a novel metric that generalizes the well-known Wasserstein metric, which is the popular choice of metric for parameter estimation. The use of this novel metric, which we call the "Orlicz-Wasserstein distance", leads us to a deeper understanding of the behavior of the posterior atoms. We show that for light-tailed kernels, the posterior atoms highly populate the neighbourhood of the true atoms with little contribution from other regions of the parameter space. As a result, the contribution to the variance mostly arises from the neighbourhoods of true atoms. Therefore, beyond a certain point, change in variance is affected marginally by change in bias.

1.2 Scalable and efficient geometric algorithms for probabilistic models

Hierarchical and latent variable models broaden the scope of inference. However, a key challenge with the use of hierarchical models is fast and efficient computation of hyperparameters, both in the parametric and nonparametric context. The meaningful inferential and methodological questions involved in hierarchical modeling are:

(I) Can we have a generic modeling scheme which encompasses a large number of data generating procedures?

(II) Can we efficiently estimate the parameters of the models under the knowledge or lack thereof of the number of latent subpopulations?

Scalable estimation for parametric models: For many complex probabilistic models, especially those with latent variables, the probability distribution of interest can be represented as an element of a convex polytope in a suitable ambient space, for which model fitting may be cast as the problem of finding the extreme points of the polytope. For instance, a mixture density can be identified as a point in a convex set of distributions whose extreme points are the mixture components. The well-acclaimed Latent Dirichlet Allocation (LDA) (cf. *Blei et al.* (2003)) model, which is used for analysis of text data, is another example of such a model. In the following, we provide a brief overview of the LDA model.

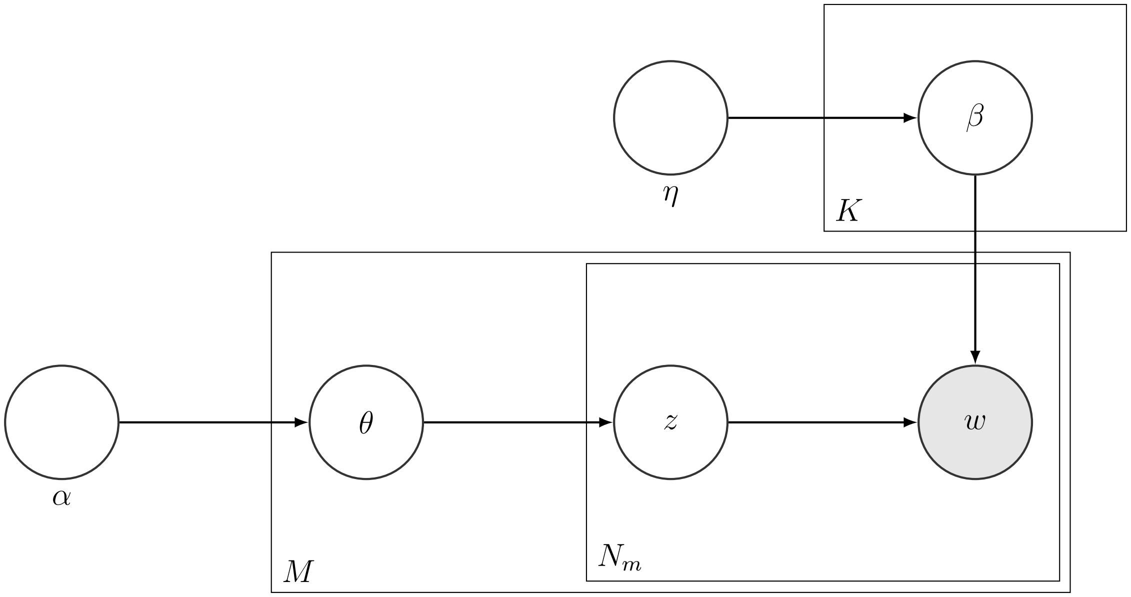
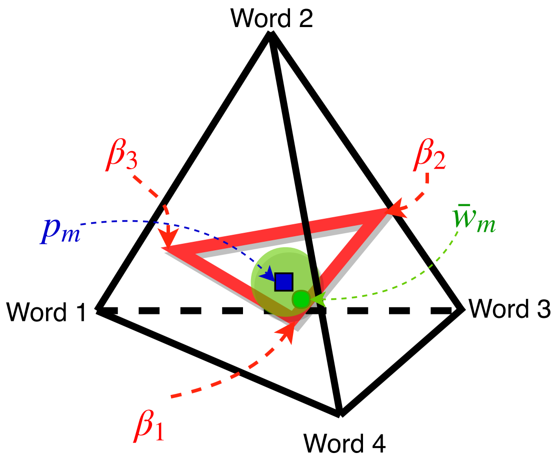


Figure 1.1: *K topics, M documents, N_m exchangeable words in each document w*

Latent Dirichlet Allocation Figure 1.1 provides a graphical model description of the LDA model. The generative process of the model can be described as follows.

Consider a corpus of M documents with V denoting the number of words in the vocabulary. Suppose there are K topics. Let $\alpha \in \mathbb{R}_+^K, \eta \in \mathbb{R}_+^V$ be the respective hyperparameters corresponding to the topic distributions in a document and the word



;

Figure 1.2: *Toy example of LDA*

distributions in a topic respectively. The topics are generated as follows:

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

For each of the M documents, generate a topic proportionality vector as:

$$\theta_m|\alpha \sim \mathrm{Dir}_V(\alpha), \quad \text{for } m = 1, \ldots, M. \tag{1.1}$$

For each of the N_m words in document m, generate a topic label z, and a sample word d from the corresponding topic as:

$$z_n|\theta_m \sim \ \mathrm{Cat}(\theta_m);\, d_n|z_n \sim \ \mathrm{Cat}(\beta_{z_n}) \quad \text{for} n = 1, \ldots, N_m. \tag{1.2}$$

Admixture model (cf. *Pritchard et al.* (2000)), which is popular in genetics is equivalent to the LDA model. The LDA model embeds the topic distributions in the probability simplex and therefore is convenient for word-document frequency distributions. This simplicial structure of the model is amenable to efficient inference and therefore prove useful for other datatypes as well. In Chapter IV, we propose Dirichlet Simplex Nest (DSN), a class of probabilistic models that generalizes the Latent Dirichlet Allocation. By viewing data as noisy observations from the low-dimensional affine hull that contains a simplex, our model shares an assumption that can be found in classical factor analysis, non-negative matrix factorization (NMF) models (cf. *Lee and Seung* (2001)), as well as in topic models (cf. *Arora et al.* (2012b)). The class of models provides a probabilistic justification for these methods, which often impose an additional geometric condition on the model known as *separability* that identifies the model parameters in a way that permits efficient estimation (cf. *Arora et al.* (2012a)). Moreover the DSN modeling provides an arguably more effective approach to archetypal analysis and non-negative matrix factorization for *non-separable* data as well.

The key challenge for inference using the DSN model lies in scalable and efficient parameter estimation. While Hamiltonian Monte Carlo lacks scalability guarantees for such complexly structured models, NMF algorthms perform poorly when the *separability* condition cannot be guaranteed. Variational Inference (cf. *Blei et al.* (2003)) forms the go-to scalable algorithm for inference with the LDA model.However, the questions of statistical efficiency with Variational Inference algorithms remain mostly unexplored. Starting with an original geometric technique of *Yurochkin and Nguyen* (2016), we provide in Chapter IV Voronoi Latent Admixture (VLAD) algorithm, a novel inference algorithm that accounts for the convex geometry and low dimensionality of the latent simplex structure endowed with a Dirichlet distribution. This allows for more effective learning of asymmetric simplicial structures and the Dirichlet's concentration parameter for the general DSN model, and hence, expands its applicability to a broad range of data distributions. More specifically, VLAD can be used for scalable and efficient estimation corresponding to the LDA model setup with greater accuracy than the state-of-the-art Variational Inference algorithms. We also establish statistical consistency and estimation error bounds for the proposed algorithm.

1.3 Evaluation of primitive scenarios for autonomous vehicles

While model-based inference schemes provide increased interpretability, they can often be computationally expensive, especially for huge datasets. Moreover, structural specificity of models may often hinder the general applicability of model-based schemes. On the other hand, model-free mechanisms find a wider variety of applications.

In autonomous vehicle research, there has been a lot of work dedicated to understanding traffic scenarios. Analysis and recognition of driving styles are profoundly important to intelligent transportation and vehicle calibration. Unfortunately, it can be hard to manually select a representative subset of scenarios or potentially computationally expensive to annotate them because of limited prior knowledge. Several recent papers employ unsupervised and empirical approaches to extract primitive driving patterns from time series driving data without prior knowledge of the number of these patterns (cf. *Wang and Zhao* (2017); *Taniguchi et al.* (2015); *Bender et al.* (2015)). However, this wide variety of unsupervised approaches leads to the obvious question of which method to choose.

With regards to that we develop a geometric invariant metric to compare different driving scenarios in Chapter V of this thesis. This novel metric is generally applicable and therefore can be easily extended to evaluate cluster efficiencies and stabilities of the existing clustering methods.

Research on traffic encounters, so far, has been primarily restricted to understanding traffic driving encounters as objects. We propose a general framework to understand distributional patterns in driving styles and provide an approximate solution for clustering the distributional behavior via unsupervised learning techniques. Additionally, the method provided is robust and model-free and therefore has a wider scope of application beyond the specific dataset considered.

1.4 Thesis Organization

The remainder of this thesis proceeds as follows:

Chapter II: Posterior contraction of parameters and interpretability in Bayesian mixture modeling This chapter addresses several key issues concerning Bayesian nonparametric mixture models such as the inestimability of the number of components with Dirichlet Process priors, and develops an understanding of the behavior of the mixture models in the misspecified regime.

Chapter III: Bayesian contraction for Dirichlet process mixtures of smooth densities This chapter proposes a solution for the problem of misspecification of the underlying parameter space via the use of sieve estimates, and develops a deeper understanding of the behavior of mixtures of smooth kernels.

Chapter IV: Dirichlet Simplex Nest and Geometric Inference This chapter introduces a general modeling framework for inference via a generalization of the well-known Latent Dirichlet Allocation Model and provides a solution for computationally and statistically efficient inference.

Chapter V: Robust Representation Learning of Temporal Dynamic Interactions This chapter proposes a framework to analyse the behavior of unsupervised clustering approaches with application to traffic encounters, and also provides a novel clustering approach for the same.

Chapter VI: Conclusions and Future Work This chapter summarizes the novel contributions of this thesis and discusses idea for future research.

Each chapter is self-contained with all the necessary background materials and can be read independently of other chapters.

CHAPTER II

Posterior Contraction of Parameters and Interpretability in Bayesian Mixture Modeling

We study posterior contraction behaviors for parameters of interest in the context of Bayesian mixture modeling, where the number of mixing components is unknown while the model itself may or may not be correctly specified. Two representative types of prior specification will be considered: one requires explicitly a prior distribution on the number of mixture components, while the other places a nonparametric prior on the space of mixing distributions. The former is shown to yield an optimal rate of posterior contraction on the model parameters under minimal conditions, while the latter can be utilized to consistently recover the unknown number of mixture components, with the help of a fast probabilistic post-processing procedure. We then turn the study of these Bayesian procedures to the realistic settings of model misspecification. It will be shown that the modeling choice of kernel density functions plays perhaps the most impactful roles in determining the posterior contraction rates in the misspecified situations. Drawing on concrete posterior contraction rates established in this paper we wish to highlight some aspects about the interesting tradeoffs between model expressiveness and interpretability that a statistical modeler must negotiate in the rich world of mixture modeling. [1].

2.1 Introduction

Mixture models are one of the most useful tools in a statistician's toolbox for analyzing heterogeneous data populations. They can be a powerful black-box modeling device to approximate the most complex forms of density functions. Perhaps more importantly, they help the statistician express the data population's heterogeneous patterns and interpret them in a useful way (*McLachlan and Basford* (1988); *Lindsay* (1995); *Mengersen et al.* (2011)). The following are common, generic and meaningful questions a practitioner of mixture modeling may ask:

(I) how many mixture components are needed to express the underlying latent subpopulations.

(II) how efficiently can one estimate the parameters representing these components.

(III) what happens to a mixture model based statistical procedure when the model is actually misspecified?

How to determine the number of mixture components is a question that has long fascinated mixture modelers. Many proposed solutions approached this as a model selection problem. The number of model parameters, hence the number of mixture components, may be selected by optimizing with respect to some regularized loss function; see, e.g., *Lindsay* (1995); *Kass and Raftery* (1995); *Dacunha-Castelle and Gassiat* (1997) and the references therein. A Bayesian approach to regularization is to place explicitly a prior distribution on the number of mixture components, e.g., *Nobile*

[1] This work has been published in *Guha et al.* (2019)

(1994); *Richardson and Green* (1997); *Nobile and Fearnside* (2007); *Miller and Harrison* (2018). A convenient aspect of separating out the modeling and inference questions considered in (I) and (II) is that once the number of parameters is determined, the model parameters concerned by question (II) can be estimated and assessed via any standard parametric estimation methods.

In a number of modern applications of mixture modeling to heterogeneous data, such as in topic modeling, the number of mixture components (the topics) may be very large and not necessarily a meaningful quantity (*Blei et al.* (2003); *Tang et al.* (2014)). In such situations, it may be appealing for the modeler to consider a nonparametric approach, where both (I) and (II) are considered concurrently. The object of inference is now the mixing measure which encapsulates all unknowns about the mixture density function. There were numerous works exemplifying this approach, for eg., *Leroux* (1992); *Figueiredo and Jain* (1993); *Ishwaran et al.* (2001). In particular, the field of Bayesian nonparametrics (BNP) has offered a wealth of prior distributions on the mixing measure based on which one can arrive at the posterior distribution of any quantity of interest related to the mixing measure (*Hjort et al.* (2010)).

A common choice of such priors is the Dirichlet process (*Ferguson* (1973); *Blackwell and MacQueen* (1973); *Sethuraman* (1994)), resulting in the famous Dirichlet process mixture models (*Antoniak* (1974); *Lo* (1984); *Escobar and West* (1995)). Dirichlet process (DP) and its variants have also been adopted as a building block for more sophisticated hierarchical modeling, thanks to the ease with which computational procedures for posterior inference via Markov Chain Monte Carlo can be implemented (*Teh et al.* (2006); *Rodriguez et al.* (2008)). Moreover, there is a well-established asymptotic theory on how such Bayesian nonparametric mixture models result in asymptotically optimal estimation procedures for the population density. See, for instance, *Ghosal et al.* (1999);

Ghosal and van der Vaart (2007); *Shen et al.* (2013) for theoretical results specifically on DP mixtures, and *Ghosal et al.* (2000); *Shen and Wasserman* (2001); *Walker et al.* (2007) for general BNP models. The rich development in both algorithms and theory in the past decades has contributed to the widespread adoption of these models in a vast array of application domains.

For some time there was a misconception among quite a few practitioners in various application domains, a misconception that may have initially contributed to their enthusiasm for Bayesian nonparametric modeling, that the use of such nonparametric models eliminates altogether the need for determining the number of mixture components, because the learning of such a quantity is "automatic" from the posterior samples of the mixing measure. The implicit presumption here is that a consistent estimate of the mixing measure may be equated with a consistent estimate of the number of mixture components. This is not correct, as has been noted, for instance, by *Leroux* (1992) in the context of mixing measure estimation. More recently, *Miller and Harrison* (2014) explicitly demonstrated that the common practice of drawing inference about the number of mixture components via the DP mixture, specifically by reading off the number of support points in the Dirichlet's posterior sample, leads to an asymptotically inconsistent estimate.

Despite this inconsistency result, it will be shown in this chapter that it is still possible to obtain a consistent estimate of the number of mixture components using samples from a Dirichlet process mixture, or any Bayesian nonparametric mixture, by applying a simple and fast post-processing procedure on samples drawn from the DP mixture's posterior. On the other hand, the parametric approach of placing an explicit prior on the number of components yields both a consistent estimate of the number of mixture components, and more notably, an optimal posterior contraction rate for component parameters, under a minimal set of conditions. It is worth emphasizing that all these results are possible only under the assumption that the model is well-specified, i.e., the true but unknown population density lies in the support of the induced prior distribution on the mixture densities.

As George Box has said, "all models are wrong", but more relevant to us, all mixture models are misspecified in some way. The statistician has a number of modeling decisions to make when it comes to mixture models, including the selection of the class of kernel densities, and the support of the space of mixing measures. The significance of question (III) comes to the fore, because if the posterior contraction behavior of model parameters is very slow due to specific modeling choices, one has to be cautious about the interpretability of the parameters of interest. A very slow posterior contraction rate in theory implies that a given data set probably has relatively very slow influence on the movement of mass from the prior to the posterior distribution.

In this chapter we study Bayesian estimation of model parameters with both well-specified and misspecified mixture models. There are two sets of results. The first results resolve several outstanding gaps that remain in the existing theory and current practice of Bayesian parameter estimation, given that the mixture model is well-specified. The second set of results describes posterior contraction properties of such procedures when the mixture model is misspecified. We proceed to describe these results, related works and implications to the mixture modeling practice.

2.1.1 Well-specified regimes

Consider discrete mixing measures $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$. Here, $\boldsymbol{p} = (p_1, \ldots, p_k)$ is a vector of mixing weights, while atoms $\{\theta_i\}_{i=1}^{k}$ are elements in a given compact space $\Theta \in \mathbb{R}^d$. Mixing measure G is combined with a likelihood function $f(\cdot|\theta)$ with respect to Lebesgue measure μ to yield a mixture density: $p_G(\cdot) = \int f(\cdot|\theta)\mathrm{d}G(\theta) = \sum_{i=1}^{k} p_i f(\cdot|\theta_i)$. When $k < \infty$, we call this a *finite mixture model* with k components. We write $k = \infty$ to denote an *infinite mixture model.* The atoms θ_i's are representatives of the underlying subpopulations.

Assume that $X_1, \ldots, X_n$ are i.i.d. samples from a mixture density $p_{G_0}(x) = \int f(x|\theta)\mathrm{d}G_0(\theta)$, where G_0 is a discrete mixing measure with *unknown* number of support points $k_0 < \infty$ residing in Θ. In the overfitted setting, i.e., an upper bound $k_0 \leq \overline{k}$ is given so that one may work with an overfitted mixture with $\overline{k}$ mixture components, *Chen* (1995) showed that the mixing measure G_0 can be estimated at a rate $n^{-1/4}$ under the L_1 metric, provided that the kernel f satisfies a second-order identifiability condition – this is a linear independence property on the collection of kernel function f and its first and second order derivatives with respect to θ.

Asymptotic analysis of Bayesian estimation of the mixing measure that arises in both finite and infinite mixtures, where the convergence is assessed under Wasserstein distance metrics, was first investigated by *Nguyen* (2013). Convergence rates of the mixing measure under a Wasserstein distance can be directly translated to the convergence rates of the parameters in the mixture model. Under the same (second-order) identifiability condition, it can be shown that either maximum likelihood estimation method or a Bayesian method with a non-informative (e.g., uniform) prior yields a $(\log n/n)^{1/4}$ rate of convergence (*Ho and Nguyen* (2016); *Nguyen* (2013); *Ishwaran et al.* (2001)). Note, however, that $n^{-1/4}$ is not the optimal *pointwise* rate of convergence. *Heinrich and Kahn* (2018) showed that a distance based estimation method can achieve $n^{-1/2}$ rate of convergence under W_1 metric, even though their method may not be easy to implement in practice. *Ho et al.* (to appear) described a minimum Hellinger distance estimator that achieves the same optimal rate of parameter estimation.

An important question in Bayesian analysis is whether there exists a suitable prior specification for mixture models according to which the posterior distribution on the mixing measure can be shown to contract toward the true mixing measure at the same fast rate $n^{-1/2}$. *Rousseau and Mengersen* (2011) provided an interesting result in this regard, which states that for overfitted mixtures with a suitable Dirichlet prior on the mixing weights $\boldsymbol{p}$, assuming that an upper bound to the number of mixture component is given, in addition to a second-order type identifiability condition, then the posterior contraction to the true mixing measure can be established by the fact that the mixing weights associated with all redundant atoms of mixing measure G vanish at the rate close to the optimal $n^{-1/2}$.

In our first main result given in Theorem 2.3.1, we show that an alternative and relatively common choice of prior also yields optimal rates of convergence of the mixing measure (up to a logarithmic term), in addition to correctly recovering the number of mixture components, under considerably weaker conditions. In particular, we study the mixture of finite mixture (MFM) prior, which places an explicit prior distribution on the number of components k and a (conditional) Dirichlet prior on the weights $\boldsymbol{p}$, given each value of k. This prior has been investigated by *Miller and Harrison* (2018). Compared to the method of *Rousseau and Mengersen* (2011), no upper bound on the true number of mixture components is needed. In addition, only first-order identifiability condition is required for the kernel density f, allowing our results to apply to popular mixture models such as location-scale Gaussian mixtures. We also note that the MFM prior is one instance in a class of modeling proposals, e.g., *Nobile* (1994); *Richardson and Green* (1997); *Nobile and Fearnside* (2007) for which the established convergence behavior continues to hold. In other words, from an asymptotic standpoint, all is good on the parametric Bayesian front.

Our second main result, given in Theorem 2.3.2, is concerned with a Bayesian nonparametric modeling practice. A Bayesian nonparametric prior on mixing measures places zero mass on measures with finite support points, so the BNP model is misspecified with respect to the number of mixture components. Indeed, when G_0 has only finite support the true density p_{G_0} lies at the boundary of the support of the class of densities produced by the BNP prior. Despite the inconsistency results mentioned earlier on the number of mixture components produced by Dirichlet process mixtures, we will show that this situation can be easily corrected by applying a post-processing procedure to the samples generated from the posterior distribution arising from the DP mixtures, or any sufficiently well-behaved Bayesian nonparametric mixture models. By "well-behaved" we mean any BNP mixtures under which the posterior contraction rate on the mixing measure can be guaranteed by an upper bound using a Wasserstein metric.

Our post-processing procedure is simple, and motivated by the observation that a posterior sample of the mixing measure tends to produce a large number of atoms with very small and vanishing weights (*Green and Richardson*, 2001; *Miller and Harrison*, 2014). Such atoms can be ignored by a suitable truncation procedure. In addition, similar atoms in the metric space Θ can also be merged in a systematic and probabilistic way. Our procedure, named Merge-Truncate-Merge algorithm, is guaranteed to not only produce a consistent estimate of the number of mixture components but also retain the posterior contraction rates of the original posterior samples for the mixing measure. Theorem 2.3.2 provides a theoretical basis for the heuristics employed in practice in dealing with mixtures with unknown number of components (*Green and Richardson* (2001); *Nobile and Fearnside* (2007)).

2.1.2 Misspecified regimes

There are several ways a mixture model can be misspecified: either in the kernel density function f, or the mixing measure G, or both. Thus, in the misspecified setting, we assume that the data samples $X_1, \ldots, X_n$ are i.i.d. samples from a mixture density p_{G_0,f_0}, namely, $p_{G_0,f_0}(x) = \int f_0(x|\theta)G_0(\mathrm{d}\theta)$, where both G_0 and f_0 are unknown. The statistician draws inference from a mixture model $p_{G,f}$, still denoted by p_G for short, where G is a mixing measure with support on compact Θ, and f is a chosen kernel density function. In particular, a Bayesian procedure proceeds by placing a prior on the mixing measure G and obtaining the posterior distribution on G given the n-data sample. In general, the true data generating density p_{G_0} lies outside the support of the induced prior on p_G. We study the posterior behavior of G as the sample size n tends to infinity.

The behavior of Bayesian procedures under model misspecification has been investigated in the foundational work of *Kleijn and van der Vaart* (2006, 2012). These papers focus primarily on density estimation. In particular, assuming that the true data generating distribution's density lies outside the support of a Bayesian prior, then the posterior distribution on the model density can be shown to contract to an element of the prior's support, which is obtained by a Kullback-Leibler (KL) projection of the true density into the prior's support (*Kleijn and van der Vaart* (2006)).

It can be established that the posterior of p_G contracts to a density p_{G_*}, where G_* is a probability measure on Θ such that p_{G_*} is the (unique) minimizer of the Kullback-Leilber divergence $K(p_{G_0,f_0}, p_G)$ among all probability measure G on Θ. This mere fact is readily deduced from the theory of *Kleijn and van der Vaart* (2006), but the outstanding and relevant issue is whether the posterior contraction behavior carries over to that of G, and if so, at what rate. In general, G_* may not be unique, so posterior contraction of G cannot be established. Under identifiability, G_* is unique, but still $G_* \neq G_0$.

This leads to the question about interpretability when the model is misspecified. Specifically, when $f \neq f_0$, it may be unclear how one can interpret the parameters that represent mixing measure G, unless f can be assumed to be a reasonable approximation of f_0. Mixing measure G, too, may be misspecified, when the true support of G_0 may not lie entirely in Θ. In practice, it is a perennial challenge to explicate the relationship between G_* and the unknown G_0. In theory, it is mathematically an interesting question to characterize this relationship, if some assumption can be made on the true G_0 and f_0, but this is beyond the scope of this chapter. Regardless of the truth about this relationship, it is important for the statistician to know how impactful a particular modeling choice on f and G can affect the posterior contraction rates of the parameters of interest.

The main results that we shall present in Theorem 2.4.1 and Theorem 2.4.2 are on the posterior contraction rates of the mixing measure G toward the limit point G_*, under very mild conditions on the misspecification of f. In particular, we shall require that the tail behavior of function f is not much heavier than that of f_0 (cf. condition (P.5) or (P.5') in Section 2.4). Specific posterior contraction rates of contraction for G are derived when f is either Gaussian or Laplace density kernel, two representatives for supersmooth and ordinary smooth classes of kernel densities (*Fan* (1991)). A key step in our proofs lies in several inequalities which provide upper bound of Wasserstein distances on mixing measures in terms of weighted Hellinger distances, a quantity that plays a fundamental role in the asymptotic characterization of misspecified Bayesian models (*Kleijn and van der Vaart* (2006)).

It is interesting to highlight that the posterior contraction rate for the misspecified

Gaussian location mixture is the same as that of well-specified setting, which is nonetheless extremely slow, in the order of $(1/\log n)^{1/2}$. On the other hand, using a misspecified Laplace location mixture results in some loss in the exponent γ of the polynomial rate $n^{-\gamma}$. Although the contrast in contraction rates for the two families of kernels is quite similar to what is obtained for well-specified deconvolution problems for both frequentist methods (*Fan* (1991); *Zhang* (1990)) and Bayesian methods (*Nguyen* (2013); *Gao and van der Vaart* (2016)), our results are given for misspecified models, which can be seen in a new light: since the model is misspecified anyway, the statistician should be "free" to choose the kernel that can yield the most favorable posterior contraction for the parameters of his/ her model. In that regard, Laplace location mixtures may be preferred to Gaussian location mixtures, provided that the limit G_* is not too far from the true G_0. When this is not the case, i.e., when the bias of the misspecified model is too large due to the use of Laplace mixtures, it is more advisable to adopt Gaussian kernels instead, despite the latter's lagging posterior contraction behavior. Although it is quite clear that the ultimate model choice under misspecification will reside on resolving the tension between aforementioned bias and contracting variance, a satisfactory formulation and solution for such a model choice problem which accounts for parameter estimation and interpretability remains an interesting and important open question.

Additionally, we note that the relatively slow posterior contraction rate for G is due to the fact that the limiting measure G_* in general may have infinite support, regardless of whether the true G_0 has finite support or not. From a practical standpoint, it is difficult to interpret the estimate of G if G_* has infinite support. However, if G_* happens to have a finite number of support points, which is bounded by a known constant, say $\overline{k}$, then by placing a suitable prior on G to reflect this knowledge we show that the posterior of G contracts to G_* at a relatively fast rate $(\log n/n)^{1/4}$. This is the same rate obtained under the well-identified setting for overfitted mixtures.

2.1.3 Further remarks

The posterior contraction theorems in this chapter provide an opportunity to re-examine several aspects of the fascinating picture about the tension between a model's expressiveness and its interpretability. They remind us once again about the tradeoffs a modeler must negotiate for a given inferential goal and the information available at hand. We enumerate a few such insights:

(1) "One size does not fit all": Even though the family of mixture models as a whole can be excellent at inferring about population heterogeneity and at density estimation as a black-box device, a specific mixture model specification cannot do a good job at both. For instance, a Dirichlet process mixture of Gaussian kernels may yield an asymptotically optimal density estimation machine but it performs poorly when it comes to learning of parameters.

(2) "Finite versus infinite": If the number of mixture components is known to be small and an object of interest, then employing an explicit prior on this quantity results in the optimal posterior contraction rate for the model parameters and thus is a preferred method. When this quantity is known to be high or not a meaningful object of inference, Bayesian nonparametric mixtures provide a more attractive alternative as it can flexibly adapt to complex forms of densities. Regardless, one can still consistently recover the true number of mixture components using a nonparametric approach.

(3) "Some forms of misspecification are more useful than others". When the mixture

model is misspecified, careful design choices regarding the (mispecified) kernel density and the support of the mixing measure can significantly speed up the posterior contraction behavior of model parameters. For instance, a heavy-tailed and ordinary smooth kernel such as the Laplace, instead of the Gaussian kernel, is shown to be especially amenable to efficient parameter estimation.

The remainder of the chapter is organized as follows. Section 2.2 provides necessary backgrounds about mixture models, Wasserstein distances and several key notions of strong identifiability. Section 2.3 presents posterior contraction theorems for well-mispecified mixture models for both parametric and nonparametric Bayesian models. Section 2.4 presents posterior contraction theorems when the mixture model is misspecified. In Section 3.5, we provide illustrations of the Merge-Truncate-Merge algorithm via a simulation study. Proofs of technical results are provided in the supplementary material.

Notation Given two densities p, q (with respect to the Lebesgue measure μ), the total variation distance is given by $V(p, q) = (1/2) \int |p(x) - q(x)| \mathrm{d}\mu(x)$. Additionally, the squared Hellinger distance is given by $h^2(p, q) = (1/2) \int (\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}\mu(x)$. Furthermore, the Kullback-Leibler (KL) divergence is given by $K(p, q) = \int \log(p(x)/q(x)) p(x) \mathrm{d}\mu(x)$ and the squared KL divergence is given by $K_2(p, q) = \int \log(p(x)/q(x))^2 p(x) \mathrm{d}\mu(x)$. For a measurable function f, let Qf denote the integral $\int f dQ$. For any $\kappa = (\kappa_1, \ldots, \kappa_d) \in \mathbb{N}^d$, we denote $\dfrac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta) = \dfrac{\partial^{|\kappa|} f}{\partial \theta_1^{\kappa_1} \ldots \partial \theta_d^{\kappa_d}}(x|\theta)$ where $\theta = (\theta_1, \ldots, \theta_d)$. For any metric d on Θ, we define the open ball of d-radius ϵ around $\theta_0 \in \Theta$ as $B_d(\epsilon, \theta_0)$. We use $D(\epsilon, \Omega, \tilde{d})$ to denote the maximal ϵ-packing number for a general set Ω under a general metric $\tilde{d}$ on Ω. Additionally, the expression $a_n \gtrsim b_n$

will be used to denote the inequality up to a constant multiple where the value of the constant is independent of n. We also denote $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Furthermore, we denote A^c as the complement of set A for any set A while $B(x, r)$ denotes the ball, with respect to the l_2 norm, of radius $r > 0$ centered at $x \in \mathbb{R}^d$. Finally, we use $\text{Diam}(\Theta) = \sup\{\|\theta_1 - \theta_2\| : \theta_1, \theta_2 \in \Theta\}$ to denote the diameter of a given parameter space Θ relative to the l_2 norm, $\|\cdot\|$, for elements in $\mathbb{R}^d$.

2.2 Preliminaries

We recall the notion of Wasserstein distance for mixing measures, along with the notions of strong identifiability and uniform Lipschitz continuity conditions that prove useful in Section 2.3.

Mixture model Throughout the chapter, we assume that $X_1, \ldots, X_n$ are i.i.d. samples from a true but unknown distribution P_{G_0} with given density function

$$p_{G_0} := \int f(x|\theta) dG_0(\theta) = \sum_{i=1}^{k_0} p_i^0 f(x|\theta_i^0)$$

where $G_0 = \sum\limits_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ is a true but unknown mixing distribution with exactly k_0 number of support points, for some unknown k_0. Also, $\left\{f(x|\theta), \theta \in \Theta \subset \mathbb{R}^d\right\}$ is a given family of probability densities (or equivalently kernels) with respect to a sigma-finite measure μ on $\mathcal{X}$ where $d \geq 1$. Furthermore, Θ is a chosen parameter space, where we empirically believe that the true parameters belong to. In a well-specified setting, all support points of G_0 reside in Θ, but this may not be the case in a misspecified setting.

Regarding the space of mixing measures, let $\mathcal{E}_k := \mathcal{E}_k(\Theta)$ and $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ respectively denote the space of all mixing measures with exactly and at most k support points, all in Θ. Additionally, denote $\mathcal{G} := \mathcal{G}(\Theta) = \bigcup\limits_{k \in \mathbb{N}_+} \mathcal{E}_k$ the set of all discrete measures with finite supports on Θ. Moreover, $\overline{\mathcal{G}}(\Theta)$ denotes the space of all discrete measures (including those with countably infinite supports) on Θ. Finally, $\mathcal{P}(\Theta)$ stands for the space of all probability measures on Θ.

Wasserstein distance As in *Nguyen* (2013); *Ho and Nguyen* (2016) it is useful to analyze the identifiability and convergence of parameter estimation in mixture models using the notion of Wasserstein distance, which can be defined as the optimal cost of moving masses transforming one probability measure to another (*Villani* (2008)). Given two discrete measures $G = \sum\limits_{i=1}^{k} p_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^{k'} p'_i \delta_{\theta'_i}$, a coupling between $\boldsymbol{p}$ and $\boldsymbol{p'}$ is a joint distribution $\boldsymbol{q}$ on $[1\ldots, k] \times [1, \ldots, k']$, which is expressed as a matrix $\boldsymbol{q} = (q_{ij})_{1 \leq i \leq k, 1\ \leq j \leq k'} \in [0,1]^{k \times k'}$ with marginal probabilities $\sum\limits_{i=1}^{k} q_{ij} = p'_j$ and $\sum\limits_{j=1}^{k'} q_{ij} = p_i$ for any $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k'$. We use $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ to denote the space of all such couplings. For any $r \geq 1$, the r-th order Wasserstein distance between G and G' is given by

$$W_r(G, G') \quad = \quad \inf_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \left(\sum_{i,j} q_{ij} \|\theta_i - \theta'_j\|^r \right)^{1/r},$$

where $\| \cdot \|$ denotes the l_2 norm for elements in $\mathbb{R}^d$. It is simple to see that if a sequence of probability measures $G_n \in \mathcal{O}_{k_0}$ converges to $G_0 \in \mathcal{E}_{k_0}$ under the W_r metric at a rate $\omega_n = o(1)$ for some $r \geq 1$ then there exists a subsequence of G_n such that the set of atoms of G_n converges to the k_0 atoms of G_0, up to a permutation of the atoms, at the same rate ω_n.

Strong identifiability and uniform Lipschitz continuity The key assumptions that will be used to analyze the posterior contraction of mixing measures include uniform Lipschitz condition and strong identifiability condition. The uniform Lipschitz condition can be formulated as follows (*Ho and Nguyen*, 2016).

Definition 2.2.1. We say the family of densities $\{f(x|\theta), \theta \in \Theta\}$ is uniformly Lipschitz up to the order r, for some $r \geq 1$, if f as a function of θ is differentiable up to the order r and its partial derivatives with respect to θ satisfy the following inequality

$$\sum_{|\kappa|=r} \left| \left(\frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta_1) - \frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta_2) \right) \gamma^{\kappa} \right| \leq C \|\theta_1 - \theta_2\|^{\delta} \|\gamma\|$$

for any $\gamma \in \mathbb{R}^d$ and for some positive constants δ and C independent of x and $\theta_1, \theta_2 \in \Theta$. Here, $\gamma^{\kappa} = \prod_{i=1}^{d} \gamma_i^{\kappa_i}$ where $\kappa = (\kappa_1, \ldots, \kappa_d)$.

The first order uniform Lipschitz condition is satisfied by many popular classes of density functions, including Gaussian, Student's t, and skew-normal family. Now, strong identifiability condition of the r^{th} order is formulated as follows,

Definition 2.2.2. For any $r \geq 1$, we say that the family $\{f(x|\theta), \theta \in \Theta\}$ (or in short, f) is *identifiable in the order* r, for some $r \geq 1$, if $f(x|\theta)$ is differentiable up to the order r in θ and the following holds

A1. For any $k \geq 1$, given k different elements $\theta_1, \ldots, \theta_k \in \Theta$. If we have $\alpha_{\eta}^{(i)}$ such that for almost all x

$$\sum_{l=0}^{r} \sum_{|\eta|=l} \sum_{i=1}^{k} \alpha_{\eta}^{(i)} \frac{\partial^{|\eta|} f}{\partial \theta^{\eta}}(x|\theta_i) = 0$$

then $\alpha_{\eta}^{(i)} = 0$ for all $1 \leq i \leq k$ and $|\eta| \leq r$.

Many commonly used families of density functions satisfy the first order identifiability condition, including location-scale Gaussian distributions and location-scale Student's t-distributions. Technically speaking, strong identifiability conditions are useful in providing the guarantee that we have some sort of lower bounds of Hellinger distance between mixing densities in terms of Wasserstein metric between mixing measures. For example, if f is identifiable in the first order, we have the following inequality (*Ho and Nguyen*, 2016)

$$h(p_G, p_{G_0}) \gtrsim W_1(G, G_0) \tag{2.1}$$

for any $G \in \mathcal{E}_{k_0}$. It implies that for any estimation method that yields the convergence rate $n^{-1/2}$ for density p_{G_0} under the Hellinger distance, the induced rate of convergence for the mixing measure G_0 is $n^{-1/2}$ under W_1 distance.

2.3 Posterior contraction under well-specified regimes

In this section, we assume that the mixture model is well-specified, i.e., the data are i.i.d. samples from the mixture density p_{G_0}, where mixing measure G_0 has k_0 support points in compact parameter space $\Theta \subset \mathbb{R}^d$. Within this section, we assume further that the true but unknown number of components k_0 is finite. A Bayesian modeler places a prior distribution Π on a suitable subspace of $\overline{\mathcal{G}}(\Theta)$. Then, the posterior distribution over G is given by:

$$\Pi(G \in B \big| X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p_G(X_i) d\Pi(G)}{\int_{\overline{\mathcal{G}}(\Theta)} \prod_{i=1}^n p_G(X_i) d\Pi(G)} \tag{2.2}$$

We are interested in the posterior contraction behavior of G toward G_0, in addition to recovering the true number of mixture components k_0.

2.3.1 Prior results

The customary prior specification for a finite mixture is to use a Dirichlet distribution on the mixing weights and another standard prior distribution on the atoms of the mixing measure. Let H be a distribution with full support on Θ. Thus, for a mixture of k components, the full Bayesian mixture model specification takes the form:

$$\begin{aligned} \boldsymbol{p} = (p_1, \ldots, p_k) &\sim \text{Dirichlet}_k(\gamma/k, \ldots, \gamma/k), \\ \theta_1, \ldots, \theta_k &\overset{iid}{\sim} H, \\ X_1, \ldots, X_n \mid G = \sum_{i=1}^{k} p_i \delta_{\theta_i} &\overset{iid}{\sim} p_G. \end{aligned} \tag{2.3}$$

Suppose for a moment that k_0 is known, we can set $k = k_0$ in the above model specification. Thus we would be in an *exact-fitted* setting. Provided that f satisfies both first-order identifiability condition and the uniform Lipschitz continuity condition, H is approximately uniform on Θ, then according to *Ho and Nguyen* (2016) it can be established that as n tends to infinity,

$$\Pi\left(G \in \mathcal{E}_{k_0}(\Theta) : W_1(G, G_0) \gtrsim (\log n/n)^{1/2} \middle| X_1, \ldots, X_n\right) \xrightarrow{p_{G_0}} 0. \tag{2.4}$$

The $(\log n/n)^{1/2}$ rate of posterior contraction is optimal up to a logarithmic term.

When k_0 is unknown, there may be a number of ways for the modeler to proceed. Suppose that an upper bound of k_0 is given, say $k_0 < \overline{k}$. Then by setting $k = \overline{k}$ in the above model specification, we have a Bayesian *overfitted* mixture model. Provided that

f satisfies the second-order identifability condition and the uniform Lipschitz continuity condition, H is again approximately uniform distribution on Θ, then it can be established that (*Ho and Nguyen*, 2016):

$$\Pi\left(G \in \mathcal{O}_{\overline{k}}(\Theta) : W_2(G, G_0) \gtrsim (\log n/n)^{1/4} \middle| X_1, \ldots, X_n\right) \xrightarrow{p_{G_0}} 0. \tag{2.5}$$

This result does not provide any guarantee about whether the true number of mixture components k_0 can be recovered. The rate (upper bound) $(\log n/n)^{1/4}$ under W_2 metric implies that under the posterior distribution the redundant mixing weights of G contracts toward zero at the rate $(\log n/n)^{1/2}$, but the posterior contraction to each of the k_0 atoms of G_0 occurs at the rate $(\log n/n)^{1/4}$ only.

Interestingly, it can be shown by *Rousseau and Mengersen* (2011) that with a more judicious choice of prior distribution on the mixing weights, one can achieve a near-optimal posterior contraction behavior. Specifically, they continued to employ the Dirichlet prior, but they required the Dirichlet's hyperparameters set to be sufficiently small: $\gamma/k \leq d/2$ in (2.3) where $k = \overline{k}$, d is the dimension of the parameter space Θ. Then, under some conditions on kernel f approximately comparable to the second-order identifiability and the uniform Lipschitz continuity condition defined in the previous section, they showed that for any $\epsilon > 0$, as n tends to infinity

$$\Pi\left(\exists I \subset \{1, \ldots, k\}, |I| = k - k_0 \text{ s.t. } \sum_{i \in I} p_i < n^{-1/2+\epsilon} \middle| X_1, \ldots, X_n\right) \xrightarrow{p_{G_0}} 1. \tag{2.6}$$

For a more precise statement along with the complete list of sufficient conditions leading to claim (2.6), we refer the reader to the original theorem of *Rousseau and Mengersen* (2011). Although their theorem is concerned with only the behavior of the redundant mixing weights p_i, where $i \in I$, which vanish at a near-optimal rate $n^{-1/2+\epsilon}$, it can be deduced from their proof that the posterior contraction for the true atoms of G_0 occurs at this near-optimal rate as well. *Rousseau and Mengersen* (2011) also showed that this performance may not hold if the Dirichlet's hyperparameters are set to be sufficiently large. Along this line, concerning the recovery of the number of mixture components k_0, *Chambaz and Rousseau* (2008) demonstrated the convergence of the posterior mode of the number of components to the true number of components k_0 at a rate $n^{-\rho}$, where ρ depends on $\overline{k} - k_0$, the number of redundant components forced upon by our model specification. Finally, we note that in addition to Dirichlet-type prior specifications, other types of prior specifications have also been taken up by other researchers (*Xie and Xu* (2017); *Fúquene et al.* (2019)).

2.3.2 Optimal posterior contraction via a parametric Bayesian mixture

We will show that optimal posterior contraction rates for mixture model parameters can be achieved by a natural Bayesian extension on the prior specification, even when the upper bound on the number of mixture component k is unknown. The modeling idea is simple and truly Bayesian in spirit: since k_0 is unknown, let K be a natural-valued random variable representing the number of mixture components. We endow K with a suitable prior distribution q_K on the positive integers. Conditioning on $K = k$, for each

k, the model is specified as before:

$$\begin{aligned}
K &\sim q_K, \\
\boldsymbol{p} = (p_1, \ldots, p_k) | K = k &\sim \text{Dirichlet}_k(\gamma/k, \ldots, \gamma/k), \\
\theta_1, \ldots, \theta_k \mid K = k &\overset{iid}{\sim} H, \\
X_1, \ldots, X_n \mid G = \sum_{i=1}^{k} p_i \delta_{\theta_i} &\overset{iid}{\sim} p_G . \qquad (2.7)
\end{aligned}$$

This prior specification is called *mixture of finite mixtures* (MFM) model (*Richardson and Green* (1997); *Stephens* (2000); *Miller and Harrison* (2018)). In the sequel we show that the application of the MFM prior leads to the optimal posterior contraction rates for the model parameters. Interestingly, such guarantees can be established under very mild conditions on the kernel density f: only the uniform Lipschitz continuity and the first-order identifiability conditions will be required. The first-order identifiability condition is the minimal condition for which the optimal posterior contraction rate can be established by the proof technique employed, since this condition is also necessary for exact-fitted mixture models to receive the $n^{-1/2}$ posterior contraction rate. We proceed to state such conditions.

(P.1) The parameter space Θ is compact, while kernel density f is first-order identifiable and admits the uniform Lipschitz property up to the first order.

(P.2) The base distribution H is absolutely continuous with respect to the Lebesgue measure μ on $\mathbb{R}^d$ and admits a density function $g(\cdot)$. Additionally, H is approximately uniform, i.e., $\min_{\theta \in \Theta} g(\theta) > c_0 > 0$.

(P.3) There exists $\epsilon_0 > 0$ such that $\displaystyle\int (p_{G_0}(x))^2 / p_G(x) d\mu(x) \leq M(\epsilon_0)$ as long as $W_1(G, G_0) \leq \epsilon_0$ for any $G \in \mathcal{O}_{k_0}$ where $M(\epsilon_0)$ depends only on ϵ_0, G_0, and Θ.

(P.4) The prior q_K places positive mass on the set of natural numbers, i.e., $q_K(k) > 0$ for all $k \in \mathbb{N}$.

Theorem 2.3.1. Under assumptions (P.1), (P.2), (P.3), and (P.4) on MFM, we have that

(a) $\Pi(K = k_0 | X_1, \ldots, X_n) \to 1$ a.s. under P_{G_0}.

(b) Moreover,

$$\Pi\bigg(G \in \overline{\mathcal{G}}(\Theta) : W_1(G, G_0) \lesssim (\log n/n)^{1/2} \bigg| X_1, \ldots, X_n \bigg) \to 1$$

in P_{G_0}-probability.

The proof of Theorem 2.3.1 is deferred to Appendix 2.6.1. We now make several remarks regarding the conditions required in the theorem.

(i) It is worth stating up front that these conditions are almost minimal in order for the optimal posterior contraction to be guaranteed, and are substantially weaker than previous works (as discussed above). In particular, assumption (P.1) is crucial in establishing that the Hellinger distance $h(p_G, p_{G_0}) \geq C_0 W_1(G, G_0)$ where C_0 is some positive constant depending only on G_0 and Θ. Assumption (P.2) and (P.4) are standard conditions on the support of the prior so that posterior consistency can be guaranteed for any unknown G_0 with unknown number of support atoms residing on Θ. The role of (P.3) is to help control the growing rate of KL neighborhood, which is central in the analysis of posterior convergence rate of mixing measures. This assumption is held for various choices of kernel f, including location families and location-scale families. Therefore, the assumptions

(P.1), (P.2),(P.3) and (P.4) are fairly general and satisfied by most common choice of kernel densities.

(ii) Condition (P.2) may be replaced by the following weaker condition:

(P.2') The base distribution H is absolutely continuous with respect to the Lebesgue measure μ on $\mathbb{R}^d$ and admits a density function $g(\cdot)$. Additionally, H must contain sufficient mass near the atoms of G_0, i.e., $\min_{\theta:\|\theta-\theta_i^0\|\leq\epsilon} g(\theta) \geq c_0 > 0$ for some $\epsilon > 0$.

We prefer (P.2) which is required for unknown G_0 and is a reasonable assumption in practice.

(iii) The contraction rate with respect to the W_1 norm for strongly identifiable family of densities is $O_P((\log(n)/n)^{1/2})$. The contraction rates relative to the L_q norms for $q \geq 1$ can be obtained by Lemma 2.9.2 and it is easy to show that the corresponding contraction rates are $O_P((\log(n)/n)^{1/2q})$ for $1 \leq q \leq 2$ and $O_P((\log(n)/n)^{1/q})$ for $q \geq 2$.

Theorem 2.3.1 provides a positive endorsement for employing the MFM prior when the number of mixture components is unknown, but is otherwise believed to be finite and an important quantity of inferential interest. The papers of *Richardson and Green* (1997); *Miller and Harrison* (2018) discuss additional favorable properties of this class of models. However, when the true number of mixture components is large, posterior inference with the MFM may still be inefficient in practice. This is because much of the computational effort needs to be expended for the model selection phase, so that the number of mixture components can be reliably ascertained. Only then does the fast asymptotic rate of parameter estimation come meaningfully into effect.

2.3.3 A posteriori processing for BNP mixtures

Instead of placing a prior distribution explicitly on the number of mixture components when this quantity is unknown, another predominant approach is to place a Bayesian nonparametric prior on the mixing measure G, resulting in infinite mixture models. Bayesian nonparametric models such as Dirichlet process mixtures and the variants have remarkably extended the reach of mixture modeling into a vast array of applications, especially those areas where the number of mixture components in the modeling is very large and difficult to fathom, or when it is a quantity of only tangential interest. For instance, in topic modeling applications of web-based text corpora, one may be interested in the most "popular" topics, the number of topics is less meaningful (*Blei et al.* (2003); *Teh et al.* (2006); *Nguyen* (2015); *Yurochkin et al.* (2017)). DP mixtures and variants can also serve as an asymptotically optimal device for estimating the population density, under standard conditions on the true density's smoothness, see, e.g., *Ghosal and van der Vaart* (2001, 2007); *Shen et al.* (2013); *Scricciolo* (2014).

Since a nonparametric Bayesian prior such as the Dirichlet process places zero probability on mixing measures with finite number of supporting atoms, the Dirichlet process mixture's posterior is inconsistent on the number of mixture components, provided the true number of mixture components is finite *Miller and Harrison* (2014). It is well known in practice that Dirichlet process mixtures tend to produce many small extraneous components around the "true" clusters, making them challenging to use to draw conclusion about the true number of mixture components when this becomes a quantity of interest (*MacEachern and Mueller* (1998); *Green and Richardson* (2001)). In this section we describe a simple posteriori processing algorithm that consistently estimates the number of components for any general Bayesian prior, even without the exact knowledge of its structure as long as the posterior for that prior contracts at some known rate to the true G_0.

Our starting point is the availability of a mixing measure sample G that is drawn from the posterior distribution $\Pi(G|X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are i.i.d. samples of the mixing density p_{G_0}. Under certain conditions on the kernel density f, it can be established that for some Wasserstein metric W_r, as $n \to \infty$

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : W_r(G, G_0) \leq \delta\omega_n \middle| X_1, \ldots, X_n\right) \xrightarrow{p_{G_0}} 1 \tag{2.8}$$

for *all* constant $\delta > 0$, while $\omega_n = o(1)$ is a vanishing rate. Thus, ω_n can be taken to be (slightly) slower than actual rate of posterior contraction of the mixing measure. Concrete examples of the posterior contraction rates in infinite and (overfitted) finite mixtures are given in *Nguyen* (2013); *Gao and van der Vaart* (2016); *Ho and Nguyen* (2016).

The posterior processing algorithm operates on an instance of mixing measure G, by suitably merging and truncating atoms that provide the support for G. The only inputs to the algorithm, which we call *Merge-Truncate-Merge* (MTM) algorithm is G, in addition to the upper bound of posterior contraction rate ω_n, and a tuning parameter $c > 0$. The tuning parameter c is useful in practice, as we shall explain, but in theory the algorithm "works" for any constant $c > 0$. Thus, the method is almost "automatic" as it does not require any additional knowledge about the kernel density f or the space of support Θ for the atoms. It is also simple and fast. We shall show that the outcome of the algorithm is a consistent estimate of both the number of mixing components and the mixing measure. The latter admits a posterior contraction rate's upper bound ω_n as well.

The detailed pseudocode of MTM algorithm is summarized in Algorithm 2.1. At a high level, it consists of two main stages. The first stage involves a probabilistic procedure for merging atoms that may be clustered near one another. The second stage involves a deterministic procedure for truncating extraneous atoms and merging them suitably with the remaining ones in a systematic way. The driving force of the algorithm lies in the asymptotic bound on the Wasserstein distance, i.e., $W_r(G, G_0) \leq c\omega_n$ with high probability. When $c\omega_n$ is sufficiently small, there may be many atoms that concentrate around each of the supporting atoms of G_0. Although G_0 is not known, such clustering atoms may be merged into one, by our first stage of probabilistic merging scheme. The second stage (truncate-merge) is also necessary in order to obtain a consistent estimate of k_0, because there remain distant atoms which carry a relatively small amount of mass. They will need to be suitably truncated and merged with the other more heavily supported atoms. In other words, our method can be viewed as a formal procedure of the common practices employed by numerous practitioners.

We proceed to present the theoretical guarantee for the outcome of Algorithm 2.1.

Theorem 2.3.2. Let G be a posterior sample from posterior distribution of any Bayesian procedure, namely, $\Pi(\cdot|X_1, \ldots, X_n)$ according to which the upper bound (2.8) holds for all $\delta > 0$. Let $\widetilde{G}$ and $\tilde{k}$ be the outcome of Algorithm 2.1 applied to G, for an arbitrary constant $c > 0$. Then the following hold as $n \to \infty$.

(a) $\Pi(\tilde{k} = k_0 | X_1, \ldots, X_n) \to 1$ in P_{G_0}-probability.

(b) For all $\delta > 0$, $\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : W_r(\tilde{G}, G_0) \leq \delta\omega_n \middle| X_1, \ldots, X_n\right) \longrightarrow 1$ in P_{G_0}-probability.

We add several comments concerning this theorem.

Algorithm 2.1 Merge-Truncate-Merge Algorithm

```
Input: Posterior sample $G = \sum_i p_i \delta_{\theta_i}$ from (2.8), rate $\omega_n$, constant $c$.
Output: Discrete measure $\widetilde{G}$ and its number of supporting atoms $\tilde{k}$.
  {Stage 1: Merge procedure:}
  Reorder atoms $\{\theta_1, \theta_2, \dots\}$ by simple random sampling without replacement with corresponding weights $\{p_1, p_2, \dots\}$.
      let $\tau_1, \tau_2, \dots$ denote the new indices, and set $\mathcal{E} = \{\tau_j\}_j$ as the existing set of atoms.
  Sequentially for each index $\tau_j \in \mathcal{E}$, if there exists an index $\tau_i < \tau_j$ such that $\|\theta_{\tau_i} - \theta_{\tau_j}\| \leq \omega_n$, then:
      update $p_{\tau_i} = p_{\tau_i} + p_{\tau_j}$, and remove $\tau_j$ from $\mathcal{E}$.
  Collect $G' = \sum_{j:\ \tau_j \in \mathcal{E}} p_{\tau_j} \delta_{\theta_{\tau_j}}$.
      write $G'$ as $\sum_{i=1}^{k} q_i \delta_{\phi_i}$ so that $q_1 \geq q_2 \geq \dots$.
  {Stage 2: Truncate-Merge procedure:}
  Set $\mathcal{A} = \{i : q_i > (c\omega_n)^r\}$, $\mathcal{N} = \{i : q_i \leq (c\omega_n)^r\}$.
  For each index $i \in \mathcal{A}$, if there is $j \in \mathcal{A}$ such that $j < i$ and $q_i \|\phi_i - \phi_j\|^r \leq (c\omega_n)^r$, then
      remove $i$ from $\mathcal{A}$ and add it to $\mathcal{N}$.
  For each $i \in \mathcal{N}$, find atom $\phi_j$ among $j \in \mathcal{A}$ that is nearest to $\phi_i$
      update $q_j = q_j + q_i$.
  Return $\widetilde{G} = \sum_{j \in \mathcal{A}} q_j \delta_{\phi_j}$ and $\tilde{k} = |\mathcal{A}|$.
```

(i) The proof of this theorem is deferred to Appendix 2.6.2, where we clarify carefully the roles played by each step of the MTM algorithm.

(ii) Although it is beyond the scope of this chapter to study the practical viability of the MTM algorithm, for interested readers we present a brief illustration of the algorithm via simulations in Section 3.5.

(iii) In practice, one may not have a mixing measure G sampled from the posterior $\Pi(\cdot|X_1, \ldots, X_n)$ but a sample from G itself, say F_n, the empirical distribution function. Then one can apply the MTM algorithm to F_n instead. Assume that F_n is sufficiently close to G, in the sense that $W_r(F_n, G) \lesssim W_r(G, G_0)$, it is straightforward to extend the above theorem to cover this scenario.

Practical implications At this point, one may look forward to some guidance regarding the modeling choices of parametrics versus nonparametrics. Even in the tight arena of Bayesian mixture modeling, the jury may still be out. The results in this section seems to provide a stronger theoretical support for the former, when it comes to the efficiency of parameter estimation and the corresponding model interpretation. However, as we will see in the next section, when the mixture model is misspecified, the fast posterior contraction rate offered by the use of the MFM prior is no longer valid. On the other hand, Bayesian nonparametric models are more versatile in adapting to complex forms of population densities. In many modern applications it is not meaningful to estimate the number of mixing components, only the most "significant" ones in a sense suitably defined. Perhaps a more meaningful question concerning a Bayesian nonparametric mixture model is whether it is capable of learning selected mixture components in an efficient way.

2.4 Posterior contraction under model misspecification

In this section, we study the posterior contraction behavior of the mixing measure under the realistic scenarios of model misspecification. There are several ways a mixture model can be misspecified, due to the misspecification of the kernel density function f, or the support of the mixing measure G, or both. From here on, we shall assume that the data population follows a mixture distribution composed of unknown kernel density f_0 and unknown mixing measure G_0 — thus, in this section the true density shall be denoted by p_{G_0,f_0} to highlight the possibility of misspecification.

To avoid heavy subscripting, we continue to use p_G instead of $p_{G,f}$ to represent the density function of the mixture model that we operate on. The kernel density f is selected by the modeler. Additionally, G is endowed with a suitable prior Π on the space of mixing measures with support belonging to compact parameter space Θ. By Bayes rule (Eq. (3.1)) one obtains the posterior distribution $\Pi(G|X_1, \ldots, X_n)$, where the n-i.i.d. sample $X_1, \ldots, X_n$ are generated by p_{G_0,f_0}. It is possible that $f \neq f_0$. It is also possible that the support of G_0 does not reside within Θ. In practice, the statistical modeler would hope that the kernel choice of f is not too different from the true but unknown f_0. Otherwise, it would be unclear how one can interpret the parameters that represent the mixing measure G. Our goal is to investigate the posterior contraction of $\Pi(G|X_1, \ldots, X_n)$ in such situations, as sample size n tends to infinity. The theory is applicable for a broad class of prior specification on the mixing measures on Θ, including the MFM prior and a nonparametric Bayesian prior such as the Dirichlet process.

A fundamental quantity that arises in the theory of Bayesian misspecification for density estimation is the minimizer of the Kullback-Leibler (KL) divergence from the true population density to a density function residing in the support of the induced prior on the space of densities p_G, which we shall assume to exist (cf. *Kleijn and van der Vaart* (2006)). Moreover, assume that the KL minimizer can be expressed as a mixture density p_{G_*}, where G_* is a probability measure on Θ. We may write

$$G_* \in \operatorname*{arg\,min}_{G \in \mathcal{P}(\Theta)} K(p_{G_0,f_0}, p_G). \tag{2.9}$$

We will see in the sequel that the existence of the KL minimizer p_{G^*} entails its uniqueness. In general, however, G_* may be non-unique. Thus, define

$$\mathcal{M}^* := \left\{ G_* \in \mathcal{P}(\Theta) : \; G_* \in \operatorname*{arg\,min}_{G \in \mathcal{P}(\Theta)} K(p_{G_0,f_0}, p_G) \right\}.$$

It is challenging to characterize the set $\mathcal{M}^*$ in general. However, a very useful technical property can be shown as follows:

Lemma 2.4.1. For any $G \in \mathcal{P}(\Theta)$ and $G_* \in \mathcal{M}^*$, it holds that $\displaystyle\int \frac{p_G(x)}{p_{G_*}(x)} p_{G_0,f_0}(x) \mathrm{d}x \leq 1$.

By exploiting the fact that the class of mixture densities is a convex set, the proof of this lemma is similar to that of Lemma 2.3 of *Kleijn and van der Vaart* (2006), so it is omitted. This leads quickly to the following fact.

Lemma 2.4.2. For any two elements $G_{1,*}, G_{2,*} \in \mathcal{M}^*$, $p_{G_{1,*}}(x) = p_{G_{2,*}}(x)$ for almost all $x \in \mathcal{X}$.

In other words, the mixture density p_{G_*} is uniquely identifiable. Under a standard identifiability condition of the kernel f, which is satisfied by the examples considered in this section, it follows that G_* is unique. Due to the model misspecification, in general $G_* \neq G_0$. The best we can hope for is that the posterior distribution of the mixing measure G contracts toward G_* as n tends to infinity. The goal of the remaining of this section is to study the posterior contraction behavior of the (misspecified) mixing measure G towards the unique G_*.

Following the theoretical framework of *Kleijn and van der Vaart* (2006), the posterior contraction behavior of the mixing measure G can be obtained by studying the relationship of a weighted version of Hellinger distance and corresponding Wasserstein distances between G and the limiting point G_*. In particular, for a fixed pair of mixture densities p_{G_0,f_0} and p_{G_*}, the weighted Hellinger $\overline{h}$ between two mixture densities is defined as follows (*Kleijn and van der Vaart*, 2006).

Definition 2.4.1. For $G_1, G_2 \in \mathcal{P}(\Theta)$,

$$\overline{h}^2(p_{G_1}, p_{G_2}) := \frac{1}{2} \int \left(\sqrt{p_{G_1}(x)} - \sqrt{p_{G_2}(x)} \right)^2 \frac{p_{G_0,f_0}(x)}{p_{G_*}(x)} dx.$$

It is clear that when $G_* = G_0$ and $f = f_0$, the weighted Hellinger distance reduces to the standard Hellinger distance. In general they are different due to misspecification. According to Lemma 2.4.1, we have $\overline{h}(p_{G_1}, p_{G_2}) \leq 1$ for all $G_1, G_2 \in \mathcal{P}(\Theta)$.

Choices of prior on mixing measures As in the previous section, we work with two representative priors on the mixing measure: the MFM prior and the Dirichlet process prior. Both prior choices may contribute to the model misspecification, if the true mixing measure G_0 lies outside of the support of the prior distribution.

Recall the MFM prior specification given in Eq. (2.7). We also need a stronger condition on q_K:

(P.4') The prior distribution q_K on the number of components satisfies $q_k \gtrsim k^{-\alpha_0}$ for some $\alpha_0 > 1$.

The $\alpha_0 > 1$ condition is placed in order to ensure that q_K is a proper distribution on natural numbers. Note that the assumption with prior on the number of components q_K is mild and satisfied by many distributions, such as Poisson distribution. In order to obtain posterior contraction rates, one needs to make sure the prior places sufficient mass on the (unknown) limiting point of interest. For the MFM prior, such a condition is guaranteed by the following lemma.

Lemma 2.4.3. Let Π denote the prior for generating G based on MFM (2.7), where H admits condition (P.2) and q_K admits (P.4'). Fix $r \geq 1$. Then the following holds, for any $G_* \in \mathcal{P}(\Theta)$

$$\begin{aligned} &\Pi\left(G : W_r^r(G, G_*) \leq (2^r + 1)\epsilon^r\right) \\ &\quad \gtrsim \frac{\gamma\Gamma(\gamma) D! q_D}{D} \left(c_0 \left(\frac{\epsilon}{\text{Diam}(\Theta)}\right)^d\right)^D \left(\frac{1}{D}\left(\frac{\epsilon}{\text{Diam}(\Theta)}\right)^r\right)^{\gamma(D-1)/D} \end{aligned} \tag{2.10}$$

for all ϵ sufficiently small so that $D(\epsilon, \Theta, \|.\|) > \gamma$. Here, $D = D(\epsilon, \Theta, \|.\|)$ and q_D stand for the maximal ϵ-packing number for Θ under $\|.\|$ norm and the prior weight $\Pi(K = D)$, respectively.

The proof of Lemma 2.4.3 is provided in Appendix 3.7.3. Alternatively, for a Dirichlet process prior, G is distributed a priori according to a Dirichlet measure with concentration parameter $\gamma > 0$ and base measure H satisfying condition (P.2). An analogous concentration bound for such a prior is given in Lemma 5 of *Nguyen* (2013).

It is somewhat interesting to note that the difference in the choices of prior under misspecification does not affect the posterior contraction bounds that we can establish. In particular, as we have seen for the definition, G_* does not depend on a specific choice of prior distribution (only its support). Due to misspecification, G_* may have infinite support, even if the true G_0 has a finite number of support points. When G_* has infinite support, the posterior contraction toward G_* becomes considerably slower compared to the well-specified setting. In addition to the structure of G_*, we will see in the sequel that the modeler's specific choice of kernel density f proves to be especially impactful on the rate of posterior contraction.

2.4.1 Gaussian location mixtures

Consider a class of kernel densities that belong to the supersmooth location family of density functions. A particular example that we focus on in this section is a class of Gaussian distributions with some fixed covariance matrix Σ. More precisely, f has the following form:

$$\left\{ f(\cdot|\theta), \theta \in \Theta \subset \mathbb{R}^d : f(x|\theta) := \frac{\exp(-(x-\theta)^\top \Sigma^{-1}(x-\theta)/2)}{|2\pi\Sigma|^{-1/2}} \right\}, \tag{2.11}$$

where $|\cdot|$ stands for matrix determinant. Note that, Gaussian kernel is perhaps the most popular choice in mixture modeling.

With the Gaussian location kernel, it is possible to obtain a lower bound on the Hellinger distance between the mixture densities in terms of the Wasserstein distance between corresponding mixing measures (*Nguyen* (2013)). More useful in the misspecified setting is a key lower bound for the weighted Hellinger distance in terms of the Wasserstein metric, which is given below in Prop. 2.4.1. In order to establish this bound we shall require a technical condition relating f to the true f_0 and G_0. This condition is stated by assumption (P.5) or a weaker version (P.5').

(P.5) The support of G_0, namely, $\text{supp}(G_0)$ is a bounded subset of $\mathbb{R}^d$. Moreover, there are some constants $C_0, C_1, \alpha > 0$ such that for any $R > 0$,

$$\sup_{x \in \mathbb{R}^d, \theta \in \Theta, \theta_0 \in \text{supp}(G_0)} \frac{f(x|\theta)}{f_0(x|\theta_0)} \mathbb{1}_{\|x\|_2 \leq R} \leq C_1 \exp(C_0 R^\alpha).$$

The condition in (P.5) that the support of G_0 has the same dimension d is purely for the sake of interpretability, if the quantity of inferential interest is the mixing measure G. This is also related to the condition in (P.5) on the density ratio $f(x/\theta)/f_0(x|\theta_0)$. In fact, both conditions on the support of G_0 and on f_0 are not strictly necessary from a technical standpoint; only a "black-box" condition directly placed on the true density p_{G_0,f_0} will be sufficient. Accordingly, (P.5) may be replaced by the following weaker condition.

(P.5') Assume that there are some constants $C_0, C_1, \alpha > 0$ such that for any $R > 0$,

$$\sup_{x \in \mathbb{R}^d, \theta \in \Theta} \frac{f(x|\theta)}{p_{G_0,f_0}(x)} \mathbb{1}_{\|x\|_2 \leq R} \leq C_1 \exp(C_0 R^\alpha).$$

It is simple to verify that (P.5) implies (P.5').

Examples In the following examples, the statistician decides to fit the data with a Gaussian location mixture model $p_{G,f}$, where the kernel $f(x|\theta)$ corresponds to a Gaussian kernel with mean parameter $\theta \in \Theta \subset \mathbb{R}^d$, a fixed non-degenerate covariance Σ as given by Eq. (2.11). In addition, the mixing measure $G \in \mathcal{P}(\Theta)$.

1. If f_0 is a Gaussian kernel with mean parameter in a bounded set $\Theta_0 \subset \mathbb{R}^d$ and fixed non-degenerate covariance Σ_0, and $G_0 \in \mathcal{P}(\mathbb{R}^d)$, then the true density p_{G_0,f_0} corresponds to a Gaussian location mixture. The model may be misspecified due to either $\Sigma_0 \neq \Sigma$, or $\text{supp}(G_0) \not\subset \Theta$, or both. In this case, (P.5) is satisfied for

$\alpha \geq 2$. The constant C_0 depends on Σ, Σ_0 as well as $\text{supp}(G_0)$ and Θ. On the other hand, C_1 depends on the eigenvalues of Σ, Σ_0 as well as the value of α.

2. If f_0 is a Gaussian kernel with both mean and covariance parameter varying in some compact subsets of $\mathbb{R}^d$ and positive definite $d \times d$ matrices, respectively, so that the true density p_{G_0,f_0} corresponds to a Gaussian location-scale mixture. In this case, (P.5) is not applicable, but (P.5') holds with $\alpha \geq 2$. The constant C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters. On the other hand, C_1 depends on the value of α chosen, Σ as well as the compact subset corresponding to the covariance parameter.

3. If f_0 is a Student's t kernel, with both mean and covariance parameter varying in some compact subsets of $\mathbb{R}^d$ and positive definite $d \times d$ matrices, respectively, then p_{G_0,f_0} corresponds to a location-scale mixture of t distributions. In this scenario too, (P.5) may not be applicable, but (P.5') is, for any $\alpha > -2$. Both C_0 and C_1 depend on the choice of α. In addition C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters, while C_1 depends on Σ as well as the compact subset corresponding to the covariance parameter.

4. If f_0 is a Laplace kernel with mean parameter in a bounded set $\Theta_0 \subset \mathbb{R}^d$, fixed covariance Σ_0, fixed scale parameter λ_0, and $G_0 \in \mathcal{P}(\mathbb{R}^d)$, (P.5) is satisfied for any $\alpha > -2$. Both C_0 and C_1 depend on the choice of α. In addition C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters, while C_1 depends on Σ as well as the compact subset corresponding to the covariance parameter.

Proposition 2.4.1. Let f be a Gaussian kernel given by (3.8), Θ a bounded subset of $\mathbb{R}^d$. Moreover, assume that f, Θ and the true data generating distribution P_{G_0,f_0} satisfy either condition (P.5) or (P.5') for $\alpha \leq 2$. Then, there exists $\epsilon_0 > 0$ depending on Θ and Σ, such that for any $G, G' \in \mathcal{P}(\Theta)$, whenever $\overline{h}(p_G, p_{G'}) \leq \epsilon_0$, the following inequality holds

$$\overline{h}(p_G, p_{G'}) \geq C \exp\biggl(- (1 + 8\lambda_{\max}(\lambda_{\min}^{-1} + C_0))/W_2^2(G, G')\biggr).$$

Here, $\lambda_{\max}$ and $\lambda_{\min}$ are respectively the maximum and minimum eigenvalue of Σ. C is a constant depending on the parameter space Θ, the dimension d, the covariance matrix Σ, G_0 and C_1 in condition (P.5) or (P.5').

The proof of Proposition 2.4.1 is provided in Appendix 2.6.4. We are ready to prove the first main result of this section.

