

Bayesian Generative Modeling of Latent Subpopulations with Nonparametric Distributions

by

Yilei Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2026

Doctoral Committee:

Professor XuanLong Nguyen, Chair
Professor Veera Baladandayuthapani
Associate Professor Yang Chen
Assistant Professor Yun Wei

Yilei Zhang

yileizh@umich.edu

ORCID iD: 0009-0008-9817-1938

© Yilei Zhang 2026

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Long, for his generous guidance, encouragement, and support throughout my Ph.D. journey. There were many moments when I was uncertain about what I would be able to accomplish, but Long believed in me more than I believed in myself. He recognized the value of my work and encouraged me even when the project had not yet produced concrete results. Long has always been hopeful, optimistic, and full of energy and passion for both research and life. From him, I learned one of the most important lessons I will carry with me as a researcher: research is an exploration of the unknown, and one must learn to recognize the value of one's work from within. An internal sense of purpose, more than external recognition, is what sustains a researcher in the long run. The past six years have not always been smooth. I began my Ph.D. during the pandemic and encountered many unexpected challenges near the end of my program. During a time when I felt anxious about almost everything, I was especially grateful for the long conversations I had with Long. His wisdom, patience, and genuine support gave me the mental strength to keep going through the most difficult moments. I am also deeply grateful for his efforts to help me secure funding during my final year. Without his intellectual, emotional, and material support, this dissertation would not have been possible.

I would also like to thank Yun, my co-advisor and one of the main collaborators on my dissertation work. I still remember the time when he went through a book on Bayesian nonparametric theory with me, explaining the theorems and the ideas behind them one by one. He is an incredible teacher who immediately understands my confusion and explains things from my perspective. Without his patience and guidance, I would have struggled with those theorems and lemmas for much longer. Discussions with Yun also produced many of the major ideas in this dissertation. I am grateful as well for his generosity in sharing his knowledge of different researchers, ideas, and bodies of work, and for his helpful advice when I was concerned about my future career and job search.

I am also deeply grateful to my collaborator Aritra, who encouraged me greatly during the early years of my Ph.D. by sharing his own Ph.D. experience with me. I am grateful to both Yun and Aritra for listening to me during several long conversations when I was struggling with anxiety. They offered sincere advice and concrete support when I needed it most. I am also thankful for Aritra's generous efforts to help me find an internship position, which led to a wonderful internship experience in New Jersey and allowed me to gain valuable industry perspective.

I would like to thank my high-school friends, Sherry and Yuanhan. In 2022, I visited them in

Washington, D.C. during the cherry blossom season. It was the first time I had seen them since high school, and it became one of the brightest memories for me during the long pandemic years. They still remembered the foods I loved in high school and made hometown dishes for me that could not be found anywhere else in the United States.

I am also deeply grateful to the friends I came to know over the past few years through labor organizing, many of whose names I may not list publicly. We are living through a period of profound change, with uncertainties affecting each of us in different ways. As an international student, the friends I met through our proud union, GEO, showed me that there are so many people in this country who sincerely care about the rights and well-being of their community members across differences in ethnicity, culture, nationality, and educational background. They devoted their time and energy to making community support concrete and accessible. Knowing that I had a union standing behind me, and that these people would be there to help when needed, brought me great relief during an uncertain time. I am also grateful to all the volunteers from our organizing resource-sharing platform. Our Sunday meetings became a steady and meaningful part of my life. Having this group of people with whom I could reflect on the events of each week, exchange ideas, and work together gave me strength, support, and a deep sense of community throughout the past years.

I am also deeply grateful to my dear friends Xianni, Yirong, and Wayne. Our trip to the Upper Peninsula was one of my best memories from my Ph.D. years. We shared many joyful moments together in Ann Arbor, and their constant love and support made this city feel warm and home-like to me. I also want to thank all the friends with whom I played board games; those gatherings brought me much happiness and comfort.

Finally, I would like to thank all the colleagues and friends I have come to know over the past few years. In particular, I am grateful to the members of our research group—Dat, Sunrit, Vincenzo, Rayleigh, Bach, Yidan, An, Khoa, Ian, Thanh, and Leo—who made Ann Arbor feel like home. I greatly enjoyed our discussions, both academic and personal, and I am thankful for the community we shared. I would also like to thank my flatmate Junting, with whom I had the good fortune to live for four years.

Getting to know these wonderful people has been one of the most precious parts of my Ph.D. life. The love, solidarity, and support I received from them will stay with me, and I will carry them with me into the future.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	ii
LIST OF FIGURES	vi
LIST OF APPENDICES	viii
LIST OF ACRONYMS	ix
ABSTRACT	x

CHAPTER

1 Introduction	1
1.1 Overview	1
1.2 Parametric finite mixture models	6
1.2.1 Gaussian mixture models	7
1.2.2 EM algorithm for Gaussian mixture models	8
1.2.3 Bayesian Gaussian mixture models	9
1.3 Dirichlet Process	11
1.3.1 Different constructions of the Dirichlet Process	13
1.3.2 Dirichlet Process mixture models	15
2 A Bayesian Approach to Learning Mixtures of Nonparametric Components	17
2.1 Introduction	17
2.2 Preliminaries	21
2.3 Problem and modeling assumptions	22
2.4 Mixture of Dirichlet Process mixtures	24
2.5 Multivariate distributions	30
2.6 Identifiability	33
2.7 Posterior contraction rates	34
2.7.1 Posterior contraction of the overall mixture density	36
2.7.2 Posterior contraction of component densities	37
2.8 Applications	39
2.8.1 Disentangling astronomical sources	39
2.8.2 Analysis of oceanic whitetip shark acceleration data	41
2.9 Conclusions	43

3 Learning Nonparametric Components Beyond Convexly Separable Latent Support .	44
3.1 Introductions	44
3.2 Identifiability	46
3.3 Modeling latent supports beyond convex separability	50
3.3.1 The HDBSCAN algorithm	50
3.3.2 Latent support recovery	53
3.3.3 Learning component densities with MDPM	56
3.3.4 Additional simulations	58
3.4 Conclusion	62
4 Discussions and Future Directions	63
4.1 Discussions	63
4.1.1 What defines a component?	63
4.1.2 How to choose K ?	64
4.2 Future directions	65
4.2.1 Methodology	65
4.2.2 Theory	66
4.2.3 Applications	68
 APPENDICES	 69
 BIBLIOGRAPHY	 136

LIST OF FIGURES

FIGURE

2.1	Graphical model representation of the MDPM (2.6) for $K = 2$. The dashed edges among c_1, r_1, c_2, r_2 represent the repulsive prior on $(\mathbf{c}, \mathbf{r})$, which enforces disjoint intervals I_1 and I_2 . For $i = 1, 2$, the base measure $H_{i0}(u, \sigma)$ is supported on $I_i \times (0, \infty)$. The dashed edge between w_1 and w_2 indicates the truncated Beta prior (equivalently, the two-parameter truncated Dirichlet) on the mixture weights $\mathbf{w} = (w_1, w_2)$. Finally, the mixture weights $\mathbf{w} = (w_1, w_2)$ combine the component distributions into the overall mixture $F = w_1 G_1 + w_2 G_2$	27
2.2	Example 1: a three-component mixture with the separation condition imposed on the location parameter u ; f_1 is generated from a random combination of Hermite functions; f_2 is a skewed exponential-power distribution, and f_3 is a Laplace distribution. Example 2: a two component mixture with the separation condition imposed on the scale parameter σ ; Both component densities are generated from random combinations of Hermite functions.	29
2.3	A bivariate two-component mixture model. Each component is a Gaussian mixture located on a circle with random covariance matrices. The left panel displays the true density, while the middle and right panels show the pointwise posterior means of the fitted component densities, weighted by their respective posterior mixture weights. . . .	32
2.4	Density estimation contours from MDPM, KDE, and mixture of King's profiles, respectively.	40
2.5	Comparison of CDFs.	40
2.6	Left, middle, and right panels show the IODBA histogram with the estimated weighted emission densities from MDPM, adSP (Chen and Finkenstädt, 2024), and fpSP (Lan-grock et al., 2015), respectively.	42
3.1	Half-moon example. The left panel (a) shows a 3D visualization of the true density. The middle panel (b) shows 10,000 samples drawn from the true density, colored by their true component labels. The right panel (c) shows the same samples, colored by the HDBSCAN cluster labels.	54
3.2	Covering squares before and after trimming.	56
3.3	Half-moon example. The first row shows the true density contours for the overall density, the first component density f_1 , and the second component density f_2 . The second row shows the corresponding fitted density contours, given by the posterior means. The true component weights and the 95% credible intervals for the posterior weights are displayed in the upper-right corners of the corresponding panels.	58

3.4	Concentric circle example. The true distribution in this example is a mixture of three components. Panel (a) shows the samples colored by their true component labels. Panel (b) shows the same samples colored by the HDBSCAN labels, along with the longest MST path and the covering rectangles for each cluster. Panels (c) and (d) display 3D visualizations of the true and fitted (posterior mean) densities, respectively. Panels (e) and (f) show the corresponding true and fitted (posterior mean) density contours, using the same density scale.	59
3.5	Concentric circle example. The first row shows the true density contours of the three components, and the second row shows the corresponding fitted (posterior mean) density contours. The true component weights and the 95% credible intervals for the posterior component weights are shown in the upper-right corners.	60
3.6	Pinwheel example. Panel (a) is the true latent manifold curves that generate the dataset. Panel (b) visualizes the 20,000 samples drawn from the true density, colored by their true component labels. Panel (c) shows the HDBSCAN clustering results.	60
3.7	The fitted longest MST path and covering cubes for each component is shown in corresponding colors.	61
3.8	Panel (a) is the isosurface plot for the true density, and Panel (b) is the isosurface plot for the fitted density.	62

LIST OF APPENDICES

APPENDIX

A	Appendix of Chapter 2	69
A.1	Algorithm	69
A.2	Simulations	72
A.2.1	Figure (2.2a) and (2.2b)	73
A.2.2	Figure (2.2c) and (2.2d)	73
A.2.3	Figure 2.3	74
A.3	Proofs for Section 2.6: Theorem 2.6.1	75
A.4	Proofs for Section 2.7.1: Theorem 2.7.1	76
A.5	Proofs for Section 2.7.2: Construction of $\widehat{f}_i$	80
A.5.1	Hermite functions	80
A.5.2	Construction of $\widehat{f}_i$	82
A.6	Proofs for Section 2.7.2: Upper bound on $\ \widehat{f}_i - f_i\ _1$	84
A.6.1	Technical lemmas	85
A.6.2	Completing the proof of Lemma A.6.1	102
A.6.3	Auxiliary lemmas	104
A.7	Proofs for Section 2.7.2: Concentration inequality for $\ \widehat{f}_i - f_i\ _1$	113
A.8	Proofs for Section 2.7.2: Lemma 2.7.3 and Theorem 2.7.2	122
B	Appendix of Chapter 3	129
B.1	Algorithms	129

LIST OF ACRONYMS

MCMC Markov Chain Monte Carlo

EM Expectation–Maximization

DP Dirichlet Process

DPM Dirichlet Process Mixtures

MDPM Mixture of Dirichlet Process Mixtures

KDE Kernel Density Estimation

HMM Hidden Markov Models

CDF Cumulative Distribution Functions

GAN Generative Adversarial Networks

VAE Variational Autoencoders

LLMs Large Language Models

ABSTRACT

Across many scientific domains, researchers increasingly collect large heterogeneous datasets that contain multiple meaningful subpopulations. In many cases, however, the subpopulation labels are unavailable. These subpopulations may be related to one another in complex ways, and each may itself exhibit a rich internal structure. These tasks typically require not only assigning observations to subpopulations, but also characterizing the distributional structure of each subpopulation. Accurately identifying and characterizing such latent subpopulations is therefore a fundamental yet challenging problem. Mixture models provide a natural framework for modeling subpopulation distributions. However, most existing work assumes that the component distributions belong to specified parametric families, which are almost always misspecified in practice. Capturing complex subpopulation structures therefore requires extending the mixture-modeling framework to accommodate nonparametric component distributions. However, this extension immediately raises fundamental challenges of identifiability and inference. Since the subpopulations are latent and only the overall population distribution is observed, it is unclear what should count as a distinct subpopulation. If each component distribution is allowed to be highly flexible, it is also unclear whether the components can be separated, particularly in regions where their densities overlap. Even when such separation is theoretically possible, a further challenge is to develop effective inference methods that can be practically implemented to estimate the latent subpopulations. In this dissertation, we address these theoretical and methodological challenges within a systematic framework.

First, we adopt a Bayesian nonparametric approach and develop a unified framework that uses mixtures of Dirichlet process mixtures (MDPMs) to model two classes of nonparametric mixture structures: one in which the high-density regions of different components are spatially differentiated, and another in which components may fully overlap spatially but are distinguished by contrasting density levels. We devise scalable algorithms for implementing the proposed framework and demonstrate their performance through extensive simulation studies and several real-data applications in both univariate and multivariate settings. The results show that, under mild conditions, the component distributions in both types of mixture structures can be accurately recovered.

Second, building on this framework, we extend the proposed approach to settings in which the

high-density regions associated with different components are spatially differentiated but cannot be contained in disjoint convex sets, a challenge that is especially relevant in multivariate problems. To handle mixture components with complex density-contour geometry, we propose a method that approximates the high-density regions by a union of hypercubes. We then construct MDPMs over these hypercube coverings, allowing the model to learn component distributions with complex latent-support geometries. Simulation studies demonstrate strong performance across a range of settings.

Third, we develop theoretical support for the proposed statistical framework, by establishing identifiability conditions for the first class of mixture structures in both univariate and multivariate settings. We then proceed to derive posterior contraction rates under the proposed MDPM framework. These results show that, under appropriate conditions, using a mixture of Dirichlet process mixtures does not reduce the efficiency of learning the overall population density relative to using a single Dirichlet process mixture. Most remarkably, we show that the latent nonparametric component distributions can be learned at a nearly polynomial rate, which is much more efficient than the typically slow rates of nonparametric deconvolution.

CHAPTER 1

Introduction

1.1 Overview

Across many scientific domains, researchers increasingly collect large heterogeneous datasets that contain multiple meaningful subpopulations. In many cases, however, the subpopulation labels are unavailable, either because they are intrinsically unobservable or because obtaining reliable subpopulation labels is prohibitively expensive. These subpopulations may be related to one another in complex ways, and each may itself exhibit a rich internal structure. Accurately identifying and characterizing such latent subpopulations is therefore a fundamental yet challenging problem. For example, in high-energy astrophysics, modern X-ray observatories record data as lists of individual photon events. These photons may originate from multiple sources, including point sources such as X-ray binaries, as well as diffuse or extended emission from supernova remnants, galaxy clusters, or merging galaxies. Since the source identity of each photon is not directly observed, photons generated by a given astrophysical source can be viewed as a latent subpopulation. The spatial and spectral distributions of these photons trace complex structures associated with astrophysical sources, and accurately recovering these latent subpopulations is essential for understanding the physics of many astronomical objects (Zezas et al., 2002, 2006; Jones et al., 2015; Wilkes and Tucker, 2019; Meyer et al., 2021; Fan et al., 2023; McKeough et al., 2026). In biology, single-cell RNA sequencing measures gene expression in individual cells, producing high-dimensional molecular profiles over thousands of genes. These profiles provide a detailed view of cellular states,

including differences among cell types, intermediate or transitional states between cell types, rare cell types and state changes induced by different experimental conditions. Identifying biologically meaningful subpopulations from these gene-expression profiles, and determining the genes and pathways that distinguish them, are therefore central tasks in single-cell analysis (Prabhakaran et al., 2016; Sun et al., 2019; Lopez-Delisle and Delisle, 2022; Lopez et al., 2023; Boyeau et al., 2023).

These tasks, arising across different scientific domains, require more than assigning each observation to a subpopulation; they also require characterizing the distributional structure of each subpopulation. Standard clustering methods that primarily partition the data space and assign observations to clusters may therefore be insufficient. In contrast, generative models provide probabilistic descriptions of the underlying subpopulations, thereby enabling a more principled characterization of subpopulation structure, variability, and uncertainty.

It’s worth noting that over the past decade, deep generative models have achieved remarkable success in modeling complex data such as text and images. Representative examples include generative adversarial networks (GANs) (Goodfellow et al., 2014), variational autoencoders (VAEs) (Kingma and Welling, 2013), normalizing-flow models (Dinh et al., 2014, 2016; Kingma and Dhariwal, 2018), diffusion models (Ho et al., 2020), and Transformer-based large language models (LLMs) (Vaswani et al., 2017; Radford et al., 2018, 2019; Brown et al., 2020). Broadly speaking, many of these deep generative models share a common architectural strategy: they learn a transformation from a simple reference distribution to a complex target distribution, with the transformation or its reverse dynamics parameterized by deep neural networks. This perspective includes normalizing flows, which learn an explicit invertible map; VAEs, which learn to generate data from latent variables while also learning an approximate encoder that maps observed data back to the latent space; GANs, which learn a pushforward map from a simple noise distribution to the data distribution; and diffusion models, which learn a reverse denoising process from noise to data. Autoregressive Transformer models instead factorize the joint distribution into a sequence of conditional distributions and use neural networks to predict the next-token probability distribu-

tion at each step. The expressive power of deep neural networks enables these models to capture remarkably complex distributions over high-dimensional data. However, despite their flexibility, the latent representations learned by these models are not in general identifiable, and the black-box nature of deep neural networks makes it difficult to impose domain-informed structure on latent subpopulations or to ensure that the learned latent representations are scientifically meaningful and interpretable (Hyvärinen and Pajunen, 1999; Locatello et al., 2019; Schölkopf et al., 2021; Roeder et al., 2021; Reizinger et al., 2024). Moreover, although uncertainty quantification can be incorporated into deep generative models, it is often not available in a calibrated or interpretable form. This limits their use in scientific applications where reliable uncertainty assessment is essential.

To develop interpretable generative models for latent subpopulations, it is natural to return to the mixture-model framework, in which the population distribution is represented as a weighted combination of subpopulation distributions (components). In classical mixture modeling, each component is typically assumed to belong to a parametric family, such as the Gaussian family. Gaussian mixture models are among the most widely used examples of this framework (Genovese and Wasserman, 2000; Ghosal et al., 1999; Xu and Jordan, 1996; Redner and Walker, 1984; Wu and Yang, 2020). More broadly, mixture models with parametric components have been studied extensively from both theoretical and computational perspectives. However, in practice, as illustrated by the examples above, meaningful subpopulations often exhibit complex distributional structures that cannot be adequately captured by standard parametric families. This limitation is especially important in mixture modeling: when the component family is misspecified—that is, when the true component distribution does not belong to the assumed parametric family—the fitted mixing measure may adapt to the misspecification by introducing artificial mixture components. Consequently, the learned mixing measure may become difficult to interpret as the true latent subpopulation structure. In practice, parametric component families are almost always misspecified, creating a substantial gap between classical parametric mixture models and the goal of recovering interpretable latent structure and subpopulation distributions. This motivates an extension of the mixture-modeling framework in which component distributions are modeled nonparametrically.

However, extending mixture models to nonparametric component families immediately raises fundamental identifiability and inference challenges. Since the subpopulations are latent and only the overall population distribution is observed, it is not obvious what should be identified as a distinct subpopulation. Moreover, if each component distribution is allowed to be highly flexible, one must ask whether the components can still be separated, especially in regions where their densities overlap. Even when such separation is theoretically possible, a further challenge is whether there exist effective inference methods that can be practically implemented to estimate the latent subpopulations. In this dissertation, we address these theoretical and methodological challenges in a systematic framework. We mainly focus on two classes of nonparametric mixture models:

- The first class is characterized by the assumption that each component has its probability mass spatially concentrated around a connected region, with distinct components concentrated in different regions. We allow overlap between component tails. Apart from this mild spatial separation condition, we do not impose any additional assumptions on the component distributions.
- The second class of models are spike-and-slab-type mixtures, in which one component exhibits one or more high-density “spikes”, while the other component has a flat, relatively low-density distribution over the entire support. The two components may share fully overlapping support.

Methodologically, mixture models are commonly fitted using either frequentist or Bayesian approaches. In parametric mixture models, frequentist methods typically estimate the unknown parameters by optimizing an objective function, such as the likelihood or a penalized likelihood function. A standard computational tool for this purpose is the expectation–maximization (EM) algorithm and its variants. Bayesian approaches, in contrast, place prior distributions on the model parameters and perform inference based on their posterior distributions given the observed data. This posterior formulation naturally provides a framework for quantifying uncertainty in model

parameters and latent allocations. Posterior inference is typically performed using Markov chain Monte Carlo (MCMC). In this dissertation, we adopt a Bayesian nonparametric modeling approach and develop a unified framework that uses mixtures of Dirichlet Process mixtures with Gaussian kernels to model both types of nonparametric mixture structures discussed above. We develop methods for the first type of mixture structure in both univariate and multivariate settings, and for the second type in the univariate setting. We also develop scalable computational algorithms for implementing the proposed methods and demonstrate their performance through extensive simulation studies and several real-data applications. The results show that, under mild conditions, the component distributions in both types of mixture structures can be accurately recovered.

On the theoretical side, we establish identifiability conditions for the first type of mixture structure using the notion of connected sets in the support of the mixing measure. These conditions are developed for both univariate and multivariate settings. We further derive posterior contraction rates for the estimated component densities under appropriate conditions in the univariate setting. The resulting contraction rate is nearly polynomial for nonparametric component-density estimation in mixture models under the same assumptions. These results provide theoretical support for the efficient recovery of complex component distributions under the proposed framework.

The rest of the dissertation is organized as follows.

In Chapter 1, we provide a brief review of the foundational concepts underlying finite mixture models, the Dirichlet process, and Dirichlet process mixture models.

In Chapter 2, we introduce the mixture-of-Dirichlet-process-mixtures framework and develop efficient algorithms for implementing it in both univariate and multivariate settings. We also establish theoretical results on the identifiability of nonparametric mixture models and posterior contraction rates under the proposed framework. These results show that, under appropriate conditions, using a mixture of Dirichlet Process mixtures does not reduce the efficiency of learning the overall population density relative to using a single Dirichlet Process mixture, while also allowing the nonparametric component distributions to be learned simultaneously at a nearly polynomial rate.

In Chapter 3, we extend the proposed approach to settings in which the high-density regions associated with different components cannot be contained in disjoint convex sets, a challenge that is especially relevant in multivariate problems. We first establish identifiability conditions for multivariate mixture models in this more general setting. We then develop a method for approximating the geometry of these high-density regions by a union of hypercubes. Building on these hypercube coverings, we construct mixtures of Dirichlet process mixtures that can learn component distributions with complex latent-support geometries.

In Chapter 4, we discuss several conceptual and practical questions that arise in mixture models with nonparametric components. We then summarize the main strengths of the proposed MDPM framework. Finally, we outline several directions for future work, including methodological extensions, further theoretical development, and potential applications to broader scientific domains.

1.2 Parametric finite mixture models

Finite mixture models are the basic tools for modeling heterogeneous data populations. The central idea is that the observed data are generated from several latent subpopulations, each represented by a probability distribution. Let $X_1, \dots, X_n \in \mathbb{R}^d$ denote independent observations. A finite mixture model with K components has density

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k f_k(\mathbf{x} \mid \boldsymbol{\theta}_k),$$

where $\pi_k \geq 0$ are the mixture weights satisfying $\sum_{k=1}^K \pi_k = 1$, and $f_k(\cdot \mid \boldsymbol{\theta}_k)$ is the density of the k th component. Equivalently, the model can be represented through latent allocation variables

$$Z_i \mid \boldsymbol{\pi} \sim \text{Categorical}(\pi_1, \dots, \pi_K), \quad X_i \mid Z_i = k \sim f_k(\cdot \mid \boldsymbol{\theta}_k),$$

where Z_i indicates the unobserved component membership of observation X_i .

Finite mixture models are widely used for density estimation, clustering, classification, and

modeling population heterogeneity. Unlike hard clustering methods, such as K -means, mixture models naturally allow each observation to have probabilistic membership across multiple components. This soft-assignment interpretation is particularly useful when clusters overlap or when the goal is to estimate the underlying density rather than only assign labels.

1.2.1 Gaussian mixture models

When the component densities are Gaussian, the resulting mixture model is called a Gaussian mixture model (GMM). For observations $X_i \in \mathbb{R}^d$, the model is

$$p(X_i | \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{k=1}^K \pi_k N_d(X_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $N_d(\cdot | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the d -dimensional Gaussian density with mean vector $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and covariance matrix $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$. Thus, each observation is generated by first drawing a latent component label Z_i and then drawing X_i from the Gaussian distribution associated with that component.

The parameters of a GMM are

$$\boldsymbol{\theta} = \{\pi_1, \dots, \pi_K, \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}.$$

The observed-data log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^n \log p(X_i | \boldsymbol{\theta}) = \sum_{i=1}^n \log \left\{ \sum_{k=1}^K \pi_k N_d(X_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

The presence of the logarithm of a sum makes direct maximum likelihood estimation difficult. If the latent labels Z_i were observed, estimation would be straightforward: each component could be estimated from the observations assigned to it. In practice, however, the labels are unobserved, and inference must account for uncertainty in component membership.

The posterior probability that observation i belongs to component k is given by:

$$r_{ik} = \Pr(Z_i = k \mid X_i, \boldsymbol{\theta}) = \frac{\pi_k N_d(X_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{\ell=1}^K \pi_\ell N_d(X_i \mid \boldsymbol{\mu}_\ell, \boldsymbol{\Sigma}_\ell)}.$$

These probabilities can be interpreted as soft cluster assignments: an observation located between two overlapping Gaussian components may have non-negligible membership probabilities for both components, rather than being assigned deterministically to only one cluster.

1.2.2 EM algorithm for Gaussian mixture models

The expectation-maximization (EM) algorithm is the standard maximum likelihood method for Gaussian mixture models. It is designed for latent-variable models in which the complete-data likelihood would be easy to optimize if the latent variables were observed. For GMMs, the latent variables are the component labels $Z_1, \dots, Z_n$.

The complete-data log-likelihood is

$$\ell_c(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{k=1}^K \mathbf{1}\{Z_i = k\} [\log \pi_k + \log N_d(X_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)].$$

If Z_i were known, the indicator $\mathbf{1}\{Z_i = k\}$ would determine which observations are assigned to each component. The EM algorithm replaces these unknown indicators by their conditional expectations under the current parameter values.

Starting from an initial value $\boldsymbol{\theta}^{(0)}$, EM alternates between the following two steps.

E-step. Given the current parameter estimate $\boldsymbol{\theta}^{(t)}$, compute the posteriors

$$r_{ik}^{(t)} = \Pr(Z_i = k \mid X_i, \boldsymbol{\theta}^{(t)}) = \frac{\pi_k^{(t)} N_d(X_i \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{\ell=1}^K \pi_\ell^{(t)} N_d(X_i \mid \boldsymbol{\mu}_\ell^{(t)}, \boldsymbol{\Sigma}_\ell^{(t)})}.$$

M-step. Update the parameters by maximizing the expected complete-data log-likelihood. Let

$$N_k^{(t)} = \sum_{i=1}^n r_{ik}^{(t)}.$$

Then the standard GMM updates are

$$\pi_k^{(t+1)} = \frac{N_k^{(t)}}{n},$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^n r_{ik}^{(t)} X_i}{N_k^{(t)}},$$

and

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^n r_{ik}^{(t)} (X_i - \boldsymbol{\mu}_k^{(t+1)})(X_i - \boldsymbol{\mu}_k^{(t+1)})^\top}{N_k^{(t)}}.$$

The EM algorithm can be understood as repeatedly constructing and maximizing a lower bound to the observed-data log-likelihood. For Gaussian mixtures, EM has an intuitive clustering interpretation. The E-step computes soft cluster assignments, while the M-step updates the mixture weights, means, and covariances using these soft assignments. Because the likelihood surface of mixture models is generally non-convex, EM may converge to a local maximum. In practice, it is common to run EM from multiple random initializations and select the solution with the largest observed-data log-likelihood.

1.2.3 Bayesian Gaussian mixture models

A Bayesian Gaussian mixture model places prior distributions on the mixture weights and component-specific parameters. The latent-variable formulation remains

$$Z_i \mid \boldsymbol{\pi} \sim \text{Categorical}(\pi_1, \dots, \pi_K), \quad X_i \mid Z_i = k, \boldsymbol{\mu}, \boldsymbol{\Sigma} \sim N_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

but the parameters are now treated as random variables. A common prior specification is

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K),$$

together with component-specific priors such as

$$\boldsymbol{\mu}_k \sim N_d(\mathbf{m}_0, \boldsymbol{\Lambda}_0), \quad \boldsymbol{\Sigma}_k \sim \text{Inverse-Wishart}(\mathbf{S}_0, \nu_0), \quad k = 1, \dots, K.$$

Here, $\mathbf{m}_0$, $\boldsymbol{\Lambda}_0$, $\mathbf{S}_0$ and ν_0 are the hyperparameters. Other conjugate or semi-conjugate choices are also possible, depending on whether the covariance matrices are known, diagonal, shared across components, or fully component-specific.

The Bayesian approach has several advantages. First, it quantifies uncertainty about the mixture weights, component means, covariance matrices, and latent allocations. Second, prior distributions can regularize estimation, which is helpful when some mixture components contain few observations or when covariance matrices are difficult to estimate. Third, Bayesian inference can be extended naturally to hierarchical mixture models and models with uncertainty about the number of components.

Posterior inference for Bayesian GMMs is commonly performed using Markov chain Monte Carlo (MCMC), especially Gibbs sampling when conjugate priors are available. A typical Gibbs sampler alternates among the following updates:

1. Sample each latent label Z_i from its full conditional distribution:

$$\Pr(Z_i = k \mid X_i, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \propto \pi_k N_d(X_i \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad k = 1, \dots, K.$$

After normalization, these probabilities have the same form as the EM responsibilities.

2. Sample the mixture weights from their full conditional distribution. If $\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K)$ and

$$n_k = \sum_{i=1}^n \mathbf{1}\{Z_i = k\},$$

then

$$\boldsymbol{\pi} \mid \mathbf{Z} \sim \text{Dirichlet}(\alpha_1 + n_1, \dots, \alpha_K + n_K).$$

3. Sample the component means and covariance matrices from their full conditional distributions. Under semi-conjugate Gaussian–inverse-Wishart priors, these full conditionals are available in closed form and depend only on the observations currently assigned to each component.

For Bayesian mixture models, the Gibbs sampler can be viewed as a stochastic counterpart to the EM algorithm, replacing deterministic updates of latent assignments and parameters with draws from their conditional posterior distributions. EM produces a point estimate, typically a local maximum likelihood estimate. In contrast, MCMC generates samples from the posterior distribution, allowing posterior summaries such as credible intervals for model parameters and uncertainty measures for cluster assignments.

1.3 Dirichlet Process

The Dirichlet process is one of the fundamental priors in Bayesian nonparametrics. It defines a prior over probability measures. To motivate the Dirichlet process, it is helpful to begin with the finite-dimensional Dirichlet distribution.

Let $(A_1, \dots, A_K)$ be a finite measurable partition of a sample space Θ . A probability measure G assigns masses

$$(G(A_1), \dots, G(A_K))$$

to these sets, where the entries are nonnegative and sum to one. In a finite Bayesian GMM, uncertainty about the mixture weights can be modeled by a Dirichlet distribution,

$$(p_1, \dots, p_K) \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_K).$$

The Dirichlet distribution is conjugate to the multinomial distribution and is therefore a natural prior for unknown probability vectors. If n_j observations fall in category j , then the posterior

distribution is

$$(p_1, \dots, p_K) \mid X_{1:n} \sim \text{Dirichlet}(\alpha_1 + n_1, \dots, \alpha_K + n_K).$$

This finite-dimensional conjugacy provides the basic intuition behind the Dirichlet process.

A Dirichlet process extends this idea from a finite probability vector to a random probability measure. Let $\alpha > 0$ be a concentration parameter and let G_0 be a probability measure on Θ , called the base measure. A random probability measure G is said to follow a Dirichlet process with parameters (α, G_0) , written

$$G \sim \text{DP}(\alpha, G_0),$$

if for every finite measurable partition $(A_1, \dots, A_K)$ of Θ ,

$$(G(A_1), \dots, G(A_K)) \sim \text{Dirichlet}(\alpha G_0(A_1), \dots, \alpha G_0(A_K)).$$

Thus, the Dirichlet process is named after the fact that the probabilities assigned by G to any finite measurable partition follow a Dirichlet distribution.

The base measure G_0 represents the prior mean of the random probability measure, since

$$\mathbb{E}\{G(A)\} = G_0(A)$$

for any measurable set A . The concentration parameter α controls how tightly G concentrates around G_0 . Larger values of α make draws from the Dirichlet process closer to G_0 , while smaller values of α produce more concentrated random measures with fewer dominant atoms. A key feature of the Dirichlet process is that, even when G_0 is continuous, a draw $G \sim \text{DP}(\alpha, G_0)$ is almost surely discrete. This discreteness makes the Dirichlet process particularly useful for clustering and mixture modeling, because repeated draws from G can take identical values with positive probability.

1.3.1 Different constructions of the Dirichlet Process

Several equivalent constructions of the Dirichlet process provide different perspectives on its behavior and lead to different computational algorithms. Two of the most important are the Pólya urn, or Chinese restaurant process representation and the stick-breaking representation.

First, consider the predictive distribution induced by a Dirichlet process. Suppose

$$G \sim \text{DP}(\alpha, G_0), \quad \theta_i \mid G \sim G, \quad i = 1, \dots, n.$$

After integrating out G , the conditional distribution of a new draw θ_{n+1} given previous draws $\theta_1, \dots, \theta_n$ is

$$\theta_{n+1} \mid \theta_{1:n} \sim \frac{\alpha}{\alpha + n} G_0 + \frac{1}{\alpha + n} \sum_{i=1}^n \delta_{\theta_i}.$$

Equivalently, if $\phi_1, \dots, \phi_K$ are the distinct values among $\theta_1, \dots, \theta_n$ and n_k is the number of observations equal to ϕ_k , then

$$\theta_{n+1} \mid \theta_{1:n} \sim \frac{\alpha}{\alpha + n} G_0 + \sum_{k=1}^K \frac{n_k}{\alpha + n} \delta_{\phi_k}.$$

This expression shows the self-reinforcing, or “rich-get-richer,” property of the Dirichlet process: a new observation either joins an existing atom with probability proportional to its current frequency, or creates a new atom drawn from G_0 with probability proportional to α . The induced random partition is commonly described by the Chinese restaurant process. In this metaphor, observations are customers, clusters are tables, and a new customer either joins an occupied table with probability proportional to the number of customers already seated there or starts a new table with probability proportional to α .

A second important construction is the stick-breaking representation. Let

$$V_k \sim \text{Beta}(1, \alpha), \quad \phi_k \sim G_0, \quad k = 1, 2, \dots,$$

independently, and define

$$\pi_1 = V_1, \quad \pi_k = V_k \prod_{\ell < k} (1 - V_\ell), \quad k \geq 2.$$

Then

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

is a draw from $\text{DP}(\alpha, G_0)$. The name “stick-breaking” comes from the interpretation of starting with a unit-length stick, breaking off a random proportion V_1 to form π_1 , and then recursively breaking the remaining stick to form $\pi_2, \pi_3, \dots$. This construction makes explicit the almost-sure discreteness of draws from a Dirichlet process.

The stick-breaking representation is particularly useful for algorithms that explicitly represent the random measure. A common approach is to truncate the infinite sum at a sufficiently large level K and approximate

$$G \approx \sum_{k=1}^K \pi_k \delta_{\phi_k},$$

often setting the final break to absorb the remaining mass. This leads to blocked Gibbs samplers and variational inference algorithms ([Ishwaran and James, 2001](#); [Blei and Jordan, 2006](#)). However, fixed truncation can introduce approximation error. Slice sampling methods avoid a fixed truncation by introducing auxiliary slice variables so that only finitely many atoms need to be represented at each Markov chain Monte Carlo iteration, while still targeting the infinite-dimensional model.

The Pólya urn and stick-breaking constructions therefore emphasize complementary aspects of the same prior. The Pólya urn representation highlights the exchangeable random partition and is natural for collapsed clustering algorithms. The stick-breaking representation highlights the random discrete measure and is natural for truncation-based, blocked, variational, and slice-sampling algorithms.

1.3.2 Dirichlet Process mixture models

Although draws from a Dirichlet process are almost surely discrete, the Dirichlet process can be used to construct flexible priors on continuous densities by mixing a smooth kernel over a random mixing distribution. This leads to the Dirichlet process mixture model. Let $K(x | \theta)$ be a parametric kernel indexed by $\theta \in \Theta$. A Dirichlet process mixture model can be written hierarchically as

$$G \sim \text{DP}(\alpha, G_0),$$

$$\theta_i | G \sim G, \quad X_i | \theta_i \sim K(\cdot | \theta_i), \quad i = 1, \dots, n.$$

Given G , the density of X_i is

$$f_G(X_i) = \int K(X_i | \theta) dG(\theta).$$

Since G is almost surely discrete under the Dirichlet process prior, this conditional density can be written as the infinite mixture

$$f_G(X_i) = \sum_{k=1}^{\infty} \pi_k K(X_i | \phi_k),$$

where $(\pi_k, \phi_k)_{k \geq 1}$ are obtained from the stick-breaking construction. Thus, the Dirichlet process mixture model can be viewed as an infinite mixture model in which the number of occupied components grows with the data rather than being fixed in advance.

This model is closely related to finite mixture models. In a finite mixture model with K components, one may take

$$\boldsymbol{\pi} \sim \text{Dirichlet}(\alpha/K, \dots, \alpha/K), \quad \theta_k \stackrel{\text{iid}}{\sim} G_0, \quad k = 1, \dots, K, \quad Z_i | \boldsymbol{\pi} \sim \text{Categorical}(\boldsymbol{\pi}),$$

and

$$X_i | Z_i, \theta_{1:K} \sim K(\cdot | \theta_{Z_i}).$$

As K becomes large, this finite mixture construction converges conceptually to a Dirichlet process mixture model.

In Dirichlet process mixture models, one can integrate out the random measure G and sample cluster assignments directly using the Chinese restaurant process predictive probabilities. When conjugacy is available between the kernel and the base measure, component parameters can also be marginalized analytically, resulting in efficient collapsed Gibbs samplers. More generally, auxiliary-variable methods can be used for nonconjugate models.

Dirichlet process mixture models are widely used for density estimation and clustering. While the Dirichlet process itself defines a prior over discrete random measures, a Dirichlet process mixture model induces clustering through the latent component parameters of the mixture; these clusters are therefore in the parameter space, not necessarily well-separated clusters in the observed data space. Multiple mixture components may jointly approximate a single non-Gaussian density feature, and a single visually apparent group may be represented by several latent atoms. This flexibility is precisely what makes Dirichlet process mixtures powerful for density estimation, but it also means that care is needed when interpreting mixture components as scientifically meaningful subpopulations.

CHAPTER 2

A Bayesian Approach to Learning Mixtures of Nonparametric Components

2.1 Introduction

Large heterogeneous datasets often contain multiple underlying subpopulations of interest. Accurately identifying and characterizing them is a fundamental yet difficult task. In practice, these subpopulation labels are typically unavailable—either because they are costly to obtain or because they are inherently unobservable—so the task falls within the domain of unsupervised learning. Clustering methods that focus on identifying subpopulations typically learn a partition of the data domain, aiming to group samples from the same subpopulation into the same cluster. However, they usually do not provide a distributional characterization for each subpopulation. Examples include K-means clustering ([McQueen, 1967](#); [Lloyd, 1982](#)), hierarchical clustering ([Hartigan, 1981](#); [Thomann et al., 2015](#)), mode clustering ([Chacón and Duong, 2013](#); [Chen et al., 2016](#)), spectral clustering ([Schiebinger et al., 2015](#)), and density-based clustering ([Rinaldo and Wasserman, 2010](#)). In comparison, model-based clustering explicitly characterizes the underlying distribution of each subpopulation and represents the overall population as a mixture—a weighted sum—of these subpopulation distributions (components), thereby yielding a generative probabilistic model for each subpopulation. Typically, each component of a mixture model is assumed to be from a parametric family, and a popular example is the Gaussian mixture model (GMM). When the parametric family satisfies identifiability conditions and the model is well-specified, meaning that the true subpop-

ulation distributions indeed belong to the specified family, the latent structure can be consistently recovered and the convergence rates of the latent mixing measure have been established in a series of work (Chen, 1995; Nguyen, 2013; Ho and Nguyen, 2016; Heinrich and Kahn, 2018; Wei and Nguyen, 2022; Wei et al., 2025). However, if the model is misspecified, the relationship between the learned and true mixing measures becomes elusive (Guha et al., 2021; Miller and Dunson, 2019), and the learned mixing measure usually fails to convey meaningful information about the underlying latent structure. In practice, the parametric families are almost always misspecified, which therefore leaves a considerable gap in the practical applications of parametric mixture models.

A variety of mixture models have been proposed to model each subpopulation with more flexible distribution families. Examples include the t-distribution (Peel and McLachlan, 2000; Stephens, 2000; Andrews and McNicholas, 2012), skew-t distribution (Lin et al., 2007), skewed power-exponential distribution (Dang et al., 2023), shifted asymmetric Laplace distribution (Franczak et al., 2013), and normal inverse Gaussian distribution (Karlis and Santourian, 2009). Levine et al. (2011) assumes each component density factorizes as the product of its marginal densities, and Rodríguez and Walker (2014) devises a family of nonparametric unimodal distributions. See also Xu et al. (2025) and the references therein for nonparametric clustering. These models enrich the component family, enabling it to accommodate heavy-tailed behavior, asymmetry, and non-elliptical structures that Gaussian mixture models cannot capture. However, each of these distribution families is tailored to capture only certain types of patterns, such as heavy tails or skewness; no single parametric family is sufficiently flexible to accommodate all of the above patterns. Moreover, these approaches require all components to come from the same specified family, an assumption that may not hold in heterogeneous real-world data. Finally, switching between different component families generally demands completely different EM or MCMC algorithms, making model selection and component-family adjustment computationally costly and practically inconvenient.

Instead of specifying a more flexible parametric family for each component, Aragam et al. (2020) proposes fitting nonparametric components by merging components from an overfitted Gaussian

mixture model, and establishes key identifiability conditions for nonparametric mixture models under this framework. However, this approach relies on a strong component-separability condition and doesn't perform well when components have overlapping tails. In other contexts, the general idea of merging components from an overfitted model has been utilized in estimating the number of components in finite mixture models ([Manole and Khalili, 2021](#); [Guha et al., 2021](#)), to learn hierarchical clustering structures ([Do et al., 2024](#)), and to mitigate model misspecification ([Dombowsky and Dunson, 2025](#)). [Tai and Aragam \(2023\)](#) obtain theoretical convergence rates for nonparametric components in a mixture model by assuming that the support of the underlying mixing measure is fixed and known. However, their estimation method is not directly applicable in practice when the support is typically unknown. [Aragam and Yang \(2025\)](#) employs Gaussian convolutions with varying scale parameters to uncover latent data structure without assuming a mixture-model form for the density. More recently, [Chakraborty and Nguyen \(2026\)](#) investigated the identifiability of a broad class of mixture models with nonparametric components, each defined as a convolution with a probability measure supported on a low-dimensional affine manifold, and developed methods for modeling a parametric subclass of these components. These approaches represent important steps toward uncovering clustering structures in data with complex subpopulations. However, they do not provide a systematic and practical framework for estimating the underlying component distributions in a fully nonparametric setting.

In this chapter we develop a first and practical Bayesian method with theoretical guarantees to estimate the mixture models of nonparametric components. We focus on two classes of nonparametric mixture models: the first class is characterized by the assumption that each component has its probability mass spatially concentrated around a connected region, with distinct components concentrated in different regions. We allow overlap between component tails. Apart from this mild spatial separation condition, we do not impose any additional assumptions on the component distributions. The second class of models are spike-and-slab-type mixtures, in which one component exhibits one or more high-density “spikes”, while the other component has a flat, relatively low-density distribution over the entire support. The two components may share fully overlapping

support. Moreover, we develop a unified framework that uses mixtures of Dirichlet process mixtures with Gaussian kernels to model both types of mixture structures. The hierarchical design leverages conjugacy to enable an efficient MCMC algorithm for fitting the posterior distributions. We apply our method to both univariate and multivariate settings through several simulation studies, and further demonstrate its scalability and practical performance in two real-data applications: disentangling astronomical sources in two-dimensional space (from approximately 0.8 million observed events) and analyzing the dynamic behavioral patterns of an oceanic whitetip shark using movement-metric data.

To enable identifiability of nonparametric mixture models while allowing for the overlap of the mixture components’ support, we develop a new separation condition defined in terms of the distances between connected regions within the support of the latent mixing measure. We note that these settings are not covered by the geometric identifiability conditions in the modeling approach of [Chakraborty and Nguyen \(2026\)](#). Furthermore, we establish theoretical guarantees showing the posterior contraction rate of the estimated component densities under appropriate conditions. To our knowledge, this is perhaps the first theoretical guarantee for a practical Bayesian method that learns nonparametric component densities within a finite mixture modeling framework. In our work, we show that the posterior contraction rate for the component densities is of the same order as the convergence rate of the point estimator studied by [Tai and Aragam \(2023\)](#). Since this rate does not admit a closed-form expression, we further derive an explicit upper bound, which reveals that the resulting rate is nearly polynomial.

The chapter is organized as follows. Preliminary background is introduced in Section 2.2. Section 2.3 introduces the problem setting for the two classes of mixture models described above. In Section 2.4, we present the framework of mixtures of Dirichlet process mixtures (MDPM) and illustrate our method through several simulation examples. Section 2.5 extends the method to multivariate distributions and provides an additional simulation study. In Section 2.6, we introduce the separation conditions based on which the identifiability of nonparametric mixture components can be established. Section 2.7 presents the posterior contraction results for both the overall mixture

density and the individual component densities under certain conditions. Finally, Section 2.8 applies our method to the oceanic whitetip shark acceleration data and the XMM-Newton astronomical dataset.

2.2 Preliminaries

In this section, we recall basic notions of nonparametric mixture models, and introduce the definition of topological connectedness that will be useful in determining the identifiability of mixture models.

Mixture models Let $\mathcal{P}(\mathbb{R}^m)$ denote the set of Borel probability measures on $\mathbb{R}^m$. Suppose $F \in \mathcal{P}(\mathbb{R}^m)$. We say that F admits a finite mixture representation if, for any Borel set $A \subseteq \mathbb{R}^m$, $F(A)$ takes the form:

$$F(A) = \sum_{i=1}^K w_i G_i(A) = \int G(A) d\Lambda(G), \quad K := |\text{supp}(\Lambda)|, \quad (w_1, \dots, w_K) \in \Delta^{K-1}, \quad (2.1)$$

where G_i 's are distinct probability measures from a subset $\tilde{\mathcal{P}} \subseteq \mathcal{P}(\mathbb{R}^m)$, and Λ is a probability measure over $\tilde{\mathcal{P}}$, supported on a finite set of K component distributions $G_1, \dots, G_K$. The choice of $G_1, \dots, G_K$ and $w_1, \dots, w_K$ do not depend on A . In this case, we refer to F as a finite mixture model with K components, where each G_i represents a distinct subpopulation of F . If the component class $\tilde{\mathcal{P}}$ is parametric, e.g. m -dimensional Gaussian distributions, then F is a mixture model of parametric components. In this work, however, we focus on a more general setting where K is finite and known, but $\tilde{\mathcal{P}}$ is a nonparametric family. The first question that arises is the identifiability of the representation (2.1). Denote by $\mathcal{P}(\tilde{\mathcal{P}})$ a set of probability measures on $\tilde{\mathcal{P}}$. The classic definition of [Teicher \(1961\)](#) specifies the probability measure F to be identifiable in $\mathcal{P}(\tilde{\mathcal{P}})$ if for any $\Lambda, \Lambda' \in \mathcal{P}(\tilde{\mathcal{P}})$ the condition $F = \int G d\Lambda(G) = \int G d\Lambda'(G)$ entails that $\Lambda = \Lambda'$.

