

Parameter Estimation and Multilevel Clustering with Mixture and Hierarchical Models

by

Nhat Pham Minh Ho

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2017

Doctoral Committee:

Associate Professor Long Nguyen, Co-Chair
Professor Ya'acov Ritov, Co-Chair
Professor Xuming He
Professor Bhramar Mukherjee



Nhat Pham Minh Ho

minhnhat@umich.edu

ORCID iD: 0000-0001-7774-1833

© Nhat Pham Minh Ho 2017

Dedicated to my family

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my great advisors, Long Nguyen and Ya'acov Ritov, for their amazing guidance and tremendous support throughout this wonderful journey. Long has been not only a fantastic mentor and teacher but also a role model for me for his professional achievements. I am greatly influenced by his knowledgeability and appreciation of statistical problems, his amazing ability to come up with interesting and important problems and to deal with difficult topics, and his invaluable advice towards my academic career. The endless encouragements from Long have been important momentums to push me forward during difficult moments of my research. I also would like to thank Ya'acov for his generosity with time and helpful advice during my job search process. Even though we had only been able to work together for about one year, your knowledgeability and deep understanding of various statistical problems have helped to shape my future research directions tremendously.

I would like to thank Bhramar Mukherjee and Xuming He for agreeing to be on my thesis defense committee. I am also grateful to Xuming for his support during my job search process as well as his invaluable advice towards my academic career. I would like to thank all the faculty members for their wonderful graduate courses that help me develop a broad perspective on statistics and machine learning. I would also like to thank Judy McDonald and David Clark for their great helps with the paperwork during my graduate study. I am very fortunate to have many friends in the department. Mikhail Yurochkin and Aritra Guha have been working with me on

several projects regarding mixture and hierarchical models. One of these work has produced fantastic results in Chapter VI of this thesis. I am also grateful to Yun-Jhong Wu, Hossein Keshavarz, Can Le, Dao Nguyen, Naveen Narisetty, Jesus Arroyo, Yingchuan Wang, Alexander Geissing, Michael Hornstein, JoonHa Park, Jun Guo, Teal Guidici, Adam Hall, Timothy Necamp, and Kam Chung Wong for interesting and helpful conversations about life and research.

Finally, I would like to thank my family for their constant love and support. To my daughter and my wife, thank you for always standing with me throughout my journey. Without you, I will not be where I have been now.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGEMENTS	iii
LIST OF FIGURES	ix
LIST OF TABLES	xi
ABSTRACT	xii
 CHAPTER	
I. Introduction	1
1.1 Statistical efficiency of parameter estimation in finite mixture models	1
1.1.1 Mixture models	1
1.1.2 Wasserstein metric	3
1.1.3 Statistical efficiency of parameter estimation	4
1.2 Robust inference with mixture and hierarchical models	7
1.3 Multi-levels clustering via optimal transport perspective	9
1.4 Thesis organization	10
II. On strong identifiability and convergence rates of parameter estimation in finite mixtures	13
2.1 Introduction	13
2.2 Preliminaries	21
2.3 General theory of strong identifiability	23
2.3.1 Definitions and general identifiability bounds	24
2.3.2 Characterization of strong identifiability	30
2.4 Minimax lower bounds, MLE rates and illustrations	34
2.4.1 Minimax lower bounds and MLE rates of convergence	34
2.4.2 Illustrations	39

2.5	Proofs of key theorems	42
2.5.1	Strong identifiability in exact-fitted mixtures	43
2.5.2	Strong identifiability in over-fitted mixtures	46
2.6	Proofs of other results	53
2.6.1	Extension to the whole domain in exact-fitted mixtures	53
2.6.2	The importance of boundedness conditions in the over-fitted setting	53
2.6.3	Characterization of strong identifiability	54
III.	Convergence rates of parameter estimation for some weakly identifiable finite mixtures	65
3.1	Introduction	65
3.1.1	Main results for Gaussian mixtures	68
3.1.2	Results for other weakly identifiable classes	72
3.2	Proof of main results for Gaussian mixtures	74
3.2.1	On the order \bar{r}	74
3.2.2	Discussion of conditions in Theorem 3.1.1	76
3.2.3	Sharp identifiability bounds	77
3.2.4	Proof of Theorem 3.1.1	78
3.3	Gamma mixtures and location extensions	82
3.4	Simulations	86
3.5	Proofs of other propositions and theorems	88
3.5.1	Proofs for over-fitted Gaussian mixtures	88
3.5.2	Mixture of Gamma distributions and location-exponential distributions	99
3.5.3	Proofs for remaining results	106
IV.	Singularity structures and impacts on parameter estimation in finite mixtures of distributions	113
4.1	Introduction	114
4.2	Background	120
4.2.1	Parameter spaces and geometries	120
4.2.2	Estimation settings	123
4.3	Singularity structure in finite mixture models	124
4.3.1	Beyond Fisher information	124
4.3.2	Behavior of likelihood in a Wasserstein neighborhood	127
4.3.3	Construction of r -minimal forms	138
4.3.4	Polynomial limits of r -minimal form coefficients . .	141
4.4	O-mixtures of skewnormal distributions	143
4.4.1	Special cases	145
4.4.2	General results	149
4.4.3	Properties of the system of limiting polynomial equa- tions	154

4.5	E-mixtures of skewnormal distributions	157
4.5.1	Singularity level of $G_0 \in \mathcal{S}_1 \cup \mathcal{S}_2$	158
4.5.2	Singularity levels of $G_0 \in \mathcal{S}_3$: a summary	162
4.6	Discussion and concluding remarks	164
4.7	Appendix A	166
4.7.1	Singularity structure of \mathcal{S}_3 : detailed analysis	166
4.8	Appendix B	179
4.8.1	Proofs for Section 3	179
4.8.2	Proofs for Section 4	180
4.8.3	Proofs for Section 5	189
4.8.4	Proofs for Section 4.7	200
4.8.5	Proofs for auxiliary results	207
V.	Robust estimation of mixing measures in finite mixture models	209
5.1	Introduction	210
5.2	Background	215
5.3	Minimum Hellinger distance estimator with non-singular Fisher information matrix	217
5.3.1	Well-specified kernel setting	221
5.3.2	Misspecified kernel setting	223
5.3.3	Analysis of WS Algorithm	228
5.4	Different approach with minimum Hellinger distance estimator	231
5.5	Extension to non-standard settings	235
5.5.1	A singular Fisher information matrix	235
5.5.2	Extension to varying true parameters	238
5.6	Empirical studies	240
5.6.1	Synthetic data	240
5.6.2	Real data	244
5.7	Summaries and discussions	246
5.8	Proofs of key results	247
5.9	Appendix A	256
5.10	Appendix B	267
VI.	Multilevel clustering via Wasserstein means	276
6.1	Introduction	276
6.2	Background	279
6.3	Clustering with multilevel structure data	282
6.3.1	Multilevel Wasserstein Means (MWM) Algorithm .	282
6.3.2	Multilevel Wasserstein Means with Sharing	288
6.4	Consistency results	289
6.5	Empirical studies	290
6.5.1	Synthetic data	290
6.5.2	Real data analysis	292

6.6	Discussion	295
6.7	Appendix A	296
6.8	Appendix B	298
6.9	Appendix C	304
6.10	Appendix D	308
VII.	Conclusions and suggestions	313
7.1	Statistical efficiency, computational complexity, and high dimensionality of mixture and hierarchical models	313
7.1.1	Statistical efficiency of parameter estimation	313
7.1.2	Computational complexity of parameter estimation	314
7.1.3	Efficient models in high dimensional clustering	315
7.1.4	Computational complexity of MCMC methods	316
7.2	Semi-parametric inference of finite mixtures of regression models	316
7.3	Statistical applications of optimal transport theory	317
BIBLIOGRAPHY	319	

LIST OF FIGURES

Figure

2.1	Mixture of Student's t-distributions. Left: Exact-fitted setting. Right: Over-fitted setting.	40
2.2	Mixture of multivariate generalized Gaussian distributions. Left: Exact-fitted setting. Right: Over-fitted setting.	40
2.3	Mixture of location-scale Gaussian distributions, which satisfy first-order identifiability but not second-order identifiability condition. Left panel: Exact-fitted setting. Middle and right panels are for over-fitted setting by one extra component. Right panel shows that $h \gtrsim W_2^2$ no longer holds as $h \rightarrow 0$	42
2.4	MLE rates for location-covariance mixtures of Gaussians. Left: Exact-fitted — $W_1 \asymp n^{-1/2}$. Right: Over-fitted by one — $W_4 \asymp n^{-1/8}$. . .	43
3.1	Location-scale Gaussian mixtures. From left to right: (1) Exact-fitted setting; (2) Over-fitted by one component; (3) Over-fitted by one component; (4) Over-fitted by two components.	86
3.2	MLE rates for location-covariance mixtures of Gaussians. L to R: (1) Exact-fitted: $W_1 \asymp n^{-1/2}$. (2) Over-fitted by one: $W_4 \asymp n^{-1/8}$. (3) Over-fitted by two: $W_6 \asymp n^{-1/12}$	87
3.3	MLE rates for shape-rate mixtures of Gamma distributions. L to R: (1) Generic/Exact-fitted: $W_1(\hat{G}_n, G_0) \asymp n^{-1/2}$. (2) Generic/Over-fitted: $W_2 \asymp n^{-1/4}$. (3) Pathological/Exact-fitted: $W_1 \approx 1/(\log n)^{1/2}$. (4) Pathological/Over-fitted: $W_1 \approx 1/(\log n)^{1/2}$	87
4.1	The illustration of the elimination steps from a complete collection of derivatives of f up to the order 3 to a reduced system of linearly independent partial derivatives, cf. Lemma 4.4.3. The circled derivatives are eliminated from the partial derivatives present in the 3-minimal form. $A \rightarrow B$ means that B is involved in the representation of A under the reduction.	145
4.2	The singularity level of G_0 relative to \mathcal{E}_{k_0} is determined by partition based on zeros of polynomials P_1, P_2 into subsets $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$. Here, "NC" stands for nonconformant.	158
4.3	The level of singularity structure of $G_0 \in \mathcal{S}_3$ when $P_1(\boldsymbol{\eta}^0) = 0$. Here, "NC" stands for nonconformant. The term $\bar{s}(G_0)$ is defined in (4.37).	161

4.4	The level of singularity structure of $G_0 \in \mathcal{S}_3$ when $P_1(\boldsymbol{\eta}^0) \neq 0$. Here, "NC" stands for nonconformant. The term $\bar{s}(G_0)$ is defined in (4.37).	162
5.1	Performance of \widehat{G}_n in Algorithm 1 under the well-specified kernel setting and fixed G_0 . Top figures - left to right: (1) $W_1(\widehat{G}_n, G_0)$ under Gaussian case. (2) $W_1(\widehat{G}_n, G_0)$ under Cauchy case. Bottom figures - left to right: (1) $W_1(\widehat{G}_n, G_0)$ under Skew normal case. (2) Percentage of time $\widehat{m}_n = 3$ obtained from 100 runs.	242
5.2	Performance of \widehat{G}_n in Algorithm 1 under the well-specified kernel setting and varied G_0 . Left to right: (1) $W_1(\widehat{G}_n, G_0)$ under Gaussian case. (2) $W_1(\widehat{G}_n, G_0)$ under Cauchy case. (3) Percentage of time $\widehat{m}_n = 3$ obtained from 100 runs.	243
5.3	Performance of \widehat{G}_n in Algorithm 1 under misspecified kernel setting. L to R: (1) $W_1(\widehat{G}_n, G_*)$ under Gaussian case. (2) $W_1(\widehat{G}_n, G_*)$ under Cauchy case. (3) Percentage of time $\widehat{m}_n = 4$ obtained from 100 runs.	244
5.4	From left to right: (1) Histogram of SLC activity data. (2) Density plot from mixture of two normals based on Algorithm 1, WS algorithm, MKE algorithm, and MLE.	246
6.1	Data with a lot of small groups: (a) NC data with constant variance; (b) NC data with non-constant variance; (c) LC data with constant variance; (d) LC data with non-constant variance	291
6.2	Data with few big groups: (a) NC data with constant variance; (b) NC data with non-constant variance; (c) LC data with constant variance; (d) LC data with non-constant variance	291
6.3	Clustering representation for two datasets: (a) Five image clusters from <i>Labelme</i> data discovered by MWMS algorithm: tag-clouds on the left are accumulated from all images in the clusters while six images on the right are randomly chosen images in that cluster; (b) StudentLife discovered network with three node groups: (1) discovered student clusters, (3) student nodes, (5) discovered activity location (from Wifi data); and two edge groups: (2) Student to cluster assignment, (4) Student involved to activity location. Node sizes (of discovered nodes) depict the number of element in clusters while edge sizes between <i>Student</i> and <i>activity location</i> represent the popularity of student's activities.	292
6.4	Examples of images used in LabelMe dataset. Each image consists of different annotated regions.	293

LIST OF TABLES

Table

5.1	Summary of parameter estimates in SLC activity data from mixture of two normal distributions with Algorithm 1, WS algorithm, MKE algorithm, and EM algorithm. Here, p_i, η_i, τ_i represents the weights, means, and variance respectively.	246
6.1	Clustering performance for LabelMe dataset.	294

ABSTRACT

Parameter Estimation and Multilevel Clustering with
Mixture and Hierarchical Models

by

Nhat Pham Minh Ho

Chairs: Long Nguyen and Ya'acov Ritov

In the big data era, data are typically collected at massive scales and often carry complex structures, which lead to unprecedented modeling and computational challenges. In numerous applications of engineering and applied sciences there are indisputable evidence of the presence of hidden subpopulations in the whole data where each subpopulation has its own features. Due to their great modeling flexibility, mixture and hierarchical models have been widely utilized by researchers to uncover these multi-level structures. However, several outstanding problems arise from these models. Firstly, it has long been observed in practice that convergence behaviors of latent variables in these models are problematic. Secondly, state of the art hierarchical models tend to perform unsatisfactorily under large-scale and complex structures settings of data. Last but not least, in many practical problems mixture and hierarchical models are strongly affected by outliers or departures from model assumptions.

The overarching themes in the thesis focus on dealing with these challenges. Our main contributions include the following. We develop a systematic understanding of statistical efficiency of parameter estimation in finite mixture models. Our studies

make explicit the deep links between model singularities, parameter estimation convergence rates, and the algebraic geometry of the parameter space for mixtures of continuous distributions. Next, we develop robust estimators of mixing measure in finite mixture models using the idea of minimum Hellinger distance estimator, model selection criteria, and super-efficiency phenomenon. Finally, we propose efficient and scalable joint optimization approaches to cluster a potentially large hierarchically structured corpus of data, which aim to simultaneously partition data in each group and discover grouping patterns among groups.

CHAPTER I

Introduction

In this chapter, we outline several key contributions of the thesis in separate sections. Background knowledge are included to make each section self-contained and transparent.

1.1 Statistical efficiency of parameter estimation in finite mixture models

1.1.1 Mixture models

Mixture models have been a popular modeling tool for making inference about heterogeneous data [Lindsay, 1995, McLachlan and Basford, 1988]. They have been used extensively in various application domains arising from biological, physical, and social sciences. Under mixture modeling, data are viewed as samples from a collection of unobserved or latent subpopulations, each positing its own distribution and associated parameters. Learning about subpopulation-specific parameters is essential to the understanding of the underlying heterogeneity. Statistical efficiency related to parameter estimation in finite mixture models, however, remain poorly understood — as noted in a recent textbook [DasGupta, 2008] (pg. 571), “mixture models are riddled with difficulties such as nonidentifiability”.

To be more concrete, let us state a specific formulation of mixture models. Assume that each subpopulation is distributed according to a density function (with respect to Lebesgue measure on an Euclidean space \mathcal{X}) that belongs to a known density class $\{f(x|\theta, \Sigma), \theta \in \Theta \subset \mathbb{R}^{d_1}, \Sigma \in \Omega \subset S_{d_2}^{++}, x \in \mathcal{X}\}$. Here, $d_1 \geq 1, d_2 \geq 0, S_{d_2}^{++}$ is the set of all $d_2 \times d_2$ symmetric positive definite matrices. A finite mixture density with k mixing components can be defined in terms of f and a discrete mixing measure $G = \sum_{i=1}^k p_i \delta_{(\theta_i, \Sigma_i)}$ with k support points as follows

$$p_G(x) = \int f(x|\theta, \Sigma) dG(\theta, \Sigma) = \sum_{i=1}^k p_i f(x|\theta_i, \Sigma_i).$$

The standard goal of learning about subpopulation-specific parameters is to understand behaviors of point estimates of the masses p_i and the support points (θ_i, Σ_i) of G according to given sample X_1, \dots, X_n from the mixture density $p_G(x)$. However, there are two fundamental challenges which had hindered the attempts of several researchers. Firstly, for any two discrete probability measures G and G_0 , typical similarity metrics between probability distributions, such as Kullback-Leibler distance, Hellinger distance, or Total variation distance, are not able to provide meaningful results when they are used to study the distance between G and G_0 . Secondly, the Fisher information matrix with respect to G may be singular, which implies that the estimation techniques such as the maximum likelihood estimator and standard Bayesian procedures do not admit root- n parametric rate of convergence. These challenges necessitate the needs to develop an appropriate similarity distance between two discrete probability measures, which is summarized in Section 1.1.2, as well as systematic methods to understand the complex singularity structures of Fisher information matrix, which are also the main contributions of Chapter II, Chapter III, and Chapter IV and are summarized in Section 1.1.3.