Theorem 2.4.1. Assume that f satisfies condition specified in Prop. 2.4.1, and Π is an MFM prior on $\mathcal{P}(\Theta)$ specified in Lemma 2.4.3. Then, as n tends to infinity, the following holds,

$$\Pi\biggl(G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_*) \lesssim \Bigl(\frac{\log\log n}{\log n}\Bigr)^{1/2} \Bigg| X_1, \ldots, X_n\biggr) \to 1$$

in p_{G_0,f_0}-probability.

The proof of Theorem 2.4.1 is given in Appendix 2.9.1. The same posterior contraction behaviors hold if we replace MFM prior by the Dirichlet process prior with no change in the proof, except that Lemma 5 of *Nguyen* (2013) is used in place of Lemma 2.4.3.

2.4.2 Laplace location mixtures

Next, we consider a class of multivariate Laplace kernel, a representative in the family of ordinary smooth density functions. It was shown by *Nguyen* (2013) that under a Dirichlet process location mixture with a Laplace kernel, assuming the model is well-specified, the posterior contraction rate of mixing measures to G_0 is of order $n^{-\gamma}$ for some constant $\gamma > 0$. Under the current misspecification setting, we will be able to derive contraction rates toward G_* in the order of $n^{-\gamma'}$ for some constant γ' dependent on γ. The density of location Laplace distributions is given by :

$$f(x|\theta) = \frac{2}{\lambda(2\pi)^{d/2}} \frac{K_{(d/2)-1}\left(\sqrt{2/\lambda}\sqrt{(x-\theta)^\top\Sigma^{-1}(x-\theta)}\right)}{\left(\sqrt{\lambda/2}\sqrt{(x-\theta)^\top\Sigma^{-1}(x-\theta)}\right)^{(d/2)-1}}, \tag{2.12}$$

where Σ and $\lambda > 0$ are respectively fixed covariance matrix and scale parameter such that $|\Sigma| = 1$. Here, K_v is a Bessel function of the second kind of order v. As discussed in *Eltoft et al.* (2006), $K_m(x) \sim \sqrt{\frac{\pi}{2x}}\exp(-x)$ as $|x| \to \infty$. Therefore, there exists $\tilde{R}$ such that as long as $\|x-\theta\| > \tilde{R}$, we have

$$f(x|\theta) \asymp \frac{\exp\left(-\sqrt{\frac{2}{\lambda}}\|x-\theta\|_{\Sigma^{-1}}\right)}{(\|x-\theta\|_{\Sigma^{-1}})^{(d-1)/2}},$$

where we use the shorthand notation $\|y\|_{\Sigma^{-1}} = \sqrt{y^\top\Sigma^{-1}y}$. To ease the ensuing presentation, we denote

$$\tau(\alpha) := \frac{\sqrt{2/(\lambda\lambda_{\max})}}{\left(\sqrt{2/(\lambda\lambda_{\min})} + \sqrt{2/(\lambda\lambda_{\max})} + C_0\right)^{1/\alpha}}.$$

The following proposition provides a key lower bound of weighted Hellinger distance in terms of the Wasserstein metric.

Proposition 2.4.2. Let f be a Laplace kernel given by (3.5) for fixed Σ and λ such that $|\Sigma| = 1$. Moreover, f, Θ and G_0 satisfy either condition (P.5) or (P.5') for some $\alpha \geq 1$. Then, there exists $\epsilon_0 > 0$ depending on Θ, λ and Σ, such that for any $G, G' \in \mathcal{P}(\Theta)$, whenever $\overline{h}(p_G, p_{G'}) \leq \epsilon_0$, the following inequality holds

$$\left(\log \frac{1}{\overline{h}(p_G, p_{G'})}\right)^{d/(2\alpha)} \exp\left(-\tau(\alpha)\left(\log \frac{1}{\overline{h}(p_G, p_{G'})}\right)^{1/\alpha}\right) \geq C W_2^{2/m}(G, G').$$

for any positive constant $m < 4/(4+5d)$. Here, $\lambda_{\max}$ and $\lambda_{\min}$ are respectively the maximum and minimum eigenvalue of Σ. The constant C depends on the parameter space Θ, the dimension d, the covariance matrix Σ, the scale parameter λ, G_0 and C_1 in (P.5) or (P.5').

The proof of Proposition 2.4.2 is provided in Appendix 2.6.5. Given the above result, the posterior contraction rate for mixing measures G in the location family of Laplace mixture distributions can be obtained from the following result:

Theorem 2.4.2. Assume that f is given by equation (3.5) for fixed Σ and λ such that $|\Sigma| = 1$. Additionally, assume that f satisfies condition specified in Prop. 2.4.2, and Π an MFM prior on $\mathcal{P}(\Theta)$ specified in Lemma 2.4.3. Then, as n tends to infinity, the following holds

(i) (Parameter estimation) We have

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_*) \lesssim \exp\left(-\frac{m\tau(\alpha)}{2}\left(\frac{\log n - \log\log n}{2(d+2)}\right)^{1/\alpha}\right) \Big| X_1, \ldots, X_n\right) \to 1$$

in p_{G_0,f_0}-probability for any positive constant $m < 4/(4+5d)$.

(ii) (Density estimation) Moreover, if f satisfies condition (P.5) or (P.5') for some $\alpha < 1$,

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : \|p_G - p_{G_*}\|_{\tilde{L}_q} \lesssim \left(\frac{(\log(n))^2}{n}\right)^{1/q(2d+1)} \middle| X_1, \ldots, X_n \right) \to 1$$

in p_{G_0,f_0}-probability for $1 \leq q \leq 2$.

When $q \geq 2$, we find that

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : \|p_G - p_{G_*}\|_{\tilde{L}_q} \lesssim \left(\frac{(\log(n))^2}{n}\right)^{2/q(2d+1)} \middle| X_1, \ldots, X_n \right) \to 1$$

in p_{G_0,f_0}-probability.

The proof of Theorem 2.4.2 is straightforward using the result in Proposition 2.4.2 and analogous to the proof argument of Theorem 2.4.1; therefore, it is omitted. Note that, identical to the Gaussian kernel case, a similar contraction behavior also holds for the Laplace kernel with the Dirichlet process prior. The proof can be obtained similar to the MFM prior as shown in the case of Gaussian kernels.

Note that we only need condition (P.5') for the proofs of Theorem 2.4.2 and Proposition 2.4.2 to hold. Condition (P.5) is a stricter condition which ensures (P.5'). Examples of condition (P.5) or (P.5') for Laplace mixtures is provided as follows.

Examples In the examples that follow, the statistician decides to fit the data with a Laplace location mixture model $p_{G,f}$, where the kernel $f(x|\theta)$ corresponds to a Laplace kernel with mean parameter $\theta \in \Theta \subset \mathbb{R}^d$, a fixed non-degenerate covariance Σ with $|\Sigma| = 1$ and a scale parameter $\lambda > 0$, as given by Eq. (3.5). In addition, the mixing measure $G \in \mathcal{P}(\Theta)$.

1. If f_0 is a Gaussian kernel with mean parameter in a bounded set $\Theta_0 \subset \mathbb{R}^d$ and fixed non-degenerate covariance Σ_0, and $G_0 \in \mathcal{P}(\mathbb{R}^d)$, then the true density p_{G_0,f_0} corresponds to a Gaussian location mixture. In this case, (P.5) is satisfied for $\alpha \geq 2$. The constant C_0 depends on $\lambda, \Sigma, \Sigma_0$ as well as $\text{supp}(G_0)$ and Θ. On the other hand, C_1 depends on the eigenvalues of λ, Σ, Σ_0 as well as the value of α.

2. If f_0 is a Gaussian kernel with both mean and covariance parameter varying in some compact subsets of $\mathbb{R}^d$ and positive definite $d \times d$ matrices, respectively, so that the true density p_{G_0,f_0} corresponds to a Gaussian location-scale mixture. In this case, (P.5) is not applicable, but (P.5') holds with $\alpha \geq 2$. The constant C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters. On the other hand, C_1 depends on the value of α chosen, Σ as well as the compact subset corresponding to the covariance parameter.

3. If f_0 is a Student's t kernel, with both mean and covariance parameter varying in some compact subsets of $\mathbb{R}^d$ and positive definite $d \times d$ matrices, respectively, then p_{G_0,f_0} corresponds to a location-scale mixture of t distributions. In this scenario too, (P.5) may not be applicable, but (P.5') is, for any $\alpha > -1$. Both C_0 and C_1 depend on the choice of α. In addition C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters, while C_1 depends on Σ as well as the compact subset corresponding to the covariance parameter.

4. If f_0 is a Laplace kernel with mean parameter in a bounded set $\Theta_0 \subset \mathbb{R}^d$, fixed covariance Σ_0, fixed scale parameter λ_0, and $G_0 \in \mathcal{P}(\mathbb{R}^d)$, (P.5) is satisfied for any

$\alpha \geq 1$. Both C_0 and C_1 depend on the choice of α. In addition C_0 depends on Σ, Θ as well as the compact subsets corresponding to the location and covariance parameters, while C_1 depends on Σ as well as the compact subset corresponding to the covariance parameter.

5. If f_0 is a Laplace kernel with mean, scale and covariance parameters varying in some compact subsets of $\mathbb{R}_+$, $\mathbb{R}^d$ and positive definite $d \times d$ matrices with determinant 1, respectively, so that the true density p_{G_0,f_0} corresponds to a Laplace location-scale mixture. In this case, (P.5') holds with $\alpha \geq 1$. The constant C_0 depends on Σ, λ, Θ as well as the compact subsets corresponding to the location, scale and covariance parameters. On the other hand, C_1 depends on the value of α chosen, Σ as well as the compact subsets corresponding to the scale and covariance parameters.

Remarks (i) It is worth noting that compared to the well-specified setting, the posterior contraction upper bound obtained for Gaussian location mixtures remains the same slow logarithmic rate $(\log\log n/\log n)^{1/2}$. For Laplace mixtures, when the truth f_0 satisfies condition (P.5) with $\alpha \leq 1$, the posterior contraction upper bound obtained under misspecification remains a polynomial rate of the form $n^{-\gamma'}$ modulo a logarithmic term. Due to misspecification there is a loss of a constant factor in the exponent γ', which is dependent on the shape of the kernel density as it is captured by the term $\tau(\alpha)$.

(ii) Although Gaussian mixtures have proved to be an asymptotically optimal density estimation device under suitable and mild conditions (cf. *Ghosal and van der Vaart* (2007)), the results obtained in this section raise some cautions for Gaussian kernels as a choice for mixture modeling under model misspecification, even if the true G_0 has finite number of support points, when the primary interest is in the quality of model parameter estimates. Mixtures of heavy-tailed and ordinary smooth kernel densities such as the Laplace prove to be more amenable to efficient parameter estimation. Thus, the modeler may be tempted to select for f, say, a Laplace kernel over a supersmooth kernel such as Gaussian kernel, provided that either condition (P.5) or (P.5') is valid.

(iii) It is interesting to consider the scenario where the true kernel f_0 happens to be a Gaussian kernel: if we use either a well-specified or a misspecified Gaussian kernel to fit the data, the posterior contraction bound is the extremely slow $(\log\log n/\log n)^{1/2}$ accordingly to Theorem 2.4.1. This rate may be too slow to be practical interpretation of parameters. If the statistician is too impatient to get to the truth G_0, because sample size n is not sufficiently large, he may well decide to select a Laplace kernel f instead. Despite the intentional misspecification, he might be comforted by the fact that the posterior distribution of G contracts at an exponentially faster rate to a G_* given by Theorem 2.4.2 for $\alpha = 2$. It is of interest, in theory at least, in this scenario to study the relation between G_* and true G_0, given certain assumptions on the true density p_{G_0,f_0}.

Practical implications All models are misspecified in practice. In particular, when the kernel is misspecified, in general the limiting mixing measure G_* would have infinite support. Since the mixing measure G is a device for representing the heterogeneity of the data population, this means when we employ (Bayesian) nonparametric models in practice, the more data we have the more heterogeneous patterns will show up via posterior estimates. As such, Theorems 2.4.1 and 2.4.2 inform us how the choice of the kernel affects the quality of the estimates for G. In the language of Bayesian inference, the theorems quantify in an asymptotic sense the role of data sample in transforming the prior distribution to the posterior distribution on the quantity of interest, whereas the matter of consistency toward the truth G_0 is left unknown (and in fact, unknowable in practice). On the other hand, these theorems should not be viewed as an endorsement of one kernel choice over another. It does not make sense to use G as a device for heterogeneity of the data population unless the kernel choice f is believed to be meaningful, i.e., f is sufficiently close to the true f_0. This is how a practitioner typically assumes. Once such a kernel choice f has been made, we have shown that some (misspecified) kernels result in more efficient estimates, and hence more conductive to interpretation, than others.

2.4.3 When G_* has finite support

The source of the deterioration in the statistical efficiency of parameter estimation under model misspecification is ultimately due to the increased complexity of the limiting point G_*. Even if the true G_0 has a finite number of support points, this is not the case for G_* in general. Unfortunately, it is very difficult to gain concrete information about G_* both in practice and in theory, due to the lack of knowledge about the true p_{G_0,f_0}. When some precious information about G_* is available, specifically, suppose that we happen to know G_* has a bounded number of support points k_* such that $k_* < \overline{k}$ for some known $\overline{k}$. Then it is possible to devise a new prior specification on the mixing measure G so that one can gain a considerably improved posterior contraction rate toward G_*. We will show that it is possible to obtain the contraction rate of the order $(\log n/n)^{1/4}$ under W_2 metric — this is the same rate of posterior contraction one would get with overfitted mixtures in the well-specified regime.

In order to analyze the convergence rate of mixing measure under that setting of k_*, we introduce a relevant notion of integral Lipschitz property, which is a generalized form of the uniform Lipschitz property for the misspecification scenarios.

Definition 2.4.2. For any given $r \geq 1$, we say that the family of densities f admits the *integral Lipschitz* property up to the order r with respect to two mixing measures

G_0 and G_* , if f as a function of θ is differentiable up to the order r and its partial derivatives with respect to θ satisfy the following inequality

$$\sum_{|\kappa|=r} \left| \left(\frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta_1) - \frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta_2) \right) \gamma^{\kappa} \right| \leq C(x) \|\theta_1 - \theta_2\|^{\delta} \|\gamma\|^{r}$$

for any $\gamma \in \mathbb{R}^d$ and for some positive constants δ independent of x and $\theta_1, \theta_2 \in \Theta$. Here, $C(x)$ is some function such that $\displaystyle\int C(x) \frac{p_{G_0, f_0}(x)}{p_{G_*}(x)} \mathrm{d}x < \infty$.

It is clear that when f has integral Lipschitz property up to the order r, for some $r \geq 1$, with respect to G_0 and G_*, then it will admit uniform Lipschitz property up to the order r. We can verify that the first order intergral Lipschitz property is satisfied by many popular kernels, including location-scale Gaussian distribution and location-scale Cauchy distribution.

In the following we shall work with the MFM prior (2.7). Moreover,

(M.0) q_K places positive masses on $K \in \left\{1, \ldots, \overline{k}\right\}$ and 0 mass elsewhere, where $\overline{k} \gg k_*$ is a fixed number.

Given that k_* is finite, we obtain a key lower bound of weighted Hellinger distance in terms of the Wasserstein metric under strong identiability of f:

Proposition 2.4.3. Assume that f is second order identifiable and admits uniform integral Lipschitz property up to the second order. Then, for any $G \in \mathcal{O}_{\overline{k}}$, the following inequality holds

$$\overline{h}(p_G, p_{G_*}) \gtrsim W_2^2(G, G_*).$$

The proof of Proposition 2.4.3 is in Appendix 2.7.2. Before stating the final theorem of this section, we will need following assumptions:

(M.1) The assumptions of Proposition 2.4.3 hold, i.e., f is second order identifiable and admits uniform integral Lipschitz property up to the second order.

(M.2) There exists $\epsilon_0 > 0$ such that $\int (p_{G_0,f_0}(x))p_{G_*}(x)/p_G(x)d\mu(x) \leq M^*(\epsilon_0)$ whenever we have $W_1(G, G_*) \leq \epsilon_0$ for any $G \in \mathcal{O}_{k_*}$ where $M^*(\epsilon_0)$ depends only on ϵ_0, G_*, G_0, and Θ.

(M.3) The parameter γ in Dirichlet distribution in MFM satisfies $\gamma < \overline{k}$. Additionally, the base distribution H satisfies Assumption (P.2).

Theorem 2.4.3. Assume $k_0 < \infty$, and assumptions (M.0),(M.1),(M.2) and (M.3) hold. Then we have that,

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_*) \lesssim (\log n/n)^{1/4} \middle| X_1, \ldots, X_n\right) \to 1$$

in p_{G_0,f_0}-probability.

The proof of Theorem 2.4.3 is deferred to Section 2.9.5.

Further remarks The above theorem raises a promising prospect for combating model misspecification, by having the modeler fit the data to an *underfitted* mixture model p_G. Unfortunately, this theorem does not address this scenario, under which the limiting mixing measure would correspond to the KL minimizer

$$G_{**} = \mathop{\arg\min}_{G \in \mathcal{O}_{\overline{k}}(\Theta)} K(p_{G_0,f_0}, p_G).$$

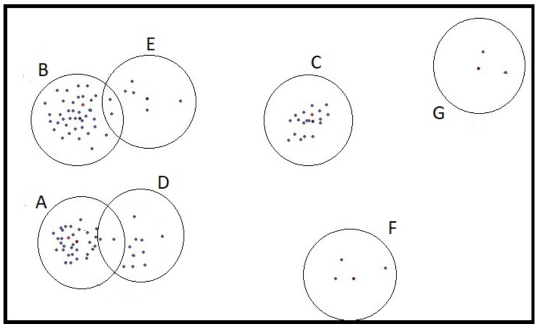


Figure 2.1: *Initial distribution G.*

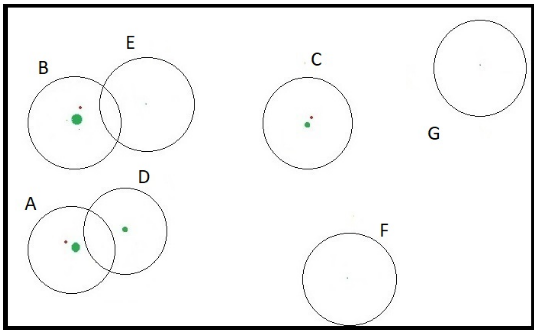


Figure 2.2: *After first stage-"merge".*

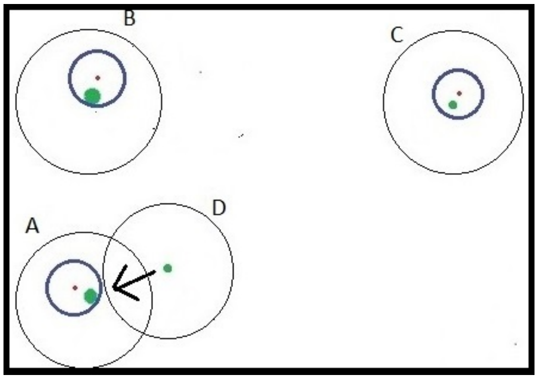


Figure 2.3: *After second stage-"truncation".*

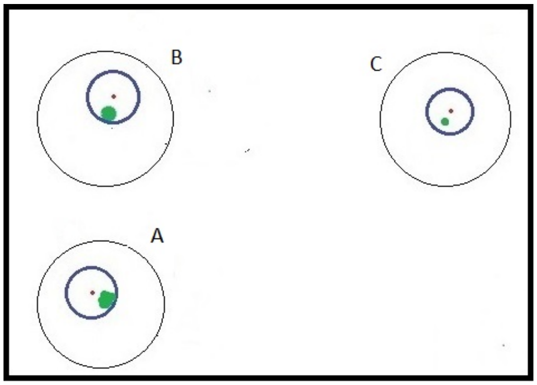


Figure 2.4: *After second stage-"merge".*

for some $\overline{k} < \infty$, provided that this quantity exists (compare this with G_* given in (2.9)). Due to the lack of convexity of the class of mixture densities with bounded number of mixture components, the theory developed in this section (tracing back to the work of *Kleijn and van der Vaart* (2006)) is not applicable. Thus, posterior contraction behaviors in an underfitted mixture models remain an interesting open question.

2.5 Simulation studies

In this section we provide an illustration of the MTM algorithm's behavior via a simple simulation study. Figures 2.1, 2.2, 2.3 and 2.4 illustrate the different stages in the application of MTM algorithm 2.1. In each figure, green dots denote the atoms in the set of "remaining atoms" at each stage, with weights proportional to their sizes. Red dots denote the supporting atoms of the true mixing measure G_0. Black circles denote balls of radius ω_n around each of the "remaining atoms". Blue circles denote balls of radius $\frac{\omega_n}{4k_0}$ around the atoms of G_0.

Starting with an input measure G represented in Fig. 2.1, the first stage of the algorithm (merge procedure, from line 1 to line 4) merges nearby atoms to produce G', which is represented by Fig. 2.2. There remains some atoms that carry very small mass, they are suitably truncated (via line 5 in the algorithm), and then merged accordingly (via line 6). Fig. 2.3 and Fig. 2.4 represent the outcome after these two steps of the algorithm. Observe how the atoms in each of the blue circles are merged to produced a reasonably accurate estimate of the corresponding atom of G_0. The number of such circles gives the correct number of the supporting atoms of G_0.

Next, we illustrate the performance of the MTM algorithm as it is applied to the samples from a Dirichlet process mixture, given the data generated by mixtures of three location Gaussian distributions:

$$p_{G_0}(\cdot) = \sum_{i=1}^{3} p_i^0 \mathcal{N}(\cdot|\mu_i^0, \Sigma^0)$$

where $\mathcal{N}(\cdot|\mu, \Sigma)$ is the Gaussian distribution with mean vector μ and covariance matrix Σ. For simulation purposes, we consider the following four different settings (n is the sample size):

1. Case A: $\mu_1^0 = (0.8, 0.8)$, $\mu_2^0 = (0.8, -0.8)$, $\mu_3^0 = (-0.8, 0.8)$, $\Sigma^0 = 0.05 I_3$, $n = 500$.

2. Case B: $\mu_1^0 = (0.8, 0.8)$, $\mu_2^0 = (0.8, -0.8)$, $\mu_3^0 = (-0.8, 0.8)$, $\Sigma^0 = 0.05 I_3$, $n = 1500$.

3. Case C: $\mu_1^0 = (1.8, 1.8)$, $\mu_2^0 = (1.8, -1.8)$, $\mu_3^0 = (-1.8, 1.8)$, $\Sigma^0 = 0.05 I_3$, $n = 500$.

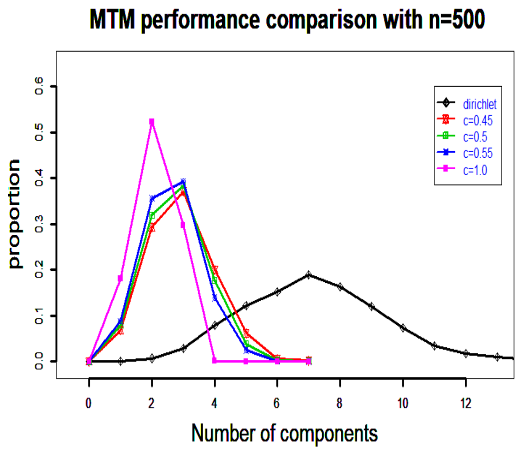


Figure 2.5: *Case A.*

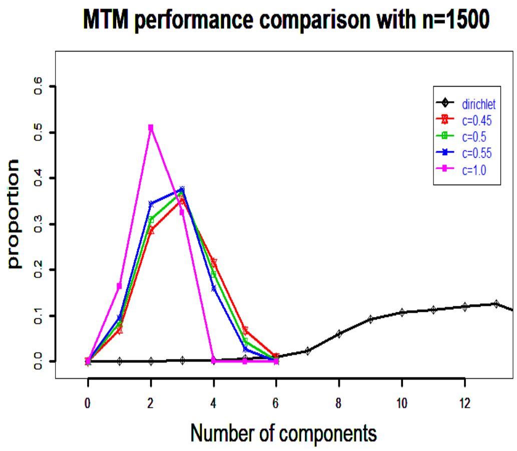


Figure 2.6: *Case B.*

4. Case D: $\mu_1^0 = (0.8, 0.8)$, $\mu_2^0 = (0.8, -0.8)$, $\mu_3^0 = (-0.8, 0.8)$, $\Sigma^0 = 0.01 I_3$, $n = 1500$.

Here, I_3 is the identity matrix of dimension 3. Additionally, the weight vector for all these cases is chosen as $p^0 = (p_1^0, p_2^0, p_3^0) = (0.4, 0.3, 0.3)$.

As mentioned above, a Dirichlet process prior with an uniform prior base measure H in the region $[-6, 6] \times [-6, 6]$, along with concentration parameter $\alpha = 1$. This choice of prior enables us to sample significantly larger numbers of components of the mixing measure than the true number of three components.

It is known that the contraction rate of mixing measures under location Gaussian DPMM is $\tilde{C}(\log(n)^{-1/2})$ with respect to the Wasserstein-2 norm, for some constant $\tilde{C}$ which depends on Σ^0(the covariance matrix), the location parameters μ_i^0 and the weights p_i^0 (*Nguyen*, 2013). For our purpose, in order for ω_n to satisfy Equation (2.8), we may choose any ω_n as long as $\frac{\omega_n}{\log(n)^{-1/2}} \to \infty$. We selected $\omega_n = \left(\frac{\log(\log(n))}{(\log(n))}\right)^{1/2}$ for all our applications of the MTM algorithm.

The MTM algorithm is provably consistent (in the asymptotic sense) for all chosen constants $c > 0$. In practice for n being fixed, the input c to Algorithm 2.1 should be chosen so that $\frac{\tilde{C}}{(\log(\log(n)))^{1/2}} \leq c$. Moreover, for finite n it is not expected that the posterior probability for $k = k_0$ is close to 1. However, for identifying the number of components the posterior mode provides a reasonable estimate. In particular, $(1 - \sum_{i=1}^{3} \frac{c}{p_i^0})$ forms a useful lower bound on the posterior mass at the true parameter as identified in Equation (2.25) in the supplement. To identify $k = k_0$ consistently using the posterior mode safely, one needs to choose $c < c_0$, with c_0 satisfying $(1 - \sum_{i=1}^{3} \frac{c_0}{p_i^0}) > 1/2$. The exact computation of the upper bound c_0 and the lower bound $\frac{\tilde{C}}{(\log(\log(n)))}$ for c may be unrealistic but a reasonable estimate may be possible. Nonetheless, we simply considered a large range of c and show there is a range where we can robustly identify the true number of components via the posterior mode.

For the DP mixture's posterior computation, we make use of the non-conjugate split-merge sampler of Jain and Neal *Jain and Neal* (2007) with $(5, 1, 1, 5)$ scheme, i.e., 5 scans to reach the split launch state, 1 split-merge move per iteration, 1 Gibbs scan per iteration, and 5 moves to reach the merge launch state. We run our experiments for two settings corresponding to sample sizes 500 and 1500. The sampler had 2000 burn-in iterations followed by 18000 sample iterations (a total 20000), with each 10th iteration being counted.

The experiments run for DP mixture-based sampler, followed by application of the MTM procedure for 4 different values of the tuning parameter c in Algorithm 2.1, namely, for $c = 0.45, 0.5, 0.55, 1.0$. The proportional frequencies are plotted in Figure 2.5 and Figure 2.6 respectively, along with the proportional frequencies for DP mixture. The uniform base measure for the Dirichlet Process prior is chosen so as to enable easier creation of newer components in the split-merge scheme. As a consequence the DP mixture's posterior yields quite bad results as far as the number of mixture components is concerned. However, even under that case, we can recover the true

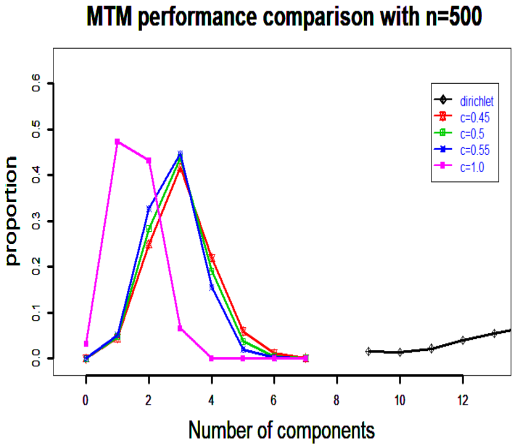


Figure 2.7: *Case C.*

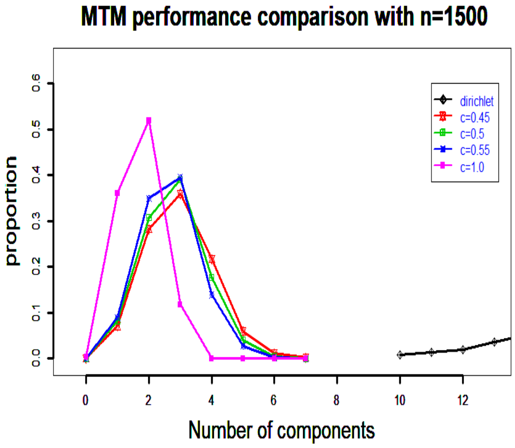


Figure 2.8: *Case D.*

number of components by considering the mode of the frequency distribution after an application of the MTM algorithm on the posterior samples, with appropriate constant c. It is expected, however, that a large choice of c would underestimate the number of components. This is also what is observed from the simulations, where the procedure breaks down when $c = 1.0$.

We perform the experiments under four different settings of data populations. In particular, figure 2.7 consists of data generated from mixture of Gaussians with more widely spread location parameter values. In this case, it is expected that the convergence to the true number of components via Algorithm 2.1 will be faster for the posterior mode, in comparison to the situation where the location parameters are closer together. This is indeed what is observed in our simulations. The value of the covariance matrix Σ^0, on the other hand does not seem to noticeably affect the results. This is again expected, since the prior support $[-6, 6] \times [-6, 6]$ is quite large in comparison to the eigenvalues of the covariance matrix chosen.

2.6 Appendix A: Proofs of key results

In this appendix, we provide proofs for several key results in the chapter.

2.6.1 Proof of Theorem 2.3.1

The proof of the theorem consists of two key parts. First, we recall a general framework for establishing posterior contraction of mixing measures. Then we proceed to apply this framework to analyze the specific setting of the MFM model.

2.6.1.1 General framework

To establish convergence rates of mixing measures under the setting of MFM, we utilize the general framework of posterior contraction of mixing measures under well-specified setting from *Nguyen* (2013). To state such results formally, we will need to introduce several key definitions in harmony with the notations in this chapter. Let G be endowed with a prior distribution Π on a measure space of discrete probability measures in $\overline{\mathcal{G}}(\Theta)$. Fix $G_0 \in \mathcal{P}(\Theta)$. For any set $\mathcal{S} \subset \overline{\mathcal{G}}(\Theta)$, we define the Hellinger information of the W_1 metric for subset $\mathcal{S}$ by the following function

$$\Psi_{\mathcal{S}}(r) := \inf_{G \in \mathcal{S}:\ W_1(G, G_0) \geq r/2} h^2(p_G, p_{G_0}).$$

Note that, the choice of first order Wasserstein metric in the above formulation is due to the lower bound of Hellinger distance between mixing densities in terms of first order Wasserstein distance between their corresponding mixing measures in (2.1). Now, for any mixing measure $G_1 \in \overline{\mathcal{G}}(\Theta)$ and $r > 0$, we define a Wasserstein ball centered at G_1

under W_1 metric as follows

$$B_{W_1}(G_1, r) = \left\{G \in \overline{\mathcal{G}}(\Theta):\; W_1(G, G_1) \le r\right\}.$$

Furthermore, for any $M > 0$, we define a Kullback-Leibler neighborhood of G_0 by

$$B_K(\epsilon, M) = \left\{G \in \overline{\mathcal{G}}(\Theta):\; K(p_{G_0}, p_G) \le \epsilon^2 \log\left(\frac{M}{\epsilon}\right), K_2(p_{G_0}, p_G) \le \epsilon^2 \left(\log\left(\frac{M}{\epsilon}\right)\right)^2\right\}$$

For the proof of Theorem 2.3.1, we use a straightforward extension of Theorem 4 in *Nguyen* (2013), adapted to the setting in this work.

Theorem 2.6.1. Fix $G_0 \in \overline{\mathcal{G}}$.
Define $M(\mathcal{G}, G, r) := D\left(\frac{\Psi_{\mathcal{G}}(r)^{1/2}}{2\text{Diam}(\Theta)^{\alpha-1}\sqrt{C_1}}, \mathcal{G} \cap B_{W_1}(G, r/2), W_1\right)$ for any $\mathcal{G} \subset \overline{\mathcal{G}}$.
Assume the following:

(a) The family of likelihood functions is finitely identifiable and satisfies $h(f(x|\theta_i), f(x|\theta'_j)) \le C_1\|\theta_i - \theta'_j\|^\alpha$ for any $\theta_i, \theta'_j \in \Theta$, for some constants $C_1 > 0$, $\alpha \ge 1$.

(b) There is a sequence $\epsilon_n \to 0$ such that $n\epsilon_n^2$ is bounded away from 0 or tending to infinity, a constant $M > 0$ sufficiently large, and a sequence M_n such that

$$\begin{aligned}
&\log D(\epsilon/2, \mathcal{G}_n \cap B_{W_1}(G_0, 2\epsilon)\backslash B_{W_1}(G_0, \epsilon), W_1) \\
&\qquad + \sup_{G \in \mathcal{G}_n} \log M(\mathcal{G}_n, G, r) \leq n\epsilon_n^2, \ \forall\ \epsilon \geq \epsilon_n, && (2.13) \\
&\frac{\Pi(\overline{\mathcal{G}}\backslash\mathcal{G}_n)}{\Pi(B_K(\epsilon_n, M))} = o\left(\exp\left(-2n\epsilon_n^2 \log\left(\frac{M}{\epsilon_n}\right)\right)\right) && (2.14) \\
&\frac{\Pi(B_{W_1}(G_0, 2j\epsilon_n)\backslash B_{W_1}(G_0, j\epsilon_n))}{\Pi(B_K(\epsilon_n, M))} \leq \exp\bigg(n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/16\bigg), \\
&\qquad\qquad\qquad\qquad \forall j \geq M_n && (2.15) \\
&\exp\left(2n\epsilon_n^2 \log\left(\frac{M}{\epsilon_n}\right)\right) \sum_{j \geq M_n} \exp\left(-n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/16\right) \to 0. && (2.16)
\end{aligned}$$

Then, we have that $\Pi(G \in \overline{\mathcal{G}} : \ W_1(G, G_0) \geq M_n\epsilon_n | X_1, \ldots, X_n) \to 0$ in P_{G_0}-probability.

2.6.1.2 Posterior contraction under MFM

Now, we apply the above result to establish the convergence rate of mixing measure under a well-specified MFM model. The constant M for part (c) of Theorem 2.6.1 is chosen later. Also, let $\epsilon_n := \overline{M}(\log n/n)^{1/2}$ where $\overline{M}$ is a sufficiently large constant that will be chosen later. Note that it is enough to show, $\Pi\bigg(G \in \mathcal{G}(\Theta) : W_1(G, G_0) \gtrsim \dfrac{(\log n)^{1/2}}{n^{1/2}}\bigg| X_1, \ldots, X_n\bigg) \to 0$, since $\Pi(G \in \overline{\mathcal{G}}(\Theta) \setminus \mathcal{G}(\Theta) | X_1, \ldots, X_n) = 0$.

With ϵ_n chosen as above, we denote $A_n := \Pi(G \in \mathcal{G}(\Theta) : W_1(G, G_0) \gtrsim \epsilon_n | X_1, \ldots, X_n)$. It is clear that

$$\begin{aligned}
A_n &= \sum_{k=1}^{\infty} \Pi(G \in \mathcal{O}_k(\Theta) : W_1(G, G_0) \gtrsim \epsilon_n | X_1, \ldots, X_n)\Pi(K = k | X_1, \ldots, X_n) \\
&\leq \Pi(G \in \mathcal{O}_{k_0}(\Theta) : W_1(G, G_0) \gtrsim \epsilon_n | X_1, \ldots, X_n) + \Pi(K \neq k_0 | X_1, \ldots, X_n).
\end{aligned}$$

Now, we divide our proof into the following key steps

Step 1: $\Pi(K = k_0 | X_1, \ldots, X_n) \to 1$ a.s. P_{G_0}. As the model is identifiable, this result is the direct application of Doob's consistency theorem *Doob* (1948).

Step 2: $P_{G_0}\biggl(\Pi(G \in \mathcal{O}_{k_0}(\Theta) : W_1(G, G_0) \gtrsim \epsilon_n | X_1, \ldots, X_n)\biggr) \to 0$ as $n \to \infty$. The proof of this result is the application of Theorem 2.6.1. In fact, as we focus on the posterior contraction of G_0 from $G \in \mathcal{O}_{k_0}(\Theta)$, we denote the prior on $G \in \mathcal{O}_{k_0}(\Theta)$ to be $\Pi = H \times Q$ where $Q \overset{d}{=} \text{Dir}(\gamma/k_0, \ldots, \gamma/k_0)$. Now, we claim that

$$\begin{aligned} \epsilon^{c_H} &\lesssim H(\|\theta_i - \theta_i^0\| \leq \epsilon, \ i = 1, \ldots, k_0) \lesssim \epsilon^{c_H} \\ \epsilon^{\gamma'} &\lesssim Q(|p_i - p_i^0| \leq \epsilon, \ i = 1, \ldots, k_0) \end{aligned} \tag{2.17}$$

where $c_H > 0$ and $\gamma' > 0$ are some positive constants and ϵ is sufficiently small. The proof of claim (2.17) can be obtained in *Ghosal and van der Vaart* (2017). To facilitate the discussion, we further divide Step 2 into two small steps.

Step 2.1: To obtain the bound for $\Pi(B_K(\epsilon_n, M))$, we utilize the result from *Wong and Shen* (1995) to bound KL divergence and squared KL divergence. In particular, from Theorem 5 of *Wong and Shen* (1995), if p and q are two densities such that $2h^2(p, q) \leq \epsilon^2$ and $\int p^2/q \leq M^2$ then we obtain that $K(p, q) \lesssim \epsilon^2 \log(M/\epsilon)$ and $K_2(p, q) \lesssim \epsilon^2(\log(M/\epsilon))^2$ where the constants in these bounds are universal.

Now, since f admits Lipschitz continuity up to the first order, we achieve that $h^2(p_G, p_{G_0}) \leq C_1 W_1(G, G_0)$ for any $G \in \mathcal{O}_{k_0}(\Theta)$ where C_1 is a positive constant depending only on Θ. Now, for any $G \in \mathcal{O}_{k_0}(\Theta)$ such that $W_1(G, G_0) \leq C\epsilon_n^2$ where $C < \epsilon_0$ is a sufficiently small constant to be chosen later, the previous bound implies that

$h^2(p_G, p_{G_0}) \leq C_1 C\epsilon_n^2$. Since $C\epsilon_n^2 \leq \epsilon_0$ for all n sufficiently large, we also have that $\int (p_{G_0}(x))^2/p_G(x)\mathrm{d}\mu(x) \leq M(\epsilon_0)$ according to assumption (P.3). Combining all the previous results, we achieve that

$$\begin{aligned} K(p_{G_0}, p_G) &\lesssim \epsilon_n^2 \log(\sqrt{M(\epsilon_0)}/\sqrt{CC_1}\epsilon_n), \\ K_2(p_{G_0}, p_G) &\lesssim \epsilon_n^2 (\log(\sqrt{M(\epsilon_0)}/\sqrt{CC_1}\epsilon_n))^2 \end{aligned}$$

when $\overline{M}$ is sufficiently large. Define $M := \sqrt{M(\epsilon_0)/CC_1}$. Therefore, we have

$$\Pi(B_K(\epsilon_n, M)) \geq \Pi(G \in \mathcal{O}_{k_0}(\Theta) : W_1(G, G_0) \leq C\epsilon_n^2).$$

For any $G = \sum_{i=1}^{k_0} p_i \delta_{\theta_i}$ such that $\|\theta_i - \theta_i^0\| \leq \overline{\epsilon}$ and $|p_i - p_i^0| \leq \overline{\epsilon}/(k_0 \mathrm{Diam}(\Theta))$ for any $1 \leq i \leq k_0$ and sufficiently small $\overline{\epsilon} > 0$, we can check that

$$W_1(G, G_0) \leq \sum_{i=1}^{k_0} p_i^0 \wedge p_i \|\theta_i - \theta_i^0\| + \sum_{i=1}^{k_0} |p_i - p_i^0| \mathrm{Diam}(\Theta) \leq 2\overline{\epsilon}.$$

Hence, by choosing universal constant C such that $C\epsilon_n^2 \leq \overline{\epsilon}$, we would have that

$$\begin{aligned} &\Pi(G \in \mathcal{O}_{k_0}(\Theta) : W_1(G, G_0) \leq C\epsilon_n^2) \\ &\quad \geq \Pi(G \in \mathcal{O}_{k_0}(\Theta) : \ \|\theta_i - \theta_i^0\| \leq C\epsilon_n^2, \ |p_i - p_i^0| \leq C\epsilon_n^2/(k_0 \mathrm{Diam}(\Theta)), \ \forall \ 1 \leq i \leq k_0) \\ &\quad \gtrsim \epsilon_n^{2(c_H + \gamma')} \end{aligned} \tag{2.18}$$

where the last inequality is due to the results from claim (2.17).

Step 2.2: To apply the posterior contraction rate result of Theorem 2.6.1, we choose $\mathcal{G}_n = \mathcal{O}_{k_0}(\Theta)$ for all n. Now, it is clear that $\Pi(\overline{\mathcal{G}} \backslash \mathcal{G}_n) = 0$. Therefore, condition (2.14) is obviously satisfied. Additionally, by means of Lemma 4 in *Nguyen* (2013), we can check that condition (2.13) is satisfied with our choice of ϵ_n as $\overline{M}$ is sufficiently large. For condition (2.15), from the bound of KL neighborhood in (2.18), we find that

$$\frac{\Pi(B_{W_1}(G_0, 2j\epsilon_n) \backslash B_{W_1}(G_0, j\epsilon_n))}{\Pi(B_K(\epsilon_n, M))} \lesssim \epsilon_n^{-2(c_H+\gamma)}.$$

Since f is first order identifiable and admits uniform Lipschitz property up to the first order, according to (2.1), we obtain that $\Psi_{\mathcal{G}_n}(r) \gtrsim Cr^2$ for any $r > 0$ where C is some positive constant that depends only on G_0 and Θ. Therefore, we have

$$\exp\biggl(n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/16\biggr) \geq \exp(nC(j\epsilon_n)^2)/16) \geq n^{CM_n^2/16}$$

for any $j \geq M_n$. By choosing M_n such that $M_n^2 \geq 32(c_H + \gamma)/C$, it is clear that condition (2.15) is satisfied.

For condition (2.16), combining with the above bound of $\Psi_{\mathcal{G}_n}(r)$, we would have that

$$\begin{aligned} \exp\left(2n\epsilon_n^2 \log\left(\frac{M}{\epsilon_n}\right)\right) \sum_{j \geq M_n} \exp\left(-n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/16\right) &\lesssim n^{\overline{M}^2} \sum_{j \geq M_n} n^{-CM_n^2/16} \\ &\lesssim n^{\overline{M}^2 - \frac{CM_n^2}{16}} \to 0 \end{aligned}$$

as long as M_n is chosen such that $M_n^2 \geq \frac{32\overline{M}^2}{C}$. Therefore, condition (2.16) is satisfied. As a consequence, we achieve the conclusion of the theorem.

Proof of claim (2.17) According to the formulation of Dirichlet distribution, we obtain that

$$\begin{aligned}
& Q(|p_i - p_i^0| \leq \epsilon,\ i = 1, \ldots, k_0) \\
= \ & \frac{\Gamma(\gamma)}{(\Gamma(\gamma/k_0))^{k_0}} \int\limits_{|p_i - p_i^0| \leq \epsilon,\ 1 \leq i \leq k_0} \prod_{i=1}^{k_0-1} p_i^{\gamma/k_0 - 1} \left(1 - \sum_{i=1}^{k_0-1} p_i\right)^{\gamma/k_0 - 1} dp_1 \ldots dp_{k_0-1} (2.19)
\end{aligned}$$

Now, for $\gamma/k_0 \leq 1$, equation (2.19) can be re-written as:

$$\begin{aligned}
& Q(|p_i - p_i^0| \leq \epsilon,\ i = 1, \ldots, k_0) \\
& \quad \geq \frac{\Gamma(\gamma)}{(\Gamma(\gamma/k_0))^{k_0}} \int\limits_{|p_i - p_i^0| \leq \epsilon,\ 1 \leq i \leq k_0} \prod_{i=1}^{k_0-1} p_i^{\gamma/k_0 - 1} dp_1 \ldots dp_{k_0-1} \\
& \quad \gtrsim \frac{\Gamma(\gamma)}{(\Gamma(\gamma/k_0))^{k_0}} \frac{1}{(\gamma/k_0)^{k_0}} \epsilon^{(k_0-1)(\gamma/k_0)}.
\end{aligned}$$

Here, the first inequality in the above display follows from the fact that $1 - \sum\limits_{i=1}^{k_0-1} p_i \leq 1$ while the second inequality is due to direct integration and the fact that ϵ is sufficiently small.

On the other hand, for $\gamma/k_0 > 1$, we can rewrite equation (2.19) as

$$\begin{aligned}
& Q(|p_i - p_i^0| \leq \epsilon,\ i = 1, \ldots, k_0) \\
& \quad \geq \frac{\Gamma(\gamma)}{(\Gamma(\gamma/k_0))^{k_0}} \left(\frac{1 - \sum\limits_{i=1}^{k_0-1} p_i^0}{2} \right)^{\gamma/k_0 - 1} \int\limits_{|p_i - p_i^0| \leq \epsilon,\ 1 \leq i \leq k_0} \prod_{i=1}^{k_0-1} p_i^{\gamma/k_0 - 1} dp_1 \ldots dp_{k_0-1}
\end{aligned}$$

where the above inequality follows due to the fact that $p_i^0 > 0$ for all $i \in \{1, \ldots, k_0\}$ and that for sufficiently small $\epsilon > 0$ such that $|p_i - p_i^0| \leq \epsilon$ for all $i \in \{1, \ldots, k_0 - 1\}$, we have $p_{k_0} \geq p^0_{k_0}/2$. Therefore, the lower bound for the Dirichlet distribution Q follows automatically from the results of these two separate conditions of $\gamma/k_0 \leq 1$ and $\gamma/k_0 > 1$.

On the other hand, to show the bounds for H, we note that $\Theta \subset \mathbb{R}^d$. Suppose $B(\theta, \epsilon) := \{\theta' \subset \Theta : \|\theta - \theta'\| \leq \epsilon\}$ denotes the ℓ_2 ball in $\Theta \cap \mathbb{R}^d$ around θ, with radius ϵ. Then it can be seen that

$$\begin{aligned}\min_{\theta' \in \Theta}(g(\theta')\mu(B(\theta', \epsilon))/\mu(\Theta))^{k_0} &\leq H(\|\theta_i - \theta^0_i\| \leq \epsilon,\ i = 1, \ldots, k_0)\\ &\leq \max_{\theta' \in \Theta}(g(\theta')\mu(B(\theta', \epsilon))/\mu(\Theta))^{k_0}\end{aligned}$$

since $B(\theta^0_i, \epsilon)$ are disjoint for all $i \leq k_0$, for sufficiently small $\epsilon > 0$. Here, $\mu(A)$ denotes the d-dimensional Lebesgue measure of the set $A \subset \Theta$ and g is the density function of H based on Assumption (P.2). Using the fact that $\epsilon^d \lesssim \mu(B(\theta', \epsilon)) \lesssim \epsilon^d$ and the condition that H is approximately uniform in Assumption (P.2), the remainder of the claim follows.

2.6.2 Posterior consistency of Merge-Truncate-Merge algorithm

The goal of this section is to both deliver a proof of Theorem 2.3.2 and clarify the role played by each of the steps of the MTM algorithm.

2.6.2.1 Probabilistic scheme for merging atoms

The first step of MTM algorithm comprises of lines from 1 to 3 in Algorithm 1. It describes a probabilistic scheme for merging atoms from an input measure G. Recall that G is a sample from the posterior distribution of a mixing measure which is assumed to be relatively close to the true G_0, per Eq. (2.8). To simplify notations within this subsection we shall remove subscript n in ω_n in (2.8), namely, we will not incorporate the randomness of data in the results in this subsection.

In that regard, suppose that we have a measure $G = \sum_j p_j \delta_{\theta_j} \in \overline{\mathcal{G}}(\Theta)$ such that $W_r(G, G_0) \leq \delta\omega$ for some $r \geq 1$. Here, δ, ω are sufficiently small such that the following two properties hold:

(B.1) $\omega < \min\{(p^0_{\min}/2)^{1/r}, \min_{u \neq v} \frac{\|\theta^0_u - \theta^0_v\|}{8}\}$.

(B.2) $\sqrt{\delta} < p^0_{\min}/(2k_0)$, where $p^0_{\min} := \min_{i=1}^{k_0} p^0_i$.

Denote by $\mathcal{A}(G)$ the set of atoms corresponding to any mixing measure G. For a given G and ω, let $g_{\omega,G}$ be the set of all discrete measures which collect the atoms from G such that all their atoms spaced apart by a distance at least ω:

$$g_{\omega,G} := \{G' = \sum_j p'_j \delta_{\theta'_j} : \theta'_j \in \mathcal{A}(G), \min_{u \neq v} \|\theta'_u - \theta'_v\| \geq \omega\}.$$

Note in this definition that any $G' \in g_{\omega,G}$ must have finite number of atoms, because Θ is compact.

The first merge step in the MTM algorithm is motivated by the following result, which establishes the existence of a probabilistic procedure that transform G into another measure $G' \in g_{\omega,G}$ that possesses some useful properties, namely, the supporting atoms of G' are well-separated from one another, while G' remains sufficiently close to G_0 in the sense of a Wasserstein metric.

Lemma 2.6.1. Assume that $W_r(G, G_0) \leq \delta\omega$ for some $r \geq 1$ where ω, δ satisfy condition (B.1) and (B.2). Then, there exists a probabilistic scheme which transform G into a

$G' = \sum_{j=1}^{k} p'_j \delta_{\theta'_j}$ such that $k \geq k_0$ and the following holds:

$$P(\{G' : G' \text{ satisfies (G.1) and (G.2)}\}|G) \geq 1 - \delta^{r/2} \sum_{i=1}^{k_0} \frac{1}{p_i^0}.$$

Here, P is the probability measure associated with the probabilistic scheme and the conditions (G.1) and (G.2) stand for

(G.1) $G' \in g_{\omega,G}$ and $W_r(G', G_0) \leq (k_0 + 2)\sqrt{\delta}\omega$.

(G.2) For each $i = 1, \ldots, k_0$ there is an index j for an atom of G' for which $|p_j - p_i^0| \leq \delta^{r/2}$ and $\|\theta'_j - \theta_i^0\| \leq \sqrt{\delta}\omega$.

Proof. The probabilistic scheme is the first merge step described in the MTM algorithm. We recall it in the following

1. Reorder the indices of components $\{\theta_1, \ldots, \theta_{|G|}\}$ by simple random sampling without replacement (SRSWOR) with corresponding weights $\{p_1, \ldots, p_{|G|}\}$.

2. Let $\tau_1, \ldots, \tau_{|G|}$ denote the new indices, and set $\mathcal{E} = \{\tau_j\}_j$ as the existing set of atoms.

3. Sequentially for each index τ_j, if there exists an index $\tau_i < \tau_j$ such that $\|\theta_{\tau_i} - \theta_{\tau_j}\| \leq \omega$, we perform the following updates

    - update $p_{\tau_i} = p_{\tau_i} + p_{\tau_j}$.
    - update $\mathcal{E}$ by removing index τ_j from $\mathcal{E}$.

4. Set $G' = \sum_{j:\ \tau_j \in \mathcal{E}} p_{\tau_j} \delta_{\theta_{\tau_j}}$.

The proof consists of two main steps. First, we shows that every atom of G_0 lies in a $\sqrt{\delta}\omega$ neighborhood of a unique atom of $G' = \sum_{i=1}^{k} p'_i \delta_{\theta'_i}$ having large mass, with high probability. This will allows us to deduce that G' satisfies (G.2) with a high probability. Next, we shall show that

$$\{G' : G' \text{ satisfies (G.2) } | G\} \subset \{G' : G' \text{ satisfies (G.1) } | G\}. \tag{2.20}$$

to conclude the lemma. Note that by the nature of construction it automatically holds that $G' \in g_{\omega,G}$.

Step 1: Let $P(B|G)$ be the probability of an event B under the SRSWOR scheme used above, conditioned the mixing measure G. Furthermore, let $G(A)$ denote the mass assigned to the set $A \subset \Theta$ by measure G. Thus, for a given $\epsilon > 0$

$$G(\mathbb{B}(\theta, \epsilon\omega)) = \sum_{i:\|\theta_i - \theta\| \leq \epsilon\omega} p_i$$

for any $\theta \in \Theta$, where $\mathbb{B}(\theta, \epsilon\omega)$ is an $\|\cdot\|$-ball of radius $\epsilon\omega$ centers at θ. Now, for calculating the Wasserstein distance, the amount of mass transfer between θ_i^0 and those atoms of G residing in $\mathbb{B}(\theta_i^0, \epsilon\omega)^{\mathsf{c}}$ is at least $|p_i^0 - G(\mathbb{B}(\theta_i^0, \epsilon\omega))|$. Therefore, as $W_r(G, G_0) \leq \delta\omega$,

$$|p_i^0 - G(\mathbb{B}(\theta_i^0, \epsilon\omega))|^{1/r}\epsilon\omega \leq \delta\omega$$

for any index $i \in \{1, \ldots, k_0\}$ and for any $2 \geq \epsilon > 0$. The upper bound of 2 arises from the consideration of selecting disjoint balls combined with the fact that $\omega < \min_{u \neq v} \frac{\|\theta_u^0 - \theta_v^0\|}{4}$.

It leads to the following inequalities

$$p_i^0 - \left(\frac{\delta}{\epsilon}\right)^r \leq G(\mathbb{B}(\theta_i^0, \epsilon\omega)) \leq p_i^0 + \left(\frac{\delta}{\epsilon}\right)^r. \tag{2.21}$$

Since $\omega < \min_{u \neq v} \frac{\|\theta_u^0 - \theta_v^0\|}{4}$, based on the standard union bound, the following inequality holds

$$G\left(\cup_{i=1}^{k_0} \mathbb{B}(\theta_i^0, \sqrt{\delta}\omega)\right) = \sum_{i=1}^{k_0} G(\mathbb{B}(\theta_i^0, \sqrt{\delta}\omega)) > 1 - k_0(\sqrt{\delta})^r > 0. \tag{2.22}$$

The last inequality in the above display holds because $1 \geq p_{\min}^0 > 2k_0\sqrt{\delta}$. Now, combining Equations (2.22) with (2.21), for specific choice of $\epsilon = \sqrt{\delta}$, we get that

$$\frac{G(\mathbb{B}(\theta_i^0, \sqrt{\delta}\omega))}{G(\mathbb{B}(\theta_i^0, \sqrt{\delta}\omega) \cup (\cup_{i=1}^{k_0} \mathbb{B}(\theta_i^0, \sqrt{\delta}\omega))^c)} \geq \frac{p_i^0 - (\sqrt{\delta})^r}{p_i^0 + (k_0+1)(\sqrt{\delta})^r} \geq 1 - \frac{\delta^{r/2}}{p_i^0} > 0.$$

Divide Θ into disjoint subsets $\Theta = A_1 \cup \ldots \cup A_{k_0+1}$, where $A_i = \mathbb{B}(\theta_i^0, \sqrt{\delta}\omega)$ for all $i \leq k_0$, and $A_{k_0+1} = (\cup_{i=1}^{k_0} \mathbb{B}(\theta_i^0, \sqrt{\delta}\omega))^c$. For each $i = 1, \ldots, k_0$, let E^i denote the set of G' obtained from G with one atom residing in A_i. The probabilistic scheme for the selection of atoms for G' from the those of G (via random sampling without replacement) will pick an atom from A_i and gives it a lower index than one from A_{k_0+1} with probability

$$P(E^i|G) = \frac{G(A_i)}{G(A_i) + G(A_{k_0+1})} \geq 1 - \frac{\delta^{r/2}}{p_i^0}.$$

Moreover, if $G' \in E^i$ and $\theta_j' \in \mathcal{A}(G')$ such that $\|\theta_j' - \theta_i^0\| \leq \sqrt{\delta}\omega$, then

$$p_i^0 + \delta^{r/2} \geq p_i^0 + \left(\frac{\delta}{2}\right)^r \geq G(\mathbb{B}(\theta_i^0, 2\omega)) \geq p_j' \geq G(\mathbb{B}(\theta_i^0, \sqrt{\delta}\omega)) \geq p_i^0 - \delta^{r/2}.$$

Thus $G' \in E^i$ satisfies (G.2).

This entails that $P(\{G' : G' \text{ satisfies (G.2)}\}|G) \geq P(\cap_{i=1}^{i_0} E^i|G) \geq 1 - \left(\sum_{i=1}^{k_0} \frac{\delta^{r/2}}{p_i^0} \right)$, which concludes the first proof step.

An useful fact to be used later is that if $\theta_i^0 \in \mathbb{B}(\theta_j', \sqrt{\delta}\omega)$ for some $j \leq k$, when $\theta_j' \in \mathcal{A}(G')$, $G' \in E^i$, then $G(\mathbb{B}(\theta_j, \omega)) \geq \omega^r$. Indeed, suppose that this claim does not hold, then by the definition of Wasserstein metric, we find that

$$|p_i^0 - \omega^r|^{1/r} \sqrt{\delta}\omega \leq W_r(G, G_0) \leq \delta\omega,$$

which is a contradiction as we have $p_{\min}^0 \geq 2k_0\sqrt{\delta}$ and $\omega^r < p_i^0/2$.

Step 2: To establish (2.20) it suffices to assume that $G' = \sum_j p_j' \delta_{\theta_j'} \in E^i$ satisfies that for every $i = 1, \ldots, k_0$, $\|\theta_i' - \theta_i^0\| \leq \sqrt{\delta}\omega$, and $|p_i^0 - G(\mathbb{B}(\theta_i^0, \sqrt{\delta}\omega))| \leq \delta^{r/2}$. Then, we have

$$p_i^0 - \delta^{r/2} \leq G(\mathbb{B}(\theta_i^0, \sqrt{\delta}\omega)) \leq G'(\mathbb{B}(\theta_i', \omega)) = p_i'.$$

The above result leads to $G'((\cup_{i=1}^{k_0} \mathbb{B}(\theta_i', \omega))^c) \leq k_0 \delta^{r/2}$.

Now, we construct an measure $\tilde{G} = \sum_l \tilde{p}_l \delta_{\phi_l}$ from G' and G, by "de-merging" all atoms of G' except for its first k_0 atoms. Specifically, for indices $l \leq k_0$, simply take $\tilde{p}_l = p_l'$ and $\phi_l = \theta_l'$. Additionally, if index $l > k_0$ is such that $\|\theta_l - \theta_i'\| > \omega$ for all $i \leq k_0$, then $\phi_l = \theta_l$ and $\tilde{p}_l = p_l$. Otherwise, let $\tilde{p}_l = 0$. By the triangle inequality with Wasserstein metric,

$$W_r(G', G_0) \leq W_r(G', \tilde{G}) + W_r(\tilde{G}, G_0) \leq k_0^{1/r} \sqrt{\delta}\omega + W_r(\tilde{G}, G_0). \tag{2.23}$$

The second inequality above holds because $\sum_{l=1}^{k_0} \tilde{p}_l = \sum_{j=1}^{k_0} p'_j \geq 1 - k_0\delta^{r/2}$. So, there exists a coupling of G' and $\tilde{G}$ such that any mass transfer occurs between atoms located at most ω in distance from each other. Moreover, the coupling can be so obtained that the total mass travelling a non-zero distance is bounded above by $k_0\delta^{r/2}$.

It remains to obtain a suitable upper bound for $W_r^r(\tilde{G}, G_0)$. From the definition of Wasserstein metric, we can write

$$W_r(\tilde{G}, G_0) = \inf_{\boldsymbol{q} \in \mathcal{Q}(\tilde{\boldsymbol{p}}, \boldsymbol{p^0})} \left(\sum_{i,l} q_{il} \|\phi_l - \theta_i^0\|^r \right)^{1/r},$$

where $\mathcal{Q}(\tilde{\boldsymbol{p}}, \boldsymbol{p^0})$ is the set of all possible couplings between $\tilde{\boldsymbol{p}} = (\tilde{p}_1, \ldots, \tilde{p}_{|G|})$ and $\boldsymbol{p^0} = (p_1^0, \ldots, p_{k_0}^0)$. Now, we consider a coupling $\boldsymbol{q} \in \mathcal{Q}(\tilde{\boldsymbol{p}}, \boldsymbol{p^0})$ such that $q_{ii} = \min\{p_i^0, \tilde{p}_i\}$ for any i. Then, the following inequalities hold

$$\begin{aligned} W_r^r(\tilde{G}, G_0) &\leq \sum_{i=1}^{k_0} \tilde{p}_i \|\phi_i - \theta_i^0\|^r + \sum_{i,l \neq i} q_{il} \|\phi_l - \theta_i^0\|^r \\ &\leq (\sqrt{\delta}\omega)^r + W_r^r(G, G_0) \leq (\sqrt{\delta}\omega)^r + (\delta\omega)^r = (1 + \delta^{r/2})\delta^{r/2}\omega^r. \end{aligned}$$

Therefore, following Equation (2.23), we have,

$$W_r(G', G_0) \leq (k_0^{1/r} + (1 + \delta^{r/2})^{1/r})\sqrt{\delta}\omega \leq (k_0 + 2)\sqrt{\delta}\omega.$$

As a consequence, we achieve the conclusion of the lemma. □

2.6.2.2 Truncate-merge scheme

In the previous subsection we studied properties of the first stage of the MTM algorithm, which is applied an arbitrary discrete measure G that is sufficiently close to G_0 under Wasserstein metric, namely, $W_r(G, G_0) \le \delta\omega$ for some small quantities $\delta > 0$ and $\omega > 0$. The next stage of the MTM algorithm comprises of lines 4 to 7 in the algorithm's description. It is applied to a measure G', which is the outcome of the algorithm's first stage. Denote $G' = \sum_{j=1}^{k} p'_j \delta_{\theta'_j}$ where $k \ge k_0$. As a consequence of Lemma 2.6.1, G' satisfies two important properties (G.1) and (G.2), which are to be restated here for the reader's convenience.

(G.1) $G' \in g_{\omega,G}$ and $W_r(G', G_0) \le \delta'\omega$, where $\delta' = (k_0 + 2)\sqrt{\delta}$.

(G.2) For each $i \le k_0$, there exists an (unique) atom of G', which is relabeled θ'_i so that $|\theta^0_i - \theta'_i\| \le \sqrt{\delta}\omega \le \omega p^0_{\min}/(2k_0) \le \omega/(2k_0)$.

By definition, $G' \in g_{\omega,G}$ implies that its atoms are well-separated, namely, for any $1 \le i < j \le k$, $\|\theta'_i - \theta'_j\| \ge \omega$. Assuming slightly stronger conditions on the two quantities ω and δ, we can say more about the structure of G', which turns out to be very useful in identifying the true number of atoms k_0 of G_0 via a truncation procedure.

(B.3) $\omega < \frac{7p^0_{\min} \min_{u \ne v} \|\theta^0_u - \theta^0_v\|}{16}$.

(B.4) $\sqrt{\delta} < p^0_{\min}/(2k_0(k_0+2))$, where $p^0_{\min} := \min_{i=1}^{k_0} p^0_i$.

Lemma 2.6.2. Suppose that ω and δ satisfy conditions (B.1), (B.3) and (B.4). Then for any G' satisfying properties (G.1) and (G.2), the following hold.

(a) For each $1 \le i \ne j \le k_0$, we obtain that $(p'_j)^{1/r}\|\theta'_i - \theta'_j\| > \omega$.

(b) For each $j > k_0$, we find that $\min_{1 \le i \le k_0} (p'_j)^{1/r}\|\theta'_i - \theta'_j\| \le \omega$.

Proof. To show (a), note for any $i, j \le k_0$

$$\|\theta'_i - \theta'_j\| \ge \|\theta^0_i - \theta^0_j\| - \|\theta'_i - \theta^0_i\| - \|\theta'_j - \theta^0_j\| \ge \|\theta^0_i - \theta^0_j\| - \frac{\omega}{k_0} \ge \frac{7}{8}\|\theta^0_i - \theta^0_j\| \quad (2.24)$$

where the first inequality follows from triangle inequality, the second inequality is due to the hypothesis with G', and the third inequality is due to (B.1).

By the definition of Wasserstein distances, for mass transport to be achieved between G' and G_0, an amount of mass at least $|p_i^0 - p_i'|$ should be transported from atom θ_i^0 of G_0 to an atom of G' other than θ_i'. Hence, for any $i \leq k_0$, $|p_i^0 - p_i'|(\omega - \|\theta_i' - \theta_i^0\|)^r \leq W_r^r(G', G_0) \leq (\delta'\omega)^r$. Invoking the hypothesis with G', these inequalities lead to $|p_i^0 - p_i'|^{1/r} \leq 2\delta'$. Combining with the condition $\sqrt{\delta} < \frac{p_i^0}{2k_0(k_0+2)}$, the above inequality leads to $p_i' > p_i^0 - \frac{(p_i^0)^2}{2k_0^2} \geq \frac{p_i^0}{2}$. Combining this with Equation (2.24) and (B.3) to conclude.

Turning to part (b), suppose for some $j > k_0$, we have $(p_j')^{1/r}\|\theta_i' - \theta_j'\| > \omega$ for all $i \leq k_0$. Then by triangle inequality and the properties of G', we find that

$$\|\theta_j' - \theta_i^0\| + \frac{\omega}{2k_0} \geq \|\theta_j' - \theta_i^0\| + \|\theta_i^0 - \theta_i'\| \geq \|\theta_i' - \theta_j'\| > \omega/{p_j'}^{1/r}.$$

Applying the triangle inequality again, $\|\theta_i^0 - \theta_j'\| \geq \|\theta_j' - \theta_i'\| - \|\theta_i^0 - \theta_i'\| \geq \omega(1 - \frac{1}{2k_0}) \geq \frac{\omega}{2}$. Combining the two preceeding bounds, we get $2\|\theta_j' - \theta_i^0\| > \omega/{p_j'}^{1/r}$ for all $i \leq k_0$.

Now, since $W_r(G', G_0) \leq \delta'\omega$, we can find a coupling $\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p'}, \boldsymbol{p}^0)$ between $\boldsymbol{p'} = (p_1', \ldots, p_k')$ and $\boldsymbol{p^0} = (p_1, \ldots, p_{k_0}^0)$ such that $\sum_{ij} q_{ij}\|\theta_j' - \theta_i^0\|^r \leq (\delta'\omega)^r$. However, based on the previous inequalities, we have for the given index j

$$\sum_{i=1}^{k_0} q_{ij}\|\theta_j' - \theta_i^0\|^r > \sum_{i=1}^{k_0} q_{ij}\left(\frac{1}{2}\right)^r \omega^r/p_j' = \left(\frac{1}{2}\right)^r \omega^r,$$

which is a contradiction as $\delta' < 1/2$ due to condition (B.4). This concludes the proof of the lemma. □

2.6.2.3 Proof of Theorem 2.3.2

Now we are ready for the proof of this theorem. It suffices to prove for the case constant $c = 1$.

Proof of part (a): Recall that we are given a (random) mixing measure for which the following holds: for each fixed $\epsilon > 0$ and $\delta > 0$, as $n \to \infty$, there holds

$$P^n_{G_0}\left\{\Pi\left(W_r(G, G_0) \geq \delta\omega_n \big| (X_1, \ldots, X_n)\right) \geq \epsilon\right\} \to 0.$$

Choose δ sufficiently small, and as n gets large ω_n also becomes so small that all conditions (B.1–4) in the preceeding sections are satisfied. Then, we can appeal to Lemma 2.6.1 and the above mentioned Equation to obtain that, measure G' as produced in the probablistic merge stage of the MTM algorithm also admits a posterior contraction toward G_0, in the sense that as $n \to \infty$

$$P^n_{G_0}\left\{\Pi\left(W_r(G', G_0) \leq (k_0 + 2)\sqrt{\delta}\omega_n \big| (X_1, \ldots, X_n)\right) \geq (1 - \epsilon)\left(1 - \sum_{i=1}^{k_0} \frac{\delta^{\frac{r}{2}}}{p^0_i}\right)\right\} \to 1.$$

Since this holds for any $\delta > 0$, we deduce that the posterior probability $\Pi\left(W_r(G', G_0) \leq (k_0 + 2)\sqrt{\delta}\omega_n \big| (X_1, \ldots, X_n)\right) \to 1$ in P_{G_0} probability.

Suppose that G' satisfies both conditions (G.1) and (G.2) (per Lemma 2.6.1), then it can be verified that if the atoms of G' are arranged in descending order of their masses, then each of the top k_0 atoms of G' lie in an $\frac{\omega_n}{2k_0}$- ball around an atom of G_0. Specifically, using the representation $G' = \sum_{i=1}^{k_n} q_i \delta_{\phi_i}$, where $q_1 \geq \cdots \geq q_{k_n}$, we have that $\|\theta^0_i - \phi_i\| \leq \frac{\omega_n}{2k_0}$ and $|q_i - p^0_i| \geq \delta^{r/2}$ for all $i \in \{1, \ldots, k_0\}$.

Recall that G' is fed into the second stage, the truncate-merge procedure, of the

MTM algorithm. Note that $|q_i - p_i^0| \leq \delta^{r/2}$ implies $q_i > p_i^0 - \delta^{r/2} > p_i^0/2 > \omega_n^r$ for n sufficiently large. By Lemma 2.6.2 that for each $j > k_0$, $\min_{i \leq k_0} (q_j)^{1/r} \|\phi_i - \phi_j\| \leq \omega_n$, but for each $i, j \leq k_0, i \neq j$, $(q_j)^{1/r} \|\phi_i - \phi_j\| \geq \omega_n$. Following the definition of $\tilde{k} = |\mathcal{A}|$, we deduce that $\tilde{k} = k_0$. The final step is to coat this guarantee with a probability statement, due to the fact that G' is random given G,

$$P_{G_0}^n \Bigg\{ \Pi\bigg(\tilde{k} = k_0 | X_1, \ldots, X_n\bigg) \ \ \geq \ \ (1-\epsilon)\left(1 - \sum_{i=1}^{k_0} \frac{\delta^{r/2}}{p_i^0}\right)\Bigg\} \longrightarrow 1 \qquad (2.25)$$

as $n \to \infty$. Let $\delta \to 0$ to conclude the proof of part (a).