In this work, we study the conditions under which F is identifiable, and develop a Bayesian method to estimate each component probability measure G_i , as well as its corresponding weight w_i for $i = 1, \dots, K$, while we only have observations from F . A fundamental concept underlying our

identifiability conditions is connectedness, therefore we first introduce the notion of connectedness for a topological space $\mathbb{X}$ (cf. [Munkres \(2020\)](#)).

Definition 2.2.1 (Connectedness). *Let $\mathbb{X}$ be a topological space. A separation of $\mathbb{X}$ is a pair of disjoint nonempty open subsets A and B of $\mathbb{X}$ s.t. $\mathbb{X} = A \cup B$. $\mathbb{X}$ is connected if there does not exist a separation of $\mathbb{X}$, otherwise it is disconnected.*

The above definition makes precise the natural intuition that a topological space is connected if it is “all in one piece”, meaning it cannot be split into distinct parts that do not touch each other. We now recall two classical Gaussian mixture families on $\mathbb{R}$. These serve as building blocks for the mixtures with nonparametric components studied in this chapter. Later we extend the definitions to the multivariate setting $\mathbb{R}^m$.

Location mixture of normals Let $g_{u,\sigma}$ stand for the normal density on $\mathbb{R}$ with mean u and standard deviation σ : $g_{u,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-u)^2}{2\sigma^2}\right)$. The density $p(x)$ is a location mixture of normals if

$$p(x) = p_{\sigma,V}(x) := \int g_{u,\sigma}(x) dV(u) \quad (2.2)$$

for some $\sigma > 0$, where $V(\cdot)$ is a mixing probability distribution on $\mathbb{R}$.

Location-scale mixture of normals $p(x)$ is a location-scale mixture of normals if

$$p(x) = p_H(x) := \int g_{u,\sigma}(x) dH(u, \sigma), \quad (2.3)$$

where $H(\cdot, \cdot)$ is a mixing probability distribution on $\mathbb{R} \times (0, \infty)$.

2.3 Problem and modeling assumptions

In this work, we study finite mixture models with nonparametric component densities. For clarity, we begin with mixture models on $\mathbb{R}$. The multivariate case follows in subsequent sections. Our

goal is to learn both the overall population distribution and the latent subpopulation distributions within the mixture model framework (2.1), assuming that the number of components K is finite, fixed and known. To this end, we develop our methodology under two different settings:

- (S2.1) The component distribution class $\tilde{\mathcal{P}}$ consists of location mixtures of normals, as defined in (2.2). The variance parameter $\sigma > 0$ is fixed but possibly unknown, and the mixing distribution $V(u)$ is assumed to have bounded support. We further restrict $\mathcal{P}(\tilde{\mathcal{P}})$ to probability measures supported on K components satisfying the following separation condition: for any two components $G_i = p_{\sigma, V_i}$ and $G_j = p_{\sigma, V_j}$, $i \neq j$, there exist disjoint bounded connected sets $I_i, I_j \subset \mathbb{R}$ (i.e., intervals on $\mathbb{R}$) such that $\text{supp}(V_i) \subseteq I_i$, and $\text{supp}(V_j) \subseteq I_j$.
- (S2.2) The component distribution class $\tilde{\mathcal{P}}$ consists of location-scale mixtures of normals, as defined in (2.3). The mixing distribution H has bounded support on (u, σ) . Accordingly, $\mathcal{P}(\tilde{\mathcal{P}})$ represents a set of probability measures supported on K components satisfying the separation condition: there exists $x \in \{u, \sigma\}$, such that for any two components $G_i = p_{H_i}$, and $G_j = p_{H_j}$, $i \neq j$, we can find disjoint bounded connected sets $I_i, I_j \subset S_x$ with $\text{supp}(\pi_x \# H_i) \subseteq I_i$, and $\text{supp}(\pi_x \# H_j) \subseteq I_j$, where $S_u = \mathbb{R}$, $S_\sigma = (0, \infty)$. Here, $\pi_x \# H$ stands the x -marginal distribution of H , where $x \in \{u, \sigma\}$.

It is important to note that the assumptions in the two settings apply only to our models and are not imposed on the actual data-generating distribution. In particular, the true component densities do not have to be a location mixture or a location–scale mixture of Gaussians. In (S2.1), we assume that each component in the finite mixture model is a location mixture of normals. This family is already quite rich, since Gaussian kernel density estimators are themselves location mixture of normals, and it is known that Gaussian kernel density estimators can approximate any continuous density arbitrarily well as $\sigma \rightarrow 0$ (Parzen (1962)). By restricting the support of mixing distribution V to be bounded and imposing a separation condition on components, (S2.1) characterizes mixture models in which each component has a dominant support region where most of its mass is concentrated, while still allowing distinct components to overlap in their tails. (S2.2),

on the other hand, generalizes the component distribution family $\tilde{\mathcal{P}}$ to location-scale mixture of normals. In (S2.2) the separation condition is imposed on either the u -marginal or the σ -marginal; thus, in addition to the spatially differentiated components, this class also accommodates spike-and-slab-type components that overlap in location but differ in scale: one component exhibits one or more high-density “spikes”, while the other has a relatively flat, low-density distribution over the entire support. The two components may have fully overlapping support. In this work, we develop a Bayesian framework for learning mixture models applicable to the general setting (S2.2) in Section 2.4. We establish theoretical results under setting (S2.1), which consists of (i) an identifiability theory in Section 2.6 and (ii) posterior contraction rates of both the mixture density and the component densities in Section 2.7.

2.4 Mixture of Dirichlet Process mixtures

Dirichlet process (DP), introduced by [Ferguson \(1973, 1974\)](#), is a foundational prior for Bayesian nonparametric modeling and has become widely used due to its analytical tractability. To model data arising from multiple distinct yet related subpopulations, a rich class of structured extensions of the DP has been developed to accommodate different data-generating mechanisms. Notable examples include the hierarchical DP ([Teh et al., 2006](#)), nested DP ([Rodríguez et al., 2008](#)), additive DP ([Müller et al., 2004](#)), and dependent DP ([MacEachern, 2000](#)), among many others. A unifying perspective on these constructions is provided by [Franzolini et al. \(2025\)](#).

As we shall see, the hierarchical Bayesian modeling framework may be extended and adapted to incorporate the model setting (S2.2). In particular, we shall describe a formulation of a mixture of Dirichlet process mixtures (MDPM), an early version of which goes back to [Antoniak \(1974\)](#). We begin by recalling the Dirichlet process mixture (DPM); see, e.g., [Ghosal and van der Vaart \(2017\)](#) and [Müller et al. \(2015\)](#). Consider the location-scale mixture of normals in (2.3): $p_H(x) := \int g_{u,\sigma}(x)dH(u,\sigma)$. If we place a Dirichlet process prior on H , i.e., $H \sim \text{DP}(\alpha H_0)$ with base measure H_0 on $\mathbb{R} \times (0, \infty)$ and concentration parameter $\alpha > 0$, then the induced random density p_H is called a Dirichlet process mixture (DPM).

In setting (S2.2), we impose a separation condition either on u or σ . For clarity, we begin with the condition on u . Since any nonempty bounded connected subset I of $\mathbb{R}$ is an interval, we parametrize it as $[c - r, c + r]$, where $c \in \mathbb{R}$ is the midpoint and $r > 0$ is the half-length. We then place priors on (c, r) to model the unknown intervals. By the separation condition in (S2.2), there exist K pairwise disjoint intervals $I_1, \dots, I_K$ such that, for each $i = 1, \dots, K$, the i -th component $G_i = p_{H_i}$ satisfies $\text{supp}(\pi_u \# H_i) \subseteq I_i$. To ensure the K intervals $I_1, \dots, I_K$ are pairwise disjoint, we adapt the repulsive priors of [Petrulia et al. \(2012\)](#); [Xie and Xu \(2020\)](#) to our setting and place the following prior on $(\mathbf{c}, \mathbf{r})$ where $\mathbf{c} = (c_1, \dots, c_K)$ and $\mathbf{r} = (r_1, \dots, r_K)$:

$$p_{\text{rp}}(\mathbf{c}, \mathbf{r}) \propto \left(\prod_{i=1}^K g_{\mu_0, \eta}(c_i) \cdot \text{Gamma}(r_i) \right) \cdot \min_{1 \leq i < j \leq K} \exp \left(-\frac{\tau}{\max(|c_i - c_j| - r_i - r_j, 0)^\nu} \right), \quad (2.4)$$

where $\tau > 0$ is a scale parameter and ν is a positive integer controlling the rate at which the exponential term approaches zero as $\max(|c_i - c_j| - r_i - r_j, 0)$ decreases, and $\eta > 0$ controls the prior spread of the centers $\mathbf{c}$. We omit the parameters of the Gamma distribution for notational convenience. The repulsion term is zero whenever intervals overlap (i.e., $|c_i - c_j| \leq r_i + r_j$), thereby assigning zero prior mass to non-disjoint configurations. Then for the i -th component, the base measure H_{i0} on (u, σ) assigns to u a normal distribution truncated to $[c_i - r_i, c_i + r_i]$, and to σ^2 an inverse-gamma distribution on $(0, \infty)$, with u and σ independent. Equivalently, H_{i0} has density

$$h_{i0}(u, \sigma) \propto g_{c_i, \sigma_0}(u) \mathbf{1}_{[c_i - r_i, c_i + r_i]}(u) \cdot \text{InvGamma}(\sigma^2) \cdot 2\sigma, \quad (2.5)$$

where $\sigma_0 > 0$ is a hyperparameter. The inverse-gamma parameters are omitted for notational convenience. We place a Dirichlet process prior $\text{DP}(\alpha H_{i0})$ on H_i , for some $\alpha > 0$. With the choice of H_{i0} above, the DPM is conditionally conjugate at the component level, enabling closed-form updates and substantially improving computational efficiency. Finally, we place a truncated Dirichlet prior ([Fang et al., 2000](#)) on the mixture weights $\mathbf{w} = (w_1, \dots, w_K) \in \Delta^{K-1}$. A random vector $\mathbf{x} \in \Delta^{K-1}$ is said to follow a truncated Dirichlet distribution $\text{TD}(\gamma_1, \dots, \gamma_K; \mathbf{a}, \mathbf{b})$, if its

density is given by

$$p(\mathbf{x}) = c \prod_{i=1}^K x_i^{\gamma_i-1}, \quad \mathbf{x} \in S_{K-1}(\mathbf{a}, \mathbf{b})$$

where c is a normalizing constant, and $S_{K-1}(\mathbf{a}, \mathbf{b}) = \{\mathbf{x} \in \Delta^{K-1} : a_i \leq x_i \leq b_i, i = 1, \dots, K\}$. The truncation set $S_{K-1}(\mathbf{a}, \mathbf{b})$ is nonempty if and only if $\sum_{i=1}^K a_i \leq 1 \leq \sum_{i=1}^K b_i$. In the special case $K = 2$, this distribution reduces to the truncated beta distribution. We specify the prior on the mixture weights $\mathbf{w}$ as $\mathbf{w} \sim \text{TDDir}(1/K, \dots, 1/K; \underline{w}\mathbf{1}_K, \overline{w}\mathbf{1}_K)$, where $\underline{w}$ and $\overline{w}$ are lower and upper bounds on each component weight, satisfying $0 < K\underline{w} < 1$ and $K\overline{w} > 1$. For simplicity we assume that all component weights share the same lower and upper bounds. In practice one may assign component-specific bounds $\underline{w}_i > 0$ and $\overline{w}_i > 0$, as long as $0 < \sum_{i=1}^K \underline{w}_i < 1$ and $\sum_{i=1}^K \overline{w}_i > 1$. Based on the construction above, we obtain the hierarchical model specified as follows:

$$\begin{aligned} \mathbf{w} = (w_1, \dots, w_K) &\sim \text{TDDir}(1/K, \dots, 1/K; \underline{w}\mathbf{1}_K, \overline{w}\mathbf{1}_K), \text{ given } K \\ z_1, \dots, z_n &\stackrel{\text{i.i.d.}}{\sim} \text{Categorical}(w_1, \dots, w_K), \\ (\mathbf{c}, \mathbf{r}) &\sim p_{\text{rp}}(\mathbf{c}, \mathbf{r}) \\ H_{i0}(u, \sigma) &\propto g_{c_i, \sigma_0^2}(u) \cdot \mathbf{1}_{[c_i-r_i, c_i+r_i]}(u) \cdot \text{InvGamma}(\sigma^2) \cdot 2\sigma \\ H_i &\sim \text{DP}(\alpha H_{i0}) \\ G_i = p_{H_i} &= \int g_{u, \sigma} dH_i(u, \sigma) \text{ for } i = 1, \dots, K \\ x_j &\sim G_{z_j} \text{ for } j = 1, \dots, n, \text{ given } G_{1:K}, z_{1:n}. \end{aligned} \tag{2.6}$$

Here, $x_1, \dots, x_n$ are the n samples observed, and $z_i, i = 1, \dots, n$ are the latent variables indicating the component generating x_i . Figure 2.1 illustrates the hierarchical model for the two-component case ($K = 2$).

In some application, we may have prior knowledge about the connected regions in which most of the component mass is concentrated. In such cases, the repulsive prior on $(\mathbf{c}, \mathbf{r})$ can be replaced by fixed parameters $(\mathbf{c}, \mathbf{r})$. When the separation condition is imposed on σ , we treat $(\mathbf{c}, \mathbf{r})$ as fixed

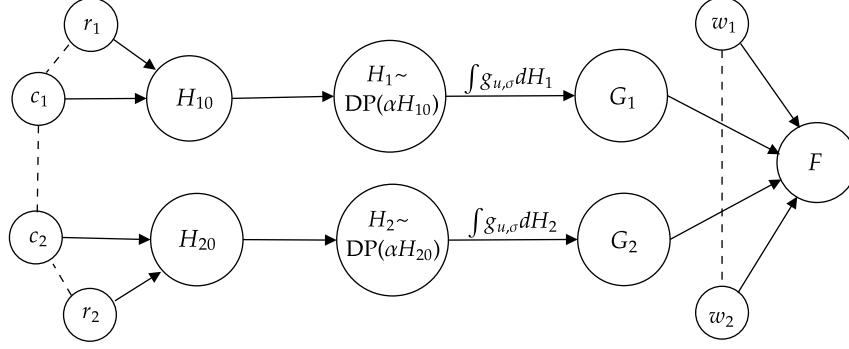


Figure 2.1: Graphical model representation of the MDPM (2.6) for $K = 2$. The dashed edges among c_1, r_1, c_2, r_2 represent the repulsive prior on $(\mathbf{c}, \mathbf{r})$, which enforces disjoint intervals I_1 and I_2 . For $i = 1, 2$, the base measure $H_{i0}(u, \sigma)$ is supported on $I_i \times (0, \infty)$. The dashed edge between w_1 and w_2 indicates the truncated Beta prior (equivalently, the two-parameter truncated Dirichlet) on the mixture weights $\mathbf{w} = (w_1, w_2)$. Finally, the mixture weights $\mathbf{w} = (w_1, w_2)$ combine the component distributions into the overall mixture $F = w_1 G_1 + w_2 G_2$.

and known, and modify the base measure $H_{i0}(u, \sigma)$, for $i = 1, \dots, K$ to

$$H_{i0}(u, \sigma) \propto g_{c_0, \sigma_0^2}(u) \cdot \text{InvGamma}(\sigma^2) \cdot 2\sigma \cdot \mathbf{1}_{[c_i - r_i, c_i + r_i]}(\sigma).$$

Note that the structure of (2.6) resembles the classical Bayesian hierarchical model for finite mixture of parametric components (Diebolt and Robert (1994); Richardson and Green (1997); Miller and Harrison (2018)), apart from replacing the prior of component parameters with a DPM prior. Given that a single DPM is already highly flexible—in particular, it achieves consistent estimation for a broad class of smooth densities at a nearly optimal rate (Ghosal and van der Vaart (2007))—one may wonder whether employing a mixture of DPMs (MDPM) is necessary. Accordingly, we explain why an MDPM is warranted and highlight its benefits over a single DPM. The first advantage lies in increased modeling effectiveness and computational efficiency. While one might expect a single DPM prior to be simpler, in practice this requires the base measure itself to be a mixture supported on several disjoint connected sets. Modeling and posterior inference with respect to complex forms of the latent base measure may be highly inefficient. In particular, constructing a base measure supported on disjoint subsets breaks conjugacy and greatly reduces

the efficiency of the MCMC sampling. By contrast, employing a mixture of DPMs—each with a truncated normal-inverse Gamma base measure—preserves conjugacy at the component level and substantially improves sampling efficiency. Second, the structural similarity between MDPM and the classical Bayesian hierarchical model for finite parametric mixtures allows us to leverage the extensive toolbox of methods developed for finite mixture modeling over the past several decades. For example, repulsive priors were originally proposed to reduce redundancy among components in Bayesian mixture models (Petrailia et al. (2012)) and are here adapted to model disjoint intervals. Therefore, the MDPM framework not only enables efficient computation but also offers considerable potential for extension using the rich set of tools developed for parametric mixture modeling. In addition, in Section 2.7 it will be shown that, under appropriate conditions, an MDPM prior achieves the same posterior contraction rate to the true mixture density as a single DPM prior. This further justifies the MDPM framework—counterintuitively, it does not add complexity but rather provides a practical framework even if one is only interested in modeling the data population, not its latent subpopulations. This leads to the fourth, and perhaps the most interesting feature of the MDPM framework, which is that it enables the inference of individual components’ distributions even when they have complex forms for which no parametric class of probability kernels can be adequate.

Inference algorithms for DPMs are typically of two types, corresponding to two representations of the DPM: the Pólya–urn representation (Blackwell and MacQueen (1973)) and the stick-breaking representation (Sethuraman (1994)). The Pólya–urn representation integrates out the random measures from the Dirichlet process and characterizes the marginal distribution of the observed data. This perspective underlies a class of Gibbs samplers (Escobar, 1994; MacEachern and Müller, 1998; Neal, 2000; Jain and Neal, 2004). In contrast, the stick-breaking representation provides an explicit constructive form of the random distributions. To deal with the potentially infinite number of components in the random distribution functions, Walker (2007); Kalli et al. (2011) developed the slice sampler. Because we aim to infer the mixture weights $\mathbf{w}$, the component distributions $\{G_i\}_{i=1}^K$, and the overall mixture F jointly, an explicit representation of the random distribution functions is preferred. Accordingly, we adapt a slice sampler to the MDPM. In order to accommodate large-

dataset applications, we adopt the slice-sampling approach proposed by [Ge et al. \(2015\)](#), which leverages a MapReduce framework to parallelize the sampler’s computations. Our algorithm for MDPM is detailed in Appendix A.1. We implement the algorithm in Julia, leveraging its fast loops and built-in support for parallel programming. The source code is available at [GitHub](#). As noted earlier, conjugacy yields closed-form conditionals, which makes the sampler highly efficient.

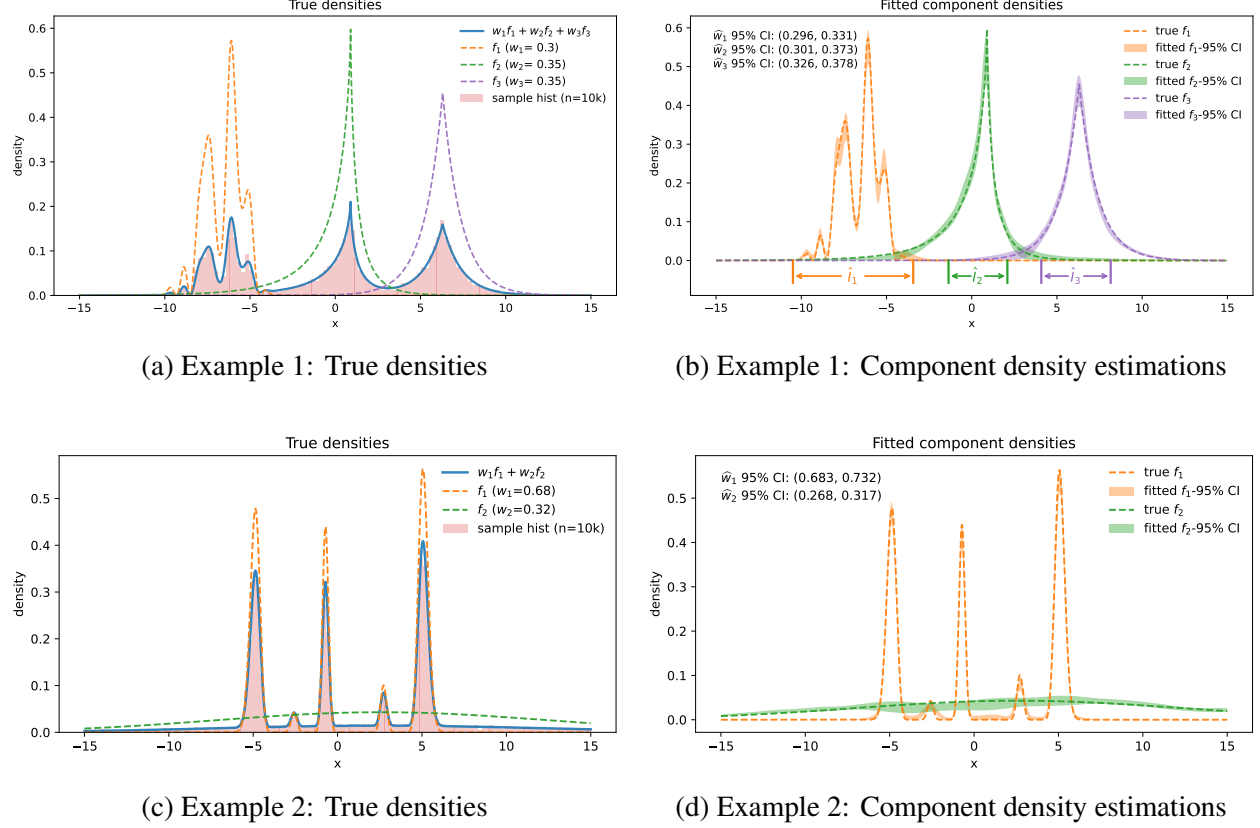


Figure 2.2: Example 1: a three-component mixture with the separation condition imposed on the location parameter u ; f_1 is generated from a random combination of Hermite functions; f_2 is a skewed exponential-power distribution, and f_3 is a Laplace distribution. Example 2: a two component mixture with the separation condition imposed on the scale parameter σ ; Both component densities are generated from random combinations of Hermite functions.

Two simulation examples are shown in Figure 2.2. Example 1 in panels (2.2a) and (2.2b) illustrates a setting in which the separation condition is imposed on the location parameter u . In (2.2b), the pointwise 95% posterior credible intervals for the component densities closely follow the true component densities, including within the overlapping tail region between f_2 and f_3 . The

three estimated latent intervals are displayed beneath the density curves and approximately identify the regions where each component places most of its mass, and the estimated weights are shown in the upper-left corner in (2.2b). Panels (2.2c) and (2.2d) show a second example in which the separation condition is imposed on the scale parameter σ . In both scenarios, the proposed method demonstrates strong performance. For the simulation in Example 1 with $n = 10,000$ observations and $K = 3$ components, 10,000 MCMC iterations without multi-threading take about 1.5 minutes (wall-clock) on a MacBook Pro (Apple M4, 10-core; 24 GB unified memory). For large datasets, enabling multi-threading typically yields an additional 10–20% speedup. We further demonstrate the scalability of our method by applying a multivariate version of our algorithm to an astronomical dataset containing approximately 0.8 million two-dimensional events in Section 2.8.

2.5 Multivariate distributions

This section extends the problem setting in section 2.3 and the method in section 2.4 to the multivariate case. We first consider location-scale mixture of normals on $\mathbb{R}^m$: For any $\boldsymbol{\mu} \in \mathbb{R}^m$, $\Sigma \in \mathbb{R}^{m \times m}$, $\Sigma \succ 0$, let $g_{\boldsymbol{\mu}, \Sigma}$ denote the m -dimensional normal density with mean $\boldsymbol{\mu}$ and covariance matrix Σ : $g_{\boldsymbol{\mu}, \Sigma}(\mathbf{x}) = \left(\sqrt{2\pi}\right)^{-m} |\Sigma|^{-1/2} \exp\left(-(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$.

We say $p(\mathbf{x})$ is a location-scale mixture of normals on $\mathbb{R}^m$ if there exists a mixing distribution $H(\cdot, \cdot)$ on $\mathbb{R}^m \times \mathbb{S}_{++}^m$ such that

$$p(\mathbf{x}) = p_H(\mathbf{x}) := \int g_{\boldsymbol{\mu}, \Sigma}(\mathbf{x}) dH(\boldsymbol{\mu}, \Sigma), \quad (2.7)$$

where $\mathbb{S}_{++}^m$ denotes the set of all $m \times m$ positive-definite matrices. In this section, we consider mixture models under the following setting:

- (S2.3) The component distribution class $\tilde{\mathcal{P}}$ consists of location-scale mixtures of normals on $\mathbb{R}^m$, as defined in (2.7). The mixing measure H satisfies: (i) its $\boldsymbol{\mu}$ -marginal $\pi_{\boldsymbol{\mu}} \# H$ has bounded support on $\mathbb{R}^m$; and (ii) its Σ -marginal $\pi_{\Sigma} \# H$ is supported on covariance matrices with eigenvalues uniformly bounded above and away from 0. $\mathcal{P}(\tilde{\mathcal{P}})$ are probability measures

supported on finitely many components satisfying the separation condition: for any two components $G_i = p_{H_i}$, and $G_j = p_{H_j}$, $i \neq j$, we can find two disjoint bounded connected sets $I_i, I_j \subset \mathbb{R}^m$ such that $\text{supp}(\pi_\mu \# H_i) \subseteq I_i$, and $\text{supp}(\pi_\mu \# H_j) \subseteq I_j$.

To extend the hierarchical model (2.6) to the multivariate setting (S2.3), the major modification is the modeling of pairwise disjoint connected sets $I_1, \dots, I_K \subset \mathbb{R}^m$. In this work, we model $I_1, \dots, I_K$ as axis-aligned hypercubes centered at c_k with half-side length $r_k > 0$, equivalently,

$$I_k = \{x \in \mathbb{R}^m : \|x - c_k\|_\infty \leq r_k\}, \quad k = 1, \dots, K.$$

Although modeling I_k 's with more flexible shapes are possible, the main computational bottleneck is efficient sampling from a truncated multivariate normal. Most existing methods address linear constraints ([Geweke et al. \(1991\)](#); [Kotecha and Djuric \(1999\)](#); [Fernández et al. \(2007\)](#); [Chopin \(2011\)](#); [Botev \(2017\)](#)), and an exact Hamiltonian Monte Carlo sampler [Pakman and Paninski \(2014\)](#) has been developed to handle both linear and quadratic inequality constraints. However, scalable samplers for more flexible constraints remain limited. Thus, extending from hypercubes to more general H-polytopes is straightforward ([Botev \(2017\)](#)), but efficient sampling under nonconvex constraints remains underexplored. Developing scalable algorithms for such cases is an interesting direction for future work.

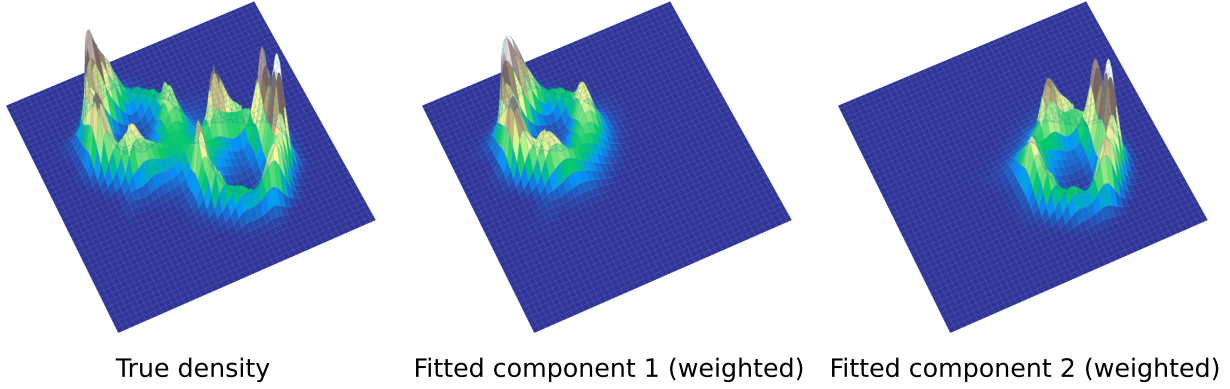


Figure 2.3: A bivariate two-component mixture model. Each component is a Gaussian mixture located on a circle with random covariance matrices. The left panel displays the true density, while the middle and right panels show the pointwise posterior means of the fitted component densities, weighted by their respective posterior mixture weights.

Accordingly, we extend the repulsive prior (2.4) to $\mathbb{R}^m$,

$$p_{mrp}(\mathbf{c}, \mathbf{r}) \propto \left(\prod_{i=1}^K g_{\mu_0, \Sigma_0}(\mathbf{c}_i) \cdot \text{Gamma}(r_i) \right) \cdot \min_{1 \leq i < j \leq K} \exp \left(-\frac{\tau}{\max(\|\mathbf{c}_i - \mathbf{c}_j\|_\infty - r_i - r_j, 0)^v} \right), \quad (2.8)$$

and change H_{i0} to its multivariate counterpart:

$$H_{i0}(\boldsymbol{\mu}, \Sigma) \propto g_{\mathbf{c}_i, \Sigma_1}(\boldsymbol{\mu}) \cdot \mathbf{1}_{[\mathbf{c}_i - r_i, \mathbf{c}_i + r_i]}(\boldsymbol{\mu}) \cdot \text{InvWishart}(\Sigma), \quad (2.9)$$

where InvWishart denotes the inverse-Wishart distribution, and hyperparameters are omitted for brevity. The rest of model (2.6) remain unchanged. A simulation example is shown in Figure 2.3.

In some applications, reasonable a priori estimates of where each component lies are available; in such settings we can fix $I_1, \dots, I_K$ and drop the repulsive prior over regions, yielding a simpler model. Code for the fixed-region algorithm is also provided. An application example is in Section 2.8.

2.6 Identifiability

In this section, we assume that the true mixture distribution satisfies (S2.1) and establish conditions for the identifiability of the mixture model's nonparametric components. We begin by introducing notions of distance between connected sets in the topological space $\mathbb{X}$.

Let I_1 and I_2 be two connected subsets of $\mathbb{X}$ such that $I_1 \cup I_2$ is disconnected. If the space $\mathbb{X}$ is equipped with some metric d , we can define the distance $d_c(I_1, I_2)$ between the two connected sets I_1 and I_2 as

$$d_c(I_1, I_2) = \inf_{x \in I_1, y \in I_2} d(x, y). \quad (2.10)$$

Any non-empty $I \subseteq \mathbb{X}$ has a unique separation S_I such that each element $I_i \in S_I$ is non-empty and connected, and if S_I contains more than one element, then for any subset $T \subseteq S_I$ with $|T| \geq 2$, the union $\bigcup_{I_i \in T} I_i$ is disconnected, and moreover $I = \bigcup_{I_i \in S_I} I_i$. This separation is unique because if there is a different separation S'_I that satisfies the same conditions, there must be a connected set $I^* \in S_I$ such that $I^* \notin S'_I$, then $\bigcup_{A \in S'_I} (I^* \cap A) = I^*$ is connected, which contradicts the conditions of S'_I . Note that S_I may be an infinite set. In particular, a singleton is itself a connected set.

For distinct $I_i, I_j \in S_I$, we say I_i and I_j are neighbors in S_I if there is no $I_k \in S_I \setminus \{I_i, I_j\}$ such that $d_c(I_i, I_k) < d_c(I_i, I_j)$ and $d_c(I_j, I_k) < d_c(I_i, I_j)$. Denote the set of all such unordered neighbor pairs by $\mathcal{N}(S_I)$. Then we can define the within-space distance d_w of a space I as

$$d_w(I) = \begin{cases} 0, & |S_I| = 1 \\ \sup_{(I_i, I_j) \in \mathcal{N}(S_I)} d_c(I_i, I_j), & |S_I| \geq 2 \end{cases} \quad (2.11)$$

Furthermore, the between-space distance d_b of two spaces I and J can be defined as

$$d_b(I, J) = \inf_{I_i \in S_I, J_j \in S_J} d_c(I_i, J_j). \quad (2.12)$$

In this section, we provide theoretical results for the identifiability of mixture models under

setting (S2.1). Let $I_1, \dots, I_K$ be pairwise disjoint compact intervals on $\mathbb{R}$, where $I_k = [c_k - r_k, c_k + r_k]$ for some $c_k \in \mathbb{R}$, and $r_k > 0$ ($k = 1, \dots, K$). Without loss of generality, index these intervals so that $c_k + r_k < c_{k+1} - r_{k+1}$, $\forall 1 \leq k \leq K-1$. Consider the mixture model

$$f = \sum_{k=1}^K w_k f_k, \text{ s.t. } f_k = \int g_{u,\sigma}(x) dV_k(u), \text{ supp}(V_k) \subseteq I_k, \text{ and } (w_1, \dots, w_K) \in \Delta^{K-1}, \quad (2.13)$$

where $\sigma \in (0, \infty)$ is a fixed but probably unknown standard deviation. For $k = 1, \dots, K$, V_k can be any probability measure supported within I_k . To ensure the identifiability of (2.13), we consider the two situations:

(C2.1) The sets $I_1, \dots, I_K$ are fixed and known.

(C2.2) K is fixed and known, but the sets $I_1, \dots, I_K$ are unknown, and

$$0 \leq \max_{1 \leq k \leq K} d_w(\text{supp}(V_k)) < \min_{1 \leq i < j \leq K} d_b(\text{supp}(V_i), \text{supp}(V_j)).$$

The following theorem shows that if either condition (C2.1) or (C2.2) is satisfied, the function f defined as in (2.13) has a unique representation as a mixture model.

Theorem 2.6.1. *For any f as defined in (2.13) such that (C2.1) or (C2.2) is satisfied, there exist a unique σ , a unique collection of densities $f_1, \dots, f_K$, and a corresponding weight vector $(w_1, \dots, w_K) \in \Delta^{K-1}$ that satisfy the conditions in (2.13) such that $f = \sum_{k=1}^K w_k f_k$.*

2.7 Posterior contraction rates

In this section, we obtain the contraction rates of the Bayesian posteriors arising from the models described in the previous sections. Our theoretical analysis focuses on the following class of mixture densities:

Let $\mathcal{F}_{K,\sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$ denote the class of densities f of the form $f = \sum_{k=1}^K w_k f_k$, where $(w_1, \dots, w_K) \in \Delta^{K-1}$, $K \geq 2$ satisfies $w_k \in [\underline{w}, \frac{1-\underline{w}}{K-1}]$, $k = 1, \dots, K$ for some $0 < \underline{w} < 1/K^2$. For

each $k = 1, \dots, K$, the component density is given by $f_k(x) = \int g_{u,\sigma}(x) dV_k(u)$, $\forall x \in \mathbb{R}$, where V_k is a probability measure with $\text{supp}(V_k) \subset [c_k - r_k, c_k + r_k]$, for some $c_k \in \mathbb{R}$, $r_k \in (0, +\infty)$. Here $\mathbf{c} = (c_1, \dots, c_K)$ and $\mathbf{r} = (r_1, \dots, r_K)$. Without loss of generality, we assume that $c_k + r_k < c_{k+1} - r_{k+1}$ for $1 \leq k \leq K - 1$. Throughout this section, $K, \sigma, \mathbf{c}, \mathbf{r}, \underline{w}$ are assumed fixed and known, and the true mixture density $f_0 = \sum_{k=1}^K w_{k0} f_{k0}$ is assumed to lie in $\mathcal{F}_{K,\sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$.

We study posterior contraction rates for Bayesian density estimation under a mixture of Dirichlet process mixture (MDPM) prior imposed on the density class $\mathcal{F}_{K,\sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$. The prior specification is given by the following assumptions:

(A2.1) Let $\mathbf{W} := (W_1, \dots, W_K)$ follow the distribution $\text{Tdir}(\alpha_1, \dots, \alpha_K; \underline{w} \mathbf{1}_K, \frac{1-\underline{w}}{K-1} \mathbf{1}_K)$. We further assume that the concentration parameters satisfy $A \leq \alpha_k \leq 1$, $k = 1, \dots, K$, for some constant $0 < A < 1/K$.

(A2.2) For $1 \leq k \leq K$, let H_k be a finite, positive measure on $[c_k - r_k, c_k + r_k]$. D_{H_k} is a Dirichlet process with base measure H_k .

(A2.3) We assume $\mathbf{W}$ and D_{H_k} 's are mutually independent. Let $\Pi_k := D_{H_k}$ and $\Pi := \sum_{k=1}^K W_k D_{H_k}$.

Π defined above induces a prior on the mixing measure $\sum_{k=1}^K W_k V_k$. By the identifiability result established in Theorem 2.6.1, there is a one-on-one correspondence between the mixing measure and the associated mixture density $f \in \mathcal{F}_{K,\sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$. Consequently, Π also uniquely induces a prior on $\mathcal{F}_{K,\sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$. With a slight abuse of notation, we henceforth regard Π as a prior on $\mathcal{F}_{K,\sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$.

Section 2.7.1 establishes the posterior contraction rate of the overall mixture density under Assumptions (A2.1)–(A2.3), while Section 2.7.2 derives contraction rates for the component densities when $K = 2$ under additional separation conditions.

Throughout this section, we introduce several notions of distances and divergences between probability densities. The Hellinger distance d between two densities f and g is defined as: $d(f, g) = \left(\int \left(f^{1/2}(x) - g^{1/2}(x) \right)^2 dx \right)^{1/2}$, and the L_1 -norm $\|f - g\|_1 = \int |f(x) - g(x)| dx$. L_1 and

Hellinger distances are related by

$$d^2(f, g) \leq \|f - g\|_1 \leq 2d(f, g). \quad (2.14)$$

Let $p_0 \in \mathcal{P}$, a class of densities and let P_0 be the probability measure with density p_0 . The Kullback-Leibler (KL) divergence is $K(p_0, p) = \int \log(p_0/p) dP_0$, and the second Kullback-Leibler variation is $K_2(p_0, p) = \int (\log(p_0/p))^2 dP_0$. Also relevant is the KL ball $B_2(p_0, \epsilon) = \{p : K(p_0, p) \leq \epsilon^2, K_2(p_0, p) \leq \epsilon^2\}$.

2.7.1 Posterior contraction of the overall mixture density

In this section we establish the posterior contraction rates of the overall mixture density.

Theorem 2.7.1. *Let $f_0 \in \mathcal{F}_{K, \sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$, where $K, \sigma, \mathbf{c}, \mathbf{r}, \underline{w}$ are fixed and known. Denote by P_{f_0} the probability measure with density f_0 . Let $X^{(n)} := (X_1, \dots, X_n)$ be n i.i.d samples from P_{f_0} . Consider the prior Π constructed under specifications (A2.1), (A2.2), and (A2.3), and let $\Pi^{(n)}(\cdot | X^{(n)})$ be the posterior distribution induced by the prior Π . Then for a large enough constant M ,*

$$\Pi^{(n)} \left(d(f, f_0) > M \frac{\log n}{\sqrt{n}} | X_1, \dots, X_n \right) \rightarrow 0 \quad (2.15)$$

in $P_{f_0}^n$ -probability.

Theorem 2.7.1 establishes that, under the assumptions (A2.1)-(A2.3) for a MDPM prior, the posterior contraction rate of our estimator for the overall mixture density matches that obtained under a single Dirichlet process mixture prior.

The proof of the above theorem follows the general framework laid out by Theorem 2.1 in [Ghosal and van der Vaart \(2001\)](#). To apply this framework, we will find a positive sequence $\tilde{\epsilon}_n \rightarrow 0$ such that $\Pi(B_2(f_0, \tilde{\epsilon}_n)) \geq c_3 e^{-c_4 n \tilde{\epsilon}_n^2}$, for some constants $c_3, c_4 > 0$. Lemma A.4.2 identifies such $\tilde{\epsilon}_n$ for the mixture of Dirichlet processes prior under the assumptions (A2.1)-(A2.3), and the proof of Theorem 2.7.1 follows.

2.7.2 Posterior contraction of component densities

The novel and considerably more demanding part of our theory lies in establishing posterior contraction rates for the individual component densities. For the ease of technical exposition we shall restrict our theorem's claim to the case $K = 2$ with the following conditions:

(C2.3) Consider the density class $\mathcal{F} = \mathcal{F}_{2,\sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$, for some $0 < \underline{w} < \frac{1}{4}$. Assume that $\mathbf{c}$ and $\mathbf{r}$ satisfy $|c_1 - c_2| > 8 \max(r_1, r_2)$. Denote $r = |c_1 - c_2|$. $\sigma, \mathbf{c}, \mathbf{r}, \underline{w}$ are fixed and known.

(C2.4) Let $f_0 = w_{10}f_{10} + w_{20}f_{20} \in \mathcal{F}_{2,\sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$.

(C2.5) Consider an MDPM prior Π on $\mathcal{F}$ satisfying (A2.1), (A2.2) and (A2.3) with $K = 2$.

Theorem 2.7.2. *Let $\mathcal{F}$, f_0 , Π be specified as in (C2.3), (C2.4), and (C2.5). Denote by P_{f_0} the probability measure with density f_0 . Let $X^{(n)} := (X_1, \dots, X_n)$ be n i.i.d. samples from P_{f_0} . Let $\Pi^{(n)}(\cdot | X^{(n)})$ denote the posterior distribution induced by the prior Π . For $i = 1, 2$, if $\epsilon_n \geq n^{-\frac{c_i}{\log \log n}}$ for some positive constant c_i , then we have*

$$\Pi^{(n)}(\|f_i - f_{i0}\|_1 > M_n \epsilon_n | X_1, \dots, X_n) \rightarrow 0$$

in $P_{f_0}^n$ -probability, for every $M_n \rightarrow \infty$.

Remark. *Because $\log \log n$ grows extremely slowly and may be treated as almost constant as $n \rightarrow \infty$, ϵ_n is almost a polynomial rate. This rate is of the same order as that established for component density estimation in [Tai and Aragam \(2023\)](#).*

A critical portion of this proof requires an existence of test argument. To establish the existence of such a test with desired controls on its type I and type II errors, we need a component density estimator $\widehat{f}_i$, and establish an upper bound on $\|\widehat{f}_i - f_i\|_1$ that vanishes as the sample size $n \rightarrow \infty$. The estimator is due to [Tai and Aragam \(2023\)](#), who also provided a convergence analysis under the l_1 norm $\|\widehat{f}_i - f_i\|_1$. For simplicity, they fixed $r_1 = r_2 = \frac{1}{2}$, and $\sigma = 1$, and provided only the order

of the convergence rate. For our purposes we extend their results to the setting where $r_i \in (0, +\infty)$, $i = 1, 2$, $\sigma \in (0, +\infty)$, and we derive the convergence rate with explicit constants. This refinement is useful, as it allows us to establish a concentration inequality for the estimate of the component densities, from which the desired existence of a test will follow.

Lemma 2.7.3. *Let $f \in \mathcal{F}$ and f_0 be as defined in (C2.3) and (C2.4). Fix a constant $M > 2$, and let $\{\epsilon_n\}$ be a positive sequence such that $\lim_{n \rightarrow \infty} \epsilon_n = 0$. For any component $i = 1$ (or 2), there exists a test function $\phi_n : \mathbb{R}^n \rightarrow \{0, 1\}$ such that for sufficiently large n and small ϵ_n ,*

$$E_{f_0}[\phi_n] \leq \exp \left(-\tilde{c} n^{4/5} \left(\frac{\epsilon_n}{\bar{c}_{1i}} \right)^{c_{2i} \log \log \left(\frac{\bar{c}_{1i}}{\epsilon_n} \right) + c_{3i}} \right),$$

and

$$\sup_{f \in \mathcal{F}: \|f_i - f_{i0}\|_1 > M\epsilon_n} E_f[1 - \phi_n] \leq \exp \left(-\tilde{c} n^{4/5} \left(\frac{(M-1)\epsilon_n}{\bar{c}_{1i}} \right)^{c_{2i} \log \log \left(\frac{\bar{c}_{1i}}{(M-1)\epsilon_n} \right) + c_{3i}} \right),$$

where $\tilde{c}$ is a positive absolute constant, $\bar{c}_{1i}$ is a positive constant depending on $i, r, \sigma, r_1, r_2, \underline{w}$, c_{2i} is a positive constant depending on i, r, σ, r_1, r_2 , and c_{3i} is a constant depending on i, r, σ, r_1, r_2 .

In Appendix A.5, we detail the construction of such a component density estimator $\widehat{f}_i$, and in Appendix A.6, we give an upper bound of $\|\widehat{f}_i - f_i\|_1$, which is a refinement of a result of [Tai and Aragam \(2023\)](#). In Appendix A.7, we further derive the tail bound of $\widehat{f}_i$. In Appendix A.8, we establish the existence of a test function ϕ_n with desired controls on type I error $E_{f_0}[\phi_n]$ and type II error $E_f[1 - \phi_n]$. Note that this test function ϕ_n cannot be used in practice, but the existence of such a test function is sufficient for the proof of the main theorem (Theorem 2.7.2) on the posterior contraction of the component density f_i .

2.8 Applications

2.8.1 Disentangling astronomical sources

XMM-Newton dataset. This dataset ([Jones et al. \(2015\)](#)) consists of X-ray events from the nearby star pair FK Aqr and FL Aqr (a visual binary) observed by ESA’s XMM-Newton telescope. The detectors recorded about 540,000 individual X-ray photons (a total of 758,869 events including noise) over a 47 ks (~ 13 hr) exposure. For each photon, we have its position (X, Y) on the detector and a spectral energy value. The two sources partly overlap in the image and the background is weak, making this a clean test case for methods that aim to separate blended sources.

Mixture models have been widely used to model X-ray events originating from different astronomical sources. [Jones et al. \(2015\)](#) modeled the spatial information $(X$ and $Y)$ using a mixture of King profiles. This and subsequent work by [Meyer et al. \(2021\)](#) extended this spatial model by incorporating additional spectral and temporal information. In this section, we employ a multivariate MDPM with fixed regions to model the spatial distribution of X-ray events. The regions are held fixed to avoid identifiability issues and are pre-defined. The two regions are determined based on the kernel density estimate (KDE) of the observed data and serve as approximate estimates of the locations of the two astronomical sources. The two rectangle regions are illustrated in the left panel of Figure 2.4. In addition to the two source components, we include a third component—a uniform distribution over the observation window—to capture the diffuse background noise.

We compare our density estimation result from MDPM (left panel), Kernel density estimate (middle panel) and Mixture of King’s profiles from [Jones et al. \(2015\)](#) (right panel) in Figure 2.4. King’s profile density is a parametric model and assumes the regular elliptical contours of each astronomical source density, however, from its comparison with KDE we know that it failed to capture the tail structure of real data sharply. In contrast, MDPM can closely recover the real data density.

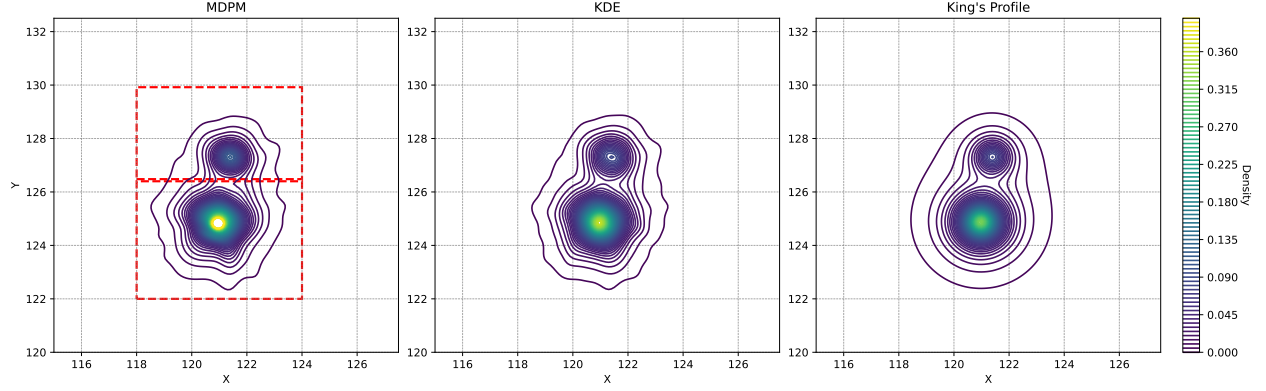


Figure 2.4: Density estimation contours from MDPM, KDE, and mixture of King's profiles, respectively.