1.1.2 Wasserstein metric

As shown by [Nguyen \[2013\]](#), the convergence of mixture model parameters can be measured in terms of a Wasserstein distance on the space of mixing measures G . In particular, let $G = \sum_{i=1}^k p_i \delta_{(\theta_i, \Sigma_i)}$ and $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, \Sigma_i^0)}$ be two discrete probability measures on parameter space $\Theta \times \Omega$, which is equipped with metric ρ . Now, the Wasserstein distance of order r , for a given $r \geq 1$ (cf. [Villani \[2009\]](#)), can be defined as

$$W_r(G, G_0) = \left(\inf_{\mathbf{q}} \sum_{i,j} q_{ij} \rho^r((\theta_i, \Sigma_i), (\theta_j^0, \Sigma_j^0)) \right)^{1/r},$$

where the infimum is taken over all joint probability distributions \mathbf{q} on $[1, \dots, k] \times [1, \dots, k_0]$ such that, when expressing \mathbf{q} as a $k \times k_0$ matrix, the marginal constraints hold: $\sum_j q_{ij} = p_i$ and $\sum_i q_{ij} = p_j^0$.

For any sequence of discrete mixing measures G_n , the convergence of mixing measures G_n in Wasserstein distances can be translated to convergence of G_n 's atoms and probability masses. In particular, suppose that a sequence of mixing measures $G_n \rightarrow G_0$ under W_r metric at a rate $\omega_n = o(1)$. If all G_n have the same number of atoms $k = k_0$ as that of G_0 , then the set of atoms of G_n converge to the k_0 atoms of G_0 at the same rate ω_n under ρ metric. If G_n have varying $k_n \in [k_0, k]$ number of atoms, where k is a fixed upper bound, then a subsequence of G_n can be constructed so that each atom of G_0 is a limit point of a certain subset of atoms of G_n — the convergence to each such limit also happens at rate ω_n . Some atoms of G_n may have limit points that are not among G_0 's atoms — the mass associated with those atoms of G_n must vanish at the generally faster rate ω_n^r (since $r \geq 1$).

1.1.3 Statistical efficiency of parameter estimation

To address parameter estimation rates, a natural approach is to study the behavior of mixing distributions that arise in the mixture models. This approach is well-developed in the context of nonparametric deconvolution [Carroll and Hall, 1988, Zhang, 1990, Fan, 1991], but these results are confined to only a specific type of model — location mixtures. Beyond location mixtures there have been far fewer results. In particular, when the parameter space $\Theta \times \Omega$ is univariate, parameter estimation in over-fitted mixture models, i.e., the settings when the true number of subpopulations (or equivalently components) is unknown and is bounded by some given number, was shown to have slow convergence rate $n^{-1/4}$ under second-order identifiability condition of kernel density function by [Chen, 1995]. For multi-dimensional parameters, the $(\log n/n)^{1/4}$ rate of posterior contraction of parameters was established by Nguyen [2013] under Wasserstein metric of second order. However, the convergence rates of parameter estimation under general setting of parameters with multiple types, including matrix-variate parameters, remain poorly understood. In Chapter II, we provide a comprehensive understanding of parameter estimation behaviors under that multiple types setting of parameters when we introduce and utilize first- and second-order identifiability condition of kernel density function, which are stronger versions of classical identifiability condition [Teicher, 1961]. In particular, under the first order identifiability of kernel density function and the exact-fitted setting of finite mixture models, the convergence rates of estimating G is $n^{-1/2}$ under W_1 metric. On the other hand, under the second order identifiability of kernel density function and the over-fitted setting of finite mixture models, the convergence rates of parameter estimation is $n^{-1/4}$ under W_2 metric. However, these convergence rates are not applicable to classical mixture models such as location-scale Gaussian mixtures or shape-rate Gamma mixtures, which belong to weakly identifiable models, i.e., those whose kernel density functions do not satisfy the strong identifiability conditions.

Location-scale Gaussian mixtures are widely regarded as one of the most utilized modeling tools in statistics. Problematic convergence behaviors exhibited by Gaussian mixtures have long been observed in practice, but a concrete understanding remains largely unavailable. In Chapter III, we demonstrated a remarkably interesting phenomenon in over-fitted Gaussian mixtures: the minimax lower bound and the MLE convergence rate of parameter estimation are characterized by the solvability of a system of polynomial equations, which ultimately depends on how much we potentially overfit these models. In particular, the estimation rate is $n^{-1/8}$ under the 4-th order Wasserstein distance, if overfitting by one extra component; $n^{-1/12}$ under the 6-th order Wasserstein distance if overfitting by two extra components. These results present a cautionary tale about the limitation of Gaussian mixtures, when it comes to assessing the quality of parameter estimation, but only when the number of mixing components is unknown. Since a tendency in practice is to "over-fit" the mixture generously with many more extra mixing components, our theory warns against this because as we have shown, the convergence rate via standard methods such as MLE for subpopulation-specific parameters deteriorates rapidly with the number of redundant components.

Even though the results obtained for Gaussian mixtures contain considerable insights about weak identifiability regarding parameter estimation in finite mixture models, they only touch upon the surface of a more general phenomenon relating to the degeneracy of Fisher information matrix. In Chapter IV, we proposed a comprehensive framework for analyzing parameter estimation behavior in finite mixture models, addressing directly the situations where the non-singularity condition of the Fisher information matrix may not hold. A fundamental notion that we introduced is called *singularity level*, a natural or infinite value given to every element in the parameter space. Fisher information singularities simply correspond to points in the parameter space whose singularity level is non-zero. Within the set of Fisher in-

formation singularities, the parameter space can be partitioned into disjoint subsets determined by different singularity levels. The singularity level describes in a precise manner the variation of the mixture likelihood with respect to changes in model parameters. This concept enables us to quantify the varying degrees of identifiability and singularity, some of which were implicitly exploited in previous chapters mentioned above.

The singularity level describes in a precise manner the variation of the mixture likelihood with respect to changes in model parameters. The statistical implication of the singularity level is easy to describe: given an i.i.d. n -sample from a (true) mixture model, a parameter value of singularity level r admits $n^{-1/2(r+1)}$ minimax lower bound for any estimator tending to the true parameter(s), as well as the same maximum likelihood estimator's convergence rate (up to a logarithmic factor and under some conditions). Thus, singularity level 0 results in root- n convergence rate for parameter estimation. Fisher singularity corresponds to singularity level 1 or greater than 1, resulting in convergence rates $n^{-1/4}, n^{-1/6}, n^{-1/8}$ or so on. The detailed picture of the distribution of singularity levels, however, can be extremely complex to capture. Remarkably, there are examples of finite mixtures for which the compact parameter space can be partitioned into disjoint subsets whose singularity level ranges from 0 to 1 to 2, . . . , up to infinity. This leads us to a story of finite mixtures of skewnormal distributions, which are rich and increasingly popular models used with asymmetric data. The singularity structure of the skewnormal mixtures is perhaps one of the more complex among the parametric mixture models that we have typically encountered in the literature. We were able to identify subsets with singularity level 0, 1, 2, . . . all the way up to infinity even in the setting of mixtures with known number of mixing components. As a result, if we were to vary the true parameter values, we would encounter a phenomenon akin to that of "phase transition" on the statistical efficiency of parameter estimation occurring within the same model class.

1.2 Robust inference with mixture and hierarchical models

In many practical problems, mixture and hierarchical models are strongly affected by outliers or (at least) some departures from model assumptions. Therefore, it is important to develop robust or semi-parametric inference of hierarchical models to better reflect the instability and complex structure in the data. In finite mixture models, apart from underlying mixing measures, true kernel density functions of each subpopulation in the data are, in many scenarios, unknown. One popular way to overcome such issue is to employ semi-parametric approach [Bickel et al., 1993]. In particular, we estimate the true kernel function from some classes of functions and achieve the estimation of mixing measure accordingly. However, parameter identifiability guarantee for such method is very difficult to establish even when the parameter space is simple [Bordes et al., 2006, Hunter et al., 2007]. Therefore, parameter estimation from semi-parametric approach is usually not reliable.

Perhaps, the most common and simplest approach to avoid the identifiability issue is to choose some kernel function that we tactically believe the data are generated from, and utilize that kernel to fit the model to obtain an estimate of the mixing measure. However, in various scenarios the chosen kernel and the true kernel are most likely different, i.e., we are under a misspecified kernel setting. Hence, the estimation of mixing measure under this approach may be highly unstable. The robustness question is unavoidable. Our principal goal in Chapter V therefore, is the construction of robust estimators of mixing measure where the estimation of both the number of components and the values of its parameters are of interest. Moreover, these estimators should achieve the best possible convergence rates under various assumptions on both the chosen kernel and the true kernel. When the true number of components is known, various robust methods had been proposed in the literature, see for example [Woodward et al., 1984, Donoho and Liu, 1988, Cutler and Cordero-

Brana, 1996]. However, there has been scarce work for robust estimators when the true number of components is unknown. Recently, Woo and Sriram [2006] proposed a robust estimator of the number of components based on the idea of minimum Hellinger distance estimator [Beran, 1977]. However, their estimator faces certain limitations as it greatly relies upon the choice of bandwidth. In Chapter V, we propose flexible robust estimators of the mixing measure that address the limitations of the estimator in [Woo and Sriram, 2006]. Not only our estimators are computationally feasible and robust but they also possess various desirable properties, such as the flexible choice of bandwidth, the consistency of the number of components, and the best possible convergence rates of the parameters. In particular, our main contributions in Chapter V can be summarized in three major points. Firstly, we treat the well-specified kernel setting and misspecified kernel setting separately. Under both settings, we achieve the consistency of our estimators regarding the true number of components for any fixed bandwidth. Furthermore, when the bandwidth vanishes to 0 at an appropriate rate, the consistency of estimating the true number of components is also guaranteed. Secondly, for any fixed bandwidth, when the true kernel is identifiable in the first order the convergence rate $n^{-1/2}$ of parameter estimation is established under the well-specified kernel setting. Additionally, when that kernel is not identifiable in the first order, we demonstrate that our estimators still achieve the best possible convergence rate of parameter estimation. Finally, under the misspecified kernel setting, we demonstrate that our estimators converge to some mixing measure G_* that is closest to the true model under the Hellinger metric for any fixed bandwidth. When the chosen kernel is first order identifiable and G_* has finite number of components, the convergence rate $n^{-1/2}$ is also established under mild conditions of both chosen kernel and true kernel. Moreover, when G_* has infinite number of components, some analyses about the consistency of our estimators are also discussed.

1.3 Multi-levels clustering via optimal transport perspective

In numerous applications in engineering and sciences, data are often organized in a multilevel structure. For instance, a typical structural view of text data in machine learning is to have words grouped into documents, documents are grouped into corpora. A prominent strand of modeling and algorithmic works in the past couple decades has been to discover latent multilevel structures from these hierarchically structured data. For specific clustering tasks, one may be interested in simultaneously partitioning the data in each group (to obtain local clusters) and partitioning a collection of data groups (to obtain global clusters). Bayesian hierarchical models provide a powerful approach, exemplified by influential works such as [Blei et al., 2003, Pritchard et al., 2000, Teh et al., 2006]. More specific to the simultaneous and multilevel clustering problem, we mention the paper of [Rodriguez et al., 2008]. In this interesting work, a Bayesian nonparametric model, namely the nested Dirichlet process (NDP) model, was introduced that enables the inference of clustering of a collection of probability distributions from which different groups of data are drawn. With suitable extensions, this modeling framework has been further developed for simultaneous multilevel clustering, see for instance, [Wulsin et al., 2016, Nguyen et al., 2014, Huynh et al., 2016]. However, these models generally rely on MCMC algorithms to compute the posterior distribution of latent variables where the computational complexity grows exponentially as the data sizes are large.

The main focus of Chapter VI is on the multilevel clustering problem motivated in the aforementioned modeling works, but we shall take a purely optimization approach to deal with the computational issues of these works. We aim to formulate optimization problems that enable the discovery of multilevel clustering structures hidden in grouped data. Our technical approach is inspired by the role of optimal transport distances in hierarchical modeling and clustering problems. The Wasserstein distances

are intimately connected to the problem of clustering — this relationship goes back at least to the work of [Pollard, 1982], where it is pointed out that the well-known K-means clustering algorithm can be directly linked to the quantization problem — the problem of determining an optimal finite discrete probability measure that minimizes its second-order Wasserstein distance from the empirical distribution of given data [Graf and Luschgy, 2000].

If one is to perform simultaneous K-means clustering for hierarchically grouped data, both at the global level (among groups), and local level (within each group), then this can be achieved by a joint optimization problem defined with suitable notions of Wasserstein distances inserted into the objective function. In particular, multilevel clustering requires the optimization in the space of probability measures defined in *different* levels of abstraction, including the space of measures of measures on the space of grouped data. In summary, the main contributions of Chapter VI can be summarized into the following major points. Firstly, we propose a novel optimization formulation to the multilevel clustering problem using Wasserstein distances defined on different levels of the hierarchical data structure. Secondly, fast algorithms by exploiting the connection of our formulation to the Wasserstein barycenter problem are introduced. Thirdly, we establish consistency theorems for proposed estimates under very mild condition of data's distributions. Last but not least, several flexible alternatives are studied by introducing constraints that encourage the borrowing of strength among local and global clusters. The demonstration of efficiency and flexibility of our approach is carried out in a number of simulated and real data sets. As a consequence, our proposed model offers an attractive alternative to popular model-based approaches such as the nested Dirichlet Process model.

1.4 Thesis organization

The remainder of this thesis is organized as follows.

Chapter II: On strong identifiability and convergence rates of parameter estimation in finite mixtures This chapter studies several notions of strong identifiability and convergence behaviors for parameters of multiple types, including matrix-variate ones, that arise in finite mixtures, and the effects of model fitting with extra mixing components. .

Chapter III: Convergence rates of parameter estimation for some weakly identifiable finite mixtures This chapter establishes minimax lower bounds and maximum likelihood convergence rates of parameter estimation for weakly identifiable models, including mean-covariance multivariate Gaussian mixtures, shape-rate Gamma mixtures, and some variants of finite mixture models.

Chapter IV: Convergence rates of parameter estimation for some weakly identifiable finite mixtures This chapter proposes a general framework for the identification of singularity structures of the parameter space of finite mixtures, and studies the impacts of the singularity levels on minimax lower bounds and rates of convergence for the maximum likelihood estimator over a compact parameter space.

Chapter V: Robust estimation of mixing measures in finite mixture models This chapter proposes simple and efficient robust estimators of the mixing measures in finite mixture models, which are inspired by the combination of minimum Hellinger distance estimators, model selection criteria, and the superefficiency phenomenon.

Chapter VI: Multilevel clustering via Wasserstein means This chapter introduces novel joint optimization approaches to the problem of multilevel clustering, which aim to simultaneously partition data in each group and discover grouping patterns among groups in a potentially large hierarchically structured corpus of data.

Chapter VII: Conclusions and suggestions This chapter summarizes the main contributions of the thesis and proposes several directions for future research.

CHAPTER II

On strong identifiability and convergence rates of parameter estimation in finite mixtures

This chapter studies identifiability and convergence behaviors for parameters of multiple types, including matrix-variate ones, that arise in finite mixtures, and the effects of model fitting with extra mixing components. We consider several notions of strong identifiability in a matrix-variate setting, and use them to establish sharp inequalities relating the distance of mixture densities to the Wasserstein distances of the corresponding mixing measures. Characterization of identifiability is given for a broad range of mixture models commonly employed in practice, including location-covariance mixtures and location-covariance-shape mixtures, for mixtures of symmetric densities, as well as some asymmetric ones. Minimax lower bounds and rates of convergence for the maximum likelihood estimates are established for such classes, which are also confirmed by simulation studies.¹

2.1 Introduction

Mixture models are a popular modeling tool for making inference about heterogeneous data [Lindsay, 1995, McLachlan and Basford, 1988]. Under mixture modeling,

¹This work has been published in [Ho and Nguyen, 2016c].

data are viewed as samples from a collection of unobserved or latent subpopulations, each positing its own distribution and associated parameters. Learning about subpopulation-specific parameters is essential to the understanding of the underlying heterogeneity. Theoretical issues related to parameter estimation in mixture models, however, remain poorly understood — as noted in a recent textbook [DasGupta, 2008] (pg. 571), “mixture models are riddled with difficulties such as nonidentifiability”.