Proof of part (b): The proof boils down to showing that the reassignment of mass as in the second stage of the MTM Algorithm only increases the Wasserstein distance by a constant factor. Denote by $\boldsymbol{p^0} = (p_1^0, \ldots, p_{k_0}^0)$ and $\boldsymbol{q} = (q_1, \ldots, q_{k_n})$ the weight vectors of G_0 and G' respectively. Suppose that $W_r(G', G_0) \leq (k_0 + 2)\sqrt{\delta}\omega_n$ as before. So can find a coupling $\boldsymbol{f} \in \mathcal{Q}(\boldsymbol{p^0}, \boldsymbol{q})$ such that

$$\left(\sum_{i,j} f_{ij} \|\theta_i^0 - \phi_j\|^r\right)^{1/r} \leq 2(k_0+2)\sqrt{\delta}\omega_n.$$

Define the set $V_{i,n} := \{\theta \in \Theta : \|\theta - \theta_i^0\| \leq \min_{u \neq v} \frac{\|\theta_u^0 - \theta_v^0\|}{2} - \frac{\omega_n}{2k_0}\}$. Furthermore, the following inequalities hold

$$\|\phi_j - \theta_i^0\| \geq \min_{u \neq v} \frac{\|\theta_u^0 - \theta_v^0\|}{2} - \frac{\omega_n}{2k_0} \geq \min_{u \neq v} \frac{\|\theta_u^0 - \theta_v^0\|}{4}$$

for all $i \leq k_0$ and $j \notin V_{i,n}$, because ω_n satisfies assumption (B.1). Therefore, we find that

$$\sum_{i,j:\phi_j \notin V_{i,n}} f_{ij} \leq 4 \left(\frac{2(k_0+2)\sqrt{\delta}\omega_n}{\min_{u \neq v} \|\theta_u^0 - \theta_v^0\|} \right)^r .$$

Notice that if $j \in V_{i,n}$, the second stage of the MTM Algorithm assigns the mass corresponding to atom j of G' to atom i of $\widetilde{G}$. We can assume henceforth that $\widetilde{G}$ is such that $|\mathcal{A}(\widetilde{G})| = k_0$ as a result of the proof of part (a) of this theorem.

Since the sets $V_{i,n}$ are disjoint, this assignment is unique. It follows that we can find $\boldsymbol{f'} \in \mathcal{Q}(\boldsymbol{p_0}, \boldsymbol{q'})$ such that

$$\begin{aligned} \sum_{i,j \neq i} f'_{ij} \|\theta_i^0 - \phi_j\|^r &\leq 4^r \left(\frac{2\text{Diam}(\Theta)(k_0+2)\sqrt{\delta}\omega_n}{\min_{u \neq v} \|\theta_u^0 - \theta_v^0\|} \right)^r , \\ \sum_i f'_{ii} \|\theta_i^0 - \phi_i\|^r &\leq ((k_0+2)\sqrt{\delta}\omega_n)^r \end{aligned} \tag{2.26}$$

where $\boldsymbol{q'} = (q'_1, \ldots, q'_{k_0})$ is the weight vector of $\widetilde{G}$. The first inequality above follows, since $\|\phi_i - \theta_i^0\| \leq \|\phi_j - \theta_i^0\|$ for all pairs i, j with $i \leq k_0$, with strict inequality for $i \neq j$. To obtain the conclusion for the second inequality in (2.26), we note that $p_i^0 = \sum_j f_{ij} = \sum_j f'_{ij}$. Therefore, $f'_{ii} = \sum_j f_{ij} - \sum_{j \neq i} f'_{ij}$. Then, if ϕ_j is an atom of G', for any $j \neq i$, we have $\|\theta_i^0 - \phi_i\| \leq \|\theta_i^0 - \phi_j\|$. Hence, we find that

$$\sum_i f'_{ii} \|\theta_i^0 - \theta_i\|^r \leq \sum_{i,j} f_{ij} \|\theta_i^0 - \phi_j\|^r \leq W_r^r(G', G_0) \leq ((k_0+2)\sqrt{\delta}\omega_n)^r.$$

By the nature of construction $\mathcal{A}(\widetilde{G}) \subset \mathcal{A}(G')$. Using the two parts of Equation (2.26), we obtain that

$$W_r(\widetilde{G}, G_0) \leq \left(1 + 4^r \left(\frac{2\text{Diam}(\Theta)}{\min_{i,l} \|\theta_i^0 - \theta_l^0\|} \right)^r \right)^{1/r} ((k_0+2)\sqrt{\delta}\omega_n).$$

The full probability statement is

$$P^n_{G_0}\biggl\{\Pi\biggl(G \in \overline{\mathcal{G}}(\Theta) : W_r(\widetilde{G}, G_0) \leq C\delta\omega_n \Big| X_{1:n}\biggr) \geq (1-\epsilon)\biggl(1 - \sum_{i=1}^{k_0} \frac{\delta^{\frac{r}{2}}}{p_i^0}\biggr)\biggr\} \to 1, \quad (2.27)$$

where $C = \left(1 + 4^r \left(\frac{2\text{Diam}(\Theta)}{\min_{i,l} \|\theta_i^0 - \theta_l^0\|}\right)^r\right)^{1/r} (k_0 + 2)$ is a constant dependent on G_0 and Θ. Finally, letting $\delta \to 0$ we obtain the desired conclusion for part (b).

2.6.3 Proof of Lemma 2.4.3

To simplify the proof argument, we specifically assume that G_* is a discrete mixture. The proof argument for other settings of G_* is similar and is omitted. Now, we consider an $\epsilon > 0$ maximal packing set of parameter space Θ. It leads to a $D-$partition $(S_1, \ldots, S_D)$ of Θ such that $\text{Diam}(S_i) \leq 2\epsilon$ for all $1 \leq i \leq D$. Choose ϵ to be sufficiently small such that $D > \gamma$.

For mixing measures $G = \sum_i p_i \delta_{\theta_i}$ and $G_* = \sum_{i=1}^{\infty} p_i^* \delta_{\theta_i^*}$, we denote $G(S_i) := \sum_{i:\theta_i \in S_i} p_i$ and $G_*(S_i) = \sum_{i:\theta_i^* \in S_i} p_i^*$. Invoking the detailed formulation of Wasserstein metric, we can check that

$$W_r^r(G, G_*) \leq (2\epsilon)^r + \text{Diam}^r(\Theta) \sum_{i=1}^{D} |G(S_i) - G_*(S_i)|.$$

Equipped with the above inequality, the following inequality holds

$$\Pi(W_r^r(G, G_*) \leq (2^r + 1)\epsilon^r) \geq \Pi\left(\sum_{i=1}^{D} |G(S_i) - G_*(S_i)| \leq (\epsilon/\text{Diam}(\Theta))^r\right).$$

For any positive constant A, we find that

$$\begin{aligned}
&\Pi\left(\sum_{i=1}^{D} |G(S_i) - G_*(S_i)| \leq A\right) \\
&\geq q_D \Pi(B \cap \{|G(S_i) - G_*(S_i)| \leq A/D, \text{for each } i\}|K = D)
\end{aligned}$$

where B stands for the event that each S_i contains exactly one atom of G.

Governed by the above observations, by substituting $A = (\epsilon/\text{Diam}(\Theta))^r$, we obtain that

$$\begin{aligned}
&\Pi(W_r^r(G, G_*) \leq (2^r + 1)\epsilon^r) \\
&\gtrsim q_D \left(c_0 \left(\frac{\epsilon}{\text{Diam}(\Theta)}\right)^d\right)^D \underbrace{\Pi(\{|G(S_i) - G_*(S_i)| \leq A/D, \text{for each } i\}|B \cap \{K = D\})}_{:=T}.
\end{aligned}$$

By means of Dirichlet probability model assumption on Δ_{D-1}, the standard simplex of dimension D, we have the following evaluations with T

$$\begin{aligned}
T &\gtrsim D! \frac{\Gamma(\gamma)}{\prod_{i=1}^{D} \Gamma(\gamma/D)} \int_{\mathcal{U}} \prod_{i=1}^{D-1} (G(S_i))^{(\gamma/D)-1} (1 - \sum_{i=1}^{D-1} G(S_i))^{(\gamma/D)-1} \text{d}(G(S_i)) \\
&\geq D! \frac{\Gamma(\gamma)}{\prod_{i=1}^{D} \Gamma(\gamma/D)} \prod_{i=1}^{D-1} \int\limits_{\max(G_*(S_i) - (\epsilon/\text{Diam}(\Theta))^r/D, 0)}^{\min(G_*(S_i) + (\epsilon/\text{Diam}(\Theta))^r/D, 1)} (G(S_i))^{(\gamma/D)-1} \text{d}(G(S_i)) \\
&\geq D! \frac{\Gamma(\gamma)\gamma/D}{\prod_{i=1}^{D} (\gamma/D)\Gamma(\gamma/D)} \left(\frac{1}{D}\left(\frac{\epsilon}{\text{Diam}(\Theta)}\right)^r\right)^{\gamma(D-1)/D}
\end{aligned}$$

where $\mathcal{U} := \Delta_{D-1} \cap |G(S_i) - G_*(S_i)| \leq (\epsilon/\text{Diam}(\Theta))^r/D$. Here, the second inequality in the above display is due to the fact that $(1 - \sum_{i=1}^{D-1} G(S_i))^{(\gamma/D)-1} > 1$ as $\gamma < D$. Invoking the basic inequality $\alpha\Gamma(\alpha) < 1$ for $0 < \alpha < 1$, we reach the conclusion of the lemma.

2.6.4 Proof of Proposition 2.4.1

We denote a sphere of radius R as $S_R := \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$ for any $R > 0$. Direct computations lead to

$$\begin{aligned}
2V(p_G, p_{G'}) &= \int\limits_{\mathbb{R}^d} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) \\
&= \int\limits_{S_R^{\mathsf{c}}} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) + \int\limits_{S_R} |p_G(x) - p_{G'}(x)| \frac{p_{G_0,f_0}(x)}{p_{G_*}(x)} \frac{p_{G_*}(x)}{p_{G_0,f_0}(x)} \mathrm{d}\mu(x) \\
&\leq \int\limits_{S_R^{\mathsf{c}}} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) \\
&\qquad\qquad + \left\| \frac{p_{G_*}(\cdot)}{p_{G_0,f_0}(\cdot)} \mathbb{1}_{S_R} \right\|_\infty \int\limits_{\mathbb{R}^d} |p_G(x) - p_{G'}(x)| \frac{p_{G_0,f_0}(x)}{p_{G_*}(x)} \mathrm{d}\mu(x)
\end{aligned} \tag{2.28}$$

where the last inequality is an application of Holder's inequality. Now, a direct evaluation yields that

$$\begin{aligned}
\int\limits_{S_R^{\mathsf{c}}} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) &\leq 2 \int\limits_{S_R^{\mathsf{c}}} \max_G p_G(x) \mathrm{d}\mu(x) \leq 2 \int\limits_{S_R^{\mathsf{c}}} \sup_\theta f(x|\theta) d\mu(x) \\
&\leq 2 \int\limits_{S_R^{\mathsf{c}}} \sup_\theta \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-\|x - \theta\|_2^2/(2\lambda_{\max})) \mathrm{d}\mu(x).
\end{aligned} \tag{2.29}$$

The last inequality is due to the fact that $(x - \theta)^\top \Sigma^{-1} (x - \theta) \geq \|x - \theta\|_2^2/\lambda_{\max}$ for all $x \in \mathbb{R}^d$ and $\theta \in \Theta$.

We now assume that $d > 2$ as the $d \leq 2$ case can be treated similarly. Since $\Theta \subset \mathbb{R}^d$ is bounded, we can find $r > 0$ such that $\|\theta\|_2 < r$ for all $\theta \in \Theta$. Now, given $R > r$, for any fixed value of $x \in S_R^c$, we can check that $\inf_\theta \|x-\theta\|_2^2 \geq (\|x\|_2 - r)^2$. Therefore equation (2.29) leads to

$$\int_{S_R^c} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) \;\; \lesssim \;\; \int_{S_R^c} \frac{1}{|2\pi\Sigma|^{1/2}} \exp(-(\|x\|_2 - r)^2/(2\lambda_{\max}))\mathrm{d}\mu(x).$$

Invoking spherical coordinates by substituting $z = \|x\|_2$, we get

$$\int_{S_R^c} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) \;\; \lesssim \;\; \int_{z>R} \frac{z^{d-1}}{|2\pi\Sigma|^{1/2}} \exp(-(z-r)^2/(2\lambda_{\max}))dz.$$

We denote $g_R(d) := \int_{z>R} z^d \exp(-(z-r)^2) dz$. By integrating by parts with some basic algebraic manipulation, we find that

$$g_R(d-1) = (d/2) g_R(d-3) + (R^{d-2}/2)\exp(-(R-r)^2) + r g_R(d-2). \qquad (2.30)$$

Observe that $(R^{d-2}/2)\exp(-(R-r)^2) \gtrsim (R^s/2)\exp(-(R-r)^2)$ for all $s \leq d-2$. Also, $\int_x^\infty \exp(-t^2/2)\mathrm{d}t \leq \frac{1}{x}\exp(-x^2/2) \lesssim x^{d-2}\exp(-x^2/2)$ and $\int_x^\infty t\exp(-t^2/2)\mathrm{d}t \lesssim \exp(-x^2/2) \lesssim x^{d-2}\exp(-x^2/2)$ follows using standard arguments for gaussian tail-bounds. Here we use the condition $d > 2$. For $d \leq 2$, the tail probability can be directly bounded using standard gaussian tailbounds.

We can expand $g_R(s)$ recursively using Equation (2.30) for all $s \leq d-2$ as well. Now, equipped with equation (2.30), and following the discussion in the previous paragraph, we can write

$$\int_{S_R^c} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) \lesssim R^{d-2} \exp(-(R-r)^2/2\lambda_{\max}). \qquad (2.31)$$

Now, we demonstrate that $\left\| \frac{p_{G_*}(\cdot)}{p_{G_0,f_0}(\cdot)} \mathbb{1}_{S_R} \right\|_\infty$ is bounded above by $c_2 \exp(\lambda_{\min}^{-1} R^2)$ for some positive constant c_2 depending only on C_1, G_0 and $\lambda_{\min}$. Recall that $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$. Here k_0 can be allowed to be ∞. The analysis follows through similar to the finite k_0 case.

The conditions on p_{G_0,f_0} imply that

$$\left\| \frac{p_{G_0,f}(\cdot)}{p_{G_0,f_0}(\cdot)} \mathbb{1}_{S_R} \right\|_\infty \leq \sup_{x \in \mathbb{R}^d, \theta \in \Theta, \theta_0 \in \text{supp}(G_0)} \frac{f(x|\theta)}{f_0(x|\theta_0)} \mathbb{1}_{\|x\|_2 \leq R} \leq C_1 \exp(C_0 R^2).$$

Then, we have the following inequalities

$$\begin{aligned}
\left\| \frac{p_{G_*}(\cdot)}{p_{G_0,f_0}(\cdot)} \mathbb{1}_{S_R} \right\|_\infty &= \left\| \frac{p_{G_*}(\cdot)}{p_{G_0,f}(\cdot)} \frac{p_{G_0,f}(\cdot)}{p_{G_0,f_0}(\cdot)} \mathbb{1}_{S_R} \right\|_\infty \\
&\lesssim \sup_{x \in S_R,\ i \in \{1,\ldots,k_0\}} C_1 \exp(C_0 R^2) \exp\left(\frac{1}{2}(x-\theta_i^0)' \Sigma^{-1} (x-\theta_i^0) \right) \\
&\leq C_1 \exp(C_0 R^2) \sup_{x \in S_R} \exp(\lambda_{\min}^{-1}(\|x\|^2 + \sup_i \|\theta_i^0\|^2)) \\
&\leq c_2 \exp((\lambda_{\min}^{-1} + C_0) R^2). \qquad (2.32)
\end{aligned}$$

The bounds apply uniformly for all $R > r$. Therefore, when $R \geq 4r$, we can bound equation (2.28) according to the bounds in (2.31) and (2.32) as follows:

$$V(p_G, p_{G'}) \lesssim \exp((\lambda_{\min}^{-1} + C_0) R^2) \overline{V}(p_G, p_{G'}) + R^{d-2} \exp(-\lambda_{\max}^{-1} R^2/4)$$

where $\overline{V}(p_G, p_{G'}) := \int_{\mathbb{R}^d} |p_G(x) - p_{G'}(x)| \frac{p_{G_0,f_0}(x)}{p_{G_*}(x)} \mu(\mathrm{d}x)$ is the weighted variational distance. Now consider $R > 0$ satisfying $R^{d-2} \leq \exp(-\lambda_{\max}^{-1} R^2/8)$. If $\overline{V}(p_G, p_{G'}) = \exp(-(\lambda_{\min}^{-1} + C_0 + \lambda_{\max}^{-1}/8) R^2)$, we obtain that

$$V(p_G, p_{G'}) \lesssim \exp(-\lambda_{\max}^{-1} R^2/8) = (\overline{V}(p_G, p_{G'}))^{\frac{1}{1 + 8\lambda_{\max}(\lambda_{\min}^{-1} + C_0)}}.$$

Note that, $\overline{V}(p_G, p_{G'}) \lesssim \overline{h}(p_G, p_{G'})$ by standard application of Holder's inequality. Also, since the kernel for location Gaussian mixtures is supersmooth (*Fan* (1991)), it follows from Theorem 2 in *Nguyen* (2013) that $V(p_G, p_{G'}) \gtrsim \exp\bigg(-1/W_2^2(G, G')\bigg)$. The proof of the proposition now follows from this fact, and the inequality connecting weighted Hellinger and variational distances.

2.6.5 Proof of Proposition 2.4.2

We denote a sphere of radius R as $S_R := \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$ for any $R > 0$. Assume $\max_{\theta \in \Theta} \|\theta\|_2 \leq r$ and also $\sup_{i \leq k_0} \|\theta_i^0\|_2 \leq r$. We consider $R > 2r$ large enough such that, $\|\theta\|_2 \leq r$ and $\|x\|_2 \geq R$ implies

$$C_{\text{lower}} \frac{\exp\left(-\sqrt{\frac{2}{\lambda}}\|x-\theta\|_{\Sigma^{-1}}\right)}{(\|x-\theta\|_{\Sigma^{-1}})^{(d-1)/2}} \leq f(x|\theta) \leq C_{\text{upper}} \frac{\exp\left(-\sqrt{\frac{2}{\lambda}}\|x-\theta\|_{\Sigma^{-1}}\right)}{(\|x-\theta\|_{\Sigma^{-1}})^{(d-1)/2}}.$$

The above inequalities can always be achieved for R large enough because of the asymptotic formulation of multivariate Laplace distributions.

Following equation (2.28) in the proof of Theorem 2.4.1, we will prove the proposition by providing upper bounds for $\left\|\frac{p_{G_*}(\cdot)}{p_{G_0, f_0}(\cdot)} \mathbb{1}_{S_R}\right\|_\infty$ and $\int_{S_R^c} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x)$. Because the Laplace density is bounded, $\|p_{G_*}(\cdot)\mathbb{1}_{S_R}\|_\infty$ is bounded by a constant. Similar to the proof of Proposition 2.4.1, we have,

$$\begin{aligned}
&\left\| \frac{p_{G_*}(\cdot)}{p_{G_0,f_0}(\cdot)} \mathbb{1}_{S_R} \right\|_\infty \\
&\lesssim C_1 \exp(C_0 R^\alpha) \max_{x,\theta: \|\theta\|_2 \le r, \|x\|_2 \le R} \exp\left(\sqrt{\frac{2}{\lambda\lambda_{\min}}} \|x-\theta\|_2 \right) (\|x-\theta\|_2)^{(d-1)/2} \\
&\lesssim \exp(C_0 R^\alpha) \exp\left(\sqrt{\frac{2}{\lambda\lambda_{\min}}} (R+r) \right) (R+r)^{(d-1)/2} \\
&\lesssim \exp\left(\left(\sqrt{\frac{2}{\lambda\lambda_{\min}}} + C_0 \right) R^\alpha \right) R^{(d-1)/2}.
\end{aligned}$$

Now, in order to minimize $\int_{S_R^\mathsf{c}} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x)$, observe that

$$\begin{aligned}
\int\limits_{S_R^\mathsf{c}} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) &\lesssim \int\limits_{S_R^\mathsf{c}} \sup_G p_G(x) d\mu(x) \lesssim \int\limits_{S_R^\mathsf{c}} \sup_\theta f(x|\theta) \mathrm{d}\mu(x) \\
&\lesssim \int\limits_{S_R^\mathsf{c}} \sup_\theta \frac{\exp\left(-\sqrt{\frac{2}{\lambda}} \|x-\theta\|_{\Sigma^{-1}} \right)}{(\|x-\theta\|_{\Sigma^{-1}})^{(d-1)/2}} \\
&\lesssim \int\limits_{S_R^\mathsf{c}} \frac{1}{(\|x\|_2 - r)^{(d-1)/2}} \exp\left(-\sqrt{2\lambda_{\max}^{-1}} (\|x\|_2 - r) \right) d\mu(x) \\
&\lesssim \int\limits_{S_R^\mathsf{c}} \frac{1}{(\|x\|_2)^{(d-1)/2}} \exp\left(-\sqrt{\frac{2}{\lambda\lambda_{\max}}} \|x\|_2 \right) \mathrm{d}\mu(x).
\end{aligned}$$

Substituting $z = \|x\|_2$ in above equation, we get

$$\int\limits_{S_R^\mathsf{c}} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) \lesssim \int\limits_{S_R^\mathsf{c}} \frac{z^{d-1}}{z^{(d-1)/2}} \exp\left(-z \sqrt{\frac{2}{\lambda\lambda_{\max}}} \right) dz.$$

Denote $g_R(s) := \int_{S_R^\mathsf{c}} z^s \exp(-z) dz$. Then, we find that

$$g_R(s) = R^s \exp(-R) + s g_R(s-1).$$

Invoking integration by parts with the above equality for $s = \frac{d-1}{2}$ leads to the following inequality

$$\int\limits_{S_R^c} |p_G(x) - p_{G'}(x)| \mathrm{d}\mu(x) \lesssim R^{\frac{d-1}{2}} \exp\left(-\sqrt{\frac{2}{\lambda\lambda_{\max}}} R\right).$$

Since the above bounds apply for all R large enough, following the approach with equation (2.28) in the proof of Theorem 2.4.1, we can write

$$V(p_G, p_{G'}) \lesssim \exp\left(\left(\sqrt{\frac{2}{\lambda\lambda_{\min}}} + C_0\right) R^\alpha\right) R^{\frac{d-1}{2}} \overline{V}(p_G, p_{G'}) + R^{\frac{d-1}{2}} \exp\left(-\sqrt{\frac{2}{\lambda\lambda_{\max}}} R\right).$$

Recall that $\overline{V}(p_G, p_{G'}) := \int_{\mathbb{R}^d} |p_G(x) - p_{G'}(x)| \frac{p_{G_0}(x)}{p_{G_*}(x)} \mu(\mathrm{d}x)$ is the weighted variational distance. By setting $\overline{V}(p_G, p_{G'}) = \exp\left(-\left[\sqrt{\frac{2}{\lambda\lambda_{\min}}} + \sqrt{\frac{2}{\lambda\lambda_{\max}}} + C_0\right] R^\alpha\right)$, as $\alpha \geq 1$, we see that

$$V(p_G, p_{G'}) \lesssim \left(\log \frac{1}{\overline{V}(p_G, p_{G'})}\right)^{d/2\alpha} \exp\left(-\tau(\alpha)\left(\log \frac{1}{\overline{V}(p_G, p_{G'})}\right)^{1/\alpha}\right),$$

where $\tau(\alpha) = \sqrt{\frac{2}{\lambda\lambda_{\max}}} \Big/ \left[\sqrt{\frac{2}{\lambda\lambda_{\max}}} + \sqrt{\frac{2}{\lambda\lambda_{\min}}} + C_0\right]^{1/\alpha}$. Now, the location family of multivariate Laplace distributions pertains to the ordinary smooth likelihood families. Therefore, from part (1) of Theorem 2 in *Nguyen* (2013), it follows that for any $m < 4/(4+5d)$, $W_2^2(G, G') \leq V(p_G, p_{G'})^m$. We note in passing that improved rates for other choices of W_r may be possible by utilizing techniques similar to *Gao and van der Vaart* (2016). Thus, by means of the inequality $\overline{V}(p_G, p_{G'}) \lesssim \overline{h}(p_G, p_{G'})$, the following inequality holds

$$\left(\log \frac{1}{\overline{h}(p_G, p_{G'})}\right)^{\frac{d}{2\alpha}} \exp\left(-\tau(\alpha)\left(\log \frac{1}{\overline{h}(p_G, p_{G'})}\right)^{1/\alpha}\right) \gtrsim W_2^{2/m}(G, G').$$

The result now follows by taking logarithms of both sides.

2.7 Appendix B: Weighted Hellinger and Wasserstein distance

In this appendix, we will establish several useful bounds between weighted Hellinger distance and Wasserstein metric that are employed in the proofs for misspecified settings of Section 2.4. See the formal setup of G_0, f_0 and G_* in the beginning of that section.

First, we start with the following lemma regarding an upper bound of weighted Hellinger distance in terms of Wasserstein metric when the kernel f satisfies first order integral Lipschitz condition.

Lemma 2.7.1. Assume that the kernel f is integral Lipschitz up to the first order. Then, for any mixing measure G_1 and G_2 in $\mathcal{P}(\Theta)$, there exists a positive constant $\overline{C}(\Theta)$ depending only on Θ such that

$$\overline{h}^2(p_{G_1}, p_{G_2}) \leq \overline{C}(\Theta) W_1(G_1, G_2).$$

Proof. Denote the weighted total variation distance as follows

$$\overline{V}(p_{G_1}, p_{G_2}) = \frac{1}{2} \int |p_{G_1}(x) - p_{G_2}(x)| \frac{p_{G_0,f_0}(x)}{p_{G_*}(x)} \mathrm{d}\mu(x)$$

for any $G_1, G_2 \in \mathcal{P}(\Theta)$. It is clear that $\overline{h}^2(p_{G_1}, p_{G_2}) \leq \overline{V}(p_{G_1}, p_{G_2})$ for any $G_1, G_2 \in \mathcal{P}(\Theta)$.

For any coupling $\boldsymbol{q}$ of the weight vectors of $G_1 = \sum\limits_{i=1}^{k_1} p_{i,1}\delta_{\theta_{i,1}}$ and $G_2 = \sum\limits_{i=1}^{k_2} p_{i,2}\delta_{\theta_{i,2}}$, we can check via triangle inequality that

$$\begin{aligned} \overline{V}(p_{G_1}, p_{G_2}) & \leq \frac{1}{2} \int \sum_{i,j} q_{ij} |f(x|\theta_{i,1}) - f(x|\theta_{j,2})| \frac{p_{G_0,f_0}(x)}{p_{G_*}(x)} \mathrm{d}\mu(x) \\ & \leq \overline{C}(\Theta) \sum_{i,j} q_{ij} \|\theta_{i,1} - \theta_{j,2}\| \end{aligned}$$

where the existence of positive constant $\overline{C}(\Theta)$ in the second inequality is due to the first order integral Lipschitz property of f. The above result implies that

$$\overline{h}^2(p_{G_1}, p_{G_2}) \leq \overline{V}(p_{G_1}, p_{G_2}) \leq \overline{C}(\Theta) W_1(G_1, G_2)$$

for any $G_1, G_2 \in \mathcal{P}(\Theta)$. We achieve the conclusion of the lemma. $\square$

2.7.1 Proof of Lemma 2.4.2

The proof is a straightforward application of Lemma 2.4.1. In fact, from that lemma, we have

$$2 \leq \int \biggr(\frac{p_{G_{1,*}}(x)}{p_{G_{2,*}}(x)} + \frac{p_{G_{2,*}}(x)}{p_{G_{1,*}}(x)}\biggr) p_{G_0,f_0}(x) \mathrm{d}\mu(x) \leq 2$$

where the first inequality is due to Cauchy inequality. The above inequality holds only when $p_{G_{1,*}}(x) = p_{G_{2,*}}(x)$ for almost all $x \in \mathcal{X}$, which concludes our lemma.

2.7.2 Proof of Proposition 2.4.3

Denote the weighted total variation distance as follows

$$\overline{V}(p_{G_1}, p_{G_2}) = \frac{1}{2} \int |p_{G_1}(x) - p_{G_2}(x)| \frac{p_{G_0,f_0}(x)}{p_{G_*}(x)} \mathrm{d}\mu(x)$$

for any $G_1, G_2 \in \mathcal{G}$. Then, by means of Holder's inequality, we can verify that

$$\begin{aligned} \overline{V}(p_{G_1}, p_{G_2}) &\leq \sqrt{2}\overline{h}(p_{G_1}, p_{G_2}) \biggr(\int (\sqrt{p_{G_1}(x)} + \sqrt{p_{G_2}(x)})^2 \frac{p_{G_0,f_0}(x)}{p_{G_*}(x)} \mathrm{d}\mu(x) \biggr)^{1/2} \\ &\leq 2\sqrt{2}\overline{h}(p_{G_1}, p_{G_2}) \end{aligned}$$

where the last inequality is due to Lemma 2.4.1. Therefore, to obtain the conclusion of the proposition, it is sufficient to demonstrate that

$$\inf_{G \in \mathcal{O}_{\overline{k}}} \overline{V}(p_G, p_{G_*})/W_2^2(G, G_*) > 0 \tag{2.33}$$

where $\overline{k} > k_*$. Firstly, we will show that

$$\lim_{\epsilon \to 0} \inf_{G_* \in \mathcal{O}_{\overline{k}}} \left\{ \frac{\overline{V}(p_G, p_{G_*})}{W_2^2(G, G_*)} : W_2(G, G_*) \leq \epsilon \right\} > 0.$$

Assume that the above inequality does not hold. It implies that there exists a sequence of $G_n \in \mathcal{O}_{\overline{k}}(\Theta)$ such that $\overline{V}(p_{G_n}, p_{G_*})/W_2^2(G_n, G_*) \to 0$ as $n \to \infty$. By means of Fatou's lemma, we have

$$0 \;=\; \liminf_{n \to \infty} \frac{\overline{V}(p_{G_n}, p_{G_*})}{W_2^2(G_n, G_*)} \geq \frac{1}{2} \int \liminf_{n \to \infty} \frac{|p_{G_n}(x) - p_{G_*}(x)| \dfrac{p_{G_0, f_0}(x)}{p_{G_*}(x)}}{W_2^2(G_n, G_*)} \mathrm{d}\mu(x).$$

Hence, for almost every $x \in \mathcal{X}$, we obtain that

$$\liminf_{n \to \infty} \frac{|p_{G_n}(x) - p_{G_*}(x)| \dfrac{p_{G_0, f_0}(x)}{p_{G_*}(x)}}{W_2^2(G_n, G_*)} = 0.$$

The above equality is equivalent to

$$\liminf_{n \to \infty} \frac{|p_{G_n}(x) - p_{G_*}(x)|}{W_2^2(G_n, G_*)} = 0$$

for almost every $x \in \mathcal{X}$. However, using the same argument as that of Theorem 3.2 in *Ho and Nguyen* (2016), the above equality cannot hold due to the second order identifiability of f, which is a contradiction. Therefore, we can find positive constant $\epsilon_0 > 0$ such that as long as $W_2(G, G_*) \leq \epsilon_0$, we achieve that $\overline{V}(p_G, p_{G_*}) \gtrsim W_2^2(G, G_*)$. As a consequence, to obtain the conclusion of (2.33), we only need to verify that

$$\inf_{G \in \mathcal{O}_{\overline{k}}:\ W_2(G, G_*) > \epsilon_0} \overline{V}(p_G, p_{G_*})/W_2^2(G, G_*) > 0.$$

Assume that the above result does not hold. It implies that we can find a sequence of $G_n \in \mathcal{O}_{\overline{k}}$ such that $W_2(G_n, G_*) > \epsilon_0$ and $\overline{V}(p_{G_n}, p_{G_*})/W_2^2(G_n, G_*) \to 0$ as $n \to \infty$. Since Θ is a bounded subset of $\mathbb{R}^d$, we can find a subsequence of G_n such that $W_1(G_n, \overline{G}) \to 0$ for some $\overline{G} \in \mathcal{O}_{\overline{k}}$ such that $W_2(\overline{G}, G_*) \geq \epsilon_0$. From our hypothesis, we will have that $\overline{V}(p_{G_n}, p_{G_*}) \to 0$. However, by virtue of Fatou's lemma, we obtain that

$$0 = \liminf_{n \to \infty} \overline{V}(p_{G_n}, p_{G_*}) \geq \overline{V}(p_{\overline{G}}, p_{G_*}).$$

The above equation leads to $p_{\overline{G}}(x) = p_{G_*}(x)$ for almost every $x \in \mathcal{X}$. Due to the identifiability of f, the previous equation leads to $\overline{G} \equiv G_*$, which is a contradiction to the assumption that $W_2(\overline{G}, G_*) \geq \epsilon_0$. We obtain the conclusion of the proposition.

2.8 Appendix C: Posterior contraction under misspecification

This appendix is devoted to the description of a general method for establishing posterior convergence rates of mixing measures under misspecified settings, extending the methods of *Kleijn and van der Vaart* (2006). Once the general method is fully developed we shall be ready to complete the proofs of the main theorems of Section 2.4, which are given in Section 2.8. Recall the weighted Hellinger distance defined in (2.4.1), which leads to the following definition.

Definition 2.8.1. For any set $\mathcal{S} \subset \overline{\mathcal{G}}$, define a real-valued function $\overline{\Psi}_{\mathcal{S}} : \mathbb{R} \to \mathbb{R}^{+}$ as follows

$$\overline{\Psi}_{\mathcal{S}}(r) = \inf_{G \in \mathcal{S}:\ W_2(G, G_*) \geq r/2} \overline{h}^2(p_G, p_{G_*})$$

for any $r \in \mathbb{R}$.

A key ingredient to establishing the posterior contraction bounds is through the existence of tests for subsets of parameters of interest. In the model misspecification setting, it is no longer appropriate to test any mixing measure G against true measure G_0. Instead, following *Kleijn and van der Vaart* (2006), it is appropriate to test any mixing measure G against G_*, which is ultimately achieved by testing $\dfrac{p_{G_0,f_0}}{p_{G_*}} p_G$ against p_{G_0,f_0}. This insight leads us to the following crucial result regarding the existence of test for discriminating G_* against a closed Wasserstein metric ball centered at G_1 for any fixed pair of mixing measures (G_*, G_1).

Lemma 2.8.1. Consider $\mathcal{S} \subset \overline{\mathcal{G}}$ such that $G_* \in \mathcal{S}$. Given $G_1 \in \mathcal{S}$ such that $W_2(G_1, G_*) \geq r$ for some $r > 0$. Assume that either one of the following two sets of conditions holds:

(1) $\mathcal{S}$ is a convex set, in which case, let $\overline{M}(\mathcal{S}, G_1, r) = 1$.

(2) $\mathcal{S}$ is a nonconvex set. In addition, f has first order integral Lipschitz property. In this case, we define that

$$\overline{M}(\mathcal{S}, G_1, r) = D\biggl(\frac{\overline{\Psi}_{\mathcal{S}}(r)}{8\overline{C}(\Theta)}, \mathcal{S} \cap B_{W_2}(G_1, r/2), W_2\biggr).$$

Then, there exists tests ϕ_n such that

$$\begin{aligned} P_{G_0,f_0}\phi_n &\leq \overline{M}(\mathcal{S}, G_1, r)\exp(-n\overline{\Psi}_{\mathcal{S}}(r)/8), && (2.34) \\ \sup_{G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)} \frac{P_{G_0,f_0}}{P_{G_*}} P_G(1-\phi_n) &\leq \exp(-n\overline{\Psi}_{\mathcal{S}}(r)/8). && (2.35) \end{aligned}$$

By means of the existence of tests in Lemma 2.8.1, we have the following result regarding testing G_* versus a complement of a closed Wasserstein ball.

Lemma 2.8.2. Assume that all the conditions in Lemma 2.8.1 hold. Let $D(\epsilon)$ be a non-decreasing function such that , for some $\epsilon_n \geq 0$ and every $\epsilon > \epsilon_n$,

$$\sup_{G \in \mathcal{S}} \overline{M}(\mathcal{S}, G, r) D(\epsilon/2, \mathcal{S} \cap B_{W_2}(G_*, 2\epsilon) \backslash B_{W_2}(G_*, \epsilon), W_2) \leq D(\epsilon).$$

Then, for every $\epsilon > \epsilon_n$ there exist tests ϕ_n (depending on $\epsilon > 0$) such that, for every $J \in \mathbb{N}$,

$$\begin{aligned} P_{G_0,f_0}\phi_n &\leq D(\epsilon) \sum_{t=J}^{[\mathrm{Diam}(\Theta)/\epsilon]} \exp(-n\overline{\Psi}_{\mathcal{S}}(t\epsilon)/8), && (2.36) \\ \sup_{G \in \mathcal{S}: W_2(G, G_*) > J\epsilon} \frac{P_{G_0,f_0}}{P_{G_*}} P_G(1-\phi_n) &\leq \exp(-n\overline{\Psi}_{\mathcal{S}}(J\epsilon)/8). && (2.37) \end{aligned}$$

For any $\epsilon > 0, M > 0$, we define a generalized Kullback-Leibler neighborhood of G_*

by

$$
\begin{aligned}
B_K^*(\epsilon, G_*, P_{G_0,f_0}, M) \ \ := \ \ & \bigg\{ G \in \overline{\mathcal{G}} : \ -P_{G_0,f_0} \log \frac{p_G}{p_{G_*}} \leq \epsilon^2 \log M/\epsilon + \epsilon, \\
& \qquad P_{G_0,f_0} \bigg(\log \frac{p_G}{p_{G_*}} \bigg)^2 \leq \epsilon^2 \left(\log(M/\epsilon)\right)^2 \bigg\}. \quad (2.38)
\end{aligned}
$$

Invoking the results in Lemma 2.8.1 and Lemma 2.8.2, we have the following theorem establishing posterior contraction rate for G_*.

Theorem 2.8.1. Suppose that for a sequence of $\{\epsilon_n\}_{n \geq 1}$ that tends to a constant (or 0) such that $n\epsilon_n^2 \to \infty$, and constants $C, M > 0$, and convex sets $\mathcal{G}_n \subset \overline{\mathcal{G}}$, we have

$$
\begin{aligned}
\log D(\epsilon_n, \mathcal{G}_n, W_2) &\leq n\epsilon_n^2, && (2.39) \\
\Pi(\overline{\mathcal{G}} \backslash \mathcal{G}_n) &\leq \exp(-n(\epsilon_n^2 \log(M/\epsilon_n) + \epsilon_n)(C+4)), && (2.40) \\
\Pi(B_K^*(\epsilon_n, G_*, P_{G_0,f_0}, M)) &\geq \exp(-n(\epsilon_n^2 \log(M/\epsilon_n) + \epsilon_n)C), && (2.41)
\end{aligned}
$$

Additionally, M_n is a sequence such that

$$
\begin{aligned}
\overline{\Psi}_{\mathcal{G}_n}(M_n\epsilon_n) &\geq 8(\epsilon_n^2 \log(M/\epsilon_n) + \epsilon_n)(C+4), && (2.42) \\
\exp(2n(\epsilon_n^2 \log(M/\epsilon_n) + \epsilon_n)) \sum_{j \geq M_n} &\exp(-n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/8) \to 0. && (2.43)
\end{aligned}
$$

Then, $\Pi(G \in \overline{\mathcal{G}} : \ W_2(G, G_*) \geq M_n\epsilon_n | X_1, \ldots, X_n) \to 0$ in P_{G_0,f_0}-probability.

The above theorem is particularly useful for establishing the convergence rate of $G_* \in \mathcal{P}(\Theta)$ for which the suitable sieves $\mathcal{G}_n$ of mixture densities are convex classes of functions. In the situation where $\mathcal{G}_n$ are non-convex, we need the following result, which is the generalization of Theorem 4 in *Nguyen* (2013).

Theorem 2.8.2. Assume that f admits the first order integral Lipschitz property. Additionally, there is a sequence ϵ_n with $\epsilon_n \to 0$ such that $n \log(1/\epsilon_n)^{-2}$ is bounded away from 0, a sequence M_n, a constant $M > 0$, and a sequence of sets $\mathcal{G}_n \subset \overline{\mathcal{G}}$ such that the following conditions hold

$$\begin{aligned}
&\log D(\epsilon/2, \mathcal{G}_n \cap B_{W_2}(G_*, 2\epsilon) \backslash B_{W_2}(G_*, \epsilon), W_2) \\
&\qquad + \sup_{G \in \mathcal{G}_n} \log \overline{M}(\mathcal{G}_n, G, r) \leq n\epsilon_n^2 \ \forall \ \epsilon \geq \epsilon_n, \qquad (2.44)
\end{aligned}$$

$$\frac{\Pi(\overline{\mathcal{G}} \backslash \mathcal{G}_n)}{\Pi(B_K^*(\epsilon_n, G_*, P_{G_0,f_0}, M))} = o\left(\exp\left(-2n\left(\epsilon_n^2 \log\left(\frac{M}{\epsilon_n}\right) + \epsilon_n\right)\right)\right), \qquad (2.45)$$

$$\frac{\Pi(B_{W_2}(G_*, 2j\epsilon_n) \backslash B_{W_2}(G_*, j\epsilon_n))}{\Pi(B_K^*(\epsilon_n, G_*, P_{G_0,f_0}, M))} \leq \exp\Big(n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/16\Big), \ \forall j \geq M_n \qquad (2.46)$$

$$\exp\left(2n\left(\epsilon_n^2 \log\left(\frac{M}{\epsilon_n}\right) + \epsilon_n\right)\right) \sum_{j \geq M_n} \exp\left(-n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/16\right) \to 0. \qquad (2.47)$$

Then, we have that $\Pi(W_2(G, G_*) \geq M_n\epsilon_n | X_1, \ldots, X_n) \to 0$ in P_{G_0,f_0}- probability.

2.9 Appendix D: Proofs of remaining results

We are now ready to complete the proof of the main posterior contraction theorems stated in Section 2.4.

2.9.1 Proof of Theorem 2.4.1

First we show part(i)

Note that the MFM prior places full mass on discrete measure with finite support, it is enough to show $\Pi\left(G \in \mathcal{G}(\Theta) : W_2(G, G_*) \gtrsim \left(\frac{\log\log n}{\log n}\right)^{1/2} \middle| X_1, \ldots, X_n\right) \to 0$.

The proof of this result is a straightforward application of Proposition 2.4.1, Lemma 2.4.3, and Theorem 2.8.1; therefore, we will only provide a sketch of this proof. Similar to the proof of Theorem 2.3.1 (for the well-specified setting), we proceed by constructing a sequence ϵ_n and sieves $\mathcal{G}_n$ that satisfy all the conditions specified in Theorem 2.8.1.

Step 1: First, we choose ϵ_n to satisfy condition (2.41) in Theorem 2.8.1. To that effect, we proceed by making use of the results from Lemma 8.1 in *Kleijn and van der Vaart* (2006). In particular, from Lemma 8.1 of *Kleijn and van der Vaart* (2006), as long as P is a probability measure and Q is a finite measure (with densities p and q respectively, with respect to Lebesgue measure on $\mathbb{R}^d$) such that $h(p,q) \leq \epsilon$ and $\int p^2/q \leq M$, we obtain that

$$\begin{aligned} P\log(p/q) &\lesssim \epsilon^2 \log(M/\epsilon) + \|p-q\|_1, \\ P(\log(p/q))^2 &\lesssim \epsilon^2(\log(M/\epsilon))^2, \end{aligned} \tag{2.48}$$

where the constants in these bounds are universal. For the purpose of our proof, we will choose $p = p_{G_0,f_0}$ and $q = p_G p_{G_0,f_0}/p_{G_*}$.

Since the Gaussian kernel satisfies the integral Lipschitz property up to the first order, by invoking the result of Lemma 2.7.1, we have

$$\|p-q\|_1 = \left\| \frac{p_{G_0,f_0}}{p_{G_*}}(p_G - p_{G_*}) \right\|_1 \lesssim W_1(G,G_*) \leq W_2(G,G_*). \tag{2.49}$$

Additionally, for the Gaussian kernel f given by (3.8), we find that

$$\overline{h}^2(p_G, p_{G_*}) \leq C_1 W_2^2(G, G^*) \tag{2.50}$$

for any $G \in \overline{\mathcal{G}}(\Theta)$. For Gaussian location mixtures as long as there exists $\epsilon_0 > 0$ such that $W_2(G,G_*) \leq \epsilon_0$, we also can check that $\int p_{G_0,f_0}(x) p_{G_*}(x)/p_G(x) \mathrm{d}\mu(x) \leq M^*(\epsilon_0)$ for some positive constant $M^*(\epsilon_0)$ depending only on ϵ_0, G_0, and Θ. Therefore, as long as $W_2(G,G_*) \leq \epsilon \leq \epsilon_0$, we have for mixing measure G,

$$\begin{aligned} -P_{G_0,f_0} \log(p_G/p_{G_*}) &\leq \epsilon^2(\log(M/\epsilon)) + \epsilon, \\ P_{G_0,f_0}[\log(p_G/p_{G_*})]^2 &\leq \epsilon^2(\log(M/\epsilon))^2, \end{aligned}$$

where $M := M^*(\epsilon_0)$. The constants in the bounds are all universal.

Governed by this result, we can write

$$\Pi(B_K^*(\epsilon_n, G_*, P_{G_0,f_0}, M)) \geq \Pi(W_2(G,G_*) \lesssim \epsilon_n)$$

for any sequence $\epsilon_n \leq \epsilon_0$. Now, the packing number $D = D(\epsilon_n)$ (with packing radius ϵ_n) in Lemma 2.4.3 satisfies $D(\epsilon_n) \asymp \left(\frac{\text{Diam}(\Theta)}{\epsilon_n}\right)^d$. Following the result in Lemma 2.4.3, we have with $r = 2$ that

$$\begin{aligned} \log(\Pi(B_K^*(\epsilon_n, G_*, P_{G_0,f_0}.M))) &\gtrsim D(\epsilon_n)(\log c_0 + \log(\epsilon_n/D(\epsilon_n))) \\ &+ \log(\epsilon_n/D(\epsilon_n))(1 + (1 + (2/d))\gamma(D(\epsilon_n) - 1)/D(\epsilon_n)). \end{aligned}$$

With $\epsilon_n \asymp (\frac{\log n}{n})^{1/(d+2)}$, one can check that condition (2.41) and (2.39) hold.

Step 2: Note that condition (2.40) holds automatically since we take $\mathcal{G}_n = \overline{\mathcal{G}}$, while condition (2.39) follows from Lemma 4 in *Nguyen* (2013).

Step 3: Next we will show condition (2.42) and condition (2.43) for some appropriate choice of M_n for the ϵ_n considered in Step 1.

Following proposition 2.4.1 we know that $\overline{\Psi}_{\mathcal{G}_n}(r) \gtrsim \exp\bigg(-(1+8\lambda_{\max}(\lambda_{\min}^{-1}+C_0))/r^2\bigg)$. Using this fact, we can check to see that M_n such that $M_n\epsilon_n \approx \left(\frac{\log\log(n)}{\log n}\right)^{1/2}$ works, with $\epsilon_n \approx (\frac{\log n}{n})^{1/(d+2)}$.

2.9.2 Proof of Lemma 2.8.2

Consider a $t\epsilon/2$ maximal-packing of the set $B_{W_1}(G_*, 2t\epsilon)\backslash B_{W_1}(G_*, t\epsilon)$. Let S_t be the corresponding set of $D(t\epsilon/2, \mathcal{S} \cap B_{W_1}(G_*, 2t\epsilon)\backslash B_{W_1}(G_*, t\epsilon)$ points obtained there in. Then as in Lemma 2.8.1 corresponding to each point G_1 in S_t, there exist $\omega_{n,t}$ which satisfies (2.34) and (2.35). Then taking ϕ_n as the supremum over all these tests $\omega_{n,t}$

over all points in S_t , all $t \geq J$ we see that by the union bound,

$$\begin{aligned} P_{G_0,f_0}\phi_n &\leq \sum_{t>J}\sum_{G_1\in S_t} \overline{M}(\mathcal{S}, G_1, t\epsilon)\exp(-n\overline{\Psi}_{\mathcal{S}}(t\epsilon)/8), \\ &\leq D(\epsilon)\sum_{t=J}^{[\mathrm{Diam}(\Theta)/\epsilon]} \exp(-n\overline{\Psi}_{\mathcal{S}}(t\epsilon)/8) \end{aligned}$$

$$\begin{aligned} \sup_{G\in\cup_{t\geq J}\{B_{W_1}(G_*,2t\epsilon)\setminus B_{W_1}(G_*,t\epsilon)\}} \frac{P_{G_0,f_0}}{P_{G_*}}P_G(1-\phi_n) &\leq \sup_{t\geq J}\exp(-n\overline{\Psi}_{\mathcal{S}}(t\epsilon)/8) \\ &\leq \exp(-n\overline{\Psi}_{\mathcal{S}}(J\epsilon)/8). \end{aligned}$$

The last inequality follows from the fact that $\overline{\Psi}_{\mathcal{S}}(\cdot)$ is an increasing function in its argument.

2.9.3 Proof of Theorem 2.8.2

The following lemma is analogous to Lemma 7.1 in *Kleijn and van der Vaart* (2006) and Lemma 8.1 in *Ghosal et al.* (2000) and can be similarly proved.

Lemma 2.9.1. For every $M, \epsilon > 0$, $C > 0$, and probability measure Π on G, we obtain that

$$\begin{aligned} P_{G_0,f_0}\bigg(\int\prod_{i=1}^{n}\frac{p_G(X_i)}{p_{G_*}(X_i)}d\Pi(G) &\leq \Pi(B_K^*(\epsilon, G_*, P_{G_0,f_0}, M)) \\ &\times\exp(-(1+C)n(\epsilon^2\log{(M/\epsilon)}+\epsilon)\bigg) \leq \frac{\log^2(M/\epsilon)}{C^2 n\,(1+\epsilon\log(M/\epsilon))^2}. \end{aligned}$$

Equipped with this lemma we can now prove the theorem as follows. Denote A_n the event such that

$$\int \prod_{i=1}^{n} \frac{p_G(X_i)}{p_{G_*}(X_i)} d\Pi(G) \leq \Pi(B_K^*(\epsilon, G_*, P_{G_0,f_0}, M)) \exp(-n(\epsilon^2 \log(M/\epsilon) + \epsilon)(1 + C)).$$

The above result indicates that $P_{G_0,f_0} \mathbb{1}_{A_n} \leq (C^2 n)^{-1} \log(M/\epsilon)^2$ for any $\epsilon > 0$ and $C > 0$.

For any sequence ϵ_n, we denote $\mathcal{U}_n = \left\{G \in \overline{\mathcal{G}} : W_2(G, G_*) \geq M_n \epsilon_n\right\}$ and $S_{n,j} = \{G \in \mathcal{G}_n :$
$W_2(G, G_*) \in [j\epsilon_n, (j+1)\epsilon_n)\}$ for any $j \geq 1$. From the result of Lemma 2.8.1 and condition (2.52), there exists a test ϕ_n such that inequality (2.36) and (2.37) hold when $D(\epsilon_n) = \exp(n\epsilon_n^2)$. Now, we have

$$\begin{aligned}
& P_{G_0,f_0} \Pi(G \in \mathcal{U}_n | X_1, \ldots, X_n) \\
& \qquad = P_{G_0,f_0} \phi_n \Pi(G \in \mathcal{U}_n | X_1, \ldots, X_n) \\
& \qquad + P_{G_0,f_0} (1 - \phi_n) \mathbb{1}_{A_n} \Pi(G \in \mathcal{U}_n | X_1, \ldots, X_n) \\
& \qquad + P_{G_0,f_0} (1 - \phi_n) \mathbb{1}_{A_n^c} \Pi(G \in \mathcal{U}_n | X_1, \ldots, X_n) \\
& \qquad \leq P_{G_0,f_0} \phi_n + P_{G_0,f_0} \mathbb{1}_{A_n} + P_{G_0,f_0} (1 - \phi_n) \mathbb{1}_{A_n^c} \Pi(G \in \mathcal{U}_n | X_1, \ldots, X_n).
\end{aligned} \tag{2.51}$$

According to Lemma 2.8.2, we have $P_{G_0,f_0} \phi_n \leq \exp(n\epsilon_n^2) \sum\limits_{j \geq M_n} \exp\left(-n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/8\right) \to 0$, which is due to condition (2.47). Additionally, from the formation of A_n, we also obtain that $P_{G_0,f_0} \mathbb{1}_{A_n} \leq (C^2 n)^{-1} \log(M/\epsilon)^2$. If $n \log(1/\epsilon)^{-2} \to \infty$, it is clear that $P_{G_0,f_0} \mathbb{1}_{A_n} \to 0$ for any $C \geq 1$. If $n \log(1/\epsilon)^{-2}$ does not tend to ∞ but is bounded away from 0, then we can choose $C > 0$ large enough such that $P_{G_0} \mathbb{1}_{A_n}$ is sufficiently close to 0. Therefore, the first two terms in (2.52) can always be made to vanish to 0. To achieve the conclusion of the theorem, it is sufficient to demonstrate that the third term in (2.51) goes to 0. In fact, we have the following equation

$$\Pi(G \in \mathcal{U}_n | X_1, \ldots, X_n) = \Big(\int\limits_{\mathcal{U}_n} \prod_{i=1}^{n} \frac{p_G(X_i)}{p_{G_*}(X_i)} d\Pi(G) \Big) / \Big(\int \prod_{i=1}^{n} \frac{p_G(X_i)}{p_{G_*}(X_i)} d\Pi(G) \Big).$$

From the formulation of A_n, we have

$$\begin{aligned}
&P_{G_0,f_0}(1-\phi_n)\mathbb{1}_{A_n^c}\Pi(G \in \mathcal{U}_n | X_1, \ldots, X_n) \qquad (2.52)\\
&\leq \Big\{ P_{G_0,f_0}(1-\phi_n) \Big(\int\limits_{\mathcal{U}_n} \prod_{i=1}^{n} \frac{p_G(X_i)}{p_{G_*}(X_i)} d\Pi(G) \Big) \Big\} \\
&\qquad\qquad \Big/ \Big\{ \Pi(B_K^*(\epsilon, G_*, P_{G_0,f_0}, M)) \exp(-(1+C)n(\epsilon^2 \log(M/\epsilon) + \epsilon) \Big\}.
\end{aligned}$$

By means of Fubini's theorem, we obtain that

$$\begin{aligned}
P_{G_0,f_0}(1-\phi_n) \Big(\int\limits_{\mathcal{U}_n \cap \mathcal{G}_n} \prod_{i=1}^{n} \frac{p_G(X_i)}{p_{G_*}(X_i)} d\Pi(G) \Big) &= \int\limits_{\mathcal{U}_n \cap \mathcal{G}_n} \frac{P_{G_0,f_0}}{p_{G_*}} p_G(1-\phi_n) d\Pi(G) \\
&\leq \sum_{j \geq M_n} \Pi(S_{n,j}) \exp(-n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon)/8) \qquad (2.53)
\end{aligned}$$

where the last inequality is due to inequality (2.37) and condition (2.52). Furthermore, by means of Fubini's theorem

$$\begin{aligned}
P_{G_0,f_0}(1-\phi_n)\Big(&\int\limits_{\mathcal{U}_n\backslash\mathcal{G}_n}\prod_{i=1}^{n}\frac{p_G(X_i)}{p_{G_*}(X_i)}d\Pi(G)\Big)\\
&\leq \;P_{G_0,f_0}\int\limits_{\mathcal{U}_n\backslash\mathcal{G}_n}\prod_{i=1}^{n}\frac{p_G(X_i)}{p_{G_*}(X_i)}d\Pi(G)\\
&= \int\limits_{\mathcal{U}_n\backslash\mathcal{G}_n}\Big(\prod_{i=1}^{n}\int\frac{p_G(x_i)}{p_{G_*}(x_i)}p_{G_0,f_0}(x_i)dx_i\Big)\Pi(G)\\
&\leq \int\limits_{\mathcal{U}_n\backslash\mathcal{G}_n}\Pi(G)=\Pi(\mathcal{U}_n\backslash\mathcal{G}_n)\leq\Pi(\overline{\mathcal{G}}\backslash\mathcal{G}_n) \qquad (2.54)
\end{aligned}$$

where the second inequality in the above result is due to Lemma 2.4.1. By combining the results of (2.52), (2.53), and (2.54), we obtain

$$\begin{aligned}
P_{G_0,f_0}&(1-\phi_n)\mathbb{1}_{A_n^c}\Pi(G\in\mathcal{U}_n|X_1,\ldots,X_n)\\
&\leq\frac{\sum\limits_{j\geq M_n}\Pi(S_{n,j})\exp(-n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon)/8)+\Pi(\overline{\mathcal{G}}\backslash\mathcal{G}_n)}{\Pi(B_K^*(\epsilon,G_*,P_{G_0,f_0},M))\exp(-(1+C)n(\epsilon^2\log(M/\epsilon)+\epsilon)}\\
&\leq\exp((1+C)n(\epsilon^2\log(M/\epsilon)+\epsilon)\sum_{j\geq M_n}\exp\left(-n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/16\right)\\
&\qquad\qquad +o(\exp((C-1)n(\epsilon^2\log M/\epsilon+\epsilon))
\end{aligned}$$

where the last inequality is due to condition (2.53) and (2.54). If $n\log(1/\epsilon_n)^{-2}$ is bounded away from 0 by choosing $C\geq 1$, the right hand side term of the above display will go to 0 due to condition (2.47). Therefore, we have $P_{G_0,f_0}(1-\phi_n)\mathbb{1}_{A_n^c}\Pi(G\in\mathcal{U}_n|X_1,\ldots,X_n)\to 0$ as $n\to\infty$.

As a consequence, $P_{G_0,f_0}\Pi(G\in\mathcal{U}_n|X_1,\ldots,X_n)\to 0$. We achieve the conclusion of the theorem.

2.9.4 Proof of Lemma 2.8.1

For the setting (1) when $\mathcal{S}$ is convex, since $B_{W_2}(G_1, r/2)$ is a convex set, we also have $\mathcal{S} \cap B_{W_2}(G_1, r/2)$ is a convex set. By means of the result of Theorem 6.1 in *Kleijn and van der Vaart* (2006), there exist tests ϕ_n such that

$$\begin{aligned} P_{G_0,f_0}\phi_n &\leq \left[1 - \frac{1}{2}\inf_{G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)} \overline{h}^2(p_G, p_{G_*})\right]^n, \\ \sup_{G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)} \frac{P_{G_0,f_0}}{p_{G_*}} p_G(1 - \phi_n) &\leq \left[1 - \frac{1}{2}\inf_{G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)} \overline{h}^2(p_G, p_{G_*})\right]^n. \end{aligned}$$

Due to inequality $(1 - x)^n \leq \exp(-nx)$ for all $0 < x < 1$ and $n \geq 1$, the above inequalities become

$$\begin{aligned} P_{G_0,f_0}\phi_n &\leq \exp\biggl(-\frac{n}{2}\inf_{G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)} \overline{h}^2(p_G, p_{G_*})\biggr), \\ \sup_{G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)} \frac{P_{G_0,f_0}}{p_{G_*}} p_G(1 - \phi_n) &\leq \exp\biggl(-\frac{n}{2}\inf_{G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)} \overline{h}^2(p_G, p_{G_*})\biggr). \end{aligned}$$

Now, since $W_2(G_1, G_*) = r$ and $W_2(G, G_1) \leq r/2$ as long as $G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)$, it implies that $W_2(G_1, G_*) \geq r/2$. Therefore, according to Definition 2.8.1, we will obtain that

$$\overline{\Psi}_{\mathcal{S}}(r) = \inf_{G \in \mathcal{S}:\ d_{W_2}(G, G_*) \geq r/2} \overline{h}^2(p_G, p_{G_*}) \leq \inf_{G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)} \overline{h}^2(p_G, p_{G_*}).$$

With the above inequality, we reach the conclusion of part (1).

Regarding part (2), we consider a maximal $c_0 r$-packing of $\mathcal{S} \cap B_{W_2}(G_1, r/2)$ under W_2 metric. It gives us a set of $\overline{M} = \overline{M}(\mathcal{S}, G_1, r) = D(c_0 r, \mathcal{S} \cap B_{W_2}(G_1, r/2), W_2)$ points $\widetilde{G}_1, \ldots, \widetilde{G}_{\overline{M}}$ in $\mathcal{S} \cap B_{W_2}(G_1, r/2)$.

Now, for any $G \in \mathcal{S} \cap B_{W_2}(G_1, r/2)$, we can find $t \in \bigl\{1, \ldots, \overline{M}\bigr\}$ such that

$W_1(G, \widetilde{G}_t) \leq c_0 r$. Due to the triangle inequality, we achieve that

$$\begin{aligned} \overline{h}(p_G, p_{G_*}) &\geq \overline{h}(p_{G_*}, p_{\widetilde{G}_t}) - \overline{h}(p_G, p_{\widetilde{G}_t}) \\ &\geq \biggr(\overline{\Psi}_{\mathcal{S}}(r)\biggr)^{1/2} - \biggr(\overline{C}(\Theta) c_0 r\biggr)^{1/2} \end{aligned}$$

where the second inequality is due to Definition 2.8.1 and Lemma 2.7.1. By choosing the positive number $c_0 = \overline{\Psi}_{\mathcal{S}}(r)/(4\overline{C}(\Theta)r)$, we obtain

$$\overline{h}(p_G, p_{\widetilde{G}_t}) \leq \overline{\Psi}_{\mathcal{S}}(r)/2 \leq \overline{h}(p_{G_*}, p_{\widetilde{G}_t})/2.$$

It eventually leads to $\overline{h}(p_G, p_{G_*}) \geq \overline{h}(p_{G_*}, p_{\widetilde{G}_t})/2$. According to the result of Theorem 6.1 in *Kleijn and van der Vaart* (2006) and inequality $(1-x)^n \leq \exp(-nx)$ for all $0 < x < 1$ and $n \geq 1$, by denoting

$$A_t := \left\{G \in \overline{\mathcal{G}} : \ \overline{h}(p_G, p_{\widetilde{G}_t}) \leq \overline{h}(p_{G_*}, p_{\widetilde{G}_t})/2\right\},$$

there exists test $\psi_n^{(t)}$ such that

$$\begin{aligned} P_{G_0, f_0} \psi_n^{(t)} &\leq \exp\biggr(-\frac{n}{2} \inf_{G \in A_t} \overline{h}^2(p_G, p_{G_*})\biggr), \\ \sup_{G \in A_t} \frac{P_{G_0, f_0}}{p_{G_*}} p_G (1 - \psi_n^{(t)}) &\leq \exp\biggr(-\frac{n}{2} \inf_{G \in A_t} \overline{h}^2(p_G, p_{G_*})\biggr). \end{aligned}$$

Since $\overline{h}^2(p_G, p_{G_*}) \geq \overline{h}(p_{G_*}, p_{\widetilde{G}_t})/2 \geq \overline{\Psi}_{\mathcal{S}}(r)/2$ for all $G \in A_t$, the above inequalities can be rewritten as

$$\begin{aligned} P_{G_0,f_0}\psi_n^{(t)} &\leq \exp(-n\overline{\Psi}_{\mathcal{S}}(r)/8), \\ \sup_{G\in A_t}\frac{P_{G_0,f_0}}{p_{G_*}}p_G(1-\psi_n^{(t)}) &\leq \exp(-n\overline{\Psi}_{\mathcal{S}}(r)/8). \end{aligned}$$

By choosing $\phi_n = \max\limits_{1\leq t\leq \overline{M}} \psi_n^{(t)}$, we quickly achieve that

$$\begin{aligned} P_{G_0,f_0}\phi_n &\leq \overline{M}(\mathcal{S}, G_1, r)\exp(-n\overline{\Psi}_{\mathcal{S}}(r)/8), \\ \sup_{G\in\mathcal{S}\cap B_W(G_1,r/2)}\frac{P_{G_0,f_0}}{p_{G_*}}p_G(1-\phi_n) &\leq \exp(-n\overline{\Psi}_{\mathcal{S}}(r)/8). \end{aligned}$$

As a consequence, we obtain the conclusion of the lemma.

2.9.5 Proof of Theorem 2.4.3

Due to the assumption on prior p_K, it is sufficient to demonstrate that

$$\Pi\biggl(G\in\mathcal{O}_{\overline{k}}(\Theta) : W_2(G,G_*) \gtrsim \frac{(\log n)^{1/4}}{n^{1/4}}|X_1,\ldots,X_n\biggr) \to 0$$

in P_{G_0,f_0}- probability. We divide our proof for the above result into the following steps

Step 1: To obtain the bound for $\Pi(B_K^*(\epsilon_n, G_*, P_{G_0,f_0}, M))$, we use Lemma 8.1 from *Kleijn and van der Vaart* (2006) to obtain a bound for weighted KL divergence and squared weighted KL divergence. Similar to the proof of Theorem 2.4.1, with the choice of $p = P_{G_0,f_0}$ and of finite measure $q = p_G p_{G_0,f_0}/p_{G_*}$, as long as $G \in \mathcal{O}_{\overline{k}}$ such that

$\overline{h}(p_G, p_{G_*}) \leq \epsilon$ and $\int p_{G_0,f_0} p_{G_*}/p_G \leq M$ then we obtain that

$$-P_{G_0,f_0} \log \frac{p_G}{p_{G_*}} \lesssim \epsilon^2 \log(M/\epsilon) + \epsilon$$
$$P_{G_0,f_0} \left(\log \frac{p_G}{p_{G_*}} \right)^2 \leq \epsilon^2 (\log(M/\epsilon))^2.$$

For the purpose of this proof we use $M = M^*(\epsilon_0)$, where $M^*(\epsilon_0)$ is as in condition (M.2) in Section 2.4.3. Now, according to Lemma 2.7.1, as f admits integral Lipschitz property up to the first order, we obtain that $\overline{h}^2(p_G, p_{G_*}) \leq \overline{C} W_1(G, G_*)$ for any $G \in \mathcal{O}_{\overline{k}}$ where $\overline{C}$ is a positive constant depending only on Θ. Now, from the discussion in the above paragraph we have

$$\begin{aligned} \Pi(B_K^*(\epsilon_n, G_*, P_{G_0,f_0}, M)) &\geq \Pi(G \in \mathcal{O}_{\overline{k}} : W_1(G, G_*) \leq \overline{C}\epsilon_n^2) \\ &\gtrsim \Pi(G \in \mathcal{E}_{\overline{k}} : W_1(G, G_*) \leq \overline{C}\epsilon_n^2) \\ &\gtrsim \epsilon_n^{2(c_H + \gamma)}. \end{aligned}$$

where the last inequality can be obtained similar to equation (2.18) based on the assumption $\gamma < \overline{k}$. Now, we note that $\overline{\Psi}_{\mathcal{G}_n(r)} \gtrsim r^4$, since f is assumed to be second order identifiable and to satisfy the integral Lipschitz property of second order. Then, by choosing $\epsilon_n = n^{-1}$ and $M_n = \overline{A} \left(\frac{\log(n)}{n^3} \right)^{1/4}$ for some sufficiently large $\overline{A}$, we can see that condition (2.46) is satisfied.

Step 2: We choose the sieves $\mathcal{G}_n = \mathcal{O}_{\overline{k}}$ for all $n \geq 1$. With these choices, it is clear that $\Pi(\overline{\mathcal{G}} \backslash \mathcal{G}_n) = 0$. Therefore, condition (2.45) is satisfied.

Step 3: For condition (2.47) to be satisfied,

$$\begin{aligned}\exp\left(2n\left(\epsilon_n^2\log\left(\frac{M}{\epsilon_n}\right)+\epsilon_n\right)\right)\sum_{j\geq M_n}\exp\left(-n\overline{\Psi}_{\mathcal{G}_n}(j\epsilon_n)/16\right)&\\ \lesssim&\sum_{j\geq M_n}\exp(-\overline{C}n^{-3}M_n^4/16)\\ \lesssim&\ 2\exp(-\overline{C}n^{-3}M_n^4/16)\\ \lesssim&\ 2n^{-\frac{\overline{AC}}{16}}\to 0.\end{aligned}$$

Lemma 2.9.2. Assume that $\sup_{\theta\in\Theta,x} f(x|\theta)<\infty$, then, for $q\geq 2$, G_1, G_2 mixing measures on Θ,

$$\|p_{G_1}-p_{G_2}\|_q\lesssim(h(p_{G_1},p_{G_2}))^{2/q}, \tag{2.55}$$

whereas for $1\leq q\leq 2$,

$$\|p_{G_1}-p_{G_2}\|_q\lesssim(h(p_{G_1},p_{G_2}))^{1/q}. \tag{2.56}$$

Proof. Assume $q\geq 2$,

$$\begin{aligned}\|p_{G_1}-p_{G_2}\|_q^q &= \int|p_{G_1}(x)-p_{G_2}(x)|^q\mathrm{d}x\\ &\leq \int|\sqrt{p_{G_1}}(x)-\sqrt{p_{G_2}(x)}|^q|\sqrt{p_{G_1}}(x)+\sqrt{p_{G_2}(x)}|^q\\ &\leq (2\sup_{\theta\in\Theta,x}f(x|\theta))^q\int|\sqrt{p_{G_1}}(x)-\sqrt{p_{G_2}(x)}|^{2+(q-2)}\mathrm{d}x\\ &\leq (2\sup_{\theta\in\Theta,x}f(x|\theta))^{2q-2}\int|\sqrt{p_{G_1}}(x)-\sqrt{p_{G_2}(x)}|^2\mathrm{d}x\\ &\lesssim h^2(p_{G_1},p_{G_2}).\end{aligned} \tag{2.57}$$

For $2 \geq q \geq 1$,

$$\begin{aligned}
\|p_{G_1} - p_{G_2}\|_q^{2q} &\leq \int |p_{G_1}(x) - p_{G_2}(x)|^{2q} \mathrm{d}x \\
&\leq \int |\sqrt{p_{G_1}}(x) - \sqrt{p_{G_2}(x)}|^{2q} |\sqrt{p_{G_1}}(x) + \sqrt{p_{G_2}(x)}|^{2q} \\
&\leq (2 \sup_{\theta \in \Theta, x} f(x|\theta))^{2q} \int |\sqrt{p_{G_1}}(x) - \sqrt{p_{G_2}(x)}|^{2+(2q-2)} \mathrm{d}x \\
&\leq (2 \sup_{\theta \in \Theta, x} f(x|\theta))^{4q-2} \int |\sqrt{p_{G_1}}(x) - \sqrt{p_{G_2}(x)}|^{2} \mathrm{d}x \\
&\lesssim h^2(p_{G_1}, p_{G_2}). \qquad (2.58)
\end{aligned}$$

□

CHAPTER III

Bayesian Contraction for Dirichlet Process Mixtures of Smooth Densities

Dirichlet process mixture models (DPMM) have been an important modeling toolbox for numerous domains arising from biological, physical, and social sciences. However, a concrete understanding of the effects of parameter space, kernel density function, and dimension on the posterior convergence rate of mixing measure with Dirichlet process prior has remained elusive. In this chapter, we study carefully the effects of these factors under both the well-specified and misspecified settings of convolution DPMM. Moreover, we develop a novel metric that generalizes the Wasserstein metric. Our technique involces establishing fundamental connections between (weighted) Hellinger distance the generalized Wasserstein metric under various regimes of parameter space, kernel density function, and dimension. To our findings, these connections also provide important insight on complex posterior convergence rate of mixing measure arising from challenging settings of other hierarchical Bayesian nonparametric models.[1].