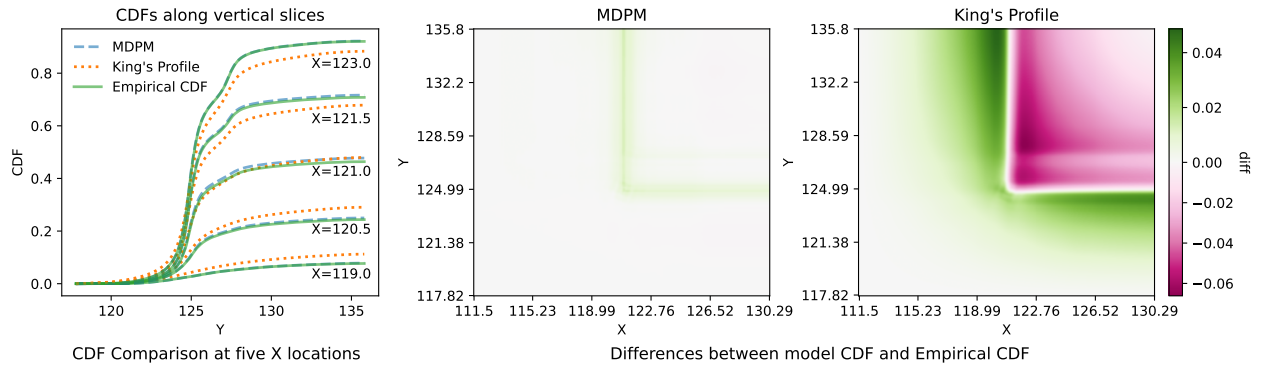


Figure 2.5: Comparison of CDFs.

Weights	MDPM		King's Profile	
	posterior mean	68% CI	posterior mean	68% CI
w_1 (lower component)	0.643	(0.641, 0.646)	0.717	(0.716, 0.718)
w_2 (upper component)	0.229	(0.227, 0.231)	0.182	(0.181, 0.182)
w_0 (background noise)	0.128	(0.127, 0.128)	0.102	(0.101, 0.102)

Table 2.1: Comparison of the estimated component weights. The 68% credible intervals are computed using 16% and 84% posterior quantiles.

Figure 2.5 compares the cumulative distribution functions (CDFs) obtained from the MDPM estimator, the mixture of King's profiles estimator, and the empirical CDF. For any location (x, y) ,

the estimated CDF is defined as $\hat{F}(x, y) = \Pr_{\hat{f}}(X \leq x, Y \leq y)$. In the left panel of Figure 2.5, we examine five X values—119.0, 120.5, 121.0, 121.5, and 123.0—and, for each value, plot the CDF as Y varies from 117.5 to 135.0. The resulting curves are shown using distinct colors and line styles for clarity. Across all five X -values, the MDPM estimator closely tracks the empirical CDF, whereas the Mixture of King’s profiles consistently overestimates the tail mass. The middle and right panels of Figure 2.5 display the pointwise differences between each model’s estimated CDF and the empirical CDF over the spatial domain, with X ranging from 111.5 to 130.29 and Y ranging from 117.8 to 135.8. The MDPM estimator matches the empirical CDF closely across most of the plane, with only minor deviations near the centers of the two astronomical sources. In contrast, the mixture of King’s profiles exhibits noticeably larger discrepancies throughout the region. These results indicate that the MDPM estimator captures the subtle tail behavior of both sources much more effectively. The estimated component weights are compared in Table 2.1.

2.8.2 Analysis of oceanic whitetip shark acceleration data

Studying the dynamic patterns of an animal’s behavioral states from animal movement metrics is a topic of interest in ecology. The Overall Dynamic Body Acceleration (ODBA) data were collected for an oceanic whitetip shark at a rate of 16 Hz over 4 days ([Langrock et al. \(2018\)](#)). The movements of the shark were recorded by accelerometers along three axes at very fine temporal scales (multiple times per second) over multiple days. The acceleration measurements from the three axes were then combined into a single representative metric. A line of work ([Langrock et al. \(2015, 2018\)](#); [Chen and Finkenstädt \(2024\)](#)) has modeled such data with spline-based hidden Markov models (HMM) to infer latent movement states (e.g., resting, foraging or migrating). [Chen and Finkenstädt \(2024\)](#) uses a smaller dataset over a timespan of 24 hours, and the raw ODBA values are averaged over nonoverlapping windows of length 15 seconds and log transformed (IODBA), resulting in a total of 5,760 observations. Prior work typically assumes that latent movement states follow a time-homogeneous Markov chain with a fixed transition matrix. This implies that, as sample size increases, the marginal distribution of the states approaches the chain’s stationary distribution

(assuming ergodicity). To assess our approach on real data, we apply the MDPM to the raw observations, estimate a component (emission) density for each movement state, and compare the results with those from existing HMM-based analyses.

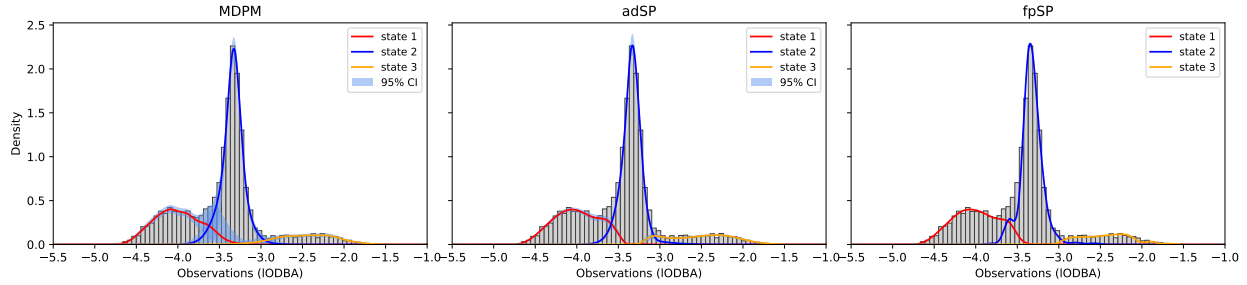


Figure 2.6: Left, middle, and right panels show the IODBA histogram with the estimated weighted emission densities from MDPM, adSP (Chen and Finkenstädt, 2024), and fpSP (Langrock et al., 2015), respectively.

We analyze the IODBA data with MDPM and set $K = 3$, as justified on biological grounds by Langrock et al. (2018); Chen and Finkenstädt (2024). In Figure 2.6, the left panel shows the MDPM posterior mean emission densities, weighted by estimated state occupancy, with pointwise 95% credible intervals; the middle panel reports the corresponding estimates from the Bayesian spline-based HMM (adSP; Chen and Finkenstädt (2024)); and the right panel shows the earlier P-spline HMM (fpSP; Langrock et al. (2015)). Although our approach does not exploit the temporal structure of the HMM and relies only on the marginal distribution of the observations, it produces state-dependent emission density estimates that are very similar to those obtained by methods that leverage the Markov structure. However, because only marginal information is used, the pointwise 95% confidence intervals in regions where the state densities overlap tend to be wider than those produced by adSP. These results validate our approach’s ability to recover the underlying subpopulation distributions accurately using only population-level observations.

2.9 Conclusions

In this chapter, we develop a Bayesian framework with theoretical guarantees to study finite mixture models of nonparametric components. Our approach employs a mixture of Dirichlet process mixtures prior and leverages conjugacy to enable an efficient MCMC algorithm for posterior inference. Through several simulation studies and real-data applications in both univariate and multivariate settings, we demonstrate that the proposed method achieves strong empirical performance and scalability. To address the fundamental challenge of identifiability in nonparametric mixture models while allowing for overlapping component support, we introduce a novel separation condition defined in terms of distances between connected regions in the support of the latent mixing measure. Under this framework, we establish posterior contraction rates for both the overall mixture density and the individual component densities. In particular, we show that the contraction rate for component density estimation attains an almost polynomial order of convergence, representing a substantial improvement over the logarithmic convergence rates typically associated with deconvolution-based estimation of mixing measures. To our best knowledge, this work provides the first theoretical guarantees for a practical Bayesian method that consistently estimates nonparametric component densities within a finite mixture modeling framework.

CHAPTER 3

Learning Nonparametric Components Beyond Convexly Separable Latent Support

3.1 Introductions

In Chapter 2, we developed the MDPM framework for mixture models with complex component distributions. The formulation in Chapter 2 focuses on the setting in which the supports of the latent mixing measures associated with different components are contained in disjoint convex sets. In practice, however, especially in multivariate settings, the latent supports of different components may be much more intertwined and need not be contained in disjoint convex sets. In this chapter, we show that the MDPM framework extends beyond the setting of Chapter 2 and can be used to model this more general class of latent support geometries in $\mathbb{R}^m$, $m \geq 2$.

Definition 3.1.1 (Convex separability). *Let S_1, S_2 be two nonempty subsets of $\mathbb{R}^m$. We say that S_1 and S_2 are convexly separable, or equivalently, that they have convex separability, if there exist disjoint convex sets $I_1, I_2 \subseteq \mathbb{R}^m$ such that $S_1 \subseteq I_1$ and $S_2 \subseteq I_2$.*

Note that convex separability does not require S_1 and S_2 themselves to be convex. In Section 2.5, we considered mixture models in $\mathbb{R}^m$ under Setting (S2.3), in which the sets $I_1, \dots, I_K$ were modeled as axis-aligned hypercubes. This modeling of $I_1, \dots, I_K$ implies that the μ -marginal supports of the mixing measures for different components are convexly separable. In this chapter, we move beyond this restriction by allowing the μ -marginal supports of different component mixing measures not

to be convexly separable, while still assuming that the number of mixture components K is finite, fixed, and known. For convenience, we restate Setting (S2.3) below.

(S2.3) The component distribution class $\tilde{\mathcal{P}}$ consists of location-scale mixtures of normals on $\mathbb{R}^m$, as defined in (2.7). The mixing measure H satisfies: (i) its μ -marginal $\pi_\mu \# H$ has bounded support on $\mathbb{R}^m$; and (ii) its Σ -marginal $\pi_\Sigma \# H$ is supported on covariance matrices with eigenvalues uniformly bounded above and away from 0. $\mathcal{P}(\tilde{\mathcal{P}})$ are probability measures supported on finitely many components satisfying the separation condition: for any two components $G_i = p_{H_i}$, and $G_j = p_{H_j}$, $i \neq j$, we can find two disjoint bounded connected sets $I_i, I_j \subset \mathbb{R}^m$ such that $\text{supp}(\pi_\mu \# H_i) \subseteq I_i$, and $\text{supp}(\pi_\mu \# H_j) \subseteq I_j$.

From the discussion in Section 2.5, the main bottleneck in modeling the sets I_k with more flexible shapes is the need to efficiently sample from a truncated normal distribution supported on I_k . Existing approaches can handle supports defined by linear or quadratic inequality constraints, but they do not extend efficiently to more general shapes, especially nonconvex sets. If the μ -marginal supports of two component mixing measures H_i and H_j are convexly separable, then we can always find two polytopes I_i and I_j such that $\text{supp}(\pi_\mu \# H_i) \subseteq I_i$, and $\text{supp}(\pi_\mu \# H_j) \subseteq I_j$. Modeling these latent supports within polytopes is a natural generalization of the axis-aligned hypercubes used in Section 2.5. The main challenge, however, is how to extend this approach to the case where the μ -marginal supports are not convexly separable.

Although directly modeling a nonconvex set and sampling from a truncated normal distribution supported on it can be difficult, a feasible alternative is to cover the latent support by many small overlapping squares. In principle, any bounded nonconvex set can be covered by a union of squares of appropriate size. This motivates modeling each component density itself as a mixture of Dirichlet Process mixtures, where each Dirichlet Process has a base measure supported on one of these squares. Following this idea, the remaining challenge is to efficiently learn the latent support and construct a covering of it by squares of appropriate size. In this chapter, we introduce a method that first builds such a square covering based on the clustering output of the HDBSCAN algorithm and then learns the component density using an MDPM defined on top of this covering.

3.2 Identifiability

In this section, we discuss the identifiability of multivariate mixture models. We first recall the definition of a location mixture of normals on $\mathbb{R}^m$: For any $\boldsymbol{\mu} \in \mathbb{R}^m$, $\Sigma \in \mathbb{R}^{m \times m}$, $\Sigma \succ 0$, let $g_{\boldsymbol{\mu}, \Sigma}$ denote the m -dimensional Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix Σ : $g_{\boldsymbol{\mu}, \Sigma}(\mathbf{x}) = (\sqrt{2\pi})^{-m} |\Sigma|^{-1/2} \exp(-(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}))$. We say that $p(\mathbf{x})$ is a location mixture of normals on $\mathbb{R}^m$ if there exists a positive definite matrix Σ and a mixing distribution H on $\mathbb{R}^m$ such that

$$p(\mathbf{x}) = p_{\Sigma, H}(\mathbf{x}) := \int g_{\boldsymbol{\mu}, \Sigma}(\mathbf{x}) dH(\boldsymbol{\mu}). \quad (3.1)$$

Let I denote a nonempty subset of $\mathbb{R}^m$. We now define the connected-component separation of I as follows.

Definition 3.2.1 (Connected-component separation). *For a nonempty subset $I \subseteq \mathbb{R}^m$, S_I is called the connected-component separation of I if $I = \cup_{I_i \in S_I} I_i$, with each $I_i \in S_I$ non-empty and connected, and, whenever $|S_I| > 1$, the union $\cup_{I_i \in T} I_i$ is disconnected for every subset $T \subseteq S_I$ with $|T| \geq 2$.*

By the definition of connectedness in Section 2.2 and the further discussions in Section 2.6, the connected-component separation of any nonempty set I is unique. Next, let d be a metric on $\mathbb{R}^m$. For $\epsilon > 0$, we define the ϵ -expansion of I as follows:

Definition 3.2.2 (ϵ -expansion). *For a nonempty subset $I \subseteq \mathbb{R}^m$, the ϵ -expansion of I , denoted by I^ϵ , is defined by*

$$I^\epsilon := \{x \in \mathbb{R}^m \mid \text{dist}(x, I) \leq \epsilon\},$$

where $\text{dist}(x, I) := \inf_{y \in I} d(x, y)$.

If we further assume that $|S_I|$ is finite for a given set I , then for $1 \leq k \leq |S_I|$, we define $\epsilon(k)$ for I to be the infimum of the positive values ϵ such that the ϵ -expansion I^ϵ has exactly k elements in its connected-component separation S_{I^ϵ} :

$$\epsilon(k) := \inf \{\epsilon > 0 \mid |S_{I^\epsilon}| = k\}. \quad (3.2)$$

Note that for some $1 \leq k \leq |S_I|$, the set $\{\epsilon > 0 \mid |S_{I^\epsilon}| = k\}$ may be empty, in which case $\epsilon(k)$ does not exist.

In Section 2.6, we established the identifiability of the mixture model (2.13) on $\mathbb{R}$ under two scenarios. Condition (C2.1) assumes that the support sets $I_1, \dots, I_K$ of the component mixing measures are fixed and known. Condition (C2.2), by contrast, allows these sets to be unknown, but requires the supports of the mixing measures from different components to be well separated, in the sense that the largest within-component distance must be strictly smaller than the smallest between-component distance, where both distances are defined through connected-set distances. In this section, we extend the identifiability conditions to mixture models on $\mathbb{R}^m$. The multivariate analogue of (C2.1) is immediate, but that of (C2.2) is less straightforward. In the univariate setting, the within-component distance was defined using the notion of neighboring pairs. In the multivariate setting, however, such a notion is not naturally available. To address this issue, we use the ϵ -expansion introduced earlier. The basic idea is that, as ϵ increases, the support set I gradually expands and its disjoint connected subsets begin to merge. Under a suitable separation condition, the connected subsets belonging to the same mixture component should merge at smaller values of ϵ than those belonging to different mixture components.

We now formalize this multivariate setting and introduce the class of mixture models to be studied. For a finite, fixed, and known $K \in \mathbb{N}_+$, let $I_1, \dots, I_K$ be pairwise disjoint compact connected sets on $\mathbb{R}^m$. Consider the mixture model

$$\begin{aligned} f &= \sum_{k=1}^K w_k f_k, \\ \text{s.t. } f_k(\mathbf{x}) &= \int g_{\mu, \Sigma}(\mathbf{x}) dH_k(\boldsymbol{\mu}), \forall \mathbf{x} \in \mathbb{R}^m, \\ \text{supp}(H_k) &\subseteq I_k, \quad k = 1, \dots, K, \\ (w_1, \dots, w_K) &\in \Delta^{K-1}. \end{aligned} \tag{3.3}$$

Note that $H_k(\boldsymbol{\mu})$ can be any probability measure supported within I_k . We establish identifiability of (3.3) under the following two scenarios:

(C3.1) The sets $I_1, \dots, I_K$ are fixed and known.

(C3.2) The number of components K is finite, fixed, and known. Let $I := \cup_{k=1}^K \text{supp}(H_k)$. Assume that $\epsilon(K)$, as defined in (3.2), exists for I and K , and that the K sets in $S_{I^{\epsilon(K)}}$ can be ordered as $I_1, \dots, I_K$ so that $\text{supp}(H_k) \subseteq I_k$ for $k = 1, \dots, K$.

We now prove that the mixture model in (3.3) is identifiable under either condition (C3.1) or (C3.2).

Theorem 3.2.3. *For any f defined by (3.3), if either (C3.1) or (C3.2) holds, then the representation*

$$f = \sum_{k=1}^K w_k f_k$$

subject to the conditions in (3.3) is unique. That is, $\Sigma \succ 0$, the component densities $f_1, \dots, f_K$, and the weight vector $(w_1, \dots, w_K) \in \Delta^{K-1}$ are uniquely determined up to a permutation of the labels.

Proof. We first show that for any multivariate function $h(\mathbf{x})$ that can be represented as $h(\mathbf{x}) = \int g_{\mu, \Sigma}(\mathbf{x}) dH(\mu)$, where H has compact support, if there also exist $\tilde{\Sigma}$ and $\tilde{H}$ such that $h(\mathbf{x}) = \int g_{\mu, \tilde{\Sigma}}(\mathbf{x}) d\tilde{H}(\mu)$, then we must have $\Sigma = \tilde{\Sigma}$ and $H = \tilde{H}$.

For any probability measure F on $\mathbb{R}^m$ and any vector $\beta \in \mathbb{R}^m$, $\beta \neq \mathbf{0}$, let $F_\beta := (\beta^T)_\# F$ denote the pushforward of F under the linear map $\mu \mapsto \beta^T \mu$. Equivalently, for any Borel set $A \subseteq \mathbb{R}$,

$$F_\beta(A) = F\left(\{\mu \in \mathbb{R}^m \mid \beta^T \mu \in A\}\right).$$

Using this notation, for any $A \subseteq \mathbb{R}$, the pushforward measure h_β is given by

$$\begin{aligned} h_\beta(A) &= \int_{\mathbb{R}^m} \mathbf{1}\{\beta^T \mathbf{x} \in A\} h(\mathbf{x}) d\mathbf{x} \\ &= \int_{\mathbb{R}^m} \mathbf{1}\{\beta^T \mathbf{x} \in A\} \int g_{\mu, \Sigma}(\mathbf{x}) dH(\mu) d\mathbf{x} \\ &= \int \left(\int_{\mathbb{R}^m} \mathbf{1}\{\beta^T \mathbf{x} \in A\} g_{\mu, \Sigma}(\mathbf{x}) d\mathbf{x} \right) dH(\mu). \end{aligned}$$

Now, since under the multivariate normal distribution $g_{\mu, \Sigma}(\mathbf{x})$, the scalar $\mathbf{x}' = \boldsymbol{\beta}^T \mathbf{x}$ has density $g_{\boldsymbol{\beta}^T \mu, \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}}(\mathbf{x}')$, it follows that

$$h_{\boldsymbol{\beta}}(A) = \int \left(\int_{\mathbb{R}} \mathbf{1}\{\mathbf{x}' \in A\} \cdot g_{\boldsymbol{\beta}^T \mu, \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}}(\mathbf{x}') d\mathbf{x}' \right) dH(\boldsymbol{\mu}),$$

therefore, under the change of variable $\boldsymbol{\mu}' = \boldsymbol{\beta}^T \boldsymbol{\mu}$, we obtain

$$\begin{aligned} h_{\boldsymbol{\beta}}(A) &= \int \left(\int_A g_{\boldsymbol{\mu}', \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}}(\mathbf{x}') d\mathbf{x}' \right) dH_{\boldsymbol{\beta}}(\boldsymbol{\mu}') \\ &= \int_A \left(\int g_{\boldsymbol{\mu}', \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}}(\mathbf{x}') dH_{\boldsymbol{\beta}}(\boldsymbol{\mu}') \right) d\mathbf{x}'. \end{aligned}$$

Therefore, $h_{\boldsymbol{\beta}}$ has density

$$h_{\boldsymbol{\beta}}(\mathbf{x}') = \int g_{\boldsymbol{\mu}', \boldsymbol{\beta}^T \Sigma \boldsymbol{\beta}}(\mathbf{x}') dH_{\boldsymbol{\beta}}(\boldsymbol{\mu}'). \quad (3.4)$$

Similarly, since $h(\mathbf{x}) = \int g_{\mu, \tilde{\Sigma}}(\mathbf{x}) d\tilde{H}(\mu)$, we likewise obtain

$$h_{\boldsymbol{\beta}}(\mathbf{x}') = \int g_{\boldsymbol{\mu}', \boldsymbol{\beta}^T \tilde{\Sigma} \boldsymbol{\beta}}(\mathbf{x}') d\tilde{H}_{\boldsymbol{\beta}}(\boldsymbol{\mu}') \quad (3.5)$$

Since H has compact support, it follows that, for any given $\boldsymbol{\beta} \in \mathbb{R}^m$, the pushforward measure $H_{\boldsymbol{\beta}}$ also has compact support. In Theorem 2.6.1 we proved that the two representations (3.4) and (3.5) of the univariate density $h_{\boldsymbol{\beta}}$ implies that $\boldsymbol{\beta}^T \Sigma \boldsymbol{\beta} = \boldsymbol{\beta}^T \tilde{\Sigma} \boldsymbol{\beta}$ and $H_{\boldsymbol{\beta}} = \tilde{H}_{\boldsymbol{\beta}}$. Therefore we must have $\Sigma = \tilde{\Sigma}$. Furthermore, by the Cramér–Wold theorem ([Cramér and Wold, 1936](#)), the fact that $H_{\boldsymbol{\beta}} = \tilde{H}_{\boldsymbol{\beta}}$ for every $\boldsymbol{\beta} \in \mathbb{R}^m$ implies that $H = \tilde{H}$.

It remains to show that, under condition (C3.1) or (C3.2), the measure H admits a unique decomposition into K components. This is immediate under condition (C3.1)). Under condition (C3.2), we may increase ϵ until the ϵ -expansion of $\text{supp}(H)$ has exactly K connected components. Labeling these K components by $I_1, \dots, I_K$, and letting w_k be the probability mass that H assigns to I_k , for $k = 1, \dots, K$, then this yields a unique representation of f as a mixture of K components. $\square$

3.3 Modeling latent supports beyond convex separability

In this section, we develop a method to learn the mixture models as specified in Setting (S2.3), in which the μ -marginal supports of two different component mixing measures, $\text{supp}(\pi_\mu \# H_i)$, and $\text{supp}(\pi_\mu \# H_j)$ may not be convexly separable. We first review the HDBSCAN algorithm (Campello et al., 2013). Next, we show how the clustering results produced by HDBSCAN can be used to estimate the μ -marginal support of the latent mixing measure. Finally, building on this learned support, we construct a mixture of Dirichlet process mixtures to estimate component densities.

3.3.1 The HDBSCAN algorithm

HDBSCAN, introduced by Campello et al. (2013), is a density-based hierarchical clustering method that performs remarkably well at identifying significant clusters lying on connected high-density regions, with different clusters separated by regions of low density. Unlike methods based on a global density threshold (Ester et al., 1996; Hinneburg and Keim, 2003), HDBSCAN can effectively characterize clusters with very different density levels by selecting the most stable clusters from the hierarchical clustering tree. We now briefly review the HDBSCAN algorithm and the core concepts underlying it.

Let $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, $\mathbf{x}_i \in \mathbb{R}^m$, for $i = 1, \dots, n$ be a dataset of n samples, and let d be a metric on $\mathbb{R}^m$. m_{pts} is a positive integer tuning parameter of the algorithm.

Definition 3.3.1 (Core distance, Campello et al. (2013)). The core distance of an object $\mathbf{x}_i \in X$ with respect to m_{pts} , denoted by $d_{\text{core}}(\mathbf{x}_i)$, is the distance from $\mathbf{x}_i$ to its m_{pts} -nearest neighbor (including $\mathbf{x}_i$).

Definition 3.3.2 (ϵ -core object, Campello et al. (2013)). An object $\mathbf{x}_i \in X$ is called an ϵ -core object if

$$d_{\text{core}}(\mathbf{x}_i) \leq \epsilon.$$

Equivalently, $\mathbf{x}_i$ is an ϵ -core object for every value of ϵ greater than or equal to its core distance

with respect to m_{pts} .

Definition 3.3.3 (Mutual reachability distance, Campello et al. (2013)). The mutual reachability distance between two objects $\mathbf{x}_i$ and $\mathbf{x}_j$ in X with respect to m_{pts} is defined by

$$d_{\text{mreach}}(\mathbf{x}_i, \mathbf{x}_j) = \max \{d_{\text{core}}(\mathbf{x}_i), d_{\text{core}}(\mathbf{x}_j), d(\mathbf{x}_i, \mathbf{x}_j)\}.$$

Definition 3.3.4 (Mutual reachability graph, Campello et al. (2013)). The mutual reachability graph with respect to m_{pts} , denoted by $G_{m_{pts}}$, is the complete graph whose vertex set is X and whose edge weight between any two objects $\mathbf{x}_i, \mathbf{x}_j \in X$ is given by their mutual reachability distance $d_{\text{mreach}}(\mathbf{x}_i, \mathbf{x}_j)$.

The HDBSCAN hierarchy is constructed as a dendrogram in the following way:

1. Compute a minimum spanning tree (MST) of the mutual reachability graph $G_{m_{pts}}$.
2. Extend the MST to obtain MST_{ext} by adding a self-edge to each vertex, with weight equal to the core distance of the corresponding object.
3. At the root of the hierarchy, all objects are assigned the same label, forming a single cluster, and ϵ is initialized as the largest edge weight in MST_{ext} . The hierarchy is then constructed by gradually decreasing ϵ and, at each step, removing from MST_{ext} all edges whose weights exceed ϵ . After each such removal, every remaining connected component containing at least one edge is treated as a cluster, while vertices that do not belong to any such component are assigned the null label (“noise”). A new level of the hierarchy is recorded whenever new clusters emerge.

After the hierarchical clustering tree is constructed, a cluster stability measure is used to extract the most stable clusters from the tree. The resulting clusters are intended to approximate the density-contour clusters, defined as the maximal connected subsets of the density level set $\{\mathbf{x} | f(\mathbf{x}) \geq \lambda\}$, for a given density $f(\mathbf{x})$ and density level λ . Intuitively, as the density level λ increases, the

density-contour clusters become smaller and smaller until they disappear or split into sub-clusters. A more stable cluster is expected to survive longer from the time it appears until it disappears. To formalize this idea, [Campello et al. \(2013\)](#) introduces the notion of Relative Excess of Mass of a density-contour cluster C_i :

$$E_R(C_i) = \int_{\mathbf{x} \in C_i} (\lambda_{\max}(\mathbf{x}, C_i) - \lambda_{\min}(C_i)) d\mathbf{x}, \quad (3.6)$$

where cluster C_i appears at density level $\lambda_{\min}(C_i)$, disappears at $\lambda_{\max}(C_i)$, and $\lambda_{\max}(\mathbf{x}, C_i) := \min\{f(\mathbf{x}), \lambda_{\max}(C_i)\}$. Then for the resulting cluster estimate $\widehat{C}_i$, defined as the set of samples in X labeled as belonging to C_i , the stability of $\widehat{C}_i$ can be obtained by adapting (3.6):

$$S(\widehat{C}_i) = \sum_{\mathbf{x}_j \in \widehat{C}_i} \left(\frac{1}{\epsilon_{\min}(\mathbf{x}_j, \widehat{C}_i)} - \frac{1}{\epsilon_{\max}(\widehat{C}_i)} \right). \quad (3.7)$$

Here, $\epsilon_{\max}(\widehat{C}_i)$ denotes the value of ϵ at which $\widehat{C}_i$ first appears in the HDBSCAN hierarchy, while $\epsilon_{\min}(\widehat{C}_i)$ denotes the value at which $\widehat{C}_i$ disappears. Moreover, $\epsilon_{\min}(\mathbf{x}_j, \widehat{C}_i) := \max\{d_{\text{core}}(\mathbf{x}_j), \epsilon_{\min}(\widehat{C}_i)\}$, which is the ϵ -value at which $\mathbf{x}_j$ leaves $\widehat{C}_i$. This adaptation from (3.6) to (3.7) is well motivated because, when $x \in \mathbb{R}$, $1/d_{\text{core}}(x)$ is a non-normalized K -NN estimate of the density $f(x)$, with $K = m_{pts}$. In $\mathbb{R}^m$, the corresponding K -NN density estimate is proportional to $1/d_{\text{core}}(x)^m$; hence $1/d_{\text{core}}(x)$ should be interpreted not as an exact density estimate, but as a monotone surrogate for local density, which is sufficient for defining the empirical stability measure used by HDBSCAN.

To extract the most significant cluster, HDBSCAN solves the following optimization problem. Let $\{\widehat{C}_2, \dots, \widehat{C}_\kappa\}$ denote the collection of all clusters that appear in the hierarchical clustering tree, excluding the root cluster $\widehat{C}_1$, and let $S(\widehat{C}_i)$ denote the stability of cluster $\widehat{C}_i$. For each $i = 2, \dots, \kappa$,

introduce a binary variable δ_i , where

$$\delta_i = \begin{cases} 1, & \text{if } \widehat{C}_i \text{ is selected in the final flat clustering,} \\ 0, & \text{otherwise.} \end{cases}$$

HDBSCAN then extracts a flat clustering by solving

$$\begin{aligned} & \max_{\delta_2, \dots, \delta_K} \quad \sum_{i=2}^K \delta_i S(\widehat{C}_i) \\ & \text{subject to} \quad \sum_{j \in I_h} \delta_j = 1, \quad \forall h \in \mathbb{L}, \end{aligned} \tag{3.8}$$

where $\mathbb{L} = \{h \mid \widehat{C}_h \text{ is a leaf cluster}\}$ is the set of indices of leaf clusters, and $I_h = \{j \mid j \neq 1 \text{ and } \widehat{C}_j \text{ lies on the path from } \widehat{C}_h \text{ to the root } \widehat{C}_1\}$. That is, for each leaf-to-root path, exactly one cluster on that path is selected. This guarantees that the selected clusters are non-overlapping and form a flat partition.

The optimization problem in (3.8) can be solved recursively in a bottom-up manner on the cluster tree. Initialize $\delta_2 = \dots = \delta_K = 1$. Then, proceeding from the leaf clusters toward the root, compare the stability $S(\widehat{C}_i)$ of each cluster $\widehat{C}_i$ with the total stability of its children, $\sum_{i_r \in \text{Ch}(i)} S(\widehat{C}_{i_r})$. If $S(\widehat{C}_i) < \sum_{i_r \in \text{Ch}(i)} S(\widehat{C}_{i_r})$, then $\widehat{C}_i$ is not selected: set $\delta_i = 0$, and replace $S(\widehat{C}_i)$ by $\sum_{i_r \in \text{Ch}(i)} S(\widehat{C}_{i_r})$. Otherwise, $\widehat{C}_i$ is retained, and all descendants of $\widehat{C}_i$ are excluded by setting their corresponding δ -values to 0.

3.3.2 Latent support recovery

Recall that our goal is to approximate the true subpopulation distributions of the data by the location-scale mixtures of normals defined in (S2.3). For the i -th component, the μ -marginal support of the latent mixing measure, $\text{supp}(\pi_\mu \# H_i)$, should therefore reflect the high-density region where the i -th component distribution places most of its mass. As described in Section 3.3.1, HDBSCAN aims to recover the most stable density-contour clusters, which is conceptually consistent with the idea of

connected high-density regions. Motivated by this connection, we use the HDBSCAN clustering results as a rough estimate of the latent support and then construct a square covering of that support for later use.

To illustrate the main idea, we consider a half-moon example. The left panel (a) of Figure 3.1 shows the true density of a two-component mixture model, where each component is itself a Gaussian mixture with components located along a half-circle. Equivalently, the support of the μ -marginal of the true mixing measure lies on the union of two one-dimensional manifolds in $\mathbb{R}^2$. The two half-circles are interlocked and therefore not convexly separable. The component weights are also unequal: the upper component has weight 0.7, while the lower component has weight 0.3. The middle panel (b) shows 10,000 samples drawn from the true density and colored by their true component labels, whereas the right panel (c) shows the clustering labels assigned by HDBSCAN to the same samples. HDBSCAN successfully identifies the two primary high-density regions, assigning cluster labels to samples within those regions while treating samples in the component tails and in lower-density transition regions as noise

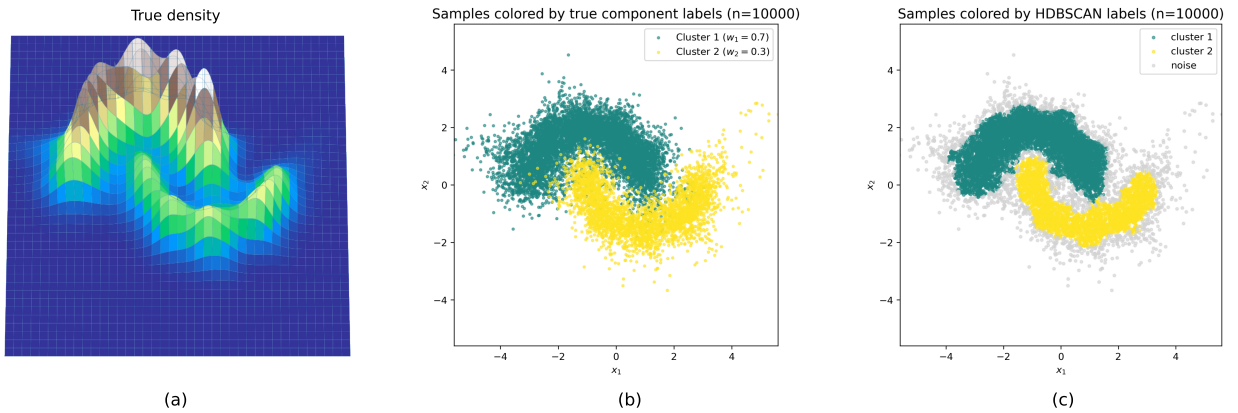


Figure 3.1: Half-moon example. The left panel (a) shows a 3D visualization of the true density. The middle panel (b) shows 10,000 samples drawn from the true density, colored by their true component labels. The right panel (c) shows the same samples, colored by the HDBSCAN cluster labels.

To further estimate the component distributions from the samples, we construct, for each component, a union of squares that approximates its high-density region. The construction must balance two competing goals: using as few squares as possible for computational efficiency, while

ensuring that the resulting union covers a sufficiently large proportion of the samples in each cluster to capture the geometry of the high-density region. To distribute the squares over this region in a structured way, we proceed as follows:

1. **Finding the diameter path of the cluster MST.** For each cluster obtained from HDBSCAN, we first construct a minimum spanning tree on the samples in that cluster using Euclidean distances. We then extract the longest path in this tree, namely its diameter path, and use it as a rough skeleton of the cluster geometry.
2. **Placing squares along the diameter path.** We compute the distances from the cluster samples to the diameter path and choose a predetermined quantile of these distances as the half-side length of the squares. Starting from the midpoint of the diameter path, we place an initial square centered at that location with the chosen half-side length. We then proceed along the path toward both ends, placing additional squares sequentially. Each new center is chosen so that the corresponding square avoids excessive overlap with existing squares, but maintains overlap with the immediately preceding square as close as possible to a prescribed target level, and covers as many previously uncovered samples from the cluster as possible.
3. **Trimming overlapping squares between different clusters.** Finally, to prevent the square unions associated with different clusters from overlapping, we iteratively trim squares involved in inter-cluster overlap. For each overlapping pair, we shrink the offending square or squares to reduce coverage of samples from other clusters while preserving as much within-cluster coverage as possible. After this trimming step, we remove squares that have become too small or redundant.

The detailed algorithms are in Appendix B.1. In Step 1, we use Euclidean distance rather than the mutual reachability distance defined in Definition 3.3.3, because the goal here is to capture the spatial geometry of the cluster in the ambient space $\mathbb{R}^m$. The diameter path then provides a simple one-dimensional skeleton along which square centers can be placed. This path-based construction is most appropriate when the true latent support is close to a one-dimensional manifold, so that a

single path can reasonably trace its overall geometry. When the latent support has higher intrinsic dimension, a single path is generally insufficient to capture the full shape of the region; we therefore discuss a more general strategy for locating the squares later.

For the same samples shown in Figure 3.1, Figure 3.2 displays the learned diameter paths and the square coverings for each cluster before and after trimming. The results show that the learned square unions capture the cluster regions very well.

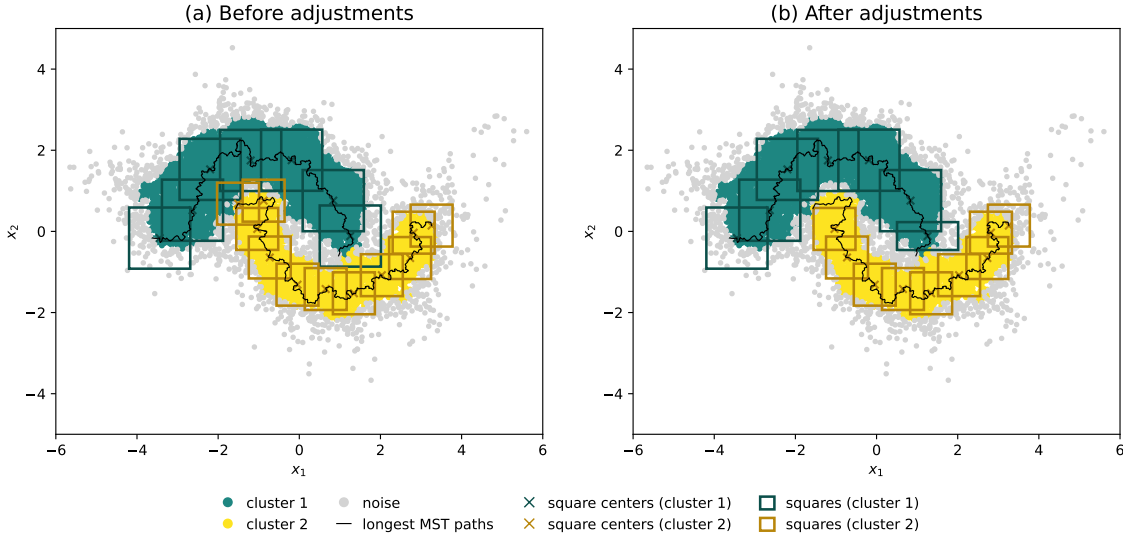


Figure 3.2: Covering squares before and after trimming.

3.3.3 Learning component densities with MDPM

After obtaining a rectangular covering for each cluster, we apply the mixture of Dirichlet process mixtures (MDPM) with fixed regions, specified in (3.9), to estimate the component densities and their corresponding mixture weights. We continue to use the half-moon example for illustration. Suppose that, after trimming, K_1 rectangles are used to cover the first cluster and K_2 rectangles are used to cover the second cluster. These $K_1 + K_2$ rectangles are encoded by two matrices, $\mathbf{c}$ and $\mathbf{r}$, each of dimension $(K_1 + K_2) \times m$, where the k -th row of $\mathbf{c}$, denoted by $\mathbf{c}_k$, gives the center of the k -th rectangle, and the k -th row of $\mathbf{r}$, denoted by $\mathbf{r}_k$, gives its half side-lengths in each coordinate direction. The K_1 rectangles associated with the first cluster are placed in the first K_1 rows of these

matrices, while the K_2 rectangles associated with the second cluster are placed in rows $K_1 + 1$ through $K_1 + K_2$. We then fit the data using an MDPM with $K = K_1 + K_2$ components, specified by the following hierarchical model:

$$\begin{aligned}
\mathbf{w} &= (w_1, \dots, w_K) \sim \text{TDDir}(1/K, \dots, 1/K; \underline{w}\mathbf{1}_K, \overline{w}\mathbf{1}_K), \quad \text{given } K, \\
z_1, \dots, z_n &\stackrel{\text{i.i.d.}}{\sim} \text{Categorical}(w_1, \dots, w_K), \\
H_{k0}(u, \sigma) &\propto g_{c_k, \sigma_0^2}(u) \cdot \mathbf{1}_{[c_k - r_k, c_k + r_k]}(u) \cdot \text{InvGamma}(\sigma^2) \cdot 2\sigma, \\
H_k &\sim \text{DP}(\alpha H_{k0}), \\
G_k = p_{H_k} &= \int g_{u, \sigma} dH_k(u, \sigma), \quad \text{for } k = 1, \dots, K, \\
x_j &\sim G_{z_j}, \quad \text{for } j = 1, \dots, n, \text{ given } G_{1:K}, z_{1:n}.
\end{aligned} \tag{3.9}$$

Finally, we obtain the posterior distributions of $\mathbf{w} = (w_1, \dots, w_K)$ and $G_1, \dots, G_K$. The mixture weights of the two original components are then recovered by aggregating the weights of the rectangles associated with each cluster:

$$\tilde{w}_1 = \sum_{k=1}^{K_1} w_k, \quad \tilde{w}_2 = \sum_{k=K_1+1}^{K_1+K_2} w_k.$$

Correspondingly, the component distributions are given by

$$F_1 = \frac{1}{\tilde{w}_1} \sum_{k=1}^{K_1} w_k G_k, \quad F_2 = \frac{1}{\tilde{w}_2} \sum_{k=K_1+1}^{K_1+K_2} w_k G_k.$$

The true and fitted density contours for the half-moon example are shown in Figure 3.3. The first row shows the true overall density and component densities, while the second row shows the posterior means of the corresponding fitted densities. Our method recovers both the overall density and the component densities well, with only small deviations in the overlapping tail regions. The 95% credible intervals for the component weights also provide estimates that are close to the true weights.

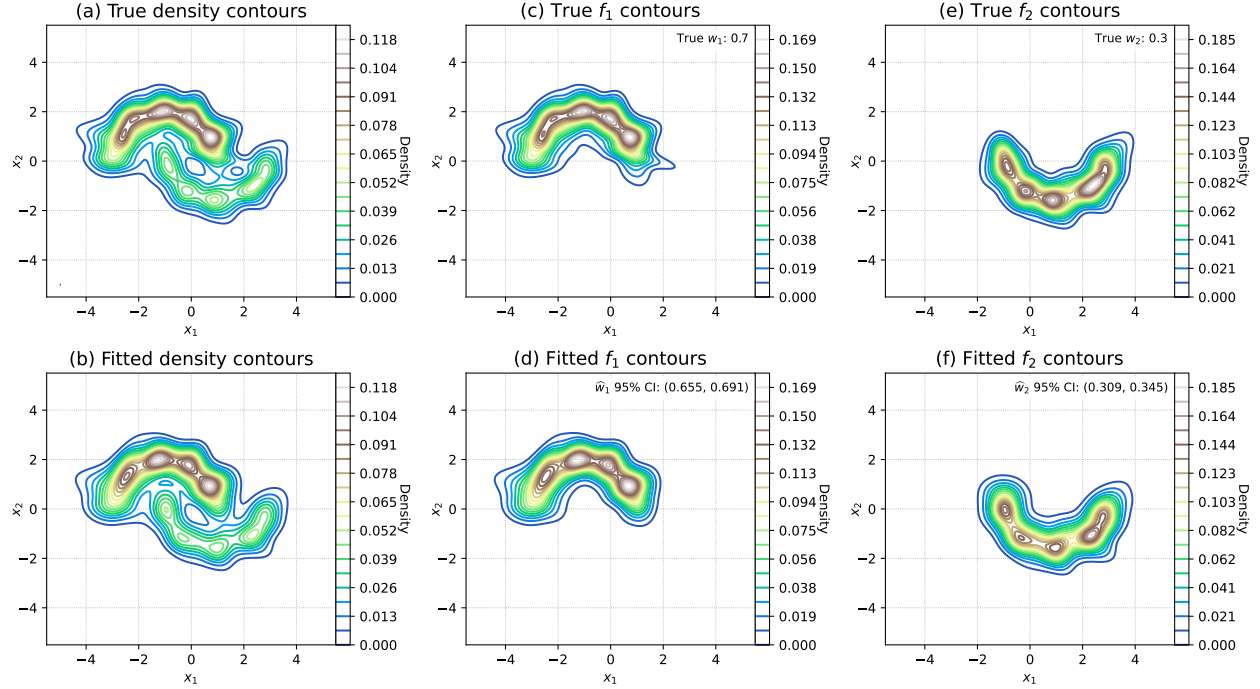


Figure 3.3: Half-moon example. The first row shows the true density contours for the overall density, the first component density f_1 , and the second component density f_2 . The second row shows the corresponding fitted density contours, given by the posterior means. The true component weights and the 95% credible intervals for the posterior weights are displayed in the upper-right corners of the corresponding panels.

3.3.4 Additional simulations

In this section, we present two additional simulation examples, one in $\mathbb{R}^2$ and one in $\mathbb{R}^3$.

First, Figure 3.4 presents a simulation example with concentric circles in $\mathbb{R}^2$. The true distribution is a three-component mixture, with each component concentrating its mass around a circle of a distinct radius. Panels (c) and (d) show 3D visualizations of the true and fitted densities, respectively, while panels (e) and (f) show the corresponding true and fitted density contours on the same density scale. The close agreement between the true and fitted densities demonstrates that our method is able to recover the true density accurately. Panel (a) shows 20,000 samples drawn from the true distribution and colored by their true component labels, while panel (b) shows that our method successfully learns the latent support shape of each component and covers it well with rectangles. The component density comparisons are shown in Figure 3.5. The plots indicate that

both the estimated component densities and the estimated component weights closely match the truth.