Research about parameter identifiability for mixture models goes back to the early work of Teicher [1961, 1963], Yakowitz and Spragins [1968] and others, and continues to attract much interest [Hall and Zhou, 2003, Hall et al., 2005, Elmore et al., 2005, Allman et al., 2009]. To address parameter estimation rates, a natural approach is to study the behavior of mixing distributions that arise in the mixture models. This approach is well-developed in the context of nonparametric deconvolution [Carroll and Hall, 1988, Zhang, 1990, Fan, 1991], but these results are confined to only a specific type of model — location mixtures. Beyond location mixtures there have been far fewer results. In particular, for finite mixture models, a notable contribution was made by Chen, who proposed a notion of strong identifiability and established the convergence of the mixing distribution for a class of over-fitted finite mixtures with scalar parameters [Chen, 1995]. Over-fitted finite mixtures, as opposed to exact-fitted ones, are mixtures that allow extra mixing components in their model specification, when the actual number of mixing components is bounded by a known constant. More recently, Nguyen showed that the convergence of the mixing distribution is naturally understood in terms of Wasserstein distance metric [Nguyen, 2013]. He established rates of convergence of mixing distributions for a number of finite and infinite mixture models with multi-dimensional parameters — the case of finite mixtures can be viewed as a generalization of Chen’s results. Rousseau and Mengersen studied over-fitted mixtures in a Bayesian estimation setting [Rousseau and Mengersen, 2011]. They did not study the convergence of all mixing parameters, focusing only on the

mixing probabilities associated with extra mixing components. Finally, we mention a related literature in computer science, which focuses almost exclusively on the analysis of computationally efficient procedures for clustering with exact-fitted Gaussian mixtures (e.g., [[Dasgupta, 1999](#), [Belkin and Sinha, 2010](#), [Kalai et al., 2012](#)]).

Setting The goal of this chapter is to establish rates of convergence for parameters of multiple types, including matrix-variate parameters, that arise in a variety of finite mixture models. Assume that each subpopulation is distributed according to a density function (with respect to Lebesgue measure on an Euclidean space \mathcal{X}) that belongs to a known density class $\{f(x|\theta, \Sigma), \theta \in \Theta \subset \mathbb{R}^{d_1}, \Sigma \in \Omega \subset S_{d_2}^{++}, x \in \mathcal{X}\}$. Here, $d_1 \geq 1, d_2 \geq 0$, $S_{d_2}^{++}$ is the set of all $d_2 \times d_2$ symmetric positive definite matrices. A finite mixture density with k mixing components can be defined in terms of f and a discrete mixing measure $G = \sum_{i=1}^k p_i \delta_{(\theta_i, \Sigma_i)}$ with k support points as follows

$$p_G(x) = \int f(x|\theta, \Sigma) dG(\theta, \Sigma) = \sum_{i=1}^k p_i f(x|\theta_i, \Sigma_i).$$

Examples for f studied in this chapter include the location-covariance family (when $d_1 = d_2 \geq 1$) under Gaussian or some elliptical families of distributions, the location-covariance-shape family (when $d_1 > d_2$) under the generalized multivariate Gaussian, skew-Gaussian or the exponentially modified Student's t-distribution, and the location-rate-shape family (when $d_1 = 3, d_2 = 0$) under Gamma or other distributions.

As shown by [Nguyen \[2013\]](#), the convergence of mixture model parameters can be measured in terms of a Wasserstein distance on the space of mixing measures G . Let $G = \sum_{i=1}^k p_i \delta_{(\theta_i, \Sigma_i)}$ and $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, \Sigma_i^0)}$ be two discrete probability measures on $\Theta \times \Omega$, which is equipped with metric ρ . Recall the Wasserstein distance of order r ,

for a given $r \geq 1$ (cf. Villani [2009])

$$W_r(G, G_0) = \left(\inf_{\mathbf{q}} \sum_{i,j} q_{ij} \rho^r((\theta_i, \Sigma_i), (\theta_j^0, \Sigma_j^0)) \right)^{1/r},$$

where the infimum is taken over all joint probability distributions \mathbf{q} on $[1, \dots, k] \times [1, \dots, k_0]$ such that, when expressing \mathbf{q} as a $k \times k_0$ matrix, the marginal constraints hold: $\sum_j q_{ij} = p_i$ and $\sum_i q_{ij} = p_j^0$.

To see how convergence of mixing measure G_n in Wasserstein distances is translated to convergence of G_n 's atoms and probability masses, suppose that a sequence of mixing measures $G_n \rightarrow G_0$ under W_r metric at a rate $\omega_n = o(1)$. If all G_n have the same number of atoms $k = k_0$ as that of G_0 , then the set of atoms of G_n converge to the k_0 atoms of G_0 at the same rate ω_n under ρ metric. If G_n have varying $k_n \in [k_0, k]$ number of atoms, where k is a fixed upper bound, then a subsequence of G_n can be constructed so that each atom of G_0 is a limit point of a certain subset of atoms of G_n — the convergence to each such limit also happens at rate ω_n . Some atoms of G_n may have limit points that are not among G_0 's atoms — the mass associated with those atoms of G_n must vanish at the generally faster rate ω_n^r (since $r \geq 1$).

In order to establish the rates of convergence for the mixing measure G , our strategy is to derive sharp bounds which relate the Wasserstein distance of mixing measures G, G' and a distance between corresponding mixture densities $p_G, p_{G'}$, such as the variational distance $V(p_G, p_{G'})$. It is relatively simple to obtain upper bounds for the variational distance of mixing densities (V for short) in terms of the Wasserstein distances $W_r(G, G')$ (shorthanded by W_r). Establishing (sharp) lower bounds for V in terms of W_r is the main challenge. Such bounds may not hold, due to a possible lack of identifiability of the mixing measures: one may have $p_G = p_{G'}$, so clearly $V = 0$ but $G \neq G'$, so that $W_r \neq 0$.

General theory of strong identifiability The classical identifiability condition requires that $p_G = p_{G'}$ entail $G = G'$. This amounts to the linear independence of elements f in the density class [Teicher, 1963]. In order to establish quantitative lower bounds on a distance of mixture densities, we employ several notions of strong identifiability, extending from the definitions employed in Chen [1995] and Nguyen [2013] to handle multiple parameter types, including matrix-variate parameters. There are two kinds of strong identifiability. One such notion involves taking the first-order derivatives of function f with respect to all parameters in the model, and insisting that these quantities be linearly independent in a sense to be precisely defined. This criterion will be called “strong identifiability in the first order”, or simply first-order identifiability. When the second-order derivatives are also involved, we obtain the second-order identifiability criterion. It is worth noting that prior studies on parameter estimation rates tend to center primarily on the second-order identifiability condition or something even stronger [Chen, 1995, Liu and Shao, 2004, Rousseau and Mengerson, 2011, Nguyen, 2013]. We show that for exact-fitted mixtures, the first-order identifiability condition (along with additional and mild regularity conditions) suffices for obtaining that

$$V(p_G, p_{G_0}) \gtrsim W_1(G, G_0), \quad (2.1)$$

when $W_1(G, G_0)$ is sufficiently small. Moreover, for a broad range of density classes, we also have $V \lesssim W_1$, for which we actually obtain $V(p_G, p_{G_0}) \asymp W_1(G, G_0)$. A consequence of this fact is that for any estimation procedure that admits the $n^{-1/2}$ convergence rate for the mixture density under V distance, the mixture model parameters also converge at the same rate under Euclidean metric.

Turning to the over-fitted setting, second-order identifiability along with mild regularity conditions would be sufficient for establishing that for any G that has *at*

most k support points where $k \geq k_0 + 1$ and k is fixed,

$$V(p_G, p_{G_0}) \gtrsim W_2^2(G, G_0). \quad (2.2)$$

when $W_2(G, G_0)$ is sufficiently small. The lower bound $W_2^2(G, G_0)$ is sharp, i.e., we cannot improve the lower bound to W_1^r for any $r < 2$ (notably, $W_2 \geq W_1$). A consequence of this result is, take any standard estimation method (such as the MLE) which yields the $n^{-1/2}$ convergence rate for p_G , the induced rate of convergence for the mixing measure G is $n^{-1/4}$ under W_2 . This means the mixing probability mass converge at $n^{-1/2}$ rate (which recovers the result of [Rousseau and Mengersen \[2011\]](#)), in addition to having that the component parameters converge at $n^{-1/4}$ rate.

We also show that there is a range of mixture models with varying parameters of multiple types that satisfies the developed strong identifiability criteria. All such models exhibit the same kinds of rate for parameter estimation. In particular, the second-order identifiability criterion (thus the first-order identifiability) is satisfied by many density families f including the multivariate Student's t-distribution, the exponentially modified multivariate Student's t-distribution. Second-order identifiability also holds for several mixture models with multiple types of (scalar) parameters. These results are presented in Section 2.3.2.

Convergence of MLE estimators and minimax lower bounds Assuming that n -iid sample X_1, \dots, X_n are generated according to p_{G_0} and let \hat{G}_n be the MLE estimate of the mixing distribution G ranging among all discrete probability distributions with at most k support points in $\Theta \times \Omega$ under the over-fitted setting or among all discrete probability distributions with exactly k_0 support points in $\Theta \times \Omega$ under the exact-fitted setting. The inequalities (2.1) and (2.2), along with the fact that these bounds are sharp enable us to easily establish the convergence rates of the mixing measures, and obtain minimax lower bounds. Such results are stated in Theorem

[2.4.2](#), Theorem [2.4.3](#), and Theorem [2.4.4](#). In particular, we obtain the minimax lower bound $n^{-1/\delta}$ under W_1 distance for the exact-fitted setting for any positive $\delta < 2$. Under the over-fitted setting, the minimax lower bound is $n^{-1/\delta}$ under W_2 distance for any positive $\delta < 4$. The MLE method can be shown to achieve both these rates, i.e., $n^{-1/2}$ and $n^{-1/4}$ up to a logarithm term, under exact-fitted and over-fitted setting, respectively. Note, however, that these are pointwise convergence rates, i.e., the constants C_1 in Theorem [2.4.2](#) and Theorem [2.4.3](#) depend on G_0 . For a fixed G_0 , we think that the MLE upper bound $n^{-1/4}$ for the over-fitted setting is tight, but we do not have a proof.

Summarizing, the technical contributions of this chapter include the following:

- (i) Convergence of parameters of multiple types, including matrix-variate parameters, for finite mixtures, under strong identifiability conditions.
- (ii) A minimax lower bound, in the sense of Wasserstein distance W_2 , for estimating mixing measures in an over-fitted setting. The maximum likelihood estimation method is shown to achieve this lower bound, up to a logarithmic term, although the convergence is pointwise.
- (iii) Characterization results showing the applicability of our theory and the convergence rates to a fairly broad range of mixture models with parameters of multiple types, including matrix-variate ones.
- (iv) Another noteworthy aspect of this work is that the settings of exact-fitted and over-fitted mixtures are treated separately: the first-order identifiability criterion is sufficient in the former setting, which attains convergence in W_1 ; while the second-order identifiability criterion is sufficient in the latter, which achieves convergence in W_2 metric.

Finally, we note in passing that both the first and second-order identifiability are in some sense *necessary* in deriving the MLE convergence rate $n^{-1/2}$ and $n^{-1/4}$ as de-

scribed above. Models such as location-scale Gaussian mixtures, shape-scale Gamma mixtures and location-scale-shape skew-Gaussian mixtures do not satisfy either or both strong identifiability conditions — we call such models “weakly identifiable”. It can be shown that such weakly identifiable models exhibit a much slower convergence behavior than the standard rates established in this chapter. Such a theory is fundamentally different from the strong identifiability theory, and will be reported elsewhere.

Chapter organization The rest of the chapter is organized as follows. Section 2.2 provides some preliminary backgrounds and facts. Section 2.3 presents a general theory of strong identifiability, by addressing the exact-fitted and over-fitted settings separately before providing a characterization of density classes for which the general theories are applicable. Section 2.4.1 contains consequences of the theory developed earlier – this includes minimax lower bounds and convergence rates of maximum likelihood estimation. The theoretical bounds are illustrated via simulations in Section 2.4.2. Self-contained proofs of the key theorems are given in Section 2.5 while proofs of the remaining results are presented in the Section 2.6.

Notation Given two densities p, q (with respect to Lebesgue measure μ), the variational distance is given by $V(p, q) = (1/2) \int |p - q| d\mu$. The Hellinger distance h is given by $h^2(p, q) = (1/2) \int (p^{1/2} - q^{1/2})^2 d\mu$.

As $K, L \in \mathbb{N}$, the first derivative of real function $g : \mathbb{R}^{K \times L} \rightarrow \mathbb{R}$ of matrix Σ is defined as a $K \times L$ matrix whose (i, j) element is $\partial g / \partial \Sigma_{ij}$. The second derivative of g , denoted by $\frac{\partial^2 g}{\partial \Sigma^2}$ is a $K^2 \times L^2$ matrix made of KL blocks of $K \times L$ matrix, whose (i, j) -block is given by $\frac{\partial}{\partial \Sigma} \left(\frac{\partial g}{\partial \Sigma_{ij}} \right)$. Additionally, as $N \in \mathbb{N}$, for function $g_2 : \mathbb{R}^N \times \mathbb{R}^{K \times L} \rightarrow \mathbb{R}$ defined on (θ, Σ) , the joint derivative between the vector component and matrix component $\frac{\partial^2 g_2}{\partial \theta \partial \Sigma} = \frac{\partial^2 g_2}{\partial \Sigma \partial \theta}$ is a $(KN) \times L$ matrix of KL

blocks for N -columns, whose (i, j) -block is given by $\frac{\partial}{\partial \theta} \left(\frac{\partial g_2}{\partial \Sigma_{ij}} \right)$.

Throughout this chapter, for any symmetric matrix $\Sigma \in \mathbb{R}^{d \times d}$, $\lambda_1(\Sigma)$ and $\lambda_d(\Sigma)$ respectively denote its smallest and largest eigenvalue. Additionally, the expression "≥" will be used to denote the inequality up to a constant multiple where the value of the constant is fixed within our setting. We write $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold.

2.2 Preliminaries

First of all, we need to define our notion of distances on the space of mixing measures. In this chapter, we restrict ourselves to the space of discrete mixing measures with exactly k_0 distinct support points on $\Theta \times \Omega$, denoted by $\mathcal{E}_{k_0}(\Theta \times \Omega)$, and the space of discrete mixing measures with at most k distinct support points on $\Theta \times \Omega$, denoted by $\mathcal{O}_k(\Theta \times \Omega)$. Consider a mixing measure $G = \sum_{i=1}^k p_i \delta_{(\theta_i, \Sigma_i)}$, where $\mathbf{p} = (p_1, p_2, \dots, p_k)$ denotes the proportion vector. Likewise, let $G' = \sum_{i=1}^{k'} p'_i \delta_{(\theta'_i, \Sigma'_i)}$. A coupling between \mathbf{p} and \mathbf{p}' is a joint distribution \mathbf{q} on $[1, \dots, k] \times [1, \dots, k']$, which is expressed as a matrix $\mathbf{q} = (q_{ij})_{1 \leq i \leq k, 1 \leq j \leq k'} \in [0, 1]^{k \times k'}$ and admits marginal constraints $\sum_{i=1}^k q_{ij} = p'_j$ and $\sum_{j=1}^{k'} q_{ij} = p_i$ for any $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, k'$. We call \mathbf{q} a coupling of \mathbf{p} and \mathbf{p}' , and use $\mathcal{Q}(\mathbf{p}, \mathbf{p}')$ to denote the space of all such couplings.

As in [Nguyen \[2013\]](#), our tool for analyzing the identifiability and convergence of parameters in a mixture model is by adopting Wasserstein distances, which can be defined as the optimal cost of moving masses from one probability measure to another [\[Villani, 2009\]](#). For any $r \geq 1$, the r -th order Wasserstein distance between G and G' is given by

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

In both occurrences in the above display, $\|\cdot\|$ denotes either the l_2 norm for elements in \mathbb{R}^d or the entrywise l_2 norm for matrices.

The central theme of the chapter is the relationship between the Wasserstein distances of mixing measures G, G' and the distances of the corresponding mixture densities $p_G, p_{G'}$. Clearly, if $G = G'$ then $p_G = p_{G'}$. Intuitively, if $W_1(G, G')$ or $W_2(G, G')$ is small, so is a distance between p_G and $p_{G'}$. This can be quantified by establishing an upper bound for the distance of p_G and $p_{G'}$ in terms of $W_1(G, G')$ or $W_2(G, G')$. There is a simple and general way to do this, by accounting for the Lipschitz property of the density class and then appealing to Jensen's inequality. We will not go into such details and refer the readers to [Nguyen \[2013\]](#) (Section 2). The followings are examples of mixture models that carry multiple types of parameter including the matrix-variate ones, along with the aforementioned upper bounds.