[1] This work has been published in *Guha et al.* (2020+)

3.1 Introduction

Mixture models are often used by statisticians as black-box methods to analyse data generated from heterogeneous subpopulations as an outcome of complex processes (*McLachlan and Basford* (1988); *Lindsay* (1995); *Mengersen et al.* (2011)). A common issue faced by all practicing mixture modelers is the choice of kernels appropriate for analyses. Smooth density families are popular choices for nonparametric inference. In that regard, families of varying smoothness such as Laplace or Gaussian kernels have been used extensively for inferential problems related to density estimation, clustering analysis etc., (cf. *Kotz et al.*; *Bailey et al.* (1994.); *Roeder and Wasserman* (1997); *Robert* (1996); *Banfield and Raftery* (1993)).

The following are some of the inferential questions to ask.

(I) How do you choose between heavy or light-tailed kernels for appropriate inference?

(II) Suppose we allow the number of components to grow with the sample size, can we efficiently estimate the parameters corresponding to components in an arbitrary region of the parameter space?

(III) Gaussian kernels are the most popular choices for model fitting, however, the parameter estimation rates for Gaussian kernels tends to be very low *Fan* (1991); *Zhang* (1990). There is therefore an inconsistency in practice and theory. Can this be resolved?

This chapter provides answers to some of these questions.

Consider discrete mixing measures $G = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}$. Here, $\boldsymbol{p} = (p_1, \ldots, p_k)$ is a vector of mixing weights, while atoms $\{\theta_i\}_{i=1}^{\infty}$ are elements in a given space $\Theta \in \mathbb{R}^d$. Mixing measure G is combined with a likelihood function $f(\cdot|\theta)$ with respect to Lebesgue measure μ to yield a mixture density: $p_G(\cdot) = \int f(\cdot|\theta) \mathrm{d}G(\theta) = \sum_{i=1}^{\infty} p_i f(\cdot|\theta_i)$. We call this an *infinite mixture model.* The atoms θ_i's are representatives of the underlying subpopulations.

This setup can also be represented in a different framework where we observe a random variable Y which is a mixture of the true signal, X and a noise component Z, i.e., the model assumes $Y = X + Z$, with $X \sim G$, $Z \sim f$. The problem of inferring about the unknown distribution G is called the deconvolution problem. The asymptotic behavior for the deconvolution problem involving smooth densities have been well-studied by several authors (cf. *Carroll and Hall* (1988); *Fan* (1991); *Zhang* (1990)).

In the Bayesian paradigm, inference on the mixing measure is carried out by the choice of a suitable prior distribution on G which provides a way to compute the posterior distributions for objects of interest. A common choice of such priors for mixing measures is the Dirichlet process (*Ferguson* (1973); *Blackwell and MacQueen* (1973); *Sethuraman* (1994)), giving rise to the famous Dirichlet process mixture models (*Antoniak* (1974); *Lo* (1984); *Escobar and West* (1995)). There is a well-established asymptotic theory on how such Bayesian nonparametric mixture models result in asymptotically optimal estimation procedures for the population density. See, for instance, *Ghosal et al.* (1999); *Ghosal and van der Vaart* (2007); *Shen et al.* (2013) for theoretical results specifically on DP mixtures, and *Ghosal et al.* (2000); *Shen and Wasserman* (2001); *Walker et al.* (2007) for general BNP models. In particular, *Ghosal and van der Vaart* (2001, 2007); *Scricciolo* (2011); *Shen et al.* (2013) have studied posterior contraction rates with convolution mixtures when the kernel in question is the gaussian kernel while *Gao and van der Vaart* (2016); *Scricciolo* (2017) consider the scenario for Laplace kernels. A number of these works deal with density estimation scenarios. However, as we have shown in Chapter II, good contraction rates for density estimation do not necessarily result in efficient estimation of parameters. This is because the efficiency in density estimation is borne out as a result of smoothness of the kernel, which itself may lead to overparametrization for parameter estimates, thereby causing an inefficiency in estimation. *Nguyen* (2013); *Gao and van der Vaart* (2016) derive contraction rates for parameter estimation of convolution kernels. In their results, they make the critical assumption that the parameter space Θ, which forms the support of the prior on atoms, is compact. However, practical application of convolution mixtures involve placing a prior on an unbounded support *Escobar and West* (1995); *Roeder and Wasserman* (1997). This results in a glaring mismatch between theoretical and practical frameworks. We address this issue in this work.

The choice of the kernel f, and the prior on the unknown distribution G, affect the outcome of inference drastically for a practitioner of Bayesian mixture models. With the advent of new methods such as Variational Inference (*Blei et al.* (2003)) and also efficient MCMC techniques, solutions to inferential questions require much lesser computational time. In addition Optimal Transport provides adequate tools for parameter estimation. However, the choice of the appropriate support for the prior still remains an open problem and this theoretical assumption has been difficult to get rid of when deadling with asymptotics. Often in practice, a flat prior is employed over a restricted region of the parameter space when one has no bias towards any specific vicinity. However, restricting the prior to a specific region of the parameter space risks misspecification of the support of the prior. On the other hand, an almost flat prior with spread out mass would mean the contraction rate suffers if the support of the prior is too "large". Being mindful of this tension between the restriction and inclusivity of the support of the prior, in this work, we design a sieve type method that allows us to expand the support of the prior at the appropriate rate so as to negligibly affect the asymptotic contraction rates. This method can be coupled with other choices of priors for more efficient results.Sieve methods have been implemented for density estimation by many authors such as *Ghosal and van der Vaart* (2001); *Shen and Wong* (1994); *Wong and Shen* (1995); *Van de Geer* (1993); *Birge and Massart* (1998). In this chapter we also show that this expansion rate is polynomial in the number of samples for supersmooth Gaussian kernels and logarithmic for ordinary smooth Laplace kernels.

The choice of the appropriate prior is another difficult decision the practitioner has to make. While Supersmooth kernels such as Gaussian kernels give fast density estimation rates (*Ghosal and van der Vaart* (2001)), they have slow logarithmic rates of convergence for parameter estimation as pointed out in Chapter II. The scenario is reversed for Laplace kernels, which show faster polynomial contraction behavior for parameter estimation, while the rates for density estimation are slower than the $n^{-1/2}$ rates obtained for Gaussian kernels. In that respect, it might be counterintuitive to note that the expansion rate of the parameter space is exponentially faster for Gaussian kernels than that for Laplace kernels. To resolve this seemingly puzzling issue, we develop in this chapter, a new theoretical framework via a novel metric to evaluate parameter estimation. The metric developed generalizes the well-known Wasserstein metric which has thus far been used successfully to study parameter estimation (*Gao and van der Vaart* (2016); *Scricciolo* (2017)). We call our metric the Orlicz-Wasserstein metric since it uses the notion of Orlicz norm for comparison of random variable along with optimal transport to compare distributions. Using the above notion of convergence we are able to show that the slow convergence rate for Gaussian kernels essentially arises from atoms of the posterior which are close to the true atoms and the posterior places vanishingly small mass on mixing measures with atoms away from the truth.

Further Remarks :

(i) "Gaussian kernels lead to coagulative mixing measures aposteriori": Even though the contraction rate is very poor for Gaussian kernels, this is mostly due to the fact that samples from the posterior place large mass on atoms close to true atoms and this mass is shared by various atoms close to the truth. Contraction rates are different for various parts of the parameter space, for example, it is almost polynomial in regions away from the true atoms, but logarithmic close to the truth.

(ii) "Gaussian kernels better for misspecified scenarios than Laplace kernels?": Laplace kernels being heavy tailed and sharp-peaked tend to have a large bias in density estimation. On the other hand, Gaussian kernels suffer from slow contraction rates for parameter estimation. However, the results of this chapter suggest the parameter estimates in regions away from the true atoms contract almost polynomially. In that context, accounting for lack of knowledge of the truth, it might be more conservative to use Gaussian mixtures for parameter estimation. Moreover, as an added advantage this would also allow us to estimate the neighborhoods regions of the truth in a fast and efficient manner.

The remainder of the chapter is organized as follows. Section 3.2 provides necessary backgrounds about mixture models, Wasserstein distances and several key notions of Dirichlet process mixture models. Section 3.3 presents presents exact lower bounds for the Hellinger metric with respect to Wasserstein distances for varying degree of smoothness of kernels as well as posterior contraction theorems corresponding to those kernels. In Section 3.4 we develop a novel metric extending the notion of Wasserstein distance and show that it helps to characterize the coagulative nature of atoms for Gaussian kernels In Section 3.5, we provide illustrations of this coagulative nature for

Gaussian kernels via a simulation study. Proofs of results are deferred to the Appendices.

Notation: For any function $f : \mathcal{X} \to \mathbb{R}$, we denote $\widetilde{f}(\omega)$ as the Fourier transformation of function f.

Given two densities p, q (with respect to the Lebesgue measure μ), the total variation distance is given by $V(p,q) = (1/2)\int |p(x) - q(x)| \mathrm{d}\mu(x)$. Additionally, the squared Hellinger distance is given by $h^2(p,q) = (1/2)\int (\sqrt{p(x)} - \sqrt{q(x)})^2 \mathrm{d}\mu(x)$. Furthermore, the Kullback-Leibler (KL) divergence is given by $K(p,q) = \int \log(p(x)/q(x)) p(x) \mathrm{d}\mu(x)$ and the squared KL divergence is given by $K_2(p,q) = \int \log(p(x)/q(x))^2 p(x) \mathrm{d}\mu(x)$. $\| \cdot \|$ is used to denote the l_2 norm, where as $\| \cdot \|_\Sigma$ denotes the scaled l_2 norm with scaling matrix, Σ. Let λ_{max} and λ_{min} respectively denote the maximum and minimum eigenvalues of Σ.

For a measurable function f, let Qf denote the integral $\int f dQ$. For any $\kappa = (\kappa_1, \ldots, \kappa_d) \in \mathbb{N}^d$, we denote $\dfrac{\partial^{|\kappa|} f}{\partial \theta^\kappa}(x|\theta) = \dfrac{\partial^{|\kappa|} f}{\partial \theta_1^{\kappa_1} \ldots \partial \theta_d^{\kappa_d}}(x|\theta)$ where $\theta = (\theta_1, \ldots, \theta_d)$. For any metric d on Θ, we define the open ball of d-radius ϵ around $\theta_0 \in \Theta$ as $B_d(\epsilon, \theta_0)$. We use $D(\epsilon, \Omega, \tilde{d})$ to denote the maximal ϵ-packing number for a general set Ω under a general metric $\tilde{d}$ on Ω. Additionally, the expression $a_n \gtrsim b_n$ will be used to denote the inequality up to a constant multiple where the value of the constant is independent of n. We also denote $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Furthermore for any set A, we denote A^c as its complement, while $B(x,r)$ denotes the ball, with respect to the l_2 norm, of radius $r > 0$ centered at $x \in \mathbb{R}^d$. The expression $D(\epsilon, \mathscr{P}, d)$ used in the chapter denotes the ϵ-packing number of the space $\mathscr{P}$ relative to the metric d. d is replaced by h to denote the hellinger norm. Finally, we use $\text{Diam}(\Theta) = \sup\{\|\theta_1 - \theta_2\| : \theta_1, \theta_2 \in \Theta\}$ to denote the diameter of a given parameter space Θ relative to the l_2 norm, $\| \cdot \|$, for elements in $\mathbb{R}^d$.

3.2 Preliminary

We recall the notion of Wasserstein distance for mixing measures, along with the notions of Dirichlet Process mixture models that prove useful in the remainder of the chapter.

Mixture model Throughout the chapter, we assume that $X_1, \ldots, X_n$ are i.i.d. samples from a true but unknown distribution P_{G_0} with given density function

$$p_{G_0} := \int f(x|\theta) dG_0(\theta) = \sum_{i=1}^{\infty} p_i^0 f(x|\theta_i^0)$$

where $G_0 = \sum\limits_{i=1}^{\infty} p_i^0 \delta_{\theta_i^0}$ is a true but unknown mixing distribution with possibly infinitely many support points. Moreover, assume that $\sup_i \|\theta_i^0\| < \infty$. Also, $\left\{f(x|\theta), \theta \in \Theta \subset \mathbb{R}^d\right\}$ is a given family of probability densities (or equivalently kernels) with respect to a sigma-finite measure μ on $\mathcal{X}$ where $d \geq 1$. For this work, f is either an ordinary smooth kernel or a supersmooth kernel. Furthermore, Θ is a chosen parameter space, where we empirically believe that the true parameters belong to. In a well-specified setting, all support points of G_0 reside in Θ, but this may not be the case in a misspecified setting.

Regarding the space of mixing measures, let $\mathcal{E}_k := \mathcal{E}_k(\Theta)$ and $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ respectively denote the space of all mixing measures with exactly and at most k support points, all in Θ. Additionally, denote $\mathcal{G} := \mathcal{G}(\Theta) = \underset{k \in \mathbb{N}_+}{\cup} \mathcal{E}_k$ the set of all discrete measures with finite supports on Θ. Moreover, $\overline{\mathcal{G}}(\Theta)$ denotes the space of all discrete measures (including those with countably infinite supports) on Θ. Finally, $\mathcal{P}(\Theta)$ stands for the space of all probability measures on Θ.

Dirichlet process mixture Models We assume that the kernel f is well-specified. A Bayesian mixture modeler places a prior distribution Π on a suitable space $\overline{\mathcal{G}}(\Theta)$. The posterior corresponding to Π_n(varying with sample size) can be computed as:

$$\Pi_n(G \in B \big| X_1, \ldots, X_n) = \frac{\int_B \prod_{i=1}^n p_G(X_i) \mathrm{d}\Pi_n(G)}{\int_{\overline{\mathcal{G}}(\Theta)} \prod_{i=1}^n p_G(X_i) \mathrm{d}\Pi_n(G)}. \tag{3.1}$$

Our primary interest is to study the posterior contraction behavior to G_0 with varying values of Π_n.

One of the most popular models that has been widely used in practice to study G_0 is the Dirichlet Process Mixture Models (DPMM), which can be stated as follows

$$\begin{aligned} G &\sim DP(\alpha, H), \\ \theta_1, \ldots, \theta_n &\overset{i.i.d.}{\sim} G, \\ X_i | \theta_i &\sim f(X_i|\theta_i), \quad \forall i = 1, \ldots, n. \end{aligned} \tag{3.2}$$

where $\{f(\cdot|\theta), \theta \in \Theta\}$ is a chosen location family of density functions and H is a prior distribution on Θ. The common assumptions/problems that people usually utilize and face with DPMM are:

(1) Some atoms of true mixing measure G_0 may not lie in the chosen parameter space Θ, i.e., we are under misspecified parameter space setting.

(2) Kernel f is potentially different from f_0, i.e., we are under misspecified kernel setting.

(3) The role of dimension d in the posterior contraction of mixing measure G. The application of studying such constant is to understand an open problem regarding

the roles of sample sizes in each group and the number of groups in complex hierarchical models, such as Hierarchical Dirichlet process (HDP).

Wasserstein distance As in *Nguyen* (2013) it is useful to analyze the convergence of parameter estimation in mixture models using the notion of Wasserstein distance, which can be defined as the optimal cost of moving masses transforming one probability measure to another *Villani* (2008). Given two discrete measures $G = \sum\limits_{i=1}^{k} p_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^{k'} p'_i \delta_{\theta'_i}$, a coupling between $\boldsymbol{p}$ and $\boldsymbol{p'}$ is a joint distribution $\boldsymbol{q}$ on $[1 \ldots, k] \times [1, \ldots, k']$, which is expressed as a matrix $\boldsymbol{q} = (q_{ij})_{1 \leq i \leq k, 1 \ \leq j \leq k'} \in [0,1]^{k \times k'}$ with marginal probabilities $\sum\limits_{i=1}^{k} q_{ij} = p'_j$ and $\sum\limits_{j=1}^{k'} q_{ij} = p_i$ for any $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k'$. We use $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ to denote the space of all such couplings of $\boldsymbol{p}$ and $\boldsymbol{p'}$. For any $r \geq 1$, the r-th order Wasserstein distance between G and G' is given by

$$W_r(G, G') \quad = \quad \inf_{\boldsymbol{q} \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \left(\sum_{i,j} q_{ij} \|\theta_i - \theta'_j\|^r \right)^{1/r},$$

where $\| \cdot \|$ denotes the l_2 norm for elements in $\mathbb{R}^d$. It is simple to see that if a sequence of probability measures $G_n \in \mathcal{O}_{k_0}$ converges to $G_0 \in \mathcal{E}_{k_0}$ under the W_r metric at a rate $\omega_n = o(1)$ for some $r \geq 1$ then there exists a subsequence of G_n such that the set of atoms of G_n converges to the k_0 atoms of G_0, up to a permutation of the atoms, at the same rate ω_n.

3.3 Sieve methods via growing parameter space

Dirichlet Process mixture models place a base measure H as a prior distribution on Θ. In practice, there is no assumption made on Θ. However, to facilitate theoretical understanding of the behavior of the posterior, it is necessary to assume Θ to be bounded

(*Gao and van der Vaart* (2016); *Ghosal and van der Vaart* (2001)). However, since the range of atoms of G_0 is unknown , it is highly possible that the true mixing measure G_0 can possibly have some of its atoms lying outside the chosen parameter space Θ. If we choose Θ to be an unbounded set of $\mathbb{R}^d$, it is known that the posterior distribution of G will not be consistent, i.e., it does not concentrate around the true mixing measure G_0. A simple yet appealing solution is to allow Θ to expand with the sample size so that ultimately it will contain all the atoms of G_0. However, to understand the effect of expansion of parameter space, we need to carefully study the precise behavior of the Wasserstein distance between two mixing measures with respect to the hellinger distance. The following subsection deals with this.

3.3.1 Lower bound of Hellinger distance based on Wasserstein metric

Throughout this subsection, we assume that $\{f(\cdot|\theta), \theta \in \Theta)\}$ is a location family of density functions such that $f(x|\theta) = f(x-\theta)$ for all $\theta \in \Theta = [-\bar{\theta}, \bar{\theta}]^d$. We will utilize two important properties regarding Fourier transform of f (*Fan* (1991)):

Definition 3.3.1. Given $\{f(\cdot|\theta), \theta \in \Theta)\}$ is a location family of density functions such that $f(x|\theta) = f(x-\theta)$ for all $\theta \in \Theta$. Then, the following holds

(1) f has ordinary smooth property with two parameters α and β if we have

$$\inf_{\omega \in \mathbb{R}^d} |\widetilde{f}(\omega) \prod_{i=1}^{d} (1 + \alpha |\omega_i|^{\beta})| > 0.$$

(2) f has supermooth property with two positive parameters α and β if we have

$$\inf_{\omega \in \mathbb{R}^d} |\widetilde{f}(\omega) \prod_{i=1}^{d} \exp(\alpha |\omega_i|^{\beta})| > 0.$$

It is clear that standard univariate Gaussian or Cauchy function is a supermooth function while univariate Laplace function is an ordinary smooth function. The product Gaussian kernel is also Gaussian. Note that the same does not apply for the multivariate Laplace kernel given in Eq. (3.5). However, the following example shows that the product Laplace kernel is also supersmooth.

Example 3.3.1. Let $f_\sigma(\cdot|\theta)$ be a univariate Laplace kernel with standard deviation σ. Then the product Laplace kernel is given by

$$f_{d,\sigma_1,\dots,\sigma_d}(x|(\theta_1,\dots,\theta_d)) = \prod_{i=1}^{d} f_{\sigma_i}(x_i|\theta_i). \tag{3.3}$$

$\tilde{f}_{\sigma_i}(\omega_i) = \dfrac{1}{1+\sigma_i^2\omega_i^2}$, and thus $\tilde{f}_{d,\sigma_1,\dots,\sigma_d}(\omega|(\theta_1,\dots,\theta_d)) = \prod_{i=1}^{d} \dfrac{1}{1+\sigma_i^2\omega_i^2}$, and thus the product of univariate Laplace kernels is also ordinary smooth with $\alpha = \max_i \sigma_i^2$ and $\beta = 2$.

The following key result provides an upper bound for Wasserstein distance of location mixing measures with respect to the Hellinger distance for corresponding mixture densities.

Theorem 3.3.1. (Ordinary smooth density) Assume that f has the ordinary smooth property with two positive parameters α and β. Denote

$$\inf_{\omega\in\mathbb{R}^d} |\widetilde{f}(\omega)\prod_{i=1}^{d}(1+\alpha|\omega_i|^\beta)| = c_f > 0.$$

Then, the following holds

(a) If $\beta > 1 + 1/d$, then

$$W_1(G, G') \leq 3 \max\left\{\sqrt{2}d, \bar{\delta}_1^{\beta d} a_d, c_d \bar{\delta}_1 \bar{\theta}^d\right\}\left\{\log\left(\frac{\overline{C}}{h(p_G, p_{G'})}\right)\right\}^{\frac{1+d/2}{1+\beta d}} h(p_G, p_{G'})^{\frac{1}{1+\beta d}}$$

where $a_d = \dfrac{\pi^{d/4}}{\sqrt{(\frac{d}{2}+1)\Gamma(d/2)}} \cdot \dfrac{2^{d+1}\sqrt{\|f\|_\infty}}{c_f} \cdot \max\{1, (\alpha \exp(-\beta/2)\beta^{\beta/2})^d\}$, $c_d = \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{(8\pi e)^{d/2}}{\sqrt{d}}$, $\bar{\delta}_1^{(\beta-1)d-1} = \beta 2^{2+d} d + (\sqrt{d}\bar{\theta})^{\frac{(\beta-1)d-1}{2}}$, and $\overline{C} = \exp\Big((\beta d+1)(\sqrt{d}\bar{\theta} + \bar{\delta}_1^2)\Big)$.

(b) If $0 < \beta \leq 1 + 1/d$, then

$$W_1(G, G') \leq 3 \max\left\{\sqrt{2}d, \bar{\delta}_2^{\beta d} a_d, \bar{c}_d\right\}\left\{\log\left(\frac{\widetilde{C}}{h(p_G, p_{G'})}\right)\right\}^{\frac{1+d/2}{1+\beta d}} h(p_G, p_{G'})^{\frac{1}{1+\beta d}}$$

where $\bar{c}_d = 2^{d+1} d \exp\left(\dfrac{d\bar{\delta}_2^2}{2} + d\bar{\theta}\right)(\bar{\theta} + \bar{\delta}_2 + \bar{\delta}_2^2)$, $\bar{\delta}_2^{\beta d+1} = \beta d(1+d)^{1+d/2}$, and $\widetilde{C} = \exp(1+d)$.

The proof of Theorem 3.3.1 is provided in Section 3.7.1. The above result is an extension of similar well-known results for ordinary smooth kernels *Gao and van der Vaart* (2016) with careful consideration of the constants relative to the boundary of the parameter space. We have the following comments regarding the results of Theorem 3.3.1:

(i) The choices of $\overline{C}$ and $\widetilde{C}$ in part (a) and part (b) are to guarantee that the functions $\epsilon\Big\{\log(\overline{C}/\epsilon)\Big\}^{1+d/2}$ and $\epsilon\Big\{\log(\widetilde{C}/\epsilon)\Big\}^{1+d/2}$ are strictly increasing functions of $\epsilon \in (0, 1]$.

(ii) When d and α are fixed, the main difference between the upper bounds of $W_1(G, G')$ under two settings of β and d is the dependence of constants on $\bar{\theta}$. When $\beta > 1+1/d$, the dependence on $\bar{\theta}$ is at the order $\bar{\theta}^{\max\{\beta d/2, d+1/2\}}$. When $\beta \leq 1 + 1/d$, the dependence on $\bar{\theta}$ is at the order $\exp(d\bar{\theta})\bar{\theta}$.

When f is a supersmooth kernel we have the following result:

Theorem 3.3.2. (Supersmooth density) Assume that f has the supersmooth property with two positive parameters α and β. Denote

$$\inf_{\omega \in \mathbb{R}^d} |\widetilde{f}(\omega) \prod_{i=1}^{d} \exp(\alpha |\omega_i|^{\beta})| = c_f > 0.$$

Then, the following holds:

$$W_1(G, G') \leq a_d \left(\frac{4d\alpha}{\log(1/h(p_G, p_{G'}))} \right)^{1/\beta} \\ + a_d d h(p_G, p_{G'})^{1/2} \left(\max\left\{ \frac{4d\alpha}{\log(1/h(p_G, p_{G'}))}, \bar{\theta} \right\} \right)^{\left(\frac{2+d}{\beta(2+d/2)} \right)},$$

where $a_d = C \max\{d, \frac{\pi^{d/4}}{\sqrt{\Gamma(d/2+1)}}\}$, with C being a universal constant independent of $d, \bar{\theta}, \beta$ and α.

The proof is provided in Section 3.6.1.

The proofs of Theorems 3.3.1 and 3.3.2 rely on the choice of mollifiers which are smoother than the kernel f, but are as less smooth as possible, the less smooth the mollifier the stricter the hellinger bound. For ordinary smooth location mixtures we use a Gaussian mollifier. We expect a sharper mollifier to give a stronger bound for ordinary smooth location mixtures. However, the primary focus for this chapter is the appropriate expansion rate for the parameter space Θ, and a sharper mollifier does not seem to affect that drastically for the Laplace kernel. Therefore, we make no attempts to secure a sharper bound.

3.3.2 Growing parameter space

Sieve methods have been used to study contraction of estimators when the parameter space in question is large (*Shen and Wong* (1994); *Wong and Shen* (1995)). As pointed out earlier, there is a mismatch between the current theoretical and practical approaches to parameter estimation for Bayesian mixture models. This section aims to reduce that gap and improve understanding for posterior contraction when the parameter space in question is unbounded. Suppose without loss of generality the parameter space $\Theta = \mathbb{R}^d$. We consider the following adaptation of Eq. (3.2) for Dirichlet Process Mixture model:

$$\begin{aligned} G_n &\sim DP(\alpha, H_n), \\ \theta_1, \ldots, \theta_n &\overset{i.i.d.}{\sim} G_n, \\ X_i|\theta_i &\sim f(X_i|\theta_i), \quad \forall i = 1, \ldots, n. \end{aligned} \tag{3.4}$$

We make the following assumption on H_n.

(P.1) The base distribution H_n is supported on $\Theta_n = [-\bar{\theta}_n, \bar{\theta}_n]^d$, is absolutely continuous with respect to the Lebesgue measure μ on Θ_n and admits a density function $g_n(\cdot)$. Additionally, H_n is approximately uniform, i.e., $\min_{\theta \in \Theta_n} g_n(\theta) > \dfrac{c_0}{\mu(\Theta_n)} > 0$.

Θ_n approximates Θ so that $\Theta_n \uparrow \Theta$ as $n \uparrow \infty$. Moreover, the posterior contraction becomes tractable with Θ_n.

The rate of expansion, $\bar{\theta}_n$ plays an important role in maintaining the efficiency of estimating G_0. In particular, if we expand the parameter space Θ_n too fast, the posterior convergence rate of G_0 will become much slower. On the other hand, if we expand Θ_n too slowly, it may take a lot of samples to avoid the misspecified setting of parameter space even though the posterior convergence rate of G_0 only becomes slightly slower. As a consequence, we want to have a good trade-off between rate of expansion and earliest time that all the true atoms of G_0 belong to Θ_n. This section explores this solution via growing parameter space and provides appropriate rates of expansion of the parameter space.

To establish convergence rates of location mixtures under the growing parameter space setting, we utilize the general framework of posterior contraction of mixing measures under well-specified setting from *Nguyen* (2013). To state such results formally, we will need to introduce several key definitions in harmony with the notations in this chapter.

Let G be endowed with the prior distribution Π_n, given by Eq. (3.4), on a measure space of discrete probability measures in $\overline{\mathcal{G}}(\Theta)$. Fix $G_0 \in \mathcal{P}(\Theta)$ such that $\Theta_0 \subset [-\bar{\theta}_0, \bar{\theta}_0]^d$ is a bounded subset of $\mathbb{R}^d$ containing all the atoms of G_0..

For any mixing measure $G_1 \in \overline{\mathcal{G}}(\Theta)$ and $r > 0$, we define a Wasserstein ball centered at G_1 under W_1 metric as follows

$$B_{W_1}(G_1, r) = \left\{ G \in \overline{\mathcal{G}}(\Theta) : \ W_1(G, G_1) \leq r \right\}.$$

Note that, the choice of first order Wasserstein metric in the above formulation is due to the lower bound of Hellinger distance between mixing densities in terms of first order Wasserstein distance between their corresponding mixing measures in Theorems 3.3.1 and 3.3.2. We restrict our attention to consider the classes of Gaussian and Laplace location mixtures.

3.3.2.1 Laplace location mixtures

For Laplace location mixtures,the Laplace kernel with covariance Σ and dimension d is given by,

$$f_{\Sigma}^{L}(x|\theta) = \frac{1}{|\Sigma|(2\pi)^{d/2}} \frac{K_{(d/2)-1}\left(\sqrt{(x-\theta)^{\top}\Sigma^{-1}(x-\theta)}\right)}{\left(\sqrt{\lambda/2}\sqrt{(x-\theta)^{\top}\Sigma^{-1}(x-\theta)}\right)^{(d/2)-1}}, \tag{3.5}$$

where $|\Sigma|$ denotes the determinant of matrix Σ. Also, K_v is a Bessel function of the second kind of order v. This form of the multivariate Laplace distribution can be seen in *Eltoft et al.* (2006).

The characteristic function $\tilde{f}_{\Sigma}^{L}$ corresponding to f_{Σ}^{L} is given by $\tilde{f}_{\Sigma}^{L}(\omega) = \frac{1}{1+\omega'\Sigma\omega}$, thus clearly satisfying the ordinary smooth property with $\beta = 2$ and $\alpha = \lambda_{max}$. The next theorem provides contraction rates corresponding to expanding parameter space under the assumption that the parameter space contains the true atoms. As noted in Example 3.3.1 the results of the theorem are applicable for both location mixtures of multivariate Laplace distributions as well as for location mixtures of product of multivariate Laplace distributions. Without loss of generality, we prove the results for location mixtures of multivariate Laplace distributions only.

Theorem 3.3.3. Assume that Π_n is the Dirichlet process prior distribution with H_n, the base distribution satisfying condition (P.1). Also, assume that $\bar{\theta}_n \uparrow \infty$, $\epsilon_n \downarrow 0$, such that $\frac{\bar{\theta}_n^d}{\epsilon_n^{2d+2}} \log\left(\frac{\exp(\bar{\theta}_n)}{\epsilon_n^2}\right) = o(n)$ and $n\epsilon_n^2 \to \infty$. Then, if f_{Σ} is of the form of Eq. (3.5), the following holds.

$$\Pi_n\left(G : W_1(G, G_0) \gtrsim \bar{\theta}_n^{d+1/2}\epsilon_n^{1/1+2d}\left\{\log\left(\frac{\exp(a_{d,\lambda_{min}}\bar{\theta}_n)}{\epsilon_n}\right)\right\}^{\dfrac{1+d/2}{1+2d}} |X_{1:n}\right) \to 0 \tag{3.6}$$

in $P^n_{G_0}$ probability, where the constant a_d and the constant of proportionality associated with $\gtrsim$ are dependent on d, λ_{min} with λ_{min} being the smallest eigenvalue of Σ.

The proof of Theorem 3.3.3 is provided in Section 3.7.2 in the Appendix.

As a consequence of this theorem, we have the following corollary.

Corollary 3.3.1. Assume that Π_n is the Dirichlet process prior distribution with H_n, the base distribution satisfying condition (P.1) with $\bar{\theta}_n = o\left(\log\left(\frac{n}{(\log n)^{2d+2}}\right)\right)$. Then, if f_Σ is of the form of Eq. (3.5), the following holds.

$$\Pi_n\left(G : W_1(G, G_0) \gtrsim a_{d,\Sigma}\left(\frac{(\log n)^{(4d^2+5d+3)/(2+4d)}}{n^{1/(2+2d)(1+2d)}}\right)\middle| X_1, \ldots, X_n\right) \to 0 \tag{3.7}$$

in $P^n_{G_0}$ probability, where $a_{d,\Sigma}$ is a constant dependent on d, Σ.

Proof. We substitute $\bar{\theta_n} = o(\log(1/\epsilon_n))$ with $\epsilon_n = \frac{\log n}{n^{1/2d+2}}$ in the result of Theorem 3.3.3 to see that all the assumptions in Theorem 3.3.3 are satisfied. This choice of ϵ_n and $\bar{\theta_n}$ gives us the result. □

Corollary 3.3.1 tells us that when the truth G_0 is mixture of location Laplace kernels, the appropriate rate of expansion of the parameter space so as to also maintain a fast contraction rate to the truth is $o(\log(n))$. In this case the parameters contract at a polynomial rate faster than the rate obtained in *Nguyen* (2013). However, for $d = 1$, it is slightly slower than the rate obtained in *Gao and van der Vaart* (2016). The next subsection deals with the expansion rate for gaussian kernels.

3.3.2.2 Gaussian location mixtures

For the Gaussian location mixtures with covariance matrix Σ, the kernel, f^G_Σ, has the following form:

$$f^G_\Sigma(x|\theta) := \frac{\exp(-(x-\theta)^\top \Sigma^{-1}(x-\theta)/2)}{|2\pi\Sigma|^{-1/2}}. \tag{3.8}$$

Moreover, if $\tilde{f}^G_\Sigma$ gives the characteristic function corresponding to f^G_Σ, then it is well-known that $\tilde{f}^G_\Sigma(\omega) = \exp(-\omega'\Sigma\omega/2)$. Therefore it has supersmooth property with $\beta = 2$ and $\alpha = \lambda_{max}$, where λ_{max} is the maximum eigenvalue of Σ.

In that regard, the next theorem provides contraction rates of parameters corresponding to the expansion rates of the parameter space:

Theorem 3.3.4. Assume that Π_n is the Dirichlet process prior distribution with H_n, the base distribution satisfying condition (P.1). Also, assume that $\bar{\theta}_n \uparrow \infty$, $\epsilon_n \downarrow 0$, such that $\frac{\bar{\theta}_n^d}{\epsilon_n^{d+2}} \log\left(\frac{\bar{\theta}_n}{\epsilon_n}\right) = o(n)$ and $n\epsilon_n^2 \to \infty$. Then, if f_Σ is of the form of Eq. (3.8), the following holds.

$$\Pi_n\left(G : W_1(G, G_0) \gtrsim a_{d,\Sigma}\left(\left(\frac{4d}{\log(1/\epsilon_n)}\right)^{1/2} + d\epsilon_n^{1/2}\bar{\theta}_n^{\left(\frac{2+d}{4+d}\right)}\right)\Bigg| X_{1:n}\right) \to 0 \tag{3.9}$$

in $P^n_{G_0}$ probability, where $a_{d,\Sigma}$ is a constant dependent on d, $Sigma$.

As a consequence of this theorem, we have the following corollary.

Corollary 3.3.2. Assume that Π_n is the Dirichlet process prior distribution with H_n, the base distribution satisfying condition (P.1) with $\bar{\theta}_n = o\left(\frac{n^{(2d+4)/(3d^2+12d+2)}}{(\log n)^{(4+d)/(2+d)}}\right)$. Then, if f_Σ is of the form of Eq. (3.8), the following holds.

$$\Pi_n\left(G : W_1(G, G_0) \gtrsim a_{d,\Sigma}\left(\frac{\log\log n}{\log(n)}\right)^{1/2}\Bigg| X_1, \ldots, X_n\right) \to 0 \tag{3.10}$$

in $P^n_{G_0}$ probability, where $a_{d,\Sigma}$ is a constant dependent on d, Σ.

Proof. We substitute $\bar{\theta_n} = o\left(\frac{1}{\epsilon_n \log(1/\epsilon_n)}\right)$ with $\epsilon_n = \frac{\log n}{n^{(2d+4)/(3d^2+12d+8)}}$ in the result of Theorem 3.3.4 to see that all the assumptions in Theorem 3.3.4 are satisfied. This choice of ϵ_n and $\bar{\theta_n}$ gives us the result.

□

Corollary 3.3.2 says that for gaussian location mixtures the appropriate rate of expansion is polynomial as opposed to the logarithmic rate of expansion for Laplace mixtures. This might be seemingly counterintuitive, since Gaussian kernel being supersmooth has a very slow logarithmic contraction rate as opposed to the polynomial contraction rate for Laplace kernels. However, this result seems to be indicative of the fact that mixing measures contract to 0 much faster for Gaussian kernels in regions of the space away from any of the atoms of the true mixing measure. This seems to suggest that the slow contraction rate for Gaussian kernels arises out of the "coagulative" nature of Gaussian mixing measures- i.e., the components of the mixing measures aposteriori tend to cluster around the components of the true mixing measure while smaller components by weight are contributed by "empty" spaces. Laplace mixtures on the other hand do not have the "coagulative tendency" courtesy the heavy tail nature which enforces aposteriori components to be away from each other. Hence, the contributing weight for "open spaces" is no different from that of the neighborhoods of true atoms. As a result mixing measures for the Laplace kernel tend to have a larger bias of estimation and a smaller variance of estimation. The opposite is observed for Gaussian kernels, where the bias is small but the variance is large. This fact was also noted in Chapter II. The above discussion suggests the following recommendation for practitioners.

Recommendation:

(i) Fit a mixture of Gaussian kernels to obtain an estimate of the "locality" of representative atoms. This gives a crude knowledge of the domain of mixing atoms while remaining conservative in approach. This is because the Gaussian kernels have slow posterior contraction rates closer to the truth and convergence behavior closer to the true atoms remains slow. Hence, Gaussian kernels may not be appropriate to estimate the number of distinct components.

(ii) To obtain an estimate of the number of components or the parameters associated with them, fit mixture of Laplace or any heavy-tail kernel with prior mass of atoms constrained/upweighted in the representative regions.

3.4 Contraction of excess mass for Gaussian mixtures

The previous section shows that the sieve estimates can be chosen suitably for Gaussian location mixtures, to obtain appropriate contraction rates. In this section we build on the discussion of the previous section. We show that when fitting with a mixture of Gaussian densities, weights of mixing components away from the truth vanish at an almost polynomial rate. In order to achieve this result we make use of a novel metric which we define below. We call this the Orlicz-Wasserstein metric. It is a generalization of the Wasserstein distance corresponding to Orlicz norms.

The Orlicz norm is defined as follows:

Definition 3.4.1. Let μ be a $\sigma-$finite measure on a space X with metric $\|\cdot\|$. Assume that $\Phi : [0,\infty) \to [0,\infty)$ be a convex function satisfying:

(i) $\frac{\Phi(x)}{x} \to \infty, \quad \text{as } x \to \infty$

(ii) $\frac{\Phi(x)}{x} \to 0, \quad \text{as } x \to 0.$

Then the Orlicz space is defined as :

$$L_\Phi := \{f : X \to \mathbb{R} | \exists\ k \in \mathbb{R}^+ \text{ so that } \int\limits_X \Phi(\|f(x)\|/k)\, d\mu(x) \leq 1\}. \tag{3.11}$$

Moreover, the Orlicz norm corresponding to $f \in L_\Phi$ is given by:

$$\|f\|_\Phi := \inf\{k \in \mathbb{R}^+ : \int\limits_X \Phi(\|f(x)\|/k)\, d\mu(x) \leq 1\} \tag{3.12}$$

The Orlicz norm generalizes the concept of L_p-norm. A coupling between two probability measures ν_1 and ν_2 on a space X is a joint distribution on $X \times X$ with corresponding marginal distributions ν_1 and ν_2. Corresponding to the Orlicz norms, we define the Orlicz-Wasserstein metric as follows which generalizes the W_r-metric. Without loss of generalisation, we will assume $X = \mathbb{R}^d$, with $\|\cdot\|$ denoting the Euclidean metric.

Lemma 3.4.1. Let ν_1, ν_2 be probability measures on $(\mathbb{R}^d, \|\cdot\|)$. Assume that $\Phi : [0,\infty) \to [0,\infty)$ is a convex function satisfying conditions (i) and (ii) in Definition 3.4.1. Suppose we define

$$W_\Phi(\nu_1,\nu_2) := \inf_{\nu \in \mathcal{Q}(\nu_1,\nu_2)} \inf\{k \in \mathbb{R}^+ : \int\limits_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x-y\|/k)\, d\nu(x,y) \leq 1\}, \tag{3.13}$$

where $\mathcal{Q}(\nu_1,\nu_2)$ is the set of all possible couplings of ν_1 and ν_2. Then, W_Φ defines a distance metric on the set of probability measures on $(\mathbb{R}^d, \|\cdot\|)$. We call W_Φ the Orlicz-Wasserstein metric on $(\mathbb{R}^d, \|\cdot\|)$

Proof. We need to show the following:

(i) $W_\Phi(\nu_1, \nu_2) = W_\Phi(\nu_2, \nu_1)$ for any probability measures ν_1, ν_2 on $(\mathbb{R}^d, \|\cdot\|)$.

(ii) $W_\Phi(\mu, \mu) = 0$ for any probability measure μ on $(\mathbb{R}^d, \|\cdot\|)$.

(iii) $W_\Phi(\nu_1, \nu_2) \leq W_\Phi(\nu_1, \nu_3) + W_\Phi(\nu_3, \nu_2)$ for any probability measures ν_1, ν_2, ν_3 on $(\mathbb{R}^d, \|\cdot\|)$.

(i) follows easily from the fact $\|x - y\|$ is symmetric with respect to $x, y \in \mathbb{R}^d$.

For (ii) consider the coupling, $\nu(x, y) = \mu(x)\mathbb{1}_{x=y}$, then it is clear to see that for any $k > 0$, $\int_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x - y\|/k)\, d\nu(x, y) = 1$ and therefore $W_\Phi(\mu, \mu) = 0$.

For part (iii), assume that $W_\Phi(\nu_1, \nu_3) = k_1, W_\Phi(\nu_3, \nu_2) = k_2$. Then, it is enough to show that there exists a coupling ν of ν_1 and ν_2 such that $\int_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x - y\|/(k_1 + k_2))\, \mathrm{d}\nu(x, y) \leq 1$.

By results from *Villani* (2003, 2008), there exists a coupling μ_1 of ν_1 and ν_3 and a coupling μ_2 of ν_2 and ν_3 such that,

$$\begin{aligned} \int\limits_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x - z\|/k_1)\, d\mu_1(x, z) &\leq 1 \\ \int\limits_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|z - y\|/k_2)\, d\mu_2(y, z) &\leq 1. \end{aligned} \tag{3.14}$$

Then, by a result in probability theory there exists a probability measure μ on

$\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d$ such that

$$\begin{aligned} \int\limits_{x\in\mathbb{R}^d} \mu(\mathrm{d}x, y, z) &= \mu_2(y, z) \\ \int\limits_{x\in\mathbb{R}^d} \mu(x, \mathrm{d}y, z) &= \mu_1(x, z) \end{aligned} \tag{3.15}$$

Define $\nu(x, y) := \int_{z\in\mathbb{R}^d} \mu(x, y, \mathrm{d}z)$.

$$\begin{aligned} &\int\limits_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/(k_1+k_2))\,\mathrm{d}\nu(x, y) \\ &= \int\limits_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/(k_1+k_2))\,\mathrm{d}\mu(x, y, z) \\ &\leq \int\limits_{\mathbb{R}^d\times\mathbb{R}^d\times\mathbb{R}^d} \Phi((\|x-z\|+\|y-z\|)/(k_1+k_2))\,\mathrm{d}\mu(x, y, z) \\ &\leq \frac{k_1}{k_1+k_2} \int\limits_{\mathbb{R}^d\times\mathbb{R}^d} \Phi\left(\frac{\|x-z\|}{k_1}\right) \mathrm{d}\mu_1(x, z) \\ &+ \frac{k_2}{k_1+k_2} \int\limits_{\mathbb{R}^d\times\mathbb{R}^d} \Phi\left(\frac{\|y-z\|}{k_2}\right) \mathrm{d}\mu_2(y, z) \leq 1. \end{aligned} \tag{3.16}$$

The first inequality follows from the triangle inequality property of $\|\cdot\|$, while the second inequality follows from the convexity of Φ. □

Example 3.4.1. When $\Phi(x) = x^r$, then $W_\Phi(\nu_1, \nu_2) = W_r(\nu_1, \nu_2)$, the usual Wasserstein norm of order r.

The notion of Orlicz-Wasserstein distance is stronger than that of the usual Wasserstein distance to compare probability measures. This is formalized in the following lemma.

Lemma 3.4.2. Let ν_1, ν_2 be probability measures on $(\mathbb{R}^d, \|\cdot\|)$. Also assume Φ, Ψ are convex functions satisfying conditions (i) and (ii) in Definition 3.4.1.

(i) Suppose that for all $x > 0$, $\Phi(x) \leq \Psi(x)$, then

$$W_\Phi(\nu_1, \nu_2) \leq W_\Psi(\nu_1, \nu_2) \tag{3.17}$$

(ii) Let $k_{\delta,d}(x_1, \ldots, x_d) = \prod_{i=1}^d k_\delta(x_i)$, where $k_\delta(x) = c\frac{1}{\delta}(\int \exp(-itx/\delta)\exp(-t^4)\mathrm{d}t)^2$, with c being the constant of proportionality. Suppose $\Phi(x) \leq \exp((7/32)x^\alpha) - 1$ for some $0 < \alpha \leq 4/3$. Moreover, also assume $\nu_2 = \nu_1 * k_{\delta,d}$. Then

$$W_\Phi(\nu_1, \nu_2) \leq C_\alpha \delta, \tag{3.18}$$

for some constant $C_{\alpha,d}$.

Proof. (i) follows easily from the fact that for each k such that $\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/k)\, d\nu(x,y), \int_{\mathbb{R}^d\times\mathbb{R}^d} \Psi(\|x-y\|/k)\, d\nu(x,y) \leq \infty$, $\int_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/k)\, d\nu(x,y) \leq \int_{\mathbb{R}^d\times\mathbb{R}^d} \Psi(\|x-y\|/k)\, d\nu(x,y)$, and thus,

$$\int\limits_{\mathbb{R}^d\times\mathbb{R}^d} \Psi(\|x-y\|/k)\, d\nu(x,y) \leq 1 \implies \int\limits_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/k)\, d\nu(x,y) \leq 1.$$

For (ii), consider the case $\alpha \geq 1$.

Following the result from (i), it is enough to show $W_\Psi(\nu_1, \nu_2) \lesssim \delta^d$, where $\Psi(x) = \exp(x^\alpha) - 1$. Consider $X \sim \nu_1$ and $Y \sim k_{\delta,d}$. Then for k such that

$\int_{\mathbb{R}} \exp((7/32)|y_i/k|^{\alpha} - (7/16)|y_i/\delta|^{4/3})\mathrm{d}y_i < \infty$, the followin holds.

$$\begin{aligned}
&\inf\left\{ \int\limits_{\mathbb{R}^d\times\mathbb{R}^d} \Phi(\|x-y\|/k)\, d\mu(x,y) : \mu \in \mathcal{Q}(\nu_1,\nu_2)\right\} \\
\leq\ & \left(\frac{1}{\delta}\right)^d \int\limits_{\mathbb{R}^d} \exp((7/32)\|y\|^{\alpha}/k^{\alpha}) \prod_{i=1}^{d} k_1(y_i/\delta) \prod_{i=1}^{d} \mathrm{d}y_i - 1 \\
\leq\ & \prod_{i=1}^{d} \left(\frac{1}{\delta}\right) \int\limits_{\mathbb{R}} \exp((7/32)|y_i|^{\alpha}/k^{\alpha}) k_1(y_i/\delta)\mathrm{d}y_i - 1 \\
=\ & \prod_{i=1}^{d} \left(\frac{1}{\delta}\right) \int\limits_{\mathbb{R}} \phi(y_i)^2 \exp((7/32)|y_i/k|^{\alpha} - (7/16)|y_i/\delta|^{4/3})\mathrm{d}y_i - 1,
\end{aligned}$$

where $\phi(\cdot)$ is the function in Lemma 3.4.4. The second inequality follows from the fact that $\|x\|_p \leq \|x\|_q$ when $p \geq q$, where $\|\cdot\|_p$ is the L_p norm. The final equality follows from Lemma 3.4.4.

Now, as $|\phi(x)| \leq C_{\phi}$ for some constant $C_{\phi} < \infty$, we have following the result in part (i),

$$W_{\Phi}(\nu_1,\nu_2) \leq C_{\alpha}\delta$$

where

$$C_{\alpha} = \inf\left\{ k > 0 : \int\limits_{\mathbb{R}} \exp(|y/k|^{\alpha} - |y|^{4/3})\mathrm{d}y - 1 \leq \frac{1}{C_{\phi}^2}\right\}.$$

C_{α} as defined above exists because $\alpha \leq 4/3$. □

Wasserstein distances are useful to quantify contraction behaviour of mixture distributions. Namely, if one mixing measure is close to another in Wasserstein distance, it provides a way to control the corresponding contraction rates of the atoms and the masses associated with them. The following lemma provides a similar result for Orlicz-Wasserstein norms.

Lemma 3.4.3. Let $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$, $G = \sum_{j=1}^{k} p_i \delta_{\theta_i}$ be mixing measures such that $\theta_j, \theta_i^0 \in \mathbb{R}^d$ for all i, j. Assume that $\Phi : [0, \infty) \to [0, \infty)$ is a convex function satisfying conditions (i) and (ii) in Definition 3.4.1. Then

$$\sum_j p_j \mathbb{1}_{\|\theta_j - \theta_i^0\| > \eta \text{ for all } i} \leq \left(\Phi \left(\frac{\eta}{W_\Phi(G, G_0)} \right) \right)^{-1}. \tag{3.19}$$

Here, k_0, k can also take the value ∞.

Proof. Suppose $\boldsymbol{q} = (q_{ij})_{1 \leq i \leq k_0, 1\ \leq j \leq k} \in [0,1]^{k_0 \times k}$ is a coupling between $\boldsymbol{p_0} = (p_1^0, \ldots, p_{k_0}^0)$ and $\boldsymbol{p} = (p_1, \ldots, p_k)$, with $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ representing the space of all such couplings of $\boldsymbol{p}$ and $\boldsymbol{p'}$.

Then, for k fixed,

$$\sum q_{ij} \Phi(\|\theta_i^0 - \theta_j\| / k) \geq \sum q_{ij} \mathbb{1}_{\|\theta_i^0 - \theta_j\| \geq \eta} \Phi(\eta / k) \geq \sum p_j \mathbb{1}_{\|\theta_i^0 - \theta_j\| \geq \eta \text{ for all } i} \Phi(\eta / k). \tag{3.20}$$

Let $K = \inf\{k \geq 0 : \sum p_j \mathbb{1}_{\|\theta_i^0 - \theta_j\| \geq \eta \text{ for all } i} \Phi(\eta/k) \leq 1\}$. Then,

$$K \geq \eta \left(\Phi^{-1} \left(\frac{1}{\sum p_j \mathbb{1}_{\|\theta_i^0 - \theta_j\| \geq \eta \text{ for all } i}} \right) \right)^{-1}, \tag{3.21}$$

where Φ^{-1} is the functional inverse to Φ. This exists and is concave as Φ is monotonic increasing and convex. Moreover, by Lemma 3.4.2(i),

$$W_\Phi(G, G_0) := \inf_{q \in \mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})} \inf\{k \geq 0 : \sum q_{ij} \Phi(\|\theta_i^0 - \theta_j\|/k) \leq 1\} \geq K \tag{3.22}$$

Combining Eqs. (3.32) and (3.22) we get the result. □

Previously, we have shown that the excess mass for atoms of G which are away from any atom of G_0 is well controlled by the Orlicz-Wasserstein norm. The next theorem obtains a contraction rate for the excess mass for Gaussian location mixtures.

The proof of Theorem 3.4.1 depends on certain key technical lemmas which we prove prior to stating the main theorem.

Lemma 3.4.4. Let $f(x) = \exp(-x^4)$, and $\tilde{f}(t) = (1/2\pi) \int_{-\infty}^{\infty} \exp(-itx) f(x) \mathrm{d}x$. Then,

$$|\tilde{f}(t)| \leq \phi(t) \exp(-7/32|t|^{4/3}), \tag{3.23}$$

where $\phi(t)$ is an absolutely bounded real-valued function.

Proof. Consider a rectangle on the complex plane, with vertices at $R, -R, R + i\zeta, -R + i\zeta$ respectively. Following Goursat's Theorem (*E.M.Stein and Shakarchi* (2010)) for integration along rectangular contours on the complex plane, the contour integral along a closed rectangle is 0.

Therefore,

$$\int_{-R}^{R} \exp(-itx)f(x)\mathrm{d}x \quad + \quad \int_{R}^{R+i\zeta} \exp(-itx)f(x)\mathrm{d}x$$
$$+ \quad \int_{-R+i\zeta}^{-R} \exp(-itx)f(x)\mathrm{d}x + \int_{R+i\zeta}^{-R+i\zeta} \exp(-itx)f(x)\mathrm{d}x = 0.$$

Now,

$$|\int_{R}^{R+i\zeta} \exp(-itx)f(x)\mathrm{d}x| = |\int_{0}^{\zeta} \exp(itR - tx)f(R+ix)i\mathrm{d}x| \leq C\exp(-R^4) \to 0,$$

as $R \to \infty$. Similarly,

$$|\int_{-R+i\zeta}^{-R} \exp(-itx)f(x)\mathrm{d}x| \to 0,$$

as $R \to \infty$.

Therefore,

$$\lim_{R\to\infty} \int_{-R+i\zeta}^{R+i\zeta} \exp(-itx)f(x)\mathrm{d}x = \lim_{R\to\infty} \int_{-R}^{R} \exp(-itx)f(x)\mathrm{d}x = 2\pi\tilde{f}(t).$$

Now,

$$\begin{aligned}
\lim_{R\to\infty}\int_{-R}^{R} \exp(-itx)f(x)\mathrm{d}x = 2\pi\tilde{f}(t) &= \lim_{R\to\infty}\int_{-R+i\zeta}^{R+i\zeta} \exp(-itx)f(x)\mathrm{d}x \\
&= \lim_{R\to\infty}\int_{-R}^{R} \exp(it(x+i\zeta))f(x+i\zeta)\mathrm{d}x. \\
&= \lim_{R\to\infty}\int_{-R}^{R} \exp(-itx-t\zeta))\exp(-(x+i\zeta)^4)\mathrm{d}x.
\end{aligned}$$

Expanding the above expression,

$$\tilde{f}(t) = (1/2\pi)\lim_{R\to\infty}\int_{-R}^{R} \exp(-itx - 4ix^3\zeta + 4ix\zeta^3 - t\zeta - (x^2-3\zeta^2)^2 + 8\zeta^4)\mathrm{d}x.$$

Substituting $\zeta = \dfrac{1}{4}\,\mathrm{sign}(t)|t|^{1/3}$ in the above equationa,

$$|\tilde{f}(t)| \leq (1/2\pi)\exp(-(7/32)|t|^{4/3})\int_{-\infty}^{\infty} \exp(-(x^2-(1/3)|t|^{1/2})^2)\mathrm{d}x. \tag{3.24}$$

The proof is complete when we note that $\phi(t) = (1/2\pi)\int_{-\infty}^{\infty}\exp(-(x^2-(1/3)|t|^{1/2})^2)\mathrm{d}x$ is an absolutely bounded function. □

Lemma 3.4.5. Let $k(t) = c\tilde{f}(t)^2$, where $\tilde{f}(t) = (1/2\pi)\int_{-\infty}^{\infty}\exp(-itx)\exp(-x^4)\mathrm{d}x$ and c is a constant of proportionality so that $\int_{-\infty}^{\infty}k(t)\mathrm{d}t = 1$. Then,

$$|\int_{-\infty}^{\infty}\exp(itx)k(t)\mathrm{d}t| \lesssim \exp(-(x/2)^4) \tag{3.25}$$

Proof. Define $f(x) = \exp(-x^4)$. Then, by a version of the Fourier inversion theorem,

$$\int_{-\infty}^{\infty} \exp(itx)k(t)\mathrm{d}t = f * f(x),$$

where $*$ is the convolution operator. Since convolution of even functions is even, it is enough to show the result for $x > 0$. Then,

$$\begin{aligned} f * f(x) &= \int_{-\infty}^{\infty} \exp(-y^4)\exp(-(y-x)^4)\mathrm{d}y \\ &= \int_{x/2}^{\infty} \exp(-y^4)\exp(-(y-x)^4)\mathrm{d}y + \int_{-\infty}^{x/2} \exp(-y^4)\exp(-(y-x)^4)\mathrm{d}y \\ &\leq \exp(-(x/2)^4)\int_{x/2}^{\infty} \exp(-(y-x)^4)\mathrm{d}y + \exp(-(x/2)^4)\int_{-\infty}^{x/2} \exp(-y^4)\mathrm{d}y \\ &\leq 2\exp(-(x/2)^4)\int_{-\infty}^{\infty} \exp(-y^4)\mathrm{d}y. \end{aligned} \tag{3.26}$$

The result holds with $C = 2\int_{-\infty}^{\infty} \exp(-y^4)\mathrm{d}y$ since $\int_{-\infty}^{\infty} \exp(-y^4)\mathrm{d}y < \infty$. □

Lemma 3.4.6. Let ν_1, ν_2 be probability measures on $(\mathbb{R}^d, \|\cdot\|)$ and let Φ be a convex function satisfying conditions (i) and (ii) in Definition 3.4.1. Then,

$$W_\Phi(\nu_1, \nu_2) \leq 2\inf\{k \in \mathbb{R}^+ : \int_{\mathbb{R}^d} \Phi(\|x\|/k)\, d|\nu_1(x) - \nu_2(x)| \leq 1\}. \tag{3.27}$$

Proof. Consider a coupling, ν between ν_1 and ν_2 that keeps fixed all the mass shared between ν_1 and ν_2, and redistributes the remaining mass independently, i.e.,

$$\nu(x,y) = (\nu_1(x) \bigwedge \nu_2(y))\mathbb{1}_{x=y} + \frac{1}{(\nu_1 - \nu_2)_+(\mathbb{R}^d)}(\nu_1(x) - \nu_2(x))_+(\nu_2(y) - \nu_1(y))_+ \tag{3.28}$$

Assume that $k_0 := \inf\{k \in \mathbb{R}^+ : \int_{\mathbb{R}^d} \Phi(\|x\|/k)\, d|\nu_1(x) - \nu_2(x)| \leq 1\}$. Then, using ν as defined in the above display we get

$$\begin{aligned}
& \int\limits_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x-y\|/2k_0)\, d\nu(x,y) \\
= & \int\limits_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x-y\|/2k_0) \frac{1}{(\nu_1 - \nu_2)_+(\mathbb{R}^d)}(\nu_1(x) - \nu_2(x))_+(\nu_2(y) - \nu_1(y))_+ \\
\leq & \int\limits_{\mathbb{R}^d \times \mathbb{R}^d} \Phi(\|x\|/k_0)(\nu_1(x) - \nu_2(x))_+ \leq 1
\end{aligned}$$

Therefore,

$$W_\Phi(\nu_1, \nu_2) \leq 2 \inf\{k \in \mathbb{R}^+ : \int\limits_{\mathbb{R}^d} \Phi(\|x\|/k)\, d|\nu_1(x) - \nu_2(x)| \leq 1\}.$$

□

Now, we are ready to state the main theorem of this section.

Theorem 3.4.1. Assume that f is such that $f(x|\theta) = f(x - \theta) \sim \mathcal{N}(\theta, \Sigma)$ for all $\theta \in \Theta = [-\bar{\theta}, \bar{\theta}]^d$. Also assume that Φ is a convex function satisfying conditions (i) and

(ii) in Definition 3.4.1 such that $\Phi(x) \leq \exp(x) - 1$. Then for mixing measures G, G',

$$\begin{aligned} W_\Phi(G, G') &\leq C_d \biggl(\frac{\bar{\theta}^{5/4}}{(\log(1/h(p_G, p_{G'})))^{1/8}} + \biggl(\frac{1}{\log(1/h(p_G, p_{G'}))} \biggr)^{11/8} \\ &+ \biggl(\frac{1}{\log(c/h(p_G, p_{G'})(\log(1/h(p_G, p_{G'})))^{d/4})} \biggr)^{1/2} \biggr) \qquad (3.29) \end{aligned}$$

for some constant $C_{d,\Sigma}$ dependent on the dimension d, Σ.

The proof of Theorem 3.4.1 is provided in Section 3.6.3.

As a consequence of this theorem we have the following corollary.

Corollary 3.4.1. Assume that $\epsilon_n \downarrow 0$, such that $\epsilon_n = \dfrac{(\log(n))^2}{n^{1/(d+2)}}$. Also assume that Π_n is the Dirichlet process prior distribution with H_n, the base distribution satisfying condition (P.1) with $\bar{\theta}_n = \bar{\theta}_0$, with $\Theta_0 \subset [-\bar{\theta}_0, \bar{\theta}_0]^d$ being bounded subset of $\mathbb{R}^d$ containing all the atoms of G_0. Then, if f_Σ is of the form of Eq. (3.8), the following holds for any $\eta > 0$.

$$\begin{aligned} \Pi_n \biggl(G &= \sum p_i \delta_{\theta_i} \in \overline{\mathcal{G}}(\Theta_n) : \sum_j p_j \mathbb{1}_{\|\theta_j - \theta_i^0\| > \eta \text{ for all } i} \\ &\geq 2 \exp\left(\frac{-\eta \log(n/(\log n)^{2d+4})^{1/8}}{(d+2)} \right) |X_1, \ldots, X_n \biggr) \overset{P_{G_0}}{\to} 0. \qquad (3.30) \end{aligned}$$

in $P_{G_0}^n$ probability.

Proof. By the proof technique of Theorem 3.3.4, we get for some sufficiently large L.

$$\Pi_n(G \in \overline{\mathcal{G}}(\Theta_n) : h(p_G, p_{G_0}) \geq \frac{L(\log(n))^2}{n^{1/(d+2)}} |X_1, \ldots, X_n) \overset{P_{G_0}}{\to} 0. \qquad (3.31)$$

From Theorem 3.4.1, we have,

$$\Pi_n \left(G \in \overline{\mathcal{G}}(\Theta_n) : W_\Phi(G, G_0) \geq \frac{1}{(\log(n/(\log n)^{2d+4})/(d+2)))^{1/8}} | X_1, \ldots, X_n \right) \overset{P_{G_0}}{\to} 0.$$

Let $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$, $G = \sum_{j=1}^{k} p_i \delta_{\theta_i}$ Suppose $\boldsymbol{q} = (q_{ij})_{1 \leq i \leq k_0, 1 \ \leq j \leq k} \in [0,1]^{k_0 \times k}$ is a coupling between $\boldsymbol{p_0} = (p_1^0, \ldots, p_{k_0}^0)$ and $\boldsymbol{p} = (p_1, \ldots, p_k)$, with $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ representing the space of all such couplings of $\boldsymbol{p_0}$ and $\boldsymbol{p}$.

Using the proof technique similar to Lemma 3.4.3, we get

$$\sum q_{ij} \exp(\|\theta_i^0 - \theta_j\|/k) \geq \sum q_{ij} \mathbb{1}_{\|\theta_i^0 - \theta_j\| \geq \eta} \exp(\eta/k) \geq \sum p_j \mathbb{1}_{\|\theta_i^0 - \theta_j\| \geq \eta \text{ for all } i} \exp(\eta/k).$$

Let $K = \inf\{k \geq 0 : \sum p_j \mathbb{1}_{\|\theta_i^0 - \theta_j\| \geq \eta \text{ for all } i} \exp(\eta/k) \leq 2\}$. Then,

$$K \geq \eta \left(\log \left(\frac{1}{\sum p_j \mathbb{1}_{\|\theta_i^0 - \theta_j\| \geq \eta \text{ for all } i}} \right) \right)^{-1},$$

$$\sum_j p_j \mathbb{1}_{\|\theta_j - \theta_i^0\| > \eta \text{ for all } i} \leq 2 \exp\left(\frac{-\eta}{W_\Phi(G, G_0)} \right).$$

Therefore,

$$\begin{aligned} \Pi_n \Bigg(G \ &= \ \sum p_i \delta_{\theta_i} \in \overline{\mathcal{G}}(\Theta_n) : \sum_j p_j \mathbb{1}_{\|\theta_j - \theta_i^0\| > \eta \text{ for all } i} \\ &\geq \ 2 \exp\left(\frac{-\eta \log(n/(\log n)^{2d+4})^{1/8}}{(d+2)} \right) | X_1, \ldots, X_n \Bigg) \overset{P_{G_0}}{\to} 0. \qquad (3.32) \end{aligned}$$

in $P_{G_0}^n$ probability. □

Corollary 3.4.1 suggests an almost polynomial rate of estimation of mixing measures in regions of the space away from the true atoms. We believe the rate of contraction can be optimized further with a more refined choice of $\Phi(\cdot)$ than the one mentioned in Theorem 3.4.1, however, we make no such attempts in this work. The Orlicz-Wasserstein metric penalizes the coupling between outlying points more heavily than those in the central region as a result providing an improved bound for mixture of Gaussian kernels. However, it does not improve on the bound for Laplace location mixtures.

3.5 Simulation Studies

This section provides simulations to compare the effect between fitting with mixtures of heavy tail distributions compared to fitting with light-tailed mixtures. We consider a simple simulation setup.

We compare how the contraction of excess mass behavior compares if we fit with light-tailed kernels such as Gaussian as opposed to fitting with heavy tailed kernels such as multivariate Student's t distribution.

We consider two settings. In the left panels of Figures 3.3 and 3.6 the data is generated by mixtures of three location Gaussian distributions:

$$p_{G_0}(\cdot) = \sum_{i=1}^{3} p_i^0 \mathcal{N}(\cdot|\mu_i^0, \Sigma^0)$$

where $\mathcal{N}(\cdot|\mu, \Sigma)$ is the Gaussian distribution with mean vector μ and covariance matrix Σ. We consider 2-dimensional distributions. On the other hand for the right panels of Figures 3.3 and 3.6, the data is generated from a mixture of Multivariate Student's t distribution of dimension 2.

$$p_{G_0}(\cdot) = \sum_{i=1}^{3} p_i^0 \mathcal{T}(\cdot|\mu_i^0, \Sigma^0).$$

For both the settings $\mu_1^0 = (0.8, 0.8)$, $\mu_2^0 = (0.8, -0.8)$, $\mu_3^0 = (-0.8, 0.8)$, $\Sigma^0 = 0.7I_2$, $n = 3500$ where n is the sample size. Here, I_2 is the identity matrix of dimension 2. Additionally, the weight vector for all these cases is chosen as $p^0 = (p_1^0, p_2^0, p_3^0) = (0.4, 0.3, 0.3)$.

A Dirichlet Process prior is chosen with a Gaussian base measure H with mean $mu = (0, 0)$ and covariance $\Sigma = 2.5I_2$ is chosen as the prior, along with concentration parameter $\alpha = 2$. This choice of prior enables us to sample significantly larger numbers of components of the mixing measure than the true number of three components and also allows us to choose more components away from the atoms of the true mixing components. .

For the DP mixture's posterior computation, we make use of the slice sampler of Walker *S.Walker* (2007); *J.Griffin and S.Walker*; *Muller et al.* (2015). The sampler had 10000 burn-in iterations followed by 20000 sample iterations (a total 30000), with each 10th iteration being counted.

This yielded sample mixing measures for the Dirichlet Posterior. We considered the mean of the posterior and smoothed it out with a Gaussian kernel of Covariance $5I_2$. The smoothing was done to make the difference between the contour plots of the posterior means apparent. The results of the experiments are denoted in Figures 3.3 and 3.6 respectively. It is clear that the contraction of excess mass when fitting with Gaussian kernel is comparable if only slightly slower than that obtained by fitting with a heavy tailed kernel. As verified by our theoretical analysis, the excess mass contracts

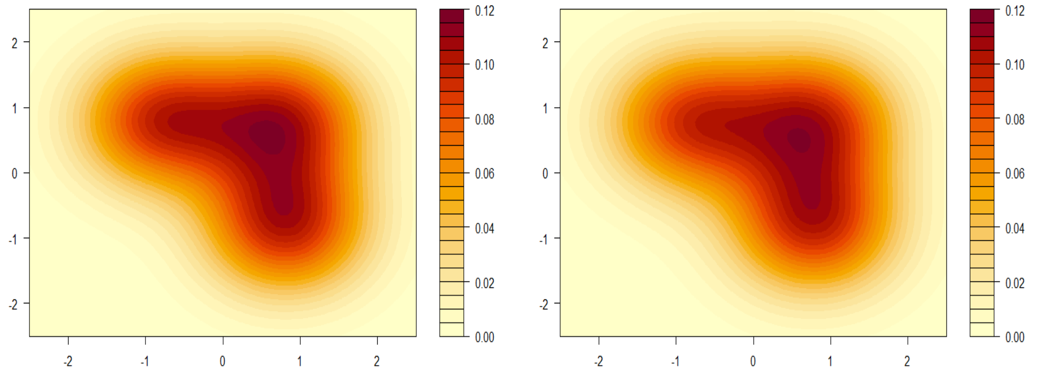


Figure 3.1: *truth is mixture of light-tail kernel, fit with mixture of light-tail.*

Figure 3.2: *truth is mixture of heavy-tail kernel, fit with mixture of light-tail*

Figure 3.3: *Fitting with Gaussian*

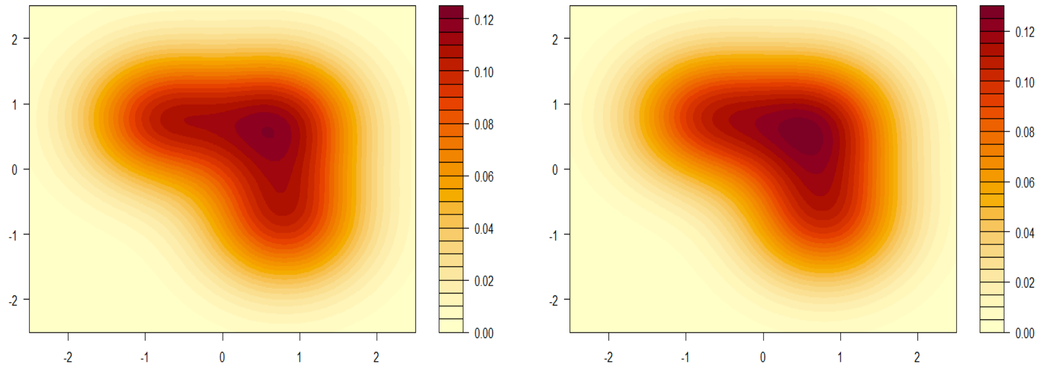


Figure 3.4: *truth is mixture of light-tail kernel, fit with mixture of heavy-tail*

Figure 3.5: *truth is mixture of heavy-tail kernel, fit with mixture of heavy-tail*

Figure 3.6: *Fitting with Student's t*

at polynomial rate for ordinary smooth kernels and slightly slower than polynomial for Gaussian kernels, although not as slow as logarithmic as was initially thought. Moreover, we can also note that the contour plots indicate a larger maximum value of the contour posterior mean density for multivariate T distribution as opposed to Gaussian. This corroborates the results obtained in the chapter.

3.6 Appendix A: Proofs

In this section, we provide the proofs for several results in the chapter.

3.6.1 Proof of Theorem 3.3.2

We have the following notations. $a \lesssim b$ for this proof implies $a \leq C \cdot b$ for a universal constant C independent of α, β, d, and $\bar{\theta}$. Also, $f * g$ will denote the outcome of convolution operation on functions f and g.