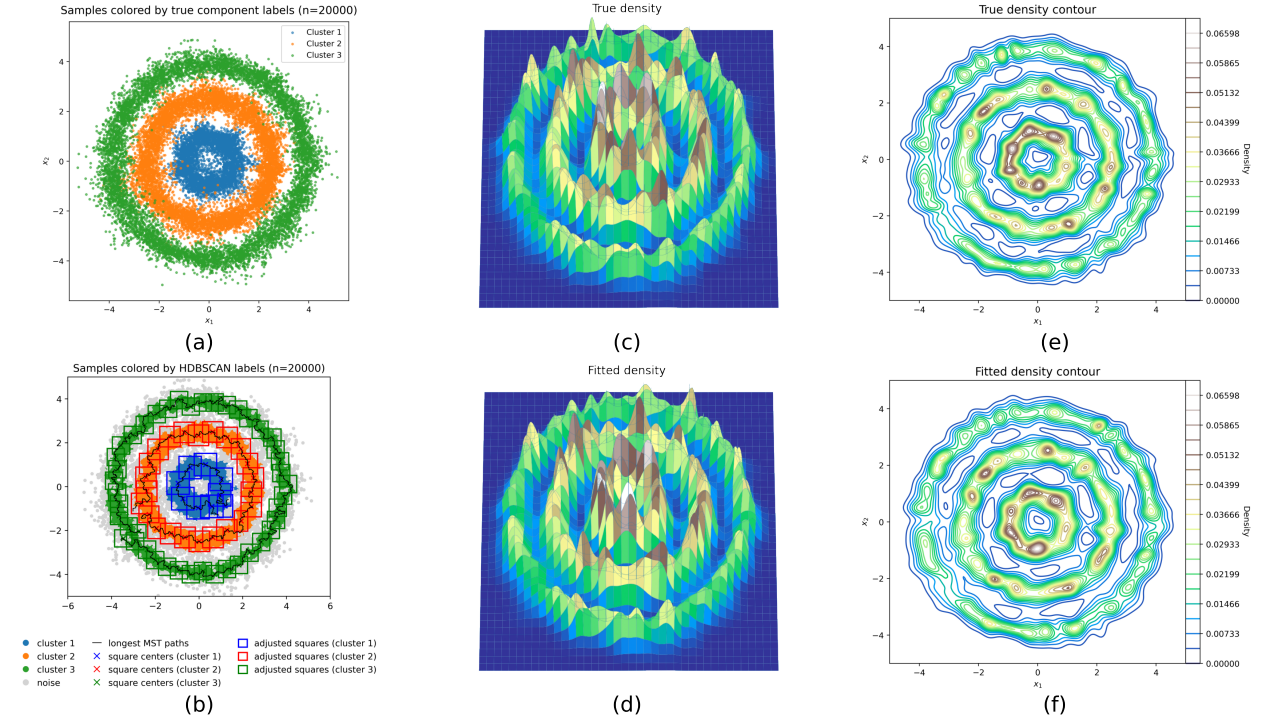


Figure 3.4: Concentric circle example. The true distribution in this example is a mixture of three components. Panel (a) shows the samples colored by their true component labels. Panel (b) shows the same samples colored by the HDBSCAN labels, along with the longest MST path and the covering rectangles for each cluster. Panels (c) and (d) display 3D visualizations of the true and fitted (posterior mean) densities, respectively. Panels (e) and (f) show the corresponding true and fitted (posterior mean) density contours, using the same density scale.

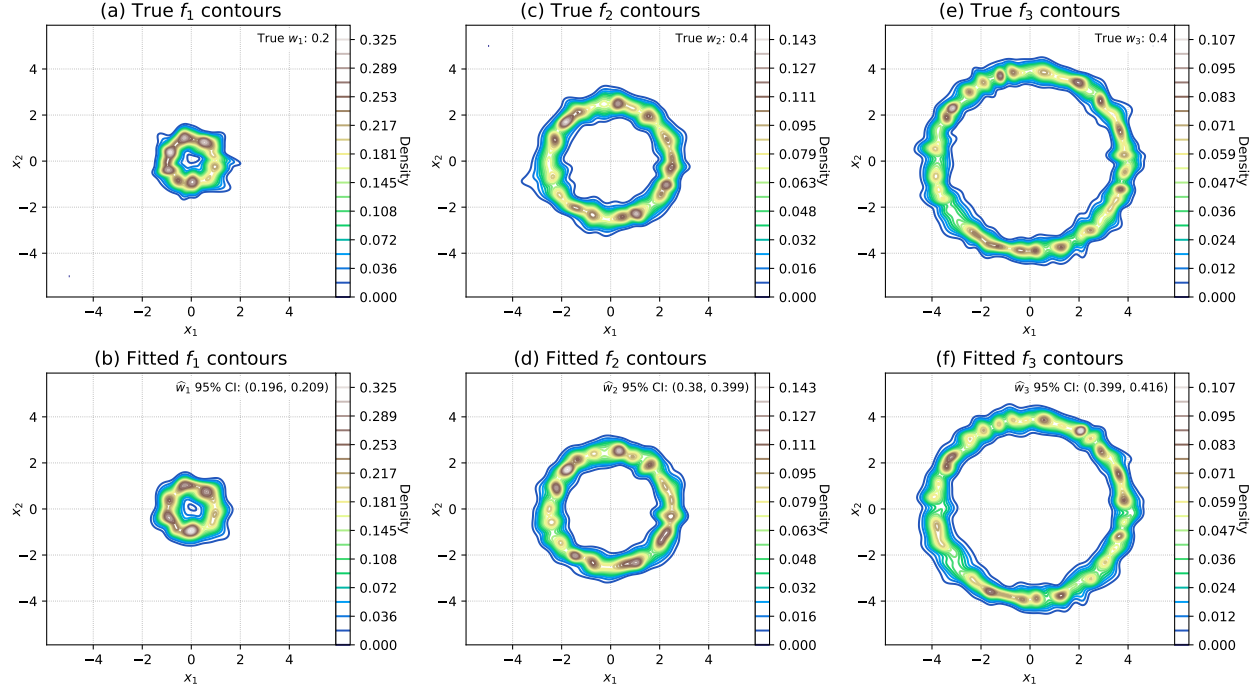


Figure 3.5: Concentric circle example. The first row shows the true density contours of the three components, and the second row shows the corresponding fitted (posterior mean) density contours. The true component weights and the 95% credible intervals for the posterior component weights are shown in the upper-right corners.

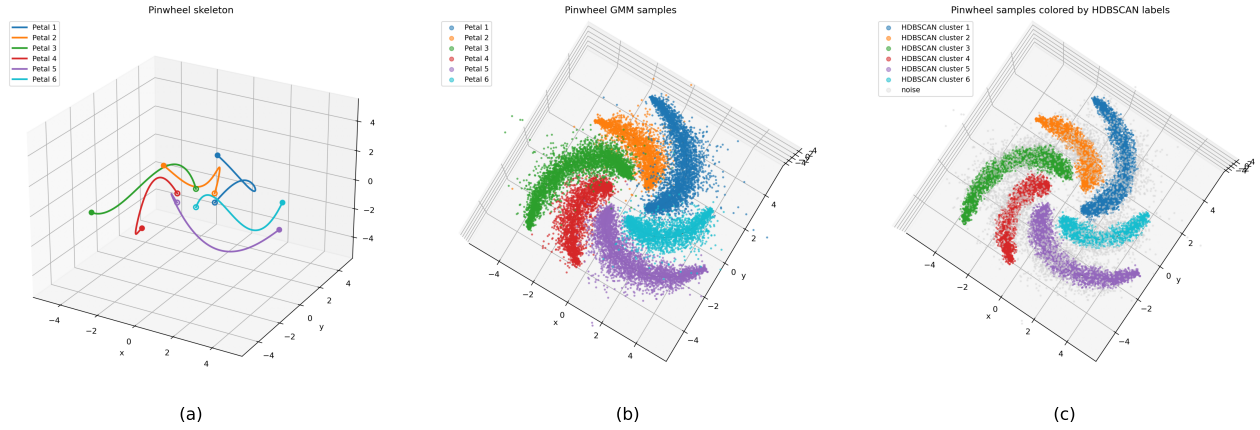


Figure 3.6: Pinwheel example. Panel (a) is the true latent manifold curves that generate the dataset. Panel (b) visualizes the 20,000 samples drawn from the true density, colored by their true component labels. Panel (c) shows the HDBSCAN clustering results.

The second example is a pinwheel dataset in $\mathbb{R}^3$ as shown in Figure 3.6. The true distribution is a mixture model with 6 components. Each component is a Gaussian mixture model with 150

Gaussian kernels located on a one-dimensional curve in $\mathbb{R}^3$. The true manifold curves of the 6 components are shown in Panel (a). Then we generate 20,000 samples from this model, which is shown in Panel (b), colored by their true component labels. Panel (c) shows the HDBSCAN clustering results. The fitted longest MST path and covering cubes are shown in Figure 3.7.

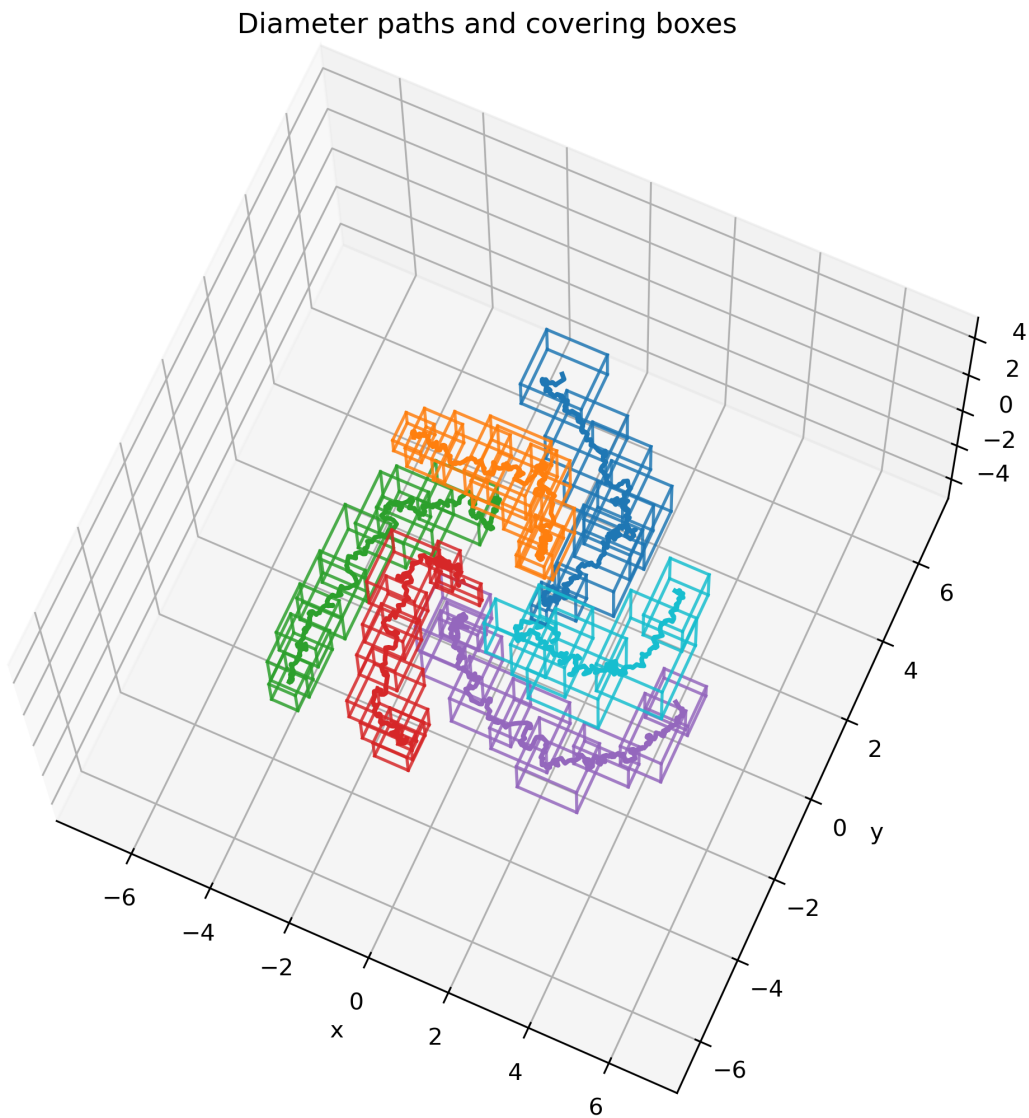


Figure 3.7: The fitted longest MST path and covering cubes for each component is shown in corresponding colors.

Finally, the true density and fitted density are compared in Figure 3.8a and Figure 3.8b. Our approach fits the true component densities quite well, even though some of the tail details are

smoothed, and not fully recovered by the approach.

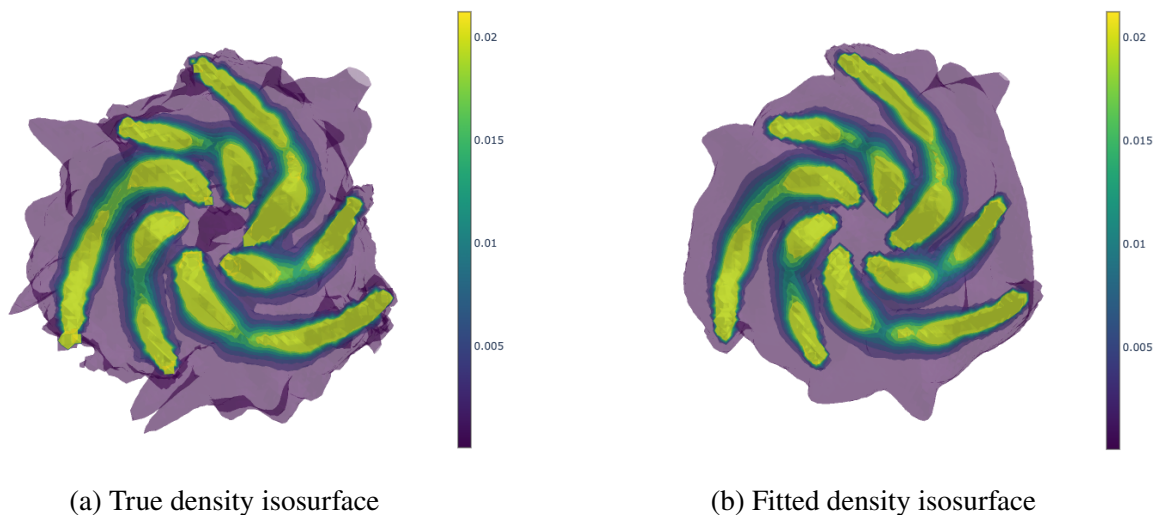


Figure 3.8: Panel (a) is the isosurface plot for the true density, and Panel (b) is the isosurface plot for the fitted density.

3.4 Conclusion

In this chapter, we extend the proposed approach to settings in which the high-density regions associated with different components cannot be contained in disjoint convex sets, a challenge that is especially relevant in multivariate problems. We first establish identifiability conditions for multivariate mixture models in this more general setting. We then develop a method for approximating the geometry of these high-density regions by a union of hypercubes. Building on these hypercube coverings, we construct mixtures of Dirichlet process mixtures that can learn component distributions with complex latent-support geometries. Our simulation results demonstrate strong performance in both $\mathbb{R}^2$ and $\mathbb{R}^3$.

CHAPTER 4

Discussions and Future Directions

4.1 Discussions

4.1.1 What defines a component?

When mixture components are modeled nonparametrically, the meaning of a “component” is no longer self-evident. This raises two related questions: what type of component do we want the model to identify, and can the proposed method recover such components? The first question is inherently application-dependent. In practice, the desired notion of a component may vary across scientific domains and analytical goals. Consequently, there is no fixed or universal definition of a true component; instead, the target notion of a component should be determined by the scientific context and the needs of practitioners. For example, in single-cell RNA-seq analysis, each observation is a high-dimensional molecular profile measured across thousands of genes. Depending on the scientific question, a meaningful component may correspond to cells sharing a common developmental stage, a common cell type, or another biologically relevant condition. The results of this dissertation suggest that component recovery is possible when the underlying components are primarily concentrated in different regions of the sample space, even if their distributions exhibit some overlap in the tails, or when the components are distinguishable through contrasting density levels or scale structures. The proposed approach also offers flexibility in modeling different notions of components according to practitioners’ scientific goals and prior knowledge. When practitioners have a specific notion of components in mind, such as components associated with prespecified re-

regions of the sample space or with distinct location or scale characteristics, the fixed-region MDPM, defined on either μ or σ , can be used to directly estimate the corresponding component densities. When such regional information is unavailable, the MDPM with a repulsive prior can instead learn the mass-concentration regions from the data and estimate the component densities associated with those learned regions. In this sense, the framework allows the definition of a component to be adapted to the needs of the application rather than imposed a priori in a single fixed form. Although in this dissertation, our identifiability results are developed for settings in which each component is modeled as a location mixture of Gaussian kernels, similar results may be established for more general classes of component distributions. We discuss this possible extension in Section 4.2.

4.1.2 How to choose K ?

Throughout this dissertation, we assume that the number of components K is fixed and known. This assumption provides a natural starting point for studying mixtures with nonparametric components. When each component is allowed to have substantial flexibility, some structural information about the mixture must be specified in order for the individual components to be learnable. In this sense, treating K as fixed and known allows us to focus on the central question of how to model and recover flexible component densities.

As discussed earlier, the hierarchical structure of the MDPM resembles that of classical Bayesian Gaussian mixture models. It is therefore possible, at least formally, to extend the framework by placing an additional prior on K to accommodate an unknown number of components. However, the more fundamental question is not only whether such an extension can be constructed, but also whether the learned mixture model has a meaningful interpretation. When the number of components is unknown and the component densities are nonparametric, additional structural constraints are needed for the components to be identifiable and scientifically interpretable. Such constraints are likely to depend on the specific application and should be formulated in response to practical modeling needs. In those settings, existing Bayesian mixture modeling tools developed for parametric models may be incorporated into the MDPM framework to accommodate uncertainty

in K , provided that the target notion of a component is clearly specified.

4.2 Future directions

4.2.1 Methodology

In this dissertation, we developed an approach for estimating univariate and multivariate mixture models with nonparametric components, covering settings in which the latent mixing-measure supports of different components are either convexly separable or not. We demonstrated the performance of the proposed method in low- to moderate-dimensional settings. The MDPM modeling framework can be extended and further developed in several directions.

One direction concerns computational efficiency in settings where the latent supports are not convexly separable. In this case, the efficiency of the algorithm can depend strongly on the number of hypercubes used to cover the latent support, since a separate DPM is fitted within each hypercube. Because there are infinitely many possible hypercube coverings of a given latent support, optimizing the number, size, and placement of the hypercubes may substantially improve subsequent posterior inference. When the number of covering hypercubes is large, each hypercube often captures only a local region of the component density. In large datasets, it may therefore be unnecessary to use all observations when fitting the local density associated with a single hypercube. Developing adaptive or subsampling-based strategies for local DPM fitting could provide another promising direction for improving computational efficiency.

A second direction concerns the recovery of the latent support. In this dissertation, we use the diameter of the minimum spanning tree to recover latent one-dimensional manifold structure and to guide the placement of covering hypercubes. This idea could be extended to higher-dimensional manifold or hypersurface structures. For example, after using HDBSCAN to identify samples belonging to a connected high-density region, one could construct a hypercube covering of the entire set of samples in that region, rather than placing hypercubes only along the diameter of the minimum spanning tree. Such a strategy would allow the covering to adapt to latent support

manifolds of arbitrary intrinsic dimension. As in the previous direction, an important open question is how to choose an efficient covering, including the number, size, and placement of hypercubes.

A third important direction is to investigate the performance of the proposed approach in high-dimensional settings, such as single-cell RNA sequencing data or image data. In such settings, the relevant probability distribution may not be well described as a density over the full ambient space. Instead, the data may concentrate near a lower-dimensional manifold or other structured subset of the ambient space. It would therefore be interesting to study how such low-dimensional geometric structure can be identified and incorporated into the MDPM framework, and to evaluate whether the proposed approach can recover meaningful nonparametric components in these settings. This direction has the potential to extend the applicability of the framework to a broader class of complex, high-dimensional data.

4.2.2 Theory

In this dissertation, we establish theoretical guarantees for the MDPM approach under a specific set of assumptions. However, there remains a substantial gap between the proposed modeling framework and a complete theoretical understanding of its behavior. This gap arises in several directions.

First, in the univariate setting, the posterior contraction rates are established under the assumptions that each component is a location mixture of Gaussian kernels and that the latent mixing-measure support is fixed, known, and satisfies suitable separation conditions. A natural extension is to allow the support regions to be unknown. One possible approach is to first derive a rate at which the latent support can be estimated to a prescribed accuracy, and then establish posterior contraction rates for the component densities conditional on this estimated support. Such results may be obtainable within a theoretical framework similar to the one developed in this dissertation, but would require additional analysis to account for the uncertainty introduced by support estimation.

Second, the general strategy for estimating a component density relies on approximating a smooth

density by a location mixture of Gaussian kernels. The overall estimation error can therefore be decomposed into two parts: the approximation error between the true component density and its best Gaussian-mixture approximation, and the statistical error associated with estimating this approximating Gaussian mixture. The theory developed in this dissertation primarily addresses the second part. A complete contraction theory for the MDPM framework would also require controlling the first part, namely the approximation error incurred by representing general smooth component densities through Gaussian mixtures. Moreover, if the true component density is not itself a location mixture of Gaussian kernels, then the latent mixing-measure support is no longer an intrinsic object associated with the true data-generating distribution. This raises the additional question of how separation between component densities should be formulated.

Third, the identifiability conditions in this dissertation are established for components represented as location mixtures of Gaussian kernels. For general smooth component densities, however, the latent mixing-measure support is not an intrinsic object. It therefore remains an open question how to formulate identifiability conditions for more general classes of smooth densities. In addition, as discussed in Section 4.1.1, the definition of a component is application-dependent. The conditions developed here represent one possible notion of identifiability, in which each component is spatially concentrated and different components are sufficiently separated. Other notions of components may require different structural assumptions to ensure unique identification.

Fourth, the posterior contraction theory for multivariate settings remains largely unexplored. A natural direction is to extend the theoretical framework used for univariate Dirichlet process mixtures to multivariate settings, while accounting for the additional complexity introduced by multivariate kernels, covariance structures, and higher-dimensional latent supports.

Fifth, the theory for spike-and-slab variants of the MDPM framework remains to be developed. Future work could investigate identifiability, posterior contraction, and computational behavior in these settings.

4.2.3 Applications

The proposed approach is already applicable to, and has been evaluated in, several low- to moderate-dimensional settings. An important direction for future work is to extend the framework to higher-dimensional data, such as single-cell RNA sequencing profiles, image data, and other applications in which learning latent subpopulations is of scientific interest. In these settings, the main challenge is not only computational scalability, but also the development of appropriate representations of component structure in high-dimensional spaces. Extending the MDPM framework to such domains would broaden its applicability to a wider range of complex data analysis problems where flexible and interpretable subpopulation learning is needed.

APPENDIX A

Appendix of Chapter 2

A.1 Algorithm

This section describes the slice sampler algorithm for the univariate MDPM hierarchical model in (2.6). The multivariate version, presented in Section 2.5, is a natural generalization of this algorithm. For each component $1 \leq i \leq K$, the random measure $H_i \sim \text{DP}(\alpha H_{i0})$ can be generated via the stick-breaking construction ([Sethuraman \(1994\)](#)):

$$H_i = \sum_{j=1}^{\infty} \beta_{ij} \delta_{(u_{ij}, \sigma_{ij})},$$

$$\beta_{ij} = v_{ij} \prod_{k=1}^{j-1} (1 - v_{ik}), \quad v_{ij} \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha), \quad (u_{ij}, \sigma_{ij}) \sim H_{i0}, \quad (\text{A.1})$$

where $\delta_{(u_{ij}, \sigma_{ij})}$ denotes the Dirac measure at (u_{ij}, σ_{ij}) . For each sample x_i ($1 \leq i \leq n$), $z_i \in \{1, \dots, K\}$ denotes its component label. Conditional on $z_i = k$, define $s_i \in \mathbb{N}$ to be the cluster (atom) label of x_i within component k , i.e., $P(s_i = j | z_i = k) = \beta_{kj}$, and $x_i | z_i = k, s_i = j \sim g_{u_{kj}, \sigma_{kj}}$. Introduce a slice variable for each observation x_i : given the atom index s_i in component z_i , denote ρ_i the slice variable $\rho_i \sim \text{Uniform}(0, \beta_{z_i s_i})$ and find the minimum $\rho^* := \min_i \rho_i$. The slice variables restrict attention to atoms whose weights exceed the global threshold ρ^* , thereby yielding a finite, iteration-specific active set. Define the truncation index in component i as $m_i = \min \left\{ j \geq 1 : 1 - \sum_{k=1}^j \beta_{ik} < \rho^* \right\}$, so that only atoms $j = 1, \dots, m_i$ need to be considered when updating allocations for data assigned to component i .

To sample $(\mathbf{c}, \mathbf{r})$ from the repulsive prior (2.4), we introduce another auxiliary slice variable ξ following [Petrulia et al. \(2012\)](#):

$$\xi = \begin{cases} 0, & \text{if } \min_{1 \leq i < j \leq K} |c_i - c_j| - r_i - r_j \leq 0 \\ \min_{1 \leq i < j \leq K} \exp\left(-\frac{\tau}{(|c_i - c_j| - r_i - r_j)^\nu}\right), & \text{else} \end{cases} \quad (\text{A.2})$$

$$\xi \sim \text{Uniform}(0, \xi)$$

Let $\mathbf{z} = (z_i)_{i=1}^n$, $\mathbf{s} = (s_i)_{i=1}^n$, $\mathbf{u} = ((u_{ij})_{j=1}^{m_i})_{i=1}^K$, $\mathbf{x} = (x_i)_{i=1}^n$, $\mathbf{c}_{-i} = (c_j)_{j \neq i}$, $\mathbf{r}_{-i} = (r_j)_{j \neq i}$, and define $\mathcal{J}_i = \{j : \exists 1 \leq l \leq n, z_l = i, s_l = j\}$. Then the full conditional of c_i is

$$\begin{aligned} p(c_i | \mathbf{c}_{-i}, \mathbf{r}, \xi, \mathbf{u}, \mathbf{z}, \mathbf{s}) &\propto g_{\mu_0, \eta}(c_i) \cdot \mathbf{1}\{\xi < \zeta\} \cdot \prod_{j \in \mathcal{J}_i} g_{c_i, \sigma_0}(u_{ij}) \mathbf{1}_{[c_i - r_i, c_i + r_i]}(u_{ij}) \\ &\propto g_{\mu', \sigma'}(c_i) \cdot \mathbf{1}\{\xi < \zeta\} \prod_{j \in \mathcal{J}_i} \mathbf{1}_{[c_i - r_i, c_i + r_i]}(u_{ij}) \end{aligned} \quad (\text{A.3})$$

where $\mu' = \frac{\mu_0 \sigma_0^2 + (\sum_{j \in \mathcal{J}_i} u_{ij}) \cdot \eta^2}{\sigma_0^2 + |\mathcal{J}_i| \cdot \eta^2}$, $\sigma' = \frac{1}{\sqrt{1/\eta^2 + |\mathcal{J}_i|/\sigma_0^2}}$. Furthermore, the full conditional of r_i is

$$p(r_i | \mathbf{c}, \mathbf{r}_{-i}, \xi, \mathbf{u}, \mathbf{z}, \mathbf{s}) \propto \text{Gamma}(r_i) \cdot \frac{1}{\left(1 - 2\Phi\left(-\frac{r_i}{\sigma_0}\right)\right)^{|\mathcal{J}_i|}} \cdot \mathbf{1}\{\xi < \zeta\} \cdot \prod_{j \in \mathcal{J}_i} \mathbf{1}_{[c_i - r_i, c_i + r_i]}(u_{ij}), \quad (\text{A.4})$$

where $\Phi(\cdot)$ is the standard normal cdf. The posterior of r_i is no longer conjugate, and this is the only step that requires a Metropolis-Hastings update.

Let $\boldsymbol{\beta} = ((\beta_{ij})_{j=1}^{m_i})_{i=1}^K$, and $\boldsymbol{\sigma} = ((\sigma_{ij})_{j=1}^{m_i})_{i=1}^K$. We update (z_i, s_i) as a block. First, z_i is sampled through the following rule:

$$p(z_i = k | \mathbf{x}, \mathbf{w}, \rho_i, \boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\sigma}) \propto w_k \cdot \left(\sum_{j=1}^{m_k} \mathbf{1}\{\beta_{kj} > \rho_i\} \cdot g_{u_{kj}, \sigma_{kj}}(x_i) \right), \quad k = 1, \dots, K. \quad (\text{A.5})$$

Given $z_i = k$, sample s_i from

$$p(s_i = j | z_i = k, \mathbf{x}, \rho_i, \boldsymbol{\beta}, \mathbf{u}, \sigma) \propto \mathbf{1}\{\beta_{kj} > \rho_i\} \cdot g_{u_{kj}, \sigma_{kj}}(x_i), \quad j = 1, \dots, m_k. \quad (\text{A.6})$$

Let $n_k = \#\{i : z_i = k\}$ for $k = 1, \dots, K$. The posterior of $\mathbf{w}$ given $\mathbf{z}$ is

$$\mathbf{w} | \mathbf{z} \sim \text{Dir}(1/K + n_1, \dots, 1/K + n_K). \quad (\text{A.7})$$

Now given $z_i = k$, $s_i = j$, let $\mathcal{I}_{kj} = \{i : z_i = k, s_i = j\}$, then the full conditional of u_{kj} is

$$p(u_{kj} | \mathbf{x}, \mathbf{z}, \mathbf{s}, c_k, r_k, \sigma_{kj}) \propto g_{\mu', \sigma'}(u_{kj}) \cdot \mathbf{1}_{[c_k - r_k, c_k + r_k]}(u_{kj}), \quad (\text{A.8})$$

where $\mu' = \frac{c_k \sigma_{kj}^2 + (\sum_{i \in \mathcal{I}_{kj}} x_i) \cdot \sigma_0^2}{\sigma_{kj}^2 + |\mathcal{I}_{kj}| \cdot \sigma_0^2}$, $\sigma' = \frac{1}{\sqrt{1/\sigma_0^2 + |\mathcal{I}_{kj}|/\sigma_{kj}^2}}$, and given u_{kj} , the full conditional of σ_{kj}^2 is

$$\sigma_{kj}^2 | \mathbf{x}, \mathbf{z}, \mathbf{s}, u_{kj} \sim \text{InvGamma}\left(\theta_1 + |\mathcal{I}_{kj}|/2, \theta_2 + \sum_{i \in \mathcal{I}_{kj}} (x_i - u_{kj})^2/2\right), \quad (\text{A.9})$$

where θ_1 and θ_2 are the hyperparameters in the inverse-Gamma prior distribution in (2.5). Finally, define the vector of active-atom weights for component k as $\boldsymbol{\beta}_k = (\beta_{k1}, \dots, \beta_{kj_k}, \beta_{k0})$, $\beta_{k0} = 1 - \sum_{j=1}^{j_k} \beta_{kj_k}$. Following [Ge et al. \(2015\)](#), the posterior of $\boldsymbol{\beta}_k$ is

$$\boldsymbol{\beta}_k | \mathbf{z}, \mathbf{s}, \alpha \sim \text{Dir}(|\mathcal{I}_{k1}|, \dots, |\mathcal{I}_{kj_k}|, \alpha). \quad (\text{A.10})$$

For each active atom $j = 1, \dots, j_k$ in component $k = 1, \dots, K$, draw the minimum slice variable

$$b_{kj} \sim \text{Beta}(1, |\mathcal{I}_{kj}|), \quad \rho_{kj}^* = \beta_{kj} b_{kj}, \quad i_{kj}^* \sim \text{Uniform}(\mathcal{I}_{kj}) \quad (\text{A.11})$$

and set the global slice threshold

$$\rho^* = \min_{1 \leq k \leq K} \min_{1 \leq j \leq j_k} \rho_{kj}^* \quad (\text{A.12})$$

We follow the MapReduce framework of [Ge et al. \(2015\)](#) to design Algorithm 1. Throughout, a superscript (t) denotes the value at the t -th MCMC iteration.

Algorithm 1: Slice Sampler for MDPM

Input : $x, \mathbf{w}^{(t)}, \mathbf{c}^{(t)}, \mathbf{r}^{(t)}, \mathbf{z}^{(t)}, \mathbf{s}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{u}^{(t)}, \boldsymbol{\sigma}^{(t)}, \{\rho_{kj}^{*(t)}\}, \{i_{kj}^{*(t)}\}, \rho^{*(t)}$
Output : $\mathbf{w}^{(t+1)}, \mathbf{c}^{(t+1)}, \mathbf{r}^{(t+1)}, \mathbf{z}^{(t+1)}, \mathbf{s}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \mathbf{u}^{(t+1)}, \boldsymbol{\sigma}^{(t+1)}, \{\rho_{kj}^{*(t+1)}\}, \{i_{kj}^{*(t+1)}\}, \rho^{*(t+1)}$

for $k \leftarrow 1$ **to** K **do**
 while $1 - \sum_{j=1}^{j_k} \beta_{kj} \geq \rho^*$ **do**
 $j_k \leftarrow j_k + 1$;
 draw β_{k,j_k}, u_{k,j_k} and σ_{k,j_k} via (A.1);
 end
end

Function Map (x_i) :
 // Global variables: $\mathbf{z}^{(t)}, \mathbf{s}^{(t)}, \mathbf{w}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{u}^{(t)}, \boldsymbol{\sigma}^{(t)}, \{\rho_{kj}^{*(t)}\}, \{i_{kj}^{*(t)}\}$
 if $i = i_{z_i, s_i}^*$ **then** $\rho_i \leftarrow \rho_{z_i, s_i}^{*(t)}$;
 else $\rho_i \leftarrow \text{Uniform}(\rho_{z_i, s_i}^{*(t)}, \beta_{z_i, s_i}^{(t)})$;
 draw $z_i^{(t+1)}$ via (A.5);
 draw $s_i^{(t+1)}$ via (A.6);
 return {key : $(z_i^{(t+1)}, s_i^{(t+1)})$, value : x_i }
draw $\mathbf{w}^{(t+1)}$ via (A.7);

for each unique key $(z^{(t+1)}, s^{(t+1)})$ **do**
 collect all $\{x_i\}$ that have this key;
 $n = \#\{x_i\}$;
 $sx = \sum x_i$; $ssx = \sum x_i^2$;
 emit {key : $(z^{(t+1)}, s^{(t+1)})$, value : (n, sx, ssx) }
end

Function Reduce ($(z^{(t+1)}, s^{(t+1)}), (n, sx, ssx)$) :
 // Global variables: $\mathbf{c}^{(t)}, \mathbf{r}^{(t)}, \mathbf{u}^{(t)}, \boldsymbol{\sigma}^{(t)}$
 draw $u_{z_i, s_i}^{(t+1)}$ via (A.8);
 draw $\sigma_{z_i, s_i}^{(t+1)}$ via (A.9);
 return $u_{z_i, s_i}^{(t+1)}, \sigma_{z_i, s_i}^{(t+1)}, n_{z_i, s_i}$
draw $\boldsymbol{\beta}^{(t+1)}$ via (A.10);
draw $\{\rho_{kj}^{*(t+1)}\}, \{i_{kj}^{*(t+1)}\}$ via (A.11), update $\rho^{*(t+1)}$ via (A.12);
draw ξ via (A.2);
draw $\mathbf{c}^{(t+1)}$ via (A.3);
draw $\mathbf{r}^{(t+1)}$ via (A.3);

A.2 Simulations

In this section, we provide the implementation details of the simulation studies presented in

Sections 2.4 and 2.5.

A.2.1 Figure (2.2a) and (2.2b)

Simulation Example 1, shown in Figures (2.2a) and (2.2b), consists of three components. The second component is a skewed exponential-power distribution

$$f(x; \mu, \sigma, p, \alpha) = \begin{cases} \frac{1}{\sigma 2p^{1/p}\Gamma(1+1/p)} \exp \left\{ -\frac{1}{2p} \left| \frac{x-\mu}{\alpha\sigma} \right|^p \right\}, & \text{if } x \leq \mu \\ \frac{1}{\sigma 2p^{1/p}\Gamma(1+1/p)} \exp \left\{ -\frac{1}{2p} \left| \frac{x-\mu}{(1-\alpha)\sigma} \right|^p \right\}, & \text{if } x > \mu \end{cases}$$

with $\mu = 0.9$, $\sigma = 1.1$, $p = 0.7$, $\alpha = 0.7$. The third component is a Laplace distribution

$$f(x; \mu, \theta) = \frac{1}{2\theta} \exp \left(-\frac{|x - \mu|}{\theta} \right)$$

with $\mu = 6.3$, and $\theta = 1.1$. The first component is a random linear combination of Hermite functions (defined in (A.20)), constructed as follows:

1. Generate the Hermite function $\psi_j(x)$ for $j = 0, \dots, L-1$.
2. Draw coefficient $a_j \sim N(0, \sigma_j^2)$, where $\sigma_j = \frac{\tau}{(j+1)^{\alpha/2}}$, for $j = 0, \dots, L-1$.
3. Define $\tilde{f}(x) = (\sum_{j=0}^{L-1} a_j \psi_j(\frac{x-\mu}{s}) s^{-1})^2$.
4. Normalize to obtain the density

$$f(x) = \begin{cases} \frac{\tilde{f}(x)}{\int_{-T}^T \tilde{f}(x) dx}, & x \in [-T, T] \\ 0, & \text{otherwise} \end{cases}, \quad (\text{A.13})$$

with $L = 14$, $\tau = 0.7$, $\alpha = 1.6$, $\mu = -7.0$, $s = 0.8$, and $T = 15$. The histogram in (2.2a) is based on 10,000 samples drawn from the mixture model.

A.2.2 Figure (2.2c) and (2.2d)

Simulation Example 2, shown in Figures (2.2c) and (2.2d), consists of two components. The

spike component is generated as follows:

1. Randomly generate two orders j_0 and j_1 from the set $\{2, 3, 5, 6\}$;
2. Draw two coefficients $a_0 \sim N(0, \sigma_0^2)$, $a_1 \sim N(0, \sigma_1^2)$, where $\sigma_0 = \frac{\tau}{(j_0+1)^{\alpha/2}}$, and $\sigma_1 = \frac{\tau}{(j_1+1)^{\alpha/2}}$;
3. Define $\tilde{f}(x) = \left(\sum_{j=0}^1 a_j \psi_j \left(\frac{x-\mu}{s} \right) s^{-1} \right)^{2\gamma}$;
4. Normalize $\tilde{f}(x)$ to obtain the density $f(x)$ as in (A.13),

where we set $\tau = 2.6$, $\alpha = 1.5$, $\mu = 0$, $s = 1.8$, $\gamma = 5.8$, and $T = 20$. The slab component is constructed according to the following procedure:

1. Draw $a_j \sim N(0, \sigma_j^2)$, where $\sigma_j = \frac{\tau}{(j+1)^{\alpha/2}}$ for $j = 0, \dots, 5$, and $\sigma_j = \frac{0.35\tau}{(j+1)^{\alpha/2}}$ for $j = 6, 7$.
2. Define $\tilde{f}(x) = (\sum_{j=0}^7 a_j \psi_j (\frac{x-\mu}{s}) s^{-1})^2$.
3. Normalize $\tilde{f}(x)$ to obtain the density $f(x)$ as in (A.13),

where $\tau = 0.9$, $\alpha = 5.2$, $\mu = 0$, $s = 14.5$, and $T = 20$.

The histogram in (2.2c) is estimated from 10,000 samples, and the density estimates in (2.2d) are obtained by MDPM with fixed intervals: $I_1 = (0, 0.5]$ and $I_2 = [3.0, \infty]$.

A.2.3 Figure 2.3

The simulation example shown in Figure 2.3 considers a two-component mixture model. Each component itself consists of a mixture of 100 bivariate Gaussian distributions. For the first component, the Gaussian means are equally spaced along a circle of radius 2 centered at $[-2, 2]$. Similarly, for the second component, the means are equally spaced along a circle of radius 2 centered at $[2, 2]$. The covariance matrices of the Gaussian distributions are generated as follows.

- For each mixture component, the 11 Gaussian distributions whose means are closest to the origin $[0, 0]$ have covariance matrices independently drawn from an inverse Wishart

distribution $\mathcal{W}^{-1}(\nu_1, \Psi_1)$, where $\nu_1 = 6$, $\Psi_1 = \begin{pmatrix} 4.5 & 0 \\ 0 & 4.5 \end{pmatrix}$. The relatively large diagonal entries induce substantial overlap between the two components in the tails near the origin.

- The remaining 89 Gaussian distributions in each component have covariance matrices independently drawn from $\mathcal{W}^{-1}(\nu_2, \Psi_2)$, where $\nu_2 = 6$ and $\Psi_2 = \begin{pmatrix} 0.8 & 0 \\ 0 & 0.8 \end{pmatrix}$.

In this simulation example, the underlying regions I_1 and I_2 are unknown and inferred from the posterior.

A.3 Proofs for Section 2.6: Theorem 2.6.1

Proof of Theorem 2.6.1. We first show that for any function $h(x)$ of the form $h(x) = \int g_{u,\sigma}(x) dG(u)$, where the mixing measure G has compact support, if $h(x) = \int g_{u,\sigma'}(x) dG'(u)$, where G' also has compact support, then we must have $\sigma = \sigma'$ and $G = G'$. If $\sigma' \neq \sigma$, w.l.o.g, assume that $\sigma' > \sigma$. Because G and G' have compact support, we can find $u_0 \in \mathbb{R}$ such that $\text{supp}(G) \subseteq (-\infty, u_0 - 1]$, and $u'_0 \in (0, +\infty)$ such that $\text{supp}(G') \subseteq [-u'_0, u'_0]$. By definition, for any $x \in \mathbb{R}$,

$$\frac{1}{\sqrt{2\pi}\sigma} \int e^{-\frac{(x-u)^2}{2\sigma^2}} dG(u) = \frac{1}{\sqrt{2\pi}\sigma'} \int e^{-\frac{(x-u)^2}{2\sigma'^2}} dG'(u). \quad (\text{A.14})$$

Multiplying both sides by $\exp\left(\frac{(x-u_0)^2}{2\sigma^2}\right)$, we get

$$\begin{aligned} \int e^{\frac{(x-u_0)^2}{2\sigma^2} - \frac{(x-u)^2}{2\sigma^2}} dG(u) &= \frac{\sigma}{\sigma'} \int e^{\frac{(x-u_0)^2}{2\sigma^2} - \frac{(x-u)^2}{2\sigma'^2}} dG'(u) \\ \int e^{\frac{(u-u_0)x}{\sigma^2}} \cdot e^{\frac{u_0^2-u^2}{2\sigma^2}} dG(u) &= \frac{\sigma}{\sigma'} \cdot e^{\left(\frac{1}{2\sigma^2} - \frac{1}{2\sigma'^2}\right)x^2 - \left(\frac{u_0}{\sigma^2}\right)x + \frac{u_0^2}{2\sigma^2}} \cdot \int e^{\frac{ux}{\sigma'^2} - \frac{u^2}{2\sigma'^2}} dG'(u). \end{aligned} \quad (\text{A.15})$$

Because $\text{supp}(G) \subseteq (-\infty, u_0 - 1]$, for any $x > 0$,

$$\text{LHS} \leq e^{-\frac{x}{\sigma^2}} \int e^{\frac{u_0^2-u^2}{2\sigma^2}} dG(u) \rightarrow 0, \text{ as } x \rightarrow +\infty,$$

However, because $\sigma' > \sigma$, and $\text{supp}(G') \subseteq [-u'_0, u'_0]$, for any sufficiently large $x > 0$,

$$\text{RHS} \geq \frac{\sigma}{\sigma'} \cdot e^{\left(\frac{1}{2\sigma^2} - \frac{1}{2\sigma'^2}\right)x^2 - \left(\frac{u_0}{\sigma^2}\right)x + \frac{u_0^2}{2\sigma^2}} \cdot e^{-\frac{u'_0 x}{\sigma'^2} - \frac{u'^2_0}{2\sigma'^2}} \rightarrow \infty, \text{ as } x \rightarrow +\infty$$

Therefore, there must exist x sufficiently large such that the left-hand side is not equal to the right-hand side, contradicting the assumption. A similar contradiction arises if $\sigma' < \sigma$. Hence we conclude that $\sigma = \sigma'$. Now assume $h(x) = \int g_{u,\sigma}(x) dG(u) = \int g_{u,\sigma}(x) dG'(u)$, then for any $t \in \mathbb{R}$,

$$\begin{aligned} \int e^{itx} h(x) dx &= \int e^{itx} \left(\int g_{u,\sigma}(x) dG(u) \right) dx \\ &= \int \left(\int e^{itx} g_{u,\sigma}(x) dx \right) dG(u) \\ &= \int e^{itu - \frac{1}{2}\sigma^2 t^2} dG(u) \\ &= e^{-\frac{1}{2}\sigma^2 t^2} \int e^{itu} dG(u). \end{aligned}$$

Similarly, we have $\int e^{itx} h(x) dx = e^{-\frac{1}{2}\sigma'^2 t^2} \int e^{itu} dG'(u)$, therefore, $\int e^{itu} dG(u) = \int e^{itu} dG'(u)$, $\forall t \in \mathbb{R}$. Since the characteristic functions of G and G' are identical, we must have $G = G'$.

Therefore, given any f as defined in (2.13), we can find unique σ and G , such that $f = \int g_{u,\sigma}(x) dG(u)$. Under condition (C2.1), G can be uniquely separated into K components. Under condition (C2.2), we first find the unique separation $S_{\text{supp}(G)}$ as defined in Section 2.2, then among all pairs of neighboring sets $(S_i, S_j) \in \mathcal{N}(S_{\text{supp}(G)})$, we identify the $K - 1$ pairs with the largest distances $d_c(S_i, S_j)$. The interval of length $d_c(S_i, S_j)$ between each of these pairs S_i and S_j will also be one of the $K - 1$ intervals that separate $\text{supp}(V_1), \dots, \text{supp}(V_K)$. Therefore $f_1, \dots, f_K$ can also be uniquely identified.

□

A.4 Proofs for Section 2.7.1: Theorem 2.7.1

In this section, Lemma A.4.1 provides a lower bound on the probability that, under the truncated

Dirichlet distribution, $(W_1, \dots, W_K)$ lies in a neighborhood of the true weights $(w_{10}, \dots, w_{K0})$. The corresponding result for a non-truncated Dirichlet prior is given in Lemma 6.1 of [Ghosal et al. \(2000\)](#).

Lemma A.4.1. *Let $(W_1, \dots, W_K)$ follow a truncated Dirichlet distribution under assumption (A2.1). Let $(w_{10}, \dots, w_{K0})$ be any point on the K -simplex such that $w_{k0} \in [\underline{w}, \frac{1-\underline{w}}{K-1}]$, for some $0 < \underline{w} < 1/K^2$, $k = 1, \dots, K$. There exist a positive constant C depending only on α and $\underline{w}$ and K such that, for $\epsilon \leq \frac{1}{K\sqrt{K-1}}$,*

$$\Pr\left(\sum_{k=1}^K |W_k - w_{k0}| \leq 2\epsilon\right) \geq C(\underline{w}, \alpha, K) \exp\left(-2(K-1) \log \frac{1}{\epsilon}\right). \quad (\text{A.16})$$

Proof. Find an index k such that $w_{k0} \leq 1/K$. By relabeling, we can assume that $k = K$. If $|w_k - w_{k0}| \leq \epsilon^2$ and $w_k \in [\underline{w}, \frac{1-\underline{w}}{K-1}]$ for $k = 1, \dots, K-1$, then because $\epsilon^2 \leq \frac{1}{K^2(K-1)}$,

$$\sum_{k=1}^{K-1} w_k \geq 1 - w_{K0} - (K-1)\epsilon^2 \geq (K-1) \left(1/K - \epsilon^2\right) \geq 1 - \frac{K+1}{K^2}.$$

So we have $w_K \leq \frac{K+1}{K^2} < \frac{1-\underline{w}}{K-1}$. In addition, because $\sum_{k=1}^{K-1} w_k \leq 1 - \underline{w}$, we also have $w_K \geq \underline{w}$. Hence there exists $w = (w_1, \dots, w_K)$ in the truncated simplex with these first $K-1$ coordinates such that $\sum_{k=1}^K w_k = 1$, $w_k \in [\underline{w}, \frac{1-\underline{w}}{K-1}]$, $\forall 1 \leq k \leq K$. Furthermore, $\sum_{k=1}^K |w_k - w_{k0}| \leq 2 \sum_{k=1}^{K-1} |w_k - w_{k0}| \leq 2\epsilon^2(K-1) \leq 2\epsilon$. Therefore the probability on the left-hand side of (A.16) is bounded below by

$$\begin{aligned} & \Pr\left(|W_K - w_{K0}| \leq \epsilon^2, W_k \in [\underline{w}, \frac{1-\underline{w}}{K-1}], k = 1, \dots, K-1\right) \\ & \geq C(\underline{w}, \alpha, K) \cdot \prod_{k=1}^{K-1} \int_{\max(w_{k0}-\epsilon^2, \underline{w})}^{\min(w_{k0}+\epsilon^2, \frac{1-\underline{w}}{K-1})} w_i^{\alpha_i-1} dw_i, \end{aligned}$$

where we use that $\left(1 - \sum_{k=1}^{K-1} w_i\right)^{\alpha_K-1} \geq 1$, since $\alpha_K \leq 1$. Similarly, since $\alpha_k \leq 1$ for every k , we can lower bound the integrand by 1. Note that, under assumption, $\underline{w} < \frac{1}{K^2}$, which implies $\frac{1-\underline{w}}{K-1} - \underline{w} > \frac{1}{K}$. Moreover, since $\epsilon^2 \leq \frac{1}{K^2(K-1)}$, the interval of integration contains at least an interval

of length ϵ^2 . We can bound the last display from below by

$$C(\underline{w}, \alpha, K) \cdot \epsilon^{2(K-1)} \geq C(\underline{w}, \alpha, K) \cdot \exp\left(-2(K-1) \log \frac{1}{\epsilon}\right).$$

This concludes the proof. $\square$

Now we are ready to prove the prior concentration inequality of MDPM in the following lemma:

Lemma A.4.2. *Let $f_0 \in \mathcal{F}_{K,\sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$, where $K, \sigma, \mathbf{c}, \mathbf{r}, \underline{w}$ are fixed and known. Consider the prior Π constructed under specifications (A2.1), (A2.2) and (A2.3). Then by taking $\tilde{\epsilon}_n = \log n / \sqrt{n}$, when n is larger than some constant, we have*

$$\Pi(\{f \mid f \in B_2(f_0, \tilde{\epsilon}_n)\}) \geq c_3 e^{-c_4 n \tilde{\epsilon}_n^2} \quad (\text{A.17})$$

where constants $c_3, c_4 > 0$.