Example 2.2.1. (Multivariate generalized Gaussian distribution [[Zhang et al., 2013](#)])
Let $f(x|\theta, m, \Sigma) = \frac{m\Gamma(d/2)}{\pi^{d/2}\Gamma(d/(2m))|\Sigma|^{1/2}} \exp(-[(x - \theta)^T \Sigma^{-1} (x - \theta)]^m)$, where $\theta \in \mathbb{R}^d, m > 0$, and $\Sigma \in S_d^{++}$. If Θ_1 is a bounded subset of \mathbb{R}^d , $\Theta_2 = \{m \in \mathbb{R}^+ : 1 \leq \underline{m} \leq m \leq \bar{m}\}$, and $\Omega = \{\Sigma \in S_d^{++} : \underline{\lambda} \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq \bar{\lambda}\}$, where $\underline{\lambda}, \bar{\lambda} > 0$, then for any mixing measures G_1, G_2 , we obtain $h^2(p_{G_1}, p_{G_2}) \lesssim W_2^2(G_1, G_2)$ and $V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2)$.

Example 2.2.2. (Multivariate Student's t-distribution [[Peel and McLachlan, 2000](#)])
Let $f(x|\theta, \Sigma) = \frac{C_\nu}{|\Sigma|^{1/2}} (\nu + (x - \theta)^T \Sigma^{-1} (x - \theta))^{-(\nu+d)/2}$, where ν is a fixed positive degree of freedom and $C_\nu = \frac{\Gamma((\nu+d)/2)\nu^{\nu/2}}{\Gamma(\nu/2)\pi^{d/2}}$. If Θ is a bounded subset of \mathbb{R}^d and $\Omega = \{\Sigma \in S_d^{++} : \underline{\lambda} \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq \bar{\lambda}\}$, then for any mixing measures G_1, G_2 , we obtain $h^2(p_{G_1}, p_{G_2}) \lesssim W_2^2(G_1, G_2)$ and $V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2)$.

Example 2.2.3. (Exponentially modified multivariate Student's t-distribution)

Let $f(x|\theta, \lambda, \Sigma)$ be the density of $X = Y + Z$, where Y follows the multivariate t-distribution with location parameter θ , covariance matrix Σ , fixed positive degree of

freedom ν , and Z is distributed by the product of d independent exponential distributions with combined shape parameter $\lambda = (\lambda_1, \dots, \lambda_d)$. If Θ is a bounded subset of $\mathbb{R}^d \times \mathbb{R}_+^d$ and $\Omega = \left\{ \Sigma \in S_d^{++} : \underline{\lambda} \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq \bar{\lambda} \right\}$, then for any mixing measures G_1, G_2 , we have $h^2(p_{G_1}, p_{G_2}) \lesssim W_2^2(G_1, G_2)$ and $V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2)$.

Example 2.2.4. (Modified Gaussian-Gamma distribution)

Let $f(x|\theta, \alpha, \beta, \Sigma)$ be the density function of $X = Y + Z$, where Y is distributed by the multivariate Gaussian distribution with mean θ , covariance matrix Σ , and Z is distributed by the product of independent Gamma distributions with combined shape vector $\alpha = (\alpha_1, \dots, \alpha_d)$ and combined rate vector $\beta = (\beta_1, \dots, \beta_d)$. If Θ is a bounded subset of $\mathbb{R}^d \times \mathbb{R}_+^d \times \mathbb{R}_+^d$, $\Omega = \left\{ \Sigma \in S_d^{++} : \underline{\lambda} \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq \bar{\lambda} \right\}$, then for any mixing measures G_1, G_2 , we have $h^2(p_{G_1}, p_{G_2}) \lesssim V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2)$.

2.3 General theory of strong identifiability

The objective of this section is to develop a general theory according to which a small distance between mixture densities p_G and $p_{G'}$ entails a small Wasserstein distance between mixing measures G and G' . The classical identifiability criteria require that $p_G = p_{G'}$ entail $G = G'$, which is essentially equivalent to a linear independence requirement for the class of density family $\{f(x|\theta, \Sigma) | \theta \in \Theta, \Sigma \in \Omega\}$. To obtain quantitative bounds, we shall need stronger notions of identifiability, ones which involve higher order derivatives of the density function f , taken with respect to the mixture model parameters. The strength of this theory is that it holds generally for a fairly broad range of mixture models, which allows for the same bounds on the Wasserstein distances. This in turn leads to “standard” rates of convergence for mixing measures.

2.3.1 Definitions and general identifiability bounds

Definition 2.3.1. *The family $\{f(x|\theta, \Sigma), \theta \in \Theta, \Sigma \in \Omega\}$ is **identifiable in the first-order** if $f(x|\theta, \Sigma)$ is differentiable in (θ, Σ) and the following holds*

A1. *For any finite k different pairs $(\theta_1, \Sigma_1), \dots, (\theta_k, \Sigma_k) \in \Theta \times \Omega$, if we have $\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^{d_1}$ and **symmetric matrices** $\gamma_i \in \mathbb{R}^{d_2 \times d_2}$ (for all $i = 1, \dots, k$) such that for almost all x*

$$\sum_{i=1}^k \alpha_i f(x|\theta_i, \Sigma_i) + \beta_i^T \frac{\partial f}{\partial \theta}(x|\theta_i, \Sigma_i) + \text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i, \Sigma_i)^T \gamma_i \right) = 0,$$

then, $\alpha_i = 0, \beta_i = \mathbf{0} \in \mathbb{R}^{d_1}, \gamma_i = \mathbf{0} \in \mathbb{R}^{d_2 \times d_2}$ for $i = 1, \dots, k$.

Remark The condition that γ_i is symmetric in Definition 2.3.1 is crucial, without which the identifiability condition would fail for many classes of density. Indeed, assume that $\frac{\partial f}{\partial \Sigma}(x|\theta_i, \Sigma_i)$ are symmetric matrices for all i , which clearly holds for elliptical distributions such as multivariate Gaussian, Student's t-distribution, and logistics distribution. Now, if we choose γ_i to be anti-symmetric matrices with zero diagonal elements, $\alpha_i = 0, \beta_i = \mathbf{0}$, then the equation in (A1) holds even when $\gamma_i \neq \mathbf{0}$ for all i .

Additionally, we say the family of densities f is uniformly Lipschitz up to the first order if the following holds: there are positive constants δ_1, δ_2 such that for any $R_1, R_2, R_3 > 0, \gamma_1 \in \mathbb{R}^{d_1}, \gamma_2 \in \mathbb{R}^{d_2 \times d_2}, R_1 \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_{d_2}(\Sigma)} \leq R_2, \|\theta\| \leq R_3, \theta_1, \theta_2 \in \Theta, \Sigma_1, \Sigma_2 \in \Omega$, there are positive constants $C(R_1, R_2)$ and $C(R_3)$ such that for all $x \in \mathcal{X}$

$$\left| \gamma_1^T \left(\frac{\partial f}{\partial \theta}(x|\theta_1, \Sigma) - \frac{\partial f}{\partial \theta}(x|\theta_2, \Sigma) \right) \right| \leq C(R_1, R_2) \|\theta_1 - \theta_2\|^{\delta_1} \|\gamma_1\|, \quad (2.3)$$

$$\left| \text{tr} \left(\left(\frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma_1) - \frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma_2) \right)^T \gamma_2 \right) \right| \leq C(R_3) \|\Sigma_1 - \Sigma_2\|^{\delta_2} \|\gamma_2\|. \quad (2.4)$$

First-order identifiability is sufficient for deriving a lower bound of $V(p_G, p_{G_0})$ in terms of $W_1(G, G_0)$, under the *exact-fitted* setting: This is the setting where G_0 has exactly k_0 support points lying in the interior of $\Theta \times \Omega$:

Theorem 2.3.1. (Exact-fitted setting) *Suppose that the density family f is identifiable in the first order and admits a uniform Lipschitz property up to the first order. Then there are positive constants ϵ_0 and C_0 , both depending on G_0 , such that as long as $G \in \mathcal{E}_{k_0}(\Theta \times \Omega)$, the set of mixing measures with exact order k_0 , and $W_1(G, G_0) \leq \epsilon_0$, we have*

$$V(p_G, p_{G_0}) \geq C_0 W_1(G, G_0).$$

Although no boundedness condition on Θ or Ω is required, this lower bound is of a local nature, in the sense that it holds only for those G sufficiently close to G_0 by a Wasserstein distance at most ϵ_0 , which again varies with G_0 . It is possible to extend this type of bound to hold globally over a compact subset of the space of mixing measures, under a mild regularity condition, as the following corollary asserts:

Corollary 2.3.1. *Suppose that all assumptions of Theorem 2.3.1 hold. Furthermore, there is a positive constant $\alpha > 0$ such that for any $G_1, G_2 \in \mathcal{E}_{k_0}(\Theta \times \Omega)$, we have $V(p_{G_1}, p_{G_2}) \lesssim W_1^\alpha(G_1, G_2)$. Then, for a fixed compact subset \mathcal{G} of $\mathcal{E}_{k_0}(\Theta \times \Omega)$, there is a positive constant $C_0 = C_0(G_0)$ such that*

$$V(p_G, p_{G_0}) \geq C_0 W_1(G, G_0) \quad \text{for all } G \in \mathcal{G}.$$

We shall verify in the sequel that the classes of densities f described in Examples 2.2.1, 2.2.2, and 2.2.3 are all identifiable in the first order. Combining with the upper bounds for V , we arrive at a remarkable fact for these density classes, that

$$V(p_G, p_{G_0}) \asymp W_1(G, G_0).$$

Moving to the *over-fitted* setting, where G_0 has exactly k_0 support points lying in the interior of $\Theta \times \Omega$, but k_0 is unknown and only an upper bound for k_0 is given, a stronger identifiability condition is required. This condition generalizes that of [Chen \[1995\]](#):

Definition 2.3.2. *The family $\{f(x|\theta, \Sigma), \theta \in \Theta, \Sigma \in \Omega\}$ is **identifiable in the second-order** if $f(x|\theta, \Sigma)$ is twice differentiable in (θ, Σ) and the following assumption holds*

A2. *For any finite k different pairs $(\theta_1, \Sigma_1), \dots, (\theta_k, \Sigma_k) \in \Theta \times \Omega$, if we have $\alpha_i \in \mathbb{R}, \beta_i, \nu_i \in \mathbb{R}^{d_1}, \gamma_i, \eta_i$ **symmetric matrices** in $\mathbb{R}^{d_2 \times d_2}$ as $i = 1, \dots, k$ such that for almost all x*

$$\begin{aligned} & \sum_{i=1}^k \left\{ \alpha_i f(x|\theta_i, \Sigma_i) + \beta_i^T \frac{\partial f}{\partial \theta}(x|\theta_i, \Sigma_i) + \nu_i^T \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i, \Sigma_i) \nu_i \right. + \\ & \quad \text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i, \Sigma_i)^T \gamma_i \right) + 2 \nu_i^T \left[\frac{\partial}{\partial \theta} \left(\text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i, \Sigma_i)^T \eta_i \right) \right) \right] + \\ & \quad \left. \text{tr} \left(\frac{\partial}{\partial \Sigma} \left(\text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i, \Sigma_i)^T \eta_i \right) \right)^T \eta_i \right) \right\} = 0, \end{aligned}$$

then, $\alpha_i = 0, \beta_i = \nu_i = \mathbf{0} \in \mathbb{R}^{d_1}, \gamma_i = \eta_i = \mathbf{0} \in \mathbb{R}^{d_2 \times d_2}$ for $i = 1, \dots, k$.

In addition, we say the family of densities f is uniformly Lipschitz up to the second order if the following holds: there are positive constants δ_3, δ_4 such that for any $R_4, R_5, R_6 > 0, \gamma_1 \in \mathbb{R}^{d_1}, \gamma_2 \in \mathbb{R}^{d_2 \times d_2}, R_4 \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_{d_2}(\Sigma)} \leq R_5, \|\theta\| \leq R_6, \theta_1, \theta_2 \in \Theta, \Sigma_1, \Sigma_2 \in \Omega$, there are positive constants C_1 depending on (R_4, R_5) and C_2 depending on R_6 such that for all $x \in \mathcal{X}$

$$|\gamma_1^T \left(\frac{\partial^2 f}{\partial \theta^2}(x|\theta_1, \Sigma) - \frac{\partial^2 f}{\partial \theta^2}(x|\theta_2, \Sigma) \right) \gamma_1| \leq C_1 \|\theta_1 - \theta_2\|_1^{\delta_3} \|\gamma_1\|_2^2,$$

$$\left| \text{tr} \left(\left[\frac{\partial}{\partial \Sigma} \left(\text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma_1)^T \gamma_2 \right) \right) - \frac{\partial}{\partial \Sigma} \left(\text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma_2)^T \gamma_2 \right) \right) \right]^T \gamma_2 \right) \right| \leq C_2 \|\Sigma_1 - \Sigma_2\|_2^{\delta_4} \|\gamma_2\|_2^2.$$

Let $k \geq 2$ and $k_0 \geq 1$ be fixed positive integers where $k \geq k_0 + 1$. $G_0 \in \mathcal{E}_{k_0}$ while G varies in \mathcal{O}_k . Then, we can establish the following results

Theorem 2.3.2. (Over-fitted setting)

(a) Assume the density family f is identifiable in the second order and admits a uniform Lipschitz property up to the second order. Moreover, Θ is a bounded subset of \mathbb{R}^{d_1} and Ω is a subset of $S_{d_2}^{++}$ such that the largest eigenvalues of elements of Ω are bounded above. Additionally, assume that for each $\theta \in \Theta$, for each $x \in \mathcal{X}$ except a finite number of values in \mathcal{X} , we have $\lim_{\lambda_1(\Sigma) \rightarrow 0} f(x|\theta, \Sigma) = 0$. Then there are positive constants ϵ_0 and C_0 depending on G_0 such that as long as $G \in \mathcal{O}_k(\Theta \times \Omega)$, the set of mixing measures with their orders bounded above by k , and $W_2(G, G_0) \leq \epsilon_0$, we have

$$V(p_G, p_{G_0}) \geq C_0 W_2^2(G, G_0).$$

(b) (Optimality of bound for variational distance) Assume f is second-order differentiable with respect to θ, Σ and all of its second derivatives are integrable uniformly for all θ, Σ . Then, for any $1 \leq r < 2$:

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_k(\Theta \times \Omega)} \left\{ V(p_G, p_{G_0}) / W_1^r(G, G_0) : W_1(G, G_0) \leq \epsilon \right\} = 0.$$

(c) (Optimality of bound for Hellinger distance) Assume f is second-order differ-

entiable with respect to θ , Σ and for some sufficiently small $c_0 > 0$,

$$\sup_{\|\theta - \theta'\| + \|\Sigma - \Sigma'\| \leq c_0} \int_{x \in \mathcal{X}} \left(\frac{\partial^2 f}{\partial \theta^{\alpha_1} \partial \Sigma^{\alpha_2}}(x|\theta, \Sigma) \right)^2 / f(x|\theta', \Sigma') dx < \infty$$

where $\alpha_1 \in \mathbb{N}^{d_1}$, $\alpha_2 \in \mathbb{N}^{d_2 \times d_2}$ in the partial derivative of f take any combination such that $|\alpha_1| + |\alpha_2| = 2$. Then, for any $1 \leq r < 2$:

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{C}_k(\Theta \times \Omega)} \left\{ h(p_G, p_{G_0}) / W_1^r(G, G_0) : W_1(G, G_0) \leq \epsilon \right\} = 0.$$

Here and elsewhere, ratio V/W_r is set to be ∞ if $W_r(G, G_0) = 0$. Some remarks:

- (i) A version of part (a) for finite mixtures with multivariate parameters was first given in [Nguyen \[2013\]](#) (Proposition 1 and Theorem 1). The original statement of Nguyen's Theorem 1 contains a mistake, as it claims something unnecessarily stronger: $V(p_{G_1}, p_{G_2})/W_2^2(G_1, G_2)$ is bounded away from 0 as both G_1 and G_2 are sufficiently close to G_0 in the sense of W_2 . This is not true, unless both G_1 and G_2 have the same number of support points as G_0 . ² This error can be corrected in the overfitted setting, by fixing G_2 to G_0 , and allowing only $G_1 \equiv G$ to vary near G_0 . This is precisely our current statement of part (a) stated for the more general matrix-variate mixture models.
- (ii) The condition $\lim_{\lambda_1(\Sigma) \rightarrow 0} f(x|\theta, \Sigma) = 0$ is important for the matrix-variate parameter Σ . In particular, it is useful for addressing the scenario when the smallest eigenvalue of matrix parameter Σ is not bounded away from 0. It is simple to see that this condition is satisfied for the examples discussed in the previous section. For instance, for the multivariate generalized Gaussian distribution, it holds for each $\theta \in \Theta$, and $x \neq \theta$. Note also that this condition can be removed

²A counterexample was pointed out to the second author by Elisabeth Gassiat, who attributed it to Jonas Kahn. A similar error is also present in Lemma 2 of [Chen \[1995\]](#), which admits the same correction described above.

if we additionally impose that all $\Sigma \in \Omega$ are positive definite matrices whose eigenvalues are bounded away from 0.