Consider the following density function in $\mathbb{R}$:

$$k(x) := \frac{96}{\pi}\left(\frac{\sin(x/4)}{x}\right)^4. \tag{3.33}$$

This density function has been used by *Caillerie et al.* (2011) to obtain bounds for Wasserstein distances of mixing measures. They show that the characteristic function corresponding to $k(\cdot)$ is given by:

$$\tilde{k}(\omega) = 3g(4|\omega|)/16, \tag{3.34}$$

where $g(\omega) = (\omega^3/2 - 2\omega^2 + 16/3)\mathbb{1}_{[0,2[}(\omega) + (-\omega^3/6 + 2\omega^2 - 8\omega + 32/3)\mathbb{1}_{[2,4[}(\omega)$. Notably, $\tilde{k}(\cdot)$ is a continuous twice differentiable symmetric function with Lipschitz second derivative and has support in $[-1, 1]$.

Our strategy to obtain upper bounds for $W_1(G, G')$ is to convolve G with mollifiers, $k_{\delta,d}(\cdot)$, of the form $k_{\delta,d}(x) = \prod_{i=1}^{d} \frac{1}{\delta} k(x_i/\delta)$ for $\delta > 0$, where $x := (x_1, \ldots, x_d)$.

By triangle inequality, we can write:

$$W_1(G, G') \leq W_1(G, G * k_{\delta,d}) + W_1(G', G' * k_{\delta,d}) + W_1(G * k_{\delta,d}, G' * k_{\delta,d}).$$

Consider a coupling of G and $G * k_{\delta,d}$, given by $(\theta, \theta+\epsilon)$, with θ, ϵ being independent with marginals G and $k_{\delta,d}$, respectively.

Then,

$$W_1(G, G * k_{\delta,d}) \leq \mathbb{E}[\|\epsilon + \theta - \theta\|] \leq \sqrt{\mathbb{E}[\|\epsilon\|^2]} \leq C\sqrt{d}\delta,$$

where C is a constant indepedent of d, δ. The last inequality above can be seen by evaluating the second derivative of the characteristic function $\tilde{k}(\cdot)$ at $t = 0$. Therefore, we can write

$$W_1(G, G') \leq C\sqrt{d}\delta + W_1(G * k_{\delta,d}, G' * k_{\delta,d}),$$

for a different constant C.

For every $M > 0$, we have,

$$\begin{aligned} W_1(G * k_{\delta,d}, G' * k_{\delta,d}) &\leq \int \|x\| \cdot |(G - G') * k_{\delta,d}(x)| dx \\ &= \underbrace{\int\limits_{\|x\|_2 \leq M} \|x\| \cdot |(G - G') * k_{\delta,d}(x)| dx}_{s_1} \\ &+ \underbrace{\int\limits_{\|x\|_2 > M} \|x\| \cdot |(G - G') * k_{\delta,d}(x)| dx,}_{s_2} \end{aligned}$$

with the first inequality following from Theorem 6.15 in *Villani* (2008).

Bounding s_1: Using Hölder's inequality, we obtain

$$\begin{aligned} s_1 &= \int\limits_{\|x\|\leq M} \|x\| \cdot |(G-G') * \phi_\delta(x)| dx \\ &\leq \Bigg(\int\limits_{\|x\|\leq M} \|x\|^2 dx \Bigg)^{1/2} \Bigg(\int\limits_{\|x\|\leq M} |(G-G') * \phi_\delta(x)|^2 dx \Bigg)^{1/2} \\ &= \frac{\pi^{d/4}}{\sqrt{(\frac{d}{2}+1)\Gamma(d/2)}} M^{1+d/2} \|(G-G') * \phi_\delta\|_2 \\ &\leq \frac{\pi^{d/4}}{\sqrt{\Gamma(d/2+1)}} M^{1+d/2} \|(G-G') * \phi_\delta\|_2. \end{aligned} \tag{3.35}$$

Let g_δ be defined as :

$$g_\delta(x) := \frac{1}{2\pi} \int\limits_{\mathbb{R}^d} e^{i\langle t,x \rangle} \frac{\tilde{k}_{\delta,d}(\omega)}{\tilde{f}(\omega)}$$

By a generalized Minkoswki's inequality, it is known that $\|\mu * g\|_2 \leq |\mu| \|g\|_2$ for a signed measure μ with total-variation $|\mu|$ and an $L_2(\mathbb{R}^d)$-integrable function g. Therefore,

$$\|(G-G') * k_{\delta,d}\|_2^2 = \|p_G * g_\delta - p_{G'} * g_\delta)\|_2^2 \leq 4V^2(p_G, p_{G'}) \|g_\delta\|_2^2 \leq 8h^2(p_G, p_{G'}) \|g_\delta\|_2^2$$

Using Plancherel's identity,

$$\|g_\delta\|_2^2 = \frac{1}{(2\pi)^d} \int \frac{\tilde{k}_{\delta,d}^2(\omega)}{\tilde{f}^2(\omega)} \mathrm{d}\omega \leq C\pi^{-d} \int\limits_{[-1/\delta, 1/\delta]^d} \frac{1}{\tilde{f}^2(\omega)} \mathrm{d}\omega.$$

The last inequality follows because $\tilde{k}_{\delta,d}(\cdot)$ is bounded with support in $[-1/\delta, 1/\delta]^d$.

Along with the condition on supersmooth densities, this leads us to the following bound on s_1.

$$s_1 \leq C\frac{\pi^{-d/4}}{\sqrt{\Gamma(d/2+1)}}M^{1+d/2}h(p_G, p_{G'})\exp(2d\alpha\delta^{-\beta}) \tag{3.36}$$

Bounding s_2: For $M > 0$,

$$\begin{aligned}
s_2 &\leq M^{-1}\int\limits_{\|x\|_2>M} \|x\|_2^2 \cdot |(G-G') * k_{\delta,d}(x)|dx \\
&\leq M^{-1}\int \|x\|_2^2 \cdot |(G-G') * k_{\delta,d}(x)|dx \\
&\leq M^{-1}\left(\int \|x\|_2^2 \cdot (G * k_{\delta,d})(x)dx + \int \|x\|_2^2 \cdot (G' * k_{\delta,d})(x)dx\right) \\
&\leq 2M^{-1}\Bigg(\int \|u\|_2^2 \cdot G(u)du + 2\int \|v\|_2^2 \cdot k_{\delta,d}(v)dv \\
&\qquad + \int \|w\|_2^2 \cdot G'(w)dw\Bigg) \qquad (3.37)
\end{aligned}$$

Now, $\int \|w\|_2^2 \cdot G'(w)dw \leq d\bar{\theta}^2$, and $\int \|v\|_2^2 \cdot k_{\delta,d}(v)dv \leq d\delta^2 \int\limits_{x\in\mathbb{R}} |z|^2 \cdot k(z)dz$.

$\int\limits_{x\in\mathbb{R}} |z|^2 \cdot k(z)dz$ is the variance pertaining to the density $k(\cdot)$ and is therefore a constant that can be computed from the second derivative of the characteristic function, $\tilde{k}(\cdot)$ at 0. Therefore, using these results we get

$$s_2 \lesssim M^{-1}d(\bar{\theta}^2 + \delta^2) \tag{3.38}$$

Combining the results of Eq. (3.36) and Eq. (3.38), we get

$$W_1(G, G') \lesssim a_d(\delta + M^{-1}(\bar{\theta}^2 + \delta^2) + M^{1+d/2}h(p_G, p_{G'})\exp(2d\alpha\delta^{-\beta})),$$

where $a_d = \max\{d, \dfrac{\pi^{d/4}}{\sqrt{\Gamma(d/2+1)}}\}$.

Limiting attention to $\delta < \bar{\theta}$, we have:

$$W_1(G, G') \lesssim a_d(\delta + M^{-1}\bar{\theta}^2 + M^{1+d/2}h(p_G, p_{G'})\exp(2d\alpha\delta^{-\beta})). \tag{3.39}$$

Differentiating the RHS of Eq. (3.39) with respect to M, we can choose

$$M = \left(\frac{\bar{\theta}^2}{(1+d/2)h(p_G, p_{G'})\exp(2d\alpha\delta^{-\beta})}\right)^{1/(2+d/2)}.$$

With regards to δ we choose $\delta = \left(\dfrac{4d\alpha}{\log(1/h(p_G, p_{G'}))}\right)^{1/\beta}$. Then we get,

$$W_1(G, G') \lesssim a_d\left(\left(\frac{4d\alpha}{\log(1/h(p_G, p_{G'}))}\right)^{1/\beta} + (2+d/2)h(p_G, p_{G'})^{1/2}(\bar{\theta})^{\left(\frac{2+d}{2+d/2}\right)}\right).$$

On the other hand, when $\delta > \bar{\theta}$, we get,

$$\begin{aligned} W_1(G, G') \ \lesssim \ a_d\bigg(& \left(\frac{4d\alpha}{\log(1/h(p_G, p_{G'}))}\right)^{1/\beta} \\ & + (2+d/2)h(p_G, p_{G'})^{1/2}\left(\frac{4d\alpha}{\log(1/h(p_G, p_{G'}))}\right)^{\left(\frac{2+d}{\beta(2+d/2)}\right)}\bigg). \end{aligned}$$

Combining the results for the cases $\bar{\theta} > \delta$ and $\bar{\theta} \leq \delta$, we get the required result.

3.6.2 Proof of Theorem 3.3.4

The proof of this result follows by an application of Lemma 3.7.2, 3.7.3 and 3.7.4 in combination with Theorem 2.1 in *Ghosal et al.* (2000).

We break the proof into several steps.

Step 1: First we compute the contraction rate relative to the Hellinger metric.

We apply Theorem 7.1 in *Ghosal et al.* (2000), with $\epsilon = L\epsilon_n$ and $D(\epsilon) = \exp\left(c_1 \left(\frac{\bar{\theta}_n}{\sqrt{\lambda_{min}}\epsilon_n}\right)^d \log\left(e + \frac{32e\bar{\theta}_n^2}{\lambda_{min}\epsilon_n^2}\right)\right)$, where $L \geq 2$ a large constant to be chosen later and c_1 is the constant in Eq. (3.70). Lemma 3.7.3 shows the validity of this choice of $D(\epsilon)$. Then there exists a test function ϕ_n that satisfies

$$\begin{aligned} P_{G_0}^n \phi_n &\leq \exp\left(c_1 \left(\frac{\bar{\theta}_n}{\sqrt{\lambda_{min}}\epsilon_n}\right)^d \log\left(e + \frac{32e\bar{\theta}_n^2}{\lambda_{min}\epsilon_n^2}\right)\right) \\ &\quad \times \exp(-KnL^2\epsilon_n^2)\frac{1}{1 - \exp(-KnL^2\epsilon_n^2)}, \\ \sup_{G \in \overline{\mathcal{G}}(\Theta_n): h(p_G, p_{G_0}) \geq L\epsilon_n} P_G^n(1 - \phi_n) &\leq \exp(-KnL^2\epsilon_n^2) \end{aligned} \tag{3.40}$$

Now,

$$\begin{aligned} &\mathbb{E}_{P_{G_0}} \Pi_n(G \in \overline{\mathcal{G}}(\Theta_n) : h(p_G, p_{G_0}) \geq L\epsilon_n | X_1, \ldots, X_n)\phi_n \\ &\quad \leq P_{G_0}^n \phi_n \leq 2\exp\left(c_1 \left(\frac{\bar{\theta}_n}{\sqrt{\lambda_{min}}\epsilon_n}\right)^d \log\left(e + \frac{32e\bar{\theta}_n^2}{\lambda_{min}\epsilon_n^2}\right) - KnL^2\epsilon_n^2\right). \end{aligned} \tag{3.41}$$

Based on computation with the posterior,

$$\begin{aligned}
& \Pi_n(G : h(p_G, p_{G_0} \geq \epsilon_n)|X_1, \ldots, X_n)(1-\phi_n) \\
= & \frac{\int_{G \in \overline{\mathcal{G}}(\Theta_n): h(p_G, p_{G_0}) \geq \epsilon_n} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)(1-\phi_n)}{\int_{G \in \overline{\mathcal{G}}(\Theta_n)} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)} \\
\leq & \frac{\int_{G \in \overline{\mathcal{G}}(\Theta_n): h(p_G, p_{G_0}) \geq \epsilon_n} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)(1-\phi_n)}{\int_{G \in \overline{\mathcal{G}}(\Theta_n): K(p_{G_0}, p_G) \lesssim \epsilon_n^2, K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2 (\log(M_n/\epsilon_n))^2} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)},
\end{aligned}$$

where $M_n = \exp(d\lambda_{min}^{-1}(5\bar{\theta}_0^2 + 4\bar{\theta}_n^2))$, with λ_{min} being the minimum eigenvalue of Σ.

Step 1.1: In this step we show that

$$\begin{aligned}
& \int_{G \in \overline{\mathcal{G}}(\Theta_n): K(p_{G_0}, p_G) \lesssim \epsilon_n^2, K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2 (\log(M_n/\epsilon_n))^2} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G) \\
& \gtrsim \exp(-(1+C)n\lambda_{min}\epsilon_n^2) \frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^{D_n}}{(2\Gamma(d/2+1))^{D_n}(2D_n)^{D_n-1}} \left(\frac{\sqrt{\lambda_{min}}\epsilon_n}{2\bar{\theta}_n}\right)^{r(D_n-1)+dD_n} \qquad (3.42) \\
& \text{with } p_{G_0}^n \text{ probability} \to 1,
\end{aligned}$$

for all $C > 0$ and $\epsilon_n > 0$ is sufficiently small, where $D_n = D(\sqrt{\lambda_{min}}\epsilon_n, \Theta_n, \|.\|) \approx \left(\frac{\bar{\theta}_n}{\epsilon_n}\right)^d$ stands for the maximal $\sqrt{\lambda_{min}}\epsilon_n$-packing number for Θ_n under $\|.\|$ norm, and $\Gamma(\cdot)$ is the gamma function. First we show that

$$\begin{aligned}
& \{G \in \overline{\mathcal{G}}(\Theta_n) : W_2(G, G_0) \lesssim \sqrt{\lambda_{min}}\epsilon_n\} \\
& \quad \subset \{G \in \overline{\mathcal{G}}(\Theta_n) : K(p_{G_0}, p_G) \lesssim \epsilon_n^2, K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2(\log(M_n/\epsilon_n))^2\},
\end{aligned} \qquad (3.43)$$

for ϵ_n sufficiently small.

Since $\int \frac{(p_{G_0}(x))^2}{p_G(x)}\mu(\mathrm{d}x) \leq M_n$ by Lemma 3.7.4 part(ii), it follows by an application of Theorem 5 in *Wong and Shen* (1995) that for $\epsilon_n < 1/2(1-e^{-1})^2$,

$$h(p_G, p_{G_0}) \lesssim \epsilon_n^2 \implies K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2(\log(M_n/\epsilon_n))^2.$$

Following Example 1 in *Nguyen* (2013), $h^2(p_G, p_{G_0}) \leq \frac{W_2^2(G, G_0)}{8\lambda_{min}}$ for Gaussian location mixtures. Moreover, $K(p_G, p_{G_0}) \leq \frac{W_2^2(G, G_0)}{2\lambda_{min}}$. Combining the above displays, Eq. (3.43) follows.

Following Lemma 8.1 in *Ghosal et al.* (2000), for every $C, \epsilon, M > 0$ and any measure Π on the set $\{G \in \overline{\mathcal{G}}(\Theta_n) : K(p_{G_0}, p_G) \lesssim \epsilon_n^2, K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2(\log(M/\epsilon_n))^2\}$, we have,

$$P_{G_0}^n\left(\int \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)}\mathrm{d}\Pi_n(G) \leq \exp(-(1+C)n\epsilon^2)\right) \leq \frac{1}{C^2 n\epsilon^2(\log(M/\epsilon))^2}. \quad (3.44)$$

The result in Eq.(3.42) now follows by an application of Lemma 3.7.2 in combination with Eq. (3.43) and (3.44) using the fact that $n\epsilon_n^2 \to \infty$.

Step 1.2: Let the event in (3.42) be denoted as T_n. Then

$$\begin{aligned} & \mathbb{E}_{P_{G_0}}\left[\Pi_n(G : h(p_G, p_{G_0}) \geq L\epsilon_n)|X_1, \ldots, X_n)(1-\phi_n)\right] \leq P_{G_0}(T_n^C) \\ + \quad & P_{G_0}(T_n)\frac{\exp((1+C)n\lambda_{min}\epsilon_n^2)}{\frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^{D_n}}{(2\Gamma(d/2+1))^{D_n}(2D_n)^{D_n-1}}\left(\frac{\sqrt{\lambda_{min}}\epsilon_n}{2\bar{\theta}_n}\right)^{r(D_n-1)+dD_n}} \\ & \times \sup_{G \in \overline{\mathcal{G}}(\Theta_n) : h(p_G, p_{G_0}) \geq L\epsilon_n} P_G^n(1-\phi_n) \\ \lesssim \quad & \frac{\exp((1+C)n\lambda_{min}\epsilon_n^2)}{\frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^{D_n}}{(2\Gamma(d/2+1))^{D_n}(2D_n)^{D_n-1}}\left(\frac{\sqrt{\lambda_{min}}\epsilon_n}{2\bar{\theta}_n}\right)^{r(D_n-1)+dD_n}}\exp(-KnL^2\epsilon_n^2) + o(1). \quad (3.45) \end{aligned}$$

The final step follows from simple computation similar to that of the Proof of Theorem 2.1 in *Ghosal et al.* (2000) and using the fact that $\frac{\bar{\theta}_n^d}{\epsilon_n^{d+2}} \log\left(\frac{\bar{\theta}_n}{\epsilon_n}\right) = o(n)$. Combining Eq. (3.41) and (3.45) and using the condition $\frac{\bar{\theta}_n^d}{\epsilon_n^{d+2}} \log\left(\frac{\bar{\theta}_n}{\epsilon_n}\right) = o(n)$, it follows that for L large enough

$$\Pi_n(G \in \overline{\mathcal{G}}(\Theta_n) : h(p_G, p_{G_0}) \geq L\epsilon_n | X_1, \ldots, X_n) \overset{P_{G_0}}{\to} 0. \tag{3.46}$$

Now using Theorem 3.3.2 for Gaussian location mixtures, the result follows by noting that as $\bar{\theta}_n \uparrow \infty$ and $\epsilon_n \downarrow 0$, $\frac{1}{\log(1/L\epsilon_n)} \leq \bar{\theta}_n$ for n sufficiently large.

3.6.3 Proof of Theorem 3.4.1

$a \lesssim b$ for this proof implies $a \leq C \cdot b$ for a universal constant C dependent on α, d, and $\bar{\theta}$. Also, $f * g$ will denote the outcome of convolution operation on functions f and g.

Consider the following density function in $\mathbb{R}$:

$$k(x) := c \left(\int_{-\infty}^{\infty} \exp(-itx) \exp(-t^4) \mathrm{d}t \right)^2, \tag{3.47}$$

where c is a proportionality constant so that $\int_{-\infty}^{\infty} k(x)\mathrm{d}x = 1$. Lemma 3.4.4 shows that $k(\cdot)$ is integrable.

Moreover, Lemma 3.4.5 shows that the characteristic function $\hat{k}(\cdot)$, corresponding to $k(\cdot)$ satisfies,

$$|\hat{k}(x)| \lesssim \exp(-(x/2)^4).$$

The strategy to obtain upper bounds for $W_\Phi(G, G')$ is to convolve G with mollifiers, $k_{\delta,d}(\cdot)$, of the form $k_{\delta,d}(x) = \prod_{i=1}^d \frac{1}{\delta} k(x_i/\delta)$ for $\delta > 0$, where $x := (x_1, \dots, x_d)$.

By triangle inequality, following Lemma 3.4.1 we can write:

$$W_\Phi(G, G') \leq W_1(G, G * k_{\delta,d}) + W_\Phi(G', G' * k_{\delta,d}) + W_\Phi(G * k_{\delta,d}, G' * k_{\delta,d}).$$

For $\Phi(x) = \exp(x) - 1$, following Lemma 3.4.2 part (ii),

$$W_\Phi(G, G * k_{\delta,d}) \leq C_\alpha \delta.$$

Therefore, we can write

$$W_\Phi(G, G') \leq 2C_\alpha \delta + W_\Phi(G * k_{\delta,d}, G' * k_{\delta,d}).$$

For every $M > 0$,

$$\begin{aligned} W_\Phi(G * k_{\delta,d}, G' * k_{\delta,d}) \quad &\leq \quad 2 \inf\{k \in \mathbb{R}^+ : \int\limits_{\mathbb{R}^d} \Phi(\|x\|/k) \cdot |(G - G') * k_{\delta,d}(x)| dx \leq 1\} \\ \leq 2 \inf\{k \in \mathbb{R}^+ : \quad & \underbrace{\int\limits_{\|x\|_2 \leq M} \Phi(\|x\|/k) \cdot |(G - G') * k_{\delta,d}(x)| dx}_{s_1} \leq 1/2 \\ \text{and} \quad & \underbrace{\int\limits_{\|x\|_2 > M} \Phi(\|x\|/k) \cdot |(G - G') * k_{\delta,d}(x)| dx}_{s_2} \leq 1/2\}, \end{aligned}$$

with the first inequality following from Lemma 3.4.6.

Bounding for s_1: Using Holder's inequality, we obtain

$$\begin{aligned}
&\inf\{k \in \mathbb{R}^+ : \int\limits_{\|x\|_2 \leq M} \Phi(\|x\|/k) \cdot |(G - G') * k_{\delta,d}(x)| dx \leq 1/2\} \\
\leq\ & \inf\{k > 0 : \int\limits_{\|x\| \leq M} \exp(\|x\|/k) \cdot |(G - G') * k_{\delta,d}(x)| dx \leq 3/2\} \\
\leq\ & \inf\Big\{k > 0 : \Big(\int\limits_{\|x\| \leq M} \exp(M/k) dx \Big)^{1/2} \Big(\int\limits_{\|x\| \leq M} |(G - G') * k_{\delta,d}(x)|^2 dx \Big)^{1/2} \leq 3/2 \Big\} \\
\leq\ & \inf\Big\{k > 0 : \frac{\pi^{d/4}}{\sqrt{(\frac{d}{2}+1)\Gamma(d/2)}} M^{d/2} \exp(M/k) \|(G - G') * k_{\delta,d}(x)\|_2 \leq 3/2 \Big\} \\
=\ & \frac{M}{\log(c_d/(\|(G - G') * k_{\delta,d}\|_2 M^{d/2}))} \qquad (3.48)
\end{aligned}$$

Since f is Gaussian, $\tilde{f}(\omega) \geq c_f \exp(-\alpha \sum_{i=1}^d \omega_i^2)$ for some $c_f, \alpha > 0$. Therefore, we have

$$\begin{aligned}
\|(G - G') * k_{\delta,d}\|_2^2 &= \int |\widetilde{G} - \widetilde{G}'|^2(\omega) |\widetilde{k}_{\delta,d}(\omega)|^2 d\omega = \int |\widetilde{f}(\widetilde{G} - \widetilde{G}')|^2(\omega) \frac{|\widetilde{k}_{\delta,d}(\omega)|^2}{|\widetilde{f}(\omega)|^2} d\omega \\
&\leq \|p_G - p_{G'}\|_2^2 \sup_{\omega \in \mathbb{R}^d} \frac{|\widetilde{k}_{\delta,d}(\omega)|^2}{|\widetilde{f}(\omega)|^2} \\
&\leq 4\|f\|_\infty h^2(p_G, p_{G_0}) \sup_{\omega \in \mathbb{R}^d} \Big\{ \frac{1}{c_f^2} \cdot \prod_{i=1}^d \exp(-\delta^4 |\omega_i|^4) \exp(2\alpha |\omega_i|^2) \Big\}.
\end{aligned}$$

Taking derivatives we obtain the maximum as,

$$\sup_{\omega_i \in \mathbb{R}} \Big\{ \exp(-\delta^4 |\omega_i|^4) \exp(2\alpha |\omega_i|^2) \Big\} = \exp(\alpha^2/\delta^4).$$

Therefore,

$$\inf\{k \in \mathbb{R}^+ : \int\limits_{\|x\|_2 \leq M} \Phi(\|x\|/k) \cdot |(G - G') * k_{\delta,d}(x)| dx \leq 1/2\} \\ \leq \frac{M}{\log(c/(h(p_G, p_{G_0}) \exp(\alpha^2 d\delta^{-4}) M^{d/2}))} \tag{3.49}$$

for a different constant c.

Bounding for s_2: For $M > 0$,
assume $k_1 = \inf\{k \in \mathbb{R}^+ : \mathbb{E}_{X \sim G-G'}(\Phi(\|X\|^{5/4}/kM^{(1/4)}) \leq 1/2\}$ and
$k_2 = \inf\{k \in \mathbb{R}^+ : \mathbb{E}_{Y \sim k_{\delta,d}}(\Phi(\|Y\|^{5/4}/kM^{(1/4)}) \leq 1/2\}$. Then by convexity of Φ it is clear that $\inf\{k \in \mathbb{R}^+ : \mathbb{E}_{X \sim G-G', Y \sim k_{\delta,d}}(\Phi(\|X + Y\|^{5/4}/kM^{(1/4)}) \leq 1/2\} \leq k_1 + k_2$. Then,

$$\begin{aligned}
&\inf\{k \in \mathbb{R}^+ : \int\limits_{\|x\|_2 > M} \Phi(\|x\|/k) \cdot |(G - G') * k_{\delta,d}(x)| dx \leq 1/2\} \\
\leq\ & \inf\{k \in \mathbb{R}^+ : \int\limits_{\|x\|_2 > M} \Phi(\|x\|^{5/4}/kM^{(1/4)}) \cdot |(G - G') * k_{\delta,d}(x)| dx \leq 1/2\} \\
\leq\ & \inf\{k \in \mathbb{R}^+ : \mathbb{E}_{X \sim G-G', Y \sim k_{\delta,d}}(\Phi(\|X + Y\|^{5/4}/kM^{(1/4)}) \leq 1/2\} \\
\leq\ & \inf\{k > 0 : \int\limits_{\mathbb{R}^d} \exp(\|x\|^{5/4}/kM^{1/4}) \cdot |(G - G')(x)| dx \leq 3/2\} \\
+\ & \inf\{k > 0 : \int\limits_{\mathbb{R}^d} \exp(\|x\|^{5/4}/kM^{1/4}) \cdot |k_{\delta,d}(x)| dx \leq 3/2\} \\
\leq\ & \frac{(d\bar{\theta})^{5/4}}{\log(3/2)M^{1/4}} + C\delta^{5/4}/M^{1/4},
\end{aligned} \tag{3.50}$$

where $C = \inf\{k > 0 : \int\limits_{\mathbb{R}^d} \exp(\|x\|^{5/4}/k) \cdot |k_{1,d}(x)| dx < \infty$ as $k_{1,d}(x) \sim O(\exp(-|x|^{4/3}))$ for large $|x|$, by Lemma 3.4.4

Therefore, using these results we get

$$
\begin{aligned}
W_\Phi(G,G') &\lesssim \delta + \max\Big\{ \frac{(d\bar{\theta})^{5/4}}{\log(3/2)M^{1/4}} + C\delta^{5/4}/M^{1/4}, \\
&\qquad\qquad \frac{M}{\log(c/(h(p_G,p_{G_0})\exp(\alpha^2 d\delta^{-4})M^{d/2}))} \Big\} \\
&\leq \frac{(d\bar{\theta})^{5/4}}{\log(3/2)M^{1/4}} + C\delta^{5/4}/M^{1/4} \\
&\qquad\qquad + \frac{M}{\log(c/(h(p_G,p_{G_0})\exp(\alpha^2 d\delta^{-4})M^{d/2}))}. \qquad (3.51)
\end{aligned}
$$

Choosing $M = (\log(1/h(p_G,p_{G_0})))^{1/2}$ and $\delta = \dfrac{2\alpha^2}{\log(1/h(p_G,p_{G_0}))}$ in Eq. (3.51) we get,

$$
\begin{aligned}
W_\Phi(G,G') &\lesssim \frac{(d\bar{\theta})^{5/4}}{(\log(1/h(p_G,p_{G_0})))^{1/8}} + \left(\frac{1}{\log(1/h(p_G,p_{G_0}))}\right)^{11/8} \\
&\qquad + \left(\frac{1}{\log(c/h(p_G,p_{G_0})(\log(1/h(p_G,p_{G_0})))^{d/4})}\right)^{1/2}
\end{aligned}
\qquad (3.52)
$$

3.7 Appendix B: Auxiliary results

In this appendix, we provide auxiliary results for posterior convergence rate of mixing measures under the growing parameter space or prior distribution settings.

3.7.1 Proof of Theorem 3.3.1

The proof of this theorem is an improvement and generalization of the argument of Lemma 7 in *Gao and van der Vaart* (2016) to multivariate settings of kernel functions where the constants in the inequalities are carefully studied. Throughout this proof, $a \lesssim b$ means that $a \leq C \cdot b$ where C is a universal constant independent of α, β, d, and $\overline{\theta}$. Denote Φ_δ to be the multivariate normal distribution with mean 0 and covariance $\delta^2 I_d$. From triangle inequality, we obtain that

$$W_1(G, G') \leq W_1(G, G * \Phi_\delta) + W_1(G * \Phi_\delta, G' * \Phi_\delta) + W_1(G' * \Phi_\delta, G').$$

By choosing two independent vectors $X \sim G$ and $Y \sim N(0, \delta^2 I)$, we have $X + Y \sim G * \Phi_\delta$. From the definition of first order Wasserstein metric, it follows that
$W_1(G, G * \Phi_\delta) \leq E\|X + Y - X\| = E\|Y\| = \dfrac{\delta d}{\sqrt{2}}$. Hence, we have

$$W_1(G, G') \leq \sqrt{2}\delta d + W_1(G * \Phi_\delta, G' * \Phi_\delta). \tag{3.53}$$

Then, for every positive constant M we obtain that

$$\begin{aligned} W_1(G * \Phi_\delta, G' * \Phi_\delta) &\leq \int \|x\| \cdot |(G - G') * \phi_\delta(x)| dx \\ &= \underbrace{\int\limits_{\|x\| \leq M} \|x\| \cdot |(G - G') * \phi_\delta(x)| dx}_{s_1} \\ &+ \underbrace{\int\limits_{\|x\| > M} \|x\| \cdot |(G - G') * \phi_\delta(x)| dx}_{s_2} . \end{aligned}$$

Bounding s_1: Following the steps as in Eq. (3.35), we can show that using Holder's inequality, we obtain

$$s_1 \leq \frac{\pi^{d/4}}{\sqrt{\Gamma(d/2+1)}} M^{1+d/2} \|(G - G') * k_{\delta,d}\|_2. \tag{3.54}$$

According to the assumption with f, we have

$$\begin{aligned} \|(G - G') * \phi_\delta\|_2^2 &= \int |\widetilde{G} - \widetilde{G}'|^2(\omega) |\widetilde{\phi}_\delta(\omega)|^2 d\omega = \int |\widetilde{f}(\widetilde{G} - \widetilde{G}')|^2(\omega) \frac{|\widetilde{\phi}_\delta(\omega)|^2}{|\widetilde{f}(\omega)|^2} d\omega \\ &\leq \|p_G - p_{G'}\|_2^2 \sup_{\omega \in \mathbb{R}^d} \frac{|\widetilde{\phi}_\delta(\omega)|^2}{|\widetilde{f}(\omega)|^2} \\ &\leq 4\|f\|_\infty h^2(p_G, p_{G_0}) \sup_{\omega \in \mathbb{R}^d} \left\{ \frac{1}{c_f^2} \cdot \exp(-\delta^2 \|\omega\|^2) \prod_{i=1}^d (1 + \alpha|\omega_i|^\beta)^2 \right\}. \end{aligned}$$

Hence, we can compute:

$$\sup_{\omega \in \mathbb{R}^d} \left\{ \exp(-\delta^2 \|\omega\|^2) \prod_{i=1}^d (1 + \alpha|\omega_i|^\beta)^2 \right\} = \big[\sup_{x \geq 0} \left\{ \exp\big(\frac{-\delta^2 x^2}{2}\big) \left(1 + \alpha x^\beta\right) \right\} \big]^{2d}$$

Since:

$$1 + \alpha x^{\beta} \leq 2 \max\{1, \alpha x^{\beta}\} = \max\{2, 2\alpha x^{\beta}\}$$

Thus:

$$\big[\sup_{x \geq 0} \big\{ \exp\big(\frac{-\delta^2 x^2}{2}\big) \left(1 + \alpha x^{\beta}\right)\big\}\big]^{2d} \leq \left(\max\Big\{2\,,\ 2\alpha \exp(\frac{-\beta}{2})\big(\frac{\beta}{\delta^2}\big)^{\beta/2}\Big\}\right)^{2d},$$

since:

$$\begin{aligned} -\delta^2 x \exp(\frac{-\delta^2 x^2}{2})(2\alpha x^{\beta}) &+ \exp(\frac{-\delta^2 x^2}{2})(2\alpha\beta x^{\beta-1}) \\ &= \exp(\frac{-\delta^2 x^2}{2})(2\alpha x^{\beta-1})(-\delta^2 x^2 + \beta) \end{aligned}$$

attains maximum at $\sqrt{\beta}/\delta$, from which we derive

$$||(G - G') * \phi_{\delta}||_2 \leq \frac{2\sqrt{\|f\|_{\infty}}}{c_f} h(p_G, p_{G'}) \cdot \left(\max\Big\{2\,,\ 2\alpha \exp(\frac{-\beta}{2})\big(\frac{\beta}{\delta^2}\big)^{\beta/2}\Big\}\right)^{d} \quad (3.55)$$

By combining the result of (3.35) and (3.55), we achieve that

$$\begin{aligned} s_1 \quad \leq \quad & \frac{2\sqrt{\|f\|_{\infty}}}{c_f} \frac{\pi^{d/4}}{\sqrt{\Gamma(d/2+1)}} M^{1+d/2} h(p_G, p_{G'}) \\ & \cdot\left(\max\Big\{2\,,\ 2\alpha \exp(\frac{-\beta}{2})\big(\frac{\beta}{\delta^2}\big)^{\beta/2}\Big\}\right)^{d} \end{aligned} \quad (3.56)$$

Bounding s_2: On the other hand, when $M > 2\sqrt{d}\bar{\theta} + 2\delta^2$, for s_2 we have

$$
\begin{aligned}
s_2 &\leq \exp(-M) \int\limits_{\|x\|>M} \|x\| \, \exp(\|x\|) \, |(G - G') * \phi_\delta(x)| dx \\
&= \exp(-M) \int\limits_{\|x\|>M} \|x\| \, \exp(\|x\|) \Big| \int\limits_{||y|| \leq \sqrt{d}\bar{\theta}} (G - G')(y) \phi_\delta(x - y) dy \Big| \, dx \\
&\leq 2 \exp(-M) \int\limits_{\|x\|>M} \int\limits_{||y|| \leq \sqrt{d}\bar{\theta}} \|x\| \, \exp(\|x\|) \, \phi_\delta(x - y) dy \, dx \\
&= \frac{2}{(2\pi)^{d/2} \delta^d} \exp(-M) \int\limits_{\|x\|>M} \int\limits_{||y|| \leq \sqrt{d}\bar{\theta}} \|x\| \, \exp(\|x\|) \, \exp\Big(- \frac{||x + y||_2^2}{2\delta^2} \Big) dy \, dx \\
&\leq \frac{2}{(2\pi)^{d/2} \delta^d} \exp(-M) \cdot \frac{\pi^{d/2}}{\Gamma(1 + \frac{d}{2})} \cdot (\sqrt{d}\bar{\theta})^d \\
&\qquad \cdot \int\limits_{\|x\|>M} \|x\| \, \exp(\|x\|) \, \exp\Big(- \frac{(||x|| - \sqrt{d}\bar{\theta})^2}{2\delta^2} \Big) \, dx \\
&= \frac{2 \exp(-M) (\sqrt{d}\bar{\theta})^d}{(2)^{d/2} \delta^d \Gamma(1 + \frac{d}{2})} \cdot \int\limits_{\|x\|>M} \|x\| \, \exp(\|x\|) \, \exp\Big(- \frac{(||x|| - \sqrt{d}\bar{\theta})^2}{2\delta^2} \Big) \, dx \\
&= \frac{2 \exp(-M) (\sqrt{d}\bar{\theta})^d}{(2)^{d/2} \delta^d \Gamma(1 + \frac{d}{2})} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)} \\
&\qquad \cdot \int\limits_{r>M} r^d \, \exp\Big(- \frac{(r - (\sqrt{d}\bar{\theta} + \delta^2))^2}{2\delta^2} \Big) \, \exp\Big(\frac{\delta^2 + 2\sqrt{d}\bar{\theta}}{2} \Big) dr \\
&= \frac{2 \exp(-M) (\sqrt{d}\bar{\theta})^d}{(2)^{d/2} \delta^d \Gamma(1 + \frac{d}{2})} \cdot \frac{2\pi^{d/2}}{\Gamma(d/2)} \\
&\qquad \cdot \int\limits_{t > M - (\sqrt{d}\bar{\theta} + \delta^2)} (2t)^d \, \exp\Big(- \frac{t^2}{2\delta^2} \Big) \, \exp\Big(\frac{\delta^2 + 2\sqrt{d}\bar{\theta}}{2} \Big) dt \\
&= \frac{4\delta \cdot 2^{d/2} \exp(-M) (\sqrt{d}\bar{\theta})^d}{\Gamma(1 + \frac{d}{2})} \cdot \frac{\pi^{d/2}}{\Gamma(d/2)} \exp\Big(\frac{\delta^2 + 2\sqrt{d}\bar{\theta}}{2} \Big) \int\limits_{u > \frac{M}{2\delta}} u^d \, \exp\Big(- \frac{u^2}{2} \Big) \, du \\
&\leq \frac{4\delta \cdot 2^{d/2} \exp(-M) (\sqrt{d}\bar{\theta})^d}{\Gamma(1 + \frac{d}{2})} \cdot \frac{\pi^{d/2}}{\Gamma(d/2)} \exp\Big(\frac{\delta^2 + 2\sqrt{d}\bar{\theta}}{2} \Big) 2^{(d-1)/2} \Gamma\left(\frac{d+1}{2} \right) \\
&\leq \frac{2\sqrt{2}\delta \cdot (4\pi d \bar{\theta}^2)^{d/2}}{\Gamma(1 + \frac{d}{2})} \exp(-M/2)
\end{aligned}
$$

Then, using Stirling's approximation along with the accurate bound from Robbins [citation], we can simplify the result to:

$$s_2 \leq \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{(8\pi e)^{d/2}}{\sqrt{d}} \cdot \delta \bar{\theta}^d \exp(-M/2).$$

Let $c_d = \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{(8\pi e)^{d/2}}{\sqrt{d}}$, we can get:

$$s_2 \leq c_d \, \delta \, \bar{\theta}^d \exp(-\frac{M}{2}). \tag{3.57}$$

By means of (3.56) and (3.58), an upper bound for $W_1(G, G')$ is

$$W_1(G, G') \leq \sqrt{2}\delta d + a_d \, M^{1+d/2} h(p_G, p_{G'}) \cdot \max\left\{1 \, , \delta^{-\beta d}\right\} + c_d \, \delta \, \bar{\theta}^d \exp(-\frac{M}{2})$$

where:

$$a_d = \frac{\pi^{d/4}}{\sqrt{(\frac{d}{2}+1)\Gamma(d/2)}} \cdot \frac{2^{d+1}\sqrt{c_1}}{c_f} \cdot \max\{1, (\alpha \exp(-\beta/2)\beta^{\beta/2})^d\}, \qquad c_d = \frac{2\sqrt{2}}{\sqrt{\pi}} \frac{(8\pi e)^{d/2}}{\sqrt{d}}$$

Assume that $\delta \in (0, \bar{\delta}]$ for some $\bar{\delta} > 1$ that will be chosen later. Then, it is clear that $\max\left\{1, \delta^{-\beta d}\right\} \leq (\bar{\delta}/\delta)^{\beta d}$ for all $\delta \in (0, \bar{\delta}]$. Hence, as $\delta \in (0, \bar{\delta}]$, we further achieve the upper bound of $W_1(G, G')$ as

$$\begin{aligned} W_1(G, G') &\leq \sqrt{2}\delta d + \bar{\delta}^{\beta d} a_d \, M^{1+d/2} h(p_G, p_{G'}) \delta^{-\beta d} + c_d \bar{\delta} \bar{\theta}^d \exp(-\frac{M}{2}) \\ &\leq C_1 \Bigg(\delta + M^{1+d/2} h(p_G, p_{G'}) \delta^{-\beta d} + \exp(-\frac{M}{2}) \Bigg) \end{aligned}$$

where $C_1 = \max\left\{\sqrt{2}d, \bar{\delta}^{\beta d} a_d, c_d \bar{\delta}\bar{\theta}^d\right\}$. By taking the derivative with respect to δ, we can choose $\delta = \biggl(\beta d M^{1+d/2} h(p_G, p_{G'})\biggr)^{1/(\beta d+1)}$. Regarding M, we choose $M = \dfrac{2}{\beta d+1}\log\biggl(\dfrac{\overline{C}}{h(p_G, p_{G'})}\biggr)$ where $\overline{C}$ is chosen to satisfy the following two properties:

- $\log(\overline{C}) \geq (\beta d+1)(\sqrt{d}\theta + \delta^2)$.
- The function $\biggl(\log(\overline{C}/\epsilon)\biggr)^{1+d/2}\epsilon$ is increasing as $\epsilon \in (0, 1]$.

The second requirement holds as long as $\overline{C} \geq \exp(1 + d/2)$. Overall, we can choose $\overline{C} = \exp\biggl((\beta d+1)(\sqrt{d}\bar{\theta} + \overline{\delta}^2)\biggr)$. Now, we need to choose $\bar{\delta}$ such that $\biggl(\beta d M^{1+d/2} h(p_G, p_{G'})\biggr)^{1/(\beta d+1)} \leq \bar{\delta}$ for all values of $h(p_G, p_{G'})$. Due to the choice of $\bar{C}$, the previous inequality is guaranteed as long as it holds for $h(p_G, p_{G'}) = 1$, i.e.,

$$\beta d\biggl(\frac{2}{\beta d+1}\biggr)^{1+d/2}(\beta d+1)^{1+d/2}(\sqrt{d}\bar{\theta} + \bar{\delta}^2)^{1+d/2} \leq \bar{\delta}^{\beta d+1},$$

which is equivalent to $\beta d 2^{1+d/2}(\sqrt{d}\bar{\theta} + \bar{\delta}^2)^{1+d/2} \leq \bar{\delta}^{\beta d+1}$. By choosing $\bar{\delta}^2 \geq \sqrt{d}\bar{\theta}$, the previous bounds holds as long as $\beta d 2^{2+d}\bar{\delta}^{2+d} \leq \bar{\delta}^{\beta d+1}$. As long as $(\beta-1)d > 1$, this inequality leads to $\bar{\delta}^{(\beta-1)d-1} \geq \beta 2^{2+d} d$. As a consequence, if we choose $\bar{\delta}^{(\beta-1)d-1} = \beta 2^{2+d} d + (\sqrt{d}\bar{\theta})^{\frac{(\beta-1)d-1}{2}}$, then all the previous conditions hold.

Combining all of the previous results, we eventually have

$$\begin{aligned}
W_1(G, G') \;&\leq\; \max\left\{\sqrt{2}d, \bar{\delta}^{\beta d} a_d, c_d \bar{\delta}\bar{\theta}^d\right\}\left\{\biggl(\frac{1}{\overline{C}}\biggr)^{1/(\beta d+1)} + \biggl(\frac{2}{\beta d+1}\biggr)^{(1+d/2)/(\beta d+1)}\right. \\
&\quad \times \left.\biggl((\beta d)^{1/(\beta d+1)} + (\beta d)^{-\beta d/(\beta d+1)}\biggr)\right\}\left\{\log\biggl(\frac{\overline{C}}{h(p_G, p_{G'})}\biggr)\right\}^{1+d/2} h(p_G, p_{G'}) \\
&\leq\; 3\max\left\{\sqrt{2}d, \bar{\delta}^{\beta d} a_d, c_d \bar{\delta}\bar{\theta}^d\right\}\left\{\log\biggl(\frac{\overline{C}}{h(p_G, p_{G'})}\biggr)\right\}^{\frac{1+d/2}{1+\beta d}} h(p_G, p_{G'})^{\frac{1}{1+\beta d}}
\end{aligned}$$

where the final inequality is due to the fact that $\beta d > 1$, which implies that

$$\left\{ \left(\frac{1}{\overline{\overline{C}}} \right)^{1/(\beta d+1)} + \left(\frac{2}{\beta d+1} \right)^{(1+d/2)/(\beta d+1)} \left((\beta d)^{1/(\beta d+1)} + (\beta d)^{-\beta d/(\beta d+1)} \right) \right\} \leq 3.$$

As a consequence, we achieve the conclusion of part (a) of the theorem.

(b) Unlike the upper bound of s_2 in part (a), the high level proof idea of this case is to achieve the upper bound of s_2 without any dependence on δ on the lower bound of M under the setting that $\beta \leq 1 + 1/d$. In particular, we obtain that

$$\begin{aligned} s_2 &\leq \exp(-M) \int\limits_{\|x\|>M} \|x\| \exp(\|x\|) \phi_\delta(x-\theta) dx \\ &\leq \exp(-M) \int \|x+\theta\| \exp(\|x+\theta\|) \phi_\delta(x) dx \\ &\leq \exp(-M) \int (\|x\| + \|\theta\|) \exp(\|x\| + \|\theta\|) \phi_\delta(x) dx \\ &\leq \exp(-M) \int (\|x\| + d\overline{\theta}) \exp(\|x\| + d\overline{\theta}) \phi_\delta(x) dx. \end{aligned} \tag{3.58}$$

It is clear that

$$\begin{aligned} \int \exp(\|x\|) \phi_\delta(x) dx &\leq \int \exp(\sum_{i=1}^{d} |x_i|) \phi_\delta(x) dx \\ &= \prod_{i=1}^{d} \int \frac{1}{\sqrt{2\pi}\delta} \exp\biggl(- \frac{|x_i|^2}{2\delta^2} + |x_i| \biggr) dx_i \\ &= \prod_{i=1}^{d} \int \exp(\delta^2/2) \frac{1}{\sqrt{2\pi}\delta} \exp\biggl(- \frac{(|x_i| - \delta^2)^2}{2\delta^2} \biggr) dx_i \\ &\leq 2^d \exp(d\delta^2/2). \end{aligned} \tag{3.59}$$

Additionally, we also have that

$$\begin{aligned}
&\int \|x\| \exp(\|x\|)\phi_\delta(x)dx \\
&\quad \leq \int (\sum_{i=1}^{d} |x_i|) \exp(\sum_{i=1}^{d} |x_i|)\phi_\delta(x)dx = \sum_{i=1}^{d} \int |x_i| \exp(\sum_{i=1}^{d} |x_i|)\phi_\delta(x)dx \\
&\quad = \sum_{i=1}^{d} \left\{ \int \frac{1}{\sqrt{2\pi}\delta} |x_i| \exp\Big(- \frac{|x_i|^2}{2\delta^2} + |x_i| \Big) dx_i \prod_{j\neq i} \int \frac{1}{\sqrt{2\pi}\delta} \exp\Big(- \frac{|x_j|^2}{2\delta^2} + |x_j| \Big) dx_j \right\} \\
&\quad \leq 2^{d-1} \exp((d-1)\delta^2/2) \sum_{i=1}^{d} \int \frac{1}{\sqrt{2\pi}\delta} |x_i| \exp\Big(- \frac{|x_i|^2}{2\delta^2} + |x_i| \Big) dx_i
\end{aligned}$$

where the final inequality is due to inequality in (3.59). For any $1 \leq i \leq d$, we can demonstrate that

$$\begin{aligned}
&\int \frac{1}{\sqrt{2\pi}\delta} |x_i| \exp\Big(- \frac{|x_i|^2}{2\delta^2} + |x_i| \Big) dx_i \\
&\quad = 2 \int\limits_{x_i > 0} \frac{1}{\sqrt{2\pi}\delta} x_i \exp\Big(- \frac{x_i^2}{2\delta^2} + x_i \Big) dx_i \\
&\quad = 2 \int\limits_{x_i > 0} \frac{1}{\sqrt{2\pi}\delta} \exp(\delta^2/2) x_i \exp\Big(- \frac{(x_i - \delta^2)^2}{2\delta^2} \Big) dx_i \\
&\quad \leq 2 \Big\{ \int\limits_{x_i > \delta^2} \frac{1}{\sqrt{2\pi}\delta} \exp(\delta^2/2)(x_i - \delta^2) \exp\Big(- \frac{(x_i - \delta^2)^2}{2\delta^2} \Big) dx_i \\
&\quad + \delta^2 \exp(\delta^2/2) \int\limits_{x_i > 0} \frac{1}{\sqrt{2\pi}\delta} \exp\Big(- \frac{(x_i - \delta^2)^2}{2\delta^2} \Big) \Big\} \\
&\quad \leq 2 \Big(\frac{1}{\sqrt{2\pi}\delta} \exp(\delta^2/2)\delta^2 + \delta^2 \exp(\delta^2/2) \Big) \\
&\quad \leq 2(\delta + \delta^2) \exp(\delta^2/2). \qquad (3.60)
\end{aligned}$$

By combining (3.58), (3.59), and (3.60), we eventually obtain that

$$\int\limits_{\|x\|>M} \|x\| \exp(\|x\|)\phi_{\delta}(x-\theta)dx \leq 2^{d} d \exp\biggr(\frac{d\delta^{2}}{2} + d\overline{\theta}\biggr)(\overline{\theta} + \delta + \delta^{2}).$$

The above result leads to the following upper bound of s_2

$$\begin{aligned} s_2 &\leq 2^{d+1} d \exp\biggr(\frac{d\delta^{2}}{2} + d\overline{\theta}\biggr)(\overline{\theta} + \delta + \delta^{2}) \exp(-M) \\ &\leq 2^{d+1} d \exp\biggr(\frac{d\overline{\delta}^{2}}{2} + d\overline{\theta}\biggr)(\overline{\theta} + \overline{\delta} + \overline{\delta}^{2}) \exp(-M) \qquad (3.61) \end{aligned}$$

for any $\delta \in (0, \overline{\delta}]$ where the choice of $\overline{\delta}$ will be chosen later. Combining the result of (3.61) with (3.56), we achieve an upper bound for $W_1(G, G')$ as follows

$$\begin{aligned} W_1(G, G') &\leq \sqrt{2}\delta d + \overline{\delta}^{\beta d} a_d M^{1+d/2} h(p_G, p_{G'}) \delta^{-\beta d} \\ &\qquad + 2^{d+1} d \exp\biggr(\frac{d\overline{\delta}^{2}}{2} + d\overline{\theta}\biggr)(\overline{\theta} + \overline{\delta} + \overline{\delta}^{2}) \exp(-M) \\ &\leq C_2 \biggr(\delta + M^{1+d/2} h(p_G, p_{G'}) \delta^{-\beta d} + \exp(-M)\biggr) \qquad (3.62) \end{aligned}$$

for any $\delta \in (0, \overline{\delta}]$ and $M > 0$ where $C_2 = \max\biggr\{\sqrt{2}d, \overline{\delta}^{\beta d} a_d, 2^{d+1} d \exp\biggr(\frac{d\overline{\delta}^{2}}{2} + d\overline{\theta}\biggr)(\overline{\theta} + \overline{\delta} + \overline{\delta}^{2})\biggr\}$. By using the same argument as that of part (a), we can choose $\delta = \biggr(\beta d M^{1+d/2} h(p_G, p_{G'})\biggr)^{1/(\beta d+1)}$ and $M = \dfrac{1}{\beta d + 1} \log\biggr(\dfrac{\widetilde{C}}{h(p_G, p_{G'})}\biggr)$ where $\widetilde{C} = \exp(1 + d)$ is chosen to guarantee that $\biggr\{\log(\widetilde{C}/\epsilon)\biggr\}^{1+d/2} \epsilon$ is a strictly increasing function of $\epsilon \in (0, 1]$. To guarantee that the above choice of $\delta \in (0, \bar{\delta}]$, it is sufficient to choose $\bar{\delta}$ such that

$$\left(\beta d M^{1+d/2} h(p_G, p_{G'})\right)^{1/(\beta d+1)} \leq \bar{\delta}$$

for all values of $h(p_G, p_{G'})$. Due to the choice of $\widetilde{C}$, we only need to consider $h(p_G, p_{G'}) = 1$, i.e., we need to choose $\bar{\delta}$ such that $\bar{\delta}^{\beta d+1} \geq \beta d(1+d)^{1+d/2}$. Therefore, by choosing $\bar{\delta}^{\beta d+1} = \beta d(1+d)^{1+d/2}$, the previous conditions hold. These choices of $\bar{\delta}$ and $\widetilde{C}$ eventually lead to

$$W_1(G, G') \leq \quad C_2 \left\{ \log\left(\frac{\widetilde{C}}{h(p_G, p_{G'})} \right) \right\}^{\frac{1+d/2}{1+\beta d}} h(p_G, p_{G'})^{\frac{1}{1+\beta d}}.$$

We achieve the conclusion of part (b) of the theorem.

3.7.2 Proof of Theorem 3.3.3

The proof of this result is similar to the proof of Theorem 3.3.4.

We break the proof into several steps.

Step 1: First we compute the contraction rate relative to the Hellinger metric.

We apply Theorem 7.1 in *Ghosal et al.* (2000), with $\epsilon = L\epsilon_n$ and $D(\epsilon) = \exp\left(c_2 \left(\frac{\tilde{\theta}_n}{\lambda_{min}\epsilon^2} \right)^d \log\left(e + \frac{16\sqrt{2}e\tilde{\theta}_n}{\lambda_{min}\epsilon^2} \right) \right)$, where $L \geq 2$ a large constant to be chosen later and c_2 is the constant in Eq. (3.71). Lemma 3.7.3 shows that this choice of

$D(\epsilon)$ is valid. Then there exists a test function ϕ_n that satisfies

$$\begin{aligned} P^n_{G_0}\phi_n &\leq \exp\left(c_2\left(\frac{\tilde{\theta}_n}{\lambda_{min}\epsilon_n^2}\right)^d \log\left(e + \frac{16\sqrt{2}e\tilde{\theta}_n}{\lambda_{min}\epsilon_n^2}\right)\right) \\ &\qquad \times \exp(-KnL^2\epsilon_n^2)\frac{1}{1-\exp(-KnL^2\epsilon_n^2)}, \\ \sup_{G\in\overline{\mathcal{G}}(\Theta_n):h(p_G,p_{G_0})\geq L\epsilon_n} P^n_G(1-\phi_n) &\leq \exp(-KnL^2\epsilon_n^2) \end{aligned} \tag{3.63}$$

Now,

$$\begin{aligned} &\mathbb{E}_{P_{G_0}}\Pi_n(G\in\overline{\mathcal{G}}(\Theta_n) : h(p_G,p_{G_0}) \geq L\epsilon_n|X_1,\ldots,X_n)\phi_n \\ \leq\ & P^n_{G_0}\phi_n \leq 2\exp\left(c_2\left(\frac{\tilde{\theta}_n}{\lambda_{min}\epsilon_n^2}\right)^d \log\left(e + \frac{16\sqrt{2}e\tilde{\theta}_n}{\lambda_{min}\epsilon_n^2}\right) - KnL^2\epsilon_n^2\right) \\ &= o(1). \end{aligned} \tag{3.64}$$

The last line follows from the condition $\frac{\bar{\theta}_n^d}{\epsilon_n^{2d+2}}\log\left(\frac{\exp(\bar{\theta}_n)}{\epsilon_n^2}\right) = o(n)$ and therefore, $\frac{\bar{\theta}_n^d}{\epsilon_n^{2d+2}}\log\left(\frac{\bar{\theta}_n}{\epsilon_n^2}\right) = o(n)$. Based on computation with the posterior,

$$\begin{aligned} &\Pi_n(G : h(p_G, p_{G_0} \geq \epsilon_n)|X_1,\ldots,X_n)(1-\phi_n) \\ &\quad = \frac{\int_{G\in\overline{\mathcal{G}}(\Theta_n):h(p_G,p_{G_0})\geq\epsilon_n} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)(1-\phi_n)}{\int_{G\in\overline{\mathcal{G}}(\Theta_n)} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)} \\ &\quad \leq \frac{\int_{G\in\overline{\mathcal{G}}(\Theta_n):h(p_G,p_{G_0})\geq\epsilon_n} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)(1-\phi_n)}{\int_{G\in\overline{\mathcal{G}}(\Theta_n):K(p_{G_0},p_G)\lesssim\epsilon_n^2\log(M_n/\epsilon_n), K_2(p_{G_0},p_G)\lesssim\epsilon_n^2(\log(M_n/\epsilon_n))^2} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)} \mathrm{d}\Pi_n(G)}, \end{aligned}$$

where $M_n = \exp\left(\frac{7}{4}d\lambda_{min}^{-1}(\bar{\theta}_0 + \bar{\theta}_n)\right)V(\bar{\theta}_0)$, where $V(\cdot)$ is the function in Lemma 3.7.4 part(i), and λ_{min} being the minimum eigenvalue of Σ.

Step 1.1: In this step we show that

$$\int\limits_{G\in\overline{\mathcal{G}}(\Theta_n):K(p_{G_0},p_G)\lesssim\epsilon_n^2\log(M_n/\epsilon_n),K_2(p_{G_0},p_G)\lesssim\epsilon_n^2(\log(M_n/\epsilon_n))^2} \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_0}(X_i)}\mathrm{d}\Pi_n(G)$$

$$\begin{aligned}&\gtrsim \exp(-(1+C)n\sqrt{\lambda_{min}}\epsilon_n^2\log(M_n/\epsilon_n))\\ &\quad\times\frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^{D_n}}{(2\Gamma(d/2+1))^{D_n}(2D_n)^{D_n-1}}\left(\frac{\sqrt{\lambda_{min}}\epsilon_n^2}{2\bar{\theta}_n}\right)^{(D_n-1)+dD_n} \qquad (3.65)\\ &\qquad\qquad\qquad\qquad\qquad\qquad\text{with } p_{G_0}^n \text{ probability} \to 1,\end{aligned}$$

for all $C>0$ and $\epsilon_n>0$ is sufficiently small, where $D_n = D(\sqrt{\lambda_{min}}\epsilon_n^2,\Theta_n,\|.\|) \approx \left(\frac{\bar{\theta}_n}{\epsilon_n^2}\right)^d$ stands for the maximal $\sqrt{\lambda_{min}}\epsilon_n$-packing number for Θ_n under $\|.\|$ norm, and $\Gamma(\cdot)$ is the gamma function. First we show that

$$\begin{aligned}&\{G\in\overline{\mathcal{G}}(\Theta_n): W_1(G,G_0)\lesssim\sqrt{\lambda_{min}}\epsilon_n^2\}\\ &\quad\subset\{G\in\overline{\mathcal{G}}(\Theta_n): K(p_{G_0},p_G)\lesssim\epsilon_n^2, K_2(p_{G_0},p_G)\lesssim\epsilon_n^2(\log(M_n/\epsilon_n))^2\}, \quad (3.66)\end{aligned}$$

for ϵ_n sufficiently small.

Since $\int \frac{(p_{G_0}(x))^2}{p_G(x)}\mu(\mathrm{d}x)\le M_n$ by Lemma 3.7.4 part(ii), it follows by an application of Theorem 5 in *Wong and Shen* (1995) that for $\epsilon_n < 1/2(1-e^{-1})^2$,

$$h(p_G,p_{G_0})\lesssim\epsilon_n^2 \implies K_2(p_{G_0},p_G)\lesssim\epsilon_n^2(\log(M_n/\epsilon_n))^2, K(p_{G_0},p_G)\lesssim\epsilon_n^2(\log(M_n/\epsilon_n))$$

Following Lemma 3.7.5, $h^2(p_G,p_{G_0})\le\frac{W_1(G,G_0)}{2\sqrt{2\lambda_{min}}}$ for Laplace location mixtures.

Combining the above displays, Eq. (3.66) follows.

The following lemma is analogous to Lemma 7.1 in *Kleijn and van der Vaart* (2006) and Lemma 8.1 in *Ghosal et al.* (2000) and can be similarly proved.

Lemma 3.7.1. For every $M, \epsilon > 0$, $C > 0$, any measure Π on the set $\{G \in \overline{\mathcal{G}}(\Theta_n) : K(p_{G_0}, p_G) \lesssim \epsilon_n^2(\log(M/\epsilon_n)), K_2(p_{G_0}, p_G) \lesssim \epsilon_n^2(\log(M/\epsilon_n))^2\}$, we obtain that

$$P_{G_0}^n\biggl(\int \prod_{i=1}^n \frac{p_G(X_i)}{p_{G_*}(X_i)} d\Pi(G) \leq \exp(-(1+C)n(\epsilon^2 \log{(M/\epsilon)})\biggr) \leq \frac{1}{C^2 n\epsilon^2}. \tag{3.67}$$

The result in Eq.(3.65) now follows by application of Lemmas 3.7.1 and 3.7.2 in combination with Eq. (3.66) and (3.67) using the fact that $n\epsilon_n^2 \to \infty$.

Step 1.2: Let the event in (3.65) be denoted as T_n. Then

$$\begin{aligned}
& \mathbb{E}_{P_{G_0}}\left[\Pi_n(G : h(p_G, p_{G_0}) \geq L\epsilon_n)|X_1, \ldots, X_n)(1-\phi_n)\right] \leq P_{G_0}(T_n^C) \\
+ \quad & P_{G_0}(T_n) \frac{\exp((1+C)n\sqrt{\lambda_{min}}\epsilon_n^2 \log{(M_n/\epsilon_n)})}{\frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^{D_n}}{(2\Gamma(d/2+1))^{D_n}(2D_n)^{D_n-1}} \left(\frac{\sqrt{\lambda_{min}}\epsilon_n}{2\bar{\theta}_n}\right)^{2(D_n-1)+dD_n}} \\
& \qquad \times \sup_{G \in \overline{\mathcal{G}}(\Theta_n): h(p_G, p_{G_0}) \geq L\epsilon_n} P_G^n(1-\phi_n) \\
\lesssim \quad & \frac{\exp((1+C)n\sqrt{\lambda_{min}}\epsilon_n^2 \log{(M_n/\epsilon_n)})}{\frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^{D_n}}{(2\Gamma(d/2+1))^{D_n}(D_n)^{D_n-1}} \left(\frac{\sqrt{\lambda_{min}}\epsilon_n}{2\bar{\theta}_n}\right)^{(D_n-1)+dD_n}} \exp(-KnL^2\epsilon_n^2) + o(1).
\end{aligned} \tag{3.68}$$

The final step above follows from simple computation similar to that of the Proof of Theorem 2.1 in *Ghosal et al.* (2000) and using the fact that $\frac{\bar{\theta}_n^d}{\epsilon_n^{2d+2}} \log\left(\frac{\bar{\theta}_n}{\epsilon_n^2}\right) = o(n)$ because $\frac{\bar{\theta}_n^d}{\epsilon_n^{2d+2}} \log\left(\frac{\exp(\bar{\theta}_n)}{\epsilon_n^2}\right) = o(n)$ as per the condition of the Theorem. Combining Eq. (3.64) and (3.68) and using the condition $\frac{\bar{\theta}_n^d}{\epsilon_n^{2d+2}} \log\left(\frac{\exp(\bar{\theta}_n)}{\epsilon_n^2}\right) = o(n)$ it follows that for L large enough

$$\Pi_n(G \in \overline{\mathcal{G}}(\Theta_n) : h(p_G, p_{G_0}) \geq L\epsilon_n|X_1, \ldots, X_n) \overset{P_{G_0}}{\to} 0. \tag{3.69}$$

Now using Theorem 3.3.1 for Laplace location mixtures with $\beta = 2$, the result follows by noting that as $\bar{\theta}_n \uparrow \infty$ and $\epsilon_n \downarrow 0$, $\dfrac{1}{\log(1/L\epsilon_n)} \leq \bar{\theta}_n$ for n sufficiently large.

3.7.3 Prior mass on Wasserstein ball

Lemma 3.7.2. Let $G \sim DP(\gamma, H_n)$, where H_n admits condition (P.1) . Fix $r \geq 1$. Then the following holds, for any $G_0 \in \mathcal{P}(\Theta_n)$

$$\Pi\left(W_r^r(G, G_0) \leq (2^r + 1)\epsilon^r\right) \geq \frac{\Gamma(\gamma)(c_0\gamma\pi^{d/2})^{D_n}}{(2\Gamma(d/2+1))^{D_n}(2D_n)^{D_n-1}} \left(\frac{\epsilon}{2\bar{\theta}_n}\right)^{r(D_n-1)+dD_n}$$

for all ϵ sufficiently small so that $D(\epsilon, \Theta_n, \|.\|) > \gamma$. Here, $D_n = D(\epsilon, \Theta_n, \|.\|)$ stands for the maximal ϵ-packing number for Θ_n under $\|.\|$ norm, and $\Gamma(\cdot)$ is the gamma function.

Proof. From Lemma 5 in *Nguyen* (2013),

$$\Pi\left(W_r^r(G, G_0) \leq (2^r + 1)\epsilon^r\right) \geq \frac{\Gamma(\gamma)\gamma^{D_n}}{(2D_n)^{D_n-1}} \left(\frac{\epsilon}{\text{Diam}(\Theta_n)}\right)^{r(D_n-1)} \sup_S \prod_{i=1}^{D_n} H_n(S_i),$$

where, $S := (S_1, ..., S_{D_n})$ denotes the D_n disjoint $\epsilon/2$-balls that form a maximal ϵ-packing of Θ_n. The supremum is taken over all such packings. Now, $H_n(A) \geq \left(\dfrac{c_0}{\mu(\Theta_n)}\right)\mu(A)$. Moreover, $\prod_{i=1}^{D_n} \mu(S_i) \geq \left(\dfrac{(\sqrt{\pi}\epsilon)^d}{2\Gamma(d/2+1)}\right)^{D_n}$. Using this, we arrive at the result. □

3.7.4 Metric Entropy with Hellinger distance

Lemma 3.7.3. Let G_0 be a discrete mixing measure with all its atoms in $\Theta = [-\tilde{\theta}, \tilde{\theta}]^d \subset \mathbb{R}^d$. Let $\mathscr{P}_{\overline{\mathcal{G}}(\Theta)} := \{p_G : G \in \overline{\mathcal{G}}(\Theta)\}$. Then,

(i) if the kernel f is multivariate Gaussian with covariance matrix Σ,

$$\log D(\epsilon/2, \{p_G \in \mathscr{P}_{\overline{\mathcal{G}}(\Theta)} : \epsilon < h(p_G, p_{G_0}) \leq 2\epsilon\}, h) \\ \leq c_1 \left(\frac{\tilde{\theta}}{\sqrt{\lambda_{min}}\epsilon}\right)^d \log\left(e + \frac{32e\tilde{\theta}^2}{\lambda_{min}\epsilon^2}\right) \quad (3.70)$$

for some universal constant c_1.

(ii) if the kernel f is multivariate Laplace with covariance matrix Σ,

$$\log D(\epsilon/2, \{p_G \in \mathscr{P}_{\overline{\mathcal{G}}(\Theta)} : \epsilon < h(p_G, p_{G_0}) \leq 2\epsilon\}, h) \\ \leq c_2 \left(\frac{\tilde{\theta}}{\lambda_{min}\epsilon^2}\right)^d \log\left(e + \frac{16\sqrt{2}e\tilde{\theta}}{\lambda_{min}\epsilon^2}\right) \quad (3.71)$$

for some universal constant c_2.

Proof. Let $N(\epsilon, \mathscr{P}, d)$ denotes the ϵ-covering number of the space $\mathscr{P}$ relative to the metric d. It is related to the packing number by the following identity:

$$N(\epsilon, \mathscr{P}, h) \leq D(\epsilon, \mathscr{P}, d) \leq N(\epsilon/2, \mathscr{P}, h). \quad (3.72)$$

The proof is similar to the proof of Lemma **??** and is therefore omitted.

(i) Using the result in Example 1 of *Nguyen* (2013), when $f_\Sigma(\cdot|\theta) \sim \mathcal{N}_d(\theta, \Sigma)$,

$$h^2(f_\Sigma(\cdot|\theta_i), f_\Sigma(\cdot|\theta'_j)) = 1 - \exp\left(-\frac{1}{8}\|\theta_i - \theta'_j\|^2_{\Sigma^{-1}}\right) \leq \frac{\|\theta_i - \theta'_j\|^2}{8\lambda_{min}}, \quad (3.73)$$

where $\|z\|_{\Sigma^{-1}} := \sqrt{z'\Sigma^{-1}z}$.

Let $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ and $G = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$ be mixing measures in $\overline{\mathcal{G}}(\Theta)$, with

$k_0, k' \in [1, \infty]$. Let $\boldsymbol{q} = (q_{ij})_{1 \le i \le k_0, 1\ \le j \le k'} \in [0,1]^{k_0 \times k'}$ denote a coupling of $\boldsymbol{p^0}$ and $\boldsymbol{p'}$.

Using Lemma 1 of *Nguyen* (2013) with $\phi(x) = \frac{1}{2}(\sqrt{x} - 1)^2$, gives us:

$$h^2(p_G, p_{G_0}) \le \inf_{\boldsymbol{q} \in Q(\boldsymbol{p_0}, \boldsymbol{p'})} \sum_{i,j} q_{ij} \frac{\|\theta_i - \theta'_j\|^2}{8\lambda_{min}} = \frac{W_2(G, G_0)^2}{8\lambda_{min}}, \tag{3.74}$$

where $Q(\boldsymbol{p_0}, \boldsymbol{p'})$ is the set of all couplings of $\boldsymbol{p_0}$ and $\boldsymbol{p'}$. Therefore, it immediately follows that:

$$\begin{aligned} & \log D(\epsilon/2, \{p_G \in \mathscr{P}_{\overline{\mathcal{G}}(\Theta)} : \epsilon < h(p_G, p_{G_0}) \le 2\epsilon\}, h) \\ \le \ & \log D(\sqrt{2\lambda_{min}}\epsilon, \{G : G \in \overline{\mathcal{G}}(\Theta)0\}, W_2) \\ & \qquad \le N\left(\sqrt{\frac{\lambda_{min}}{8}}\epsilon, \Theta, \|\cdot\|\right) \log\left(e + \frac{32e\tilde{\theta}^2}{\lambda_{min}\epsilon^2}\right). \end{aligned}$$

The last inequality follows by applying Eq. (3.72) followed by Lemma 4 part (b) of *Nguyen* (2013). The result then follows immediately.

(ii) Consider the multivariate Laplace kernel of dimension d. Let $Y \sim f_\Sigma(\cdot|\theta)$, where $f_\Sigma(\cdot|\theta)$ is the multivariate Laplace density with location parameter θ and covariance matrix Σ as in Eq. (3.79). Then *Eltoft et al.* (2006) shows that :

$$Y = \theta + \sqrt{Z}X, \tag{3.75}$$

for some random variables $X \sim \mathcal{N}(0, \Sigma), Z \sim Exp(1)$.

Following Lemma 1 in *Nguyen* (2013),
$h^2(f_\Sigma(\cdot|\theta_i), f_\Sigma(\cdot|\theta'_j)) \le \mathbb{E}_Z h^2(g(\theta_i, Z\Sigma), g(\theta'_j, Z\Sigma))$, where $g(\theta, \Sigma) \sim \mathcal{N}(\theta, \Sigma)$.