Remark. Lemma A.4.2 shows that, for any finite number of components K , a mixture of K Dirichlet process priors, under the specified conditions, attains the same prior concentration rate around the true mixture density f_0 as a single Dirichlet process prior.

Proof of Lemma A.4.2. As established in the proof of Theorem 5.1 in Ghosal and van der Vaart (2001), the probability lower bound for the first set in equation (5.12) of Ghosal and van der Vaart (2001) applies in our setting as well. The only difference is that σ is fixed here; in this case, the first factor on the right-hand side of equation (5.14) in Ghosal and van der Vaart (2001) can be replaced by 1. Therefore the conclusion in (5.16) still holds, which in the proof is implied by the following: For any $k = 1, \dots, K$, $0 < \epsilon < \frac{1}{K}$, and some positive constant d_k ,

$$\Pi_k \left(\left\{ V_k \mid \|f_k - f_{k0}\|_1 \leq d_k \epsilon \left(\log \frac{1}{\epsilon} \right)^{1/2} \right\} \right) \geq C_k \exp \left[-c_k \left(\log \frac{1}{\epsilon} \right)^2 \right], \quad (\text{A.18})$$

for some positive constants C_k and c_k . By Lemma A.4.1, when $\epsilon < \min(\frac{1}{K\sqrt{K-1}}, e^{-1})$, we have

$\log \frac{1}{\epsilon} > 1$, therefore

$$\begin{aligned} \Pr \left(\sum_{k=1}^K |W_k - w_{k0}| \leq 2\epsilon \left(\log \frac{1}{\epsilon} \right)^{1/2} \right) &\geq \Pr \left(\sum_{k=1}^K |W_k - w_{k0}| \leq 2\epsilon \right) \\ &\geq C_w \exp \left(-2(K-1) \log \frac{1}{\epsilon} \right) \end{aligned}$$

for some positive constants C_w . Furthermore,

$$\begin{aligned} \|f - f_0\|_1 &= \left\| \sum_{k=1}^K (W_k f_k - w_{k0} f_{k0}) \right\|_1 \\ &= \left\| \sum_{k=1}^K (W_k - w_{k0})(f_k - f_{k0}) + (W_k - w_{k0})f_{k0} + (f_k - f_{k0})w_{k0} \right\|_1 \\ &\leq \sum_{k=1}^K |W_k - w_{k0}| \cdot \|f_k - f_{k0}\|_1 + \sum_{k=1}^K |W_k - w_{k0}| + \sum_{k=1}^K \|f_k - f_{k0}\|_1 \\ &\leq 2 \sum_{k=1}^K \|f_k - f_{k0}\|_1 + \sum_{k=1}^K |W_k - w_{k0}|, \end{aligned}$$

and because $\mathbf{W}$ and Π_k 's are mutually independent,

$$\begin{aligned} &\Pi \left(\left\{ f \mid \|f - f_0\|_1 \leq 2 \left(\sum_{k=1}^K d_k + 1 \right) \epsilon \log \left(\frac{1}{\epsilon} \right)^{1/2} \right\} \right) \\ &\geq \Pr \left(\sum_{k=1}^K |W_k - w_{k0}| \leq 2\epsilon \log \left(\frac{1}{\epsilon} \right)^{1/2}, \text{ and } \|f_k - f_{k0}\|_1 \leq d_k \epsilon \left(\log \frac{1}{\epsilon} \right)^{1/2}, k = 1, \dots, K \right) \\ &= \Pr \left(\sum_{k=1}^K |W_k - w_{k0}| \leq 2\epsilon \log \left(\frac{1}{\epsilon} \right)^{1/2} \right) \cdot \prod_{k=1}^K \Pi_k \left(\|f_k - f_{k0}\|_1 \leq d_k \epsilon \left(\log \frac{1}{\epsilon} \right)^{1/2} \right) \\ &= \prod_{k=1}^K C_k \cdot C_w \exp \left[- \left(\sum_{k=1}^K c_k \right) \left(\log \frac{1}{\epsilon} \right)^2 - 2(K-1) \log \frac{1}{\epsilon} \right] \\ &\geq \tilde{C} \exp \left[-\tilde{c} \left(\log \frac{1}{\epsilon} \right)^2 \right] \end{aligned}$$

for some positive constants $\tilde{C}$ and $\tilde{c}$. By (2.14) and Lemma 4.1 in [Ghosal and van der Vaart \(2001\)](#)

and analogous results for location-scale mixtures, we know when ϵ is sufficiently small such that

$$2 \left(\sum_{k=1}^K d_k + 1 \right) \epsilon \log \left(\frac{1}{\epsilon} \right)^{1/2} < \frac{1}{4},$$

$$\left\{ f \mid \|f - f_0\|_1 \leq 2 \left(\sum_{k=1}^K d_k + 1 \right) \epsilon \log \left(\frac{1}{\epsilon} \right)^{1/2} \right\} \subset B_2 \left(f_0, c \epsilon^{1/2} \left(\log \frac{1}{\epsilon} \right)^{5/4} \right),$$

for some constant c , therefore

$$\Pi \left(\left\{ f = p_{\sigma_0, \sum_{k=1}^K W_k V_k} \mid f \in B_2 \left(f_0, c \epsilon^{1/2} \left(\log \frac{1}{\epsilon} \right)^{5/4} \right) \right\} \right) \geq \tilde{C} \exp \left[-\tilde{c} \left(\log \frac{1}{\epsilon} \right)^2 \right]$$

Putting $\epsilon' = c \epsilon^{1/2} (\log \frac{1}{\epsilon})^{5/4}$ and noting that $\log \frac{1}{\epsilon'} \sim \log \frac{1}{\epsilon}$, we have

$$\Pi(\{f \mid f \in B_2(f_0, \epsilon')\}) \geq c_3 \exp \left[-c_4 \left(\log \frac{1}{\epsilon'} \right)^2 \right] \quad (\text{A.19})$$

for some positive constants c_3 and c_4 . Take $\tilde{\epsilon}_n = \log n / \sqrt{n}$, we obtain

$$\Pi(\{f \mid f \in B_2(f_0, \tilde{\epsilon}_n)\}) \geq c_3 e^{-c_4 n \tilde{\epsilon}_n^2}.$$

□

We now proceed to establish the posterior contraction rate around the true density f_0 . To this end, we apply Theorem 2.1 of [Ghosal and van der Vaart \(2001\)](#).

Proof of Theorem 2.7.1. We verify the conditions (2.8), (2.9) and (2.10) of Theorem 2.1 in [Ghosal and van der Vaart \(2001\)](#) are satisfied for $\Pi_n = \Pi$, $\mathcal{P} = \mathcal{F}_{K, \sigma}(\mathbf{c}, \mathbf{r}, \underline{w})$, $\mathcal{P}_n = \mathcal{P}$ with $\bar{\epsilon}_n = \log n / \sqrt{n}$ and $\tilde{\epsilon}_n = \log n / \sqrt{n}$. Condition (2.8) follows from Theorem 3.1 in [Ghosal and van der Vaart \(2001\)](#) with $\bar{\epsilon}_n = \log n / \sqrt{n}$. Lemma A.4.2 establishes that (2.10) holds for $\tilde{\epsilon}_n = \log n / \sqrt{n}$. □

A.5 Proofs for Section 2.7.2: Construction of $\widehat{f}_i$

In this section, we briefly describe the point estimate $\widehat{f}_i$ given in [Tai and Aragam \(2023\)](#). To this end, we first introduce Hermite functions in section A.5.1.

A.5.1 Hermite functions

We review some useful basics of Hermite orthonormal basis to be used in the sequel. Let ψ_j be the j th (physicist's) Hermite functions, i.e.

$$\psi_j(x) = (-1)^j \frac{1}{\sqrt{2^j j! \sqrt{\pi}}} h_j(x) e^{-\frac{1}{2}x^2}, \quad \forall x \in \mathbb{R}, \quad (\text{A.20})$$

where $h_j(x)$ is the (physicist's) Hermite polynomials, i.e.

$$h_j(x) = (-1)^j e^{x^2} \frac{d^j e^{-\xi^2}}{d\xi^j} \Big|_{\xi=x}, \quad \forall x \in \mathbb{R}.$$

The Hermite functions are orthonormal in $L^2(\mathbb{R})$, i.e. $\|\psi_i\|_2 = 1$ and $\langle \psi_i, \psi_j \rangle = 0 \quad \forall i, j \in \mathbb{N}_0$, where $\langle f, g \rangle = \int_{\mathbb{R}} f(x)g(x)dx$ denotes the inner product with respect to Lebesgue measure. For any $\mu \in \mathbb{R}$ and $j \in \mathbb{N}_0$, we denote $\psi_{j,\mu}(x) = \psi_j(x - \mu)$. It is known that

$$h_j(x - \mu) = \sum_{k=0}^j \binom{j}{k} (-2\mu)^{j-k} h_k(x). \quad (\text{A.21})$$

This can be proved by using the generating function for Hermite polynomials: $e^{-t^2+2tx} = \sum_{j=0}^{\infty} h_j(x) \frac{t^j}{j!}$, $\forall x, t \in \mathbb{R}$. Substitute x by $x - \mu$ in the generating function: $e^{-t^2+2t(x-\mu)} = \sum_{j=0}^{\infty} h_j(x - \mu) \frac{t^j}{j!}$, and rewrite $e^{-t^2+2t(x-\mu)} = e^{-t^2+2tx} \cdot e^{-2t\mu} = \left(\sum_{j=0}^{\infty} h_j(x) \frac{t^j}{j!} \right) \left(\sum_{k=0}^{\infty} \frac{(-2\mu)^k}{k!} t^k \right)$, then by comparing coefficients of $\frac{t^j}{j!}$, we get (A.21). Based on (A.21), it can be checked that for any $\mu \in \mathbb{R}$ and any $i, j \in \mathbb{N}_0$, the inner product (with respect to Lebesgue measure) of $\psi_{i,0}$ and $\psi_{j,\mu}$ takes the form

$$\begin{aligned} \langle \psi_{i,0}, \psi_{j,\mu} \rangle &= \int_{-\infty}^{+\infty} \psi_{i,0}(x) \psi_{j,\mu}(x) dx \\ &= \frac{(-1)^{i+j}}{\sqrt{2^{i+j} i! j! \pi}} \int_{-\infty}^{+\infty} h_i(x) h_j(x - \mu) e^{-x^2/2} e^{-(x-\mu)^2/2} dx \end{aligned}$$

$$\begin{aligned}
&= \frac{(-1)^{i+j}}{\sqrt{2^{i+j} i! j! \pi}} e^{-\mu^2/4} \int_{-\infty}^{+\infty} h_i(x) h_j(x - \mu) e^{-(x-\mu/2)^2} dx \\
&\stackrel{t=x-\mu/2}{=} \frac{(-1)^{i+j}}{\sqrt{2^{i+j} i! j! \pi}} e^{-\mu^2/4} \int_{-\infty}^{+\infty} h_i(t + \mu/2) h_j(t - \mu/2) e^{-t^2} dt.
\end{aligned}$$

Using the equation (A.21), we expand both $h_i(t + \mu/2)$ and $h_j(t - \mu/2)$ in terms of $\{h_k(t)\}$. By orthogonality $\int h_m(x) h_n(x) e^{-x^2} dx = \sqrt{\pi} 2^n n! \cdot \mathbf{1}_{n=m}$,

$$\langle \psi_{i,0}, \psi_{j,\mu} \rangle = \sum_{k=0}^{i \wedge j} \left(e^{-\frac{1}{8}\mu^2} (-1)^i \sqrt{\frac{k!}{i!}} \binom{i}{k} \left(\frac{\mu}{\sqrt{2}} \right)^{i-k} \right) \left(e^{-\frac{1}{8}\mu^2} (-1)^j \sqrt{\frac{k!}{j!}} \binom{j}{k} \left(\frac{-\mu}{\sqrt{2}} \right)^{j-k} \right). \quad (\text{A.22})$$

Let $g_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ be the density of normal distribution with mean μ and standard deviation σ . Since $g_{\mu,1} = \frac{1}{\sqrt{2\sqrt{\pi}}} \psi_{0,\mu}$, we have

$$\langle \psi_{j,0}, g_{\mu,1} \rangle = \frac{1}{\sqrt{2\sqrt{\pi}}} \langle \psi_{j,0}, \psi_{0,\mu} \rangle = \frac{(-1)^j}{\sqrt{2^{j+1} j! \sqrt{\pi}}} e^{-\frac{1}{4}\mu^2} \mu^j.$$

Similarly, if we denote $\psi_{j,\mu,\sigma} = \frac{1}{\sqrt{\sigma}} \psi_j\left(\frac{x-\mu}{\sigma}\right)$, then $\|\psi_{i,\mu,\sigma}\|_2 = 1$, $\langle \psi_{i,\mu,\sigma}, \psi_{j,\mu,\sigma} \rangle = 0$ for $i \neq j$, and

$$\begin{aligned}
&\langle \psi_{i,\mu_1,\sigma}, \psi_{j,\mu_2,\sigma} \rangle \\
&= \sum_{k=0}^{i \wedge j} \left(e^{-\frac{1}{8\sigma^2}(\mu_2-\mu_1)^2} (-1)^i \sqrt{\frac{k!}{i!}} \binom{i}{k} \left(\frac{\mu_2 - \mu_1}{\sqrt{2}\sigma} \right)^{i-k} \right) \times \\
&\quad \left(e^{-\frac{1}{8\sigma^2}(\mu_2-\mu_1)^2} (-1)^j \sqrt{\frac{k!}{j!}} \binom{j}{k} \left(\frac{\mu_1 - \mu_2}{\sqrt{2}\sigma} \right)^{j-k} \right). \quad (\text{A.23})
\end{aligned}$$

Here the inner product is with regard to Lebesgue measure. In particular, since $g_{\mu,\sigma} = \frac{1}{\sqrt{2\sigma\sqrt{\pi}}} \psi_{0,\mu,\sigma}$, we have

$$\langle \psi_{j,\mu_1,\sigma}, g_{\mu_2,\sigma} \rangle = \frac{(-1)^j}{\sqrt{2^{j+1} j! \sigma \sqrt{\pi}}} e^{-\frac{(\mu_2-\mu_1)^2}{4\sigma^2}} \left(\frac{\mu_2 - \mu_1}{\sigma} \right)^j. \quad (\text{A.24})$$

A.5.2 Construction of $\widehat{f}_i$

For any $f \in \mathcal{F}$ as defined in (C2.3), assume $X_1, \dots, X_n$ are n i.i.d. samples from P_f . Let us

represent each component f_i in terms of the Hermite basis functions: $f_i = \sum_{j=0}^{\infty} \alpha_{i,j} \psi_{j,c_i,\sigma}$, where $\alpha_{i,j} = \langle f_i, \psi_{j,c_i,\sigma} \rangle$, for $i = 1, 2$ and $j \in \mathbb{N}_0$. Recall $f = w_1 f_1 + w_2 f_2$. Let $\lambda_{i,j} := w_i \alpha_{i,j}$, for $i = 1, 2$, and $j \in \mathbb{N}_0$. It follows that

$$f = \sum_{j=0}^{\infty} \lambda_{1,j} \psi_{j,c_1,\sigma} + \sum_{j=0}^{\infty} \lambda_{2,j} \psi_{j,c_2,\sigma},$$

and by projecting f onto each $\psi_{k,c_i,\sigma}$, for $i = 1, 2$ and $k \in [\ell]$, where ℓ is a nonnegative integer and $[\ell] := \{0, 1, \dots, \ell - 1\}$,

$$\langle f, \psi_{k,c_i,\sigma} \rangle = \sum_{j=0}^{\infty} \lambda_{1,j} \langle \psi_{j,c_1,\sigma}, \psi_{k,c_i,\sigma} \rangle + \sum_{j=0}^{\infty} \lambda_{2,j} \langle \psi_{j,c_2,\sigma}, \psi_{k,c_i,\sigma} \rangle.$$

Thus, we obtain

$$y = A\lambda + \sum_{j=\ell}^{\infty} \lambda_{1,j} z_{1,j} + \sum_{j=\ell}^{\infty} \lambda_{2,j} z_{2,j},$$

where y is the 2ℓ -dimensional vector whose entries are given by $\langle f, \psi_{j,c_i,\sigma} \rangle$ for $i = 1, 2$; $j \in [\ell]$ (the entries indexed by (i, j) are ordered by i , and then by j), likewise, λ is the 2ℓ -dimensional vector whose entries are given by $\lambda_{i,j} = w_i \alpha_{i,j}$ for $i = 1, 2$; $j \in [\ell]$, A is a 2ℓ -by- 2ℓ matrix whose entries are given by $\langle \psi_{j_1,c_{i_1},\sigma}, \psi_{j_2,c_{i_2},\sigma} \rangle$, for $i_1, i_2 = 1, 2$; $j_1, j_2 \in [\ell]$, e.g., when $\ell = 2$:

$$A = \begin{pmatrix} 1 & 0 & \langle \psi_{0,c_2,\sigma}, \psi_{0,c_1,\sigma} \rangle & \langle \psi_{1,c_2,\sigma}, \psi_{0,c_1,\sigma} \rangle \\ 0 & 1 & \langle \psi_{0,c_2,\sigma}, \psi_{1,c_1,\sigma} \rangle & \langle \psi_{1,c_2,\sigma}, \psi_{1,c_1,\sigma} \rangle \\ \langle \psi_{0,c_1,\sigma}, \psi_{0,c_2,\sigma} \rangle & \langle \psi_{1,c_1,\sigma}, \psi_{0,c_2,\sigma} \rangle & 1 & 0 \\ \langle \psi_{0,c_1,\sigma}, \psi_{1,c_2,\sigma} \rangle & \langle \psi_{1,c_1,\sigma}, \psi_{1,c_2,\sigma} \rangle & 0 & 1 \end{pmatrix},$$

and

$$z_{1,j} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \langle \psi_{j,c_1,\sigma}, \psi_{0,c_2,\sigma} \rangle \\ \vdots \\ \langle \psi_{j,c_1,\sigma}, \psi_{\ell-1,c_2,\sigma} \rangle \end{bmatrix}, \quad z_{2,j} = \begin{bmatrix} \langle \psi_{j,c_2,\sigma}, \psi_{0,c_1,\sigma} \rangle \\ \vdots \\ \langle \psi_{j,c_2,\sigma}, \psi_{\ell-1,c_1,\sigma} \rangle \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \text{for } j \geq \ell.$$

For the rest of the paper, the 2ℓ items will be indexed by (i, k) , $i = 1, 2$, $k = 0, \dots, \ell - 1$, i.e., the $1, 2, \dots, 2\ell$ -th columns in A will be indexed by $(1, 0), (1, 1), \dots, (1, \ell-1), (2, 0), (2, 1), \dots, (2, \ell-1)$, respectively.

Next, let $\widehat{f}$ denote an estimate of f based on the n samples $X_1, \dots, X_n$. For instance, a kernel density estimator can be used for this purpose. Then, take the 2ℓ -dimensional vector $\widehat{y}$ whose entries are given by $\langle \widehat{f}, \psi_{j,c_i,\sigma} \rangle$ for $i = 1, 2$ and $j \in [\ell]$. Now we define

$$\widehat{\lambda} = A^{-1}\widehat{y}. \quad (\text{A.25})$$

Consider a truncation using the first l Hermite basis functions:

$$\widetilde{f}_i := \sum_{j=0}^{\ell-1} \widehat{\lambda}_{i,j} \psi_{j,c_i,\sigma},$$

and then normalizing the $\widetilde{f}_i$, we arrive at the following estimate for the probability density function associated with the component indexed by $i = 1$ and 2 :

$$\widehat{f}_i := \frac{(\widetilde{f}_i)_+}{\|(\widetilde{f}_i)_+\|_1}, \quad (\text{A.26})$$

where for any function $q : \mathbb{R} \rightarrow \mathbb{R}$, $(q(x))_+ := \max\{q(x), 0\}$, $\forall x \in \mathbb{R}$.

A.6 Proofs for Section 2.7.2: Upper bound on $\|\widehat{f}_i - f_i\|_1$

The goal of this section of the Appendix is to prove Lemma A.6.1.

Lemma A.6.1. *Let $f \in \mathcal{F}$ be as defined in (C2.3). Denote by P_f the probability measure with density f . Take n i.i.d. samples from P_f , and let $\widehat{f}$ be any density estimator of f from the n samples. Define $\Delta f := \widehat{f} - f$. For $i = 1, 2$, if $\ell \geq \lfloor 2er_i^2/\sigma^2 \rfloor + 2$, then the estimator $\widehat{f}_i$ defined in (A.26) satisfies*

$$\begin{aligned} \|\widehat{f}_i - f_i\|_1 &\leq \frac{1}{w_i} c_1(r, \sigma, r_1, r_2) \cdot \left(\frac{8 \max(r_1, r_2)}{r} \right)^\ell \cdot \ell^{5/4} + c_2(r_i, \sigma) \left(\frac{1}{2} \right)^\ell \\ &\quad + \frac{1}{w_i} c_3(r, \sigma) \cdot \|\Delta f\|_2 \cdot \exp(\ell \log \ell + c_4(r, \sigma) \ell + c_5 \log \ell) \end{aligned}$$

for some constants $c_1(r, \sigma, r_1, r_2) > 0$, $c_2(r_i, \sigma) > 0$, $c_3(r, \sigma) > 0$, $c_4(r, \sigma) > 0$, and $c_5 > 0$, where the dependence on parameters is indicated in parentheses.

This Lemma is a refinement of a result of [Tai and Aragam \(2023\)](#), who provide the order of the convergence rate and fix $r_1 = r_2 = \frac{1}{2}$, and $\sigma = 1$. We extend their result by making explicit the constants and their dependence on $r_i \in (0, +\infty)$, $i = 1, 2$, $\sigma \in (0, +\infty)$.

A.6.1 Technical lemmas

Recall that $f_i = \int_{c_i-r_i}^{c_i+r_i} g_{u,\sigma}(x) dV_i(u)$, for some $\sigma > 0$, where $V_i(u)$ is the mixing measure supported within $[c_i - r_i, c_i + r_i]$. We can expand f_i in terms of the Hermite functions: $f_i = \sum_{j=0}^{\infty} \alpha_{i,j} \psi_{j,c_i,\sigma}$, where $\alpha_{i,j} = \langle f_i, \psi_{j,c_i,\sigma} \rangle$, for $i = 1, 2$ and $j \in \mathbb{N}_0$. We obtain the following (cf. Lemma 16 of [Tai and Aragam \(2023\)](#)).

Lemma A.6.2. *For $i = 1, 2$, and any $j \in \mathbb{N}_0$, we have $|\alpha_{i,j}| \leq \left(\frac{\sqrt{2}r_i}{2\sigma} \right)^j \cdot \frac{1}{\sqrt{2\sigma j! \sqrt{\pi}}}$.*

Proof. We have

$$\begin{aligned}
\alpha_{i,j} &= \langle f_i, \psi_{j,c_i,\sigma} \rangle = \int_{-\infty}^{+\infty} \left(\int_{c_i-r_i}^{c_i+r_i} g_{u,\sigma}(x) dV_i(u) \right) \psi_{j,c_i,\sigma}(x) dx \\
&= \int_{c_i-r_i}^{c_i+r_i} \left(\int_{-\infty}^{+\infty} g_{u,\sigma}(x) \psi_{j,c_i,\sigma}(x) dx \right) dV_i(u) \\
&= \int_{-r_i}^{r_i} \langle g_{u,\sigma}(x), \psi_{j,0,\sigma} \rangle dV_i(c_i + u),
\end{aligned}$$

so, by (A.24),

$$|\alpha_{i,j}| \leq \int_{-r_i}^{r_i} |\langle \psi_{j,0,\sigma}, g_{u,\sigma} \rangle| dV_i(c_i + u) \leq \frac{1}{\sqrt{2\sigma}\sqrt{\pi}} \cdot \frac{\left(\frac{r_i}{\sqrt{2\sigma}}\right)^j}{\sqrt{j!}}. \quad (\text{A.27})$$

□

Recall the definition of $\widehat{\lambda}$ in (A.25) that $\widehat{\lambda} = A^{-1}\widehat{y}$, where $\widehat{y}$ is a $2l$ -dimensional vector whose entries are given by $\langle \widehat{f}, \psi_{j,c_i,\sigma} \rangle$ for $i = 1, 2$ and $j \in [\ell]$, while $\widehat{f}$ is a density estimator of f . Let $y^\Delta := \widehat{y} - y$, then we have

$$\begin{aligned}
\widehat{\lambda} &= A^{-1}\widehat{y} = A^{-1}y + A^{-1}y^\Delta \\
&= A^{-1} \left(A\lambda + \sum_{j=\ell}^{\infty} \lambda_{1,j} z_{1,j} + \sum_{j=\ell}^{\infty} \lambda_{2,j} z_{2,j} \right) + A^{-1}y^\Delta \\
&= \lambda + \underbrace{\sum_{j=\ell}^{\infty} \lambda_{1,j} A^{-1} z_{1,j} + \sum_{j=\ell}^{\infty} \lambda_{2,j} A^{-1} z_{2,j}}_{\mathcal{E}_t} + \underbrace{A^{-1}y^\Delta}_{\mathcal{E}_a}. \quad (\text{A.28})
\end{aligned}$$

Let

$$\mathcal{E}_t = \sum_{j=\ell}^{\infty} \lambda_{1,j} A^{-1} z_{1,j} + \sum_{j=\ell}^{\infty} \lambda_{2,j} A^{-1} z_{2,j}, \quad (\text{A.29})$$

$$\mathcal{E}_a = A^{-1}y^\Delta. \quad (\text{A.30})$$

$\mathcal{E}_t$ is the truncation error and $\mathcal{E}_a$ is the approximation error. As $\ell \rightarrow \infty$ and sample size $n \rightarrow \infty$, we expect $\mathcal{E}_t$ and $\mathcal{E}_a$ to vanish.

Recall that $\lambda_{i,j} = w_i \alpha_{i,j}$, hence the function

$$\widetilde{f}_i := \sum_{j=0}^{\ell-1} \widehat{\lambda}_{i,j} \psi_{j,c_i,\sigma} \quad (\text{A.31})$$

is expected to be an approximation of $w_i f_i$. This motivates the following lemma (which refines Lemma 17 of [Tai and Aragam \(2023\)](#)).

Lemma A.6.3. *Let $\Delta > 0$ and ℓ be a nonnegative integer. If $|\lambda_{i,j} - \widehat{\lambda}_{i,j}| < \Delta$, $\forall i = 1, 2, j \in [\ell]$, then $\|w_i f_i - \widetilde{f}_i\|_1 \leq \frac{2^{17/4} c_\psi \sqrt{\sigma}}{\Gamma(\frac{1}{4})^2} \cdot \Delta \cdot \ell^{5/4} + \frac{2^{11/2} c_\psi r_i w_i}{\Gamma(\frac{1}{4})^2 \pi^{1/4} \sigma} \cdot \left(\frac{1}{2}\right)^\ell$, for some $c_\psi > 0$ and any $\ell \geq \lfloor 2er_i^2/\sigma^2 \rfloor + 2$.*

Proof. From [Aptekarev et al. \(2012\)](#) and [Aptekarev et al. \(1995\)](#) we know $\forall j \in \mathbb{N}_+$,

$$\|\psi_{j-1}\|_1 = \frac{2^{17/4}}{\Gamma\left(\frac{1}{4}\right)^2} \cdot j^{1/4} (1 + o(1)),$$

so there exists a constant $c_\psi > 0$, s.t., $\|\psi_{j-1}\|_1 \leq (2^{17/4} c_\psi / \Gamma(\frac{1}{4})^2) j^{1/4}$, $\forall j \in \mathbb{N}_+$. Because $\|\psi_{j,\mu,\sigma}\|_1 = \|\frac{1}{\sqrt{\sigma}} \psi_j\left(\frac{x-\mu}{\sigma}\right)\|_1 = \sqrt{\sigma} \|\psi_j\|_1$, we obtain $\|\psi_{j-1,\mu,\sigma}\|_1 \leq \left(2^{17/4} c_\psi \sqrt{\sigma} / \Gamma(\frac{1}{4})^2\right) j^{1/4}$. It follows that

$$\begin{aligned} \|w_i f_i - \widetilde{f}_i\|_1 &= \left\| \sum_{j=0}^{\ell-1} (\lambda_{i,j} - \widehat{\lambda}_{i,j}) \psi_{j,c_i,\sigma} + \sum_{j=\ell}^{\infty} \lambda_{i,j} \psi_{j,c_i,\sigma} \right\|_1 \\ &\leq \sum_{j=0}^{\ell-1} |\lambda_{i,j} - \widehat{\lambda}_{i,j}| \cdot \|\psi_{j,c_i,\sigma}\|_1 + \sum_{j=\ell}^{\infty} |\lambda_{i,j}| \cdot \|\psi_{j,c_i,\sigma}\|_1. \end{aligned} \quad (\text{A.32})$$

For the first sum in the RHS of (A.32),

$$\begin{aligned} \sum_{j=0}^{\ell-1} |\lambda_{i,j} - \widehat{\lambda}_{i,j}| \cdot \|\psi_{j,c_i,\sigma}\|_1 &\leq \Delta \cdot \sum_{j=0}^{\ell-1} \|\psi_{j,c_i,\sigma}\|_1 \\ &\leq \Delta \cdot \ell \cdot \left(2^{17/4} c_\psi \sqrt{\sigma} / \Gamma\left(\frac{1}{4}\right)^2\right) \ell^{1/4} \end{aligned}$$

$$= \frac{2^{17/4} c_\psi \sqrt{\sigma}}{\Gamma\left(\frac{1}{4}\right)^2} \cdot \Delta \cdot \ell^{5/4} \quad (\text{A.33})$$

when ℓ is large enough. For the second sum in the RHS of (A.32),

$$\begin{aligned} \sum_{j=\ell}^{\infty} |\lambda_{i,j}| \cdot \|\psi_{j,c_i,\sigma}\|_1 &\leq \sum_{j=\ell}^{\infty} w_i |\alpha_{i,j}| \cdot \frac{2^{17/4} c_\psi \sqrt{\sigma}}{\Gamma\left(\frac{1}{4}\right)^2} \cdot (j+1)^{1/4} \\ &\leq \sum_{j=\ell}^{\infty} w_i \cdot \left(\frac{r_i}{\sqrt{2}\sigma}\right)^j \cdot \frac{1}{\sqrt{2\sigma j! \sqrt{\pi}}} \cdot \frac{2^{17/4} c_\psi \sqrt{\sigma}}{\Gamma\left(\frac{1}{4}\right)^2} \cdot 2^{1/4} j^{1/4} \\ &= \frac{2^{9/2} c_\psi w_i}{\Gamma\left(\frac{1}{4}\right)^2} \cdot \frac{1}{\sqrt{2\sqrt{\pi}}} \cdot \sum_{j=\ell}^{\infty} \frac{\left(\frac{r_i}{\sqrt{2}\sigma}\right)^j}{\sqrt{j!}} \cdot j^{1/4}. \end{aligned}$$

The series on the right hand side absolutely converges. Stirling's approximation yields $j! \geq \left(\frac{j}{e}\right)^j$,

so we have

$$\begin{aligned} \sum_{j=\ell}^{\infty} \frac{\left(\frac{r_i}{\sqrt{2}\sigma}\right)^j}{\sqrt{j!}} \cdot j^{1/4} &= \sum_{j=\ell}^{\infty} \sqrt{\frac{j^{1/2}}{j!}} \cdot \left(\frac{r_i}{\sqrt{2}\sigma}\right)^j \\ &\leq \sum_{j=\ell}^{\infty} \frac{1}{\sqrt{(j-1)!}} \cdot \left(\frac{r_i}{\sqrt{2}\sigma}\right)^j \\ &\leq \frac{r_i}{\sqrt{2}\sigma} \sum_{j=\ell}^{\infty} \frac{\left(\frac{r_i}{\sqrt{2}\sigma}\right)^{j-1}}{\sqrt{\left(\frac{j-1}{e}\right)^{j-1}}} \\ &= \frac{r_i}{\sqrt{2}\sigma} \sum_{j=\ell-1}^{\infty} \left(\frac{r_i \sqrt{e}}{\sqrt{2j}\sigma}\right)^j \end{aligned}$$

and when $j > 2er_i^2/\sigma^2$, $\frac{r_i \sqrt{e}}{\sqrt{2j}\sigma} < \frac{1}{2}$, therefore when $\ell \geq \lfloor 2er_i^2/\sigma^2 \rfloor + 2$,

$$\frac{r_i}{\sqrt{2}\sigma} \sum_{j=\ell-1}^{\infty} \left(\frac{r_i \sqrt{e}}{\sqrt{2j}\sigma}\right)^j \leq \frac{r_i}{\sqrt{2}\sigma} \cdot \left(\frac{1}{2}\right)^{\ell-2},$$

so the second sum in the RHS of (A.32) is upper bounded by

$$\sum_{j=\ell}^{\infty} |\lambda_{i,j}| \cdot \|\psi_{j,c_i,\sigma}\|_1 \leq \frac{2^{11/2} w_i r_i c_\psi}{\Gamma\left(\frac{1}{4}\right)^2 \pi^{1/4} \sigma} \cdot \left(\frac{1}{2}\right)^\ell. \quad (\text{A.34})$$

By combining (A.33) and (A.34), we get

$$\|w_i f_i - \tilde{f}_i\|_1 \leq \frac{2^{17/4} c_\psi \sqrt{\sigma}}{\Gamma\left(\frac{1}{4}\right)^2} \cdot \Delta \cdot \ell^{5/4} + \frac{2^{11/2} c_\psi r_i w_i}{\Gamma\left(\frac{1}{4}\right)^2 \pi^{1/4} \sigma} \cdot \left(\frac{1}{2}\right)^\ell.$$

□

Lemma A.6.3 gives an upper bound for $\|w_i f_i - \tilde{f}_i\|_1$ when $|\lambda_{i,j} - \widehat{\lambda}_{i,j}| < \Delta$, $\forall i = 1, 2, j \in [\ell]$.

Next we want to find the upper bound Δ for all $|\lambda_{i,j} - \widehat{\lambda}_{i,j}|$'s. Recall from (A.28),

$$\widehat{\lambda} - \lambda = \underbrace{\sum_{j=\ell}^{\infty} \lambda_{1,j} A^{-1} z_{1,j} + \sum_{j=\ell}^{\infty} \lambda_{2,j} A^{-1} z_{2,j}}_{\mathcal{E}_t} + \underbrace{A^{-1} y^\Delta}_{\mathcal{E}_a},$$

and we need to bound the entries of $\mathcal{E}_t$ and $\mathcal{E}_a$ respectively.

A.6.1.1 Bounding the entries of $\mathcal{E}_t$

We first bound the entries of $\mathcal{E}_t$. Let $\mathcal{E}_{t,i,j}$ be the vector $A^{-1} z_{i,j}$ for $i = 1, 2$ and $j \geq \ell$. Namely, we have

$$\mathcal{E}_t = \sum_{j=\ell}^{\infty} \lambda_{1,j} \mathcal{E}_{t,1,j} + \sum_{j=\ell}^{\infty} \lambda_{2,j} \mathcal{E}_{t,2,j}.$$

We only need to analyze the first sum $\sum_{j=\ell}^{\infty} \lambda_{1,j} \mathcal{E}_{t,1,j}$ and by symmetry we can conclude a similar bound for the second sum $\sum_{j=\ell}^{\infty} \lambda_{2,j} \mathcal{E}_{t,2,j}$. By Cramer's rule, the entry of $\mathcal{E}_{t,1,j} = A^{-1} z_{1,j}$ indexed at (i, k) is given by

$$\frac{\det\left(A^{(i,k) \rightarrow (1,j)}\right)}{\det(A)},$$

where $A^{(i,k) \rightarrow (1,j)}$ is the $2l$ -by- $2l$ matrix same as A except that the column indexed at (i, k) is replaced with $z_{1,j}$ for $i = 1, 2, k \in [\ell]$ and $j \geq \ell$. Lemma A.6.5 gives a bound on $|\det(A^{(i,k) \rightarrow (1,j)})|$ when comparing to $\det(A)$. Before proving this, we first state the Cauchy-Binet formula in the following lemma.

Lemma A.6.4 (Cauchy–Binet formula). *Let $b_0, b_1, \dots$ and $c_0, c_1, \dots$ be two infinite sequences of n dimensional vectors. For any (ordered) subset $S = \{s_0 < s_1 < \dots < s_{n-1}\}$ of $\mathbb{N}_0$, let B_S (resp. C_S) be the matrix that the column indexed at i is b_{s_i} (resp. c_{s_i}) for $i \in [n]$. The determinant of $A = \sum_{i=0}^{\infty} b_i c_i^\top$ is*

$$\det A = \sum_{S \subseteq \mathbb{N}_0, |S|=n} (\det B_S) \cdot (\det C_S)$$

if the RHS converges.

Lemma A.6.5 (Lemma 30, [Tai and Aragam \(2023\)](#)). *We can assume $c_2 > c_1$ w.l.o.g, and denote $r = c_2 - c_1$, Then for any $\ell \in \mathbb{N}_0, k \in [\ell]$ and $j \geq \ell$, we have*

$$\begin{aligned} \left| \det(A^{(1,k) \rightarrow (1,j)}) \right| &\leq e^{\frac{r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \cdot \det(A), \quad \text{and} \\ \left| \det(A^{(2,k) \rightarrow (1,j)}) \right| &\leq e^{\frac{5r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \cdot \det(A). \end{aligned}$$

Hence, $\forall k \in [\ell]$, the absolute values of the entries of $\lambda_{1,j} \mathcal{E}_{t,1,j}$ are bounded by,

$$\begin{aligned} |\lambda_{1,j}(\mathcal{E}_{t,1,j})_{1,k}| &\leq w_1 e^{\frac{r^2}{4\sigma^2}} \frac{1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \left(\frac{2r_1}{r} \right)^j \cdot 4^\ell, \\ |\lambda_{1,j}(\mathcal{E}_{t,1,j})_{2,k}| &\leq w_1 e^{\frac{5r^2}{4\sigma^2}} \frac{1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \left(\frac{2r_1}{r} \right)^j \cdot 4^\ell. \end{aligned}$$

Proof. We first observe that the matrices $A, A^{(1,k) \rightarrow (1,j)}$ and $A^{(2,k) \rightarrow (1,j)}$ can be decomposed as

$$A = V^\top V, \quad A^{(1,k) \rightarrow (1,j)} = V^\top V^{(1,k) \rightarrow (1,j)} \quad \text{and} \quad A^{(2,k) \rightarrow (1,j)} = V^\top V^{(2,k) \rightarrow (1,j)}$$

where we abuse the notation to define $V, V^{(1,k) \rightarrow (1,j)}$ and $V^{(2,k) \rightarrow (1,j)}$ as follows. Let V be the

$|\mathbb{N}_0|$ -by- 2ℓ matrix whose column indexed at (i, k) is the $|\mathbb{N}_0|$ dimensional vector $v^{(i,k)}$ for $i = 1, 2$ and $k \in [\ell]$. Here, for $i = 1, 2$ and $j \in \mathbb{N}_0$, $v^{(i,j)}$ is the $|\mathbb{N}_0|$ dimensional vector whose k -th entry is $\langle \psi_{j,c_i,\sigma}, \psi_{k,c_1,\sigma} \rangle$ for $k \in \mathbb{N}_0$. In particular, $v^{(1,j)}$ is the zero vector except that the j -th entry is 1. For example, when $\ell = 2$,

$$V = \begin{bmatrix} 1 & 0 & \langle \psi_{0,r,\sigma}, \psi_{0,0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{0,0,\sigma} \rangle \\ 0 & 1 & \langle \psi_{0,r,\sigma}, \psi_{1,0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{1,0,\sigma} \rangle \\ 0 & 0 & \langle \psi_{0,r,\sigma}, \psi_{2,0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{2,0,\sigma} \rangle \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

It is easy to check that, by the orthogonality of Hermite functions,

$$V^T V = \begin{bmatrix} 1 & 0 & \langle \psi_{0,r,\sigma}, \psi_{0,0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{0,0,\sigma} \rangle \\ 0 & 1 & \langle \psi_{0,r,\sigma}, \psi_{1,0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{1,0,\sigma} \rangle \\ \langle \psi_{0,0,\sigma}, \psi_{0,r,\sigma} \rangle & \langle \psi_{1,0,\sigma}, \psi_{0,r,\sigma} \rangle & 1 & 0 \\ \langle \psi_{0,0,\sigma}, \psi_{1,r,\sigma} \rangle & \langle \psi_{1,0,\sigma}, \psi_{1,r,\sigma} \rangle & 0 & 1 \end{bmatrix} = A.$$

Define $V^{(i,k) \rightarrow (1,j)}$ to be the same matrix as V except that the column indexed at (i, k) is replaced with $v^{(1,j)}$ for $i = 1, 2$, $k \in [\ell]$ and $j \geq \ell$. For example, when $\ell = 2$, $i = 1$, $k = 1$ and $j = 3$,

$$V^{(i,k) \rightarrow (1,j)} = \begin{bmatrix} 1 & 0 & \langle \psi_{0,r,\sigma}, \psi_{0,0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{0,0,\sigma} \rangle \\ 0 & 0 & \langle \psi_{0,r,\sigma}, \psi_{1,0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{1,0,\sigma} \rangle \\ 0 & 0 & \langle \psi_{0,r,\sigma}, \psi_{2,0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{2,0,\sigma} \rangle \\ 0 & 1 & \langle \psi_{0,r,\sigma}, \psi_{3,0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{3,0,\sigma} \rangle \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

and when $\ell = 2$, $i = 2$, $k = 1$ and $j = 3$,

$$V^{(i,k) \rightarrow (1,j)} = \begin{bmatrix} 1 & 0 & \langle \psi_{0,r,\sigma}, \psi_{0,0,\sigma} \rangle & 0 \\ 0 & 1 & \langle \psi_{0,r,\sigma}, \psi_{1,0,\sigma} \rangle & 0 \\ 0 & 0 & \langle \psi_{0,r,\sigma}, \psi_{2,0,\sigma} \rangle & 0 \\ 0 & 0 & \langle \psi_{0,r,\sigma}, \psi_{3,0,\sigma} \rangle & 1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Again, we can check that $A^{(i,k) \rightarrow (1,j)} = V^T V^{(i,k) \rightarrow (1,j)}$.

By the Cauchy-Binet formula (Lemma A.6.4) and expanding the determinant along the columns indexed at $(1, k)$, we have

$$\det(A) = \sum_{K \in \mathcal{K}} \det(U_K)^2,$$

where $\mathcal{K}$ is the set of subsets of $\mathbb{N}_0$ of size ℓ whose elements are larger than or equal to ℓ , i.e. $\mathcal{K} = \{K \mid K = \{\ell \leq a_0 < \dots < a_{\ell-1}\}\}$ and U_K is the ℓ -by- ℓ matrix whose (i, j) -entry is $\langle \psi_{c_j, r, \sigma}, \psi_{b_i, 0, \sigma} \rangle$ for $b_i \in K$ and $c_j \in [\ell]$. For example, when $\ell = 2$,

$$\begin{aligned} \det(A) = & \left(\det \begin{bmatrix} \langle \psi_{0,r,\sigma}, \psi_{\mathbf{2},0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{\mathbf{2},0,\sigma} \rangle \\ \langle \psi_{0,r,\sigma}, \psi_{\mathbf{3},0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{\mathbf{3},0,\sigma} \rangle \end{bmatrix} \right)^2 + \left(\det \begin{bmatrix} \langle \psi_{0,r,\sigma}, \psi_{\mathbf{2},0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{\mathbf{2},0,\sigma} \rangle \\ \langle \psi_{0,r,\sigma}, \psi_{\mathbf{4},0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{\mathbf{4},0,\sigma} \rangle \end{bmatrix} \right)^2 + \dots \\ & + \left(\det \begin{bmatrix} \langle \psi_{0,r,\sigma}, \psi_{\mathbf{3},0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{\mathbf{3},0,\sigma} \rangle \\ \langle \psi_{0,r,\sigma}, \psi_{\mathbf{4},0,\sigma} \rangle & \langle \psi_{1,r,\sigma}, \psi_{\mathbf{4},0,\sigma} \rangle \end{bmatrix} \right)^2 + \dots \\ & + \dots \end{aligned}$$

where the numbers in red represent the set K .