- (iii) Part (b) demonstrates the sharpness of the bound in part (a). In particular, we cannot improve the lower bound in part (a) to any quantity $W_1^r(G, G_0)$ for any $r < 2$. For any estimation method that yields $n^{-1/2}$ convergence rate under the Hellinger distance for p_G , part (a) induces $n^{-1/4}$ convergence rate under W_2 for G . A consequence of part (c) is a minimax lower bound $n^{-1/4}$ for estimating the mixing measure G . See Section 2.4.1 for formal statements of such results.
- (iv) It is also worth mentioning that the boundedness of Θ , as well as the boundedness from above of the eigenvalues of elements of Ω are both necessary conditions for the conclusion of part (a) to hold. Indeed, it is possible to show that if one of these two conditions is not met, we are not able to obtain the lower bound of $V(p_G, p_{G_0})$ as established, because the distance $h \geq V$ can vanish much faster than the distance $W_r(G, G_0)$, as can be seen by:

Proposition 2.3.1. *Let Θ be a subset of \mathbb{R}^{d_1} and $\Omega = S_{d_2}^{++}$. Then for any $r \geq 1$ and $\beta > 0$ we have*

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_k(\Theta \times \Omega)} \left\{ \exp \left(\frac{1}{W_r^\beta(G, G_0)} \right) h(p_G, p_{G_0}) : W_r(G, G_0) \leq \epsilon \right\} = 0.$$

Finally, as in the exact-fitted setting, to establish the bound $V \gtrsim W_2^2$ in a global manner, we simply add a compactness condition on the subset within which G varies:

Corollary 2.3.2. *Suppose that all assumptions of Theorem 2.3.2 (part (a)) hold. Furthermore, there is a positive constant $\alpha \leq 2$ such that for any $G_1, G_2 \in \mathcal{O}_k(\Theta \times \Omega)$, we have $V(p_{G_1}, p_{G_2}) \lesssim W_2^\alpha(G_1, G_2)$. Then, for a fixed compact subset \mathcal{O} of $\mathcal{O}_k(\Theta \times \Omega)$ there is a positive constant $C_0 = C_0(G_0)$ such that*

$$V(p_G, p_{G_0}) \geq C_0 W_2^2(G, G_0) \quad \text{for all } G \in \mathcal{O}.$$

A consequence of this result is, take any standard estimation method such as the MLE, which yields the $n^{-1/2}$ convergence rate for p_G , the induced rate of convergence for the mixing measure G is $n^{-1/4}$ under W_2 . This also entails that the mixing probability masses converge at the $n^{-1/2}$ rate (which recovers the result of [Rousseau and Mengersen \[2011\]](#)), in addition to having that the component parameters converge at the $n^{-1/4}$ rate.

2.3.2 Characterization of strong identifiability

In this subsection we identify a fairly broad range of density classes for which the strong identifiability conditions developed previously hold either in the first or the second order. Then we also present general results which show how strong identifiability conditions continue to be preserved under certain transformations with respect to the parameter space. First, we consider univariate density functions with parameters of multiple types:

Theorem 2.3.3. (*Densities with multiple scalar parameters*)

- (a) *Generalized univariate logistic densities:* Let $f(x|\theta, \sigma) := \frac{1}{\sigma} f((x - \theta)/\sigma)$, where $f(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} \frac{\exp(px)}{(1 + \exp(x))^{p+q}}$, and p, q are fixed in \mathbb{N}_+ . Then the family $\{f(x|\theta, \sigma), \theta \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$ is identifiable in the second order.
- (b) *Generalized Gumbel densities:* Let $f(x|\theta, \sigma, \lambda) := \frac{1}{\sigma} f((x - \theta)/\sigma, \lambda)$, where $f(x, \lambda) = \frac{\lambda^\lambda}{\Gamma(\lambda)} \exp(-\lambda(x + \exp(-x)))$ as $\lambda > 0$. Then we have the family $\{f(x|\theta, \sigma, \lambda), \theta \in \mathbb{R}, \sigma \in \mathbb{R}_+, \lambda \in \mathbb{R}_+\}$ is identifiable in the second order.
- (c) *Weibull densities:* Let $f(x|\nu, \lambda) = \frac{\nu}{\lambda} \left(\frac{x}{\lambda}\right)^{\nu-1} \exp\left(-\left(\frac{x}{\lambda}\right)^\nu\right)$, for $x \geq 0$, where $\nu, \lambda > 0$ are shape and scale parameters, respectively. Then the family $\{f(x|\nu, \lambda), \nu \in \mathbb{R}_+, \lambda \in \mathbb{R}_+\}$ is identifiable in the second order.
- (d) *Von Mises densities* [[Hsu et al., 1981](#), [Kent, 1983](#), [Mardia, 1975](#)]: Let $f(x|\mu, \kappa) =$

$\frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(x - \mu)).1_{\{x \in [0, 2\pi)\}}$, where $\mu \in [0, 2\pi)$, $\kappa > 0$, and $I_0(\kappa)$ is the modified Bessel function of order 0. Then the family $\{f(x|\mu, \kappa), \mu \in [0, 2\pi), \kappa \in \mathbb{R}_+\}$ is identifiable in the second order.

Next, we turn to the density function classes with matrix-variate parameter spaces, as introduced in Section 2.2:

Theorem 2.3.4. (*Densities with matrix-variate parameters*)

- (a) The family $\{f(x|\theta, \Sigma, m), \theta \in \mathbb{R}^d, \Sigma \in S_d^{++}, m \geq 1\}$ of multivariate generalized Gaussian distribution is identifiable in the first order.
- (b) The family $\{f(x|\theta, \Sigma), \theta \in \mathbb{R}^d, \Sigma \in S_d^{++}\}$ of multivariate t-distribution with fixed odd degree of freedom is identifiable in the second order.
- (c) The family $\{f(x|\theta, \Sigma, \lambda), \theta \in \mathbb{R}^d, \Sigma \in S_d^{++}, \lambda \in \mathbb{R}_+^d\}$ of exponentially modified multivariate t-distribution with fixed odd degree of freedom is identifiable in the second order.
- (d) The family $\{f(x|\theta, \Sigma, a, b), \theta \in \mathbb{R}^d, \Sigma \in S_d^{++}, a \in \mathbb{R}_+^d, b \in \mathbb{R}_+^d\}$ of modified Gaussian-Gamma distribution is not identifiable in the first order.

These theorems are the matrix-variate or multiple parameter-type counterparts of results proven for density classes with a single scalar parameter [Chen, 1995]. As the proofs of these results are technically involved, we present only the proof of Theorem 2.3.4 in the Section 2.6. A useful insight one can draw from these proofs is that the strong identifiability of these density classes are established by exploiting how the corresponding characteristic functions and moment generating functions behave at infinity. Thus it can be concluded that the common feature in establishing strong identifiability hinges on the smoothness of the density f in question.

Some additional details: regarding part (a), as the class of multivariate Gaussian distribution ($m = 1$) is not identifiable in the second order, the conclusion of this

part only holds for the first-order identifiability. However, if we impose the constraint $m > 1$, the class of multivariate generalized Gaussian distributions is identifiable in the second order. The condition *odd degree of freedom* in part (b) and (c) of Theorem 2.3.4 is mainly due to our proof technique. We believe both (b) and (c) hold for any fixed positive degree of freedom, but do not have a proof. Finally, the conclusion of part (d) is due to the fact that family of Gamma distribution is not identifiable in the first order.

The results of Theorem 2.3.4 shed light on which classes of distribution satisfy the inequality being established in Theorem 2.3.1 and Theorem 2.3.2. A consequence is the following: take any standard estimation method (such that the MLE) which yields the $n^{-1/2}$ convergence rate for p_G , the induced rate of convergence for the mixing measure G is $n^{-1/2}$ under W_1 for the exact-fitted setting or $n^{-1/4}$ under W_2 for the over-fitted setting. From the definition of Wasserstein distances, under the MLE, the mixing probabilities' estimate converge at the $n^{-1/2}$ rate; while the component parameters converge at the rate $n^{-1/2}$ for the exact-fitted setting, and $n^{-1/4}$ for the over-fitted setting.

Before ending this section, we state a result in response to a question posed by Xuming He on strong identifiability in transformed parameter spaces. The following theorem asserts that the first-order identifiability with respect to a transformed parameter space is preserved under some regularity conditions of the transformation operator. Let T be a bijective mapping from $\Theta^* \times \Omega^*$ to $\Theta \times \Omega$ such that

$$T(\eta, \Lambda) = (T_1(\eta, \Lambda), T_2(\eta, \Lambda)) = (\theta, \Sigma)$$

for all $(\eta, \Lambda) \in \Theta^* \times \Omega^*$, where $\Theta^* \subset \mathbb{R}^{d_1}$, $\Omega^* \subset S_{d_2}^{++}$. Define the class of density functions $\{g(x|\eta, \Lambda), \eta \in \Theta^*, \Lambda \in \Omega^*\}$ by

$$g(x|\eta, \Lambda) := f(x|T(\eta, \Lambda)).$$

Additionally, for any $(\eta, \Lambda) \in \Theta^* \times \Omega^*$, let $J(\eta, \Lambda) \in \mathbb{R}^{(d_1+d_2^2) \times (d_1+d_2^2)}$ be the modified Jacobian matrix of $T(\eta, \Lambda)$, i.e. the usual Jacobian matrix when (η, Λ) is taken as a $d_1 + d_2^2$ vector.

Theorem 2.3.5. *Assume that $\{f(x|\theta, \Sigma), \theta \in \Theta, \Sigma \in \Omega\}$ is identifiable in the first order. Then the class of density functions $\{g(x|\eta, \Lambda), \eta \in \Theta^*, \Lambda \in \Omega^*\}$ is identifiable in the first order if and only if the modified Jacobian matrix $J(\eta, \Lambda)$ is non-singular for all $(\eta, \Lambda) \in \Theta^* \times \Omega^*$.*

The conclusion of Theorem 2.3.5 still holds if we replace the first-order identifiability by the second-order identifiability. This type of results allows us to construct more examples of strongly identifiable density classes.

As we have seen previously, strong identifiability (either in the first or second order) yields sharp lower bounds of $V(p_G, p_{G_0})$ in terms of Wasserstein distances $W_r(G, G_0)$. It is useful to know that in the transformed parameter space, one may still enjoy the same inequality. Specifically, for any discrete probability measure $Q = \sum_{i=1}^{k_0} p_i \delta_{(\eta_i, \Lambda_i)} \in \mathcal{E}_{k_0}(\Theta^* \times \Omega^*)$, denote

$$p'_Q(x) = \int g(x|\eta, \Lambda) dQ(\eta, \Lambda) = \sum_{i=1}^{k_0} p_i g(x|\eta_i, \Lambda_i).$$

Let Q_0 to be a fixed discrete probability measure on $\mathcal{E}_{k_0}(\Theta^* \times \Omega^*)$, while probability measure Q varies in $\mathcal{E}_{k_0}(\Theta^* \times \Omega^*)$.

Corollary 2.3.3. *Assume that the conditions of Theorem 2.3.5 hold. Further, suppose that the first derivative of f in terms of θ, Σ and the first derivative of T in terms of η, Λ are α -Hölder continuous and bounded where $\alpha > 0$. Then there are positive constants $\epsilon_0 := \epsilon_0(Q_0)$ and $C_0 := C_0(Q_0)$ such that as long as $Q \in \mathcal{E}_{k_0}(\Theta^* \times \Omega^*)$ and $W_1(Q, Q_0) \leq \epsilon_0$, we have*

$$V(p'_Q, p'_{Q_0}) \geq C_0 W_1(Q, Q_0).$$

Remark. If Θ and Ω are bounded sets, the condition on the boundedness of the first derivative of f in terms of θ, Σ and the first derivative of g in terms of η, Λ can be left out. Additionally, the restriction that these derivatives be α -Hölder continuous can be relaxed to only that the first derivative of f and the first derivative of g are α_1 -Hölder continuous and α_2 -Hölder continuous where $\alpha_1, \alpha_2 > 0$ can be different. Finally, the conclusion of Corollary 2.3.3 still holds for the lower bound $W_2^2(Q, Q_0)$ if we impose the second-order identifiability on the kernel density f as well as the additional structures on the second derivative of T .

2.4 Minimax lower bounds, MLE rates and illustrations

2.4.1 Minimax lower bounds and MLE rates of convergence

Given n -iid sample X_1, X_2, \dots, X_n distributed according to the mixture density p_{G_0} , where G_0 is an unknown true mixing distribution with exactly k_0 support points, and the class of densities $\{f(x|\theta, \Sigma), \theta \in \Theta, \Sigma \in \Omega\}$ is assumed known. Given $k \in \mathbb{N}$ such that $k \geq k_0 + 1$. The support points of G_0 lie in $\Theta \times \Omega$. In this section we shall assume that Θ is a compact subset of \mathbb{R}^{d_1} and $\Omega = \{\Sigma \in S_{d_2}^{++} : \underline{\lambda} \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_{d_2}(\Sigma)} \leq \bar{\lambda}\}$, where $0 < \underline{\lambda}, \bar{\lambda}$ are known and $d_1 \geq 1, d_2 \geq 0$. We denote $\Theta^* = \Theta \times \Omega$. The maximum likelihood estimator for G_0 in the over-fitted mixture setting is given by

$$\widehat{G}_n = \arg \max_{G \in \mathcal{O}_k(\Theta \times \Omega)} \sum_{i=1}^n \log(p_G(X_i)).$$

For the exact-fitted mixture setting, \mathcal{O}_k is replaced by \mathcal{E}_{k_0} .

The inequalities established by Theorem 2.3.1 and Theorem 2.3.2 allow us to translate existing results on convergence rates (under Hellinger distance) of maximum likelihood density estimation to that of the mixing measure (under Wasserstein distance metrics). Under standard assumptions, the convergence rate for density estimation using finite mixture densities is $(\log n/n)^{1/2}$. Then it follows that the con-

vergence rate for the mixing measure under W_1 distance in the exact-fitted setting is also $(\log n/n)^{1/2}$. For the over-fitted setting, the rate is $(\log n/n)^{1/4}$ under W_2 distance.

To state such results formally, we need to introduce several standard notions, which will allow us to appeal to a general convergence theorem for the MLE (e.g., [van de Geer, 2000]). For any positive integer number k_1 , define several classes of mixture densities $\mathcal{P}_{k_1}(\Theta^*) = \{p_G : G \in \mathcal{O}_{k_1}(\Theta^*)\}$, $\bar{\mathcal{P}}_{k_1}(\Theta^*) = \left\{p_{\frac{G+G_0}{2}} : G \in \mathcal{O}_{k_1}(\Theta^*)\right\}$, and $\bar{\mathcal{P}}_{k_1}^{1/2}(\Theta^*) = \left\{\left(p_{\frac{G+G_0}{2}}\right)^{1/2} : G \in \mathcal{O}_{k_1}(\Theta^*)\right\}$. For any $\delta > 0$, define the intersection with a Hellinger ball centered at p_{G_0} via

$$\bar{\mathcal{P}}_{k_1}^{1/2}(\Theta^*, \delta) = \left\{\left(p_{\frac{G+G_0}{2}}\right)^{1/2} \in \bar{\mathcal{P}}_{k_1}^{1/2} : h(p_{\frac{G+G_0}{2}}, p_{G_0}) \leq \delta\right\}.$$

The size of this set is captured by the entropy integral:

$$\mathcal{J}_B(\delta, \bar{\mathcal{P}}_{k_1}^{1/2}(\Theta^*, \delta), \mu) = \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(u, \bar{\mathcal{P}}_{k_1}^{1/2}(\Theta^*, u), \mu) du \vee \delta,$$

where H_B denotes the bracketing entropy of $\bar{\mathcal{P}}_{k_1}^{1/2}(\Theta^*)$ under L_2 distance (cf. van de Geer [2000] for a definition of the bracket entropy).