Using the result in Eq. (3.73), we can show,

$$\begin{aligned}
h^2(f_\Sigma(\cdot|\theta_i), f_\Sigma(\cdot|\theta'_j)) &\leq 1 - \int_0^\infty \exp\left(-\frac{1}{8Z}\|\theta_i - \theta'_j\|^2_{\Sigma^{-1}}\right)\exp(-Z)\mathrm{d}Z \\
&\leq 1 - \int_b^\infty \exp\left(-\frac{1}{8Z}\|\theta_i - \theta'_j\|^2_{\Sigma^{-1}}\right)\exp(-Z)\mathrm{d}Z \\
&\leq 1 - \exp\left(-\frac{1}{8b}\|\theta_i - \theta'_j\|^2_{\Sigma^{-1}}\right)\exp(-b).
\end{aligned}$$

The above equation holds for any $b > 0$. Minimizing w.r.t. b we obtain,

$$h^2(f_\Sigma(\cdot|\theta_i), f_\Sigma(\cdot|\theta'_j)) \leq 1 - \exp\left(-\frac{1}{\sqrt{2}}\|\theta_i - \theta'_j\|_{\Sigma^{-1}}\right) \leq \frac{\|\theta_i - \theta'_j\|}{\sqrt{2}\lambda_{min}}.$$

It, therefore, immediately follows similar to Eq. (3.74) that

$$h^2(p_G, p_{G_0}) \leq \frac{W_1(G, G_0)}{\sqrt{2}\lambda_{min}}. \tag{3.76}$$

The result then follows similar to the steps in part (i).

□

3.7.5 Computation of M corresponding to KL ball

Lemma 3.7.4. Let G be a discrete mixing measure with all its atoms in $\left[-\tilde{\theta}, \tilde{\theta}\right]^d$ for some $\tilde{\theta} > 0$. Furthermore, assume the atoms of G_0 lie in $\left[-\bar{\theta}, \bar{\theta}\right]^d$ where $\bar{\theta} > 0$ is given. Then, the following holds:

(i) if the kernel f is multivariate Laplace,

$$\int \frac{(p_{G_0}(x))^2}{p_G(x)} \mu(\mathrm{d}x) \leq \exp\left(\frac{7}{4} d\lambda_{min}^{-1}(\bar{\theta} + \tilde{\theta})\right) V(\bar{\theta}), \tag{3.77}$$

for some continuous function $V(\cdot)$ independent of $\tilde{\theta}$.

(ii) if the kernel f is multivariate Gaussian,

$$\int \frac{(p_{G_0}(x))^2}{p_G(x)} \mu(\mathrm{d}x) \leq \exp(d\lambda_{min}^{-1}(5\bar{\theta}^2 + 4\tilde{\theta}^2)). \tag{3.78}$$

Here μ is the Lebesgue measure on $\mathbb{R}^d$.

Proof. (i) Following *Eltoft et al.* (2006), we consider the multivariate Laplace kernel of dimension $d > 1$, with mean parameter θ and variance Σ, given by:

$$f_\Sigma(x|\theta) := \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \frac{K_{(d/2)-1}(\sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)})}{\left[\sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}\right]^{(d/2)-1}}. \tag{3.79}$$

Here, $|\cdot|$ denotes the determinant of a matrix. Moreover, $K_n(x)$ denotes the modified Bessel function of the second kind with parameter n, evaluated at x.

For $d = 1$, the Laplace kernel of mean parameter θ and standard deviation parameter σ is given by:

$$f_\sigma(x|\theta) := \frac{1}{2\sigma} \exp\left(-\frac{|x-\theta|}{\sigma}\right) \tag{3.80}$$

We provide the solution for $d > 1$. The case $d = 1$ is trivial and will not be shown here. The function $K_n(z)$ does not possess a closed form but for $n > (-1/2)$ an integral form can be obtained as follows:

$$K_n(z) = \frac{\sqrt{\pi}}{(n-1/2)!}\left(\frac{z}{2}\right)^n \int\limits_1^\infty \exp(-zt)(t^2-1)^{n-1/2}\mathrm{d}t. \tag{3.81}$$

Let $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ and $G = \sum_{j=1}^{k'} p'_j \delta_{\theta'_j}$. Let $\boldsymbol{q} = (q_{ij})_{1\le i\le k_0, 1\ \le j\le k'} \in [0,1]^{k_0\times k'}$ denote a coupling of $\boldsymbol{p^0}$ and $\boldsymbol{p'}$.

Using lemma 2 of *Nguyen* (2013) with $\phi(x) = \dfrac{1}{x}$, gives us:

$$\int \frac{(p_{G_0}(x))^2}{p_G(x)}\mu(\mathrm{d}x) \le \inf_{\boldsymbol{q}\in Q(\boldsymbol{p_0},\boldsymbol{p'})} \sum_{i,j} q_{ij} \int \frac{(f_\Sigma(x|\theta_i^0))^2}{f_\Sigma(x|\theta'_j)}\mu(\mathrm{d}x), \tag{3.82}$$

where $Q(\boldsymbol{p_0},\boldsymbol{p'})$ is the set of all couplings of $\boldsymbol{p_0}$ and $\boldsymbol{p'}$.

Denote $\sqrt{(x-\theta)'\Sigma^{-1}(x-\theta)}$ as $\|x-\theta\|_{\Sigma^{-1}}$. By explicit computation,

$$\begin{aligned}
\frac{f_\Sigma(x|\theta_i^0)}{f_\Sigma(x|\theta'_j)} &= \left[\frac{K_{(d/2)-1}\left(\|x-\theta_i^0\|_{\Sigma^{-1}}\right)}{K_{(d/2)-1}\left(\|x-\theta'_j\|_{\Sigma^{-1}}\right)}\right]\left[\frac{\|x-\theta'_j\|_{\Sigma^{-1}}}{\|x-\theta_i^0\|_{\Sigma^{-1}}}\right]^{(d/2)-1} \\
&= \frac{\int_1^\infty \exp(-t\|x-\theta_i^0\|_{\Sigma^{-1}})(t^2-1)^{(d/2)-(3/2)}\mathrm{d}t}{\int_1^\infty \exp(-t\|x-\theta'_j\|_{\Sigma^{-1}})(t^2-1)^{(d/2)-(3/2)}\mathrm{d}t} \\
&\le \frac{\int_1^\infty \exp(-t\|x-\theta_i^0\|_{\Sigma^{-1}})(t^2-1)^{(d/2)-(3/2)}\mathrm{d}t}{\int_{3/2}^{7/4} \exp(-t\|\theta_i^0-\theta'_j\|_{\Sigma^{-1}})\exp(-t\|x-\theta_i^0\|_{\Sigma^{-1}})(t^2-1)^{(d/2)-(3/2)}\mathrm{d}t} \\
&\le \exp(\frac{7}{4}\|\theta_i^0-\theta'_j\|_{\Sigma^{-1}})\underbrace{\left(\frac{\int_1^\infty \exp(-t\|x-\theta_i^0\|_{\Sigma^{-1}})(t^2-1)^{(d/2)-(3/2)}\mathrm{d}t}{\int_{3/2}^{7/4} \exp(-\frac{7}{4}\|x-\theta_i^0\|_{\Sigma^{-1}})(t^2-1)^{(d/2)-(3/2)}\mathrm{d}t}\right)}_{:=T_\Sigma(x|\theta_i^0)}
\end{aligned}$$

The second equality in the above equation follows from Eq. (3.81) for $n = d/2-1$, while the second inequality follows by the triangle inequality for the norm $\|\cdot\|_{\Sigma^{-1}}$ and the fact that $\int_1^\infty \exp(-t\|x-\theta'_j\|_{\Sigma^{-1}})(t^2-1)^{n-1/2}\mathrm{d}t \geq \int_{3/2}^{7/4} \exp(-t\|x-\theta'_j\|_{\Sigma^{-1}})(t^2-1)^{n-1/2}\mathrm{d}t$.

Eltoft et al. (2006) shows that the Bessel function satisfies

$$K_d(x) \sim \sqrt{\frac{\pi}{2x}}\exp(-x) \tag{3.83}$$

as $|x| \to \infty$. Using this fact, and the fact that $\theta_i^0 \in [-\bar{\theta},\bar{\theta}]^d$ it is easy to show that

$$\int f_\Sigma(x|\theta_i^0)T_\Sigma(x|\theta_i^0)\mu(\mathrm{d}x) < V(\bar{\theta}) < \infty$$

for some function V depending only on $\bar{\theta}$.

It therefore follows from Eq. (3.82) that

$$\begin{aligned}\int \frac{(p_{G_0}(x))^2}{p_G(x)}\mu(\mathrm{d}x) &\leq \inf_{\boldsymbol{q}\in Q(\boldsymbol{p_0},\boldsymbol{p'})}\sum_{i,j} q_{ij}\exp\left(\frac{7}{4}\|\theta_i^0-\theta'_j\|_{\Sigma^{-1}}\right)V(\bar{\theta}) \\ &< \exp\left(\frac{7}{4}d\lambda_{min}^{-1}(\bar{\theta}+\tilde{\theta})\right)V(\bar{\theta}).\end{aligned}$$

The result for $d=1$ follows similarly.

(ii) For the multivariate Gaussian kernel with covariance matrix Σ, similar to the multivariate Laplace case, we get:

$$\int \frac{(p_{G_0}(x))^2}{p_G(x)}\mu(\mathrm{d}x) \leq \inf_{\boldsymbol{q}\in Q(\boldsymbol{p_0},\boldsymbol{p'})}\sum_{i,j} q_{ij}\int \frac{(f_\Sigma(x|\theta_i^0))^2}{f_\Sigma(x|\theta'_j)}\mu(\mathrm{d}x), \tag{3.84}$$

where $Q(\boldsymbol{p_0},\boldsymbol{p'})$ is the set of all couplings of $\boldsymbol{p_0}$ and $\boldsymbol{p'}$, and $f_\Sigma(\cdot|\theta)$ is the multivariate Gaussian kernel with covariance parameter Σ and mean parameter θ.

$$\begin{aligned}\int \frac{(f_\Sigma(x|\theta_i^0))^2}{f_\Sigma(x|\theta_j')}\mu(\mathrm{d}x) &= \int f_\Sigma(x|\theta_i^0)\exp\left(\frac{-\|x-\theta_i^0\|_{\Sigma^{-1}}^2 + \|x-\theta_j'\|_{\Sigma^{-1}}^2}{2}\right)\mu(\mathrm{d}x) \qquad (3.85)\\ &= \int f_\Sigma(x|\theta_i^0)\exp\left(\frac{-\|\theta_j'-\theta_i^0\|_{\Sigma^{-1}}^2}{2} + \langle x-\theta_j', \Sigma^{-1}\theta_j'-\theta_i^0\rangle\right)\mu(\mathrm{d}x),\end{aligned}$$

where the second equality follows by simple calculation using $x-\theta_i^0 = (x-\theta_j')+(\theta_j'-\theta_i^0)$.

If $M_\Sigma(t|\theta)$ is the moment generating function of the Gaussian distribution with mean θ and covariance Σ, then

$$M_\Sigma(t|\theta) = \exp(\langle\theta, t\rangle + \frac{1}{2}\langle t, \Sigma t\rangle).$$

Using this result , we can rewrite Eq. (3.86) as

$$\int \frac{(f_\Sigma(x|\theta_i^0))^2}{f_\Sigma(x|\theta_j')}\mu(\mathrm{d}x) \quad = \quad \exp(\langle\theta_j'-\theta_i^0, \Sigma^{-1}\theta_i^0+\theta_j'\rangle) \le \exp(2d\lambda_{min}^{-1}(\tilde{\theta}+\bar{\theta})^2 + d\lambda_{min}^{-1}\bar{\theta}^2),$$

The bound on $\int(p_{G_0}(x))^2/p_G(x)\mu(\mathrm{d}x)$ then follows immediately. □

3.7.6 Equivalence of Hellinger and Wasserstein metrics for Laplace location mixtures

Lemma 3.7.5. Let G, G' be discrete mixing measures. Also assume that f is the multivariate Laplace kernel as given in Eq. (3.5). Then the following holds:

$$h^2(p_G, p_{G'}) \le \frac{1}{2\sqrt{2\lambda_{min}}}W_1 G, G'), \qquad (3.86)$$

where λ_{min} is the minimum eigenvalue of Σ.

Proof. Suppose f_i, f_i' are Gaussian kernels with location parameter θ_i, θ_i' respectively, and Covariance matrix Σ. Then, as in proof of Lemma 3.7.3,

$$h^2(f_i, f_i') = 1 - \exp\left(-\frac{1}{8\lambda_{min}}\|\theta_i - \theta_i'\|^2\right).$$

Moreover, from *Eltoft et al.* (2006) we know that if a random variable Y has a Laplace distribution with location parameter θ and covariance matrix Σ, then Y is distributionally equivalent to:

$$Y \overset{d}{=} \theta + \sqrt{Z}X, \tag{3.87}$$

where X is a Gaussian random variable with mean 0 and covariance Σ and $Z \sim Exp(1)$ and is independent of X.

Using this formulation, if $g(\cdot|\theta, \Sigma)$ is the Laplace kernel then,

$$g(x|\theta, \Sigma) = \int_{\mathbb{R}} \exp(-z) f(x|\theta, z\Sigma) \mathrm{d}z. \tag{3.88}$$

Using the techniques similar to proof of 3.7.3 combined with Eq. (3.86), we obtain

$$\begin{aligned} h^2(g(\cdot|\theta_i,\Sigma),g(\cdot|\theta_i',\Sigma)) &\leq 1-\int\limits_{\mathbb{R}}\exp(-z)\exp\left(-\frac{1}{8z\lambda_{min}}\|\theta_i-\theta_i'\|^2\right)\mathrm{d}z \\ &\leq 1-\int\limits_{0}^{\frac{\|\theta_i-\theta_i'\|}{2\sqrt{2\lambda_{min}}}}\exp(-z)\mathrm{d}z \\ &\quad -\exp\left(-\frac{\|\theta_i-\theta_i'\|}{2\sqrt{2\lambda_{min}}}\right)\int\limits_{\frac{\|\theta_i-\theta_i'\|}{2\sqrt{2\lambda_{min}}}}^{\infty}\exp(-z)\mathrm{d}z \\ &\leq \exp\left(-\frac{\|\theta_i-\theta_i'\|}{2\sqrt{2\lambda_{min}}}\right)\left(1-\exp\left(-\frac{\|\theta_i-\theta_i'\|}{2\sqrt{2\lambda_{min}}}\right)\right) \\ &\leq \frac{\|\theta_i-\theta_i'\|}{2\sqrt{2\lambda_{min}}}. \end{aligned} \tag{3.89}$$

The conclusion of the lemma follows similar to Lemma 3.7.3. □

CHAPTER IV

Dirichlet Simplex Nest and Geometric Inference

We propose Dirichlet Simplex Nest, a class of probabilistic models suitable for a variety of data types, and develop fast and provably accurate inference algorithms by accounting for the model's convex geometry and low dimensional simplicial structure. By exploiting the connection to Voronoi tessellation and properties of Dirichlet distribution, the proposed inference algorithm is shown to achieve consistency and strong error bound guarantees on a range of model settings and data distributions. The effectiveness of our model and the learning algorithm is demonstrated by simulations and by analyses of text and financial data.[1] [2].

4.1 Introduction

For many complex probabilistic models, especially those with latent variables, the probability distribution of interest can be represented as an element of a convex polytope in a suitable ambient space, for which model fitting may be cast as the problem of finding the extreme points of the polytope. For instance, a mixture density can be

[1] Code: `https://github.com/moonfolk/VLAD`

[2] This work has been published in *Yurochkin* et al.* (2019)

identified as a point in a convex set of distributions whose extreme points are the mixture components. In the well-known topic model (*Blei et al.* (2003)) for text analysis, a document corresponds to a point drawn from the topic polytope, its extreme points are the topics to be inferred. This convex geometric viewpoint provides the basis for posterior contraction behavior analysis of topic models, as well as developing fast geometric inference algorithms (*Nguyen* (2015); *Tang et al.* (2014); *Yurochkin and Nguyen* (2016); *Yurochkin et al.* (2017)).

The basic topic model – the Latent Dirichlet Allocation (LDA) of *Blei et al.* (2003), as well as the comparable finite admixtures developed in population genetics (*Pritchard et al.* (2000)) were originally designed for categorical data. However, there are many real world applications in which the convex geometric probabilistic modeling continues to be a sensible approach, even if observed measurements are no longer discrete-valued, but endowed with a variety of distributions. To expand the scope of admixture modeling for a variety of data types, we propose to study Dirichlet Simplex Nest (DSN), a class of probabilistic models that generalizes the LDA, and to develop fast and provably accurate inference algorithms by accounting for the model's convex geometry and its low dimensional simplicial structure.

The generative process given by a DSN is simple to describe: starting from a simplex $\mathscr{B}$ of K vertices embedded in a high-dimensional ambient space $\mathcal{S}$, one draws random points from the $\mathscr{B}$'s relative interior according to a Dirichlet distribution. Given each such point, a data point is generated according to a suitable probability kernel F. For the general simplex nest, $\mathcal{S}$ can be any vector space of dimensions $D \geq K - 1$, while the probability kernel F can be taken to be Gaussian, Multinomial, Poisson, etc, depending on the nature of the observed data (continuous, categorical or counts, resp.). If $\mathcal{S}$ is standard probability simplex, and F a Multinomial distribution over categories, then the model is reduced to the familiar LDA model of *Blei et al.* (2003).

Although several geometric aspects of the DSN can be found in a vast array of well-known models in the literature, they were rarely treated together. First, viewing data as noisy observations from the low-dimensional affine hull that contains $\mathscr{B}$, our model shares an assumption that can be found in both classical factor analysis and non-negative matrix factorization (NMF) models (*Lee and Seung* (2001)), as well as the work of *Anandkumar et al.* (2012); *Arora et al.* (2012b) arising in topic models. Second, the convex constraints (i.e., linear weights of a convex combination are non-negative and sum to one) are present in all latent variable probabilistic modeling, even though most dominant computational approaches to inference such as MCMC sampling (*Griffiths and Steyvers* (2004)) and variational inference (*Blei et al.* (2003); *Hoffman et al.* (2013); *Kucukelbir et al.* (2017)) do not appear to take advantage of the underlying convex geometry.

As is the case with topic models, scalable parameter estimation is a key challenge for the Dirichlet Simplex Nest. Thus, our main contribution is a novel inference algorithm that accounts for the convex geometry and low dimensionality of the latent simplex structure endowed with a Dirichlet distribution. Starting with an original geometric technique of *Yurochkin and Nguyen* (2016), we present several new ideas allowing for more effective learning of asymmetric simplicial structures and the Dirichlet's concentration parameter for the general DSN model, thereby expanding its applicability to a broad range of data distributions. We also establish statistical consistency and estimation error bounds for the proposed algorithm.

The chapter proceeds as follows. Section 4.2 describes Dirichlet Simplex Nest models and reviews existing geometric inference techniques. Section 4.3 elucidates the convex geometry of the DSN via its connection to the Voronoi Tessellation of simplices and the

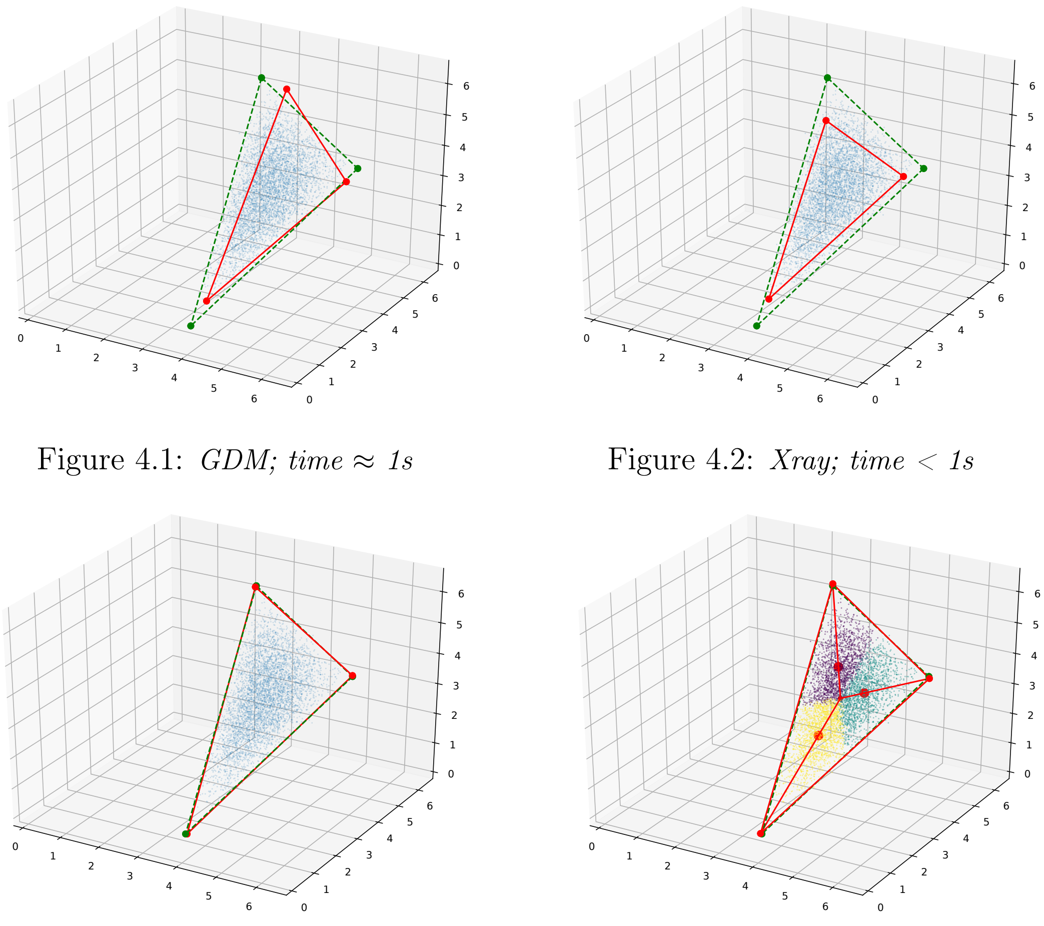


Figure 4.1: *GDM; time ≈ 1s*

Figure 4.2: *Xray; time < 1s*

Figure 4.3: *HMC; time ≈ 10m*

Figure 4.4: *VLAD; time < 1s*

Figure 4.5: *Toy simplex learning:* $n = 5000, D = 3, K = 3, \alpha = 2.5, \sigma = 0.1$.

structure of Dirichlet distribution on low-dimensional simplices. This helps motivate the proposed Voronoi Latent Admixture (VLAD) algorithm. Theoretical analysis of VLAD is given in Section 4.4. Section 4.5 presents an exhaustive comparative study on simulated and real data. We conclude with a discussion in Section 4.6.

4.2 Dirichlet Simplex Nest

We proceed to formally describe Dirichlet Simplex Nest as a generative model. Let $\beta_1, \ldots, \beta_K \in \mathcal{S}$ be K elements in a D-dimensional vector space $\mathcal{S}$, and define $\mathscr{B} = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$ as their convex hull. When $K \leq D+1$, $\mathscr{B}$ is a simplex in general positions. Next, for each $i = 1, \ldots, n$, generate a random vector $\mu_i \in \mathscr{B}$ by taking $\mu_i := \sum_{k=1}^{K} \theta_{ik}\beta_k$, where the corresponding coefficient vector $\theta_i = (\theta_{i1}, \ldots, \theta_{iK}) \in \Delta^{K-1}$ is generated by letting $\theta_i \sim \mathrm{Dir}_K(\alpha)$ for some concentration parameter $\alpha \in \mathbb{R}_+^K$. Now, given μ_i the data point x_i is generated by $x_i|\mu_i \sim F(\cdot \mid \mu_i)$, where F is a given probability kernel such that $\mathbb{E}[x_i \mid \theta_i] = \mu_i$ for any $i = 1, \ldots, n$.

Relation to existing models The DSN encompasses several existing models in the literature. If we set $\mathcal{S} := \Delta^{D-1}$ and likelihood kernel $F(\cdot)$ to Multinomial, then we recover the LDA model (*Blei et al.* (2003)). Other specific instances include Gaussian-Exponential (*Schmidt et al.* (2009)) and Poisson-Gamma models (*Cemgil* (2009)).

Estimating $\mathscr{B}$ is a challenging task for the general Dirichlet Simplex Nest model. Taking the perspective of Bayesian inference, a standard MCMC implementation for the DSN is likely computationally inefficient. In the case of LDA, as noted in *Yurochkin and Nguyen* (2016), the inefficiency of posterior inference can be traced to the need for approximating the posterior distributions of the large number of latent variables representing the topic labels. With the DSN model, we bypass the representation of such latent variables by integrating them out, but doing so at the cost of losing conjugacy. An alternative technique is variational inference (*Blei et al.* (2017); *Paisley et al.* (2014)). While very fast, this powerful method may be inaccurate in practice and does not carry a strong theoretical guarantee.

Relation to NMF and archetypal analysis The DSN provides a probabilistic justification for these methods, which often impose an additional geometric condition on the model known as *separability* that identifies the model parameters in a way that permits efficient estimation (*Donoho and Stodden* (2003); *Arora et al.* (2012a); *Gillis and Vavasis* (2014)). Separability is somewhat related to a control on the Dirichlet's concentration parameter α, by setting α be sufficiently small. The DSN allows for a probabilistic description of the nature of the separation. Moreover, by addressing *also* the case where α is large, the DSN modeling provides an arguably more effective approach to archetypal analysis and non-negative matrix factorization for *non-separable* data. We remark that an approach proposed by *Huang et al.* (2016) also permits a more general geometric identification condition called sufficiently scattered, but this generality comes at the expense of efficient estimation.

Geometric inference Geometric Dirichlet Means (GDM) algorithm of *Yurochkin and Nguyen* (2016) is a geometric technique for estimating the (topic) simplex $\mathscr{B}$ that arises in the Latent Dirichlet Allocation model. The basic idea of GDM is simple: performing the K-means clustering algorithm on the n points μ_i (or their estimates) to obtain K centroids. These centroids cannot be a good estimate for $\mathscr{B}$'s vertices, but they provide reasonable directions toward the vertices. Starting from the simplex's estimated centroid, the GDM constructs K line segments connecting to the K centroids and suitably extends the rays to provide an estimate for the K vertices. The GDM method is shown to be accurate when either $\mathscr{B}$ is equilateral, or the Dirichlet concentration parameter α is very small, i.e., most of the points μ_is are concentrated near the vertices. The quality of the estimates deteriorates in the absence of such conditions.

The deficiency of the GDM algorithm can be attributed to several factors: first, for a general simplex, the K-means centroids and the simplex's vertices do not line up. Fortunately, we will see that they may be lined up in a straight line by a suitable affine transformation of the simplex structure. Second, the nature of the Dirichlet distribution on the simplex is not pro-actively exploited, including that of parameter α. Third, typically $K \ll D$, the affine hull of $\mathscr{B}$ is a very low-dimensional structure, a fact not utilized by the GDM algorithm. It turns out that these shortcomings may be overcome by a careful consideration of the geometric structure of the simplex and the Dirichlet distribution.

For illustrations, we consider a toy problem of learning extreme points of simplex $\mathscr{B}$, given Gaussian data likelihood $x_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma^2 I_D)$ and $D = K = 3$. The triangle is chosen to be non-equilateral and Dirichlet concentration parameter is set to $\alpha = 2.5$. Figure 4.1 illustrates the deteriorating performance of the GDM. In Figure 4.2, we also observe Xray (*Kumar et al.* (2013)), another recent NMF algorithm, failing to solve the problem, as the aforementioned separability assumption is violated for large α. On the other hand, Figure 4.3 demonstrates the high accuracy of the posterior mean obtained by Hamiltonian Monte Carlo (HMC) (*Neal et al.* (2011); *Hoffman and Gelman* (2014)) implemented using Stan (*Carpenter et al.* (2017)), albeit at the cost of 10 minutes training time. Lastly our new algorithm (VLAD) in Fig. 4.4, exhibits an accuracy comparable to that of the HMC and the run-time of the GDM algorithm.

4.3 Inference of the Dirichlet Simplex Nest

4.3.1 Simplicial Geometry

In order to motivate our algorithm, we elucidate the geometry of the DSN through the concept of Centroidal Voronoi Tessellation (CVT) (*Du et al.* (1999)) of a simplex

$\mathscr{B}$, a convex subset of D-dimensional metric space $\mathcal{S}$.

Definition 4.3.1 (Centroidal Voronoi Tessellation)**.** Let $\Omega \subset \mathcal{S}$ be an open set equipped with a distance function d and a probability density ρ. For a set of K points $c_1, \ldots, c_K$, the Voronoi cell corresponding to c_k is the set

$$V_k = \{x \in \Omega : d(x, c_k) < d(x, c_l) \text{ for any } l \neq k\}.$$

The collection of Voronoi cells $V_1, \ldots, V_K$ is a tessellation of Ω; i.e. the cells are disjoint and $\cup_k V_k = \Omega$. If $c_1, \ldots, c_K$ are also the centroids of their respective Voronoi cells, i.e.,

$$c_k = \frac{1}{\int_{V_k} \rho(x) \mathrm{d}x} \int\limits_{V_k} x \rho(x) \mathrm{d}x$$

the tessellation is a Centroidal Voronoi Tessellation.

CVTs are special: any set of k points induces a Voronoi tessellation, but these points are generally not the centroids of their associated cells. One can check that a CVT minimizes

$$J(c_1, \ldots, c_K) = \int\limits_{V_k} d(x, c_k)^2 \rho(x) \mathrm{d}x.$$

It is a fact that J has a unique global minimizer as long as ρ vanishes on a set of measure zero, the Voronoi cells are convex, and the distance function is convex in each argument (*Du et al.* (1999)). Moreover, it can be seen that the centroids of the CVT of an equilateral simplex equipped with the $\mathrm{Dir}_K(\alpha)$ distribution fall on the line segments between the centroid of the simplex and the extreme points of the simplex, but this is not the case when the simplex shape is non-equilateral (cf. Fig. 4.1).

The following lemma formalizes the aforementioned insight to a simplex of arbitrary shape $\mathscr{B}$ by considering a suitably modified distance function $d(\cdot,\cdot)$ of the CVT. (In Fig. 4.4, the blue, purple and yellow dots are the sample versions of the Voronoi cells of the CVT under the new distance function and the corresponding centroids are in red.)

Lemma 4.3.1. Let $B \in \mathbb{R}^{D\times K}$ denote the matrix form of simplex $\mathscr{B}$. Suppose it has full (column) rank, equipped with distance function $\|\cdot\|_{(BB^T)^\dagger}$ and the probability distribution $\mathbb{P}_B$ defined as

$$\mathbb{P}_B(S) = \mathrm{Prob}(\{\theta \in \Delta^{K-1} : B\theta \in S\}),$$

where θ is distributed by symmetric Dirichlet density $\rho_\alpha := \mathrm{Dir}_K(\alpha)$, for any $S \subset \mathsf{int}(\mathscr{B})$, and $A^\dagger$ denotes a pseudo-inverse of A. The centroids of its CVT fall on the line segments connecting the centroid of $\mathscr{B}$ to $\beta_1,\ldots,\beta_K$.

Proof. Let $c_1,\ldots,c_K$ and $V_1,\ldots,V_K$ be the centroids and cells of the CVT of Δ^{K-1} equipped with Euclidean distance and $\mathrm{Dir}_K(\alpha)$ density ρ_α. It suffices to verify that $Bc_1,\ldots,Bc_K$ and $BV_1,\ldots,BV_K$ are the centroids and cells of the CVT of $\mathscr{B} = B\Delta^{K-1}$. By a change of variables formula,

$$\begin{aligned}
&\mathrm{argmin}\left\{\frac{\int_{BV_k}\|x-Bv\|^2_{(BB^T)^\dagger}\rho_\alpha(B^\dagger x)|\det(B^\dagger)|\mathrm{d}x}{\int_{V_k}\rho_\alpha(B^\dagger x)|\det(B^\dagger)|\mathrm{d}x} : v \in V_k\right\}\\
&= \mathrm{argmin}\left\{\frac{\int_{V_k}\|B\theta-Bv\|^2_{(BB^T)^\dagger}\rho_\alpha(\theta)\mathrm{d}\theta}{\int_{V_k}\rho_\alpha(\theta)\mathrm{d}\theta} : v \in V_k\right\}\\
&= \mathrm{argmin}\left\{\frac{\int_{V_k}\|\theta-v\|_2^2\rho_\alpha(\theta)\mathrm{d}\theta}{\int_{V_k}\rho_\alpha(\theta)\mathrm{d}\theta} : v \in V_k\right\},
\end{aligned}$$

which we recognize as the centroids of the CVT of Δ^{K-1} under ℓ_2 metric. Since Δ^{K-1} is a standard simplex and therefore equilateral, the centroids of the CVT of equilateral simplex fall on the line segments connecting the centroid of the simplex to its extreme points. □

Lemma 4.3.1 suggests an algorithm to estimate the extreme points of $\mathscr{B}$. First, estimate the centroids of the CVT of $\mathscr{B}$ (equipped with scaled Euclidean norm $\|\cdot\|_{(BB^T)^\dagger}$) and search along the rays extending from the centroid of $\mathscr{B}$ through the CVT centroids for the simplicial vertices.

4.3.2 The Voronoi Latent Admixture (VLAD) Algorithm

We first consider the noiseless problem, $F(\cdot \mid \mu) = \delta_\mu$. That is, $x_i = \mu_i$s are observed. In this case, Lemma 4.3.1 suggests estimating the CVT centroids by scaled K-means optimization:

$$\underset{c_1,\ldots,c_K}{\operatorname{argmin}}\big\{\tfrac{1}{2}\textstyle\sum_{k=1}^{K}\sum_{x_i\in V_k}(x_i - c_k)^T(BB^T)^\dagger(x_i - c_k)\big\}, \tag{4.1}$$

Unfortunately, the scaled Euclidean norm $\|\cdot\|_{(BB^T)^\dagger}$ is unknown. We propose an equivalent approach that does not depend on knowledge of BB^T.

In the noiseless case, observe that the population covariance matrix of the samples takes the form $\Sigma = BSB^T$, where S is the covariance matrix of a $\text{Dir}(\alpha)$ random variable on Δ^{K-1}. By the standard properties of the $\text{Dir}(\alpha)$ distribution, it can be seen that $S = \frac{1}{K(K\alpha+1)}P$, where $P = I_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T$ is the centering matrix. Hence, knowledge of Σ will be sufficient because the centered data points x fall in $\mathsf{span}(\Sigma) = \mathsf{span}(BPB^T)$: For each (θ, x) pair,

$$\bar{x} := \underbrace{B\theta}_{x} - \underbrace{\tfrac{1}{K}B\mathbf{1}}_{\mathbb{E}[x]} = B\theta - \tfrac{1}{K}B\mathbf{1}(\underbrace{\mathbf{1}^T\theta}_{=1}) = BP\theta := B\bar{\theta}. \tag{4.2}$$

This suggests that the centroids of the CVT may be recovered by clustering the centered data points in the $\|\cdot\|_{\Sigma^\dagger}$-norm. This insight is formalized by

Lemma 4.3.2. The centroids of the CVT of simplex $\mathscr{B}$ under $\|\cdot\|_{(BB^T)^\dagger}$-norm are given by $\{c_k^* + c_0 | k = 1, \ldots, K\}$, where $(c_1^*, \ldots, c_K^*)$ solves the minimization

$$\min_{\substack{c_1,\ldots,c_K \\ V_1,\ldots,V_K}} \frac{1}{2} \sum_{k=1}^{K} \int\limits_{x \in BV_k} (\bar{x} - c_k)^T \Sigma^\dagger (\bar{x} - c_k) \rho(x) \mathrm{d}x \tag{4.3}$$

and $c_0 = \int x\rho(x)\mathrm{d}x$ is the centroid of simplex $\mathscr{B}$.

Proof. We first show that (4.3) is equivalent to (unscaled) K-means clustering on Δ^{K-1}. Note that $\Sigma = \delta BPB^T$ for some $\delta > 0$. Without loss of generality, we restrict to c_k's in $\mathsf{span}\{BPB^T\}$. Write $c_k = BPv_k$ for $v_k \in \mathbb{R}^K$, for $k = 1, \ldots, K$. Recalling (4.2) and the fact P is a projector,

$$\begin{aligned}
& (1/\delta) \textstyle\sum_{k=1}^{K} \int_{x \in BV_k} (\bar{x} - c_k)^T \Sigma^\dagger (\bar{x} - c_k) \rho(x) \mathrm{d}x \\
= \;& \textstyle\sum_{k=1}^{K} \int_{\theta \in V_k} (\bar{\theta} - v_k)^T PB^T \Sigma^\dagger BP (\bar{\theta} - v_k) \rho_\alpha(\theta) \mathrm{d}\theta \\
= \;& \textstyle\sum_{k=1}^{K} \int_{\theta \in V_k} (\bar{\theta} - v_k)^T P (\bar{\theta} - v_k) \rho_\alpha(\theta) \mathrm{d}\theta \\
= \;& \textstyle\sum_{k=1}^{K} \int_{\theta \in V_k} \|\bar{\theta} - Pv_k\|_2^2 \rho_\alpha(\theta) \mathrm{d}\theta. \qquad (4.4)
\end{aligned}$$

Since θ is distributed by the symmetric Dirichlet $\rho_\alpha = \mathrm{Dir}(\alpha)$ on Δ^{K-1}, the last equality entails that the optimal v_k's are the points which represent the barycentric coordinate of the centroids of the CVT of Δ^{K-1}. Thus, the optimal solution for $c_k = BPv_k$ represents the centroids of the CVT of simplex $\mathscr{B}$ under $\|\cdot\|_{(BB^T)^\dagger}$-norm (using the coordinating system that is centered at origin c_0). □

We proceed to address the optimization (4.3) applied to empirical data to arrive at Voronoi Latent Admixture (VLAD) algorithm in Algorithm 4.1. We utilize the singular value decomposition (SVD) of the centered data points to simplify computation. Let

$\bar{X} \in \mathbb{R}^{n\times D}$ be the matrix whose rows are the centered data points and $\bar{X} = U\Lambda W^T$ be its SVD. Each term in the objective of (4.3) is equivalent to, with Σ being replaced by its empirical version, $\Sigma_n = \frac{1}{n} W\Lambda^2 W^T$:

$$\begin{aligned}&(\bar{x}_i - c_k)^T \Sigma_n^{\dagger} (\bar{x}_i - c_k) = \\ &n(u_i - \eta_k)^T \Lambda W^T W \Lambda^{-2} W^T W \Lambda (u_i - \eta_k) = n\|u_i - \eta_k\|_2^2,\end{aligned}$$

where $\bar{x}_i = W\Lambda u_i$, and set $c_k = W\Lambda\eta_k$. Thus, instead of performing scaled K-means clustering in $\mathcal{S}$, it suffices to perform standard K-means in the low $(K-1)$ dimensional space. This yields a significant computational speed-up. After applying VLAD, the weights θ_i's can be obtained by projecting the data points onto $\mathscr{B}$ and compute the barycentric coordinates of the projected points.

Algorithm 4.1 Voronoi Latent Admixture (VLAD)

```
Input: data $x_1, \ldots, x_n$; $K$; extension parameter $\gamma$.
Output: simplex vertices $\beta_1, \ldots, \beta_K$
$\widehat{c}_0 \leftarrow \frac{1}{n}\sum_i x_i$ {find data center}
$\bar{x}_i \leftarrow x_i - \widehat{c}_0$, $i = 1, \ldots, n$ {centering}
compute top $K-1$ singular factors of the centered data matrix $\bar{X} \in \mathbb{R}^{n\times D}$: $\bar{X} = U\Lambda W^T$
$\eta_1, \ldots, \eta_K \leftarrow$ K-means$(u_1, \ldots, u_n)$, where the $u_i$'s are the rows of $U \in \mathbb{R}^{n\times(K-1)}$
$\widehat{c}_k \leftarrow W\Lambda\eta_k + \widehat{c}_0$
$\widehat{\beta}_k \leftarrow \widehat{c}_0 + \gamma(\widehat{c}_k - \widehat{c}_0)$
```

It remains to estimate the extreme points β_ks given the CVT centroids c_ks. This task is simplified by two observations: First, the CVT centroids reside on the line segment between the centroid of simplex $\mathscr{B}$ and its extreme points, per Lemma 4.3.1. Thus we merely need to estimate the ratio of the distance from the extreme point to the centroids of $\mathscr{B}$ and the distance from the CVT centroids to the centroid of $\mathscr{B}$. Due to the symmetry of $\text{Dir}_K(\alpha)$ distribution on Δ^{K-1}, this ratio is identical for all extreme points – we refer to this ratio as the extension parameter γ. Secondly, γ does not depend on the geometry of $\mathscr{B}$, only that of the Dirichlet distribution. Thus, γ can be easily estimated by appealing to a Monte Carlo technique on Dir_K. This subroutine is summarized in Algorithm 4.2, provided that α is given.

Algorithm 4.2 Evaluating extension parameters

```
1: generate $\theta_1, \dots, \theta_m \sim \mathsf{Dir}_K(\alpha)$, where $m$ is the number of Monte Carlo samples
2: $v_1, \dots, v_K \leftarrow$ K-means$(\theta_1, \dots, \theta_m)$
3: $\gamma \leftarrow \sqrt{K^2 - K} \left( \sum_{l=1}^{K} \|v_l - \frac{1}{K}\mathbf{1}_K\|_2 \right)^{-1}$
```

4.3.3 Estimating the Dirichlet Concentration Parameter

Next, we describe how to estimate concentration parameter α from the data, by employing a moment-based approach. Recall from the previous section that there is an one-to-one mapping between α and the extension parameter γ. For each $\alpha > 0$, let $\gamma(\alpha) > 0$ denote the corresponding extension parameter and $B(\gamma) \in \mathbb{R}^{D \times K}$ the estimator of B output by VLAD with extension parameter γ. In the absence of noise, the covariance matrix of the DSN model has the form $BS(\alpha)B^T$, where $S(\alpha) \in \mathbb{R}^{K \times K}$ is the covariance matrix of a $\mathrm{Dir}(\alpha)$ random variable on Δ^{K-1}. This suggests we estimate α by a generalized method of moments approach:

$$\hat{\alpha} = \operatorname*{argmin}_{\alpha > 0} \|\hat{B}(\gamma(\alpha))S(\alpha)\hat{B}(\gamma(\alpha))^T - \hat{\Sigma}\|, \tag{4.5}$$

where $\hat{\Sigma}$ is the sample covariance matrix $\hat{\Sigma} = \frac{1}{n}\bar{X}^T\bar{X}$. We remark that there is no need to run VLAD multiple times to evaluate the objective in (4.5) at multiple α-values. After VLAD is run once, we may evaluate $\gamma(\alpha)$ for any value of γ by affinely transforming the output of VLAD. Further, (4.5) is a scalar optimization problem, so the computational cost of solving (4.5) is negligible.

In the presence of noise, the covariance matrix of the DSN model no longer has the form $BS(\alpha)B^T$. We need to add a correction term to ensure a consistent estimator of $BS(\alpha)B^T$. For example, if the noise is Gaussian, a consistent estimator of $BS(\alpha)B^T$ is

$$\tilde{\Sigma} = \hat{\Sigma} - \hat{\sigma}^2 I_D,$$

where $\hat{\sigma}^2$ is an estimate of the noise variance. In Supplement 4.7.1.3, we give consistent estimators of $BS(\alpha)B^T$ for multinomial and Poisson noise. With a good estimator $\tilde{\Sigma}$ of $BS(\alpha)B^T$ in place, we replace $\hat{\Sigma}$ in (4.5) by $\tilde{\Sigma}$ and then solve (4.5) to obtain an estimate of α.

4.4 Consistency and Estimation Error Bounds

In this section we establish consistency properties and error bound guarantees of the VLAD procedure.

For $c = (c_1, \ldots, c_K) \in \mathbb{R}^{K \times D}$, define $\phi_A : \mathbb{R}^D \times \mathbb{R}^{K \times D} \to \mathbb{R}$ as

$$\phi_A(x, c) = \min_{k \in \{1, \ldots, K\}} \|x - c_k\|^2_{A^\dagger}$$

where A is a positive semidefinite matrix. Recall Σ as the covariance matrix of the data generating distribution and Σ_n its empirical counterpart. In the algorithm, we work with the best rank $K-1$ approximation of Σ_n, which we denote by $(\Sigma_n)^K$. Let $\mathbb{Q}$ denote the distribution for μ_is. Recall that $X_i|\mu_i \sim F(\cdot|\mu_i)$. Let $\mathbb{P}$ be the induced distribution corresponding to $\tilde{X}_i$, which is the projection of X_i on the affine space of dimension $K-1$ spanned by the top $K-1$ eigenvectors of Σ. We also use $\mathbb{P}_n$ to denote the empirical distribution of the data represented by random variables X_i.

Since K-means clustering is a subroutine of our algorithm, we expect at least some sort of condition requiring that the K-means clustering routine be well-behaved in some sense. To that end we need the following standard condition on the population K-means objective (*Pollard* (1981)).

(a.1) Pollard's regularity criterion (PRC): The Hessian matrix of the function $c \mapsto \mathbb{Q}\phi_{BSB^T}(\cdot, c)$ evaluated at c^* for all optimizers c^* of $\mathbb{Q}\phi_{BSB^T}(\cdot, c)$ is positive definite, with minimum eigenvalue $\lambda_0 > 0$.

It turns out that this will be all we need for the following theorem in the noiseless setting, where we have $\Sigma = BSB^T = (\Sigma)^K$ has rank $K-1$ and so, $\mathbb{P} = \mathbb{Q}$ and $\tilde{X}_i \overset{\mathscr{L}}{=} X_i$.

Theorem 4.4.1. Consider the noiseless setting, i.e., $F(\cdot \mid \mu) = \delta_\mu$. Suppose that $\mathscr{B} = \text{Conv}(\beta_1, \ldots, \beta_K)$ is the true topic simplex, while $(\beta_{1n}, \ldots, \beta_{Kn})$ are the vertex estimates obtained by VLAD algorithm. Moreover, assume the error due to Monte Carlo estimates of the extension parameter is negligible. Provided that condition (a.1) holds,

$$\min_\pi \|(\beta_{\pi_{(1)}n}, \ldots, \beta_{\pi_{(K)}n}) - (\beta_1, \ldots, \beta_K)\| = O_{\mathbb{P}}(n^{-1/2}),$$

where the minimization is taken over all permutations π of $\{1, \ldots, K\}$.

Note that the constant corresponding to the rate $O_{\mathbb{P}}(n^{-1/2})$ is dependent on the Hessian matrix of the function $c \mapsto \mathbb{P}\phi_\Sigma(\cdot, c)$. The proof for Theorem 4.4.1 is in Supplement 4.7.1.1.

In general, $F(\cdot \mid \mu)$ is not degenerate. Due to the presence of "noise" in the $K-1$ SVD subspace, the estimates of the CVT centroids may be inconsistent, which entails inconsistency of the VLAD's estimate for $\mathscr{B}$. The following theorem provides an error bound in the general setting. We need a strengthening of Pollard's Regularity Criterion. Let $(\Sigma)^K$ denote the best $K-1$ rank approximation of Σ with respect to the Frobenius norm. Assume:

(a.2) The Hessian matrix of the function $c \mapsto \mathbb{P}\phi_{(\Sigma)^K}(\cdot, c)$ evaluated at c^* for all optimizers c^* of $\mathbb{P}\phi_{(\Sigma)^K}(\cdot, c)$ is uniformly positive definite with minimum eigenvalue $\lambda_0 > 0$, for all $(\Sigma)^K$ such that $(\Sigma - BSB^T) \leq \tilde{\epsilon} I_D$, for some $\tilde{\epsilon} > 0$.

The noise level is formalized by the following conditions:

(b) There is $\epsilon_0 > 0$ such that $\epsilon_0 I_D - Cov(X|\theta)$ is positive semi-definite uniformly over $\theta \in \Delta^{K-1}$.

(c) There exists M_0 such that for all $M > M_0$, $\int_{\mathcal{B}(\sqrt{M}, c_0)^c} \|x - c_0\|_2^2 g(x) \mathrm{d}x \leq \frac{k_1}{M}$, for some universal constant k_1, where $\mathcal{B}(\sqrt{M}, c_0)$ is a ball of radius $\sqrt{M}$ around population centroid c_0 and $g(\cdot)$ is the density of $\mathbb{P}$ with respect to the Lebesgue measure on the $K-1$ dimensional space which contains the top $K-1$ eigenvectors of $BSB^T + \epsilon_0 I_D$.

Theorem 4.4.2. Suppose that $\mathscr{B} = \mathrm{Conv}(\beta_1, \ldots, \beta_K)$ is simplex corresponding to extreme points of the DSN. Let $(\beta_{1n}, \ldots, \beta_{Kn})$ be the corresponding extreme point estimates obtained by the VLAD algorithm. Assume the error in the Monte Carlo estimates of the extension parameter is negligible. Provided that (a.2), (b) and (c) hold, then

$$\min_\pi \|(\beta_{\pi_{(1)}n}, \ldots, \beta_{\pi_{(K)}n}) - (\beta_1, \ldots, \beta_K)\|_2 = O\left(\sqrt{\epsilon_0^{1/3}/\lambda_0}\right) + O_{\mathbb{P}}(n^{-1/2}), \tag{4.6}$$

where π ranges over permutations of $\{1, \ldots, K\}$.

The constant corresponding to the rate $O_{\mathbb{P}}(n^{-1/2})$ in the theorem depends on the Hessian matrix of the function $c \mapsto \mathbb{P}\phi_{\Sigma}(\cdot, c)$; constant corresponding to the $O\left(\sqrt{\epsilon_0^{1/3}/\lambda_0}\right)$ depends on the minimum and maximum eigenvalues of the matrix BSB^T. Proof is in Supplement 4.7.1.2.

The preceding results control the error incurred by the VLAD algorithm when the concentration parameter α is known. When α is unknown, our proposed solution in Section 4.3.3 performs well in both simulated and real-data experiments. We do not know in theory whether the concentration parameter α is identifiable, we shall present empirical results in Supplement 4.7.1.5 which suggest identifiability. Assuming a condition which guarantees model identifiability, we can establish that the estimate obtained by the VLAD algorithm via (4.5) is in fact consistent.

Theorem 4.4.3. Assume that function $\varphi(\tilde{\alpha}) = \frac{\gamma(\tilde{\alpha})^2}{K(K\tilde{\alpha}+1)}$ is monotonically increasing in $\tilde{\alpha}$, where $\gamma(\tilde{\alpha})$ is the extension parameter corresponding to $\tilde{\alpha}$. Let $\alpha_0 \in \mathscr{C}$ be the true concentration parameter for some compact set $\mathscr{C}$. Let $\hat{\alpha}_n = \operatorname{argmin}_{\alpha \in \mathscr{C}} \|\hat{B}(\gamma(\alpha))S(\alpha)\hat{B}(\gamma(\alpha))^T - \tilde{\Sigma}_n\|$, where $\tilde{\Sigma}_n$ is a consistent estimator of $BS(\alpha)B^T$. Then,

$$\|\hat{\alpha}_n - \alpha_0\| \xrightarrow{\mathbb{P}} 0. \tag{4.7}$$

See Supplement 4.7.1.4 for the proof.

4.5 Experiments

The goal of our experimental studies is to demonstrate the applicability and efficiency of our algorithm for a number of choices of the DSN probability kernel: Gaussian, Poisson and Multinomial (i.e. LDA). We summarize all competing estimation procedures in our comparative study and their corresponding underlying assumptions in Table 4.1.

We remark that Gibbs sampler (*Griffiths and Steyvers* (2004)), Stan implementation of No U-Turn HMC (*Hoffman and Gelman* (2014); *Carpenter et al.* (2017)) and Stochastic Variational Inference (SVI) (*Hoffman et al.* (2013)) may be augmented with techniques such as empirical Bayes to estimate hyperparameter α, although it may slow down convergence. We instead allow these baselines to use true values of α in all simulated experiments to their advantage; when latent simplex is of general geometry (i.e. non-equilateral), GDM (*Yurochkin and Nguyen* (2016)) requires $\alpha \to 0$ to perform well, which is alike separability. Not all baselines are suitable for all three probability kernels, i.e. Gibbs sampler and SVI rely on (local) conjugacy and are only applicable in the LDA scenario; RecoverKL by *Arora et al.* (2013) is an algorithm that relies on a separability condition (i.e. anchor words) designed for topic models.

In simulated experiments we will consider both VLAD with estimated concentration parameter α following our results in Section 4.3.3 and VLAD trained with the knowledge of true data generating α (VLAD-α). For real data analysis, we estimate the concentration parameter by (4.5) and apply VLAD to a text corpus and stock market data set.

4.5.1 Comparative Simulation Studies

Convergence behavior We investigate the convergence of the estimates of the DSN extreme points for the three likelihood kernels under the increasing sample size. The hyperparameter settings are $D = 500, K = 10, \alpha = 2$ (for LDA vocabulary size $D = 2000$). To ensure non-trivial geometry of the DSN we rescale extreme points towards their mean by uniform random factors between 0.5 and 1. We use the Minimum Matching distance - a metric previously studied in the context of polytopes estimation (*Nguyen*

Table 4.1: *Baselines and required conditions*

Method	Conjugacy	True α	Separability
VLAD (this work)	$\times$	$\times$	$\times$
VLAD-α (this work)	$\times$	$\surd$	$\times$
Gibbs (*Griffiths and Steyvers* (2004))	$\surd$	$\surd^\star$	$\times$
Stan-HMC (*Carpenter et al.* (2017))	$\times$	$\surd^\star$	$\times$
SVI (*Hoffman et al.* (2013))	$\surd$	$\surd^\star$	$\times$
GDM (*Yurochkin and Nguyen* (2016))	$\times$	$\times$	$\surd^\star$
RecoverKL (*Arora et al.* (2013))	$\times$	$\times$	$\surd$
SPA (*Gillis and Vavasis* (2014))	$\times$	$\times$	$\surd$
MVES (*Chan et al.* (2009))	$\times$	$\times$	$\surd$
Xray (*Kumar et al.* (2013))	$\times$	$\times$	$\surd$

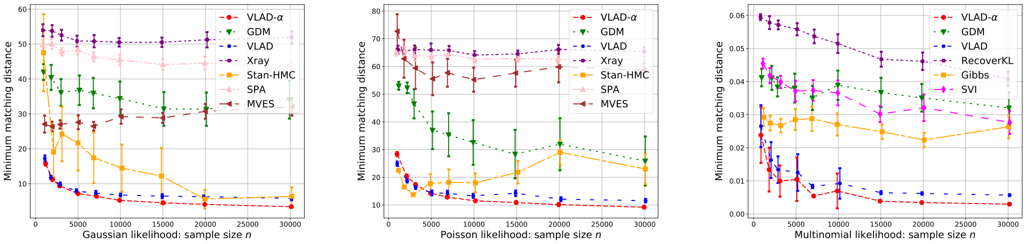


Figure 4.6: *Gaussian data* Figure 4.7: *Poisson data* Figure 4.8: *Categorical data*

Figure 4.9: *Minimum matching distance for increasing n*

(2015)) to compare the quality of the fitted DSN model returned by a variety of inference algorithms. We defer additional details to Supplement 4.7.2.

In Fig. 4.9 we see that VLAD and VLAD-α significantly outperform all baselines. Further, the estimation error reduces with increased sample size verifying statements of Theorems 4.4.2 and 4.4.3. We note that Stan HMC may also achieve good performance, however it is very costly to fit (e.g., 40 HMC iterations for Poisson case and $n = 30000$ took 14 hours compared to 7 seconds for VLAD), therefore we had to restrict number of iterations, which explains its wider error bars across experiments.

Geometry of the DSN To study the role of geometry of the DSN we rescale

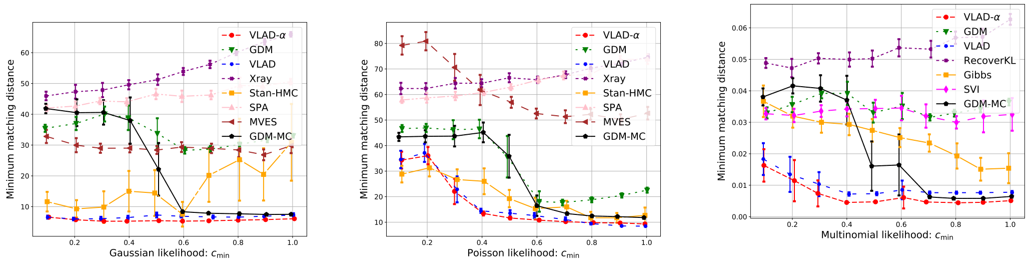


Figure 4.10: *Gaussian data* Figure 4.11: *Poisson data* Figure 4.12: *Categorical data*

Figure 4.13: *Minimum matching distance for varying DSN geometry.*

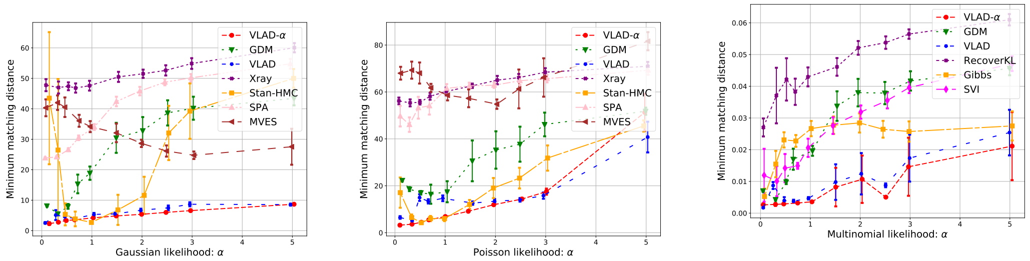


Figure 4.14: *Gaussian data* Figure 4.15: *Poisson data* Figure 4.16: *Categorical data*

Figure 4.17: *Minimum matching distance for increasing α.*

extreme points towards their mean by uniform random factors $c_k \sim \text{Unif}(c_{\min}, 1)$ for $k = 1, \ldots, K$ and vary $c_{\min}$ in Fig. 4.13 (smaller values imply more severe skewness of the latent simplex). To isolate the effect of the geometry of the DSN, we compare to GDM combined with knowledge of true α and extension parameter estimation using Algorithm 4.2 (GDM-MC). If the underlying simplex is equilateral, GDM-MC will be equivalent to VLAD-α.

In Fig. 4.13 we see that VLAD and VLAD-α are robust to varying skewness of the DSN. On the contrary, GDM-MC is only accurate when the latent simplex becomes closer to equilateral. This experiment verifies geometric motivation of our work — in practice we can not expect latent geometric structure to be necessarily equilateral and geometrically robust method such as VLAD is more reliable.

Table 4.2: *NYT topic modeling (categorical data)*

	Perplexity	Coherence	Time
VLAD	1767	0.86	**6min**
GDM	1777	0.88	30min
Gibbs || HMC	**1520**	0.80	5.3hours
RecoverKL || MVES	2365	0.70	17min
SVI || SPA	1669	0.81	40min

Table 4.3: *Stock data factor analysis (continuous data)*

	Frobenius norm	Volume	Time
VLAD	0.300	**0.14**	**1s**
GDM	0.294	1499	**1s**
Gibbs || HMC	0.299	1.95	10min
RecoverKL || MVES	0.287	5.39×10^9	3min
SVI || SPA	0.392	3.31×10^7	**1s**

Varying Dirichlet prior To complete our simulation studies we verify α estimation procedure proposed in Section 4.3.3 and analyzed in Theorem 4.4.3. It is also interesting to compare performance of other baselines for larger α — scenario often overlooked in the literature.

In Fig. 4.17 (and in previous experiments) we see that performance gap between VLAD and VLAD-α is very small, supporting effectiveness of our α estimation procedure across probability kernels. Additionally, we see that higher values of α lead to degrading performance of all considered methods, however VLAD degrades more gracefully.

4.5.2 Real Data Analysis

Topic modeling We analyze a collection of news articles from the New York Times. After preprocessing, we have 5320 unique words and 100k training documents with 25k left out for perplexity evaluation. We also compare semantic coherence of the topics (*Newman et al.* (2010)).

In Table 4.2 (left) we present results for $K = 80$ topics. The Gibbs sampler has the best perplexity score, but it falls behind in topic coherence. VLAD estimated $\alpha = 0.05$ and has approximately same perplexity and coherence as GDM, while being 5 times faster. VLAD identified contextually meaningful topics, as can be seen from good coherence score and by eye-balling the topics — they cover a variety of concepts from fishing and cooking to the Enron scandal and cancer. The top 20 words for each of the VLAD topics are provided along with the code.

Stock market analysis We collect variations (closure minus opening price) for 3400 days and 55 companies. We train several algorithms on data from the first 3000 days and report the average distance between the data points from the last 400 days and fitted simplices (i.e., Frobenius norm). This metric alone might be misleading since stretching any simplex will always reduce the score, therefore we also report the volumes of corresponding simplices. Results are summarized in Table 4.3 (right) — our method (estimated $\alpha = 0.05$) achieves comparable fit in terms of the Frobenius norm with a more compact simplex. Among the factors identified by VLAD, we notice a growth component related to banks (e.g., Bank of America, Wells Fargo). Another factor suggests that the performance of fuel companies like Valero Energy and Chevron are inversely related to the performance of defense contractors (Boeing, Raytheon).

4.6 Summary and Discussion

The Dirichlet Simplex Nest model generalizes a number of popular models in machine learning applications, including LDA and several variants of non-negative matrix factorization (NMF). We also develop an algorithm that exploits the geometry of the DSN to perform fast and accurate inference. We demonstrate the superior statistical and computational properties of the algorithm on several real datasets and verify its accuracy through simulations.

One of the key distinctions between the DSN model and NMF models is we replace the separability assumption by a Dirichlet prior on the weights. The main benefit of this approach is it enables us to model data that does not contain archetypal points (*Cutler and Breiman* (1994)). Among the limitations of our approach is the reliance on the Dirichlet distribution assumption in a crucial way, that the Dirichlet distribution is symmetric on the standard probability simplex Δ^{K-1}. In theory, the algorithm breaks down when the Dirichlet distribution is asymmetric. Surprisingly, in simulations at least, we found that VLAD seems quite robust in recovering the correct direction of extreme points, even as most existing methods break down in such situations. These findings are reported in Supplement 4.7.3.

4.7 Appendix

4.7.1 Proofs of Theorems

In this section we present the proofs of main theorems in Section 4.4. We will first reintroduce some notations for the reader's convenience.

Notation Let $\lambda_{\max}(A)$ and $\lambda_{min}(A)$ denote the largest and smallest non-zero singular values of the matrix A. We use $f(\cdot)$ to denote the density of $\mathbb{Q}$ with respect to Lebesgue measure on the $K-1$ dimensional subspace containing the simplex $\mathscr{B}$. Let $g(\cdot)$ be the density of $\mathbb{P}$ with respect to the Lebesgue measure on the $K-1$ dimensional space containing the eigenvectors of Σ_{tot}^K, where Σ_{tot}^K is best $K-1$-rank approximation matrix of $\Sigma_{tot} := BSB^T + \epsilon_0 I_D$ and $\epsilon_0 I_D$ is a uniform upper bound on $\mathrm{Cov}[x_i \mid \theta]$. Let Σ be the population covariance matrix with Σ^K as the best $K-1$ rank approximation. Note that

$$\Sigma = \mathrm{Cov}(X_i) = \mathbb{E}[\mathrm{Cov}(X_i|\mu_i)] + \mathrm{Cov}(\mathbb{E}[X_i|\mu_i]) \leq \epsilon_0 I_D + BSB^T. \tag{4.8}$$

4.7.1.1 Proof of Theorem 1

The following is a standard assumption to ensure the consistency of the k-means procedure embedded in our algorithm:

(a.1) Pollard's regularity criterion (PRC): The Hessian matrix of the function $c \mapsto \mathbb{Q}\phi_{BSB^T}(\cdot, c)$ evaluated at c^* for all optimizer c^* of $\mathbb{Q}\phi_{BSB^T}(\cdot, c)$ is positive definite, with minimum eigenvalue $\lambda_0 > 0$.

Proof. First, we note that under the assumption of the noiseless setting, by following along the lines of the proof of Lemma 4.3.2, it can be seen that if $c^* = (c_1^*, \ldots, c_K^*)$ optimize Eq. (4.1) and v_k's are such that $(v_1, \ldots, v_K)$ form the empirical CVT centroids of Δ^{K-1}, then $c_i^* = BPv_i + c_0$, where c_0 is the population centroid.

Next, the convergence of the empirical CVT centroids to the corresponding population CVT centroids occurs at rate $O_{\mathbb{P}}(\frac{1}{\sqrt{n}})$ rate following *Pollard* (1982b). The consistency of the extreme points of the Dirichlet Simplex Nest follows by the continuous mapping theorem since

$$\frac{\|Pe_k\|_2}{\|Pv_k\|_2} = \frac{\|e_k - \frac{1}{K}\mathbf{1}_K\|_2}{\|v_k - \frac{1}{K}\mathbf{1}_K\|_2} = \frac{\|B(e_k - \frac{1}{K}\mathbf{1}_K)\|_2}{\|B(v_k - \frac{1}{K}\mathbf{1}_K)\|_2}, \tag{4.9}$$

where $e_1, \ldots, e_K$ are the canonical basis vectors on $\mathbb{R}^K$ denoting the vertices of Δ^{K-1}.

Finally, the knowledge of α enables us to compute $\frac{\|e_k - \frac{1}{K}\mathbf{1}_K\|_2}{\|v_k - \frac{1}{K}\mathbf{1}_K\|_2}$. This concludes the proof. □

4.7.1.2 Proof of Theorem 2

It is considerably more challenging to establish the error bounds for our algorithm in the general setting where the observations are noisy. First, let us define the following:

$$\begin{aligned}
\mathscr{C}_{\mathbb{P}_n} =& \{c^* : c^* = \operatorname*{argmin}_{c \in \mathbb{R}^{kD}} \mathbb{P}_n \phi_{(\Sigma_n)^K}(\cdot, c) = \operatorname*{argmin}_{c \in \mathbb{R}^{kD}} \frac{1}{n} \sum_{i=1}^{n} \phi_{(\Sigma_n)^K}(\tilde{X}_i, c)\},\\
\mathscr{C}_{\mathbb{Q}} =& \{c^* : c^* = \operatorname*{argmin}_{c \in \mathbb{R}^{kD}} \mathbb{Q}\phi_{BSB^T}(\cdot, c)\}.
\end{aligned}$$

Recall the following assumptions from the main text:

(a.2) The Hessian matrix of the function $c \mapsto \mathbb{P}\phi_{(\Sigma)^K}(\cdot, c)$ evaluated at c^* for all optimizer c^* of $\mathbb{P}\phi_{(\Sigma)^K}(\cdot, c)$ is uniformly positive definite with minimum eigenvalue bounded below from some $\lambda_0 > 0$, for all $(\Sigma)^K$ such that $(\Sigma - BSB^T) \leq \tilde{\epsilon} I_D$, for some $\tilde{\epsilon} > 0$.

(b) There exists $\epsilon_0 > 0$ such that $\epsilon_0 I_D - \text{Conv}(X|\theta)$ is positive semi-definite uniformly over $\theta \in \Delta^{K-1}$.

(c) There exists M_0 such that for all $M > M_0$,

$$\int_{\mathcal{B}(\sqrt{M}, c_0)^c} \|x - c_0\|_2^2 g(x) \mathrm{d}x \leq \frac{k_1}{M},$$

for some universal constant k_1, where $\mathcal{B}(\sqrt{M}, c_0)$ is a ball of radius $\sqrt{M}$ around the population centroid, c_0.

The assumptions (b) and (c) are very general assumptions and satisfied by a vast array of noise distributions, especially those with subexponential tails. In particular, the noise distributions considered in this work all satisfy these assumptions.

Proof. The proof proceeds by the following steps:

First, in **Step 1**, we show that it is enough to restrict attention to the population estimates instead of empirical estimates. Next, in **Step 2**, we show that the k-means objectives for distributions of μ_i's and x_i's are close. **Step 3** shows that the objective values at the respective minimizers are also close to each other for the distributions considered in **Step 2**. Finaly, **Step 4** uses the strong convexity condition of (a.2) to bound the distance between respective k-means centers, and **Step 5** translates this bound to the estimation of the simplex vertices.

In that regard,

Step 1: Following *Pollard* (1982b), the empirical estimates of CVT centroids optimizing $\mathbb{P}\phi_{\Sigma^K}(\cdot, c)$ converges to the corresponding population estimate at rate $O_{\mathbb{P}_n}(n^{-1/2})$. Thus it is enough to restrict attention to the population estimates.

Step 2: We will show that for all ϵ_0 sufficiently small,

$$|\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{P}\phi_{\Sigma^K}(\cdot, c)| = O(\epsilon_0^{1/3})$$

uniformly over $c \in \mathscr{B}^K$.

Since $\mathbb{Q}$ denotes the distribution corresponding to μ_i's, this distribution places its entire mass inside the simplex, therefore all minimizers of the function $\mathbb{Q}\phi_{BSB^T}(\cdot, c)$ lie inside $\mathscr{B}^K$. We can hence restrict our attention to $c \in \mathscr{B}^K$. By assumption (b), we have $BSB^T \leq \Sigma^K$. Thus, it is enough to establish a bound for $|\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{P}\phi_{\Sigma^K}(\cdot, c)| \;\; \forall \; c \; \in \mathscr{B}^K$.

$$\begin{aligned}|\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{P}\phi_{\Sigma^K}(\cdot, c)| \leq |\mathbb{P}\phi_{\Sigma^K}(\cdot, c) - \mathbb{Q}\phi_{\Sigma^K}(\cdot, c)|& \\ +|\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{Q}\phi_{\Sigma^K}(\cdot, c)|&.\end{aligned} \tag{4.10}$$

Step 2.1: Now, to bound the second term on the right hand side of Eq. (4.10) we use,

$$\begin{aligned}|\mathbb{Q}\phi_{BSB^T}(\cdot, c) - \mathbb{Q}\phi_{\Sigma^K}(\cdot, c)| \quad & \leq \int |\phi_{BSB^T}(x, c) - \phi_{\Sigma^K}(x, c)| f(x) \mathrm{d}x \\ & \leq \lambda_{\max}([BSB^T]^\dagger - [\Sigma^K]^\dagger) \\ & \leq \lambda_{\max}([BSB^T]^\dagger - [(BSB^T + \epsilon_0 I_D^K]^\dagger) \\ & \leq \frac{\epsilon_0}{\lambda_{\min}(BSB^T)\lambda_{\min}(BSB^T + \epsilon_0 I_D^K)},\end{aligned}$$

where $B^\dagger$ denotes the pseudo-inverse of B, and I_D^K is the matrix with top $K-1$ diagonal elements as 1, the rest zeros.