We first give a bound on $|\det(A^{(1,k) \rightarrow (1,j)})|$. By a similar argument, we have

$$|\det(A^{(1,k) \rightarrow (1,j)})| \leq \sum_{K \in \mathcal{K}_j} |\det(U_{K \cup \{k\}})| |\det(U_{K \cup \{j\}})|$$

where $\mathcal{K}_j$ is the set of subsets of $\mathbb{N}_0$ of size $\ell - 1$ whose elements are larger than or equal to ℓ

and not equal to j , i.e. $\mathcal{K}_j = \{K \mid K = \{\ell \leq a_0 < \dots < a_{\ell-2} \text{ and } a_i \neq j\}\}$. Furthermore, by Cauchy-Schwarz inequality, we have

$$\begin{aligned} \left| \det(A^{(1,k) \rightarrow (1,j)}) \right| &\leq \sqrt{\left(\sum_{K \in \mathcal{K}_j} \det(U_{K \cup \{k\}})^2 \right) \left(\sum_{K \in \mathcal{K}_j} \det(U_{K \cup \{j\}})^2 \right)} \\ &\leq \sqrt{\left(\sum_{K \in \mathcal{K}_j} \det(U_{K \cup \{k\}})^2 \right) \cdot \det(A)}. \end{aligned} \quad (\text{A.35})$$

The last line is due to the fact that the subset $K \cup \{j\} \in \mathcal{K}$, $\forall K \in \mathcal{K}_j$. By Lemma A.6.8, we have

$$\sum_{K \in \mathcal{K}_j} \det(U_{K \cup \{k\}})^2 \leq \left(e^{\frac{r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \right)^2 \cdot \det(A).$$

and, by plugging it into (A.35), we have

$$\left| \det(A^{(1,k) \rightarrow (1,j)}) \right| \leq e^{\frac{r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \cdot \det(A). \quad (\text{A.36})$$

We now give a bound on $\left| \det(A^{(2,k) \rightarrow (1,j)}) \right|$. By a similar argument, we have

$$\left| \det(A^{(2,k) \rightarrow (1,j)}) \right| \leq \sum_{K \in \mathcal{K}_j} \left| \det U_K^{(-k)} \right| \cdot \left| \det(U_{K \cup \{j\}}) \right|$$

where $U_K^{(-k)}$ is the $(\ell - 1)$ -by- $(\ell - 1)$ matrix whose (i, j) -entry is $\langle \psi_{c_j, r, \sigma}, \psi_{b_i, 0, \sigma} \rangle$ for $b_i \in K$ and $c_j \in [\ell] \setminus \{k\}$. Furthermore, by Cauchy-Schwarz inequality, we have

$$\begin{aligned} \left| \det(A^{(2,k) \rightarrow (1,j)}) \right| &\leq \sum_{K \in \mathcal{K}_j} \left| \det U_K^{(-k)} \right| \cdot \left| \det(U_{K \cup \{j\}}) \right| \\ &\leq \sqrt{\sum_{K \in \mathcal{K}_j} \det(U_K^{(-k)})^2 \cdot \det(A)}. \end{aligned} \quad (\text{A.37})$$

The last line is due to the fact that the subset $K \cup \{j\}$ is in $\mathcal{K}$ for each $K \in \mathcal{K}_j$. By Lemma A.6.9

below, we have

$$\sum_{K \in \mathcal{K}_j} \det \left(U_K^{(-k)} \right)^2 \leq \left(e^{\frac{5}{4\sigma^2} r^2} 4^\ell \cdot \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \right)^2 \cdot \det(A),$$

and, by plugging it into (A.37), we have

$$\left| \det \left(A^{(2,k) \rightarrow (1,j)} \right) \right| \leq e^{\frac{5}{4\sigma^2} r^2} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \cdot \det(A). \quad (\text{A.38})$$

By combining the above results with the upper bound of $\alpha_{i,j}$ given in (A.27), we obtain $\forall k \in [\ell]$

$$\begin{aligned} |\lambda_{1,j}(\mathcal{E}_{t,1,j})_{1,k}| &= |\lambda_{1,j}| \cdot \left| \left(A^{-1} z_{1,j} \right)_{1,k} \right| = w_1 \cdot |\alpha_{1,j}| \cdot \frac{\left| \det \left(A^{(1,k) \rightarrow (1,j)} \right) \right|}{|\det(A)|} \\ &\leq w_1 \cdot \left(\frac{r_1}{\sqrt{2}\sigma} \right)^j \cdot \frac{1}{\sqrt{2j!\sigma\sqrt{\pi}}} \cdot e^{\frac{r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \\ &= w_1 e^{\frac{r^2}{4\sigma^2}} \frac{1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \left(\frac{2r_1}{r} \right)^j \cdot 4^\ell. \end{aligned}$$

and

$$\begin{aligned} |\lambda_{1,j}(\mathcal{E}_{t,1,j})_{2,k}| &= |\lambda_{1,j}| \cdot \left| \left(A^{-1} z_{1,j} \right)_{2,k} \right| = w_1 \cdot |\alpha_{1,j}| \cdot \frac{\left| \det \left(A^{(2,k) \rightarrow (1,j)} \right) \right|}{|\det(A)|} \\ &\leq w_1 \cdot \left(\frac{r_1}{\sqrt{2}\sigma} \right)^j \cdot \frac{1}{\sqrt{2j!\sigma\sqrt{\pi}}} \cdot e^{\frac{5r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \\ &= w_1 e^{\frac{5r^2}{4\sigma^2}} \frac{1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \left(\frac{2r_1}{r} \right)^j \cdot 4^\ell. \end{aligned}$$

This concludes the proof of the lemma. □

By the similar argument in Lemma A.6.5, for any $\ell \in \mathbb{N}_0$, $k \in [\ell]$ and $j \geq \ell$, we have

$$\begin{aligned} \left| \det \left(A^{(1,k) \rightarrow (2,j)} \right) \right| &\leq e^{\frac{5r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \cdot \det(A), \quad \text{and} \\ \left| \det \left(A^{(2,k) \rightarrow (2,j)} \right) \right| &\leq e^{\frac{r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \cdot \det(A). \end{aligned}$$

Hence, $\forall k \in [\ell]$, the absolute values of the entries of $\lambda_{2,j}\mathcal{E}_{t,2,j}$ are bounded by,

$$|\lambda_{2,j}(\mathcal{E}_{t,2,j})_{1,k}| \leq w_2 e^{\frac{5r^2}{4\sigma^2}} \frac{1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \left(\frac{2r_2}{r}\right)^j \cdot 4^\ell, \quad (\text{A.39})$$

$$|\lambda_{2,j}(\mathcal{E}_{t,2,j})_{2,k}| \leq w_2 e^{\frac{r^2}{4\sigma^2}} \frac{1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \left(\frac{2r_2}{r}\right)^j \cdot 4^\ell. \quad (\text{A.40})$$

Now we are ready to discuss under what conditions each entry of $\mathcal{E}_t$ has an upper bound, as well as the conditions under which this upper bound approaches zero as $\ell \rightarrow \infty$.

Lemma A.6.6. *Define $\mathcal{E}_{t,1} := \sum_{j=\ell}^{\infty} \lambda_{1,j}\mathcal{E}_{t,1,j}$ and $\mathcal{E}_{t,2} := \sum_{j=\ell}^{\infty} \lambda_{2,j}\mathcal{E}_{t,2,j}$. When $2\max(r_1, r_2) < r$, we have $\forall k \in [\ell]$,*

$$\begin{aligned} |(\mathcal{E}_{t,1})_{1,k}| &\leq \frac{w_1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \frac{r e^{\frac{1}{4\sigma^2}r^2}}{r - 2r_1} \cdot \left(\frac{8r_1}{r}\right)^\ell, \\ |(\mathcal{E}_{t,1})_{2,k}| &\leq \frac{w_1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \frac{r e^{\frac{5}{4\sigma^2}r^2}}{r - 2r_1} \cdot \left(\frac{8r_1}{r}\right)^\ell, \\ |(\mathcal{E}_{t,2})_{1,k}| &\leq \frac{w_2}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \frac{r e^{\frac{5}{4\sigma^2}r^2}}{r - 2r_2} \cdot \left(\frac{8r_2}{r}\right)^\ell, \\ |(\mathcal{E}_{t,2})_{2,k}| &\leq \frac{w_2}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \frac{r e^{\frac{1}{4\sigma^2}r^2}}{r - 2r_2} \cdot \left(\frac{8r_2}{r}\right)^\ell. \end{aligned}$$

Furthermore, when $8\max(r_1, r_2) < r$, we have for $i = 1, 2$, and $\forall k \in [\ell]$,

$$|(\mathcal{E}_t)_{i,k}| \leq |(\mathcal{E}_{t,1})_{i,k}| + |(\mathcal{E}_{t,2})_{i,k}| \leq \frac{r e^{\frac{5}{4\sigma^2}r^2}}{\sqrt{2\sigma\sqrt{\pi}}(r - 2\max(r_1, r_2))} \left(\frac{8\max(r_1, r_2)}{r}\right)^\ell \rightarrow 0, \text{ as } \ell \rightarrow \infty.$$

Proof. By Lemma A.6.5, when $r_1 < \frac{r}{2}$, $\forall k \in [\ell]$,

$$\begin{aligned} |(\mathcal{E}_{t,1})_{1,k}| &\leq \sum_{j=\ell}^{\infty} |\lambda_{1,j}| \cdot |(\mathcal{E}_{t,1,j})_{1,k}| \\ &\leq w_1 e^{\frac{r^2}{4\sigma^2}} \frac{4^\ell}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \sum_{j=\ell}^{\infty} \left(\frac{2r_1}{r}\right)^j \end{aligned}$$

$$\begin{aligned}
&= w_1 e^{\frac{r^2}{4\sigma^2}} \frac{4^\ell}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \frac{r}{r-2r_1} \cdot \left(\frac{2r_1}{r}\right)^\ell \\
&= \frac{w_1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \frac{r e^{\frac{1}{4\sigma^2}r^2}}{r-2r_1} \cdot \left(\frac{8r_1}{r}\right)^\ell,
\end{aligned}$$

and similarly,

$$|(\mathcal{E}_{t,1})_{2,k}| \leq \sum_{j=\ell}^{\infty} |\lambda_{1,j}| \cdot |(\mathcal{E}_{t,1,j})_{2,k}| \leq \frac{w_1}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \frac{r e^{\frac{5}{4\sigma^2}r^2}}{r-2r_1} \cdot \left(\frac{8r_1}{r}\right)^\ell.$$

When $r_2 < \frac{r}{2}$, by (A.39) and (A.40), we obtain $\forall k \in [\ell]$,

$$\begin{aligned}
|(\mathcal{E}_{t,2})_{1,k}| &\leq \sum_{j=\ell}^{\infty} |\lambda_{2,j}| \cdot |(\mathcal{E}_{t,2,j})_{1,k}| = \frac{w_2}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \frac{r e^{\frac{5}{4\sigma^2}r^2}}{r-2r_2} \cdot \left(\frac{8r_2}{r}\right)^\ell \\
|(\mathcal{E}_{t,2})_{2,k}| &\leq \sum_{j=\ell}^{\infty} |\lambda_{2,j}| \cdot |(\mathcal{E}_{t,2,j})_{2,k}| = \frac{w_2}{\sqrt{2\sigma\sqrt{\pi}}} \cdot \frac{r e^{\frac{1}{4\sigma^2}r^2}}{r-2r_2} \cdot \left(\frac{8r_2}{r}\right)^\ell.
\end{aligned}$$

Therefore, for $i = 1, 2$, and $\forall k \in [\ell]$, we have

$$|(\mathcal{E}_t)_{i,k}| \leq |(\mathcal{E}_{t,1})_{i,k}| + |(\mathcal{E}_{t,2})_{i,k}| \leq \frac{r e^{\frac{5}{4\sigma^2}r^2}}{\sqrt{2\sigma\sqrt{\pi}}(r-2\max(r_1, r_2))} \left(\frac{8\max(r_1, r_2)}{r}\right)^\ell,$$

and when $8\max(r_1, r_2) < r$, the right hand side goes to zero as $\ell \rightarrow \infty$.

□

A.6.1.2 Bounding the entries of $\mathcal{E}_a$

Next, we turn to $\mathcal{E}_a$. Recall that $\widehat{f}$ is a density estimator of the true density $f = w_1 f_1 + w_2 f_2$, and we define y^Δ to be the 2ℓ -dimensional vector whose entries are given by $\langle \Delta f, \psi_{j,c_i,\sigma} \rangle$ for $i = 1, 2$ and $j \in [\ell]$, where $\Delta f = \widehat{f} - f$. We defined in (A.30) that:

$$\mathcal{E}_a = A^{-1} y^\Delta.$$

By Cramer's rule, the entry of $\mathcal{E}_a$ indexed at (i, k) is

$$\frac{\det \left(A^{(i,k) \rightarrow \Delta} \right)}{\det(A)},$$

where $A^{(i,k) \rightarrow \Delta}$ is the 2ℓ -by- 2ℓ matrix same as A except that the column indexed at (i, k) is replaced with y^Δ for $i = 1, 2$ and $k \in [\ell]$. Lemma A.6.7 gives a bound on $|\det(A^{(1,k) \rightarrow \Delta})|$ when comparing to $\det(A)$.

Lemma A.6.7 (Lemma 31, [Tai and Aragam \(2023\)](#)). *For any $\ell \in \mathbb{N}_0$, $i = 1, 2$ and $k \in [\ell]$, we have*

$$|(\mathcal{E}_a)_{i,k}| = \frac{|\det \left(A^{(i,k) \rightarrow \Delta} \right)|}{\det(A)} \leq \|\Delta f\|_2 \cdot c_1(r, \sigma) \exp(\ell \log \ell + c_2(r, \sigma)\ell + c_3 \log \ell),$$

for some constants $c_1(r, \sigma) > 0$ and $c_2(r, \sigma)$ depending on r and σ , and some independent constant $c_3 > 0$.

Proof. We first observe that the matrix $A^{(i,k) \rightarrow \Delta}$ can be decomposed as

$$A^{(i,k) \rightarrow \Delta} = V^\top V^{(i,k) \rightarrow \Delta}$$

where $V^{(i,k) \rightarrow \Delta}$ is the $|\mathbb{N}_0|$ -by- 2ℓ matrix whose column indexed at (i, k) is replaced with the $|\mathbb{N}_0|$ dimensional vector v^Δ for $i = 1, 2$ and $k \in [\ell]$. Here, v^Δ is the $|\mathbb{N}_0|$ dimensional vector whose k -th entry is $\langle \Delta f, \psi_{k,c_1,\sigma} \rangle$ for $k \in \mathbb{N}_0$. Recall that $\Delta f = \widehat{f} - f$. For example, when $\ell = 2$, $i = 1$ and $k = 1$,

$$V^{(i,k) \rightarrow \Delta} = \begin{bmatrix} 1 & \langle \Delta f, \psi_{0,c_1,\sigma} \rangle & \langle \psi_{0,c_2,\sigma}, \psi_{0,c_1,\sigma} \rangle & \langle \psi_{1,c_2,\sigma}, \psi_{0,c_1,\sigma} \rangle \\ 0 & \langle \Delta f, \psi_{1,c_1,\sigma} \rangle & \langle \psi_{0,c_2,\sigma}, \psi_{1,c_1,\sigma} \rangle & \langle \psi_{1,c_2,\sigma}, \psi_{1,c_1,\sigma} \rangle \\ 0 & \langle \Delta f, \psi_{2,c_1,\sigma} \rangle & \langle \psi_{0,c_2,\sigma}, \psi_{2,c_1,\sigma} \rangle & \langle \psi_{1,c_2,\sigma}, \psi_{2,c_1,\sigma} \rangle \\ 0 & \langle \Delta f, \psi_{3,c_1,\sigma} \rangle & \langle \psi_{0,c_2,\sigma}, \psi_{3,c_1,\sigma} \rangle & \langle \psi_{1,c_2,\sigma}, \psi_{3,c_1,\sigma} \rangle \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Recall that V is the $|\mathbb{N}_0|$ -by- 2ℓ matrix whose column indexed at (i, k) is the $|\mathbb{N}_0|$ dimensional vector $v^{(i,k)}$ for $i = 1, 2$ and $k \in [\ell]$. Here, for $i = 1, 2$ and $j \in \mathbb{N}_0$, $v^{(i,j)}$ is the $|\mathbb{N}_0|$ dimensional vector whose k -th entry is $\langle \psi_{j,c_i,\sigma}, \psi_{k,c_1,\sigma} \rangle$ for $k \in \mathbb{N}_0$.

We first give a bound on $|\det(A^{(1,k) \rightarrow \Delta})|$. By Cauchy-Binet formula (Lemma A.6.4) and expanding the determinant along the columns with a single 1, we have

$$|\det(A^{(1,k) \rightarrow \Delta})| \leq \sum_{K \in \mathcal{K}} |\det(U_K)| |\det(U_K^\Delta)|,$$

where $K \in \mathcal{K} = \{K | K = \{\ell \leq a_0 < \dots < a_{\ell-1}\}\}$, U_K^Δ is the $(\ell + 1)$ -by- $(\ell + 1)$ matrix whose (i, j) -entry is
$$\begin{cases} \langle \psi_{j-1,c_2,\sigma}, \psi_{b_i,c_1,\sigma} \rangle, & j \in [\ell + 1] \setminus \{0\} \\ \langle \Delta f, \psi_{b_i,c_1,\sigma} \rangle, & j = 0 \end{cases}$$
 for $b_i \in \{k\} \cup K$. Furthermore, by Cauchy-Schwarz inequality, we have

$$\begin{aligned} |\det(A^{(1,k) \rightarrow \Delta})| &\leq \sqrt{\left(\sum_{K \in \mathcal{K}} \det(U_K)^2\right) \left(\sum_{K \in \mathcal{K}} \det(U_K^\Delta)^2\right)} \\ &= \sqrt{\det(A) \cdot \sum_{K \in \mathcal{K}} \det(U_K^\Delta)^2}. \end{aligned}$$

We first expand the determinant $\det(U_K^\Delta)$ along the column indexed at $j = 0$.

$$\begin{aligned} &\det(U_K^\Delta)^2 \\ &\leq \left(|\langle \Delta f, \psi_{k,c_1,\sigma} \rangle| \cdot |\det(U_K)| + \sum_{c=0}^{\ell-1} |\langle \Delta f, \psi_{a_c,c_1,\sigma} \rangle| \cdot |\det(U_{K \cup \{k\} \setminus \{a_c\}})| \right)^2 \\ &\leq (\ell + 1) \left(|\langle \Delta f, \psi_{k,c_1,\sigma} \rangle|^2 \cdot |\det(U_K)|^2 + \sum_{c=0}^{\ell-1} |\langle \Delta f, \psi_{a_c,c_1,\sigma} \rangle|^2 \cdot |\det(U_{K \cup \{k\} \setminus \{a_c\}})|^2 \right). \quad (\text{A.41}) \end{aligned}$$

The last inequality is obtained by Cauchy-Schwarz inequality. Now we bound the second term on

the right hand side.

$$\begin{aligned}
\sum_{K \in \mathcal{K}} \sum_{c=0}^{\ell-1} |\langle \Delta f, \psi_{a_c, c_1, \sigma} \rangle|^2 \cdot |\det(U_{K \cup \{k\} \setminus \{a_c\}})|^2 &= \sum_{K \in \mathcal{K}'} \left(\sum_{j \geq \ell, j \notin K} \langle \Delta f, \psi_{j, c_1, \sigma} \rangle^2 \right) |\det(U_{K \cup \{k\}})|^2 \\
&\leq \|\Delta f\|_2^2 \sum_{K \in \mathcal{K}'} |\det(U_{K \cup \{k\}})|^2, \tag{A.42}
\end{aligned}$$

where $\mathcal{K}'$ is the set of subsets of $\mathbb{N}_0$ of size $\ell - 1$ whose elements are larger than or equal to ℓ , i.e.,

$\mathcal{K}' = \{K | K = \{\ell \leq a_0 < \dots < a_{\ell-2}\}\}$. by combining (A.41) and (A.42), we obtain

$$\begin{aligned}
&\det(A) \cdot \sum_{K \in \mathcal{K}} \det(U_K^\Delta)^2 \\
&\leq (\ell + 1) \det(A) \left(|\langle \Delta f, \psi_{k,0} \rangle|^2 \cdot \sum_{K \in \mathcal{K}} |\det(U_K)|^2 + \|\Delta f\|_2^2 \sum_{K \in \mathcal{K}'} |\det(U_{K \cup \{k\}})|^2 \right) \\
&= (\ell + 1) \det(A) \left(|\langle \Delta f, \psi_{k,0} \rangle|^2 \det(A) + \|\Delta f\|_2^2 \sum_{K \in \mathcal{K}'} |\det(U_{K \cup \{k\}})|^2 \right).
\end{aligned}$$

Since each $K \in \mathcal{K}'$ has $\ell - 1$ elements, then $\{\ell, \dots, 2\ell - 1\} \not\subseteq K$. We have

$$\sum_{K \in \mathcal{K}'} |\det(U_{K \cup \{k\}})|^2 \leq \sum_{j=\ell}^{2\ell-1} \sum_{K \in \mathcal{K}_j} |\det(U_{K \cup \{k\}})|^2.$$

Recall that $\mathcal{K}_j$ is the set of subsets of $\mathbb{N}_0$ of size $\ell - 1$ whose elements are larger than or equal to ℓ and not equal to j , i.e. $\mathcal{K}_j = \{K | K = \{\ell \leq a_0 < \dots < a_{\ell-2} \text{ and } a_i \neq j\}\}$. By Lemma A.6.8, we have

$$\begin{aligned}
\sum_{K \in \mathcal{K}'} |\det(U_{K \cup \{k\}})|^2 &\leq \sum_{j=\ell}^{2\ell-1} \left(e^{\frac{r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \right)^2 \cdot \det(A) \\
&= \left(e^{\frac{1}{4\sigma^2} r^2} 4^\ell \right)^2 \cdot \det(A) \cdot \sum_{j=\ell}^{2\ell-1} (j!) \cdot \left(\frac{8\sigma^2}{r^2} \right)^j \\
&\leq \left(e^{\frac{1}{4\sigma^2} r^2} 4^\ell \right)^2 \det(A) \cdot (2\ell)! \cdot \sum_{j=\ell}^{2\ell-1} \left(\frac{8\sigma^2}{r^2} \right)^j.
\end{aligned}$$

From Stirling's approximation we have $(2\ell)! \leq \sqrt{2\pi \cdot (2\ell)} \cdot e^{\frac{1}{12 \cdot (2\ell)}} \left(\frac{2\ell}{e}\right)^{2\ell}$, hence

$$\begin{aligned}
& \sum_{K \in \mathcal{K}'} |\det(U_{K \cup \{k\}})|^2 \\
& \leq \begin{cases} \left(e^{\frac{1}{4\sigma^2} r^2} 4^\ell\right)^2 \det(A) \cdot \sqrt{4\pi\ell} \cdot e^{\frac{1}{24\ell} - 2\ell} (2\ell)^{2\ell} \cdot \left(\frac{8\sigma^2}{r^2}\right)^\ell \left(1 - \left(\frac{8\sigma^2}{r^2}\right)^\ell\right) \cdot \frac{r^2}{r^2 - 8\sigma^2}, & r \neq 2\sqrt{2}\sigma \\ \left(e^{\frac{1}{4\sigma^2} r^2} 4^\ell\right)^2 \det(A) \cdot \sqrt{4\pi\ell} \cdot e^{\frac{1}{24\ell} - 2\ell} (2\ell)^{2\ell} \cdot \ell, & r = 2\sqrt{2}\sigma \end{cases} \\
& \leq \begin{cases} 2\sqrt{\pi} e^{\frac{r^2}{2\sigma^2} + 1} \left(\frac{r^2}{r^2 - 8\sigma^2}\right) \det(A) \cdot \exp\left(2\ell \log \ell + \left(\log \frac{512\sigma^2}{r^2} - 2\right)\ell + \frac{1}{2} \log \ell\right), & r > 2\sqrt{2}\sigma \\ 2\sqrt{\pi} e^{\frac{r^2}{2\sigma^2} + 1} \left(\frac{r^2}{8\sigma^2 - r^2}\right) \det(A) \cdot \exp\left(2\ell \log \ell + \left(\log \frac{4096\sigma^4}{r^4} - 2\right)\ell + \frac{1}{2} \log \ell\right), & r < 2\sqrt{2}\sigma \\ 2\sqrt{\pi} e^{\frac{r^2}{2\sigma^2} + 1} \det(A) \cdot \exp\left(2\ell \log \ell + (\log 64 - 2)\ell + \frac{3}{2} \log \ell\right), & r = 2\sqrt{2}\sigma. \end{cases}
\end{aligned}$$

By combining all above results, we obtain

$$\begin{aligned}
& \det(A) \cdot \sum_{K \in \mathcal{K}} \det(U_K^\Delta)^2 \\
& \leq (\ell + 1) \det(A)^2 \|\Delta f\|_2^2 + (\ell + 1) \det(A)^2 \|\Delta f\|_2^2 c_1(r, \sigma) \exp(2\ell \log \ell + c_2(r, \sigma)\ell + c_3 \log \ell) \\
& \leq \det(A)^2 \|\Delta f\|_2^2 \cdot c_1(r, \sigma) \exp(2\ell \log \ell + c_2(r, \sigma)\ell + c_3 \log \ell),
\end{aligned}$$

for some constants $c_1(r, \sigma) > 0$ and $c_3 > 0$. Therefore

$$\begin{aligned}
|\det(A^{(1,k) \rightarrow \Delta})| & \leq \sqrt{\det(A) \cdot \sum_{K \in \mathcal{K}} \det(U_K^\Delta)^2} \\
& \leq \det(A) \cdot \|\Delta f\|_2 \cdot c_1(r, \sigma) \exp(\ell \log \ell + c_2(r, \sigma)\ell + c_3 \log \ell). \quad (\text{A.43})
\end{aligned}$$

We now give a bound on $|\det(A^{(2,k) \rightarrow \Delta})|$. By a similar argument, we have

$$|\det(A^{(2,k) \rightarrow \Delta})| \leq \sum_{K \in \mathcal{K}} |\det(U_K)| |\det(U_K^{k \rightarrow \Delta})|$$

where, for any $k \in [\ell]$, $U_K^{k \rightarrow \Delta}$ is the ℓ -by- ℓ matrix whose (i, j) -entry is

$$\begin{cases} \langle \psi_{j,c_2,\sigma}, \psi_{b_i,c_1,\sigma} \rangle & \text{if } j \in [\ell] \setminus k \\ \langle \Delta f, \psi_{b_i,c_1,\sigma} \rangle & \text{if } j = k \end{cases}$$

for $b_i \in K$. Furthermore, by the Cauchy-Schwarz inequality, we have

$$|\det(A^{(2,k) \rightarrow \Delta})| \leq \sqrt{\left(\sum_{K \in \mathcal{K}} \det(U_K)^2 \right) \left(\sum_{K \in \mathcal{K}} \det(U_K^{k \rightarrow \Delta})^2 \right)} \leq \sqrt{\sum_{K \in \mathcal{K}} \det(U_K^{k \rightarrow \Delta})^2 \cdot \det(A)}. \quad (\text{A.44})$$

For each $K = \{\ell \leq a_0 < \dots < a_{\ell-1}\} \in \mathcal{K}$, we expand the determinants along the column indexed at Δf .

$$\begin{aligned} \det(U_K^{k \rightarrow \Delta})^2 &\leq \left(\sum_{c=0}^{\ell-1} |\langle \Delta f, \psi_{a_c,c_1,\sigma} \rangle| |\det(U_{K \setminus \{a_c\}}^{(-k)})| \right)^2 \\ &\leq \ell \cdot \left(\sum_{c=0}^{\ell-1} |\langle \Delta f, \psi_{a_c,c_1,\sigma} \rangle|^2 |\det(U_{K \setminus \{a_c\}}^{(-k)})|^2 \right). \end{aligned}$$

We now consider the summation $\sum_{K \in \mathcal{K}} \sum_{c=0}^{\ell-1} |\langle \Delta f, \psi_{a_c,c_1,\sigma} \rangle|^2 |\det(U_{K \setminus \{a_c\}}^{(-k)})|^2$ and we have

$$\begin{aligned} \sum_{K \in \mathcal{K}} \sum_{c=0}^{\ell-1} |\langle \Delta f, \psi_{a_c,c_1,\sigma} \rangle|^2 |\det(U_{K \setminus \{a_c\}}^{(-k)})|^2 &= \sum_{K \in \mathcal{K}'} \left(\sum_{j \geq \ell, j \notin K} \langle \Delta f, \psi_{j,c_1,\sigma} \rangle^2 \right) |\det(U_K^{(-k)})|^2 \\ &\leq \|\Delta f\|_2^2 \sum_{K \in \mathcal{K}'} |\det(U_K^{(-k)})|^2, \end{aligned}$$

where $\mathcal{K}'$ is the set of subsets of $\mathbb{N}_0$ of size $\ell - 1$ whose elements are larger than or equal to ℓ , i.e., $\mathcal{K}' = \{K \mid K = \{\ell \leq a_0 < \dots < a_{\ell-2}\}\}$. Since each $K \in \mathcal{K}'$ has $\ell - 1$ elements, then $\{\ell, \dots, 2\ell - 1\} \not\subseteq K$. We have

$$\sum_{K \in \mathcal{K}'} |\det(U_K^{(-k)})|^2 \leq \sum_{j=\ell}^{2\ell-1} \sum_{K \in \mathcal{K}_j} |\det(U_K^{(-k)})|^2.$$

Recall that $\mathcal{K}_j$ is the set of subsets of $\mathbb{N}_0$ of size $\ell - 1$ whose elements are larger than or equal to ℓ and not equal to j , i.e., $\mathcal{K}_j = \{K \mid K = \{\ell \leq a_0 < \dots < a_{\ell-2} \text{ and } a_i \neq j\}\}$. By Lemma A.6.9, we have

$$\begin{aligned} \sum_{K \in \mathcal{K}'} |\det(U_K^{(-k)})|^2 &\leq \sum_{j=\ell}^{2\ell-1} \left(e^{\frac{5}{4\sigma^2} r^2} 4^\ell \cdot \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \right)^2 \cdot \det(A) \\ &\leq c_1(r, \sigma) \exp(2\ell \log \ell + c_2(r, \sigma)\ell + c_3 \log \ell) \cdot \det(A), \end{aligned}$$

for some constants $c_1(r, \sigma) > 0$ and $c_3 > 0$, which means

$$\sum_{K \in \mathcal{K}} \det(U_K^{k \rightarrow \Delta})^2 \leq \|\Delta f\|^2 \cdot \det(A) \cdot c_1(r, \sigma) \exp(2\ell \log \ell + c_2(r, \sigma)\ell + c_3 \log \ell)$$

By plugging it into (A.44),

$$|\det(A^{(2,k) \rightarrow \Delta})| \leq \|\Delta f\|_2 \cdot \det(A) \cdot c_1(r, \sigma) \exp(\ell \log \ell + c_2(r, \sigma)\ell + c_3 \log \ell). \quad (\text{A.45})$$

By combining (A.43) and (A.45), we obtain that for $i = 1, 2$, and $\forall k \in [\ell]$,

$$|(\mathcal{E}_a)_{i,k}| = \frac{|\det(A^{(i,k) \rightarrow \Delta})|}{\det(A)} \leq \|\Delta f\|_2 \cdot c_1(r, \sigma) \exp(\ell \log \ell + c_2(r, \sigma)\ell + c_3 \log \ell),$$

for some constants $c_1(r, \sigma) > 0$ and $c_3 > 0$. □

A.6.2 Completing the proof of Lemma A.6.1

We are ready to complete the proof of this key lemma.

Proof of Lemma A.6.1. By Lemma A.6.3 we know if $|\lambda_{i,k} - \widehat{\lambda}_{i,k}| < \Delta$, $\forall i = 1, 2$, $k \in [\ell]$, and $\ell \geq \lfloor 2er_i^2/\sigma^2 \rfloor + 2$, then

$$\|w_i f_i - \widetilde{f}_i\|_1 \leq \frac{2^{17/4} c_\psi \sqrt{\sigma}}{\Gamma\left(\frac{1}{4}\right)^2} \cdot \Delta \ell^{5/4} + \frac{2^{11/2} c_\psi r_i w_i}{\Gamma\left(\frac{1}{4}\right)^2 \pi^{1/4} \sigma} \cdot \left(\frac{1}{2}\right)^\ell. \quad (\text{A.46})$$

By Lemma A.6.6, we know when $8 \max(r_1, r_2) < r$, for $i = 1, 2$, and $\forall k \in [\ell]$, as $\ell \rightarrow \infty$,

$$|(\mathcal{E}_t)_{i,k}| \leq |(\mathcal{E}_{t,1})_{i,k}| + |(\mathcal{E}_{t,2})_{i,k}| \leq \frac{r e^{\frac{5}{4\sigma^2} r^2}}{\sqrt{2\sigma\sqrt{\pi}}(r - 2 \max(r_1, r_2))} \left(\frac{8 \max(r_1, r_2)}{r} \right)^\ell \rightarrow 0, \quad (\text{A.47})$$

Furthermore, by Lemma A.6.7, we know for any $\ell \in \mathbb{N}_0$, $i = 1, 2$ and $k \in [\ell]$, we have

$$|(\mathcal{E}_a)_{i,k}| = \frac{\left| \det \left(A^{(i,k) \rightarrow \Delta} \right) \right|}{\det(A)} \leq \|\Delta f\|_2 \cdot c_1(r, \sigma) \exp(\ell \log \ell + c_2(r, \sigma)\ell + c_3 \log \ell). \quad (\text{A.48})$$

From (A.48) and (A.47) we know $\forall i = 1, 2, k \in [\ell]$,

$$\begin{aligned} |\lambda_{i,k} - \widehat{\lambda}_{i,k}| &\leq |(\mathcal{E}_t)_{i,k}| + |(\mathcal{E}_a)_{i,k}| \\ &\leq c_1(r, \sigma, r_1, r_2) \cdot \left(\frac{8 \max(r_1, r_2)}{r} \right)^\ell + c_2(r, \sigma) \|\Delta f\|_2 \exp(\ell \log \ell + c_3(r, \sigma)\ell + c_4 \log \ell). \end{aligned}$$

We plug this into Δ in (A.46), then we obtain

$$\begin{aligned} \|w_i f_i - \widetilde{f}_i\|_1 &\leq c_1(r, \sigma, r_1, r_2) \cdot \left(\frac{8 \max(r_1, r_2)}{r} \right)^\ell \cdot \ell^{5/4} + c_2(r_i, \sigma) w_i \left(\frac{1}{2} \right)^\ell \\ &\quad + c_3(r, \sigma) \cdot \|\Delta f\|_2 \cdot \exp(\ell \log \ell + c_4(r, \sigma)\ell + c_5 \log \ell), \end{aligned}$$

Finally, by triangle inequality, we get

$$\begin{aligned} \|\widehat{f}_i - f_i\|_1 &= \left\| \frac{(\widetilde{f}_i)_+}{\|(\widetilde{f}_i)_+\|_1} - \frac{(\widetilde{f}_i)_+}{w_i} + \frac{(\widetilde{f}_i)_+}{w_i} - f_i \right\|_1 \\ &\leq \left\| \frac{(\widetilde{f}_i)_+}{\|(\widetilde{f}_i)_+\|_1} - \frac{(\widetilde{f}_i)_+}{w_i} \right\|_1 + \left\| \frac{(\widetilde{f}_i)_+}{w_i} - f_i \right\|_1 \\ &= \left| 1 - \frac{\|(\widetilde{f}_i)_+\|_1}{w_i} \right| \cdot \left\| \frac{(\widetilde{f}_i)_+}{\|(\widetilde{f}_i)_+\|_1} \right\|_1 + \frac{1}{w_i} \|(\widetilde{f}_i)_+ - w_i f_i\|_1 \\ &\leq \frac{2}{w_i} \|w_i f_i - (\widetilde{f}_i)_+\|_1 \end{aligned}$$

$$\begin{aligned}
&\leq \frac{2}{w_i} \|w_i f_i - \tilde{f}_i\|_1 \\
&\leq \frac{1}{w_i} c_1(r, \sigma, r_1, r_2) \cdot \left(\frac{8 \max(r_1, r_2)}{r} \right)^\ell \cdot \ell^{5/4} + c_2(r_i, \sigma) \left(\frac{1}{2} \right)^\ell \\
&\quad + \frac{1}{w_i} c_3(r, \sigma) \cdot \|\Delta f\|_2 \cdot \exp(\ell \log \ell + c_4(r, \sigma) \ell + c_5 \log \ell)
\end{aligned}$$

for some constants $c_1(r, \sigma, r_1, r_2) > 0$, $c_2(r_i, \sigma) > 0$, $c_3(r, \sigma) > 0$, $c_4(r, \sigma) > 0$, and $c_5 > 0$. $\square$

A.6.3 Auxiliary lemmas

For completeness, in this subsection we collect technical lemmas required in the proof from the previous subsection.

Lemma A.6.8 (Lemma 32, [Tai and Aragam \(2023\)](#)). *For any $\ell \geq 1$, $k \in [\ell]$ and $j \geq \ell$, we have*

$$\sum_{K \in \mathcal{K}_j} \det(U_{K \cup \{k\}})^2 \leq \left(e^{\frac{r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \right)^2 \cdot \det(A).$$

Recall that, for any $K = \{\ell \leq a_1 < \dots < a_{\ell-1} \text{ and } a_i \neq j\} \in K_j$ and $k \in [\ell]$, $U_{K \cup \{k\}}$ is the ℓ -by- ℓ matrix whose (i, j) -entry is $\langle \psi_{c_j, r, \sigma}, \psi_{b_i, 0, \sigma} \rangle$ for $b_i \in \{k\} \cup K$ and $c_j \in [\ell]$.

Proof. For each $K = \{\ell \leq a_0 < \dots < a_{\ell-2} \text{ and } a_i \neq j\} \in \mathcal{K}_j$, we have the following. By Lemma A.6.10, we have

$$\begin{aligned}
\frac{|\det(U_{K \cup \{k\}})|}{|\det(U_{K \cup \{j\}})|} &= \sqrt{\frac{j!}{k!}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{k-j} \left(\prod_{c=0}^{\ell-2} \frac{|a_c - k|}{|a_c - j|} \right) \cdot \frac{\Gamma_{K \cup \{k\}, \ell}}{\Gamma_{K \cup \{j\}, \ell}} \\
&\leq \sqrt{\frac{j!}{k!}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{k-j} \left(\prod_{c=0}^{\ell-2} \frac{|a_c - k|}{|a_c - j|} \right)
\end{aligned}$$

The second inequality follows from Lemma A.6.14, which states that $\Gamma_{K, b}$ is an increasing function w.r.t any $a_i \in K$.

By Lemma A.6.13, we have $\prod_{c=0}^{\ell-2} \frac{|a_c - k|}{|a_c - j|} \leq \prod_{c=0}^{\ell-2} \frac{|a_c|}{|a_c - j|} \leq 2^j \cdot 4^\ell$ and hence

$$\begin{aligned}
|\det(U_{K \cup \{k\}})| &\leq 4^\ell \cdot 2^j \sqrt{\frac{j!}{k!}} \left(\frac{r}{\sqrt{2}\sigma}\right)^{k-j} |\det(U_{K \cup \{j\}})| \\
&\leq e^{\frac{1}{4\sigma^2}r^2} \cdot 4^\ell \cdot \sqrt{j!} \cdot \left(\frac{2\sqrt{2}\sigma}{r}\right)^j |\det(U_{K \cup \{j\}})|
\end{aligned}$$

since $\frac{1}{\sqrt{k!}} \left(\frac{r}{\sqrt{2}\sigma}\right)^k \leq \sqrt{\sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{r^2}{2\sigma^2}\right)^i} = e^{\frac{1}{4\sigma^2}r^2}$. For each $K \in \mathcal{K}_j$, the set $K \cup \{j\}$ is in $\mathcal{K}$ since $j \notin K$. Hence, we conclude that

$$\begin{aligned}
\sum_{K \in \mathcal{K}_j} \det(U_{K \cup \{k\}})^2 &\leq \left(e^{\frac{1}{4\sigma^2}r^2} \cdot 4^\ell \cdot \sqrt{j!} \cdot \left(\frac{2\sqrt{2}\sigma}{r}\right)^j \right)^2 \sum_{K \in \mathcal{K}_j} \det(U_{K \cup \{j\}})^2 \\
&\leq \left(e^{\frac{r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r}\right)^j \sqrt{j!} \right)^2 \cdot \det(A).
\end{aligned}$$

□

Before we show the lemmas below, we first define the following notations to simplify the expressions in our proof. For any $r \in \mathbb{R}$, and $\sigma > 0$, let $W_{r,\sigma}$ be the double-indexed sequence such that

$$(W_{r,\sigma})_{s,t} = e^{-\frac{1}{8\sigma^2}r^2} (-1)^s \sqrt{\frac{t!}{s!}} \binom{s}{t} \left(\frac{r}{\sqrt{2}\sigma}\right)^{s-t}$$

for $s, t \in \mathbb{N}_0$. Here, $\binom{s}{t}$ is the binomial coefficient which is equal to

$$\binom{s}{t} = \begin{cases} 0 & \text{if } s < t \\ \frac{s!}{t!(s-t)!} & \text{if } s \geq t \end{cases}$$

For any subsets $S, T \subset \mathbb{N}_0$ of same sizes, let $(W_r)_{S,T}$ be the matrix whose (i, j) -entry is $(W_r)_{s_i, t_j}$ for $s_i \in S$ and $t_j \in T$. For any (ordered) set K of m nonnegative integers $a_0, \dots, a_{m-1}$ such that

$0 \leq a_0 < \cdots < a_{m-1}$, we define

$$\Sigma_K = \sum_{c=0}^{m-1} a_c, \quad F_K = \prod_{c=0}^{m-1} a_c!, \quad C_K = \prod_{0 \leq c_2 < c_1 \leq m-1} (a_{c_1} - a_{c_2}).$$

Finally, for any $b \in \mathbb{N}_0$, we define

$$\Gamma_{K,b} = \frac{1}{b!} \sum_{d=0}^b (-1)^d \binom{b}{d} \prod_{c=0}^{m-1} (a_c - d).$$

Lemma A.6.9 (Lemma 33, [Tai and Aragam \(2023\)](#)). *For any $\ell \geq 1$, $k \in [\ell]$ and $j \geq \ell$, we have*

$$\sum_{K \in \mathcal{K}_j} \det \left(U_K^{(-k)} \right)^2 \leq \left(e^{\frac{5}{4\sigma^2} r^2} 4^\ell \cdot \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \right)^2 \cdot \det(A).$$

Recall that, for any $K = \{\ell \leq a_0 < \cdots < a_{\ell-2} \text{ and } a_i \neq j\} \in \mathcal{K}_j$ and $k \in [\ell]$, $U_K^{(-k)}$ is the $(\ell-1)$ -by- $(\ell-1)$ matrix whose (i, j) -entry is $\langle \psi_{c_j, r, \sigma}, \psi_{b_i, 0, \sigma} \rangle$ for $b_i \in K$ and $c_j \in [\ell] \setminus \{k\}$.

Proof. For each $K = \{\ell \leq a_0 < \cdots < a_{\ell-2} \text{ and } a_i \neq j\} \in \mathcal{K}_j$, we have the following. By Lemma A.6.10 and A.6.11, we have

$$\begin{aligned} \frac{|\det(U_K^{(-k)})|}{|\det(U_{K \cup \{j\}})|} &\leq \frac{e^{-\frac{\ell-1}{4\sigma^2} r^2} 2^k \sqrt{\frac{1}{k! F_{[\ell]} F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell-1)/2 + k} C_K \prod_{c=0}^{\ell-2} a_c}{e^{-\frac{\ell}{4\sigma^2} r^2} \sqrt{\frac{1}{F_{[\ell]} F_{K \cup \{j\}}}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_{K \cup \{j\}} - \ell(\ell-1)/2} C_{K \cup \{j\}} \Gamma_{K \cup \{j\}, \ell}} \\ &= e^{\frac{r^2}{4\sigma^2}} 2^k \sqrt{\frac{j!}{k!}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{k-j} \prod_{c=0}^{\ell-2} \frac{a_c}{|a_c - j|} \cdot \frac{1}{\Gamma_{K \cup \{j\}, \ell}} \\ &\leq e^{\frac{r^2}{4\sigma^2}} 2^k \sqrt{\frac{j!}{k!}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{k-j} \prod_{c=0}^{\ell-2} \frac{a_c}{|a_c - j|} \end{aligned}$$

The last inequality is because $\Gamma_{K \cup \{j\}, \ell} \geq \Gamma_{[\ell+1] \setminus \{0\}, \ell} = 1$ by Lemma A.6.14. In addition, by Lemma A.6.13, we have

$$\frac{|\det(U_K^{(-k)})|}{|\det(U_{K \cup \{j\}})|} \leq e^{\frac{r^2}{4\sigma^2}} \sqrt{j!} \left(\frac{r}{\sqrt{2}\sigma} \right)^{-j} \sqrt{\frac{\left(\frac{2r^2}{\sigma^2} \right)^k}{k!}} \cdot 2^j \cdot 4^{\ell-1}$$

$$= e^{\frac{5r^2}{4\sigma^2}} 4^\ell \sqrt{j!} \left(\frac{2\sqrt{2}\sigma}{r} \right)^j$$

since $\sqrt{(2r^2/\sigma^2)^k / k!} \leq \sqrt{e^{2r^2/\sigma^2}} = e^{r^2/\sigma^2}$. For each $K \in \mathcal{K}_j$, the set $K \cup \{j\}$ is in $\mathcal{K}$ since $j \notin K$.

Hence,

$$\begin{aligned} \sum_{K \in \mathcal{K}_j} \det(U_K^{(-k)})^2 &\leq \left(e^{\frac{5r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \right)^2 \sum_{K \in \mathcal{K}_j} \det(U_{K \cup \{j\}})^2 \\ &\leq \left(e^{\frac{5r^2}{4\sigma^2}} 4^\ell \left(\frac{2\sqrt{2}\sigma}{r} \right)^j \sqrt{j!} \right)^2 \det(A) \end{aligned}$$

□

Lemma A.6.10 (Lemma 34, [Tai and Aragam \(2023\)](#)). *Let $a_0, \dots, a_{\ell-1}$ be ℓ nonnegative integers and K be the set $\{0 \leq a_0 < \dots < a_{\ell-1}\}$. Then, we have*

$$|\det(U_K)| = e^{-\frac{\ell}{4\sigma^2} r^2} \sqrt{\frac{1}{F_{[\ell]} F_K} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell-1)/2}} C_K \cdot \Gamma_{K, \ell}.$$

Recall that, for any $K = \{0 \leq a_0 < \dots < a_{\ell-1}\}$, U_K is the ℓ -by- ℓ matrix whose (i, j) -entry is $\langle \psi_{c_j, r, \sigma}, \psi_{b_i, 0, \sigma} \rangle$ for $b_i \in [K]$ and $c_j \in [\ell]$.

Proof. Recall that, from (A.23),

$$\begin{aligned} &\langle \psi_{i, 0, \sigma}, \psi_{j, r, \sigma} \rangle \\ &= \sum_{k=0}^{i \wedge j} \left(e^{-\frac{r^2}{8\sigma^2}} (-1)^i \sqrt{\frac{k!}{i!}} \binom{i}{k} \left(\frac{r}{\sqrt{2}\sigma} \right)^{i-k} \right) \left(e^{-\frac{r^2}{8\sigma^2}} (-1)^j \sqrt{\frac{k!}{j!}} \binom{j}{k} \left(\frac{-r}{\sqrt{2}\sigma} \right)^{j-k} \right) \end{aligned}$$

It means that U_K can be decomposed as

$$U_K = (W_{r, \sigma})_{K, [\ell]} (W_{-r, \sigma})_{[\ell], [\ell]}^\top.$$

By Lemma A.6.12, the determinant of $(W_{-r,\sigma})_{[\ell],[\ell]}^\top$ is

$$e^{-\frac{\ell}{8\sigma^2}r^2}(-1)^{\Sigma_{[\ell]}}\sqrt{\frac{1}{F_{[\ell]}F_{[\ell]}}}\left(-\frac{r}{\sqrt{2}\sigma}\right)^{\Sigma_{[\ell]}-\ell(\ell-1)/2}C_{[\ell]}\cdot\Gamma_{[\ell],\ell}=e^{-\frac{\ell}{8\sigma^2}r^2}(-1)^{\ell(\ell-1)/2},$$

since $\Sigma_{[\ell]} = \ell(\ell-1)/2$, $C_{[\ell]} = F_{[\ell]}$ and $\Gamma_{[\ell],\ell} = 1$. Also, the determinant of $(W_{r,\sigma})_{K,[\ell]}$ is

$$\det((W_{r,\sigma})_{K,[\ell]}) = e^{-\frac{\ell}{8\sigma^2}r^2}(-1)^{\Sigma_K}\sqrt{\frac{1}{F_{[\ell]}F_K}}\left(\frac{r}{\sqrt{2}\sigma}\right)^{\Sigma_K-\ell(\ell-1)/2}C_K\cdot\Gamma_{K,\ell}.$$

Hence, we have

$$\begin{aligned}\det(U_K) &= \det((W_{r,\sigma})_{K,[\ell]}) \cdot \det((W_{-r,\sigma})_{[\ell],[\ell]}) \\ &= e^{-\frac{\ell}{8\sigma^2}r^2}(-1)^{\ell(\ell-1)/2} \cdot e^{-\frac{\ell}{8\sigma^2}r^2}(-1)^{\Sigma_K}\sqrt{\frac{1}{F_{[\ell]}F_K}}\left(\frac{r}{\sqrt{2}\sigma}\right)^{\Sigma_K-\ell(\ell-1)/2}C_K\cdot\Gamma_{K,\ell} \\ &= e^{-\frac{\ell}{4\sigma^2}r^2}(-1)^{\Sigma_K+\ell(\ell-1)/2}\sqrt{\frac{1}{F_{[\ell]}F_K}}\left(\frac{r}{\sqrt{2}\sigma}\right)^{\Sigma_K-\ell(\ell-1)/2}C_K\cdot\Gamma_{K,\ell}.\end{aligned}$$

□

Lemma A.6.11 (Lemma 35, [Tai and Aragam \(2023\)](#)). *Let $a_0, \dots, a_{\ell-2}$ be $\ell-1$ nonnegative integers and K be the set $\{0 \leq a_0 < \dots < a_{\ell-2}\}$. Then, for any $k \in [\ell]$, we have*

$$\left|\det\left(U_K^{(-k)}\right)\right| \leq e^{-\frac{\ell-1}{4\sigma^2}r^2}2^k\sqrt{\frac{1}{k!F_{[\ell]}F_K}}\left(\frac{r}{\sqrt{2}\sigma}\right)^{\Sigma_K-\ell(\ell-1)/2+k}C_K\prod_{c=0}^{\ell-2}a_c$$

Recall that, for any $K = \{0 \leq a_0 < \dots < a_{\ell-2}\} \in \mathcal{K}$ and $k \in [\ell]$, $U_K^{(-k)}$ is the $(\ell-1)$ -by- $(\ell-1)$ matrix whose (i, j) -entry is $\langle \psi_{c_j, r, \sigma}, \psi_{b_i, 0, \sigma} \rangle$ for $b_i \in K$ and $c_j \in [\ell] \setminus \{k\}$.