Before arriving at the main results in this section, we state the result of Theorem 7.4 of van de Geer [2000] with the adaption of notations as those in our paper

Theorem 2.4.1. *Take $\Psi(\delta) \geq J_B(\delta, \bar{\mathcal{P}}_{k_1}^{1/2}(\Theta^*, \delta), \mu)$ in such a way that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ . Then, for a universal constant c and for*

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n),$$

we have for all $\delta \geq \delta_n$

$$P(h(p_{G_n}, p_{G_0}) > \delta) \leq c \exp \left[-\frac{n\delta^2}{c^2} \right].$$

Now, we are ready to state a general result on the MLE under the exact-fitted mixture setting:

Theorem 2.4.2. (Exact-fitted mixtures) *Assume that f satisfies the conditions of Theorem 2.3.1. Take $\Psi(\delta) \geq \mathcal{J}_B(\delta, \bar{\mathcal{P}}_{k_0}^{1/2}(\Theta^*, \delta), \mu_0)$ in such a way that $\frac{\Psi(\delta)}{\delta^2}$ is a non-increasing function of δ . Then for a universal constant c , constant $C_1 = C_1(\Theta^*)$, a non-negative sequence $\{\delta_n\}$ such that*

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n),$$

and for all $\delta \geq \frac{\delta_n}{\sqrt{C_1}}$, we have

$$P(W_1(\hat{G}_n, G_0) > \delta) \leq c \exp \left(-\frac{nC_1^2\delta^2}{c^2} \right).$$

Proof. By Theorem 2.3.1,

$$C_1(\Theta^*)W_1^2(G, G_0) \leq V^2(p_G, p_{G_0}) \leq 2h^2(p_G, p_{G_0}) \text{ for all } G \in \mathcal{E}_{k_0}(\Theta^*), \quad (2.5)$$

where $C_1(\Theta^*)$ is a positive constant depending only on Θ^* and G_0 . Now, invoking Theorem 2.4.1, as $\delta \geq \delta_n$, there is a universal constant $c > 0$ such that

$$P(h(p_{\hat{G}_n}, p_{G_0}) > \delta) \leq c \exp \left(-\frac{n\delta^2}{c^2} \right). \quad (2.6)$$

Combining (2.5) and (2.6), we readily achieve the conclusion of our theorem. \square

Using the same argument we arrive at a general convergence rate result of the

MLE under the over-fitted setting:

Theorem 2.4.3. (Over-fitted mixtures) *Assume that f satisfies the conditions in part (a) of Theorem 2.3.2. Take $\Psi(\delta) \geq \mathcal{J}_B(\delta, \bar{\mathcal{P}}_k^{1/2}(\Theta^*, \delta), \mu_0)$ in such a way that $\frac{\Psi(\delta)}{\delta^2}$ is a non-increasing function of δ . Then for a universal constant c , constant $C_1 = C_1(\Theta^*)$, $\{\delta_n\}$ is a non-negative sequence such that*

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n),$$

and for all $\delta \geq \frac{\delta_n}{\sqrt{C_1}}$, we have

$$P(W_2(\hat{G}_n, G_0) > \delta^{1/2}) \leq c \exp\left(-\frac{nC_1^2\delta^2}{c^2}\right).$$

To derive the concrete rates δ_n , one simply need to verify the conditions on the bracket entropy integral \mathcal{J}_B for a given model class. As a concrete example, the following results are concerned with the finite mixtures of multivariate generalized Gaussian distributions.

Corollary 2.4.1. (Mixtures of multivariate generalized Gaussian distributions) *Given $\Theta = [-a_n, a_n]^d \times [\underline{m}, \bar{m}]$ where $a_n \leq L(\log(n))^\gamma$ as L is some positive constant, $\gamma > 0$, and $1 < \underline{m} \leq \bar{m}$ are two known positive numbers. Let $\{f(x|\theta, m, \Sigma)|(\theta, m) \in \Theta, \Sigma \in \Omega\}$ to be the class of multivariate generalized Gaussian distributions.*

(a) (Exact-fitted case) There holds $P(W_1(\hat{G}_n, G_0) > \delta_n) \lesssim \exp(-c\log(n))$, where δ_n is a sufficiently large multiple of $(\log(n)/n)^{1/2}$ and c is positive constant depending only on $L, \gamma, \underline{m}, \bar{m}, \underline{\lambda}, \bar{\lambda}$.

(b) (Over-fitted case) There holds $P(W_2(\hat{G}_n, G_0) > \delta'_n) \lesssim \exp(-c\log(n))$, where δ'_n is a sufficiently large multiple of $(\log(n)/n)^{1/4}$ and c is positive constant depending only on $L, \gamma, \underline{m}, \bar{m}, \underline{\lambda}, \bar{\lambda}$.

Remark (i) The condition $\underline{m} > 1$ can be relaxed to $\underline{m} \geq 1$ under the exact-fitted setting; however, it is crucial under the over-fitted setting that $\underline{m} > 1$. In fact, the location-covariance Gaussian mixtures (which correspond to $m = 1$) have to be excluded from the class of generalized Gaussian mixtures for the above results to hold. This is a consequence of the fact that the (sub)class of location-covariance multivariate Gaussian distributions is not identifiable in the second order. In fact, the failure to satisfy the second-order identifiability leads to very slow convergence rate of the MLE under the over-fitted location-scale Gaussian mixture setting, as we noted briefly in the introduction. (ii) The conclusions of this corollary also hold for mixtures of multivariate Student's t-distribution as well as all the classes of distributions considered in Theorem 2.3.3 with suitable boundedness conditions on the parameter spaces.

Finally, we shall show that the convergence rates $n^{-1/2}$ and $n^{-1/4}$ for the exact-fitted and over-fitted finite mixtures, respectively, are in fact minimax lower bounds. Under the exact-fitted finite mixture setting, it is simple to establish the standard $n^{-1/2}$ minimax lower bound:

$$\inf_{\widehat{G}_n \in \mathcal{E}_{k_0}} \sup_{G_0 \in \mathcal{E}_{k_0}} E_{p_{G_0}} W_1(\widehat{G}_n, G_0) \gtrsim n^{-1/2},$$

where the infimum is taken over all possible sequences of estimate \widehat{G}_n based on n -samples. Perhaps more interesting is the following minimax lower bound result for the over-fitted mixture setting.

Theorem 2.4.4. (Minimax lower bound for over-fitted mixtures) *If the class of densities f satisfies condition (c) of Theorem 2.3.2, then for any positive $r < 4$ and any $n \geq 1$,*

$$\inf_{\widehat{G}_n \in \mathcal{O}_k} \sup_{G_0 \in \mathcal{O}_k \setminus \mathcal{O}_{k_0-1}} E_{p_{G_0}} W_1(\widehat{G}_n, G_0) \gtrsim n^{-1/r}.$$

Proof. The proof is almost immediate following a standard argument for establishing

minimax lower bounds. Fix a $G_0 \in \mathcal{E}_{k_0}$ and $r \in [1, 2)$. Let $C_0 > 0$ be any fixed constant. According to Theorem 2.3.2, part (c), for any sufficiently small $\epsilon > 0$, there exists $G'_0 \in \mathcal{O}_k$ such that $W_1(G_0, G'_0) = 2\epsilon$ and $h(p_{G_0}, p_{G'_0}) \leq C_0\epsilon^r$. Applying Lemma 1 from Yu [1997], for any sequence of estimates \widehat{G}_n ranging in \mathcal{O}_k , we obtain that

$$\sup_{G \in \{G_0, G'_0\}} E_{p_G} W_1(\widehat{G}_n, G) \geq \epsilon \left(1 - V(p_{G_0}^n, p_{G'_0}^n)\right),$$

where $p_{G_0}^n$ denotes the density of the n -iid sample X_1, \dots, X_n . Now,

$$\begin{aligned} V(p_{G_0}^n, p_{G'_0}^n) &\leq h(p_{G_0}^n, p_{G'_0}^n) \\ &= \sqrt{1 - (1 - h^2(p_{G_0}, p_{G'_0}))^n} \\ &\leq \sqrt{1 - (1 - C_0^2\epsilon^{2r})^n}. \end{aligned}$$

As a consequence, we obtain

$$\sup_{G \in \{G_0, G'_0\}} E_{p_G} W_1(\widehat{G}_n, G) \geq \epsilon \left(1 - \sqrt{1 - (1 - C_0^2\epsilon^{2r})^n}\right).$$

By choosing $\epsilon^{2r} = \frac{1}{C_0^2 n}$, the right hand side of the above inequality is bounded below by $C_1\epsilon \asymp n^{-1/2r}$ for any $r < 2$ where C_1 is some positive constant. We achieve the conclusion of our theorem. Noting that $G_0, G'_0 \in \mathcal{O}_k \setminus \mathcal{O}_{k_0-1}$, this concludes the proof of our theorem. \square

2.4.2 Illustrations

For the remainder of this section, we shall illustrate via simulations the strong identifiability bounds established in Section 2.3 for several classes of distributions with matrix-variate parameter space for which strong identifiability conditions in both the first and second order hold. In addition, we also present some simulations

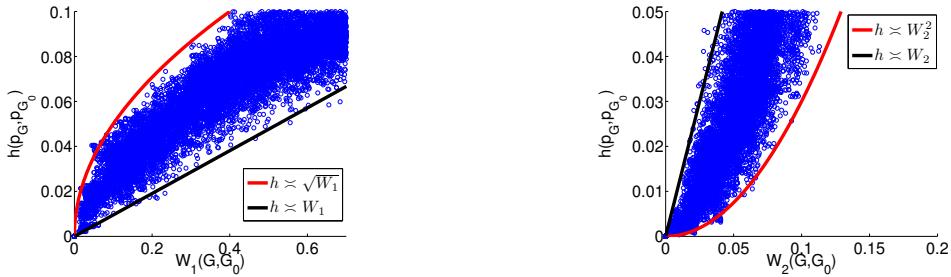


Figure 2.1: Mixture of Student's t-distributions. Left: Exact-fitted setting. Right: Over-fitted setting.

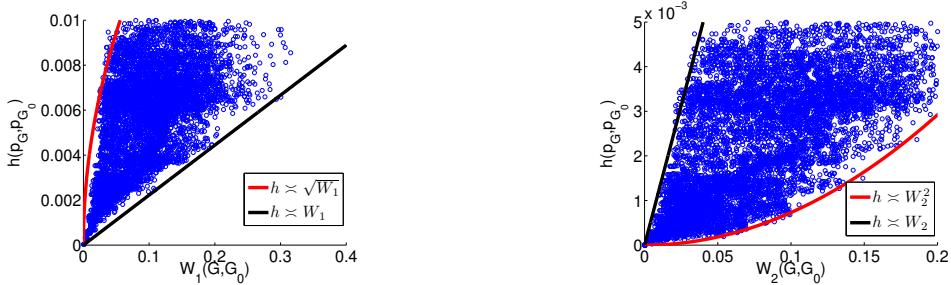


Figure 2.2: Mixture of multivariate generalized Gaussian distributions. Left: Exact-fitted setting. Right: Over-fitted setting.

for the well-known location-scale Gaussian finite mixtures, which satisfy the first-order identifiability but not the second-order identifiability.

Strong identifiability bounds The inequalities $V \gtrsim W_1$ for exact-fitted mixtures and $V \gtrsim W_2^2$ for over-fitted mixtures are illustrated for the class of Student's t-distributions and the class of multivariate generalized Gaussian distributions, both of which satisfy first and second-order identifiability. See Figure 2.1 and Figure 2.2. Here we plot h against W_1 and W_2^2 , but note the relation $h \geq V \geq h^2$. The upper bounds of V and h in terms of W_1 were given in Section 2.2.

For details, we choose $\Theta = [-10, 10]^2$ for Student's t-distribution (Gaussian distribution) or $\Theta = [-10, 10]^2 \times [1.5, 5]$ for multivariate generalized Gaussian distribution, $\Omega = \left\{ \Sigma \in S_2^{++} : \sqrt{2} \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq 2 \right\}$. Note that closed interval $[1.5, 5]$ is chosen for m to exclude Gaussian distributions, which corresponds to $m = 1$. Now, the true mixing probability measure G_0 has exactly $k_0 = 2$ support points with loca-

tions $\theta_1^0 = (-2, 2)$, $\theta_2^0 = (-4, 4)$, covariances $\Sigma_1^0 = \begin{pmatrix} 9/4 & 1/5 \\ 1/5 & 13/6 \end{pmatrix}$, $\Sigma_2^0 = \begin{pmatrix} 5/2 & 2/5 \\ 2/5 & 7/3 \end{pmatrix}$, $m_1^0 = 2$, $m_2^0 = 3$ (for the setting of multivariate generalized Gaussian distribution), and $p_1^0 = 1/3, p_2^0 = 2/3$. 10000 random samples of discrete mixing measures $G \in \mathcal{E}_2(\Theta \times \Omega)$, 10000 samples of $G \in \mathcal{O}_3(\Theta \times \Omega)$ were generated to construct these plots. Note that, since we focus on obtaining the lower bound of Hellinger distance in terms of small Wasserstein distances, we generate G by making small perturbations of G_0 (that is, adding small random noise ϵ to the mixing coefficients and support points of G_0).

It can be observed that both lower bounds and upper bounds match exactly that of our theorems for strongly identifiable classes of distributions such as the t-distribution and the generalized Gaussian distribution. Turning to mixtures of location-covariance Gaussian distributions (Figure 3.1), the bounds $\sqrt{W_1} \gtrsim h \gtrsim W_1$ continue to hold under the exact-fitted setting, but under the over-fitted setting it can be observed that the lower bound $h \gtrsim W_2^2$ no longer holds. In fact, if the Gaussian mixture is over-fitted by one extra component, it can be shown that $h \gtrsim W_4^4 \geq W_2^4$ (see illustrations in the middle and right panels), and that this bound is sharp. This has a drastic consequence on the convergence rate of the mixing measure, which we discuss next.

Convergence rates of MLE First, we generate n -iid samples from a bivariate location-covariance Gaussian mixture with three components with an arbitrarily fixed choice of G_0 . The true parameters for the mixing measure G_0 are: $\theta_1^0 = (0, 3)$, $\theta_2^0 = (1, -4)$, $\theta_3^0 = (5, 2)$, $\Sigma_1^0 = \begin{pmatrix} 4.2824 & 1.7324 \\ 1.7324 & 0.81759 \end{pmatrix}$, $\Sigma_2^0 = \begin{pmatrix} 1.75 & -1.25 \\ -1.25 & 1.75 \end{pmatrix}$, $\Sigma_3^0 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$, and $p_1^0 = 0.3, p_2^0 = 0.4, p_3^0 = 0.3$. The parameter spaces Θ, Ω are iden-

tical to those of multivariate Student's t-distribution setting. MLE \widehat{G}_n is obtained by the EM algorithm as we assume that the data come from a mixture of k Gaussians where $k \geq k_0 = 3$. See Figure 3.2 where the Wasserstein distances between \widehat{G}_n and G_0 are plotted against increasing sample size n ($n \leq 30000$). The error bars were obtained by running the experiment 7 times for each n . The simulation results under the exact-fitted case match quite well with the standard $n^{-1/2}$ rate. If we fit the data to a mixture of $k = k_0 + 1 = 4$ Gaussian distributions, however, we observe empirically that the convergence rate of $W_4(\widehat{G}_n, G_0)$ (thus W_2 distance) is almost $n^{-1/8}$ up to a logarithmic term. This result is much slower than the “standard” convergence rate $n^{-1/4}$ under W_2 , should second-identifiability condition holds. A rigorous theory for weakly identifiable mixture models such as location-covariance Gaussian mixtures will be reported elsewhere.

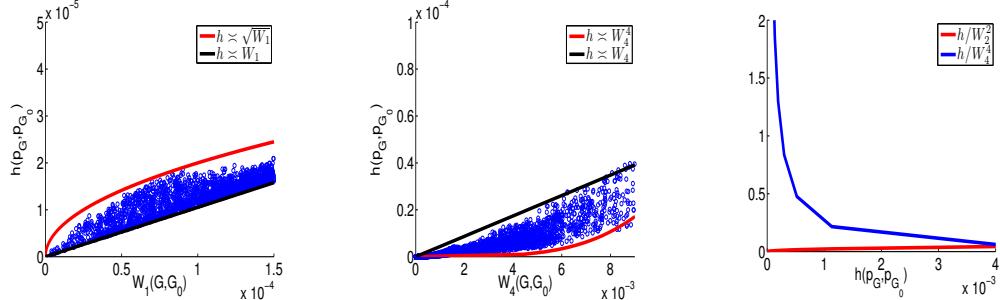


Figure 2.3: Mixture of location-scale Gaussian distributions, which satisfy first-order identifiability but not second-order identifiability condition. Left panel: Exact-fitted setting. Middle and right panels are for over-fitted setting by one extra component. Right panel shows that $h \gtrsim W_2^2$ no longer holds as $h \rightarrow 0$.

2.5 Proofs of key theorems

In this section, we present self-contained proofs for two key theorems: Theorem 2.3.1 for strongly identifiable mixtures in the exact-fitted setting and Theorem 2.3.2 for strongly identifiable mixtures in the over-fitted setting. These moderately long proofs carry useful insights underlying the theory — they are organized in a

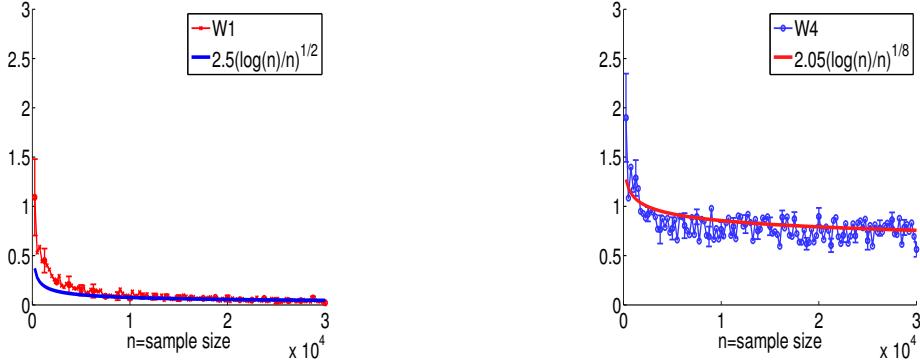


Figure 2.4: MLE rates for location-covariance mixtures of Gaussians. Left: Exact-fitted — $W_1 \asymp n^{-1/2}$. Right: Over-fitted by one — $W_4 \asymp n^{-1/8}$.

sequence of steps to help the reader. The proofs of the remaining results are deferred to Section 2.6.