Step 2.2: Turning to the first term on right hand side of Eq. (4.10), we note that $\|\beta_i - \beta_j\|^2 \leq \frac{K-1}{K}\lambda_{\max}(BSB^T)$. Therefore a compact ball of radius $a\lambda_{\max}(BSB^T)$ around the centroid c_0 of the simplex $\mathscr{B}$ for all sufficiently large constants $a > \frac{K-1}{K}$ contains the simplex completely. Consider a ball $\mathcal{B}(\sqrt{M}, c_0)$ of radius $\sqrt{M}$, with $M = a\lambda_{\max}(BSB^T)$ around the centroid c_0, the scalar a to be chosen later. For any $M > 0$,

$$\begin{aligned}|\mathbb{P}\phi_{\Sigma^K}(\cdot, c) - \mathbb{Q}\phi_{\Sigma^K}(\cdot, c)| \leq & \left| \int\limits_{\mathcal{B}(\sqrt{M}, c_0)^c} \phi_{\Sigma^K}(x, c) g(x) \mathrm{d}x \right| \\ & + \left| \int\limits_{\mathcal{B}(\sqrt{M}, c_0)} \phi_{\Sigma^K}(x, c)[g(x) - f(x)] \mathrm{d}x \right|. \end{aligned} \tag{4.11}$$

Step 2.2.1: For the first term on the right hand side of Eq. (4.11), we see that,

$$\begin{aligned} & \int\limits_{\mathcal{B}(\sqrt{M}, c_0)^c} \phi_{\Sigma^K}(x, (c_1, \ldots, c_K)) g(x) \mathrm{d}x \\ & \leq \min_{i \in \{1, \ldots, K\}} \int\limits_{\mathcal{B}(\sqrt{M}, c_0)^c} \|x - c_i\|^2_{\Sigma^K} g(x) \mathrm{d}x \\ & \leq \max 2\|c_i - c_0\|_2^2 \mathbb{P}(X \in \mathcal{B}(\sqrt{M}, c_0)^c) \\ & \qquad + \frac{2}{\lambda_{\min}(BSB^T)} \int\limits_{\mathcal{B}(\sqrt{M}, c_0)^c} \|x - c_0\|_2^2 g(x) \mathrm{d}x. \end{aligned} \tag{4.12}$$

The first inequality follows from Fatou's lemma, while the second follows from the fact that $\|a + b\|_2^2 \leq 2(\|a\|_2^2 + \|b\|_2^2)$.

Suppose that the noise distribution is subexponential for all latent locations $\theta \in \mathscr{B}$. Combining this with the Chebyshev inequality and condition (c), Eq. (4.12) can be re-written as:

$$\begin{aligned}
&\int\limits_{\mathcal{B}(\sqrt{M},c_0)^c} \phi_{\Sigma^K}(x,(c_1,\ldots,c_K))g(x)\mathrm{d}x \\
&\leq \tilde{C}\lambda_{\max}(BSB^T)\frac{Var(X)}{M} + \frac{2k_1}{\lambda_{\min}(BSB^T)M} \\
&\leq \tilde{C}\frac{2(K-1)\lambda_{\max}^2(BSB^T)}{M} + \frac{2k_1}{\lambda_{\min}(BSB^T)M}
\end{aligned} \tag{4.13}$$

for some universal constant k_1.

Step 2.2.2: For the second term on the right hand side on Eq. (4.11), we use the following result.

Claim 1. For $M = a\lambda_{\max}(BSB^T)$, when centroids $c_i \in \mathscr{B}\ \forall\ i$, $\phi_{\Sigma^K}(x, c = (c_1,\ldots,c_K))$ as a function of x is Lipschitz on $\mathcal{B}(\sqrt{M}, c_0)$, with Lipschitz constant $\frac{4\sqrt{M}}{\lambda_{\min}(BSB^T)}$.

Now using the above result, we can easily extend $\phi_{\Sigma^K}(x, c = (c_1,\ldots,c_K))$ to a Lipschitz function on the entire domain. For the particular choice of a,

$$\begin{aligned}
&\left|\int\limits_{B(\sqrt{M},c_0)} \phi_{(\Sigma)^K}(x,c)(g(x)-f(x))\mathrm{d}x\right| \\
&\leq \frac{2\sqrt{a\lambda_{\max}(BSB^T)}}{\lambda_{\min}(BSB^T)} \sup_{\|l\|_{Lip}\leq 1}\left|\int l(x)(g(x)-f(x))\mathrm{d}x\right| \\
&\leq \frac{2\sqrt{a\lambda_{\max}(BSB^T)}}{\lambda_{\min}(BSB^T)} W_1(g,f) \\
&\leq \frac{2\sqrt{a\lambda_{\max}(BSB^T)}}{\lambda_{\min}(BSB^T)} \sqrt{(K-1)\epsilon_0}.
\end{aligned} \tag{4.14}$$

In the above, $\|l\|_{Lip}$ denotes the Lipschitz constant of the function $l(\cdot)$. The second inequality in the above equation follows from Kantorovich-Rubinstein duality while for the last inequality, we use the definition of the Wasserstein distance and take (X,μ) as the coupling with densities $X \sim g$ and $\mu \sim f$ marginally (*Villani* (2008)). Then, for any upper bound M_1 on the variance of $\|X-\mu\|_2$, $W_2(g,f) \leq M_1$, and we use the fact that $\sqrt{(K-1)\epsilon_0}$ forms such an upper bound.

Now, for the noise level $\epsilon_0 > 0$ sufficiently small, there exists $\epsilon > 0$, which is dependent on ϵ_0, such that the open interval $\left(C'\frac{(K-1)\lambda^2_{\max}(BSB^T)}{\epsilon}, \frac{\lambda^2_{\min}(BSB^T)}{\lambda_{\max}(BSB^T)(K-1)\epsilon_0}\epsilon^2/16\right)$ is non-empty for any fixed constant C'. Whenever a is chosen in this range, $|\mathbb{Q}\phi_{BSB^T}(\cdot,c) - \mathbb{P}\phi_{\Sigma^K}(\cdot,c)| \leq \epsilon$. Note that we can choose $\epsilon = O(\epsilon_0^{1/3})$ and $a = O(\epsilon_0^{-1/3})$ to satisfy the above condition.

Step 3: In this step, we show that objective function values for k-means corresponding to that of the population distributions of x_i's and μ_i's are close. Notice that the bounds obtained in Step 2 are uniform over $c \in \mathscr{B}$. For ease of writing, we denote $R_q(c) = \mathbb{Q}\phi_{BSB^T}(\cdot,c)$ and $R_p(c) = \mathbb{P}\phi_{\Sigma^K}(\cdot,c)$. Also, let $\operatorname{argmin} R_p(c) = c_p$ and $\operatorname{argmin} R_q(c) = c_q$. Then, for ϵ_0 sufficiently small, it follows from the discussion above that

$$\begin{aligned} &|R_q(c_p) - R_q(c_q)| \\ &= |R_q(c_p) - R_p(c_p) + R_p(c_q) - R_q(c_q) + R_p(c_p) - R_p(c_q)| \\ &\leq |R_q(c_p) - R_p(c_p) + R_p(c_q) - R_q(c_q)| = O(\epsilon_0^{1/3}). \end{aligned} \tag{4.15}$$

Step 4: In this step, we show that $\|\operatorname{argmin}_c \mathbb{P}\phi_{\Sigma^K}(\cdot,c) - \operatorname{argmin}_c \mathbb{Q}\phi_{BSB^T}(\cdot,c)\|_2 \to 0$ as $\epsilon_0 \to 0$. The intuition behind this is that since the functions $\mathbb{Q}\phi_{BSB^T}(\cdot,c)$ and $R_p(c) = \mathbb{P}\phi_{\Sigma^K}(\cdot,c)$ are point-wise close, and their minimized values are also close to one another, therefore, the points of minima must also be close. By a standard strong convexity argument, employing condition (a.2), for ϵ_0 sufficiently small, we get,

$$\| \operatorname*{argmin}_c \mathbb{P}\phi_{\Sigma^K}(\cdot, c) - \operatorname*{argmin}_c \mathbb{Q}\phi_{BSB^T}(\cdot, c)\|_2 = O\left(\sqrt{\epsilon_0^{1/3}/\lambda_0}\right). \tag{4.16}$$

Step 5 : Finally, the error bound for the simplex vertices follows from a continuous mapping theorem's argument in a similar manner to that of the proof for Theorem 4.4.1. □

Claim 1. For $M = a\lambda_{\max}(BSB^T)$, when centroids $c_i \in \mathscr{B}\ \forall\ i$, $\phi_{\Sigma^K}(x, c = (c_1, \ldots, c_K))$ as a function of x is Lipschitz on $\mathcal{B}(\sqrt{M}, c_0)$, with Lipschitz constant $\frac{4\sqrt{M}}{\lambda_{\min}(BSB^T)}$.

Proof of Claim 1.

$$\begin{aligned}
&\frac{|\phi_{\Sigma^K}(x, c = c_1, \ldots, c_K) - \phi_{\Sigma^K}(y, c = c_1, \ldots, c_K)|}{\|x - y\|} \\
&\leq \max_{i \in \{1, \ldots, K\}} \frac{|\|x - c_i\|_{\Sigma^K} - \|y - c_i\|_{\Sigma^K}|}{\|x - y\|_2} \\
&\leq \sup \frac{2\|x - y\|}{\lambda_{\min}(BSB^T)} \leq \frac{4\sqrt{M}}{\lambda_{\min}(BSB^T)}.
\end{aligned} \tag{4.17}$$

□

4.7.1.3 Consistent estimation of concentration parameter

In this section we first provide several easy calculations required for the estimating equations for some commonly used noise distributions.

Lemma 4.7.1. Depending on the data generating distribution, the covariance matrix of the DSN model is given as follows.

(a) Gaussian data: $\Sigma = BS(\alpha)B^T + \sigma^2 I_d$, provided that $x_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma^2 I_D)$.

(b) Poisson data: $\Sigma = BS(\alpha)B^T + \text{Diag}(\sum_i B_i/K)$, provided that $x_{ij}|\mu_i \overset{ind}{\sim} Poi(\mu_{ij})$, where B_i denotes the i^{th} column of B and $\text{Diag}(a)$ is a diagonal matrix with the i^{th} diagonal element denoting the i^{th} element of the vector a. Here, $\mu_i = (\mu_{i1}, \dots, \mu_{iD})$.

(c) Multinomial data:
$\Sigma = (1 - \frac{1}{N})BS(\alpha)B^T + \frac{1}{N}\text{Diag}(\sum_i B_i/K) - \frac{1}{N}(\sum_i B_i/K)(\sum_i B_i/K)^T$, provided that $x_i|\mu_i \sim \text{Multinomial}(N, \mu_{i1}, \mu_{iD})$. Here, $\mu_i = (\mu_{i1}, \dots, \mu_{iD})$ is a probability vector. (N resembles the number of words per document in the LDA model).

Proof. We compute $Cov(x_i)$ for each of the models. Note that $Cov(X_i) = \mathbb{E}(Cov(x_i|\mu_i)) + Cov(\mathbb{E}(x_i|\mu_i))$ from the tower property of conditional covariance, and $Cov(\mathbb{E}(x_i|\mu_i)) = BS(\alpha)B^T$ for all the models. Therefore we just need the computation for $\mathbb{E}(Cov(x_i|\mu_i))$ for each of the models.

For the Gaussian model, $\mathbb{E}(Cov(x_i|\mu_i)) = \sigma^2 I_D$.

For the Poisson model, $\mathbb{E}(Cov(x_i|\mu_i)) = \mathbb{E}(\mu_i) = B\mathbb{E}(\theta_i) = \text{Diag}(\sum_i B_i/K)$, where the second equality follows as $\mu_i = B\theta_i$ by the model, and the last equality follows because $\theta_i \sim \text{Dir}(\alpha)$.

For the multinomial model,

$$\begin{aligned} \mathbb{E}(Cov(x_i|\mu_i)) &= \frac{1}{N}\mathbb{E}(\text{Diag}(\mu_i)) - \frac{1}{N}Cov(\mu_i\mu_i^T) \\ &= \frac{1}{N}(\text{Diag}(\sum_i B_i/K) - BS(\alpha)B^T), \end{aligned}$$

from which the result follows.

□

Equation 4.6, for estimating α uses the data covariance matrix, $\hat{\Sigma}_n$. While this gives the correct estimating equation in the noiseless scenario, but for the noisy version we need to use $\tilde{\Sigma}_n$ instead where $\tilde{\Sigma}_n$ is a consistent estimator for $BS(\alpha)B^T$. The estimator estimator for different noise distributions can be obtained via the above lemma.

4.7.1.4 Proof of Theorem 3

The proof of consistency of the proposed estimate for the Dirichlet concentration parameter is given as follows.

Proof. Notice that $\|\tilde{\Sigma}_n - BS(\alpha_0)B^T\| = o_P(1)$. Also, $\|\hat{B}(\gamma(\alpha)) - B(\gamma(\alpha))\| = O_P(n^{-1/2})$ for all $\alpha \in \mathscr{C}$. Therefore $\|\hat{B}(\gamma(\alpha))S(\alpha)\hat{B}(\gamma(\alpha))^T - B(\gamma(\alpha))S(\alpha)B(\gamma(\alpha))^T\| = O_P(n^{-1})$ for all $\alpha \in \mathscr{C}$. By monotonicity of the function φ, $BS(\alpha_0)B^T - B(\gamma(\alpha))S(\alpha)B(\gamma(\alpha))^T$ as a function of α is injective for all $\alpha \in \mathscr{C}$. Therefore, $\|\hat{B}(\gamma(\alpha_0))S(\alpha_0)\hat{B}(\gamma(\alpha_0))^T - \tilde{\Sigma}_n\| = o_P(1)$, by triangle inequality. The statement of the theorem then follows by employing a subsequence argument. □

4.7.1.5 Identifiability of the concentration parameter

In the statement of Theorem 3, we require a condition which amounts to a identifiability condition of the parameter α. In this section, we provide empirical evidence that the DSN model with unknown concentration parameter α is identifiable from second moments.

As we shall see, the identifiability of α boils to the invertibility of a scalar function. Recall the covariance matrix of a $\mathsf{Dir}(\alpha)$ distribution is

$$S(\alpha) = \frac{I_K - P_K}{K(K\alpha + 1)},$$

where $P_K = \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T$ is the projector onto $\mathsf{span}\{\mathbf{1}_K\}$. Let $B(\gamma) = \gamma(C - \mu) + \mu$ be the γ-extension of the (scaled) K-means centroids C from the center of the DSN $\mu = \frac{1}{K}B\mathbf{1}_K$.

The question of the identifiability of the concentration parameter boils down to whether there are distinct α_1 and α_2 such that

$$B(\gamma(\alpha_1))S(\alpha_1)B(\gamma(\alpha_1))^T = B(\gamma(\alpha_2))S(\alpha_2)B(\gamma(\alpha_2))^T, \tag{4.18}$$

where $\gamma(\alpha)$ is the extension parameter that corresponds to concentration parameter α. As long as C has full column rank, we may pre and post-multiply (4.18) by $C^\dagger$ and $(C^\dagger)^T$ respectively to see that (4.18) is equivalent to

$$\begin{aligned}&(\gamma(\alpha_1)(I_K - P_K) + P_K)S(\alpha_1)(\gamma(\alpha_1)(I_K - P_K) + P_K)\\ &\quad = (\gamma(\alpha_2)(I_K - P_K) + P_K)S(\alpha_2)(\gamma(\alpha_2)(I_K - P_K) + P_K).\end{aligned}$$

Recalling $S(\alpha)$ is a scalar multiple of $I_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^T$, we see that (4.18) is equivalent to whether there are distinct α_1 and α_2 such that

$$\frac{\gamma(\alpha_1)^2}{K(K\alpha_1+1)} = \frac{\gamma(\alpha_2)^2}{K(K\alpha_2+1)}.$$

This is equivalent to the invertibility of the function

$$\varphi(\alpha) = \frac{\gamma(\alpha)^2}{K(K\alpha+1)}. \tag{4.19}$$

Figure 4.18 shows this function for $K = 10$ over a range of reasonable values of α. We see that the function is in fact invertible.

Although Figure 4.18 suggests (4.19) is invertible, we do not have a rigorous proof. The main challenge is obtaining precise control on the growth of (4.18). Inspecting Figure 4.18 shows that $\varphi(\alpha)$ is almost flat as soon as α exceeds $\frac{5}{2}$. Intuitively, this

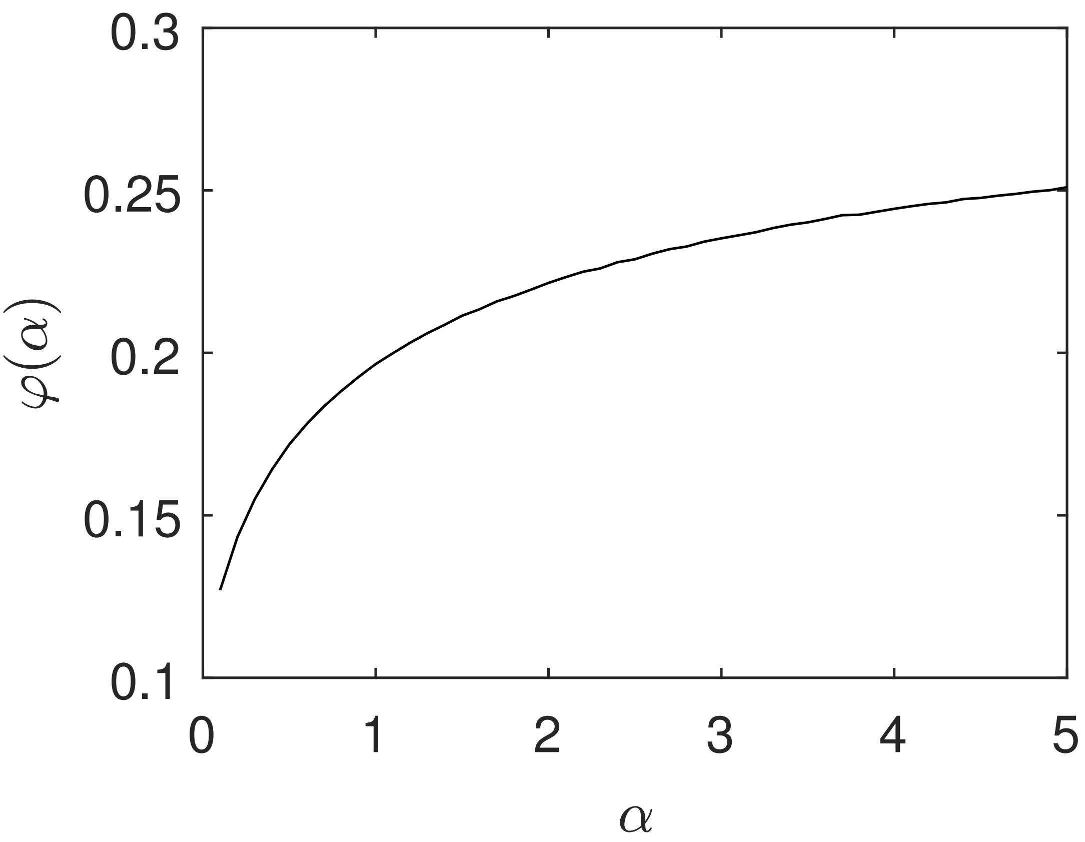


Figure 4.18: *Empirical study of α identifiability.*

is a consequence of the hardness of distinguishing between DSNs with large α's (and correspondingly large extension parameters). Mathematically, it is hard to obtain precise control on the growth of $\varphi(\alpha)$ because it is not possible to evaluate $\gamma(\alpha)$ explicitly. Although it is possible to show that

$$\gamma(\alpha) = \frac{1 - \frac{1}{K}}{\int_{V_k} e_k^T \theta p_\alpha(x) dx - \frac{1}{K}}, \tag{4.20}$$

where $V_k = \{\theta \in \Delta^{K-1} : \text{argmax}\{\theta_l : l \in [K]\} = k\}$ is the k-th Voronoi cell in a centroidal Voronoi tessellation of Δ^{K-1}, e_k is the k^{th} canonical basis vector and p_α is

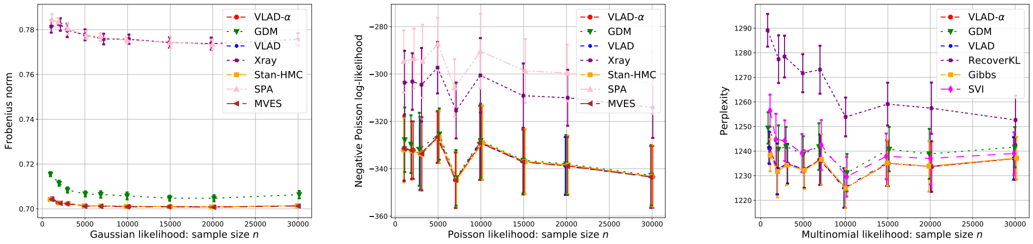


Figure 4.19: *Frobenius norm for Gaussian data*

Figure 4.20: *Negative log-likelihood for Poison data*

Figure 4.21: *Perplexity for LDA data*

Figure 4.22: *Held out data performance for increasing sample size n*

the $\mathsf{Dir}(\alpha)$ density, it is hard to evaluate the integral. We defer an investigation of the identifiability of the concentration parameter to future work.

4.7.2 Experimental Details

4.7.2.1 Computational cost of VLAD

In this section, we tally up the computational cost of VLAD. The dominant cost it that of computing the top K singular factors of the centered data matrix $\bar{X}$. This costs $O(DKn)$ floating point operations (FLOP's). The cost of the subsequent clustering step is asymptotically negligible compared to the cost of the SVD. Assuming each step of the K-means algorithm costs $O(Kn)$ FLOP's and the algorithm converges linearly, we see that the cost of obtaining an $O(\frac{1}{n})$-suboptimal solution is $O(Kn\log n)$. We discount the cost of Monte Carlo estimates of the extension parameter because it can be tabulated. Thus the computational cost of the algorithm is dominated by the cost of computing the SVD.

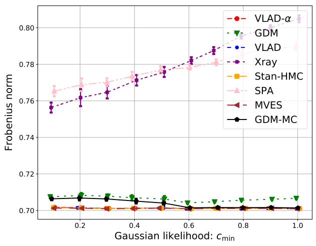


Figure 4.23: *Frobenius norm for Gaussian data*

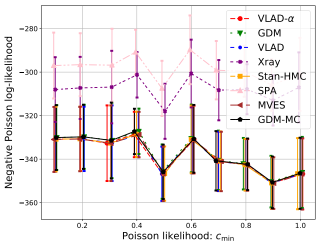


Figure 4.24: *Negative log-likelihood for Poison data*

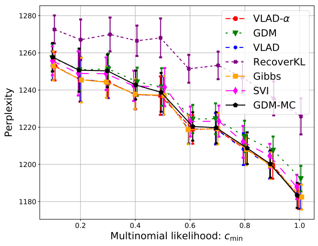


Figure 4.25: *Perplexity for LDA data*

Figure 4.26: *Held out data performance for varying DSN geometry*

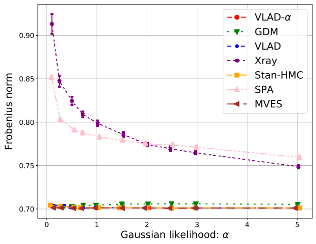


Figure 4.27: *Frobenius norm for Gaussian data*

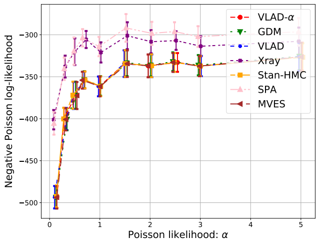


Figure 4.28: *Negative log-likelihood for Poison data*

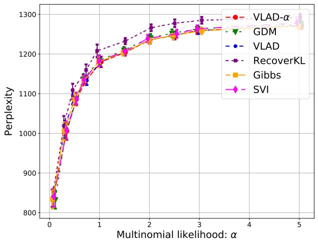


Figure 4.29: *Perplexity for LDA data*

Figure 4.30: *Held out data performance for increasing α*

4.7.2.2 Additional results

Additional results for convergence behavior We complement the results presented in Fig. 4.9 with the corresponding plots of the likelihood evaluated on a set of held out data. These results are summarized in Fig. 4.9. For all plots, the smaller value is better. We see that VLAD shows performance as good as HMC and Gibbs sampler at a much lower computational time.

Additional results for geometry of the DSN Again, we further support our results of Fig. 4.13 with the corresponding held out data likelihood scores. Fig. 4.26 summarizes the results - VLAD shows competitive performance.

Additional results for varying Dirichlet prior In Figure 4.30 we demonstrate held out data likelihood corresponding to experiments of Fig. 4.17. We see that VLAD performs well in the whole range of analyzed values and likelihood kernels.

Data generation for simulations studies For all experiments, unless otherwise specified, we set $D = 500, K = 10, \alpha = 2, n = 10000$ (for LDA vocabulary size $D = 2000$). To generate DSN extreme points, for Gaussian data we sample $\beta_1, \ldots, \beta_K \sim \mathcal{N}(0, K)$; for Poisson data $\beta_1, \ldots, \beta_K \sim \text{Gamma}(\mathbf{1}, K\mathbf{1})$; for the LDA $\beta_1, \ldots, \beta_K \sim \text{Dir}_D(\eta)$ with $\eta = 0.1$. To ensure skewed geometry we further rescale extreme points towards their mean by uniform random factors between 0.5 and 1. To do so first compute the mean of extreme points $C = \frac{1}{K}\sum_k \beta_k$ and then rescale each one with $\beta_k = C + c_k(\beta_k - C)$, where $c_k \sim \text{Unif}(c_{\min}, 1)$. Except for the DSN geometry experiment, we set $c_{\min} = 0.5$.

Then we sample weights $\theta_i \sim \text{Dir}_K(\alpha)$ and data mean $\mu_i = \sum_k \theta_{ik}\beta_k$. For Gaussian data $x_i|\mu_i \sim \mathcal{N}(\mu_i, \sigma^2 I_D)$, $\sigma = 1$; for Poisson data $x_i|\mu_i \sim \text{Pois}(\mu_i)$; for LDA we follow standard generating process of *Blei et al.* (2003) with 3000 words per document. All experiments were run for 20 repetitions and mean was used in the plots along with half standard deviation error bars.

Baseline methods and algorithms setups We considered four separability based NMF algorithms: Xray (*Kumar et al.* (2013)) with code from `https://github.com/arbenson/mrnmf`; MVES (*Chan et al.* (2009)) with code from `http://www.ee.nthu.edu.tw/cychi/source_code_download-e.php`; Sequential Projection Algorithm (*Gillis and Vavasis* (2012)) that we implemented in Python; RecoverKL (*Arora et al.* (2013)) for the LDA case with code from `https://github.com/MyHumbleSelf/anchor-baggage`.

Bayesian NMF approaches often assume positive weights without the simplex constraint imposed by the Dirichlet prior on weights. Incorporating the simplex constraint complicates the inference (*Paisley et al.* (2014)) as Dirichlet distribution is not conjugate to popular choices of data likelihood such as Gaussian or Poisson. Therefore we are not aware of any implementation for DSN type of models outside of the LDA scenario. We instead chose to compare to automated Bayesian inference methods. We implemented DSN inference with Poison and Gaussian likelihoods in Stan (*Carpenter et al.* (2017)) and considered all three supported estimation procedures: HMC with No U-Turn Sampler (*Hoffman and Gelman* (2014)), MAP optimization and (*Kucukelbir et al.* (2017)) Automatic Differentiation Variational Inference. MAP optimization and ADVI performed poorly and we did not report their performance. HMC was always trained with true value of α and with knowledge of $\sigma = 1$ for the Gaussian scenario. Number of iterations was set to 80 for $n < 3000$, 60 for $n = 3000$ and 40 for $n > 3000$. We had to restrict number of iterations due to prohibitively long running time (40 iterations for $n = 30000$ took 3.5 hours for Gaussian likelihood and 14 hours for Poisson likelihood; VLAD took 7 seconds in both cases). For the LDA, we used Gibbs sampler (*Griffiths and Steyvers* (2004)) from `https://github.com/lda-project/lda` trained for 1000 iterations (1000 iterations for $n = 30000$ took 3.6 hours; VLAD took 3min). Gibbs sampler was trained with true values of α and η. We used Stochastic Variational Inference (*Hoffman et al.* (2013)) implementation from scikit-learn (*Pedregosa et al.* (2011)) and trained it with true values of α and η.

For the Geometric Dirichlet Means (*Yurochkin and Nguyen* (2016)) we used implementation from `https://github.com/moonfolk/Geometric-Topic-Modeling` with 8 K-means restarts and ++ initialization.

VLAD was implemented in Python using numpy SVD package and scikit-learn (*Pedregosa et al.* (2011)) K-means clustering with 8 restarts and ++ initialization. The code is available at `https://github.com/moonfolk/VLAD`.

For the NYT data `https://archive.ics.uci.edu/ml/datasets/bag+of+words` we trained Gibbs sampler with $\alpha = 0.1$ and $\eta = 0.1$ for 1000 iterations and SVI with default settings. For the stock data we trained HMC for 100 iterations with $\alpha = 0.05$.

4.7.3 On asymmetric Dirichlet prior

In our work we assumed that $\theta_i \sim \mathrm{Dir}_K(\alpha)$, where $\alpha \in \mathbb{R}_+$. When α is a scalar, the corresponding Dirichlet distribution is referred to as symmetric. More generally, $\alpha \in \mathbb{R}_+^K$ is a vector of parameters. Our algorithmic guarantees, such as alignment of CVT centroids of $\mathscr{B}$, extreme points and centroid of $\mathscr{B}$ and equivalence of extension parameters for all extreme points directions, fail for the general asymmetric case. *Wallach et al.* (2009) showed that more careful treatment of the parameter α can improve the quality of the LDA topics. Geometric treatment of the asymmetric Dirichlet distribution remains to be the question of future studies. To facilitate the discussion, here we visualize the problem using toy $D = 3, K = 3$ example (similar to Fig. 4.5) with $\alpha = (0.5, 1.5, 2.5)$. Results of the four different algorithms are shown in Fig. 4.35. Note that for VLAD (Fig. 4.34) we only show the directions of the line segments of the obtained sample CVT centroids and the data center, since we do not have a procedure for extension parameter estimation in the asymmetric Dirichlet case. We see that all of the algorithms fail with various degrees of error and notice that the directions obtained by VLAD no longer appear consistent, however do not deviate drastically from the truth. We propose to call such toy triangle experiment a *triangle test* and hope to "pass" the asymmetric Dirichlet *triangle test* in the future work.

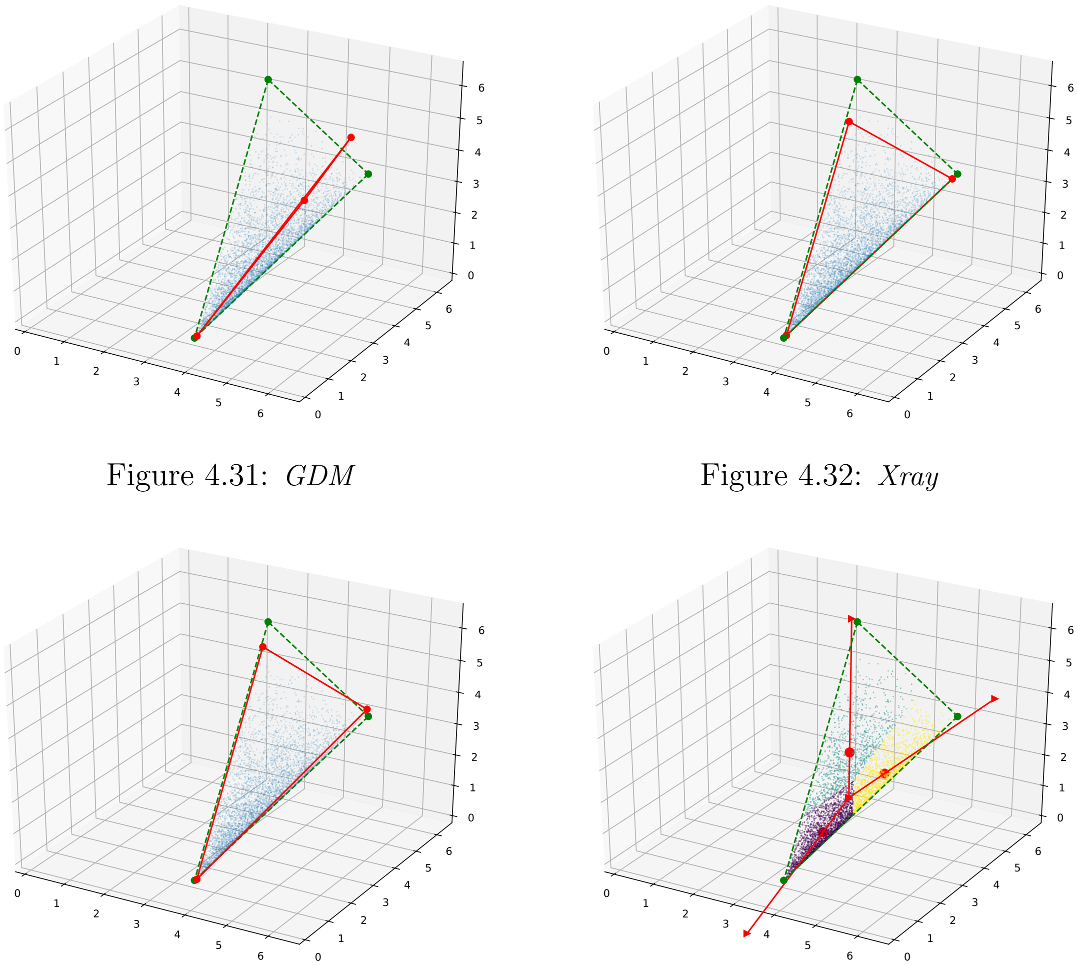


Figure 4.31: *GDM*

Figure 4.32: *Xray*

Figure 4.33: *HMC*

Figure 4.34: *VLAD*

Figure 4.35: *Asymmetric Dirichlet toy simplex learning:* $n = 5000, D = 3, K = 3, \alpha = (0.5, 1.5, 2.5)$

CHAPTER V

Robust Representation Learning of Temporal Dynamic Interactions

Robust representation learning of temporal dynamic interactions is an important problem in robotic learning in general and automated unsupervised learning in particular. Temporal dynamic interactions can be described by (multiple) geometric trajectories in a suitable space over which unsupervised learning techniques may be applied to extract useful features from raw and high-dimensional data measurements. Taking a geometric approach to robust representation learning for temporal dynamic interactions, it is necessary to develop suitable metrics and a systematic methodology for comparison and for assessing the stability of an unsupervised learning method with respect to its tuning parameters. Such metrics must account for the (geometric) constraints in the physical world as well as the uncertainty associated with the learned patterns. [1]In this chapter we introduce a model-free metric based on the Procrustes distance for robust representation learning of interactions, and an optimal transport based distance metric for comparing between distributions of interaction primitives. These distance metrics can serve as an objective for assessing the stability of an interaction learning algorithm. They are

[1] This work has been published in *Guha et al.* (2020)

also used for comparing the outcomes produced by different algorithms. Moreover, they may also be adopted as an objective function to obtain clusters and representative interaction primitives. These concepts and techniques will be introduced, along with mathematical properties, while their usefulness will be demonstrated in unsupervised learning of vehicle-to-vechicle interactions extracted from the Safety Pilot database, the world's largest database for connected vehicles.

5.1 Introduction

Advances in large scale data processing and computation enables the application of sophisticated learning algorithms to robotic design in complex and dynamic environments. In many applications a fundamental challenge lies not only in learning about the interaction between the ego agent and the environment, but also interactions between multiple agents. Due to the high dimensionality and typically noisy nature of the data required for such learning tasks, a standard approach is to utilize strong modeling assumptions on the interactions. For example, the interaction between a robotic agent and the environment can be represented by instantaneous physical variables such as positions, velocities, a time series of which are then endowed with a stationary distribution for mathematical convenience and interpretability (e.g., via a Markov process framework). While such approach is useful in highly controlled environments, the strong modeling assumption are usually violated in domains where the interactions among agents and with the environment are highly dynamic *Foerster et al.* (2017). Such domains require the development of more robust and data-driven representation learning approaches.

As a concrete example which serves as a primary motivation for this work, take the interaction between two intelligent vehicles that approach each other in a typical intersection. What the two vehicles proceed to do next depend on what they can learn of their encounter in real-time. The two cars may come toward the intersection in varying speeds at perhaps slightly different time points. They may or may not signal their intention. For example, one plans to go straight while the other plans to take a turn cutting through the other's path. Not only do the two agents have to learn their temporally varying interaction, they have to do so quickly and accurately while continually negotiating the traffic. In this type of applications where the interaction is highly dynamic, a promising approach to robust interaction learning is by decomposing the interaction in terms of simpler elements *Frazzoli et al.* (2005); *Wang and Zhao* (2017). For traffic applications, such interaction elements are called traffic *primitives.* These primitives can be learned, labeled, and effectively utilized for subsequent tasks such as vehicle trajectory prediction *Zhu et al.* (2019), traffic data generation *Ding et al.* (2018); *Zhang et al.* (2019), or anomaly detection *Zhang and Wang* (2019).

Stripping away the language of vehicle-to-vehicle (V2V) interactions, the temporal dynamic interaction between two agents comprises of a pair of well-aligned trajectories defined on a suitable space that satisfy constraints presented by the environment and agents' behaviors. Thus, the goal of robust representation learning of a pairwise interaction between the two dynamic agents boils down to the learning of pairs of functions or curves which describe the aligned car physical movements and/or driving behaviors. Such a mathematical viewpoint can be generalized to interactions among three or more vehicles. In this chapter we will focus on the learning of interactions in two-agent dynamic scenarios. Although our work is motivated by the learning of multi-agent traffic interaction's primitives, we believe that the techniques developed here can be utilized to other settings of multi-agent temporal dynamic interaction learning.

Within the context of real-world traffic learning, both rule-based methods *Frazzoli* *et al.* (2005), supervised learning *Pervez et al.* (2017), and unsupervised learning *Wang and Zhao* (2017) have been applied to identify the interaction primitives. Due to the heterogeneity and complexity of traffic scenarios, unsupervised learning is a powerful tool to identify latent structures in unlabeled traffic scenario time series data; the goal is to organize the data into homogeneous groups/ clusters *Bender et al.* (2015); *Hamada et al.* (2016); *Taniguchi et al.* (2015); *Wang et al.* (2017); *Liao* (2005). Within automatically learned clusters, interpretable and typical driving behaviors can be obtained and analyzed, e.g., left/right turns along with multiple attributes including speed, acceleration, yaw rate and side-slip angle using Dynamic Time Wrapping (DTW) as a similarity measure *Yao et al.* (2019). Statistical model-based approaches that can learn complex driving behaviors while allowing for encoding domain-knowledge are also available. For instance, primitive segments extracted from time series traffic data can be obtained without specifying the number of categories via Bayesian nonparametric methods based on Dirichlet processes. They include hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) *Taniguchi et al.* (2016); *Wang et al.* (2017). Dirichlet process mixtures of Gaussian processes were also successfully employed to identify complex multi-vehicle velocity fields *Guo et al.* (2019); *Joseph et al.* (2011).

Given the plethora of methods and the need for learning complex interaction patterns in dynamic domains, it is natural to ask which method one should use. For unsupervised learning, this question is particularly challenging because one typically works with unlabelled data and without immediately available objective functions for the quality of learned clusters of interactions, especially ones which are mathematically represented as a collection of two or more curves taking values in a suitable space, as discussed above. In addition, while the problem of devising techniques which are free of any tuning parameters is an important one, parameter-free algorithms tend to be not robust. A typical unsupervised learning method still requires some prior knowledge or pre-defined parameters (tuning knobs). As a result, clustering results may still be sensitive to these choices. Thus, even when a method is settled on, it is still an important issue how to handle the various tuning knobs and to assess their sensitivity, or stability with respect to changes in the tuning parameters.

Identifying suitable clustering criteria and analyzing learning stability/sensitivity have received much attention in data mining and statistical learning literatures. For clustering criteria, there are broadly two categories: internal and external criteria *Xu and Wunsch* (2009). Internal criteria relies on a similarity or dissimilarity measure that may be applied to the data samples. Such measures evaluate how alike the members of the same cluster are, how different the members of different clusters are, or some combination of thereof *Rokach and Maimon* (2005); *Xu and Tian* (2015). On the other hand, there is a priori structure how the data should be partitioned, external criteria allow one to compare the clustering results against this structure *Rokach and Maimon* (2005). Examples include Rand index, mutual information and model-based likelihood-type objectives *Xu and Tian* (2015).

Meanwhile, there is a rich literature on sensitivity analysis that focuses on the impacts of changes in model/method specification on the learning outcomes, see, e.g. textbooks *Chatterjee and Hadi* (1988); *Saltelli et al.* (2004, 2008). If we focus on Bayesian methods or model-based methods, the key issue is on the effect of the prior/ model specification. While there are a number of variations, most sensitivity analysis techniques involve model fitting with varying prior/ model specifications, and assessing the impacts on posterior distributions or estimates of parameters of interest. A model is said to be robust if the estimates are relatively insensitive to such varying specifications *Gustafson* (2000); *Sivaganesan* (2000). Alternatively, instead of varying the model parameters one may consider perturbing data: a geometric framework was developed to conduct sensitivity analysis with respect to the perturbation of the data, the prior and the sampling distribution for a class of statistical models. Within this framework, various geometric quantities were studied to characterize the intrinsic structure and effect of the perturbation *Zhu et al.* (2011).

To assess the quality of unsupervised learning methods for temporal dynamic interactions, at a high level one may consider the aforementioned methods and frameworks. Moreover, it is necessary to develop a set of suitable metrics for interaction comparison and for assessing the stability of an unsupervised learning method with respect to its tuning parameters. Motivated by the representation of dynamic vehicle-to-vehicle interactions that arise in the traffic learning domain, one has to effectively deal with pairs of aligned functions, i.e., trajectories taking values in a suitable space, which is typically non-Euclidean and has high or infinite dimensions. Such metrics must account for the geometric constraints in the physical world as well as the uncertainty associated with the learned patterns.

To this end, we introduce a model-free metric on pairs of functions based on a Procrustes-type distance, and an optimal transport based Wasserstein distance metric for comparing between distributions of such pairs of functions. The former metric is critical because it preserves translation and rotation invariance, key properties required for capturing the essence of the temporal dynamic between two autonomous or semi-autonomous agents (e.g., vehicles or robots). The latter metric is also appropriate because the result of a clustering algorithm can be mathematically represented as the solution of an optimal transport problem *Graf and Luschgy* (2000); *Ho et al.* (2017). In addition to some connection to optimal transport based clustering, it is worth noting how our technical contributions are also inspired by several other prior lines of work.

In particular, Procrustes-type metrics have been employed in generalized Procrustes analysis which solves the problem of reorienting points to a fixed configuration *Gower* (1975). Similar metrics have also been successfully used in literature to study such problems of shape preservation *Srivastava and Klassen* (2016) as well for alignment of manifolds *Wang and Mahadevan* (2008); *Lipman et al.* (2013). In our work, we use it to solve the clustering problem by comparing pairs of curves, each of which may be viewed as manifolds on $\mathbb{R}^2$.

Finally, we note that the introduced distance metrics can serve as an objective for assessing the stability of an interaction primitive learning algorithm. They are also used for comparing the outcomes produced by different algorithms. Furthermore, they may also be adopted as an objective function to obtain clusters of interactions, and the representative interactions. These concepts and techniques will be introduced in this chapter, along with mathematical properties, while their usefulness will be demonstrated in the analysis of vehicle-to-vehicle interactions that arise in the Safety Pilot database **?**, the world's largest database for connected vehicles.

The chapter is organized as follows. In Section 5.2, we describe a distance metric for pairs of trajectories and explicate its useful mathematical properties. Building on this, Section 5.3 studies distributions of trajectory pairs, which lead to methods for obtaining and assessing clusters of interactions. Finally, Section 5.4 illustrates our methods on the clustering analysis of vehicle-to-vehicle interactions data.

5.2 A distance metric on temporal interactions

Because a temporal interaction between two agents is composed of trajectories, we need to first formally define a trajectory. Let $f : \mathbb{R} \to \mathbb{R}^2$ denote a trajectory of an object (e.g., vehicles, robots). In particular, $f(t)$ represents the location of the object at time-point t. It suffices for our purpose to restrict to $t \geq 0$.

We can consider all possible trajectories in a similar manner. Define the set of all possible trajectories as $\mathbb{F} = \{f : [0, \infty) \to \mathbb{R}^2 : f \text{ is continuous}\}$. The set of all possible trajectories up to time-point t starting from time-point s is denoted by $\mathcal{F}_{[s,t)} = \{f : [s, t) \to \mathbb{R}^2 | f \in \mathbb{F}\}$. Similarly for $(t_1, \ldots, t_k) \in \mathbb{R}_+^k$ we will use $\mathcal{F}_{t_1,\ldots,t_k} := \{(f(t_1), \ldots, f(t_k)) : f \in \mathbb{F}\}$. Also, we define $\mathcal{F} := \cup_{s,t \in \mathbb{R}^+} \mathcal{F}_{[s,t)}$.

Next, operations can be defined on these trajectories. For any $c \in \mathbb{R}^2$, and $f \in \mathbb{F}$ we define $f + c \in \mathbb{F}$ as $(f + c)(x) = f(x) + c$ for all $x \in [0, \infty)$. Similarly, for any orthogonal matrix $O \in \mathbb{R}^{2\times2}$, define the function $O \odot f \in \mathbb{F}$ as $(O \odot f)(x) = O \cdot f(x)$ for all $x \in [0, \infty)$, where $O \cdot f(x)$ is the usual matrix product between matrix O and vector $f(x)$ which have matching dimensions.

With these definitions in place, we now define an interaction and operations on these interactions. An interaction is an ordered pair (f_1, f_2) such that $f_1, f_2 \in \mathbb{F}$. We also define operations on interactions as well. Let $SO(n)$ be the group of $n \times n$ orthogonal matrices with determinant $+1$. For a pair $f_1, f_2 \in \mathbb{F}$, we define $\mathcal{O}_{(f_1,f_2)} = \{(O \odot f_1, O \odot f_2) : O \in SO(2)\}$. Similarly, define $\mathcal{C}_{(f_1,f_2)} = \{(f_1 + c, f_2 + c) : c \in \mathbb{R}^2\}$.

5.2.1 Rotation and translation-invariant metrics on curves

To evaluate the stability and overall quality of clustering, we want a distance metric, $d : \mathbb{F}^2 \times \mathbb{F}^2 \to \mathbb{R}_+$, where ($\mathbb{F}^2 = \mathbb{F} \times \mathbb{F}$), that has the following properties:

(a) Distance between two interactions is invariant with respect to the re-ordering of corresponding trajectories, i.e., for $f_{11}, f_{12}, f_{21}, f_{22} \in \mathbb{F}$, the following holds:

$$d((f_{11}, f_{12}), (f_{21}, f_{22})) = d((f_{12}, f_{11}), (f_{21}, f_{22})).$$

(b) Distance between a pair of interactions is invariant of starting points of the trajectories composing the interactions, given the knowledge of the relative distance of the starting points of trajectories comprising each interaction. Specifically, if $(f_1', f_2') \in \mathcal{C}_{(f_1,f_2)} \cup \mathcal{O}_{(f_1,f_2)}$, then,

$$d((f_1', f_2'), (f_1, f_2)) = 0.$$

Condition (a) enables the removal of order in a pair of curves in an interaction, while condition (b) in essence characterizes **rotational** and **translational invariance** of interactions. We will henceforth use $(\mathcal{O}, \mathcal{C})_{(f_1,f_2)}$ to denote the set $\{(O \odot f_1 + c, O \odot f_2 + c) : O \in SO(2), c \in \mathbb{R}^2\}$. As shown in Lemma 5.6.1, condition (b) implies that $d((f_{11}, f_{12}), (f_{21}, f_{22})) = d((O \odot f_{11} + c, O \odot f_{12} + c), (f_{21}, f_{22}))$ for all $c \in \mathbb{R}^2$, $O \in SO(2)$. This appears to be a reasonable requirement since the exact location and orientation of interactions should not affect the classification of different interactions into clusters characterized by "primitives". Note that throughout this chapter we only consider non-reflective rotational transforms, i.e., transforms involving orthogonal matrices, O, such that $\det(O) = +1$.

Let ρ be a distance metric for $\mathbb{F}^2$. We will then construct a metric d satisfying (a) and (b) from ρ. Definition 5.2.1 shows how we can define d in terms of ρ.

Definition 5.2.1. Define Procrustes distance

$$\begin{aligned} & d((f_{11}, f_{12}), (f_{21}, f_{22})) \qquad (5.1) \\ := & \inf_{(f_1', f_2') \in (\mathcal{O}, \mathcal{C})_{(f_{21}, f_{22})}} \left\{ \min\Big\{ \rho((f_{11}, f_{12}), (f_1', f_2')), \rho((f_{12}, f_{11}), (f_1', f_2')) \Big\} \right\}. \end{aligned}$$

From the definition of metric d above, it is clear that $(f_{21}, f_{22}) \in (\mathcal{O}, \mathcal{C})_{(f_{11},f_{12})} \cup$

$(\mathcal{O}, \mathcal{C})_{(f_{12}, f_{11})} \iff d((f_{11}, f_{12}), (f_{21}, f_{22})) = 0$. With that knowledge, we can define an equivalence relation, $\sim$, as

$$(f_{11}, f_{12}) \sim (f_{21}, f_{22}) \iff d((f_{11}, f_{12}), (f_{21}, f_{22})) = 0. \tag{5.2}$$

Although d is not a proper metric on $\mathbb{F}^2$, as Proposition 5.2.1 shows, d does define a metric on the quotient space relative to the equivalence relation.

Proposition 5.2.1. Let ρ be a distance metric on $\mathbb{F}^2$ such that for all $f_{11}, f_{12}, f_{21}, f_{22} \in \mathbb{F}$,

(i) ρ satisfies, for some function h,

$$\rho((f_{11}, f_{12}), (f_{21}, f_{22})) = h((f_{11}, f_{12}) - (f_{21}, f_{22})).$$

(ii) ρ is an inner-product norm.

Then d given by Eq. (5.1) is a distance metric on the quotient space $\mathbb{F}^2/\sim$.

Proposition 5.2.1 also provides a method to build a metric satisfying conditions (a) and (b) above. One way to do so is from a probability measure perspective. In fact, let μ be a probability measure on $[0, \infty)$. We consider the set of trajectories with integrable Euclidean norm on $[0, \infty)$, i.e., we restrict attention to the following set of trajectories:

$$\mathbb{F}_2(\mu) = \left\{ f : [0, \infty) \to \mathbb{R}^2 \middle| f \text{ is continuous}, \int_0^\infty \|f(x)\|_2^2 \mu(\mathrm{d}x) < \infty, \right\}$$

where $\|\cdot\|_2$ is the Euclidean norm in $\mathbb{R}^2$. For our purposes, we use ρ as the usual Euclidean metric on $\mathbb{F}_2^2(\mu) := \mathbb{F}_2(\mu) \times \mathbb{F}_2(\mu)$. Namely, for $(f_{11}, f_{12}), (f_{21}, f_{22}) \in \mathbb{F}_2^2(\mu)$, we use

$$\rho((f_{11}, f_{12}), (f_{21}, f_{22}))^2 := \|f_{11} - f_{21}\|_2^2 + \|f_{12} - f_{22}\|_2^2, \tag{5.3}$$

where $\|f_{1i} - f_{2i}\|_2^2 = \int_0^\infty \|f_{1i}(x) - f_{2i}(x)\|_2^2 \mu(\mathrm{d}x), i = 1, 2$. Here and henceforth we assume that the trajectories $f_{11}, f_{12}, f_{21}, f_{22}$ all span across the same length of time. Note that this choice of metric satisfies the criteria in Proposition 5.2.1. Also, equivalently, to define similar rotation and translation invariant metrics on $\mathcal{F}_{[s,t)}$, for any $s < t \neq \infty$, we can simply choose any probability measure μ with support on $[s, t)$.

Proposition 5.2.2 below provides a simple method to explicitly compute the metric d between interactions, when ρ is given by (5.3). We will need the following notation:

(A1) For $(f_{11}, f_{12}), (f_{21}, f_{22}) \in \mathbb{F}_2^2(\mu)$, let UDV^T be the singular value decomposition for the matrix given by

$$\sum_{i=1}^{2} \int_0^\infty \left(f_{2i}(x) - \bar{f_{2\cdot}}(x)\right) \left(f_{1i}(x) - \bar{f_{1\cdot}}(x)\right)^T \mu(\mathrm{d}x),$$

where $\bar{f_{2\cdot}}(x) = \int_0^\infty (f_{21}(x) + f_{22}(x))/2 \ \mu(\mathrm{d}x)$ and
$\bar{f_{1\cdot}}(x) = \int_0^\infty (f_{11}(x) + f_{12}(x))/2 \ \mu(\mathrm{d}x)$.

Each of the summands in (A1) form a 2×2 dimensional matrix. Here, $\bar{f_1}(x)$ denotes the elementwise integration of the 2×1 vector $(f_{11}(x) + f_{12}(x))/2$. Moreover, the outer-integral in each of the summands in (A1) is an elementwise integral of the 2×2 matrix integrand formed by matrix multiplication of the 2×1 vector $\left(f_{21}(x) - \bar{f_{2\cdot}}(x)\right)$ and the 1×2 vector $\left(f_{11}(x) - \bar{f_{1\cdot}}(x)\right)^T$.

Proposition 5.2.2. Assume $f_{11}, f_{12}, f_{21}, f_{22} \in \mathbb{F}_2(\mu)$. Let UDV^T be the singular value decomposition as in (A1). Then,

$$\inf_{(f'_1,f'_2)\in(\mathcal{O},\mathcal{C})_{(f_{21},f_{22})}} (\rho((f_{11},f_{12}),(f'_1,f'_2)))^2 = -2\,\mathrm{trace}\left(D \begin{bmatrix} 1 & 0 \\ 0 & \det(V^T U) \end{bmatrix} \right) + \sum_{i=1}^{2} \int_0^\infty \left\| f_{2i}(x) - \bar{f}_{2\cdot}(x) \right\|_2^2 + \left\| f_{1i}(x) - \bar{f}_{1\cdot}(x) \right\|_2^2 \mu(\mathrm{d}x).$$

The optimal $\mathcal{O}, \mathcal{C}$ that define the infimum are given by :

$$\begin{aligned} \tilde{\mathcal{O}} &= V^T \begin{bmatrix} 1 & 0 \\ 0 & \det(V^T U) \end{bmatrix} U, \\ \tilde{\mathcal{C}} &= \bar{f}_{1\cdot}(x) - \tilde{\mathcal{O}} \cdot \left(\bar{f}_{2\cdot}(x) \right). \end{aligned} \tag{5.4}$$

The proof of the proposition is discussed in Section 5.6.1.2. The problem discussed in Propostion 5.2.2 is a version of the well-known **least root mean square deviation** problem. It was first solved by the Kabsch algorithm *Kabsch* (1976, 1978). A more computationally efficient method to compute the optimal $\mathcal{O}, \mathcal{C}$ was later obtained using the theory of quarternions *Horn* (1986); *Coutsias et al.* (2004).

5.3 Quantifying distributions of primitives

The metric defined above can be used to obtain clusters of interactions, in addition to evaluating the overall quality and stability of a particular clustering method. Our starting point is to note that the problem of clustering or summarizing interactions can be formalized as finding a discrete distribution on the space of interactions. More specifically, one needs to obtain a discrete probability distribution on interactions, where each supporting atom represents a typical interaction (namely, an interaction *primitive*) and the mass associated with each atom represents the proportion of a cluster. From this perspective, an objective that naturally arises is to minimize a distance from the empirical distribution of interactions to a discrete probability measures with a fixed number, say k, of supporting atoms, which represent the primitives. An useful tool for defining distance metrics on the space of distributions arises from the theory of optimal transport *Villani* (2003).

Optimal transport distances enable comparisons of distributions in arbitrary structured and metric spaces by accounting for the underlying metric structure. They have been increasingly adopted to address clustering in a number of contexts *Pollard* (1982a); *Graf and Luschgy* (2000); *Ho et al.* (2017). For instance, it is well-known that the problem of determining an optimal finite discrete probability measure minimizing the second-order Wasserstein distance W_2 to the empirical distribution of the data is directly connected to the k-means clustering problem (discussed in Section III in details). Inspired by this connection, we will seek to summarize the distribution of interactions appropriately. To this end, we will define Wasserstein distances for distributions of interactions as follows, by accounting for the metric structure developed in the previous section.

Let d be a distance metric on $\mathbb{F}_2^2(\mu)/\sim$, where $\sim$ is the equivalence relation defined in Eq. (5.2).

Fix $[(f_{11}, f_{12})] \in \mathbb{F}_2^2(\mu)/\sim$. Here, $[(f_{11}, f_{12})]$ denotes the equivalence class corresponding to interaction (f_{11}, f_{12}) relative to the equivalence relation $\sim$ and $\mathbb{F}_2^2(\mu)/\sim$ denotes the collection of all such classes of interactions. Let $P(\mathbb{F}_2^2(\mu)/\sim)$ denote all probability measures on $\mathbb{F}_2^2(\mu)/\sim$. For a fixed order $r \geq 1$, define the following subset of $P(\mathbb{F}_2^2(\mu)/\sim)$ subject to a moment-type condition using the metric d:

$$\mathcal{P}_r(\mathbb{F}_2^2(\mu)/\sim) := \Big\{ G \in P(\mathbb{F}_2^2(\mu)/\sim) \mid \\ \int d^r([(f_{21}, f_{22})], [(f_{11}, f_{12})]) \mathrm{d}G([(f_{21}, f_{22})]) < \infty \Big\}.$$

This class of probability measures can be shown to be independent of the choice of $[(f_{11}, f_{12})]$ and therefore the collection of order-r integrable probability measures on the quotient space $\mathbb{F}_2^2(\mu)/\sim$ is independent of the choice of the base class $[(f_{11}, f_{12})]$. We arrive at the following distance metric to compare between probability measures on the quotient space $\mathbb{F}_2^2(\mu)/\sim$. This is an instantiation of Wasserstein distances that arise in the theory of optimal transport in metric spaces *Villani* (2003).

Definition 5.3.1 (Wasserstein distances). Let $F, G \in \mathcal{P}_r(\mathbb{F}_2^2(\mu)/\sim)$. The Wasserstein distance of order r between F and G is defined as:

$$W_r(F, G) := \left(\inf_{\pi \in \Pi(F,G)} \int d^r([(f_{11}, f_{12})], [(f_{21}, f_{22})]) \mathrm{d}\pi([(f_{11}, f_{12})], [(f_{21}, f_{22})]) \right)^{1/r},$$

where $\Pi(F, G)$ is the collection of all joint distributions on $\mathbb{F}_2^2(\mu)/\sim \times \mathbb{F}_2^2(\mu)/\sim$ with marginals F and G.

5.3.1 Wasserstein barycenter and k-means clustering

In this section, we present the Wasserstein barycenter problem and highlight its connection to the k-means formulation.

Wasserstein barycenter problem Fixing the order $r = 2$, let $P_1, P_2, \ldots, P_N \in \mathcal{P}_2(\mathbb{F}_2^2(\mu)/\sim)$ be probability measures on $\mathbb{F}_2^2(\mu)/\sim$. Their second-order Wasserstein barycenter is a probability measure $\bar{P}_{N,\lambda}$ such that

$$\bar{P}_{N,\lambda} = \underset{P \in P_2(\mathbb{F}_2^2(\mu)/\sim)}{\operatorname{argmin}} \sum_{i=1}^{N} \lambda_i W_2^2(P, P_i).$$

The Wasserstein barycenter problem was first studied by *Agueh and Carlier* (2011). When P_i are themselves finite discrete probability measures on arbitrary metric spaces, efficient algorithms are available for obtaining locally optimal solutions to the above *Cuturi and Doucet* (2014).

k-means clustering problem The k-means clustering problem, when adapted to obtaining clusters in a non-Euclidean space of interactions, can be viewed as solving for the set S of k elements $[(g_{11}, g_{12})], \ldots, [(g_{k1}, g_{k2})] \in \mathbb{F}_2^2(\mu)/\sim$ such that, given samples $(f_{11}, f_{12}), \ldots, (f_{n1}, f_{n2}) \in \mathcal{F}^2(\mu)$

$$S = \underset{T:|T| \leq k}{\operatorname{argmin}} \sum_{i=1}^{n} \inf_{[(f_1', f_2')] \in T} d^2([(f_{i1}, f_{i2})], [(f_1', f_2')]). \tag{5.5}$$

It can be shown that this is equivalent to finding a discrete measure P which solves the following for the choice $r = 2$:

$$\inf_{P \in \mathcal{O}_k(\mathbb{F}_2^2(\mu)/\sim)} W_r(P, P_n), \tag{5.6}$$

where P_n is the empirical measure on $\mathbb{F}_2^2(\mu)/\sim$, i.e., P_n places mass $1/n$ on equivalence class sample $[(f_{i1}, f_{i2})]$ for all $i = 1, \ldots, n$, and $\mathcal{O}_k(\mathbb{F}_2^2(\mu)/\sim)$ is the set of all measures in $\mathbb{F}_2^2(\mu)/\sim$ with at most k support points. (It is interesting to note that Eq. (5.6) is a special case of the Wasserstein barycenter problem for $N = 1$ and $r = 2$.)

At the high level our approach is simple: we seek to summarize the empirical data distribution of interactions using a k-means-like approach, but there are several challenges due to the complex metric structure exhibited by the non-Euclidean space of interactions. Finding the exact solution even in the simplest cases is an NP-hard problem. The most common method to approximate the solution is the use of iterative steps similar to Lloyd's algorithm *Lloyd* (1982) for solving the Euclidean k-means problem. However, the computation of cluster centroids at each iteration of Lloyd's algorithm when applied to the non-Euclidean metric d is non-trivial. Moreover, the computation of pairwise distances between equivalence classes of interactions is non-trivial. In the next subsection we present some approximate solutions to Eq. (5.5).

5.3.2 Approximations for non-Euclidean k-means clustering

The primary objective for this section is to obtain a robust representation for the distribution over interaction primitives. Although the empirical distribution of interactions provides an estimate of the distribution over primitives, it suffers from lack of robustness guarantees. A robust k-approximation for the empirical distribution is formalized by Eq. (5.6). For order $r = 2$ this is equivalent to solving the k-means problem given by Eq. (5.5) for the interaction scenarios. The computational problem for computing exact centroids of k-means clusters is cumbersome and generally not solvable for arbitrary distance metrics d. To overcome such challenges we propose three separate methods to obtain approximate solutions to Eq. (5.5). The first approach is a standard application of multi-dimensional scaling technique. The second and third approaches are based on other geometric ideas to be described in the sequel.

5.3.2.1 Multidimensional Scaling

Multi-dimensional scaling (MDS) provides a way to obtain a lower dimensional representation of high-dimensional and/or non-Euclidean space elements while approximately preserving some distance measure among data points. Given a distance (a.k.a. dissimilarity) matrix $D = (d_{ij})_{1 \leq i,j \leq n}$, which collects all pairwise distance among the n data points using a notion of distance such as metric d described earlier, MDS finds points $x_1, \ldots, x_n \in \mathbb{R}^m$, for some small dimension m, such that

$$\{x_1, \ldots, x_n\} = \underset{y_1,\ldots,y_n \in \mathbb{R}^m}{\operatorname{argmin}} \sum_{i,j=1}^{n} (\|y_i - y_j\| - d_{ij})^2 \tag{5.7}$$

In order to apply the k-means clustering technique to our MDS representation, the following implicit assumption is required:

(C1) Each of the cluster centroids for the k-means problem corresponds to an interaction in the data sample.

Given (C1), Eq. (5.5) can be reformulated as follows.

Approximate k-means Given interaction samples $(f_{11}, f_{12}), \ldots, (f_{n1}, f_{n2}) \in \mathcal{F}^2(\mu)$, find a set $S \subset \{1, \ldots, n\}$ such that,

$$S = \underset{T:|T| \leq k}{\operatorname{argmin}} \sum_{i=1}^{n} \min_{j \in T} d^2([(f_{i1}, f_{i2})], [(f_{j1}, f_{j2})]). \tag{5.8}$$

The approximate k-means problem in Eq. (5.8) differs from the k-means problem (5.5) in that instead of finding primitives that are the global minimizer (and hence correspond to the cluster means), we look for the primitive that is closest to all other interactions in its cluster. The advantage of this approach is that we do not need explicitly the inverse map that goes from the MDS representation back to the interaction space. We summarize this approach as Algorithm 5.1 in the following.

Algorithm 5.1 Clustering interactions

```
Input: interaction sample $\{(f_{i1}, f_{i2})\}_{i=1}^{n}$
Output: $k$ interaction primitives
Obtain $x_1, \ldots, x_n$ as solution of MDS Eq. (5.7) with $d_{ij} = d([(f_{i1}, f_{i2})], [(f_{j1}, f_{j2})])$.
Perform k-means on $x_1, \ldots, x_n$ to obtain the centroids.
Approximate the centroids with points $x_i \in \mathbb{R}^m$ which are closest in $\|\cdot\|$ distance to
    the centroids, $\Gamma_1, \Gamma_2, \ldots, \Gamma_k$.
Return as primitives the $k$ interaction sample corresponding to these approximate
    centroids, $\{(g_{j1}, g_{j2})\}_{j=1}^{k}$.
```

5.3.2.2 Geometric Approximations

A major computational challenge to solving Eq. (5.8) lies in the SVD decomposition of the Procrustes distances (Eq. (5.1)) relative to each pair of interactions. There require $O(n^2)$ such decomposition. To avoid this, we instead consider a geometric approximation of the Procrustes distance, inspired by work from the field of morphometrics *Stegmann and Gomez*.

Consider two interactions (f_{i1}, f_{i2}) and (f_{j1}, f_{j2}). Then, by an application of triangle inequality,

$$\begin{aligned} & \inf_{(f_1', f_2') \in (\mathcal{O}, \mathcal{C})_{(f_{j1}, f_{j2})}} \rho((f_{i1}, f_{i2}), (f_1', f_2')) \\ = & \inf_{O_1 \in SO(2), c_1 \in \mathbb{R}^2} \rho((f_{i1}, f_{i2}), O_1 \odot (f_{j1}, f_{j2}) + c_1) \\ \leq & \inf_{O_1 \in SO(2), c_1 \in \mathbb{R}^2} \rho((f_{11}, f_{12}), O_1 \odot (f_{j1}, f_{j2}) + c_1) \\ + & \inf_{O_2 \in SO(2), c_2 \in \mathbb{R}^2} \rho((f_{11}, f_{12}), O_2 \odot (f_{i1}, f_{i2}) + c_2). \end{aligned} \tag{5.9}$$

Eq. (5.9) shows that knowledge of optimal rotational matrices and translation vectors for computing the distances $d([(f_{i1}, f_{i2})], [(f_{11}, f_{12})])$ and $d([(f_{j1}, f_{j2})], [(f_{11}, f_{12})])$ can provide an upper bound for computing the distance between the i^{th} and j^{th} pair of interactions. Therefore, we can provide a reasonable upper bound for all the n^2 pairwise distances by simply performing only $O(n)$ SVD decompositions. This approach, which we call the *first geometric approximation*, is summarized in Algorithm 5.2.

Algorithm 5.2 First Geometric Approximation

```
Input: ${(f_{i1}, f_{i2})}_{i=1}^{n}$
Output: $k$ centroids
for $i = 1, 2, \ldots, n$ do
    Center and reorient $(f_{i1}, f_{i2})$ to $(f_{11}, f_{12})$ using Algorithm 5.4.
end for
Perform k-means on the centered and oriented $\{(f_{i1}, f_{i2})\}_{i=1}^{n}$ to obtain the centroids, $\{(\Gamma_{j1}, \Gamma_{j2})\}_{j=1}^{k}$.
Return the centroids, $\{(\Gamma_{j1}, \Gamma_{j2})\}_{j=1}^{k}$.
```

However, this gain in computation efficiency is also accompanied by a loss of statistical efficiency. To mitigate this tension between computational and statistical efficiency we propose a *second geometric approximation* which performs the approximation of Algorithm 5.2 in batch form, where the batches comprise of the respective clusters. This procedure is described in Algorithm 5.3.

5.4 Experimental Results

In this section we provide a demonstration of our methods for unsupervised learning of vehicle interactions. In particular, we will evaluate the quality and stability of clustered primitives extracted from vehicle-to-vehicle interactions based on real-world experiments conducted in Ann Arbor, Michigan. In the literature for this application domain, a real-time interaction between two vehicles is also alternatively referred to

Algorithm 5.3 Second Geometric Approximation

```
Input: $\{(f_{i1}, f_{i2})\}_{i=1}^{n}$
Output: $k$ centroids
Randomly assign interaction samples $\{(f_{i1}, f_{i2})\}_{i=1}^{n}$ to $k$ clusters. Let $z_i$ indicate the cluster assignment.
while k-means convergence criterion has not been met do
    for $k' = 1, 2, \ldots, k$ do
        Center and orient all interaction samples $(f_{i1}, f_{i2})$ to $(f_{i_{k'}1}, f_{i_{k'}2})$ using Algorithm 5.4 if $(f_{i_{k'}1}, f_{i_{k'}2})$ is the first interaction sample such that $z_i = k'$ for $i = 1, 2, \ldots n$. Denote these oriented and centered samples as $(f'_{i1}, f'_{i2})(t)$.
        Compute the centroid for cluster $j$, $(\Gamma_{j1}, \Gamma_{j2})$, such that for $t = 1, 2, \ldots, t_m$,
            $(\Gamma_{j1}, \Gamma_{j2})(t) = \frac{1}{\#\,(z_i = k)} \sum_{i:z_i=k} (f'_{i1}, f'_{i2})(t)$
    end for
    for $i = 1, 2, \ldots, N$ do
        for $j = 1, 2, \ldots, k$ do
            Center and orient $(f_{i1}, f_{i2})$ to $(\Gamma_{j1}, \Gamma_{j2})$.
            Compute the $L_2$ distance between the centered and oriented $(f_{i1}, f_{i2})$ and $(\Gamma_{j1}, \Gamma_{j2})$.
        end for
        Set $z_i = j$ if the smallest computed distance is from the centroid of cluster $j$.
    end for
end while
Return the centroids, $\{(\Gamma_{j1}, \Gamma_{j2})\}_{j=1}^{k}$.
```

as an encounter. In practice, the interactions between vehicles are represented by multi-dimensional time series of varying duration, which need to be further segmented into shorter time duration via suitable data processing techniques.