Proof. Recall that, from (A.23),

$$\begin{aligned}\langle \psi_{i, 0, \sigma}, \psi_{j, r, \sigma} \rangle &= \sum_{k=0}^{i \wedge j} \left(e^{-\frac{r^2}{8\sigma^2}}(-1)^i \sqrt{\frac{k!}{i!}} \binom{i}{k} \left(\frac{r}{\sqrt{2}\sigma} \right)^{i-k} \right) \left(e^{-\frac{r^2}{8\sigma^2}}(-1)^j \sqrt{\frac{k!}{j!}} \binom{j}{k} \left(\frac{-r}{\sqrt{2}\sigma} \right)^{j-k} \right)\end{aligned}$$

It means that $U_K^{(-k)}$ can be decomposed as

$$U_K^{(-k)} = (W_{r,\sigma})_{K,[\ell]} (W_{-r,\sigma})_{[\ell],[\ell]\setminus\{k\}}.$$

By Cauchy-Binet formula (Lemma A.6.4), we have

$$\det \left(U_K^{(-k)} \right) = \sum_{b=0}^{\ell-1} \det((W_{r,\sigma})_{K,[\ell]\setminus\{b\}}) \cdot \det((W_{-r,\sigma})_{[\ell]\setminus\{b\}, [\ell]\setminus\{k\}}). \quad (\text{A.49})$$

By Lemma A.6.12, the determinant of $(W_{-r,\sigma})_{[\ell]\setminus\{b\}, [\ell]\setminus\{k\}}$ is

$$\begin{aligned} & e^{-\frac{\ell-1}{8\sigma^2}r^2} (-1)^{\frac{(\ell-1)\ell}{2}-b} \cdot \sqrt{\frac{k!}{F_{[\ell]\setminus\{b\}} \cdot F_{[\ell]}}} \cdot \left(-\frac{r}{\sqrt{2}\sigma} \right)^{-b+k} \cdot C_{[\ell]\setminus\{b\}} \cdot \Gamma_{[\ell]\setminus\{b\},k} \\ &= \begin{cases} 0, & \text{if } k < b \\ e^{-\frac{\ell-1}{8\sigma^2}r^2} (-1)^{\ell(\ell-1)/2+k} \frac{\sqrt{k!b!}}{b!(k-b)!} \left(\frac{r}{\sqrt{2}\sigma} \right)^{k-b}, & \text{if } k \geq b \end{cases} \end{aligned}$$

since $C_{[\ell]\setminus\{b\}} = \frac{F_{[\ell]}}{(\ell-1-b)!b!}$ and $\Gamma_{[\ell]\setminus\{b\},k} = \begin{cases} 0 & \text{if } k < b \\ \frac{(\ell-1-b)!}{(k-b)!} & \text{if } k \geq b \end{cases}$. Also, the determinant of

$(W_{r,\sigma})_{K,[\ell]\setminus\{b\}}$ is

$$e^{-\frac{\ell-1}{8\sigma^2}r^2} (-1)^{\Sigma_K} \sqrt{\frac{b!}{F_{[\ell]} F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell-1)/2+b} \cdot C_K \cdot \Gamma_{K,b}.$$

Hence, when $b \leq k$, we have

$$\begin{aligned} & \left| \det(W_{r,\sigma})_{K,[\ell]\setminus\{b\}} \det((W_{-r,\sigma})_{[\ell]\setminus\{b\}, [\ell]\setminus\{k\}}) \right| \\ &= e^{-\frac{\ell-1}{4\sigma^2}r^2} \frac{1}{(k-b)!} \sqrt{\frac{k!}{F_{[\ell]} F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell-1)/2+k} C_K \cdot \Gamma_{K,b} \\ &\leq e^{-\frac{\ell-1}{4\sigma^2}r^2} \frac{1}{(k-b)!b!} \sqrt{\frac{k!}{F_{[\ell]} F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell-1)/2+k} C_K \prod_{c=0}^{\ell-2} a_c \end{aligned}$$

The last inequality is due to Lemma A.6.14. By plugging it into (A.49), we have

$$\begin{aligned} \left| \det(U_K^{(-k)}) \right| &= \sum_{b=0}^k \left| \det((W_{r,\sigma})_{K,[\ell] \setminus \{b\}}) \det((W_{-r,\sigma}^\top)_{[\ell] \setminus \{b\}, [\ell] \setminus \{k\}}) \right| \\ &\leq \sum_{b=0}^k e^{-\frac{\ell-1}{4\sigma^2}r^2} \frac{1}{(k-b)!b!} \sqrt{\frac{k!}{F_{[\ell]}F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell-1)/2+k} C_K \prod_{c=0}^{\ell-2} a_c. \end{aligned}$$

Since $\sum_{b=0}^k \frac{1}{(k-b)!b!} = \frac{2^k}{k!}$, we obtain

$$\left| \det(U_K^{(-k)}) \right| \leq e^{-\frac{\ell-1}{4\sigma^2}r^2} 2^k \sqrt{\frac{1}{k!F_{[\ell]}F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell-1)/2+k} C_K \prod_{c=0}^{\ell-2} a_c$$

□

Lemma A.6.12 (Lemma 37, [Tai and Aragam \(2023\)](#)). *Let $a_0, \dots, a_{\ell-1}$ be ℓ nonnegative integers such that $0 \leq a_0 < \dots < a_{\ell-1}$ and K be the ordered set $\{a_0, \dots, a_{\ell-1}\}$. Then, for any integer $b \in [\ell + 1]$, we have*

$$\det((W_{r,\sigma})_{K,[\ell+1] \setminus \{b\}}) = e^{-\frac{\ell}{8\sigma^2}r^2} (-1)^{\Sigma_K} \sqrt{\frac{b!}{F_{[\ell+1]}F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell+1)/2+b} C_K \cdot \Gamma_{K,b},$$

and in particular,

$$\det((W_{r,\sigma})_{K,[\ell]}) = e^{-\frac{\ell}{8\sigma^2}r^2} (-1)^{\Sigma_K} \sqrt{\frac{1}{F_{[\ell]}F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell-1)/2} C_K \cdot \Gamma_{K,\ell}.$$

Proof. By factoring the common terms in each row and column, we have

$$\begin{aligned} &\det((W_{r,\sigma})_{K,[\ell+1] \setminus \{b\}}) \\ &= e^{-\frac{\ell}{8\sigma^2}r^2} (-1)^{\Sigma_K} \sqrt{\frac{b!}{\prod_{c=0}^{\ell} c! \prod_{c=0}^{\ell-1} a_c!}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell+1)/2+b} \det((W')_{K,[\ell+1] \setminus \{b\}}) \\ &= e^{-\frac{\ell}{8\sigma^2}r^2} (-1)^{\Sigma_K} \sqrt{\frac{b!}{F_{[\ell+1]}F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell+1)/2+b} \det((W')_{K,[\ell+1] \setminus \{b\}}) \end{aligned}$$

where $W'_{K, [\ell+1] \setminus \{b\}}$ is a ℓ -by- ℓ matrix that

$$W'_{K, [\ell+1] \setminus \{b\}} = \begin{bmatrix} 1 & \cdots & \prod_{d=0}^{b-2} (a_0 - d) & \prod_{d=0}^b (a_0 - d) & \cdots & \prod_{d=0}^{\ell-1} (a_0 - d) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & \cdots & \prod_{d=0}^{b-2} (a_{\ell-1} - d) & \prod_{d=0}^b (a_{\ell-1} - d) & \cdots & \prod_{d=0}^{\ell-1} (a_{\ell-1} - d) \end{bmatrix}$$

We will apply column operations on $\det(W'_{K, [\ell+1] \setminus \{b\}})$. For the column indexed at $c \neq 0, b+1$, subtract the column indexed at $c-1$ multiplied by $(a_0 - c)$ to it and, for the column indexed at $b+1$, subtract the column indexed at $b-1$ multiplied by $(a_0 - b)(a_0 - (b-1))$ to it. We have the first row to be a zero row except the first entry is 1 and expand the determinant along the first row. Then, the row indexed at a_c has a factor $a_c - a_0$ and in particular the entry indexed at $b+1$ has an extra factor $a_c + a_0 - (2b-1) = (a_c - (b-1)) + (a_0 - b)$. By factoring out $a_c - a_0$, we have

$$\begin{aligned} \det(W'_{K, [\ell+1] \setminus \{b\}}) &= \left(\prod_{c=1}^{\ell} (a_c - a_0) \right) \left(\det((W'(r))_{K \setminus \{a_0\}, [\ell] \setminus \{b-1\}}) + (a_0 - b) \det((W'(r))_{K \setminus \{a_0\}, [\ell] \setminus \{b\}}) \right). \end{aligned}$$

By induction, we conclude

$$\det(W'_{K, [\ell+1] \setminus \{b\}}) = C_K \cdot \Gamma_{K,b}$$

Therefore,

$$\det((W_{r,\sigma})_{K, [\ell+1] \setminus \{b\}}) = e^{-\frac{\ell}{8\sigma^2} r^2} (-1)^{\Sigma_K} \sqrt{\frac{b!}{F_{[\ell+1]} F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell+1)/2 + b} C_K \cdot \Gamma_{K,b},$$

and in particular,

$$\det((W_{r,\sigma})_{K, [\ell]}) = e^{-\frac{\ell}{8\sigma^2} r^2} (-1)^{\Sigma_K} \sqrt{\frac{1}{F_{[\ell]} F_K}} \left(\frac{r}{\sqrt{2}\sigma} \right)^{\Sigma_K - \ell(\ell-1)/2} C_K \cdot \Gamma_{K,\ell}.$$

□

Lemma A.6.13 (Lemma 36, [Tai and Aragam \(2023\)](#)). *Let $a_0, \dots, a_{\ell-1}$ be ℓ nonnegative integers*

such that $0 \leq a_0 < \cdots < a_{\ell-1}$ and $a_c \neq j$ for some $j \in \mathbb{N}_0$. Then, we have

$$\prod_{c=0}^{\ell-1} \left| \frac{a_c}{a_c - j} \right| \leq 2^j \cdot 4^\ell.$$

Proof. Suppose $a_0, \dots, a_{b-1}$ are the integers less than j and the rest of them are larger than j . For $c \leq b-1$, we have

$$\left| \frac{a_c}{a_c - j} \right| = \frac{a_c}{j - a_c} = \frac{j}{j - a_c} - 1 \leq \frac{j}{b - c} - 1 = \frac{j - (b - c)}{b - c}$$

since $a_c \leq j - (b - c)$. For $c \geq b$, we have

$$\left| \frac{a_c}{a_c - j} \right| = \frac{a_c}{a_c - j} = 1 + \frac{j}{a_c - j} \leq 1 + \frac{j}{c - b + 1} = \frac{j + 1 + (c - b)}{c - b + 1}$$

since $a_c \geq j + 1 + (c - b)$. Hence, we have

$$\begin{aligned} \prod_{c=0}^{\ell-1} \left| \frac{a_c}{a_c - j} \right| &\leq \prod_{c=0}^{b-1} \frac{j - (b - c)}{b - c} \cdot \prod_{c=b}^{\ell-1} \frac{j + 1 + (c - b)}{c - b + 1} \\ &= \frac{(j - 1)!}{(j - b - 1)!b!} \cdot \frac{(j + \ell - b)!}{(\ell - b)!j!} \\ &\leq \frac{(j + \ell - b)!}{(\ell - b)!b!(j - b)!} = \frac{(j + \ell - b)!}{\ell!(j - b)!} \cdot \frac{\ell!}{(\ell - b)!b!} \\ &\leq 2^{j+\ell-b} \cdot 2^\ell \leq 2^j \cdot 4^\ell. \end{aligned}$$

□

Lemma A.6.14 (Lemma 38, [Tai and Aragam \(2023\)](#)). Let $a_0, \dots, a_{\ell-1}$ be ℓ nonnegative integers such that $0 \leq a_0 < \cdots < a_{\ell-1}$ and K be the set $\{a_0, \dots, a_{\ell-1}\}$. Then, for any $b \in \mathbb{N}_0$, we have

$$0 \leq \Gamma_{K,b} \leq \frac{1}{b!} \prod_{\alpha=0}^{\ell-1} a_\alpha,$$

and

$$\frac{\partial}{\partial a_i} \Gamma_{K,b} \geq 0, \forall a_i \in K$$

which means that $\Gamma_{K,b}$ is an increasing function with regard to any $a_i \in K$.

Proof. We first prove $\Gamma_{K,b} \geq 0$ and will use induction on ℓ to prove the statement. Suppose $\ell = 1$. When $b = 0$, we have $\Gamma_{K,0} = a_0 \geq 0$. When $b = 1$, we have $\Gamma_{K,1} = a_0 - (a_0 - 1) = 1 > 0$. When $b \geq 2$, we have $\Gamma_{K,b} = \frac{1}{b!} \sum_{d=0}^b (-1)^d \binom{b}{d} (a_0 - d) = 0$ since both $\sum_{d=0}^b (-1)^d \binom{b}{d}$ and $\sum_{d=0}^b (-1)^d \binom{b}{d} d = b \cdot \sum_{d=1}^b (-1)^d \binom{b-1}{d-1}$ are zero. Suppose $\ell \geq 2$. We view $\Gamma_{K,b}$ as a function of $a_0, a_1, \dots, a_{\ell-1}$. The partial derivative of $\Gamma_{K,b}$ is

$$\frac{\partial}{\partial a_i} \Gamma_{K,b} = \frac{1}{b!} \sum_{d=0}^b (-1)^d \binom{b}{d} \prod_{c=0, c \neq i}^{\ell-1} (a_c - d) = \Gamma_{K \setminus \{a_i\}, b} \geq 0$$

by the induction assumption. It means that $\Gamma_{K,b}$ is an increasing function and it implies $\Gamma_{K,b} \geq \Gamma_{[\ell],b} = 0$ by direct calculation.

Now, we will prove $\Gamma_{K,b} \leq \frac{1}{b!} \prod_{c=0}^{\ell-1} a_c$ and will use induction on b to prove the statement. Suppose $b = 1$. We have $\Gamma_{K,0} = \prod_{c=0}^{\ell-1} a_c$. Suppose $b \geq 2$. Note that $\binom{b}{d} = \binom{b-1}{d} + \binom{b-1}{d-1}$. We have

$$\begin{aligned} \Gamma_{K,b} &= \frac{1}{b!} \left(\sum_{d=0}^{b-1} (-1)^d \binom{b-1}{d} \prod_{c=0}^{\ell-1} (a_c - d) - \sum_{d=1}^b (-1)^{d-1} \binom{b-1}{d-1} \prod_{c=0}^{\ell-1} (a_c - d) \right) \\ &\leq \frac{1}{b!} \sum_{d=0}^{b-1} (-1)^d \binom{b-1}{d} \prod_{c=0}^{\ell-1} (a_c - d) \leq \frac{1}{b} \frac{1}{(b-1)!} \prod_{c=0}^{\ell-1} a_c = \frac{1}{b!} \prod_{c=0}^{\ell-1} a_c. \end{aligned}$$

The inequalities are due to the induction assumption. □

A.7 Proofs for Section 2.7.2: Concentration inequality for $\|\widehat{f}_i - f_i\|_1$

Lemma A.6.1 gives an upper bound of $\|\widehat{f}_i - f_i\|_1$ for any density estimator $\widehat{f}$. In this section, let us consider a specific form of the estimator $\widehat{f}$, namely the standard kernel density estimator. Then we can further derive the tail behavior of $\widehat{f}_i$, as given in the following lemma:

Lemma A.7.1. *Let $f \in \mathcal{F}$ be as defined in (C2.3). We can verify that all such densities f satisfy the regularity conditions in Lemma A.7.3: there exist $C_1 > 0$, $C_2 > 0$, $C_3 > 0$, $C_4 > 0$, such that $\forall f \in \mathcal{F}$, $\|f\|_2^2 < C_1$, $|\int f'(x)dx| < C_2$, $\|f''\|_2^2 < C_3$, $|\int f''(x)dx| < C_4$. Let kernel K be a standard Gaussian density. Suppose $X_1, \dots, X_n$ are n i.i.d. samples from P_f , and let $\hat{f}(x; h)$ be the kernel density estimator defined as*

$$\hat{f}(x; h) = n^{-1} \sum_{i=1}^n K_h(x - X_i),$$

where $h > 0$ is the bandwidth, and $K_h(x) := h^{-1}K(x/h)$. Choose $h = c_K n^{-1/5}$, where c_K is a positive constant depending only on K . Then by taking

$$\begin{cases} \alpha = \frac{2}{5} \log \frac{r}{8 \max(r_1, r_2)}, \beta = \frac{1}{2} \log \frac{r}{8 \max(r_1, r_2)}, & \text{if } 1 < \frac{r}{8 \max(r_1, r_2)} \leq 2, \\ \alpha = \frac{3}{5} \log \frac{r}{16 \max(r_1, r_2)}, \beta = \log 2, & \text{if } \frac{r}{8 \max(r_1, r_2)} > 2, \end{cases}$$

and letting ϵ_α be a constant such that $0 < \epsilon_\alpha < e^{-1}$, and $\forall 0 < \epsilon < \epsilon_\alpha$, $\log\left(\frac{1}{\epsilon}\right) < \left(\frac{1}{\epsilon}\right)^\alpha$, we have that for $i = 1, 2$, $\forall 0 < \eta < \bar{c}_{1i} \epsilon_\alpha^\beta$, and $n > (2\varphi_K)^{5/2} (\frac{\bar{c}_{1i}}{\eta})^{5/(2\beta)}$, take $\ell = \max(\lfloor \beta^{-1} \log(\bar{c}_{1i}/\eta) \rfloor, \lfloor 2er_i^2/\sigma^2 \rfloor) + 2$, then $\hat{f}_i$ defined in (A.26) satisfies

$$P_f^n \left(\|\hat{f}_i - f_i\|_1 \geq \eta \right) \leq \exp \left(-\tilde{c}_K n^{4/5} \left(\frac{\eta}{\bar{c}_{1i}} \right)^{\frac{2}{\beta} (b_i \log \log \left(\frac{\bar{c}_{1i}}{\eta} \right) - b_i \log \beta + c_{2i})} \right),$$

where φ_K and $\tilde{c}_K$ are positive constants depending only on K , and $\bar{c}_{1i} = \bar{c}_{1i}(r, \sigma, r_1, r_2, \underline{w})$ is a positive constant depending on $i, r, \sigma, r_1, r_2, \underline{w}$, $c_{2i} = c_{2i}(r, r_1, r_2, \sigma)$ is a positive constant depending on i, r, r_1, r_2, σ , and $b_i = b(r_i, \sigma) \geq 1$ is a constant depending on r_i and σ .

This section of the Appendix is devoted to the proof of Lemma A.7.1. We will need the following standard tool.

Lemma A.7.2 (Bounded difference inequality). *Let χ be some set, and let $f : \chi^n \rightarrow R$ be a*

function such that there exist $c_1, \dots, c_n > 0$, $\forall k = 1, \dots, n$, $\forall x_1, \dots, x_n \in \mathcal{X}$,

$$\sup_{x'_k \in \mathcal{X}} |f(x_1, \dots, x_k, \dots, x_n) - f(x_1, \dots, x'_k, \dots, x_n)| \leq c_k.$$

Let $X_1, \dots, X_n$ be independent random variables on $\mathcal{X}$, then $\forall t > 0$,

$$\Pr(f(X_1, \dots, X_n) - E f(X_1, \dots, X_n) > t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n c_i^2}}.$$

Lemma A.7.3. Let the kernel K be a probability density on $\mathbb{R}$ that is symmetric around the origin, bounded and has a finite fourth moment. Let $\mathcal{F}$ be a density class on $\mathbb{R}$ such that there exist $C_1 > 0$, $C_2 > 0$, $C_3 > 0$, $C_4 > 0$, $\forall f \in \mathcal{F}$, $\|f\|_2^2 < C_1$, $|\int f'(x)dx| < C_2$, $\|f''\|_2^2 < C_3$, $|\int f''(x)dx| < C_4$.

Proof.

$$E \left[\|\widehat{f} - f\|_2^2 \right] = E \int (\widehat{f}(x; h) - f(x))^2 dx = \int E(\widehat{f}(x; h) - f(x))^2 dx.$$

$E(\widehat{f}(x; h) - f(x))^2$ on the RHS is the mean square error (MSE) of $\widehat{f}$ at x , and the MSE can be decomposed into variance and squared bias

$$E(\widehat{f}(x; h) - f(x))^2 = \text{Var}(\widehat{f}(x; h)) + (E\widehat{f}(x; h) - f(x))^2.$$

Furthermore, we have (see e.g. Chapter 2 of [Wand and Jones \(1994\)](#)),

$$E\widehat{f}(x; h) = \int K(z)f(x - hz)dz,$$

and

$$\text{Var}\{\widehat{f}(x; h)\} = (nh)^{-1} \int K^2(z)f(x - hz)dz - n^{-1}(E\widehat{f}(x; h))^2.$$

By Taylor expansion of $f(x - hz)$, we obtain

$$E\widehat{f}(x; h) = \int K(z)f(x - hz)dz$$

$$= \int K(z) \left[f(x) + f'(x)(-hz) + h^2 z^2 \int_0^1 (1-t) f''(x-thz) dt \right] dz.$$

Since $\int K(z) dz = 1$, and $\int zK(z) dz = 0$,

$$E\widehat{f}(x; h) = f(x) + h^2 \int_0^1 (1-t) \left(\int f''(x-thz) z^2 K(z) dz \right) dt. \quad (\text{A.50})$$

By Cauchy-Schwarz's inequality,

$$\int f''(x-thz) z^2 K(z) dz \leq \sqrt{\left(\int f''(x-thz)^2 dz \right) \cdot \left(\int z^4 K^2(z) dz \right)} < \infty,$$

because K is bounded and has finite fourth moment, and $\int f''(x)^2 dx < C_3$. Thus, the second term on the RHS of (A.50) is finite. Let $R_1(f, K; h) := \int_0^1 (1-t) \left(\int f''(x-thz) z^2 K(z) dz \right) dt$, then

$$E\widehat{f}(x; h) = f(x) + h^2 R_1(f, K; h).$$

Similarly,

$$\begin{aligned} & \text{Var}(\widehat{f}(x; h)) \\ &= (nh)^{-1} \int K^2(z) \left[f(x) + f'(x)(-hz) + h^2 z^2 \int_0^1 (1-t) f''(x-thz) dt \right] dz \\ & \quad - n^{-1} \left(f(x) + h^2 R_1(f, K; h) \right)^2 \\ &= (nh)^{-1} \|K\|_2^2 \cdot f(x) - n^{-1} \left(f'(x) R_2(K) + f^2(x) \right) + \frac{h}{n} R_3(f, K; h) \\ & \quad - \frac{2h^2}{n} f(x) R_1(f, K; h) - \frac{h^4}{n} R_1(f, K; h)^2 \end{aligned}$$

where $R_2(K) := \int zK(z)^2 dz$, $R_3(f, K; h) := \int_0^1 (1-t) \left(\int f''(x-thz) z^2 K^2(z) dz \right) dt$. By the regularity conditions of K and f'' , $R_2(K)$ and $R_3(f, K; h)$ are both finite. Thus, for each x ,

$$E(\widehat{f}(x; h) - f(x))^2$$

$$\begin{aligned}
&= \text{Var}(\widehat{f}(x; h)) + (E\widehat{f}(x; h) - f(x))^2 \\
&= (nh)^{-1} \|K\|_2^2 \cdot f(x) - n^{-1} \left(f'(x) R_2(K) + f^2(x) \right) + \frac{h}{n} R_3(f, K; h) \\
&\quad - \frac{2h^2}{n} f(x) R_1(f, K; h) - \frac{h^4}{n} R_1(f, K; h)^2 + h^4 R_1(f, K; h)^2,
\end{aligned}$$

so, by Fubini's theorem,

$$\begin{aligned}
E \left[\|\widehat{f} - f\|_2^2 \right] &= \int E(\widehat{f}(x; h) - f(x))^2 dx \\
&= h^4 \int R_1(f, K; h)^2 dx + (nh)^{-1} \|K\|_2^2 - n^{-1} \left(R_2(K) \int f'(x) dx + \int f^2(x) dx \right) \\
&\quad + \frac{h}{n} \int R_3(f, K; h) dx - \frac{2h^2}{n} \int f(x) R_1(f, K; h) dx - \frac{h^4}{n} \int R_1(f, K; h)^2 dx.
\end{aligned} \tag{A.51}$$

Define $A_t(x) := \int f''(x - thz) z^2 K(z) dz$. By Cauchy-Schwarz inequality, for any fixed x ,

$$R_1(f, K; h)^2 = \left| \int_0^1 (1-t) A_t(x) dt \right|^2 \leq \left(\int_0^1 (1-t)^2 dt \right) \cdot \left(\int_0^1 |A_t(x)|^2 dt \right).$$

Integrate over x and use Fubini's theorem,

$$\int R_1(f, K; h)^2 dx \leq \left(\int_0^1 (1-t)^2 dt \right) \cdot \int_0^1 \|A_t\|_2^2 dt = \frac{1}{3} \int_0^1 \|A_t\|_2^2 dt. \tag{A.52}$$

Now we rewrite A_t as a convolution. Let $k_{t,h}(u) := \frac{1}{th} \left(\frac{u}{th} \right)^2 K \left(\frac{u}{th} \right)$, then with the change of variable $u = thz$,

$$A_t(x) = \int f''(x - u) k_{t,h}(u) du = (f'' * k_{t,h})(x).$$

By Young's inequality ($L^2 * L^1 \rightarrow L^2$),

$$\begin{aligned}
\|A_t\|_2 &\leq \|f''\|_2 \cdot \|k_{t,h}\|_1 = \|f''\|_2 \cdot \int \frac{1}{th} \left(\frac{u}{th} \right)^2 K \left(\frac{u}{th} \right) du \\
&= \|f''\|_2 \cdot \int z^2 K(z) dz
\end{aligned}$$

$$\leq \sqrt{C_3} \int z^2 K(z) dz.$$

Plug this back to (A.52),

$$\int R_1(f, K; h)^2 dx \leq \frac{1}{3} C_3 \left(\int z^2 K(z) dz \right)^2 \leq \frac{1}{3} C_3 \int z^4 K(z) dz =: \tilde{R}_1(K),$$

where the last inequality is by Jensen's inequality. Furthermore,

$$\begin{aligned} \int R_3(f, K; h) dx &= \int \left(\int_0^1 (1-t) \left(\int f''(x - thz) z^2 K^2(z) dz \right) dt \right) dx \\ &= \int_0^1 (1-t) \left(\int z^2 K^2(z) \left(\int f''(x - thz) dx \right) dz \right) dt \\ &\leq \frac{1}{2} C_4 \int z^2 K^2(z) dz =: \tilde{R}_3(K), \end{aligned}$$

and

$$\begin{aligned} \int f(x) R_1(f, K; h) dx &= \int f(x) \left(\int_0^1 (1-t) \left(\int f''(x - thz) z^2 K(z) dz \right) dt \right) dx \\ &= \int_0^1 (1-t) \left(\int \left(\int f(x) f''(x - thz) dx \right) z^2 K(z) dz \right) dt \\ &\leq \frac{1}{2} \sqrt{C_1 C_3} \int z^2 K(z) dz =: R_4(K) \end{aligned}$$

Because K is bounded and has finite fourth moment, $\tilde{R}_1(K)$ and $\tilde{R}_3(K)$ are both finite. Plug these two terms back into (A.51), we obtain that $\forall f \in \mathcal{F}$,

$$\begin{aligned} E[\|\hat{f} - f\|_2^2] &\leq h^4 \tilde{R}_1(K) + (nh)^{-1} \|K\|_2^2 + n^{-1} (R_2(K) C_2 + C_1) + \frac{h}{n} \tilde{R}_3(K) - \frac{2h^2}{n} R_4(K) - \frac{h^4}{n} \tilde{R}_1(K). \end{aligned}$$

Because $E[\|\hat{f} - f\|_2^2] = h^4 \tilde{R}_1(K) + (nh)^{-1} \|K\|_2^2 + o(h^4 + (nh)^{-1})$, we can find h_{opt} such that the

dominant order $O(h^4 + (nh)^{-1})$ is minimized:

$$h_{opt} = \left(\frac{\|K\|_2^2}{\widetilde{R}_1(K)n} \right)^{1/5},$$

and

$$\begin{aligned} E[\|\hat{f} - f\|_2^2] &\leq 2\|K\|_2^{8/5} \widetilde{R}_1(K)^{1/5} \cdot n^{-4/5} + (R_2(K)C_2 + C_1)n^{-1} + \|K\|_2^{2/5} \widetilde{R}_3(K) \cdot \widetilde{R}_1(K)^{-1/5} \cdot n^{-6/5} \\ &\quad - 2R_4(K)\|K\|_2^{4/5} \widetilde{R}_1(K)^{-2/5} \cdot n^{-7/5} - \widetilde{R}_1(K)^{1/5} \|K\|_2^{8/5} \cdot n^{-9/5}. \end{aligned}$$

Note that the coefficients on the RHS only depend on K , so it is a uniform upper bound $\forall f \in \mathcal{F}$.

There must exist $\psi_K > 0$, s.t. $E[\|\hat{f} - f\|_2^2] \leq \psi_K n^{-4/5}, \forall n \in \mathbb{N}, \forall f \in \mathcal{F}$. $\square$

Now we are ready to prove Lemma A.7.1.

Proof of Lemma A.7.1. For any ϵ such that $0 < \epsilon < e^{-1}$, take $\ell = \max(\lfloor \log(1/\epsilon) \rfloor, \lfloor 2er_i^2/\sigma^2 \rfloor) + 2$, then there exists a constant $b(r_i, \sigma) \geq 1$, such that $\log(1/\epsilon) < \ell < b(r_i, \sigma) \log(1/\epsilon)$. By Lemma A.6.1, and noting by assumption that $8 \max(r_1, r_2)/r < 1$,

$$\begin{aligned} \|\hat{f}_i - f_i\|_1 &\leq \frac{1}{w_i} c_1(r, \sigma, r_1, r_2) \cdot \left(\frac{8 \max(r_1, r_2)}{r} \right)^{\log(\frac{1}{\epsilon})} \cdot \left(b(r_i, \sigma) \log\left(\frac{1}{\epsilon}\right) \right)^{5/4} + c_2(r_i, \sigma) \left(\frac{1}{2} \right)^{\log(\frac{1}{\epsilon})} \\ &\quad + \frac{1}{w_i} c_3(r, \sigma) \cdot \|\Delta f\|_2 \cdot \exp \left[b(r_i, \sigma) \log\left(\frac{1}{\epsilon}\right) \log\left(b(r_i, \sigma) \log\left(\frac{1}{\epsilon}\right)\right) \right. \\ &\quad \left. + c_4(r, \sigma) \cdot b(r_i, \sigma) \log\left(\frac{1}{\epsilon}\right) + c_5 \log\left(b(r_i, \sigma) \log\left(\frac{1}{\epsilon}\right)\right) \right]. \end{aligned}$$

By altering the bases and exponents in the right hand side, we obtain

$$\|\hat{f}_i - f_i\|_1 \leq \frac{1}{w_i} c_1(r, \sigma, r_1, r_2) \cdot \epsilon^{\log \frac{r}{8 \max(r_1, r_2)}} \left(\log\left(\frac{1}{\epsilon}\right) \right)^{5/4} + c_2(r_i, \sigma) \epsilon^{\log 2}$$

$$+ \frac{1}{w_i} c_3(r, \sigma, r_i) \cdot \|\Delta f\|_2 \cdot \left(\frac{1}{\epsilon}\right)^{b(r_i, \sigma) \cdot \log \log \left(\frac{1}{\epsilon}\right)} \cdot \left(\frac{1}{\epsilon}\right)^{c_4(r, r_i, \sigma)} \cdot \left(\log \left(\frac{1}{\epsilon}\right)\right)^{c_5}. \quad (\text{A.53})$$

Here, some terms involving $b(r_i, \sigma)$ are absorbed into the constants c_1, c_3 and c_4 . Since $\forall \alpha > 0$, there exists a small enough $0 < \epsilon_\alpha < e^{-1}$, s.t. $\forall 0 < \epsilon < \epsilon_\alpha$, $\log \left(\frac{1}{\epsilon}\right) < \left(\frac{1}{\epsilon}\right)^\alpha$, for $0 < \epsilon < \epsilon_\alpha$, we can replace $\log(1/\epsilon)$ in (A.53) with $(1/\epsilon)^\alpha$ for some $\alpha > 0$ to have

$$\begin{aligned} \|\widehat{f}_i - f_i\|_1 &\leq \frac{1}{w_i} c_1(r, \sigma, r_1, r_2) \cdot \epsilon^{\log \frac{r}{8 \max(r_1, r_2)} - \frac{5\alpha}{4}} + c_2(r_i, \sigma) \epsilon^{\log 2} \\ &\quad + \frac{1}{w_i} c_3(r, \sigma, r_i) \cdot \|\Delta f\|_2 \cdot \left(\frac{1}{\epsilon}\right)^{b(r_i, \sigma) \cdot \log \log \left(\frac{1}{\epsilon}\right) + c_4(r, r_i, \sigma) + c_5 \alpha}. \end{aligned}$$

Let $\beta := \min \left(\log \frac{r}{8 \max(r_1, r_2)} - \frac{5\alpha}{4}, \log 2 \right)$. Then, if $\|\Delta f\|_2 \leq \epsilon^{b(r_i, \sigma) \cdot \log \log \left(\frac{1}{\epsilon}\right) + c_4(r, r_i, \sigma) + c_5 \alpha + \beta}$, we have

$$\|\widehat{f}_i - f_i\|_1 \leq \left(\frac{1}{w_i} c_1(r, \sigma, r_1, r_2) + c_2(r_i, \sigma) + \frac{1}{w_i} c_3(r, \sigma, r_i) \right) \epsilon^\beta.$$

Note that by assumption, $\forall f \in \mathcal{F}$, $w_i \in [\underline{w}, 1 - \underline{w}]$, for $i = 1, 2$, so if we denote $\bar{c}_{1i}(r, \sigma, r_1, r_2, \underline{w}) = \frac{1}{\underline{w}} c_1(r, \sigma, r_1, r_2) + c_2(r_i, \sigma) + \frac{1}{\underline{w}} c_3(r, \sigma, r_i)$, we have $\|\widehat{f}_i - f_i\|_1 \leq \bar{c}_{1i}(r, \sigma, r_1, r_2, \underline{w}) \epsilon^\beta$.

We will select α in the following way so β can be in an asymptotic sense as small as possible.

- If $\frac{r}{8 \max(r_1, r_2)} \leq 2$, we can take α as small as possible such that $0 < \alpha < \frac{4}{5} \log \frac{r}{8 \max(r_1, r_2)}$. However, the smaller α is, the smaller ϵ_α becomes. It suffices to set $\alpha = \frac{2}{5} \log \frac{r}{8 \max(r_1, r_2)}$, then $\beta = \frac{1}{2} \log \frac{r}{8 \max(r_1, r_2)}$.
- If $\frac{r}{8 \max(r_1, r_2)} > 2$, we take $0 < \alpha < \frac{4}{5} \left(\log \frac{r}{8 \max(r_1, r_2)} - \log 2 \right) = \frac{4}{5} \log \frac{r}{16 \max(r_1, r_2)}$, and $\beta = \log 2$.

Therefore, when $0 < \epsilon < \epsilon_\alpha$, we have that for $i = 1, 2$,

$$P_f^n \left(\|\Delta f\|_2 < \epsilon^{b(r_i, \sigma) \cdot \log \log \left(\frac{1}{\epsilon}\right) + c_{2i}(r, r_1, r_2, \sigma)} \right) \leq P_f^n \left(\|\widehat{f}_i - f_i\|_1 < \bar{c}_{1i}(r, \sigma, r_1, r_2, \underline{w}) \epsilon^\beta \right),$$

which is equivalent to,

$$P_f^n \left(\|\widehat{f}_i - f_i\|_1 \geq \bar{c}_{1i}(r, \sigma, r_1, r_2, \underline{w}) \epsilon^\beta \right) \leq P_f^n \left(\|\Delta f\|_2 \geq \epsilon^{b(r_i, \sigma) \cdot \log \log(\frac{1}{\epsilon}) + c_{2i}(r, r_1, r_2, \sigma)} \right). \quad (\text{A.54})$$

Note that $\Delta f = \widehat{f} - f$, and now we account for the fact that $\widehat{f}$ is chosen to be a kernel density estimator:

$$\widehat{f}_{X_{1:n}, h}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad \forall x \in R,$$

where $K(x)$ is the standard Gaussian kernel. Then we define $q_n(X_1, \dots, X_n) = \|\widehat{f}_{X_{1:n}, h} - f\|_2$, and we have

$$\begin{aligned} & \left| q_n(X_1, \dots, X_k, \dots, X_n) - q_n(X_1, \dots, X'_k, \dots, X_n) \right| \\ &= \left| \|\widehat{f}_{X_{1:n}, h} - f\|_2 - \|\widehat{f}_{X_{\{1:n\} \setminus \{i\}} \cup X'_k, h} - f\|_2 \right| \\ &\leq \|\widehat{f}_{X_{1:n}, h} - \widehat{f}_{X_{\{1:n\} \setminus \{i\}} \cup X'_k, h}\|_2 \\ &= \frac{1}{nh} \left\| K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x - X'_i}{h}\right) \right\|_2 \\ &= \frac{1}{nh} \left(\int \left| K\left(\frac{x - X_i}{h}\right) - K\left(\frac{x - X'_i}{h}\right) \right|^2 dx \right)^{1/2} \\ &\leq \frac{1}{nh} \left(2 \int \left| K\left(\frac{x - X_i}{h}\right) \right|^2 + \left| K\left(\frac{x - X'_i}{h}\right) \right|^2 dx \right)^{1/2} \quad (\text{Cauchy-Schwarz}) \\ &= \frac{2}{n (2\sqrt{\pi}h)^{1/2}}. \end{aligned}$$

Therefore by the bounded difference inequality (Lemma A.7.2) we get:

$$P_f^n(\|\Delta f\|_2 - E[\|\Delta f\|_2] \geq \epsilon) = P_f^n(q_n - E[q_n] \geq \epsilon) \leq \exp\left(-\sqrt{\pi}hn\epsilon^2\right).$$

By Lemma A.7.3, if the density class $\mathcal{F}$ and the choice of kernel K satisfy the regularity conditions, then choosing bandwidth $h = c_K n^{-1/5}$, where c_K is a constant depending only on K , yields $\forall f \in \mathcal{F}, \forall n, E[\|\Delta f\|_2^2] \leq \psi_K n^{-4/5}$, for some constant ψ_K that only depends on the kernel K . Because $E[\|\Delta f\|_2]^2 \leq E[\|\Delta f\|_2^2]$, we have $E[\|\Delta f\|_2] \leq \varphi_K n^{-2/5}$ for some constant $\varphi_K > 0$.

When $n > \left(\frac{2\varphi_K}{\epsilon}\right)^{5/2}$, we have $E[\|\Delta f\|_2] < \frac{\epsilon}{2}$, $\forall f \in \mathcal{F}$, then

$$P_f^n(\|\Delta f\|_2 \geq \epsilon) \leq P_f^n\left(\|\Delta f\|_2 - E[\|\Delta f\|_2] \geq \frac{\epsilon}{2}\right) \leq \exp\left(-\tilde{c}_K n^{4/5} \epsilon^2\right), \quad (\text{A.55})$$

for some positive constant $\tilde{c}_K$ depending only on K .

Now we combine (A.54) and (A.55). We can take $\eta = \bar{c}_{1i}\epsilon^\beta$, and plug $\epsilon = (\frac{\eta}{\bar{c}_{1i}})^{1/\beta}$ in the right hand side of (A.54). For ϵ small enough such that $0 < \epsilon < \epsilon_\alpha$, and n large enough such that $n > (\frac{2\varphi_K}{\epsilon})^{5/2}$, or equivalently, for η small enough such that $\eta < \bar{c}_{1i}\epsilon_\alpha^\beta$, and $n > (2\varphi_K)^{5/2}(\frac{\bar{c}_{1i}}{\eta})^{5/(2\beta)}$, take $\ell = \max\left(\lfloor \beta^{-1} \log(\bar{c}_{1i}/\eta) \rfloor, \lfloor 2er_i^2/\sigma^2 \rfloor\right) + 2$, then we have

$$\begin{aligned} P_f^n\left(\|\widehat{f}_i - f_i\|_1 \geq \eta\right) &\leq P_f^n\left(\|\Delta f\|_2 \geq \left(\frac{\eta}{\bar{c}_{1i}}\right)^{\frac{1}{\beta}}\left(b_i \log \log\left(\frac{\bar{c}_{1i}}{\eta}\right)^{1/\beta} + c_{2i}\right)\right) \\ &\leq P_f^n\left(\|\Delta f\|_2 \geq \left(\frac{\eta}{\bar{c}_{1i}}\right)^{\frac{1}{\beta}}\left(b_i \log \log\left(\frac{\bar{c}_{1i}}{\eta}\right) - b_i \log \beta + c_{2i}\right)\right) \\ &\leq \exp\left(-\tilde{c}_K n^{4/5} \left(\frac{\eta}{\bar{c}_{1i}}\right)^{\frac{2}{\beta}}\left(b_i \log \log\left(\frac{\bar{c}_{1i}}{\eta}\right) - b_i \log \beta + c_{2i}\right)\right), \end{aligned}$$

where $\tilde{c}_K$ is a positive constant depending only on K , and $\bar{c}_{1i} = \bar{c}_{1i}(r, \sigma, r_1, r_2, \underline{w})$ is a positive constant depending on $i, r, \sigma, r_1, r_2, \underline{w}$, $c_{2i} = c_{2i}(r, r_1, r_2, \sigma)$ is a positive constant depending on i, r, r_1, r_2, σ , and $b_i = b(r_i, \sigma) \geq 1$ is a constant depending on r_i and σ . $\square$

A.8 Proofs for Section 2.7.2: Lemma 2.7.3 and Theorem 2.7.2

The last piece of the argument is to establish the existence of a test function with desired controls on type I and type II errors. Suppose that f_0 satisfies condition (C2.4). Define the test function $\phi_n : \mathbb{R}^n \rightarrow \{0, 1\}$ by

$$\phi_n = \mathbf{1}\{\|\widehat{f}_i - f_{i0}\|_1 > \epsilon_n\}, \quad (\text{A.56})$$

where $\widehat{f}_i$ is as defined in (A.26), and we take $\ell = \max\left(\lfloor \beta^{-1} \log(\bar{c}_{1i}/\epsilon_n) \rfloor, \lfloor 2er_i^2/\sigma^2 \rfloor\right) + 2$ as

specified in Lemma A.7.1. Although ϕ_n depends on i . We suppress this dependence in the notation for simplicity. The following lemma establish upper bounds of type I error, $E_{f_0}[\phi_n]$, and type II error, $E_f[1 - \phi_n]$ when n is large enough and ϵ_n is small enough. Note that f_{i0} is unknown, thus this test function cannot be used in practice, but the existence of such a test function is sufficient for the proof of the main theorem (Theorem 2.7.2) on the posterior contraction of the component density f_i .

Lemma A.8.1 (Restatement of Lemma 2.7.3). *Let $f \in \mathcal{F}$ and f_0 be as defined in (C2.3) and (C2.4). Fix a constant $M > 2$, and let $\{\epsilon_n\}$ be a positive constant sequence such that $\lim_{n \rightarrow \infty} \epsilon_n = 0$. Let $\hat{f}_i$ be the component density estimator defined in (A.26). Let $X_1, \dots, X_n$ be n i.i.d. samples from P_f . For a component $i = 1$ (or 2), consider the test function $\phi_n : \mathbb{R}^n \rightarrow \{0, 1\}$ as given by (A.56). Then for $\epsilon_n < \frac{\bar{c}_{1i}\epsilon_n^\beta}{M-1}$ and $n > (2\varphi_K)^{5/2} \left(\frac{\bar{c}_{1i}}{\epsilon_n} \right)^{5/(2\beta)}$,*

$$E_{f_0}[\phi_n] \leq \exp \left(-\tilde{c}_K n^{4/5} \left(\frac{\epsilon_n}{\bar{c}_{1i}} \right)^{\frac{2}{\beta} \left(b_i \log \log \left(\frac{\bar{c}_{1i}}{\epsilon_n} \right) - b_i \log \beta + c_{2i} \right)} \right),$$

and

$$\sup_{f \in \mathcal{F}: \|f_i - f_{i0}\|_1 > M\epsilon_n} E_f[1 - \phi_n] \leq \exp \left(-\tilde{c}_K n^{4/5} \left(\frac{(M-1)\epsilon_n}{\bar{c}_{1i}} \right)^{\frac{2}{\beta} \left(b_i \log \log \left(\frac{\bar{c}_{1i}}{(M-1)\epsilon_n} \right) - b_i \log \beta + c_{2i} \right)} \right),$$

where φ_K , $\tilde{c}_K$, α , β , $\bar{c}_{1i}$, b_i and c_{2i} are as defined in Lemma A.7.1.