2.5.1 Strong identifiability in exact-fitted mixtures

PROOF OF THEOREM 2.3.1 It suffices to show that

$$\liminf_{\epsilon \rightarrow 0} \left\{ V(p_G, p_{G_0}) / W_1(G, G_0) \mid W_1(G, G_0) \leq \epsilon \right\} > 0, \quad (2.7)$$

where the infimum is taken over all $G \in \mathcal{E}_{k_0}(\Theta \times \Omega)$.

Step 1 Suppose that (2.7) does not hold, which implies that we have a sequence of $G_n = \sum_{i=1}^{k_0} p_i^n \delta_{(\theta_i^n, \Sigma_i^n)} \in \mathcal{E}_{k_0}(\Theta \times \Omega)$ converging to G_0 in the W_1 distance such that $V(p_{G_n}, p_{G_0}) / W_1(G_n, G_0) \rightarrow 0$ as $n \rightarrow \infty$. As $W_1(G_n, G_0) \rightarrow 0$, the support points of G_n must converge to that of G_0 . By permutation of the labels i , it suffices to assume that for each $i = 1, \dots, k_0$, $(\theta_i^n, \Sigma_i^n) \rightarrow (\theta_i^0, \Sigma_i^0)$. For each pair (G_n, G_0) , let $\{q_{ij}^n\}$ denote the corresponding probabilities of the optimal coupling for the pair (G_n, G_0) , so we can write:

$$W_1(G_n, G_0) = \sum_{1 \leq i, j \leq k_0} q_{ij}^n (\|\theta_i^n - \theta_j^0\| + \|\Sigma_i^n - \Sigma_j^0\|).$$

Since $(\theta_i^n, \Sigma_i^n) \rightarrow (\theta_i^0, \Sigma_i^0)$ and G_n and G_0 have the same number of support points, it is an easy observation that for sufficiently large n , $q_{ii}^n = \min(p_i^n, p_i^0)$. And so, $\sum_{i \neq j} q_{ij}^n = \sum_{i=1}^{k_0} |p_i^n - p_i^0|$. Adopting the notations that $\Delta\theta_i^n := \theta_i^n - \theta_i^0$, $\Delta\Sigma_i^n := \Sigma_i^n - \Sigma_i^0$, and $\Delta p_i^n := p_i^n - p_i^0$ for all $1 \leq i \leq k_0$, we have

$$\begin{aligned} W_1(G_n, G_0) &= \sum_{i=1}^{k_0} q_{ii}^n (\|\Delta\theta_i^n\| + \|\Delta\Sigma_i^n\|) + \sum_{i \neq j} q_{ij}^n (\|\theta_i^n - \theta_j^0\| + \|\Sigma_i^n - \Sigma_j^0\|) \\ &\lesssim \sum_{i=1}^{k_0} p_i^n (\|\Delta\theta_i^n\| + \|\Delta\Sigma_i^n\|) + |\Delta p_i^n| =: d(G_n, G_0). \end{aligned}$$

The inequality in the above display is due to $q_{ii}^n \leq p_i^n$, and the observation that $\|\theta_i^n - \theta_j^0\|, \|\Sigma_i^n - \Sigma_j^0\|$ are bounded for all $1 \leq i, j \leq k_0$ for sufficiently large n . Thus, we have $V(p_{G_n}, p_{G_0})/d(G_n, G_0) \rightarrow 0$.

Step 2 Now, consider the following important identity:

$$p_{G_n}(x) - p_{G_0}(x) = \sum_{i=1}^{k_0} \Delta p_i^n f(x|\theta_i^0, \Sigma_i^0) + \sum_{i=1}^{k_0} p_i^n (f(x|\theta_i^n, \Sigma_i^n) - f(x|\theta_i^0, \Sigma_i^0)).$$

For each x , applying Taylor expansion to function f to the first order to obtain

$$\begin{aligned} \sum_{i=1}^{k_0} p_i^n (f(x|\theta_i^n, \Sigma_i^n) - f(x|\theta_i^0, \Sigma_i^0)) &= \sum_{i=1}^{k_0} p_i^n \left\{ (\Delta\theta_i^n)^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0) + \right. \\ &\quad \left. \text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \Delta\Sigma_i^n \right) \right\} + R_n(x), \end{aligned}$$

where $R_n(x) = O \left(\sum_{i=1}^{k_0} p_i^n (\|\Delta\theta_i^n\|^{1+\delta_1} + \|\Delta\Sigma_i^n\|^{1+\delta_2}) \right)$, where the appearance of δ_1 and δ_2 are due the assumed Lipschitz conditions, and the big-O constant does not depend on x . It is clear that $\sup_x |R_n(x)/d(G_n, G_0)| \rightarrow 0$ as $n \rightarrow \infty$.

Denote $A_n(x) = \sum_{i=1}^{k_0} p_i^n \left[(\Delta\theta_i^n)^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0) + \text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \Delta\Sigma_i^n \right) \right]$, $B_n(x) =$

$\sum_{i=1}^k \Delta p_i^n f(x|\theta_i^0, \Sigma_i^0)$. Then, we can rewrite

$$(p_{G_n}(x) - p_{G_0}(x))/d(G_n, G_0) = (A_n(x) + B_n(x) + R_n(x))/d(G_n, G_0).$$

Step 3 We see that $A_n(x)/d(G_n, G_0)$ and $B_n(x)/d(G_n, G_0)$ are linear combinations of the scalar elements of $f(x|\theta, \Sigma)$, $\frac{\partial f}{\partial \theta}(x|\theta, \Sigma)$ and $\frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma)$ such that the coefficients do not depend on x . We shall argue that *not* all such coefficients in the linear combination converge to 0 as $n \rightarrow \infty$. Indeed, if the opposite is true, then the summation of the absolute values of these coefficients must also tend to 0:

$$\left\{ \sum_{i=1}^{k_0} |\Delta p_i^n| + p_i^n (\|\Delta \theta_i^n\|_1 + \|\Delta \Sigma_i^n\|_1) \right\} / d(G_n, G) \rightarrow 0.$$

Since we have the entrywise ℓ_1 and ℓ_2 norms are equivalent, the above entails

$$\left\{ \sum_{i=1}^{k_0} |\Delta p_i^n| + p_i^n (\|\Delta \theta_i^n\| + \|\Delta \Sigma_i^n\|) \right\} / d(G_n, G_0) \rightarrow 0,$$

which contradicts with the definition of $d(G_n, G_0)$. As a consequence, we can find at least one coefficient of the elements of $A_n(x)/d(G_n, G_0)$ or $B_n(x)/d(G_n, G_0)$ that does not vanish as $n \rightarrow \infty$.

Step 4 Let m_n be the maximum of the absolute value of the scalar coefficients of $A_n(x)/d(G_n, G_0)$, $B_n(x)/d(G_n, G_0)$ and $d_n = 1/m_n$, then d_n is uniformly bounded from above for all n . Thus, as $n \rightarrow \infty$,

$$\begin{aligned} d_n A_n(x)/d(G_n, G_0) &\rightarrow \sum_{i=1}^{k_0} \beta_i^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0) + \text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \gamma_i \right), \\ d_n B_n(x)/d(G_n, G_0) &\rightarrow \sum_{i=1}^{k_0} \alpha_i f(x|\theta_i^0, \Sigma_i^0), \end{aligned}$$

such that *not* all scalar elements of α_i, β_i and γ_i vanish. Moreover, γ_i are symmetric matrices because Σ_i^n are symmetric matrices for all n, i . Note that

$$\begin{aligned} d_n V(p_{G_n}, p_{G_0}) / d(G_n, G_0) &= \int d_n |p_{G_n}(x) - p_{G_0}(x)| / d(G_n, G_0) \\ &= \int d_n |A_n(x) + B_n(x) + R_n(x)| / d(G_n, G_0) dx \rightarrow 0. \end{aligned}$$

By Fatou's lemma, the integrand in the above display vanishes for almost all x . Thus, for almost all x

$$\sum_{i=1}^{k_0} \alpha_i f(x|\theta_i^0, \Sigma_i^0) + \beta_i^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0) + \text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \gamma_i \right) = 0.$$

By the first-order identifiability criteria of f , we have $\alpha_i = 0, \beta_i = \mathbf{0} \in \mathbb{R}^{d_1}$, and $\gamma_i = \mathbf{0} \in \mathbb{R}^{d_2 \times d_2}$ for all $i = 1, 2, \dots, k$, which is a contradiction. Hence, (2.7) is proved.

2.5.2 Strong identifiability in over-fitted mixtures

PROOF OF THEOREM 2.3.2 (a) We only need to establish that

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_k(\Theta)} \left\{ \sup_{x \in \mathcal{X}} |p_G(x) - p_{G_0}(x)| / W_2^2(G, G_0) : W_2(G, G_0) \leq \epsilon \right\} > 0. \quad (2.8)$$

The conclusion of the theorem follows from an application of Fatou's lemma in the same manner as Step 4 in the proof of Theorem 2.3.1.

Step 1 Suppose that (2.8) does not hold, then we can find a sequence $G_n \in \mathcal{O}_k(\Theta)$ tending to G_0 in W_2 distance and $\sup_{x \in \mathcal{X}} |p_{G_n}(x) - p_{G_0}(x)| / W_2^2(G_n, G_0) \rightarrow 0$ as $n \rightarrow \infty$. Since k is finite, there is some $k^* \in [k_0, k]$ such that there exists a subsequence of G_n having exactly k^* support points. We cannot have $k^* = k_0$, due to Theorem 2.3.1 and the fact that $W_2^2(G_n, G_0) \lesssim W_1(G_n, G_0)$ for all n . Thus, $k_0 + 1 \leq k^* \leq k$.

Write $G_n = \sum_{i=1}^{k^*} p_i^n \delta_{(\theta_i^n, \Sigma_i^n)}$ and $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, \Sigma_i^0)}$. Since $W_2(G_n, G_0) \rightarrow 0$, there

exists a subsequence of G_n such that each support point (θ_i^0, Σ_i^0) of G_0 is the limit of a subset of $s_i \geq 1$ support points of G_n . There may also a subset of support points of G_n whose limits are not among the support points of G_0 — we assume there are $m \geq 0$ such limit points. To avoid notational cluttering, we replace the subsequence of G_n by the whole sequence $\{G_n\}$. By re-labeling the support points, G_n can be expressed by

$$G_n = \sum_{i=1}^{k_0+m} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(\theta_{ij}^n, \Sigma_{ij}^n)} \xrightarrow{W_2} G_0 = \sum_{i=1}^{k_0+m} p_i^0 \delta_{(\theta_i^0, \Sigma_i^0)}$$

where $(\theta_{ij}^n, \Sigma_{ij}^n) \rightarrow (\theta_i^0, \Sigma_i^0)$ for each $i = 1, \dots, k_0 + m$, $j = 1, \dots, s_i$, $p_i^0 = 0$ for $i < k_0$, and we have that $p_i^n := \sum_{j=1}^{s_i} p_{ij}^n \rightarrow p_i^0$ for all i . Moreover, the constraint $k_0 + 1 \leq \sum_{i=1}^{k_0+m} s_i \leq k$ must hold.

We note that if matrix Σ is (strictly) positive definite whose maximum eigenvalue is bounded (from above) by constant M , then Σ is also bounded under the entrywise ℓ_2 norm. However if Σ is only positive semidefinite, it can be singular and its ℓ_2 norm potentially unbounded. In our context, for $i \geq k_0 + 1$ it is possible that the limiting matrices Σ_i^0 can be singular. It comes from the fact that the some eigenvalues of Σ_{ij}^n can go to 0 as $n \rightarrow \infty$, which implies $\det(\Sigma_{ij}^n) \rightarrow 0$ and hence $\det(\Sigma_i^0) = 0$. By re-labeling the support points, we may assume without loss of generality that $\Sigma_{k_0+1}^0, \dots, \Sigma_{k_0+m_1}^0$ are (strictly) positive definite matrices and $\Sigma_{k_0+m_1+1}^0, \dots, \Sigma_{k_0+m}^0$ are singular and positive semidefinite matrices for some $m_1 \in [0, m]$. For those singular matrices, we shall make use of the assumption that for each $\theta \in \Theta$, except a finite number of values of $x \in \mathcal{X}$, we have $\lim_{\lambda_1(\Sigma) \rightarrow 0} f(x|\theta, \Sigma) = 0$ and the fact that θ_{ij}^n as $k_0 + m_1 + 1 \leq i \leq k_0 + m$ will converge to at most $m - m_1 \leq k - k_0$ limit points: accordingly, for all x except a finite number of values in \mathcal{X} , $f(x|\theta_{ij}^n, \Sigma_{ij}^n) \rightarrow 0$ as $n \rightarrow \infty$ for all $k_0 + m_1 + 1 \leq i \leq k_0 + m$, $1 \leq j \leq s_i$. Here, we denote $f(x|\theta_i^0, \Sigma_i^0) = 0$ for all $k_0 + m_1 + 1 \leq i \leq k_0 + m$.

Step 2 Using shorthand notations $\Delta\theta_{ij}^n := \theta_{ij}^n - \theta_i^0$, $\Delta\Sigma_{ij}^n := \Sigma_{ij}^n - \Sigma_i^0$ for $i = 1, \dots, k_0 + m_1$ and $j = 1, \dots, s_i$, it is simple to see that

$$W_2^2(G_n, G_0) \lesssim d(G_n, G_0) := \sum_{i=1}^{k_0+m_1} \sum_{j=1}^{s_i} p_{ij}^n (\|\Delta\theta_{ij}^n\|^2 + \|\Delta\Sigma_{ij}^n\|^2) + \sum_{i=1}^{k_0+m} |p_{i.}^n - p_i^0|,$$

because $W_2^2(G_n, G_0)$ is the optimal transport cost with respect to ℓ_2^2 , while $d(G_n, G_0)$ corresponds to a multiple of the cost of a possibly non-optimal transport plan, which is achieved by coupling the atoms $(\theta_{ij}^n, \Sigma_{ij}^n)$ for $j = 1, \dots, s_i$ with (θ_i^0, Σ_i^0) by mass $\min(p_{i.}^n, p_i^0)$, while the remaining masses are coupled arbitrarily. From the assumption, $\sup_{x \in \mathcal{X}} |p_{G_n}(x) - p_{G_0}(x)|/W_2^2(G_n, G_0)$ vanishes in the limit, it also implies that $\sup_{x \in \mathcal{X}} |p_{G_n}(x) - p_{G_0}(x)|/d(G_n, G_0) \rightarrow 0$.

For each x , we make use of the key identity:

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{k_0+m_1} \sum_{j=1}^{s_i} p_{ij}^n (f(x|\theta_{ij}^n, \Sigma_{ij}^n) - f(x|\theta_i^0, \Sigma_i^0)) \\ &\quad + \sum_{i=1}^{k_0+m_1} (p_{i.}^n - p_i^0) f(x|\theta_i^0, \Sigma_i^0) \\ &\quad + \sum_{i=k_0+m_1+1}^{k_0+m} \sum_{j=1}^{s_i} p_{ij}^n f(x|\theta_{ij}^n, \Sigma_{ij}^n) \\ &:= A_n(x) + B_n(x) + C_n(x). \end{aligned} \tag{2.9}$$

Step 3 By means of Taylor expansion up to the second order:

$$\begin{aligned} A_n(x) &= \sum_{i=1}^{k_0+m_1} \sum_{j=1}^{s_i} p_{ij}^n (f(x|\theta_{ij}^n, \Sigma_{ij}^n) - f(x|\theta_i^0, \Sigma_i^0)) = \sum_{i=1}^{k_0+m_1} \sum_{\alpha} A_{\alpha_1, \alpha_2}^n(\theta_i^0, \Sigma_i^0) \\ &\quad + R_n(x), \end{aligned}$$

where $\alpha = (\alpha_1, \alpha_2)$ such that $\alpha_1 + \alpha_2 \in \{1, 2\}$. Specifically,

$$\begin{aligned}
A_{1,0}^n(\theta_i^0, \Sigma_i^0) &= \sum_{j=1}^{s_i} p_{ij}^n (\Delta\theta_{ij}^n)^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0), \\
A_{0,1}^n(\theta_i^0, \Sigma_i^0) &= \sum_{j=1}^{s_i} p_{ij}^n \text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \Delta\Sigma_{ij}^n \right), \\
A_{2,0}^n(\theta_i^0, \Sigma_i^0) &= \frac{1}{2} \sum_{j=1}^{s_i} p_{ij}^n (\Delta\theta_{ij}^n)^T \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^0, \Sigma_i^0) \Delta\theta_{ij}^n, \\
A_{0,2}^n(\theta_i^0, \Sigma_i^0) &= \frac{1}{2} \sum_{j=1}^{s_i} p_{ij}^n \text{tr} \left(\frac{\partial}{\partial \Sigma} \left(\text{tr} \left(\frac{\partial}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \Delta\Sigma_{ij}^n \right) \right)^T \Delta\Sigma_{ij}^n \right), \\
A_{1,1}^n(\theta_i^0, \Sigma_i^0) &= 2 \sum_{j=1}^{s_i} (\Delta\theta_{ij}^n)^T \left[\frac{\partial}{\partial \theta} \left(\text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \Delta\Sigma_{ij}^n \right) \right) \right].
\end{aligned}$$

In addition, $R_n(x) = O\left(\sum_{i=1}^{k_0+m_1} \sum_{j=1}^{s_i} p_{ij}^n (\|\Delta\theta_{ij}^n\|^{2+\delta} + \|\Delta\Sigma_{ij}^n\|^{2+\delta})\right)$ due to the second-order Lipschitz condition. It is clear that $\sup_x |R_n(x)|/d(G_n, G_0) \rightarrow 0$ as $n \rightarrow \infty$.