5.4.1 Vehicle-to-vehicle (V2V) interaction data processing

We work with a real-world V2V interaction data set which is extracted from the naturalistic driving database generated by experiments conducted as part of the University of Michigan Safety Pilot Model Development (SPMD) program. In these experiments,

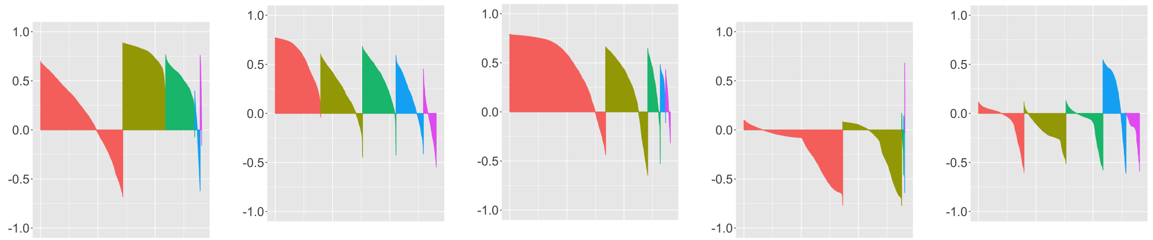


Figure 5.1: *Multidim. Scaling (cf. Section 5.3.2.1)*

Figure 5.2: *First geometric approx. (cf. Section 5.3.2.2)*

Figure 5.3: *Second geometric approx. (cf. Section 5.3.2.2)*

Figure 5.4: *Polynomial coefficients (cf. Section 5.4.2)*

Figure 5.5: *DTW cost matrix (cf. Wang and Zhao (2017))*

Figure 5.6: *Silhouette plots for 5 clusters obtained under various approaches:*

	Total within 3*Square Distance	**Cluster 1 Average within 3*Square Distance**	**Cluster 5 Average within 3*Square Distance**	**Cluster 1 Average between 3*Square Distance**	**Cluster 5 Average between 3*Square Distance**
MDS	10.68	1.74e-03 (1.53e-05)	1.23e-02 (4.60e-04)	1.55e-02 (4.57e-04)	1.17e-01 (5.13e-03)
First Geometric Approx.	13.32	9.50e-04 (2.96e-05)	6.33e-03 (3.92e-04)	5.01e-02 (4.97e-03))	1.12e-02 (2.92e-04)
Second Geometric Approx.	274.27	3.97e-03 (1.42e-04)	8.01e-02 (1.38e-03)	1.86e-02 (8.43e-04)	1.84e-02 (1.51e-03)
Spline Coefficients	222.56	5.02e-02 (3.91e-03)	9.40e-02 (1.05e-02)	5.95e-02 (4.27e-03)	2.05e-01 (2.33e-02)
DTW Matrices	201.12	3.12e-02 (4.51e-03)	9.00e-03 (3.41e-04)	5.01e-02 (6.30e-03)	9.38e-03 (3.79e-04)

Table 5.1: *A table of the quantities from Eq.* (5.8)*, Eq.* (5.11)*, and Eq.* (5.13)*) for each method's cluster with the most interaction (Cluster 1) and cluster with the fewest (Cluster 5). Variance of these distances are included in parentheses. Note that the Procrustes distances were normalized so that the maximum distance between any interaction is 1.*

dedicated short range communications (DSRC) technology was utilized for the communication between two vehicles. Approximately 3,500 equipped vehicles have collected data for more than 3 years. Latitude and longitude data of each vehicle was recorded by the by-wire speed sensor. The on-board sensor records data in 10Hz.

To investigate basic V2V interaction behaviors, a subset of 1400 driving scenarios was further filtered out from the SPMD's database. Each scenario consists of a time series of GPS locations and speeds of a pair of vehicles, which are mutually less than 100 metres apart. For our purposes, it is natural to posit that each scenario is inclusive of multiple shorter encounters through different time duration. Pre-processing of the data was therefore aimed at segmenting each scenario into more basic driving segments. These segments constitute basic building blocks from which we can meaningfully learn interaction primitives using a variety of clustering algorithms. The issue of segmentation is akin to identifying change points on functional curves embedded in a higher dimensional space. We consider two different segmentation schemes for V2V interaction data processing.

The first segmentation scheme is detailed in Appendix 5.6.2. It will be called a *two-step spline* approach, which goes as follows. Given an encounter, we fit it with cubic splines in two main steps. Here, the change points act as the knots. The first step involves identifying a large number of probable change points via a binary search approach to add change points if adding change points reduced the squared error between the fitted values and the observed data. The next step involves a single forward pass to remove excess change points from consideration in order to minimize the squared error with a penalty for the number of change points. We then segment each interaction at the knots. This segmentation technique created a set of 5622 basic V2V interactions to work with.

The second segmentation scheme is considerably more complex, as it is derived from a nonparametric Bayesian model for time series data, the sticky Hierarchical Dirichlet Process Hidden Markov Model (HDP-HMM) *Fox et al.* (2009). This model extends the basic HMM by allowing the number of hidden Markov states to be unbounded, while encouraging the Markov process to be "sticky", that is, the state tends to be constant for a period of time (e.g., a car tends to go straight after a long period of time). For model selection, as we will elaborate later, one of the hyperparameters of sticky HDP-HMM is varied. Consequently, the number of basic V2V interactions varied from 8779 to 8829 with an average of 8799 interactions.

5.4.2 Cluster analysis of V2V interactions

We evaluate the clustering of primitives qualitatively and quantitatively. For the former, silhouette plots are useful – the silhouette, $s(i)$, for interaction i is defined as following:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}.$$

Here, $a(i)$ is the average Procrustes distance between interaction i and all other interactions in the same cluster as interaction i, while $b(i)$ is the average Procrustes distance from interaction i to those in another cluster. The cluster used for $b(i)$ is the one that minimizes this average distance. By definition, the silhouette ranges from -1 to 1. It will be close to 1 if $b(i)$ is significantly larger than $a(i)$ and -1 if $a(i)$ is significantly larger than $b(i)$. Thus, the quality of the clustering for interaction i decreases as $s(i)$ decreases. Plotting the silhouettes for all interactions provides a qualitative way to determine how the clustering is performing because if most silhouettes are close to 1, the clustering is performing well.

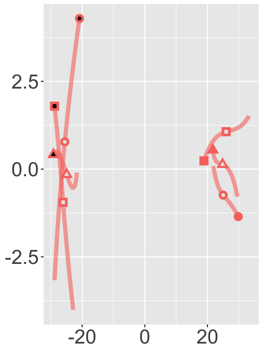


Figure 5.7: *Cluster 1*

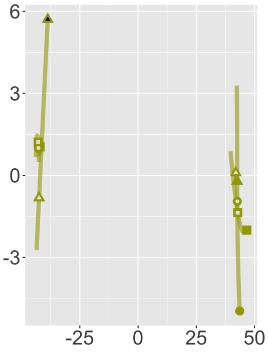


Figure 5.8: *Cluster 2*

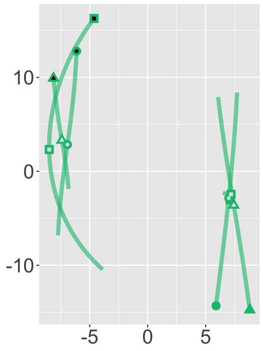


Figure 5.9: *Cluster 3*

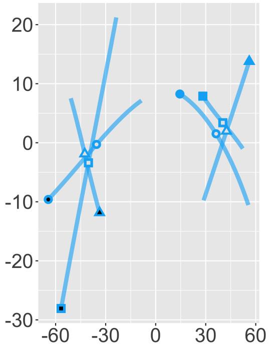


Figure 5.10: *Cluster 4*

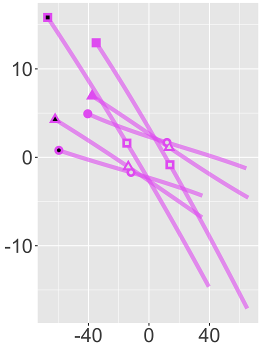


Figure 5.11: *Cluster 5*

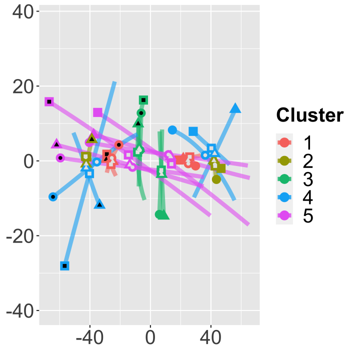


Figure 5.12: *All clusters*

Figure 5.13: *Plot of the three most typical interactions organized from the cluster with the most interaction to the cluster with the fewest for clustering using Multidimensional scaling (cf. Section 5.3.2). The interactions are centered and oriented using Algortihm 5.4 to* $(t, 2t-1, -t, 1-2t)$ *for* $t = 0, 0.01, \ldots, 1$. *The solid shapes and shapes with a black interior indicate the starting location of each interaction. The dot with the black interior indicates the second trajectory. Midpoints are indicated by dots filled in with a grey interior. Different shapes indicate different V2V interactions. Note that the individual cluster interactions plots are placed on their own scales.*

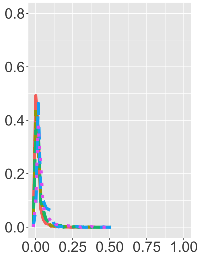


Figure 5.14: *In-cluster distances from mean interactions*

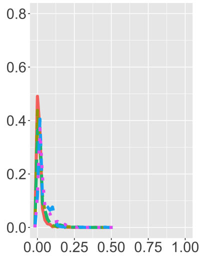


Figure 5.15: *In-cluster distances from typical interactions*

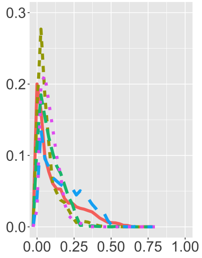


Figure 5.16: *All distances from mean interactions*

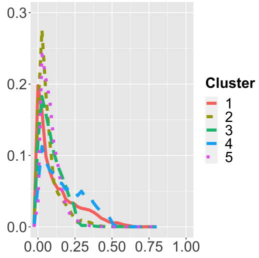


Figure 5.17: *All distances from typical interactions*

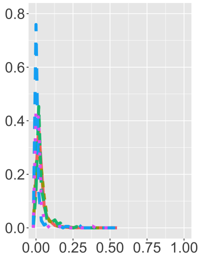


Figure 5.18: *In-cluster distances from the mean interactions*

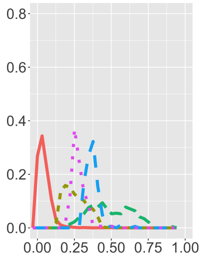


Figure 5.19: *In-cluster distances from the typical interactions*

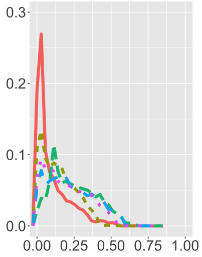


Figure 5.20: *All distances from mean interactions*

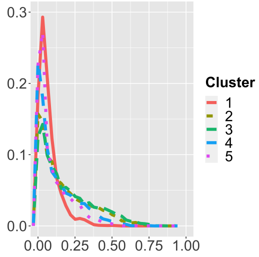


Figure 5.21: *All distances from the typical interactions*

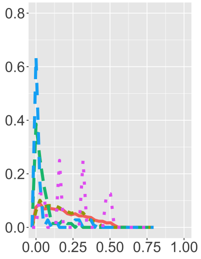


Figure 5.22: *In-cluster from mean interactions*

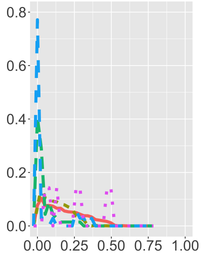


Figure 5.23: *In-cluster distances from typical interactions*

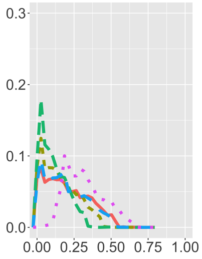


Figure 5.24: *All distances from mean interactions*

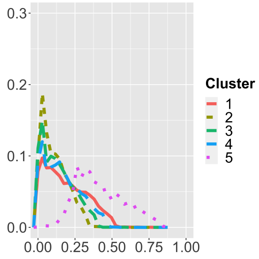


Figure 5.25: *All distances from typical interactions*

Figure 5.26: *Line plots showing the distribution (frequency) of interaction distance to either the cluster mean or the typical interaction. Clusters are obtained by the first geometric method in row 1, the second geometric method in row 2, and the cubic spline coefficients based method in row 3 (cf. Section 5.4.2). The clusters are numbered according to the number of interactions so that Cluster 1 has the most and Cluster 5 has the fewest. Note that the range for the y-axis are much larger on the left plots compared to the right plots.*

forming because if most silhouettes are close to 1, the clustering is performing well.

For a more quantitative way to examine the clustering, we look at the quantity in Eq. (5.5) and Eq. (5.8). Naturally, the method that reduces that quantity the most should be selected. We can also break down that quantity further by the contribution of each cluster. Specifically, suppose z_i indicates the cluster membership. If $(\Gamma_{j1}, \Gamma_{j2})$ is the cluster's mean, then we report the following:

$$\frac{1}{\#\left(z_i = k\right)} \sum_{i:z_i=k} d^2([(f_{i1}, f_{i2})], [(\Gamma_{j1}, \Gamma_{j2})]). \tag{5.10}$$

Alternatively, if interaction j minimizes $d^2([(f_{i1}, f_{i2})], [(g_{j1}, g_{j2})])$ for all $z_i, z_j = k$, the approximate version is the following:

$$\frac{1}{\#\left(z_i = k\right)} \sum_{i:z_i=k} d^2([(f_{i1}, f_{i2})], [(g_{j1}, g_{j2})]). \tag{5.11}$$

To compare clusters with different number of interactions, we choose to divide it by the size of the cluster. Finally, like the silhouette, it might also be helpful to compare this against the average square Procrustes distance of one cluster's mean V2V interaction and the interactions of all other clusters. In other words, we report the following if the cluster's mean interactions are recoverable:

$$\frac{1}{\#\left(z_i \neq k\right)} \sum_{i:z_i\neq k} d^2([(f_{i1}, f_{i2})], [(\Gamma_{j1}, \Gamma_{j2})]). \tag{5.12}$$

Again, we can report the approximate version instead:

$$\frac{1}{\#\left(z_i \neq k\right)} \sum_{i:z_i\neq k} d^2([(f_{i1}, f_{i2})], [(g_{j1}, g_{j2})]). \tag{5.13}$$

Note that the silhouette is more stringent because for the silhouette, we average only the distance from interactions of the nearest cluster for an observation.

We then made these silhouette plots and calculated the quantity for five different methods. First, we wanted to evaluate how well the three clustering approaches, namely the Multidimensional Scaling approximation and the first and second geometric approximations of the Procrustes distance, introduced in Section 5.3, performed. Next, because we segmented encounters using splines, we wanted to examine the quality of k-means clustering based on the coefficients of the cubic splines fitted to these interactions. In other words, suppose for interaction i, we fit the cubic spline $c^1_{i10} + c^1_{i11}t + c^1_{i12}t^2 + c^1_{i13}t^3$ to $(g_{i1})_1$. We do the same for $(g_{i2})_2$, $(g_{i2})_1$, and $(g_{i2})_2$. Then, we perform k-means on the vectors, $\{\{c^1_{i1\ell}\}^3_{\ell=0}, \{c^2_{i1\ell}\}^3_{\ell=0}, \{c^1_{i2\ell}\}^3_{\ell=0}, \{c^2_{i2\ell}\}^3_{\ell=0}\}^n_{i=1}$. We call this approach *spline coefficient clustering.* Finally, dynamic time warping (DTW) is a standard approach to match curves – in Wang et. al *Wang and Zhao* (2017), k-means clustering is performed on the DTW matrices that match one trajectory to another for each V2V interaction. This is another approach we wish to evaluate.

We focus on reporting for the case $k = 5$ for the moment, while the analysis can be replicated on other choices of k. The results for encounters segmented by the two step approach can be seen in Figure 5.6 and Table 5.1. Accordingly, the MDS approach outlined in Algorithm 5.1 appears to perform the best whereas the spline coefficients and DTW matrices perform the worst. The total within square distance from Eq. (5.8) for the MDS approach in Table 5.1 is smallest and the silhouette plots in Figure 5.6 look reasonable. Indeed, even though the first geometric approximation's average within square distance for the clusters with most and fewest interactions is smaller and the average between square distance is comparable or larger, the silhouette plot shows us that the MDS approach does significantly better with the cluster with the second and third largest cluster. The silhouette values are much higher for that cluster than the first geometric approximation.

For interpretability, one may be interested in visualizing typical interactions from each clusters. Take the MDS method. There are various interesting observations in Fig. 5.13. For instance, the two clusters with most interactions are interactions in which the vehicles do not move far from each other. On the other hand, the other clusters have interactions in which the opposite is true. Further, while only the cluster with the fewest interactions have vehicles going in the same direction, there are variation in how the vehicles are moving in opposite directions. Figure 5.41 in the Appendix shows the three most typical encounters for all clustering methods.

One can look more deeply into the distribution of interactions in each cluster, which is revealed by Figure 5.26. The left two plots show the proportion of interactions in each cluster a certain distance away from the mean or typical interaction for each cluster. On the other hand, the right two plots show the proportion of interactions from the entire data set a certain distance away for each cluster's mean or typical interaction. The first geometric approximation plot is ideal. While the other methods have a cluster that peaks higher near zero, the left two plots show higher peaks near zero across all clusters, indicating that most interactions in the cluster are close to the typical or mean interaction. On the other hand, the right two plots show a peak near zero and then plateau for a bit before decreasing to zero. This supports what we see in the silhouette plot. The plateau demonstrates that the clusters are well separated because interactions outside the cluster are further away. Because of the left two plots, the peak near zero likely comes from the interactions assigned to that cluster. It is likely that the plots for the MDS will look similar to the first geometric approximation plots. For the second geometric approximation, the interaction plots for the mean interaction are ideal.

However, outside of the largest cluster, the typical interaction to cluster interaction plots peak at values not near zero or plateau for ranges of distances. Meanwhile, the plots for polynomial coefficient exhibit peculiar peaks or plateaus in the left plots. These peaks are slightly dampened when using the typical interaction in place of the mean interaction.

5.4.3 Stability Evaluation

We had to develop a statistic for stability based on our distance metric. Consider the k-means problem in a Euclidean space. Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be points in an Euclidean space belonging to clusters $\{1, \ldots, K\}$. The cost function relative to the k-means problem is given by

$$\min_{\{\Gamma_1,\ldots,\Gamma_k,z_1,\ldots,z_n\}} \frac{1}{n} \sum_{j=1}^{k} \sum_{i:z_i=j} \|x_i - \Gamma_j\|^2.$$

The above cost function can also be written as:

$$\min_{\{z_1,\ldots,z_n\}} \frac{1}{2n} \sum_{k=1}^{K} \sum_{i,j:z_i,z_j=k} \|x_i - x_j\|^2. \tag{5.14}$$

Eq. (5.14) provides a way to partition the dataset $\{x_1, \ldots, x_n\}$ so as to optimize the within cluster distance. We then use a measure equivalent to (5.14) to evaluate the stability of algorithms. Namely, if we use the same notation as before, then the stability of the algorithm is measured by computing

$$\frac{1}{2n} \sum_{k=1}^{K} \sum_{i,j:z_i,z_j=k} d^2((f_{i1}, f_{i2}), (f_{j1}, f_{j2}))), \tag{5.15}$$

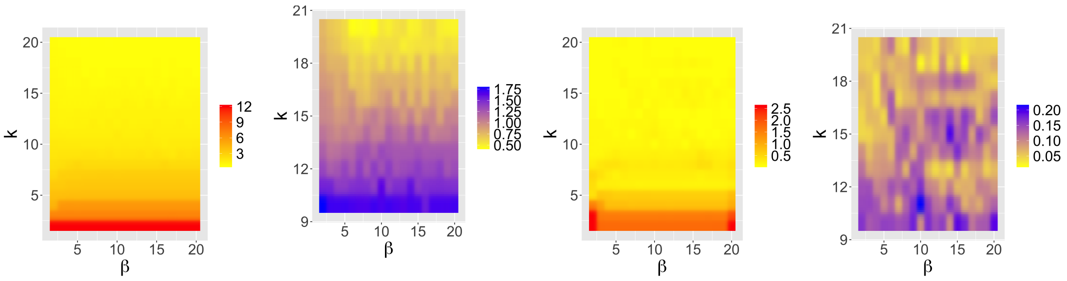


Figure 5.27: *All k and β between 2 and 20*

Figure 5.28: *All β between* 10 *and* 20 *and k between* 10 *and* 20

Figure 5.29: *All β between 2 and 20 and k between 10 and 20*

Figure 5.30: *All β between* 2 *and* 20 *and k between* 10 *and* 20

Figure 5.31: *The left two plots show heatmaps of the statistic introduced in Eq.* (5.15) *for encounters segmented by the two-step spline approach (cf. Appendix 5.6.2) and then clustered via MDS (cf. Section 5.3.2). The right two show changes in this statistic.*

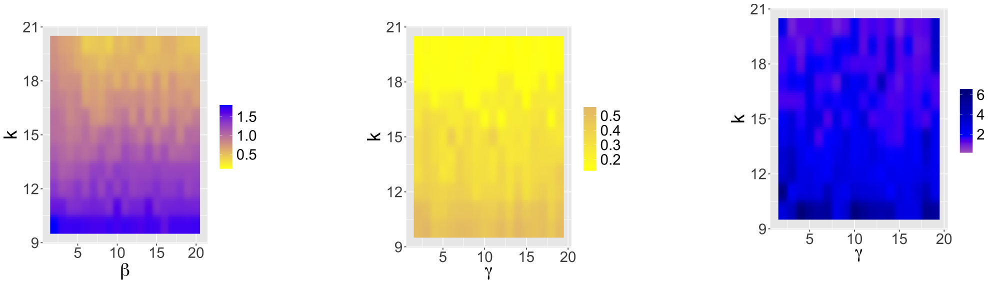


Figure 5.32: *MDS clustering (cf. Section 5.3.2) applied to two-step spline segmented encounters (cf. Appendix 5.6.2).*

Figure 5.33: *DTW matrix clustering applied to encounters segmented by BNP (cf. Wang and Zhao (2017)).*

Figure 5.34: *Two step spline segmented encounters clustered using primitives extracted from BNP segmented encounters clustered using DTW matrices (cf. 5.4.3).*

Figure 5.35: *Heatmaps of the statistic given by Eq.* (5.15) *for non-reflective Procrustes distance for $k \geq$* 10 *across different methods.*

for varying values of tuning parameters, where $d((f_{i1}, f_{i2}), (f_{j1}, f_{j2})))$ is the metric introduced in Eq. (5.2).

We then calculated this statistic for the MDS approach outlined in Section 5.3.2. Applying k-means to the MDS projection of the (non-reflective) Procrustes distance requires the specification of the following parameters: dimension of the projection, β, and the number of clusters, k. The results for $\beta \in [2, 20]$ and $k \in [2, 20]$ can be seen in Figure 5.31. From the heatmap, the MDS approach is particularly stable for $k \geq 15$ and $\beta > 5$. This is further supported by examining the change in the statistic introduced in Eq. (5.15). This makes sense because increasing the dimension for the MDS representation provides a better representation of the pairwise distance. On the other hand, increasing k also leads to greater stability. However, the scales in the figure suggests that most of the instability occurs when $k \leq 10$.

As before, we proceed to compare among methods and data sets. First, following Wang and Zhou *Wang and Zhao* (2017), we examined the stability of the DTW approach for the encounters segmented by sticky HMM-HDP. While there are more parameters to consider, we empirically investigated the results with α and c fixed to 2 and 100 respectively and allowed γ and k to vary between $[2, 19]$ and $[2, 20]$. Because changing γ gives us new primitives, we had to interpolate and recalculate the Procrustes distance for each set of primitives. Second, we wanted to inspect the stability of "transferring" primitives. In other words, let $\{(g'_{j1}, g'_{j2})\}_{j=1}^{k}$ be the primitives derived from applying BNP to segment encounters and using the DTW matrices to cluster them and $\{(f_{i1}, f_{i2})\}_{i=1}^{n}$ be the interactions extracted from the encounters using our two-step approach outlined in Appendix 5.6.2. We assign interaction i to cluster j if

$$j = \text{argmin}_{j'=1:k} d((f_{i1}, f_{i2}), (g'_{j1}, g'_{j2})).$$

Here, $d((f_{i1}, f_{i2}), (g'_{j1}, g'_{j2}))$ is the distance introduced in Eq. (5.1). The results can be seen in Figure 5.35. For DTW, the results are similar to the results before with respect to k and may even be better. On the other hand, as seen in the scales in Figure 5.35, we see that there is greater instability in both the range and the pattern when we "transfer" primitives. Further, unlike before, this instability persists even as k increases. This could be due to the more extreme values in the BNP primitive data set. As a result, there might be primitives that do not exist in the data set segmented by the two step spline approach. This could mean that as we increase k, we might not be adding centroids used to cluster the data. In addition, the ones that do exist might be influenced by these more extreme values. This might be why the values are unstable for lower values of k.

5.5 Conclusion

We developed a distance metric for the space of trajectory pairs that is invariant under translation and rotation. By using it to measure the distance between distributions, we could also use this metric for clustering and for evaluating a variety of unsupervised techniques for interaction learning. The distance metric and geometric approximation methods that we introduced help to address the challenges for robust learning of non-Euclidean quantities that represent temporally dynamic interactions. These techniques were demonstrated by the unsupervised learning of vehicle-to-vehicle interactions. An interesting direction for our work is to extend the metric based representation and geometric algorithms to the multiple-vehicle interaction setting, and general multi-agent settings. The challenge is the find a right metric or a family of metrics which are both meaningful and computationally tractable for a number of learning tasks of interests.

5.6 Appendix

5.6.1 Proofs

5.6.1.1 Proof of Proposition 5.2.1

We need to establish

(a) For any $f_{11}, f_{12}, f_{21}, f_{22} \in \mathbb{F}$,$d((f_{11}, f_{12}), (f_{21}, f_{22})) = 0$ if and only if $(f_{11}, f_{12}) \sim (f_{21}, f_{22})$.

(b) For any $f_{11}, f_{12}, f_{21}, f_{22} \in \mathbb{F}$, $d((f_{11}, f_{12}), (f_{21}, f_{22})) = d((f_{21}, f_{22}), (f_{11}, f_{12}))$.

(c) For any $f_{11}, f_{12}, f_{21}, f_{22}, f_{31}, f_{32} \in \mathbb{F}$, $d((f_{11}, f_{12}), (f_{21}, f_{22}))$
$\leq d((f_{31}, f_{32}), (f_{21}, f_{22})) + d((f_{11}, f_{12}), (f_{31}, f_{32}))$.

Condition (a) follows by definition. To establish (b), note $\rho((f_{11}, f_{12}), (f_1', f_2')) = \rho((f_1', f_2'), (f_{11}, f_{12}))$, so

$$\begin{aligned} & \rho((f_{11}, f_{12}), O_1 \odot (f_{21}, f_{22}) + c_1) \\ = \ & \rho(O_1 \odot (f_{21}, f_{22}) + c_1, (f_{11}, f_{12})) \qquad (5.16) \\ = \ & \rho((f_{21}, f_{22}), O_1^* \odot (f_{11}, f_{12}) - O_1^* \odot c_1), \end{aligned}$$

where the second equality is due to property (i) and (ii) in the proposition, with O_1^* being the conjugate transpose of O_1, which is also orthogonal when O_1 is. Now taking infimum over C_1 and O_1 the conclusion of part (b) is achieved.

For condition (c), notice that it is easy to see, following the argument similar to

Eq. (5.16), that

$$
\begin{aligned}
& \inf_{O_1,O_2\in SO(2);C_1,C_2\in\mathbb{R}^2} \rho(O_2\odot(f_{11},f_{12})+C_2, O_1\odot(f_{21},f_{22})+C_1) && (5.17)\\
= & \inf_{O_1\in SO(2),C_1\in\mathbb{R}^2} \rho((f_{11},f_{12}), O_1\odot(f_{21},f_{22})+C_1).
\end{aligned}
$$

Now for any $f_{31}, f_{32} \in \mathbb{F}$,

$$
\begin{aligned}
& \rho(O_2\odot(f_{11},f_{12})+C_2, O_1\odot(f_{21},f_{22})+C_1) && (5.18)\\
\leq & \rho(O_2\odot(f_{11},f_{12})+C_2,(f_{31},f_{32}))+\rho((f_{31},f_{32}),O_1\odot(f_{21},f_{22})+C_1),
\end{aligned}
$$

by triangle inequality applied to ρ. Taking infimum wrt $O_1, O_2 \in SO(2); C_1, C_2 \in \mathbb{R}^2$, the rest follows immediately.

5.6.1.2 Proof of Proposition 5.2.2

Note that

$$
\begin{aligned}
& d((f_{11},f_{12}),O\odot(f_{21},f_{22})+c))^2 := && (5.19)\\
& \int_0^\infty \Big(\|f_{11}(x)-O\cdot f_{21}(x)-c\|_2^2+\|f_{12}(x)-O\cdot f_{22}(x)-c\|_2^2\Big)\mu(\mathrm{d}x).
\end{aligned}
$$

Minimizing Eq. (5.19) with respect to c, for fixed O, we get

$$
c = \int_0^\infty \frac{f_{11}(x)+f_{12}(x)}{2}\mu(\mathrm{d}x) - O\cdot\left(\int_0^\infty \frac{f_{21}(x)+f_{22}(x)}{2}\mu(\mathrm{d}x)\right).
$$

Substituting this value of c, we obtain Eq. (5.20).

$$
\begin{aligned}
&\inf_{c\in\mathbb{R}^2}(\rho((f_{11},f_{12}),O\odot(f_{21},f_{22})+c))^2\\
&=-2\int\limits_0^\infty\left(f_{11}(x)-\int\limits_0^\infty\frac{f_{11}(x)+f_{12}(x)}{2}\mu(\mathrm{d}x)\right)^T\\
&\qquad\cdot O\cdot\left(f_{21}(x)-\int\limits_0^\infty\frac{f_{21}(x)+f_{22}(x)}{2}\mu(\mathrm{d}x)\right)\mu(\mathrm{d}x)\\
&-2\int\limits_0^\infty\left(f_{12}(x)-\int\limits_0^\infty\frac{f_{11}(x)+f_{12}(x)}{2}\mu(\mathrm{d}x)\right)^T\\
&\qquad\cdot O\cdot\left(f_{22}(x)-\int\limits_0^\infty\frac{f_{21}(x)+f_{22}(x)}{2}\mu(\mathrm{d}x)\right)\mu(\mathrm{d}x)\\
&+\sum_{i=1}^2\int\limits_0^\infty\left\|f_{2i}(x)-\int\limits_0^\infty\frac{f_{21}(x)+f_{22}(x)}{2}\mu(\mathrm{d}x)\right\|_2^2\mu(\mathrm{d}x)\\
&+\sum_{i=1}^2\int\limits_0^\infty\left\|f_{1i}(x)-\int\limits_0^\infty\frac{f_{11}(x)+f_{12}(x)}{2}\mu(\mathrm{d}x)\right\|_2^2\mu(\mathrm{d}x).
\end{aligned}
\tag{5.20}
$$

Minimizing Eq. (5.20) with respect to O is same as maximizing Eq. (5.21) with respect to $O\in SO(2)$.

$$
\begin{aligned}
&2\mathrm{trace}\Big(\int\limits_0^\infty\Big(f_{11}(x)-\int\limits_0^\infty\frac{f_{11}(x)+f_{12}(x)}{2}\mu(\mathrm{d}x)\Big)^T\cdot\\
&O\cdot\left(f_{21}(x)-\int\limits_0^\infty\frac{f_{21}(x)+f_{22}(x)}{2}\mu(\mathrm{d}x)\right)\mu(\mathrm{d}x)\Big)\\
&+2\mathrm{trace}\Big(\int\limits_0^\infty\left(f_{12}(x)-\int\limits_0^\infty\frac{f_{11}(x)+f_{12}(x)}{2}\mu(\mathrm{d}x)\right)^T\cdot\\
&O\cdot\Big(f_{22}(x)-\int\limits_0^\infty\frac{f_{21}(x)+f_{22}(x)}{2}\mu(\mathrm{d}x)\Big)\mu(\mathrm{d}x)\Big)\\
&=2\mathrm{trace}(UDV^T\cdot O)=2trace(D(U^T\cdot O^T\cdot V)^T).
\end{aligned}
\tag{5.21}
$$

Now, this is maximized for $O \in SO(2)$, when $O = V^T \begin{bmatrix} 1 & 0 \\ 0 & \det(V^T U) \end{bmatrix} U$. Plugging in this minimizing value for O, we get the solution for $\inf_{(f'_1, f'_2) \in (\mathcal{O}, \mathcal{C})_{(f_{21}, f_{22})}} (\rho((f_{11}, f_{12}), (f'_1, f'_2)))^2$ as required.

Lemma 5.6.1. Assume that $(f', g') \in \mathcal{C}_{(f,g)} \cup \mathcal{O}_{(f,g)} \implies d((f', g'), (f, g)) = 0$. Then, for all $c \in \mathbb{R}^2,\ O \in SO(2)$,

$$\begin{aligned} &d((f_{11}, f_{12}), (f_{21}, f_{22})) \\ &\quad = d((O \odot f_{11} + c, O \odot f_{12} + c), (f_{21}, f_{22})). \end{aligned} \tag{5.22}$$

Proof. By triangle inequality, $d((f_{11}, f_{12}), (f_{21}, f_{22})) \leq d((O \odot f_{11} + c, O \odot f_{12} + c), (f_{21}, f_{22})) + d((O \odot f_{11} + c, O \odot f_{12} + c), (f_{11}, f_{12}))$, so by assumption $d((f_{11}, f_{12}), (f_{21}, f_{22})) \leq d((O \odot f_{11} + c, O \odot f_{12} + c), (f_{21}, f_{22}))$. The lemma follows by considering the reverse inequality. □

5.6.2 Obtaining primitives via splines

Our two-step procedure to extract primitives is as follows.

1. Add change points for each trajectory via the following steps. (a) Test whether using the midpoint as a change point reduces the squared error of the fitted polynomial (b) If it does, return the midpoint. (c) Otherwise, test whether using the midpoint of the valid interval of the half with the larger square error as a change point reduces the squared error. Rule out the other half as a site for change points. (d) Repeat (b)-(c) until either a change point is found or no further candidates exist. (e) If a change point was added previously, repeat (a)-(d) for the

two segments and any subsequent segments. Stop when no more change points are added.

2. Combine the change points from all trajectories in the following manner. Remove change points via a forward search in the following way. Suppose that we have a set, $\mathcal{C}$, of L ordered change points, $c_1, c_2, ..., c_L$, across all trajectories. Let c_0 denote the start point and c_{L+1} denote the end point. Define ϵ to be our tolerence. Proceed in these steps: (a) Set $\ell = 0$, $\ell' = 1$, and $\ell'' = 2$; (b) Fit a polynomial to each trajectory from c_ℓ to $c_{\ell''}$; (c) If the sum of the squared error of the fitted polynomials is below ϵ or there are only 4 observations between c_ℓ and $c_{\ell''}$, remove $c_{\ell'}$ from the set of $\mathcal{C}$. Otherwise, increment ℓ. Increase ℓ' and ℓ'' by one and go back to (b) if $\ell'' \leq L+1$; (d) Set L to be the size of $\mathcal{C}$. Re-index the change points in $\mathcal{C}$ from one to L and return $\mathcal{C}$.

   To select ϵ from a set of potential tolerances, we set it to be the value that after running (2), minimizes

$$\sum_{i=1}^{n}\sum_{\ell=1}^{L+1}\Big(f(t_i) - \hat{f}_\ell(t_i)\Big)^2 \mathbb{1}_{(t_i \leq c_\ell)} + L + 2.$$

5.6.3 Algorithm for centering and reorienting primitives

Algorithm 5.4 provides a way to reorient one set of interactions to another and is embedded in Algorithms 5.2 and 5.3.

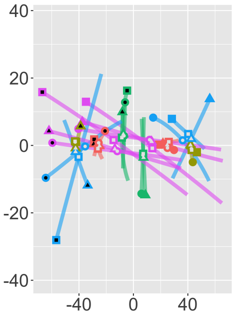


Figure 5.36: *Multidim. scaling approach (cf. Section 5.3.2.1)*

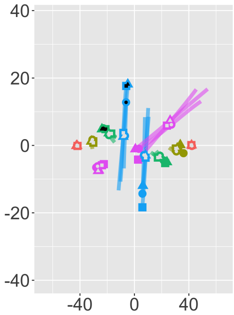


Figure 5.37: *First geometric approx. (cf. Section 5.3.2.2)*

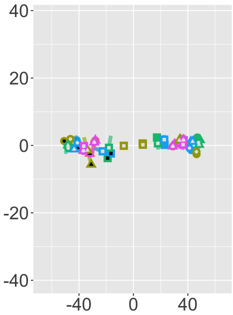


Figure 5.38: *Second geometric approx. (cf. Section 5.3.2.2)*

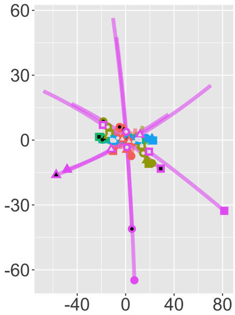


Figure 5.39: *Polynomial coefficients (cf. Section 5.4.2)*

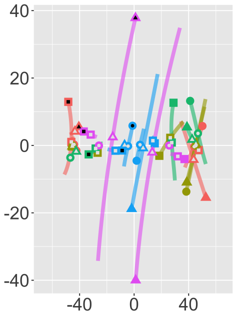


Figure 5.40: *DTW cost matrix (cf. Wang and Zhao (2017))*

Figure 5.41: *Plot of the three most typical interactions organized from the cluster with the most interaction to the cluster with the fewest for various methods. See Figure 5.13 for the legend and for how the interactions are oriented.*

Algorithm 5.4 Centering and reorienting interactions

```
Input: Two interaction samples, $(f_{i1}, f_{i2})$ and $(f_{j1}, f_{j2})$
Output: A centered $(f_{i1}, f_{i2})$ and a centered $(f_{j1}, f_{j2})$ reoriented to the centered $(f_{i1}, f_{i2})$
Set $\widetilde{f}_i(t) \in \mathbb{R}^{2t_m}\mathbb{R}^2$ to be the concatenation of $f_{i1}$ and $f_{i2}$ such that $\widetilde{f}_i(t) = f_{i1}(t)$ for $t = 1, 2, \ldots t_m$ and $\widetilde{f}_i(t) = f_{i2}(t)$ for $t = t_m + 1, t_m + 2, \ldots 2t_m$ and $\widetilde{f}_j(t) \in \mathbb{R}^{2t_m}\mathbb{R}^2$ be the same concatenation of $f_{i1}$ and $f_{i2}$.
For $\overline{f}_i(t) \in \mathbb{R}^{2t_m}\mathbb{R}^2$ and $\overline{f}_i(t) \in \mathbb{R}^{2t_m}\mathbb{R}^2$, set
    $\overline{f}_i(t) = \widetilde{f}_i(t) - \frac{1}{2t_m}\sum_{t'=1}^{2t_m} \widetilde{f}_i(t')$
    $\overline{f}_j(t) = \widetilde{f}_j(t) - \frac{1}{2t_m}\sum_{t'=1}^{2t_m} \widetilde{f}_j(t').$
Perform singular value decomposition to get the matrices $U$, $D$, $V$ such that $UDV^T = \overline{f}_j(t)^T\overline{f}_i(t)$.
From before, let
    $\tilde{\mathcal{O}} = V^T \begin{bmatrix} 1 & 0 \\ 0 & \det(V^T U) \end{bmatrix} U.$
Then, set $\overline{f}'_{i1}(t), \overline{f}'_{i2}(t) \in \mathbb{R}^T\mathbb{R}^2$ to be the matrices such that for $t = 1, 2, \ldots, T$,
    $\overline{f}'_{i1}(t) = \overline{f}_i(t)$
    $\overline{f}'_{i2}(t) = \overline{f}_i(t + t_m).$
On the other hand, set $\overline{f}'_{j1}(t), \overline{f}'_{j2}(t) \in \mathbb{R}^T\mathbb{R}^2$ to be the matrices such that for $t = 1, 2, \ldots, T$,
    $\overline{f}'_{j1}(t) = (\tilde{\mathcal{O}}\overline{f}_j)(t)$
    $\overline{f}'_{j2}(t) = (\tilde{\mathcal{O}}\overline{f}_j)(t + t_m).$
Return $(\overline{f}'_{i1}, \overline{f}'_{i2})$ and $(\overline{f}'_{j1}, \overline{f}'_{j2})$.
```

CHAPTER VI

Conclusions and Future Work

This dissertation develops a deeper understanding of the behaviors of latent structural models and provides algorithms for scalable and statistically efficient inference with them. Our main contributions are summarized as follows:

- An algorithmic approach to estimating the number of components in Bayesian mixture models along with an exploration of the consistency behavior of the same.
- A comprehensive understanding of the behavior of Bayesian nonparametric models under various misspecified settings and the impact of the choice of kernel.
- A general modeling framework for data arising from heterogeneous populations and a fast parametric geometric algorithm for inference.
- An unsupervised learning approach for evaluating clustering algorithms with application to traffic encounters.

We provide a summary of future endeavors in the next sections.

6.1 A velocity flow model to analyse traffic movements

In Chapter V of this thesis, we develop a distributional viewpoint to analyse traffic driving patterns. The primary motivation for the problem arises from the endeavour to develop an extensive understanding of autonomous vehicles. While such a viewpoint is useful to model one-on-one traffic interactions, practical applications also demand a broader perspective. For example, autonomous vehicles need to navigate through various surroundings at different points in time. The number of active on-road vehicles could be potentially infinite with varying degrees of correlations between the vehicles. This raises a challenging question of modeling multi-vehicle interactions under different on-road traffic scenarios. *Guo et al.* (2019) provide a very useful model via the use of Gaussian random fields. However, their model relies on the key assumption that traffic flow patterns are independent over time within the same spatial domain. While this assumption may not be unreasonable for freeway traffic, presence of traffic signals within city limits makes it impractical.

A possible solution may be obtained by employing an Hierarchical Dirichlet process-Hidden Markov model (HDP-HMM) (*Fox et al.* (2009)) over the Gaussian random field framework. However, this makes the inference problem quite challenging. While MCMC algorithms may not be feasible for such large datasets, Variational inference algorithms are not guaranteed to be statistically efficient under such complex modeling setups. To provide a solution for this problem, we aim to explore a statistically and computationally efficient solution by the use of geometric algorithms based on the framework of *Yurochkin et al.* (2019).

6.2 Misspecification in the number of mixture model components

In Chapter II and III we provide a systematic understanding of Bayesian nonparametric mixture models under various scenarios of misspecification. One common form of misspecification is that in the number of components. Consider the following model.

Let $X_1, \ldots, X_n$ be i.i.d samples from true distribution P_{G_0} with density function

$$p_{G_0} := \int f(x|\theta) dG_0(\theta) = \sum_{i=1}^{k_0} p_i^0 f(x|\theta_i^0)$$

where $G_0 = \sum\limits_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ is a true but unknown mixing distribution with exactly k_0 number of support points. Here, $\left\{f(x|\theta), \theta \in \Theta \subset \mathbb{R}^d\right\}$ is a given family of probability densities with respect to a sigma-finite measure μ on $\mathcal{X}$ and Θ is a bounded set of $\mathbb{R}^d$ where $d \geq 1$. We use $\mathcal{E}_k := \mathcal{E}_k(\Theta)$ and $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ respectively to denote the space of all mixing measures with exactly and at most k components in Θ.

Motivation: In practice, the knowledge of the kernel f is not available to the statistician. The choice of the kernel may, therefore be completely subjective. This may give rise to situations when $k_0 = \infty$. Under such circumstances, there may be some atoms θ_i^0 such that their associated mass (weight) p_i^0 is very small. Additionally, there may also be some atoms θ_i^0 and θ_j^0 such that their distance is very close. Due to these issues, the estimation of the parameters and weights under these situations suffers from very slow convergence rates. Moreover the number of true components may be underestimated due to lack of representation of some of the heterogeneous components. For practical application, therefore, it is common custom (*Ishwaran and Zarepour* (2002)) to choose the number of components K to be large enough so as to overestimate the true number of components. *Rousseau and Mengersen* (2011) provides a critical analysis of overfitted mixtures. This, however, raises a different concern. If K is chosen to be too large, the rate of contraction suffers, while the choice of a small K risks underestimation scenarios, where the contraction properties are not well-understood. As it is often not very hard to estimate the closed atoms and very small weights, we would like to estimate some mixing measures G_* from a model with probability measures that has $k < k_0$ components. The choice of k ensures that we obtain the most informative atoms and weights while the remaining atoms and weights may be mixed together. In particular, we consider the MLE estimation of mixture models as follows

$$\widehat{G}_n = \mathop{\arg\max}_{G \in \mathcal{O}_k} \sum_{i=1}^{n} \log p_G(X_i).$$

Assume that there exists a discrete mixing measure G_* that minimizes the KL divergence between p_{G_0} and p_G, i.e.,

$$G_* = \mathop{\arg\min}_{G \in \mathcal{O}_k} KL(p_{G_0}, p_G).$$

Since $\mathcal{O}_k$ is not convex set in terms of mixing measures, G_* may not be unique. This also raises an issue since parameter or density estimation is not well-studied when the underlying space of probability densities is non-convex. Given this theoretical challenge, our future goals are two-fold: providing an in-depth analysis for estimation under non-convexity of the parameter space and using the results developed to promote an understanding of asymptotic and finite sample behaviors for under-fitted mixtures with $k < k_0$.

6.3 Statistical efficiency of Variational Inference for hierarchical Bayesian models

While Markov Chain Monte Carlo methods are statistically efficient, computationally they may not be as effective with complex models especially under time-constraints, for example, when dealing with online learning problems with hierarchical models. Variational Inference (*Blei et al.* (2003); *Hoffman et al.* (2013)) provides a computationally efficient alternative for inference to MCMC algorithms, especially when the underlying model is complicated and involves numerous latent variables. However, statistical efficiency of Variational Inference algorithms remain poorly studied. Even though some of the recent works by *Wang and Blei* (2018, 2019); *Yang et al.* (2017) provide a theoretical understanding of the statistical efficiencies of VI algorithms, general frameworks involving complex hierarchical models such as Hierarchical Dirichlet Process (*Teh et al.* (2006)) remain largely under-studied. Moreover, an understanding of the estimation of the number of components when the underlying model is a mixture model is also not well understood. This provides an opportunity to extensively explore the statistical and theoretical properties of VI algorithms and is one of the avenues of research we aim to explore in the near future.

6.4 Geometric inference for hierarchical models

Chapter IV of this thesis introduces a general modeling framework for hierarchical models, while also providing a statistically and computationally efficient solution for the same. The inference algorithm developed in the chapter relies on two key assumptions: (a) the knowledge of the number of topics K, and (b) assumption of symmetry of the

Dirichlet parameter α.

The asymmetry of α vastly extends the applicability of the model to arbitrary datasets. While K, the number of topics can be estimated using nonparametric techniques such as in *Yurochkin et al.* (2017), the question of inference for asymmetric α yet remains unresolved. Potential approaches could involve the use of a probabilistic or differential gemetric transform instead of a linear transform as employed in this thesis. Secondly, the consistency results are obtained under asymptotic conditions as the number of words in each document increases to ∞. An explicit computation of the error bounds of inferential parameters under finite sample cases could also prove to be a potential avenue of research. While *Nguyen* (2015) provides some similar results, the bounds are not guaranteed to be sharp. An in-depth analysis of sharp finite sample bounds could prove to be of interest for both theoretical and practical reasons.

BIBLIOGRAPHY

BIBLIOGRAPHY

Agueh, M., and G. Carlier (2011), Barycenters in the Wasserstein space, *Journal on Mathematical Analysis*, *43*(2), 904–924.

Anandkumar, A., D. P. Foster, D. J. Hsu, S. M. Kakade, and Y.-K. Liu (2012), A spectral algorithm for Latent Dirichlet Allocation, in *Advances in Neural Information Processing Systems*, pp. 917–925.

Anderson, T., and H. Rubin (1956), Statistical inference in factor analysis, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, V, U. Cal, Berkeley*, 111–150.

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

Arora, S., R. Ge, R. Kannan, and A. Moitra (2012a), Computing a nonnegative matrix factorization–provably, in *Proceedings of the 44th annual ACM symposium on Theory of computing*, pp. 145–162, ACM.

Arora, S., R. Ge, and A. Moitra (2012b), Learning topic models – going beyond SVD, in *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pp. 1–10, IEEE.

Arora, S., R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu (2013), A practical algorithm for topic modeling with provable guarantees, in *Proceedings of the 30th International Conference on Machine Learning*, pp. 280–288.

Bailey, T. L., C. Elkan, et al. (1994.), Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Banfield, J., and A. Raftery (1993), Model based gaussian and non-gaussian clustering., *Biometrics*, *49*, 803–821.

Baum, L., and T. Petrie (1966), Statistical inference for probabilistic functions of finite state markov chains, *Ann. Math. Statist.*, *37*, 1554–1563.

Baum, L., T. Petrie, G. Soules, and N. Weiss (1970), A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains, *Ann. Math. Statist.*, *41*, 164–171.

Bender, A., G. Agamennoni, J. R. Ward, S. Worrall, and E. M. Nebot (2015), An unsupervised approach for inferring driver behavior from naturalistic driving data, *IEEE Trans. Intell. Transport. Syst.*, *16*(6), 3325–3336.

Birge, L. , and P. Massart (1998), Minimum contract estimators on sieves: exponential bounds and rates of convergence, *Bernoulli*, *4*, 329–375.

Blackwell, D., and J. MacQueen (1973), Ferguson distributions via Polya urn schemes, *Annals of Statistics*, *1*, 353–355.

Blei, D., and M. Jordan (2006), Variational inference for dirichlet process mixtures, *Bayesian Analysis*, *1*, 121–144.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003), Latent Dirichlet allocation, *J. Mach. Learn. Res*, *3*, 993–1022.

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

Caillerie, C., F. Chazal, J. Dedecker, and B. Michel (2011), Deconvolution for the wasserstein metric and geometric inference, *Electron. J. Statist.*, *5*, 1394–1423.

Carpenter, B., et al. (2017), Stan: A probabilistic programming language, *Journal of statistical software*, *76*(1).

Carroll, R. J., and P. Hall (1988), Optimal rates of convergence for deconvolving a density, *Journal of American Statistical Association*, *83*, 1184–1186.

Cemgil, A. T. (2009), Bayesian inference for nonnegative matrix factorisation models, *Computational intelligence and neuroscience*, *2009*.

Chambaz, A., and J. Rousseau (2008), Bounds for Bayesian order identification with application to mixtures, *Annals of Statistics*, *36*(2), 938–962.

Chan, T.-H., C.-Y. Chi, Y.-M. Huang, and W.-K. Ma (2009), A convex analysis-based minimum-volume enclosing simplex algorithm for hyperspectral unmixing, *IEEE Transactions on Signal Processing*, *57*(11), 4418–4432.

Chatterjee, S., and A. S. Hadi (Eds.) (1988), *Sensitivity Analysis in Linear Regression*, Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., Hoboken, NJ, USA, doi:10.1002/9780470316764.

Chen, J. (1995), Optimal rate of convergence for finite mixture models, *Annals of Statistics*, *23*(1), 221–233.

Coutsias, E. A., C. Seok, and K. A. Dill (2004), Using quaternions to calculate rmsd, *Journal of Computational Chemistry*, *25*, 1849–1857.

Cutler, A., and L. Breiman (1994), Archetypal analysis, *Technometrics*, *36*(4), 338–347.

Cuturi, M., and A. Doucet (2014), Fast computation of wasserstein barycenters, in *International Conference on Machine Learning*.

Dacunha-Castelle, D., and E. Gassiat (1997), The estimation of the order of a mixture model, *Bernoulli*, *3*, 279–299.

Ding, W., W. Wang, and D. Zhao (2018), Multi-vehicle trajectories generation for vehicle-to-vehicle encounters, *arXiv preprint arXiv:1809.05680*.

Donoho, D., and V. Stodden (2003), When does non-negative matrix factorization give a correct decomposition into parts?, in *Advances in neural information processing systems*.

Doob, J. (1948), Applications of the theory of martingales, *Le calcul des Probabilites et ses Applications, Colloques Internationales du CNRS, Paris*, pp. 22–28.

Du, Q., V. Faber, and M. Gunzburger (1999), Centroidal Voronoi Tessellations: Applications and algorithms, *SIAM Review*, *41*(4), 637–676.

Eltoft, T., T. Kim, and T. Lee (2006), On the multivariate laplace distribution, *IEEE Signal Processing Letters*, *13*, 300–303.

E.M.Stein, and R. Shakarchi (2010), *Complex Analysis*, Princeton University Press.

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

Fan, J. (1991), On the optimal rates of convergence for nonparametric deconvolution problems, *Annals of Statistics*, *19*, 1257–1272.

Ferguson, T. (1973), A Bayesian analysis of some nonparametric problems, *Ann. Statist.*, *1*, 209–230.

Figueiredo, M., and A. K. Jain (1993), Unsupervised learning of finite mixture models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*.

Foerster, J., N. Nardelli, G. Farquhar, T. Afouras, P. H. S. Torr, P. Kohli, and S. Whiteson (2017), Stabilising experience replay for deep multi-agent reinforcement learning.

Fox, E., E. Sudderth, M. I. Jordan, and A. Willsky (2009), The sticky HDP-HMM: Bayesian nonparametric hidden Markov models with persistent states, *Tech. Rep. P-2777*, MIT LIDS.

Frazzoli, E., M. A. Dahleh, and E. Feron (2005), Maneuver-based motion planning for nonlinear systems with symmetries, *IEEE Transactions on Robotics*, *21*(6), 1077–1091.

Frazzoli, E., M. A. Dahleh, and E. Feron (2005), Maneuver-based motion planning for nonlinear systems with symmetries, *IEEE transactions on robotics*, *21*(6), 1077–1091.

Fúquene, J., M. Steel, and D. Rossell (2019), On choosing mixture components via non-local priors, *Journal of the Royal Statistical Society Series B*.

Gao, F., and A. W. van der Vaart (2016), Posterior contraction rates for deconvolution of Dirichlet-Laplace mixtures, *Electronic Journal of Statistics*, *10*, 608–627.

Ghahramani, Z. (2004), Unsupervised learning, in *Advanced lectures on machine learning*, pp. 72–112, Springer.

Ghosal, S., and A. van der Vaart (2007), Posterior convergence rates of Dirichlet mixtures at smooth densities, *Ann. Statist.*, *35*, 697–723.

Ghosal, S., and A. van der Vaart (2017), *Fundamentals of nonparametric Bayesian inference, vol. 44 of Cambridge Series in Statistical and Probabilistic Mathematics.*, Cambridge University Press, Cambridge.

Ghosal, S., and A. W. van der Vaart (2001), Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities, *Ann. Statist.*, *29*, 1233–1263.

Ghosal, S., J. K. Ghosh, and R. V. Ramamoorthi (1999), Posterior consistency of Dirichlet mixtures in density estimation, *Ann. Statist.*, *27*, 143–158.

Ghosal, S., J. K. Ghosh, and A. W. van der Vaart (2000), Convergence rates of posterior distributions, *Ann. Statist.*, *28*(2), 500–531.

Gillis, N., and S. Vavasis (2012), Fast and robust recursive algorithms for separable nonnegative matrix factorization, *arXiv preprint arXiv:1208.1237*.

Gillis, N., and S. A. Vavasis (2014), Fast and Robust Recursive Algorithms for Separable Nonnegative Matrix Factorization, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(4), 698–714, doi:10.1109/TPAMI.2013.226.

Gower, J. (1975), Generalized procrustes analysis, *Psychometrika*, *40(1)*, 33–51.

Graf, S., and H. Luschgy (2000), *Foundations of quantization for probability distributions*, Springer-Verlag, New York.

Green, P., and S. Richardson (2001), Modelling heterogeneity with and without the Dirichlet process, *Scandinavian Journal of Statistics*, *28*, 355—-377.

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

Guha, A., N. Ho, and X. Nguyen. (2019), On posterior contraction of parameters and interpretability in bayesian mixture modeling, *Arxiv. https://arxiv.org/abs/1901.05078. Under Review,* Bernoulli.

Guha, A., R. Lei, J. Zhu, X. Nguyen, and D. Zhao (2020), Robust unsupervised representation learning of temporal dynamic interactions, *Arxiv. https://arxiv.org/abs/2006.10241. Under Review,* IEEE Transactions on Neural Networks and Learning Systems .

Guha, A., C.-Y. Yang, N. Ho, X. Nguyen, and M. I. Jordan. (2020+), Bayesian contraction for dirichlet process mixtures of smooth densities, *Under Review,* Annals of Statistics.

Guo, Y., V. V. Kalidindi, M. Arief, W. Wang, J. Zhu, H. Peng, and D. Zhao (2019), Modeling multi-vehicle interaction scenarios using gaussian random field, *arXiv preprint arXiv:1906.10307*.

Gustafson, P. (2000), Local Robustness in Bayesian Analysis, in *Robust Bayesian Analysis*, edited by D. R. Insua and F. Ruggeri, Lecture Notes in Statistics, pp. 71–88, Springer, New York, NY, doi:10.1007/978-1-4612-1306-2_4.

Hamada, R., T. Kubo, K. Ikeda, Z. Zhang, T. Shibata, T. Bando, K. Hitomi, and M. Egawa (2016), Modeling and prediction of driving behaviors using a nonparametric bayesian method with ar models, *IEEE Trans. Intell. Veh.*, *1*(2), 131–138.

Heinrich, P., and J. Kahn (2018), Strong identifiability and optimal minimax rates for finite mixture estimation, *Annals of Statistics*, *46*, 2844–2870.

Hjort, N., C. Holmes, P. Mueller, and S. Walker (2010), *Bayesian Nonparametrics: Principles and Practice*, Cambridge University Press.

Ho, N., and X. Nguyen (2016), On strong identifiability and convergence rates of parameter estimation in finite mixtures, *Electronic Journal of Statistics*, *10 (1)*, 271–307.

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

Ho, N., X. Nguyen, and Y. Ritov (to appear), Robust estimation of mixing measures in finite mixture models, *Bernoulli*.

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

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

Horn, B. (1986), Closed-form solution of absolute orientation using unit quaternions, *Journal of the Optical Society of America*, *4*, 629–642.

Huang, K., X. Fu, and N. D. Sidiropoulos (2016), Anchor-Free Correlated Topic Modeling: Identifiability and Algorithm, *arXiv:1611.05010 [cs, stat]*.

Hyvärinen, A., J. Karhunen, and E. Oja (2004), *Independent component analysis*, vol. 46, John Wiley & Sons.

Ishwaran, H., and M. Zarepour (2002), Dirichlet prior sieves in finite normal mixtures, *Statistica Sinica*, *12*, 941–963.

Ishwaran, H., L. James, and J. Sun (2001), Bayesian model selection in finite mixtures by marginal density decompositions, *Journal of American Statistical Association*, *96*(456), 1316–1332.

Jain, S., and R. M. Neal (2007), Splitting and merging components of a nonconjugate Dirichlet process mixture model, *Bayesian Analysis*, *2*, 445–472.

J.Griffin, and S.Walker (), Posterior simulation of normalized random measure mixtures, *Journal of Computational and Graphical Statistics*, *20*.

Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999), An introduction to variational methods for graphical models, *Machine Learning*, *37*(2), 183–233.

Joseph, J., F. Doshi-Velez, A. S. Huang, and N. Roy (2011), A bayesian nonparametric approach to modeling motion patterns, *Autonomous Robots*, *31*(4), 383.

Kabsch, W. (1976), A solution for the best rotation to relate two sets of vectors, *Acta Crystallographica*, *32*, 922–923.

Kabsch, W. (1978), A discussion of the solution for the best rotation to relate two sets of vectors, *Acta Crystallographica*, *34*, 827–828.

Kass, R., and A. Raftery (1995), Bayes factors, *Journal of the American Statistical Association*, *90*, 773–795.

Kleijn, B., and A. van der Vaart (2006), Misspecification in infinite-dimensional Bayesian statistics, *Annals of Statistics*, *34*, 837–877.

Kleijn, B., and A. van der Vaart (2012), The misspecified Bernstein-Von Mises theorem, *Electron. J. Statist.*, *6*, 354–381.

Kotz, S., T. J. Kozubowski, and K. Podgorski (), *The Laplace distribution and generalizations*, Birkhauser Boston, Inc., Boston, MA.

Kucukelbir, A., D. Tran, R. Ranganath, A. Gelman, and D. M. Blei (2017), Automatic differentiation variational inference, *Journal of Machine Learning Research*, *18*(1), 430–474.

Kumar, A., V. Sindhwani, and P. Kambadur (2013), Fast conical hull algorithms for near-separable non-negative matrix factorization, in *International Conference on Machine Learning*, pp. 231–239.

Lee, D. D., and H. S. Seung (2001), Algorithms for non-negative matrix factorization, in *Advances in neural information processing systems*, pp. 556–562.

Leroux, B. (1992), Consistent estimation of a mixing distribution, *Annals of Statistics*, *20*(3), 1350–1360.

Liao, T. W. (2005), Clustering of time series data—a survey, *Pattern recognition*, *38*(11), 1857–1874.

Lindsay, B. (1995), *Mixture models: Theory, geometry and applications*, In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA.

Lipman, Y., R. Al-Aifari, and I. Daubechies (2013), The continuous procrustes distance between two surfaces, *https://doi.org/10.1002/cpa.21444*.

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

Lo, A. Y. (1984), On a class of Bayesian nonparametric estimates : I. Density estimates, *Annals of Statistics*, *12*(1), 351–357.

MacEachern, S., and P. Mueller (1998), Estimating mixture of Dirichlet process models, *Journal of Computational and Graphical Statistics*, *7*, 223—-238.

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

McLachlan, G., and K. Basford (1988), *Mixture models: Inference and Applications to Clustering*, Marcel-Dekker, New York.

McLachlan, G., and D. McGriffin (1994), On the role of finite mixture models in survival analysis, *Statistical Methods in Medical Research*, *3*, 211–226.

Mengersen, K. L., C. Robert, and M. Titterington (2011), *Mixtures: Estimation and Applications*, Wiley.

Miller, J., and M. Harrison (2014), Inconsistency of Pitman-Yor process mixtures for the number of components, *Journal of Machine Learning Research*, *15*, 3333–3370.

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

Muller, P., F. Quintana, A. Jara, and T. Hanson (2015), *Bayesian Nonparametric Data Analysis*, Springer.

Neal, R. (2000), Markov chain sampling methods for Dirichlet process mixture models, *Journal of Computational and Graphical Statistics*, *9*, 249—-265.

Neal, R. M., et al. (2011), Mcmc using hamiltonian dynamics, *Handbook of Markov Chain Monte Carlo*, *2*(11).

Newman, D., J. H. Lau, K. Grieser, and T. Baldwin (2010), Automatic evaluation of topic coherence, in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108, Association for Computational Linguistics.

Nguyen, X. (2013), Convergence of latent mixing measures in finite and infinite mixture models, *Annals of Statistics*, *4*(1), 370–400.

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

Nobile, A. (1994), Bayesian analysis of finite mixture distributions, Ph.D. thesis, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.

Nobile, A., and A. Fearnside (2007), Bayesian finite mixtures with an unknown number of components: The allocation sampler, *Statistics and Computing*, *17(2)*, 147–162.

Paisley, J. W., D. M. Blei, and M. I. Jordan (2014), Bayesian nonnegative matrix factorization with stochastic variational inference.

Pedregosa, F., et al. (2011), Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Pervez, A., Y. Mao, and D. Lee (2017), Learning deep movement primitives using convolutional neural networks, in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pp. 191–197, IEEE.

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

Pollard, D. (1982a), Quantization and the method of k-means, *IEEE Transactions on Information Theory*, *28*(2), 199–204.

Pollard, D. (1982b), A central limit theorem for k-means clustering, *The Annals of Probability*, *10*(4), 919–926.

Pritchard, J., M. Stephens, and P. Donnelly (2000), Inference of population structure using multilocus genotype data, *Genetics*, *155*, 945–959.

Richardson, S., and P. Green (1997), On Bayesian analysis of mixtures with an unknown number of components., *Journal of the Royal Statistical Society, B*, *59*, 731–792.

Robert, C. (1996), *Mixtures of distributions: inference and estimation. In Markov Chain Monte Carlo in Practice (W. Gilks, S. Richardson and D. Spiegelhalter, eds.).*

Rodriguez, A., D. Dunson, and A. E. Gelfand (2008), The nested Dirichlet process, *J. Amer. Statist. Assoc.*, *103*(483), 1131–1154.

Roeder, K., and L. Wasserman (1997), Practical bayesian density estimation using mixtures of normals, *J. Amer. Statist. Assoc.*, *92*, 894—-902.

Rokach, L., and O. Maimon (2005), Clustering Methods, in *Data Mining and Knowledge Discovery Handbook*, edited by O. Maimon and L. Rokach, pp. 321–352, Springer US, Boston, MA, doi:10.1007/0-387-25465-X_15.

Rousseau, J., and K. Mengersen (2011), Asymptotic behaviour of the posterior distribution in overfitted mixture models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 689–710.

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

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto (2004), *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*, John Wiley & Sons.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola (2008), *Global Sensitivity Analysis: The Primer*, John Wiley & Sons.

Schmidt, M. N., O. Winther, and L. K. Hansen (2009), Bayesian non-negative matrix factorization, in *International Conference on Independent Component Analysis and Signal Separation*, pp. 540–547, Springer.

Scricciolo, C. (2011), Posterior rates of convergence for dirichlet mixtures of exponential power densities, *Electron. J. Stat.*, *5*, 270—308.

Scricciolo, C. (2014), Adaptive Bayesian density estimation in l_p-metrics with pitman-yor or normalized inverse-gaussian process kernel mixtures, *Bayesian Analysis*, *9*, 457–520.

Scricciolo, C. (2017), Bayes and maximum likelihood for l1-wasserstein deconvolution of laplace mixtures, *Arxiv preprint.*

Sethuraman, J. (1994), A constructive definition of Dirichlet priors, *Statistica Sinica*, *4*, 639–650.

Shen, W., S. Tokdar, and S. Ghosal (2013), Adaptive Bayesian multivariate density estimation with Dirichlet mixtures, *Biometrika*, pp. 1–18.

Shen, X., and L. Wasserman (2001), Rates of convergence of posterior distributions, *Ann. Statist.*, *29*, 687–714.

Shen, X., and W. H. Wong (1994), Convergence rate of sieves estimates., *Annals of Statistics*, *22*, 580–615.

Sivaganesan, S. (2000), Global and Local Robustness Approaches: Uses and Limitations, in *Robust Bayesian Analysis*, edited by D. R. Insua and F. Ruggeri, Lecture Notes in Statistics, pp. 89–108, Springer, New York, NY, doi:10.1007/978-1-4612-1306-2_5.

Srivastava, A., and E. Klassen (2016), *Functional and Shape Data Analysis*, Springer Series in Statistics.

Stegmann, M. B., and D. D. Gomez (), A brief introduction to statistical shape analysis, accessed from `https://graphics.stanford.edu/courses/cs164-09-spring/Handouts/paper_shape_spaces_imm403.pdf`.

Stephens, M. (2000), Bayesian analysis of mixture models with an unknown number of components—An alternative to reversible jump methods, *Annals of Statistics*, *28*, 40–74.

S.Walker (2007), Sampling the dirichlet mixture model with slices, *Communications in Statistics - Simulation and Computation*, *36*.

Tang, J., Z. Meng, X. Nguyen, Q. Mei, and M. Zhang (2014), Understanding the limiting factors of topic modeling via posterior contraction analysis, in *Proceedings of the International Conference on Machine Learning*.

Taniguchi, T., S. Nagasaka, K. Hitomi, K. Takenaka, and T. Bando (2015), Unsupervised hierarchical modeling of driving behavior and prediction of contextual changing points, *IEEE Trans. Intell. Transport. Syst.*, *16*(4), 1746–1760.

Taniguchi, T., S. Nagasaka, K. Hitomi, N. P. Chandrasiri, T. Bando, and K. Takenaka (2016), Sequence prediction of driving behavior using double articulation analyzer, *IEEE Trans. Syst., Man, and Cyber.: Syst.*, *46*(9), 1300–1313.

Teh, Y., M. Jordan, M. Beal, and D. Blei (2006), Hierarchical Dirichlet processes, *J. Amer. Statist. Assoc.*, *101*, 1566–1581.

Tipping, M., and C. Bishop (1999), Probabilistic principal component analysis, *Journal of Royal Statistical Society: Series B*, *61*, 611–622.

Van de Geer, S. (1993), Hellinger consistency of certain nonparametric maximum likelihood estimators, *Ann. Statist.*, *21*, 14–44.

Villani, C. (2003), *Topics in Optimal Transportation*, American Mathematical Society.

Villani, C. (2008), *Optimal transport: Old and New*, Springer.

Walker, S., A. Lijoi, and I. Prunster (2007), On rates of convergence for posterior distributions in infinite-dimensional models, *Ann. Statist.*, *35*(2), 738–746.

Wallach, H. M., D. M. Mimno, and A. McCallum (2009), Rethinking lda: Why priors matter, in *Advances in neural information processing systems*, pp. 1973–1981.

Wang, C., and S. Mahadevan (2008), Manifold alignment using procrustes analysis, in *ICML*.

Wang, W., and D. Zhao (2017), Extracting traffic primitives directly from naturalistically logged data for self-driving applications, *IEEE Robotics and Automation Letters*.

Wang, W., J. Xi, and D. Zhao (2017), Driving style analysis using primitive driving patterns with bayesian nonparametric approaches, arXiv:1708.08986.

Wang, Y., and D. Blei (2019), Variational bayes under model misspecification, in *Neural Information Processing Systems*.

Wang, Y., and D. M. Blei (2018), Frequentist consistency of variational bayes, *Journal of the American Statistical Association*, pp. 1147–1161.

Wong, W. H., and X. Shen (1995), Probability inequalities for likelihood ratios and convergences of sieves mles, *Annals of Statistics*, *23*, 339–362.

Xie, F., and Y. Xu (2017), Bayesian repulsive Gaussian mixture model, *arXiv:1703.09061v2 [stat.ME].*

Xu, D., and Y. Tian (2015), A Comprehensive Survey of Clustering Algorithms, *Annals of Data Science*, *2*(2), 165–193, doi:10.1007/s40745-015-0040-1.

Xu, R., and D. C. Wunsch (2009), *Clustering*, IEEE Press Series on Computational Intelligence, Wiley ; IEEE Press, Hoboken, N.J. : Piscataway, NJ, oCLC: ocn216937130.

Yang, Y., D. Pati, and A. Bhattacharya (2017), -variational inference with statistical guarantees, *Arxiv preprint. arXiv:1710.03266.*

Yao, H., Y. Liu, Y. Wei, X. Tang, and Z. Li (2019), Learning from multiple cities: A meta-learning approach for spatial-temporal prediction, *The World Wide Web Conference on - WWW '19*, doi:10.1145/3308558.3313577.

Yurochkin, M., and X. Nguyen (2016), Geometric Dirichlet Means Algorithm for topic inference, in *Advances in Neural Information Processing Systems*, pp. 2505–2513.

Yurochkin, M., A. Guha, and X. Nguyen (2017), Conic scan and cover algorithms for nonparametric topic modeling, in *NIPS 31.*

Yurochkin, M., Z. Fan, A. Guha, P. Koutris, and X. Nguyen (2019), Scalable inference of topic evolution via models for latent geometric structures, in *Advances in Neural Information Processing Systems, 32.*

Yurochkin*, M., A. Guha*, Y. Sun, and X. Nguyen (2019), Dirichlet simplex nest and geometric inference, in *Proceedings of the 36th International Conference on Machine Learning*, vol. 97, pp. 7262–7271.

Zhang, C. (1990), Fourier methods for estimating mixing densities and distributions, *Annals of Statistics*, *18*(2), 806–831.

Zhang, W., and W. Wang (2019), Learning v2v interactive driving patterns at signalized intersections, *Transportation Research Part C: Emerging Technologies*, *108*, 151–166.

Zhang, W., W. Wang, and D. Zhao (2019), Multi-vehicle interaction scenarios generation with interpretable traffic primitives and gaussian process regression, *arXiv preprint arXiv:1910.03633.*

Zhu, H., J. G. Ibrahim, and N. Tang (2011), Bayesian influence analysis: A geometric approach, *Biometrika*, *98*(2), 307–323, doi:10.1093/biomet/asr009.

Zhu, J., S. Qin, W. Wang, and D. Zhao (2019), Probabilistic trajectory prediction for autonomous vehicles with attentive recurrent neural process.