Proof of Lemma A.8.1. By Lemma A.7.1, if we assume $\epsilon_n < \bar{c}_{1i}\epsilon_n^\beta$, and $n > (2\varphi_K)^{5/2} \left(\frac{\bar{c}_{1i}}{\epsilon_n} \right)^{5/(2\beta)}$, then

$$\begin{aligned} E_{f_0}[\phi_n] &= P_{f_0}^n(\|\hat{f}_i - f_{i0}\|_1 > \epsilon_n) \\ &\leq \exp \left(-\tilde{c}_K n^{4/5} \left(\frac{\epsilon_n}{\bar{c}_{1i}} \right)^{\frac{2}{\beta} \left(b_i \log \log \left(\frac{\bar{c}_{1i}}{\epsilon_n} \right) - b_i \log \beta + c_{2i} \right)} \right). \end{aligned}$$

For any $f \in \mathcal{F}$ such that $\|f_i - f_{i0}\|_1 > M\epsilon_n$, because

$$\begin{aligned}\|\widehat{f}_i - f_{i0}\|_1 &= \|\widehat{f}_i - f_i + f_i - f_{i0}\|_1 \\ &\geq \|f_i - f_{i0}\|_1 - \|\widehat{f}_i - f_i\|_1,\end{aligned}$$

we have

$$\begin{aligned}E_f[1 - \phi_n] &= P_f^n(\|\widehat{f}_i - f_{i0}\|_1 \leq \epsilon_n) \\ &\leq P_f^n(\|f_i - f_{i0}\|_1 - \|\widehat{f}_i - f_i\|_1 \leq \epsilon_n) \\ &= P_f^n(\|\widehat{f}_i - f_i\|_1 \geq \|f_i - f_{i0}\|_1 - \epsilon_n) \\ &\leq P_f^n(\|\widehat{f}_i - f_i\|_1 \geq (M - 1)\epsilon_n) \\ &\leq \exp\left(-\tilde{c}_K n^{4/5} \left(\frac{(M - 1)\epsilon_n}{\bar{c}_{1i}}\right)^{\frac{2}{\beta}} \left(b_i \log \log \left(\frac{\bar{c}_{1i}}{(M - 1)\epsilon_n}\right) - b_i \log \beta + c_{2i}\right)\right),\end{aligned}\tag{A.57}$$

if $(M - 1)\epsilon_n < \bar{c}_{1i}\epsilon_n^\beta$, and $n > (2\varphi_K)^{5/2} \left(\frac{\bar{c}_{1i}}{(M - 1)\epsilon_n}\right)^{5/(2\beta)}$. Note that (A.57) holds for all $f \in \mathcal{F}$ such that $\|f_i - f_{i0}\|_1 > M\epsilon_n$. Therefore we obtain for $i = 1, 2$,

$$\sup_{f \in \mathcal{F}: \|f_i - f_{i0}\|_1 > M\epsilon_n} E_f[1 - \phi_n] \leq \exp\left(-\tilde{c}_K n^{4/5} \left(\frac{(M - 1)\epsilon_n}{\bar{c}_{1i}}\right)^{\frac{2}{\beta}} \left(b_i \log \log \left(\frac{\bar{c}_{1i}}{(M - 1)\epsilon_n}\right) - b_i \log \beta + c_{2i}\right)\right).$$

□

Proof of Theorem 2.7.2. Let $X_1, \dots, X_n$ be n i.i.d. samples from P_{f_0} . Let $\tilde{\epsilon}_n$ be a positive constant sequence that goes to zero as $n \rightarrow \infty$. Define the event A_n as

$$A_n = \left\{ (X_1, \dots, X_n) \mid \int \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f) \geq e^{-(2+D)n\tilde{\epsilon}_n^2} \right\},\tag{A.58}$$

where $p_f(X_1, \dots, X_n)$ is the density of $X_1, \dots, X_n$ under P_f , and $p_0(X_1, \dots, X_n)$ is the density of $X_1, \dots, X_n$ under P_{f_0} . D is a positive constant. Let ϕ_n be given by (A.56). Let $M > 2$ be a constant. Then, the posterior probability of the set $\{f_i \mid \|f_i - f_{i0}\|_1 > M\epsilon_n\}$ can be expressed as

follows

$$\begin{aligned}
\Pi^{(n)}(\|f_i - f_{i0}\|_1 > M\epsilon_n | X_1, \dots, X_n) &= \frac{\int_{f: \|f_i - f_{i0}\|_1 > M\epsilon_n} p_f(X_1, \dots, X_n) d\Pi(f)}{\int p_f(X_1, \dots, X_n) d\Pi(f)} \\
&= (\phi_n + 1 - \phi_n) \frac{\int_{f: \|f_i - f_{i0}\|_1 > M\epsilon_n} \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f)}{\int \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f)} \\
&= (\phi_n + \mathbf{1}_{A_n^c} \cdot (1 - \phi_n) + \mathbf{1}_{A_n} \cdot (1 - \phi_n)) \frac{\int_{f: \|f_i - f_{i0}\|_1 > M\epsilon_n} \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f)}{\int \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f)} \\
&\leq \phi_n + \mathbf{1}_{A_n^c} + \mathbf{1}_{A_n} \cdot (1 - \phi_n) \frac{\int_{f: \|f_i - f_{i0}\|_1 > M\epsilon_n} \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f)}{\int \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f)} \\
&\leq \phi_n + \mathbf{1}_{A_n^c} + e^{(2+D)n\tilde{\epsilon}_n^2} (1 - \phi_n) \int_{f: \|f_i - f_{i0}\|_1 > M\epsilon_n} \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f)
\end{aligned}$$

The last inequality is due to the definition of the event A_n . It follows that

$$\begin{aligned}
&E_{f_0}[\Pi^{(n)}(\|f_i - f_{i0}\|_1 > M\epsilon_n | X_1, \dots, X_n)] \\
&\leq E_{f_0}[\phi_n] + E_{f_0}[\mathbf{1}_{A_n^c}] + e^{(2+D)n\tilde{\epsilon}_n^2} E_{f_0} \left[(1 - \phi_n) \int_{f: \|f_i - f_{i0}\|_1 > M\epsilon_n} \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f) \right]
\end{aligned}$$

By Lemma A.8.1, we have if $\epsilon_n < \frac{\bar{c}_{1i}\epsilon_\alpha^\beta}{M-1}$ and $n > (2\varphi_K)^{5/2} \left(\frac{\bar{c}_{1i}}{\epsilon_n} \right)^{5/(2\beta)}$, then

$$E_{f_0}[\phi_n] \leq \exp \left(-\tilde{c}_K n^{4/5} \left(\frac{\epsilon_n}{\bar{c}_{1i}} \right)^{\frac{2}{\beta}} \left(b_i \log \log \left(\frac{\bar{c}_{1i}}{\epsilon_n} \right) - b_i \log \beta + c_{2i} \right) \right). \quad (\text{A.59})$$

Furthermore, by Lemma 8.1 in [Ghosal et al. \(2000\)](#), and a similar argument used in Lemma 8.10 of [Ghosal and van der Vaart \(2017\)](#), for any $\epsilon > 0$, and any $C > 0$,

$$P_{f_0}^n \left(\int \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f) \geq \Pi(B_2(f_0, \epsilon)) e^{-(1+C)n\epsilon^2} \right) \geq 1 - \frac{1}{C^2 n \epsilon^2}.$$

By Lemma A.4.2, if we take $\epsilon = \tilde{\epsilon}_n = \log n / \sqrt{n}$, when n is larger than a constant threshold, we

have $\Pi(B_2(f_0, \tilde{\epsilon}_n)) \geq c_3 e^{-c_4 n \tilde{\epsilon}_n^2}$, therefore

$$P_{f_0}^n \left(\int \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f) \geq c_3 e^{-c_4 n \tilde{\epsilon}_n^2} e^{-(1+C)n \tilde{\epsilon}_n^2} \right) \geq 1 - \frac{1}{C^2 n \tilde{\epsilon}_n^2}.$$

Take $C = 1$, we can find some $D > 1$ such that

$$P_{f_0}^n \left(\int \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f) \geq e^{-(2+D)n \tilde{\epsilon}_n^2} \right) \geq 1 - \frac{1}{n \tilde{\epsilon}_n^2}.$$

Recall the definition of A_n in (A.58), then we obtain

$$E_{f_0} [\mathbf{1}_{A_n^c}] \leq \frac{1}{n \tilde{\epsilon}_n^2} = \frac{1}{(\log n)^2} \rightarrow 0, \text{ as } n \rightarrow \infty. \quad (\text{A.60})$$

Finally, by Lemma A.8.1,

$$\begin{aligned} & E_{f_0} \left[(1 - \phi_n) \int_{f: \|f_i - f_{i0}\|_1 > M \epsilon_n} \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} d\Pi(f) \right] \\ & \stackrel{\text{Fubini}}{=} \int_{f: \|f_i - f_{i0}\|_1 > M \epsilon_n} \left(\int p_0(X_1, \dots, X_n) \cdot \frac{p_f(X_1, \dots, X_n)}{p_0(X_1, \dots, X_n)} (1 - \phi_n) dX_{1:n} \right) d\Pi(f) \\ & = \int_{f: \|f_i - f_{i0}\|_1 > M \epsilon_n} E_f [1 - \phi_n] d\Pi(f) \\ & \leq \exp \left(-\tilde{c}_K n^{4/5} \left(\frac{(M-1)\epsilon_n}{\bar{c}_{1i}} \right)^{\frac{2}{\beta}} \left(b_i \log \log \frac{\bar{c}_{1i}}{(M-1)\epsilon_n} - b_i \log \beta + c_{2i} \right) \right). \end{aligned} \quad (\text{A.61})$$

Combining (A.59), (A.60) and (A.61),

$$\begin{aligned} & E_{f_0} [\Pi^{(n)}(\|f_i - f_{i0}\|_1 > M \epsilon_n | X_1, \dots, X_n)] \\ & \leq \exp \left(-\tilde{c}_K n^{4/5} \left(\frac{\epsilon_n}{\bar{c}_{1i}} \right)^{\frac{2}{\beta}} \left(b_i \log \log \frac{\bar{c}_{1i}}{\epsilon_n} - b_i \log \beta + c_{2i} \right) \right) + \frac{1}{(\log n)^2} \\ & \quad + \exp \left((2+D)n \tilde{\epsilon}_n^2 - \tilde{c}_K n^{4/5} \left(\frac{(M-1)\epsilon_n}{\bar{c}_{1i}} \right)^{\frac{2}{\beta}} \left(b_i \log \log \frac{\bar{c}_{1i}}{(M-1)\epsilon_n} - b_i \log \beta + c_{2i} \right) \right). \end{aligned}$$

For the the first and third term to go to zero as $n \rightarrow \infty$, there must exist a constant $C_M > (2+D)/\tilde{c}_K$

such that when n is large enough,

$$n^{4/5} \left(\frac{(M-1)\epsilon_n}{\bar{c}_{1i}} \right)^{\frac{2}{\beta}} \left(b_i \log \log \frac{\bar{c}_{1i}}{(M-1)\epsilon_n} - b_i \log \beta + c_{2i} \right) \geq C_M n \tilde{\epsilon}_n^2.$$

Take logarithm on both sides, we obtain

$$\frac{2}{\beta} \left(b_i \log \log \frac{\bar{c}_{1i}}{(M-1)\epsilon_n} - b_i \log \beta + c_{2i} \right) \cdot \left(\log \frac{\bar{c}_{1i}}{(M-1)\epsilon_n} \right) \leq \frac{4}{5} \log n - 2 \log \log n - \log C_M.$$

Denote $x = \frac{\bar{c}_{1i}}{(M-1)\epsilon_n}$. When $\epsilon_n < \frac{\bar{c}_{1i}}{(M-1)e^e}$, we have $x > e^e$, and hence $\log \log x > 1$. Therefore it suffices to solve the following

$$\frac{2}{\beta} (b_i - b_i \log \beta + c_{2i}) \cdot \left(\log \log \frac{\bar{c}_{1i}}{(M-1)\epsilon_n} \right) \cdot \left(\log \frac{\bar{c}_{1i}}{(M-1)\epsilon_n} \right) \leq \frac{4}{5} \log n - 2 \log \log n - \log C_M.$$

Let $c_{\beta,i} = 2(b_i - b_i \log \beta + c_{2i})/\beta$, which is positive by Lemma A.7.1, and let $y = \log \log x$, then we obtain

$$ye^y \leq \frac{4}{5c_{\beta,i}} \log n - \frac{2}{c_{\beta,i}} \log \log n - \frac{1}{c_{\beta,i}} \log C_M, \quad y > 0. \quad (\text{A.62})$$

The function $W(z)$, defined as the solution to $W(z)e^{W(z)} = z$, is known as the Lambert W function (see, e.g., [Corless et al. \(1996\)](#)). We consider the portion of the principle branch of $W(z)$ where $W(z)$ is real and positive, then (A.62) is equivalent to:

$$y \leq W \left(\frac{4}{5c_{\beta,i}} \log n - \frac{2}{c_{\beta,i}} \log \log n - \frac{1}{c_{\beta,i}} \log C_M \right). \quad (\text{A.63})$$

From equation (4.19) in [Corless et al. \(1996\)](#), we know as $z \rightarrow \infty$,

$$W(z) \geq \log z - \log \log z + \frac{\log \log z}{2 \log z},$$

so applying this lower bound to the RHS of (A.63) and recalling that $y = \log \log x$, it is sufficient

to have

$$\begin{aligned}
\log \log x &\leq \log \frac{z}{\log z} + \frac{\log \log z}{2 \log z} \\
\log x &\leq \exp \left(\log \frac{z}{\log z} \right) \cdot \exp \left(\frac{\log \log z}{2 \log z} \right) \\
&= \frac{z}{\log z} \cdot (\log z)^{\frac{1}{2 \log z}} \\
x &\leq \exp(z)^{(\log z)^{\frac{1}{2 \log z}} \cdot \frac{1}{\log z}}.
\end{aligned}$$

Because $(\log z)^{\frac{1}{2 \log z}} \downarrow 1$ as $z \rightarrow \infty$, it suffices to have that $x \leq \exp(z)^{\frac{1}{\log z}}$. Now plug in z and x . Because there must exist some constant $c_{M,\beta,i} > 1$ such that $z < c_{M,\beta,i} \log n$, we have $\log z < c_{i1} \log \log n$ for some constant $c_{i1} > 0$, so it is sufficient to have

$$\begin{aligned}
\frac{\bar{c}_{1i}}{(M-1)\epsilon_n} &\leq \left(n^{\frac{4}{5c_{\beta,i}}} \cdot (\log n)^{-\frac{2}{c_{\beta,i}}} \cdot C_M^{-\frac{1}{c_{\beta,i}}} \right)^{\frac{1}{c_{i1} \log \log n}} \\
&= n^{\frac{4}{5c_{\beta,i}c_{i1} \log \log n}} \cdot (\log n)^{\frac{1}{\log \log n} \cdot \left(-\frac{2}{c_{\beta,i}c_{i1}}\right)} \cdot C_M^{-\frac{1}{c_{\beta,i}c_{i1} \log \log n}}.
\end{aligned}$$

Observe that $(\log n)^{\frac{1}{\log \log n}} = \exp \left(\frac{1}{\log \log n} \cdot \log \log n \right) = e$. Moreover, since $C_M^{-\frac{1}{c_{\beta,i}c_{i1} \log \log n}} \rightarrow 1$ as $n \rightarrow \infty$, there exists a constant $0 < c_{i2} < 1$ such that $C_M^{-\frac{1}{c_{\beta,i}c_{i1} \log \log n}} > c_{i2}$ for all $n > 10$. Consequently, it suffices to solve

$$\begin{aligned}
\frac{\bar{c}_{1i}}{(M-1)\epsilon_n} &\leq c_{i2} n^{\frac{c_{i1}}{\log \log n}} \\
\epsilon_n &\geq c_{i2} \cdot n^{-\frac{c_{i1}}{\log \log n}},
\end{aligned}$$

for some positive constants c_{i1} and c_{i2} .

□

APPENDIX B

Appendix of Chapter 3

B.1 Algorithms

In this section, we describe the algorithms used in Section 3.3 in detail.

Notation. Let $\mathbf{X} \in \mathbb{R}^{n \times m}$ be the data matrix in ambient dimension $m \geq 1$, and let $\boldsymbol{\ell} \in \{-1, 1, 2, \dots, K\}^n$ be the HDBSCAN cluster-label vector ($\ell_i = -1$ marks noise). For each non-noise cluster $c \in \{1, \dots, K\}$, write $\mathcal{I}_c = \{i = 1, \dots, n \mid \ell_i = c\}$ and $\mathbf{X}_c$ for the corresponding sub-matrix. The L^∞ distance is $\|\mathbf{u} - \mathbf{v}\|_\infty = \max_{j=1, \dots, m} |u_j - v_j|$.

A box is parametrized by a centre $\mathbf{c} \in \mathbb{R}^m$ and a half-side vector $\mathbf{h} = (h^{(1)}, \dots, h^{(m)}) \in \mathbb{R}_+^m$,

$$\text{Box}(\mathbf{c}, \mathbf{h}) = \{\mathbf{x} \in \mathbb{R}^m : |x_j - c_j| \leq h^{(j)}, j = 1, \dots, m\}.$$

We write $|x|$ for componentwise absolute value, $\text{vol}(\mathbf{h}) = \prod_{j=1}^m 2h^{(j)}$ for the box volume, and use $\langle \cdot, \cdot \rangle$ for the Euclidean inner product on $\mathbb{R}^m$. For a path $\mathcal{P} = (p_1, \dots, p_K)$ in $\mathbf{X}$, the polyline distance from $\mathbf{x}$ to $\mathcal{P}$ is

$$d_{\mathcal{P}}(\mathbf{x}) = \min_{k=1, \dots, K-1} \min_{t \in [0, 1]} \|\mathbf{x} - (1-t)\mathbf{x}_{p_k} - t\mathbf{x}_{p_{k+1}}\|_2,$$

i.e. the minimum Euclidean distance from $\mathbf{x}$ to any point on a linear segment of $\mathcal{P}$. Throughout,

$\text{Quantile}_q(S)$ denotes the empirical q -quantile of a finite set $S \subset \mathbb{R}$.

User parameters: quantile $q \in (0, 1)$ for the global radius ; overlap target $\rho \in [0, 1)$; spacing tolerance $\tau \in [0, 1]$; error tolerance $\eta \in [0, 1]$ for the candidate-selection tie-breaker; local half-length quantile $q_{\text{loc}} \in (0, 1]$; padding multiplier $\beta \geq 1$; maximum-expand multiplier $M \geq 1$; inter-cluster gap $g \geq 0$; maximum overlap-resolution passes T_{max} ; minimum allowed half-side $h_{\text{min}} > 0$; minimum volume fraction $v_{\text{min}} \in (0, 1)$ for keeping a box after pruning.

Algorithm 2: CLUSTERBOXCOVER_m: covering one HDBSCAN cluster by L^∞ boxes in $\mathbb{R}^m$ with anisotropic per-box half-sides

Input: Data $X \in \mathbb{R}^{n \times m}$; labels ℓ ; cluster id c ; parameters $q, \rho, \tau, \eta, q_{\text{loc}}, \beta, M, h_{\text{min}}$.

Output: Diameter path $\mathcal{P}_c$, global radius h_c , set of boxes $\mathcal{B}_c = \{(\mathbf{x}_{i_b}, \mathbf{h}_b)\}_b$.

(1) Build the Euclidean MST of the cluster.

Form $G_c = (\mathcal{I}_c, E_c)$ where the edge weight of (i, j) is $\|\mathbf{x}_i - \mathbf{x}_j\|_2$ (Prim's algorithm).

(2) Extract the longest path.

Pick any $v_0 \in \mathcal{I}_c$. Let $a = \arg \max_v d_{G_c}(v_0, v)$ and $b = \arg \max_v d_{G_c}(a, v)$; the diameter path $\mathcal{P}_c = (p_1, \dots, p_K)$ is the unique tree path from $p_1 = a$ to $p_K = b$, recovered from parent pointers in the second search.

(3) Compute point-to-path distances.

For every $i \in \mathcal{I}_c$, compute $\delta_i = d_{\mathcal{P}_c}(\mathbf{x}_i)$ — the Euclidean distance from $\mathbf{x}_i$ to the polyline $\mathcal{P}_c$.

(4) Choose the global radius adaptively.

$h_c \leftarrow \text{Quantile}_q(\{\delta_i : i \in \mathcal{I}_c\})$.

(5) Place box centres along the diameter.

$C_c \leftarrow \text{PLACECENTERSALONGPATH}_m(X, \mathcal{P}_c, h_c, \rho, \tau, \eta, \mathcal{I}_c)$ (Algorithm 4).

(6) Assign cluster points to nearest centre.

For every $i \in \mathcal{I}_c$, set $\pi(i) = \arg \min_{j \in C_c} \|\mathbf{x}_i - \mathbf{x}_j\|_\infty$, breaking ties by the smaller index.

Let $\mathcal{N}_j = \{i \in \mathcal{I}_c : \pi(i) = j\}$ for each $j \in C_c$.

(7) Compute anisotropic per-box half-side lengths.

For each centre $j \in C_c$, set $\mathbf{h}_j \leftarrow \text{LOCALHALFLENGTHS}(X, j, \mathcal{N}_j, q_{\text{loc}}, \beta, M, h_c, h_{\text{min}})$ (Algorithm 3).

return diameter path $\mathcal{P}_c$, global radius h_c , and box set $\mathcal{B}_c = \{(\mathbf{x}_j, \mathbf{h}_j) : j \in C_c\}$, each box tagged with cluster id c .

Algorithm 3: LOCALHALFLENGTHS: anisotropic per-box half-side lengths from the points assigned to a centre

Input: Data X ; centre index j ; assigned points $\mathcal{N}_j \subseteq \mathcal{I}_c$; quantile q_{loc} , padding β , max-expand M , global radius h_c , minimum half-side h_{min} .

Output: Half-side vector $\mathbf{h}_j \in \mathbb{R}_+^m$.

if $\mathcal{N}_j = \emptyset$ **then**

return $\mathbf{h}_j = (h_{\text{min}}, \dots, h_{\text{min}})$

for $\ell = 1, \dots, m$ **do**

$h_j^{(\ell)} \leftarrow \text{Quantile}_{q_{\text{loc}}}(\{|X_{i,\ell} - X_{j,\ell}| : i \in \mathcal{N}_j\})$

$h_j^{(\ell)} \leftarrow \max(\beta h_j^{(\ell)}, h_{\text{min}})$ /* padding */

$h_j^{(\ell)} \leftarrow \min(h_j^{(\ell)}, M h_c)$ /* cap relative to global radius */

return $\mathbf{h}_j$.

Algorithm 4: PLACECENTERSALONGPATH_m: middle-out greedy selection of box centres on the diameter path with L^∞ overlap control, generalised to $\mathbb{R}^m$

Input: Data X ; ordered diameter path $\mathcal{P} = (p_1, \dots, p_K)$; global radius h ; overlap target ρ , tolerance τ , error tolerance η ; cluster index set $\mathcal{I}$.

Output: Ordered subset of centre indices $C \subseteq \mathcal{P}$.

Spacing targets.

$$d^* \leftarrow 2h(1 - \rho), \quad d_{\min} \leftarrow \max(d^* - \tau h, 0), \quad \varepsilon \leftarrow \eta h.$$

Locate the middle of the path.

$$s_1 \leftarrow 0;$$

$$\text{for } k = 2, \dots, K, s_k \leftarrow s_{k-1} + \|\mathbf{x}_{p_k} - \mathbf{x}_{p_{k-1}}\|_2.$$

$$m \leftarrow \arg \min_k |s_k - s_K/2|;$$

$$C \leftarrow \{m\};$$

mark all points covered by $\text{Box}(\mathbf{x}_{p_m}, h\mathbf{1})$.

for direction $\text{dir} \in \{\text{left}, \text{right}\}$ **do**

$\text{cur} \leftarrow m$

while cur has unvisited path indices in direction dir **do**

 /* Build candidates: indices on the current side of the path whose L^∞ distance to all already-chosen centres is at least $d_{\min}$. */

$$\mathcal{A} \leftarrow \{j \in \mathcal{P} \text{ on the dir-side of cur} : \min_{p \in C} \|\mathbf{x}_{p_j} - \mathbf{x}_{p_p}\|_\infty \geq d_{\min}\}$$

if $\mathcal{A} = \emptyset$ **then break**

 /* For each candidate, compute spacing error to the current centre and the count of new cluster points it would cover. */

for $j \in \mathcal{A}$ **do**

$$e_j \leftarrow \|\mathbf{x}_{p_j} - \mathbf{x}_{p_{\text{cur}}}\|_\infty - d^*$$

$$n_j \leftarrow \#\{i \in \mathcal{I} : i \text{ uncovered and } \|\mathbf{x}_i - \mathbf{x}_{p_j}\|_\infty \leq h\}$$

 /* Tie-break: among candidates whose spacing error is within ε of the best, prefer the one with the most new coverage; break further ties by the smaller spacing error. */

$$e^* \leftarrow \min_{j \in \mathcal{A}} e_j; \mathcal{A}^* \leftarrow \{j \in \mathcal{A} : e_j \leq e^* + \varepsilon\}$$

$$j^* \leftarrow \arg \max_{j \in \mathcal{A}^*} (n_j, -e_j)$$

$$C \leftarrow C \cup \{j^*\}; \text{ mark points covered by } \text{Box}(\mathbf{x}_{p_{j^*}}, h\mathbf{1}); \text{ cur} \leftarrow j^*$$

return C ordered by position along $\mathcal{P}$.

Algorithm 5: RESOLVEINTERCLUSTEROVERLAPS_m: post-processing pass that shrinks pairs of overlapping boxes from different clusters, one axis at a time

Input: Data X ; labels ℓ ; box list $\mathcal{B} = \bigcup_{c=1}^K \mathcal{B}_c$ (each box tagged with its cluster id $c(b)$); gap g , max passes $T_{\max}$, $h_{\min}$, $v_{\min}$.

Output: Cleaned box list $\mathcal{B}'$.

```

for pass  $t = 1, \dots, T_{\max}$  do
  changed  $\leftarrow$  false
  for  $(i, j) : 1 \leq i < j \leq |\mathcal{B}|, c(b_i) \neq c(b_j)$  do
    if not  $\text{BoxOverlap}_g(b_i, b_j)$  then continue
    /* Step (a): for each box, check whether it covers
       any points from the other cluster ('contamination')
       and shrink it along the best axis to remove them. */
    for  $(b, b') \in \{(b_i, b_j), (b_j, b_i)\}$  do
       $F \leftarrow \{k \in \mathcal{I}_{c(b')} : \mathbf{x}_k \in \text{Box}(\mathbf{x}_{i_b}, \mathbf{h}_b)\}$  /* foreign points inside  $b$  */
      if  $F \neq \emptyset$  then
         $b \leftarrow \text{SHRINKTOREDUCECONTAMINATION}(b, F)$  (Algorithm 6);
        changed  $\leftarrow$  true
    /* Step (b): if the two boxes still overlap after
       contamination cleanup, shrink the smaller-volume one
       to enforce the gap  $g$  on the offending axes. */
    if  $\text{BoxOverlap}_g(b_i, b_j)$  then
      if  $\text{vol}(\mathbf{h}_{b_i}) \leq \text{vol}(\mathbf{h}_{b_j})$  then  $b_i \leftarrow \text{SHRINKTOENFORCEGAP}(b_i, b_j, g)$ 
      else  $b_j \leftarrow \text{SHRINKTOENFORCEGAP}(b_j, b_i, g)$ 
      changed  $\leftarrow$  true
  if not changed then break

```

Pruning.

$\mathcal{B}' \leftarrow \{b \in \mathcal{B} : \text{vol}(\mathbf{h}_b) > v_{\min} \cdot \text{vol}(\mathbf{h}_b^{(0)}) \text{ and } \exists i \in \mathcal{I}_{c(b)} \text{ with } \mathbf{x}_i \in \text{Box}(\mathbf{x}_{i_b}, \mathbf{h}_b)\}$

where $\mathbf{h}_b^{(0)}$ are the original (pre-shrink) half-sides recorded for box b .

return $\mathcal{B}'$.

Predicate.

$\text{BoxOverlap}_g(b, b') := \bigwedge_{\ell=1}^m (|X_{i_b, \ell} - X_{i_{b'}, \ell}| < h_b^{(\ell)} + h_{b'}^{(\ell)} + g)$ /* L^∞ -cube overlap with margin */

Algorithm 6: SHRINKTOREDUCECONTAMINATION: greedy single-axis shrink to evict foreign points from a box

Input: Box $b = (\mathbf{x}_{i_b}, \mathbf{h}_b)$; foreign points $F \subset \mathcal{I}$; bound $h_{\min}$; tolerance $\epsilon > 0$ (default 10^{-9}).
Output: Shrunk box b' .
for $\ell = 1, \dots, m$ **do**
 /* Smallest ℓ -th half-side that excludes every foreign point along axis ℓ . */
 $h_\ell^\dagger \leftarrow \max\left(\min_{k \in F} |X_{k,\ell} - X_{i_b,\ell}| - \epsilon, h_{\min}\right)$
 $h_\ell^\dagger \leftarrow \min(h_\ell^\dagger, h_b^{(\ell)})$
 $b_\ell \leftarrow$ box obtained from b by replacing $h_b^{(\ell)}$ with $h_\ell^\dagger$
 $f_\ell \leftarrow$ foreign-cluster points still inside b_ℓ ; $o_\ell \leftarrow$ own-cluster points still inside b_ℓ ;
 $s_\ell \leftarrow h_b^{(\ell)} - h_\ell^\dagger$
return $b' \leftarrow b_{\ell^\star}$ where $\ell^\star = \arg \min_\ell (|f_\ell|, -|o_\ell|, s_\ell)$ (lexicographic).

Algorithm 7: SHRINKTOENFORCEGAP: shrink box b along one axis so the pair (b, b') satisfies the gap constraint

Input: Boxes b, b' ; gap g ; bound $h_{\min}$, tolerance ϵ .
Output: Shrunk box b' .
Find offending axes.
For $\ell = 1, \dots, m$, let $v_\ell \leftarrow h_b^{(\ell)} + h_{b'}^{(\ell)} + g - |X_{i_b,\ell} - X_{i_{b'},\ell}|$.
Let $D = \{\ell : v_\ell > 0\}$.
if $D = \emptyset$ **then return** b
for $\ell \in D$ **do**
 $h_\ell^\dagger \leftarrow \max(h_b^{(\ell)} - v_\ell - \epsilon, h_{\min})$
 $b_\ell \leftarrow$ box obtained from b by replacing $h_b^{(\ell)}$ with $h_\ell^\dagger$
 $o_\ell \leftarrow$ own-cluster points still inside b_ℓ ; $s_\ell \leftarrow h_b^{(\ell)} - h_\ell^\dagger$
return $b_{\ell^\star}$ where $\ell^\star = \arg \min_{\ell \in D} (-|o_\ell|, s_\ell)$ (lexicographic).

Algorithm 8: COVERALLCLUSTERS $_m$: outer driver

Input: Data X , labels ℓ , all parameters.
Output: Final box list $\mathcal{B}^\star$ and per-cluster coverage ratios $\{\gamma_c\}$.
for $c \leftarrow 1$ **to** K **do**
 Run CLUSTERBOXCOVER $_m$ (Algorithm 2) to obtain box set $\mathcal{B}_c$
 $\mathcal{B} \leftarrow \bigsqcup_{c=1}^K \mathcal{B}_c$
 $\mathcal{B}^\star \leftarrow \text{RESOLVEINTERCLUSTEROVERLAPS}_m(X, \ell, \mathcal{B}, g, T_{\max}, h_{\min}, v_{\min})$ (Algorithm 5)
for $c \leftarrow 1$ **to** K **do**
 $\gamma_c \leftarrow |\mathcal{I}_c|^{-1} \#\{i \in \mathcal{I}_c : \exists b \in \mathcal{B}^\star \text{ with } c(b) = c \text{ and } \mathbf{x}_i \in \text{Box}(\mathbf{x}_{i_b}, \mathbf{h}_b)\}$
return $(\mathcal{B}^\star, \{\gamma_c\}_{c=1}^K)$.

Remarks.

1. From scalar to vector half-sides. In ambient dimension m , each box has its own half-side vector $\mathbf{h} \in \mathbb{R}^m$ rather than a single scalar. `LOCALHALFLENGTHS` (Algorithm 3) determines this vector from the points assigned to the box’s centre using a per-axis quantile, so that boxes are tight along thin directions and broad along the cluster’s elongated direction. The cap $h_j^{(\ell)} \leq M h_c$ prevents any single box from growing far beyond the global radius set in Step (4) of Algorithm 2.
2. Polyline (segment) distance vs. vertex distance. In Step (3) of Algorithm 2, the distance from a point to the diameter path is the minimum distance to the segments, not just to the path vertices. This matters when the path has sparsely sampled vertices: for diffuse clusters in higher dimension a vertex-only distance would over-estimate δ_i and inflate the global radius h_c via the quantile rule.
3. Inter-cluster overlap resolution as two-step shrinking. Algorithm 5 resolves each cross-cluster collision in two greedy steps. First, it attempts to evict every foreign point from each box by shrinking the box along its single most effective axis; this is the contamination step. If overlap remains after both boxes have been cleaned, the second step shrinks the smaller-volume box on its offending axes only enough to open a gap of g . The pruning stage at the end discards any box whose volume has shrunk below $v_{\min}$ of its original volume, or whose own cluster no longer has a point inside it. The result is a deterministic procedure: in each pass, every box pair is examined once, and the loop exits as soon as a full pass finishes without modification.
4. Contamination axis is chosen by a fixed sequence of tie-breaks. `SHRINKTOREDUCECONTAMINATION` (Algorithm 6) first selects the axis that minimizes the number of foreign points still inside. If multiple axes achieve the same minimum, it chooses among them the one that preserves the largest number of own-cluster points. If ties still remain, it chooses the axis requiring the smallest shrink amount. This ordered tie-breaking rule matches the `sortperm(... by = i → (...))` clause in the source. The `SHRINKTOENFORCEGAP` sub-procedure uses a similar but simpler tie-breaking rule, since it does not need to count

foreign points: it operates only on pairs that are already non-overlapping.

BIBLIOGRAPHY

- Andrews, J. L. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: the t eigen family. Statistics and Computing, 22(5):1021–1029.
- Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. The Annals of Statistics, 2(6):1152–1174.
- Aptekarev, A., Buyarov, V., and Degeza, I. (1995). Asymptotic behavior of the lp-norms and the entropy for general orthogonal polynomials. Sbornik: Mathematics, 82(2):373.
- Aptekarev, A., Dehesa, J., Sánchez-Moreno, P., and Tulyakov, D. (2012). Asymptotics of lp-norms of hermite polynomials and rényi entropy of rydberg oscillator states. Contemp. Math, 578:19–29.
- Aragam, B., Dan, C., Xing, E. P., and Ravikumar, P. (2020). Identifiability of nonparametric mixture models and bayes optimal clustering. The Annals of Statistics, 48(4):2277–2302.
- Aragam, B. and Yang, R. (2025). Model-free estimation of latent structure via multiscale nonparametric maximum likelihood. arXiv:2410.22248 [stat.ME].
- Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. The Annals of Statistics, 1(2):353–355.
- Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. Bayesian Analysis, 1(1):121 – 143.
- Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. Journal of the Royal Statistical Society Series B: Statistical Methodology, 79(1):125–148.
- Boyeau, P., Regier, J., Gayoso, A., Jordan, M. I., Lopez, R., and Yosef, N. (2023). An empirical bayes method for differential expression analysis of single cells with deep generative models. Proceedings of the National Academy of Sciences, 120(21):e2209124120.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33:1877–1901.
- Campello, R. J., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Pacific-Asia conference on knowledge discovery and data mining, pages 160–172. Springer.

- Chacón, J. E. and Duong, T. (2013). Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting. Electronic Journal of Statistics, 7:499 – 532.
- Chakraborty, S. and Nguyen, X. (2026). Learning mixtures of nonparametric and convolutional measures on effectively low-dimensional affine spaces. arXiv:2604.17236 [math.ST].
- Chen, J. (1995). Optimal rate of convergence for finite mixture models. The Annals of Statistics, 23(1):221–233.
- Chen, S. and Finkenstädt, B. (2024). Bayesian spline-based hidden markov models with applications to actimetry data and sleep analysis. Journal of the American Statistical Association, 119(548):2833–2843.
- Chen, Y.-C., Genovese, C. R., and Wasserman, L. (2016). A comprehensive approach to mode clustering. Electronic Journal of Statistics, 10(1):210 – 241.
- Chopin, N. (2011). Fast simulation of truncated gaussian distributions. Statistics and Computing, 21(2):275–288.
- Corless, R. M., Gonnet, G. H., Hare, D. E., Jeffrey, D. J., and Knuth, D. E. (1996). On the lambert w function. Advances in Computational mathematics, 5(1):329–359.
- Cramér, H. and Wold, H. (1936). Some theorems on distribution functions. Journal of the London Mathematical Society, s1-11(4):290–294.
- Dang, U. J., Gallagher, M. P., Browne, R. P., and McNicholas, P. D. (2023). Model-based clustering and classification using mixtures of multivariate skewed power exponential distributions. Journal of Classification, 40(1):145–167.
- Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through bayesian sampling. Journal of the Royal Statistical Society: Series B (Methodological), 56(2):363–375.
- Dinh, L., Krueger, D., and Bengio, Y. (2014). Nice: Non-linear independent components estimation. arXiv:1410.8516.
- Dinh, L., Sohl-Dickstein, J., and Bengio, S. (2016). Density estimation using real nvp. arXiv:1605.08803 [cs.LG].
- Do, D., Do, L., McKinley, S. A., Terhorst, J., and Nguyen, X. (2024). Dendrogram of mixing measures: Hierarchical clustering and model selection for finite mixture models. arXiv:2403.01684 [stat.ME].
- Dombowsky, A. and Dunson, D. B. (2025). Bayesian clustering via fusing of localized densities. Journal of the American Statistical Association, 120(551):1775–1786.
- Escobar, M. D. (1994). Estimating normal means with a dirichlet process prior. Journal of the American Statistical Association, 89(425):268–277.

- Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96, page 226–231. AAAI Press.
- Fan, M., Wang, J., Kashyap, V. L., Lee, T. C., van Dyk, D. A., and Zezas, A. (2023). Identifying diffuse spatial structures in high-energy photon lists. The Astronomical Journal, 165(2):66.
- Fang, K.-T., Geng, Z., and Tian, G.-L. (2000). Statistical inference for truncated dirichlet distribution and its application in misclassification. Biometrical Journal: Journal of Mathematical Methods in Biosciences, 42(8):1053–1068.
- Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. The Annals of Statistics, 1(2):209–230.
- Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. The Annals of Statistics, 2(4):615–629.
- Fernández, P. J., Ferrari, P. A., and Grynberg, S. P. (2007). Perfectly random sampling of truncated multinormal distributions. Advances in Applied Probability, 39(4):973–990.
- Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2013). Mixtures of shifted asymmetric laplace distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(6):1149–1157.
- Franzolini, B., Lijoi, A., Prünster, I., and Rebaudo, G. (2025). Multivariate species sampling models. arXiv:2503.24004 [math.ST].
- Ge, H., Chen, Y., Wan, M., and Ghahramani, Z. (2015). Distributed inference for dirichlet process mixture models. In International Conference on Machine Learning, pages 2276–2284. PMLR.
- Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the gaussian mixture sieve. The Annals of Statistics, 28(4):1105–1127.
- Geweke, J. et al. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In Computing science and statistics: Proceedings of the 23rd symposium on the interface, volume 571, page 578. Fairfax, Virginia: Interface Foundation of North America, Inc.
- Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999). Posterior consistency of dirichlet mixtures in density estimation. The Annals of Statistics, 27(1):143–158.
- Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. The Annals of Statistics, 28(2):500 – 531.
- Ghosal, S. and van der Vaart, A. W. (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. The Annals of Statistics, 29(5):1233 – 1263.
- Ghosal, S. and van der Vaart, A. W. (2007). Posterior convergence rates of dirichlet mixtures at smooth densities. The Annals of Statistics, 35(2):697 – 723.

- Ghosal, S. and van der Vaart, A. W. (2017). Fundamentals of Nonparametric Bayesian Inference, volume 44. Cambridge University Press.
- Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.
- Guha, A., Ho, N., and Nguyen, X. (2021). On posterior contraction of parameters and interpretability in bayesian mixture modeling. Bernoulli, 27(4):2159–2188.
- Hartigan, J. A. (1981). Consistency of single linkage for high-density clusters. Journal of the American Statistical Association, 76(374):388–394.
- Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. The Annals of Statistics, 46(6A):2844 – 2870.
- Hinneburg, A. and Keim, D. A. (2003). A general approach to clustering in large databases with noise. Knowledge and Information Systems, 5(4):387–415.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33:6840–6851.
- Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. The Annals of Statistics, 44(6):2726 – 2755.
- Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. Neural Networks, 12(3):429–439.
- Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. Journal of the American statistical Association, 96(453):161–173.
- Jain, S. and Neal, R. M. (2004). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. Journal of Computational and Graphical Statistics, 13(1):158–182.
- Jones, D. E., Kashyap, V. L., and Van Dyk, D. A. (2015). Disentangling overlapping astronomical sources using spatial and spectral information. The Astrophysical Journal, 808(2):137.
- Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. Statistics and Computing, 21(1):93–105.
- Karlis, D. and Santourian, A. (2009). Model-based clustering with non-elliptically contoured distributions. Statistics and Computing, 19(1):73–83.
- Kingma, D. P. and Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. Advances in Neural Information Processing Systems, 31.
- Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. arXiv:1312.6114 [stat.ML].

- Kotecha, J. H. and Djuric, P. M. (1999). Gibbs sampling approach for generation of truncated multivariate gaussian random variables. In 1999 IEEE international conference on acoustics, speech, and signal processing. Proceedings. ICASSP99 (Cat. No. 99CH36258), volume 3, pages 1757–1760. IEEE.
- Langrock, R., Adam, T., Leos-Barajas, V., Mews, S., Miller, D. L., and Papastamatiou, Y. P. (2018). Spline-based nonparametric inference in general state-switching models. Statistica Neerlandica, 72(3):179–200.
- Langrock, R., Kneib, T., Sohn, A., and DeRuiter, S. L. (2015). Nonparametric inference in hidden markov models using p-splines. Biometrics, 71(2):520–528.
- Levine, M., Hunter, D. R., and Chauveau, D. (2011). Maximum smoothed likelihood for multivariate mixtures. Biometrika, 98(2):403–416.
- Lin, T. I., Lee, J. C., and Hsieh, W. J. (2007). Robust mixture modeling using the skew t distribution. Statistics and Computing, 17(2):81–92.
- Lloyd, S. (1982). Least squares quantization in pcm. IEEE transactions on information theory, 28(2):129–137.
- Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In International Conference on Machine Learning, pages 4114–4124. PMLR.
- Lopez, R., Tagasovska, N., Ra, S., Cho, K., Pritchard, J., and Regev, A. (2023). Learning causal representations of single cells via sparse mechanism shift modeling. In Conference on Causal Learning and Reasoning, pages 662–691. PMLR.
- Lopez-Delisle, L. and Delisle, J.-B. (2022). baredsc: Bayesian approach to retrieve expression distribution of single-cell data. BMC bioinformatics, 23(1):36.
- MacEachern, S. N. (2000). Dependent dirichlet processes. Unpublished manuscript, Department of Statistics, The Ohio State University.
- MacEachern, S. N. and Müller, P. (1998). Estimating mixture of dirichlet process models. Journal of Computational and Graphical Statistics, 7(2):223–238.
- Manole, T. and Khalili, A. (2021). Estimating the number of components in finite mixture models via the group-sort-fuse procedure. The Annals of Statistics, 49(6):3043–3069.
- McKeough, K., Kashyap, V. L., Siemiginowska, A., Van Dyk, D. A., Yang, S., Meng, X.-L., Martin, B., and Zezas, A. (2026). The lira-ising model: Estimating the boundaries of irregularly shaped x-ray sources. The Astrophysical Journal Supplement Series, 283(1):14.
- McQueen, J. B. (1967). Some methods of classification and analysis of multivariate observations. In Proc. of 5th Berkeley Symposium on Math. Stat. and Prob., pages 281–297.

- Meyer, A. D., van Dyk, D. A., Kashyap, V. L., Campos, L. F., Jones, D. E., Siemiginowska, A., and Zezas, A. (2021). ebascs: Disentangling overlapping astronomical sources ii, using spatial, spectral, and temporal information. Monthly Notices of the Royal Astronomical Society, 506(4):6160–6180.
- Miller, J. W. and Dunson, D. B. (2019). Robust bayesian inference via coarsening. Journal of the American Statistical Association, 114(527):1113–1125. PMID: 31942084.
- Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. Journal of the American Statistical Association, 113(521):340–356.
- Müller, P., Quintana, F., and Rosner, G. (2004). A method for combining inference across related nonparametric bayesian models. Journal of the Royal Statistical Society Series B: Statistical Methodology, 66(3):735–749.
- Müller, P., Quintana, F. A., Jara, A., and Hanson, T. (2015). Bayesian nonparametric data analysis, volume 1. Springer.
- Munkres, J. R. (2020). Topology. Pearson, 2nd (updated) edition.
- Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9(2):249–265.
- Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. The Annals of Statistics, 41(1):370 – 400.
- Pakman, A. and Paninski, L. (2014). Exact hamiltonian monte carlo for truncated multivariate gaussians. Journal of Computational and Graphical Statistics, 23(2):518–542.
- Parzen, E. (1962). On estimation of a probability density function and mode. The Annals of Mathematical Statistics, 33(3):1065–1076.
- Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. Statistics and Computing, 10(4):339–348.
- Petralia, F., Rao, V., and Dunson, D. B. (2012). Repulsive mixtures. In Advances in Neural Information Processing Systems, volume 25. Curran Associates, Inc.
- Prabhakaran, S., Azizi, E., Carr, A., and Pe’er, D. (2016). Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In International Conference on Machine Learning, pages 1070–1079. PMLR.
- Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.
- Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. SIAM review, 26(2):195–239.

- Reizinger, P., Guo, S., Huszár, F., Schölkopf, B., and Brendel, W. (2024). Identifiable exchangeable mechanisms for causal structure and representation learning. [arXiv:2406.14302 \[stat.ML\]](https://arxiv.org/abs/2406.14302).
- Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). Journal of the Royal Statistical Society Series B: Statistical Methodology, 59(4):731–792.
- Rinaldo, A. and Wasserman, L. (2010). Generalized density clustering. The Annals of Statistics, 38(5):2678 – 2722.
- Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008). The nested dirichlet process. Journal of the American statistical Association, 103(483):1131–1154.
- Rodríguez, C. E. and Walker, S. G. (2014). Univariate bayesian nonparametric mixture modeling with unimodal kernels. Statistics and Computing, 24(1):35–49.
- Roeder, G., Metz, L., and Kingma, D. (2021). On linear identifiability of learned representations. In International Conference on Machine Learning, pages 9030–9039. PMLR.
- Schiebinger, G., Wainwright, M. J., and Yu, B. (2015). The geometry of kernelized spectral clustering. The Annals of Statistics, 43(2):819–846.
- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5):612–634.
- Sethuraman, J. (1994). A constructive definition of dirichlet priors. Statistica Sinica, 4(2):639–650.
- Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components— an alternative to reversible jump methods. The Annals of Statistics, 28(1):40–74.
- Sun, Z., Chen, L., Xin, H., Jiang, Y., Huang, Q., Cillo, A. R., Tabib, T., Kolls, J. K., Bruno, T. C., Lafyatis, R., et al. (2019). A bayesian mixture model for clustering droplet-based single-cell transcriptomic data from population studies. Nature Communications, 10(1):1649.
- Tai, W. M. and Aragam, B. (2023). Tight bounds on the hardness of learning simple nonparametric mixtures. In The Thirty Sixth Annual Conference on Learning Theory, pages 2849–2849. PMLR.
- Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581.
- Teicher, H. (1961). Identifiability of mixtures. The Annals of Mathematical Statistics, 32(1):244–248.
- Thomann, P., Steinwart, I., and Schmid, N. (2015). Towards an axiomatic approach to hierarchical clustering of measures. Journal of Machine Learning Research, 16:1949–2002.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

- Walker, S. G. (2007). Sampling the dirichlet mixture model with slices. Communications in Statistics—Simulation and Computation®, 36(1):45–54.
- Wand, M. P. and Jones, M. C. (1994). Kernel smoothing. CRC press.
- Wei, Y., Mukherjee, S., and Nguyen, X. (2025). Minimum ϕ -distance estimators for finite mixing measures. arXiv:2304.10052 [math.ST].
- Wei, Y. and Nguyen, X. (2022). Convergence of de finetti’s mixing measure in latent structure models for observed exchangeable sequences. The Annals of Statistics, 50(4):1859–1889.
- Wilkes, B. and Tucker, W., editors (2019). The Chandra X-ray Observatory: Exploring the high energy universe. 2514-3433. IOP Publishing.
- Wu, Y. and Yang, P. (2020). Optimal estimation of Gaussian mixtures via denoised method of moments. The Annals of Statistics, 48(4):1981 – 2007.
- Xie, F. and Xu, Y. (2020). Bayesian repulsive gaussian mixture model. Journal of the American Statistical Association, 115(529):187–203.
- Xu, L. and Jordan, M. I. (1996). On convergence properties of the em algorithm for gaussian mixtures. Neural Computation, 8(1):129–151.
- Xu, Y., Luo, K., and Ma, L. (2025). A tree-based kernel for densities and its applications in clustering dnase-seq profiles. arXiv:2509.15480 [stat.ME].
- Zezas, A., Fabbiano, G., Baldi, A., Schweizer, F., King, A., Ponman, T., and Rots, A. (2006). Chandra monitoring observations of the antennae galaxies. i. catalog of source properties. The Astrophysical Journal Supplement Series, 166(1):211–248.
- Zezas, A., Fabbiano, G., Rots, A., and Murray, S. (2002). Chandra observations of “the antennae” galaxies (ngc 4038/4039). ii. detection and analysis of galaxian x-ray sources. The Astrophysical Journal Supplement Series, 142(2):239–260.