Step 4 Write $D_n := d(G_n, G_0)$ for short. Note that $(p_{G_n}(x) - p_{G_0}(x))/D_n$ is a linear combination of the scalar elements of $f(x|\theta, \Sigma)$ and its derivatives taken with respect to θ and Σ up to the second order, and evaluated at the distinct pairs (θ_i^0, Σ_i^0) for $i = 1, \dots, k_0 + m$. (To be specific, the elements of $f(x|\theta, \Sigma)$, $\frac{\partial f}{\partial \theta}(x|\theta, \Sigma)$, $\frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma)$, $\frac{\partial^2 f}{\partial \theta^2}(x|\theta, \Sigma)$, $\frac{\partial^2 f}{\partial \theta^2}(x|\theta, \Sigma)$, $\frac{\partial^2 f}{\partial \Sigma^2}(x|\theta, \Sigma)$, and $\frac{\partial^2 f}{\partial \theta \partial \Sigma}(x|\theta, \Sigma)$). In addition, the coefficients associated with these elements do not depend on x . As in the proof of Theorem 2.3.1, we shall argue that *not* all such coefficients vanish as $n \rightarrow \infty$. Indeed, if this is not true, then by taking the summation of all the absolute value of the coefficients associated with the elements of $\frac{\partial^2 f}{\partial \theta_l^2}$ as $1 \leq l \leq d_1$ and $\frac{\partial^2 f}{\partial \Sigma_{uv}^2}$ for $1 \leq u, v \leq d_2$, we obtain

$$\sum_{i=1}^{k_0+m_1} \sum_{j=1}^{s_i} p_{ij}^n (\|\Delta\theta_{ij}^n\|^2 + \|\Delta\Sigma_{ij}^n\|^2)/D_n \rightarrow 0.$$

Therefore, $\sum_{i=1}^{k_0+m} |p_i^n - p_i^0|/D_n \rightarrow 1$ as $n \rightarrow \infty$. It implies that we should have at least one coefficient associated with an element of $f(x|\theta, \Sigma)$ (appearing in $B_n(x)/D_n$, $C_n(x)/D_n$) not converging to 0 as $n \rightarrow \infty$, which is a contradiction. As a consequence, not all the coefficients vanish to 0.

Step 5 Let m_n be the maximum of the absolute value of the aforementioned coefficients. and set $d_n = 1/m_n$. Then, d_n is uniformly bounded above when n is sufficiently large. Therefore, as $n \rightarrow \infty$, we obtain

$$\begin{aligned}
d_n B_n(x)/D_n &\rightarrow \sum_{i=1}^{k_0+m_1} \alpha_i f(x|\theta_i^0, \Sigma_i^0), \\
d_n \sum_{i=1}^{k_0+m_1} A_{1,0}^n(\theta_i^0, \Sigma_i^0)/D_n &\rightarrow \sum_{i=1}^{k_0+m_1} \beta_i^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0), \\
d_n \sum_{i=1}^{k_0+m_1} A_{0,1}^n(\theta_i^0, \Sigma_i^0)/D_n &\rightarrow \sum_{i=1}^{k_0+m_1} \text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \gamma_i \right), \\
d_n \sum_{i=1}^{k_0+m_1} A_{2,0}^n(\theta_i^0, \Sigma_i^0)/D_n &\rightarrow \sum_{i=1}^{k_0+m_1} \sum_{j=1}^{s_i} \nu_{ij}^T \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^0, \Sigma_i^0) \nu_{ij}, \\
d_n \sum_{i=1}^{k_0+m_1} A_{1,1}^n(\theta_i^0, \Sigma_i^0)/D_n &\rightarrow \sum_{i=1}^{k_0+m_1} \sum_{j=1}^{s_i} \nu_{ij}^T \left[\frac{\partial}{\partial \theta} \left(\text{tr} \left(\frac{\partial}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \eta_{ij} \right) \right) \right], \\
d_n \sum_{i=1}^{k_0+m_1} A_{0,2}^n(\theta_i^0, \Sigma_i^0)/D_n &\rightarrow \sum_{i=1}^{k_0+m_1} \sum_{j=1}^{s_i} \text{tr} \left(\frac{\partial}{\partial \Sigma} \left(\text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \eta_{ij} \right) \right) \right)^T \times \\
&\quad \times \eta_{ij},
\end{aligned}$$

where $\alpha_i \in \mathbb{R}$, $\beta_i, \nu_{i1}, \dots, \nu_{is_i} \in \mathbb{R}^{d_1}$, $\gamma_i, \eta_{i1}, \dots, \eta_{is_i}$ are symmetric matrices in $\mathbb{R}^{d_2 \times d_2}$ for all $1 \leq i \leq k_0 + m_1$, $1 \leq j \leq s_i$. Additionally,

$$d_n C_n(x)/D_n = D_n^{-1} \sum_{i=k_0+m_1+1}^{k_0+m} \sum_{j=1}^{s_i} d_n p_{ij}^n f(x|\theta_{ij}^n, \Sigma_{ij}^n) \rightarrow 0$$

due to the fact that for almost all x , $f(x|\theta_{ij}^n, \Sigma_{ij}^n) \rightarrow 0$ for all $k_0 + m_1 + 1 \leq i \leq k_0 + m$, $1 \leq j \leq s_i$ and the fact that $d_n p_{ij}^n / D_n \leq 1$ for all $k_0 + m_1 + 1 \leq i \leq k_0 + m$, $1 \leq j \leq s_i$. As a consequence, we obtain for almost all x that

$$\begin{aligned} & \sum_{i=1}^{k_0+m_1} \left\{ \alpha_i f(x|\theta_i^0, \Sigma_i^0) + \beta_i^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0) + \sum_{j=1}^{s_i} \nu_{ij}^T \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^0, \Sigma_i^0) \nu_{ij} \right. + \\ & \quad \text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \gamma_i \right) + 2 \sum_{j=1}^{s_i} \nu_{ij}^T \left[\frac{\partial}{\partial \theta} \left(\text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \eta_{ij} \right) \right) \right] + \\ & \quad \left. \sum_{j=1}^{s_i} \text{tr} \left(\frac{\partial}{\partial \Sigma} \left(\text{tr} \left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \eta_{ij} \right) \right)^T \eta_{ij} \right) \right\} = 0. \quad (2.10) \end{aligned}$$

Now, in this paragraph we will argue that not all coefficients in (2.10) go to 0 as $n \rightarrow \infty$. There are two scenarios. Case 1: If m_n , the maximum of all the coefficients considered in Step 4, does not lie in the set $\{p_{ij}^n/D_n\}$ as $k_0 + m_1 + 1 \leq i \leq k_0 + m$, $1 \leq j \leq s_i$ for infinitely many n . Then, it indicates that at least one coefficient in (2.10) should be 1. Our observation is proved. Case 2: Otherwise, m_n lies in the set $\{p_{ij}^n/D_n\}$ as $k_0 + m_1 + 1 \leq i \leq k_0 + m$, $1 \leq j \leq s_i$ for infinitely many n . This means that we can find two indices $i^* \in [k_0 + m_1 + 1, k_0 + m]$, $j^* \in [1, s_{i^*}]$ such that $m_n = p_{i^*j^*}^n/D_n$. Assume now that all of the coefficients in (2.10) vanish to 0. Therefore, $d_n |p_i^n - p_i^0|/D_n = |p_i^n - p_i^0|/p_{i^*j^*}^n \rightarrow 0$ for all $1 \leq i \leq k_0 + m_1$. Since we have $p_{i^*j^*}^n \leq \sum_{i=k_0+m_1+1}^{k_0+m} \sum_{j=1}^{s_i} p_{ij}^n \leq \sum_{i=1}^{k_0+m_1} |p_i^n - p_i^0|$, this leads to $|p_i^n - p_i^0|/\sum_{i=1}^{k_0+m_1} |p_i^n - p_i^0| \rightarrow 0$ for all $1 \leq i \leq k_0 + m_1$ as $n \rightarrow \infty$, which is a contradiction. Our observation is proved.

Therefore, at least one coefficient in (2.10) is different from 0. However, from the second-order identifiability of $\{f(x|\theta, \Sigma), \theta \in \Theta, \Sigma \in \Omega\}$, we obtain $\alpha_i = 0, \beta_i = \nu_{i1} = \dots = \nu_{is_i} = \mathbf{0} \in \mathbb{R}^{d_1}, \gamma_i = \eta_{i1} = \dots = \eta_{is_i} = \mathbf{0} \in \mathbb{R}^{d_2 \times d_2}$ for all $1 \leq i \leq k_0 + m_1$, which is a contradiction. This concludes the proof of Eq. (2.8) and that of the theorem.

(b) Recall $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, \Sigma_i^0)}$. Construct a sequence of probability measures G_n

having exactly $k_0 + 1$ support points as follows: $G_n = \sum_{i=1}^{k_0+1} p_i^n \delta_{(\theta_i^n, \Sigma_i^n)}$, where $\theta_1^n = \theta_1^0 - \frac{1}{n} \mathbf{1}_{d_1}$, $\theta_2^n = \theta_1^0 + \frac{1}{n} \mathbf{1}_{d_1}$, $\Sigma_1^n = \Sigma_1^0 - \frac{1}{n} I_{d_2}$ and $\Sigma_2^n = \Sigma_1^0 + \frac{1}{n} I_{d_2}$. Here, I_{d_2} denotes the identity matrix in $\mathbb{R}^{d_2 \times d_2}$ and $\mathbf{1}_n$ a vector with all elements being equal to 1. In addition, $(\theta_{i+1}^n, \Sigma_{i+1}^n) = (\theta_i^0, \Sigma_i^0)$ for all $i = 2, \dots, k_0$. Also, $p_1^n = p_2^n = \frac{p_1^0}{2}$ and $p_{i+1}^n = p_i^0$ for all $i = 2, \dots, k_0$. It is simple to verify that $E_n := W_1^r(G_n, G_0) = \frac{(p_1^0)^r}{2^r} (\|\theta_1^n - \theta_1^0\| + \|\theta_2^n - \theta_2^0\| + \|\Sigma_1^n - \Sigma_1^0\| + \|\Sigma_2^n - \Sigma_1^0\|)^r = \frac{(p_1^0)^r}{2^r} (\sqrt{d_1} + \sqrt{d_2})^r \frac{1}{n^r} \asymp \frac{1}{n^r}$.

By means of Taylor's expansion up to the first order, we get that as $n \rightarrow \infty$

$$\begin{aligned} V(p_{G_n}, p_{G_0}) &\asymp \int_{x \in \mathcal{X}} \left| \sum_{i=1}^2 \sum_{\alpha_1, \alpha_2} (\Delta \theta_{1i}^n)^{\alpha_1} (\Delta \Sigma_{1i}^n)^{\alpha_2} \frac{\partial f}{\partial \theta^{\alpha_1} \partial \Sigma^{\alpha_2}}(x | \theta_1^0, \Sigma_1^0) + \right. \\ &\quad \left. + R_1(x) \right| dx \\ &= \int_{x \in \mathcal{X}} |R_1(x)| dx, \end{aligned}$$

where $\alpha_1 \in \mathbb{N}^{d_1}$, $\alpha_2 \in \mathbb{N}^{d_2 \times d_2}$ in the sum such that $|\alpha_1| + |\alpha_2| = 1$, $R_1(x)$ is Taylor expansion's remainder. The second equality in the above equation is due to $\sum_{i=1}^2 (\Delta \theta_{1i}^n)^{\alpha_1} (\Delta \Sigma_{1i}^n)^{\alpha_2} = 0$ for each α_1, α_2 such that $|\alpha_1| + |\alpha_2| = 1$. Since f is second-order differentiable with respect to θ, Σ , $R_1(x)$ takes the form

$$\begin{aligned} R_1(x) &= \sum_{i=1}^2 \sum_{|\alpha|=2} \frac{2}{\alpha!} (\Delta \theta_{1i}^n)^{\alpha_1} (\Delta \Sigma_{1i}^n)^{\alpha_2} \times \\ &\quad \times \int_0^1 (1-t) \frac{\partial^2 f}{\partial \theta^{\alpha_1} \partial \Sigma^{\alpha_2}}(x | \theta_1^0 + t \Delta \theta_{1i}^n, \Sigma_1^0 + t \Delta \Sigma_{1i}^n) dt, \end{aligned}$$

where $\alpha = (\alpha_1, \alpha_2)$. Note that, $\sum_{i=1}^2 |\Delta_{1i}^n|^{\alpha_1} |\Delta \Sigma_{1i}^n|^{\alpha_2} = O(n^{-2})$. Additionally, from the hypothesis, $\sup_{t \in [0, 1]} \int_{x \in \mathcal{X}} \left| \frac{\partial^2 f}{\partial \theta^{\alpha_1} \partial \Sigma^{\alpha_2}}(x | \theta_1^0 + t \Delta \theta_{1i}^n, \Sigma_1^0 + t \Delta \Sigma_{1i}^n) \right| dx < \infty$. It follows that $\int |R_1(x)| dx = O(n^{-2})$. So for any $r < 2$, $V(p_{G_n}, p_{G_0}) = o(W_1^r(G_n, G_0))$. This concludes the proof.

(c) Continuing with the same sequence G_n constructed in part (b), we have

$$h^2(p_{G_n}, p_{G_0}) \leq \frac{1}{2p_1^0} \int_{x \in \mathcal{X}} \frac{(p_{G_n}(x) - p_{G_0}(x))^2}{f(x|\theta_1^0, \Sigma_1^0)} dx \lesssim \int_{x \in \mathcal{X}} \frac{R_1^2(x)}{f(x|\theta_1^0, \Sigma_1^0)} dx.$$

where first inequality is due to $\sqrt{p_{G_n}(x)} + \sqrt{p_{G_0}(x)} > \sqrt{p_{G_0}(x)} > \sqrt{p_1^0 f(x|\theta_1^0, \Sigma_1^0)}$ and the second inequality is because of Taylor expansion taken to the first order. The proof proceeds in the same manner as that of part (b).

2.6 Proofs of other results

2.6.1 Extension to the whole domain in exact-fitted mixtures

PROOF OF COROLLARY 2.3.1 By Theorem 2.3.1, there are positive constants $\epsilon = \epsilon(G_0)$ and $C_0 = C_0(G_0)$ such that $V(p_G, p_{G_0}) \geq C_0 W_1(G, G_0)$ when $W_1(G, G_0) \leq \epsilon$. It remains to show that $\inf_{G \in \mathcal{G}: W_1(G, G_0) > \epsilon} V(p_G, p_{G_0}) / W_1(G, G_0) > 0$. Assume the contrary, then we can find a sequence of $G_n \in \mathcal{G}$ and $W_1(G_n, G_0) > \epsilon$ such that $\frac{V(p_{G_n}, p_{G_0})}{W_1(G_n, G_0)} \rightarrow 0$ as $n \rightarrow \infty$. Since \mathcal{G} is a compact set, we can find $G' \in \mathcal{G}$ and $W_1(G', G_0) > \epsilon$ such that $G_n \rightarrow G'$ under W_1 metric. It implies that $W_1(G_n, G_0) \rightarrow W_1(G', G_0)$ as $n \rightarrow \infty$. As $G' \not\equiv G_0$, we have $\lim_{n \rightarrow \infty} W_1(G_n, G_0) > 0$. As a consequence, $V(p_{G_n}, p_{G_0}) \rightarrow 0$ as $n \rightarrow \infty$.

From the hypothesis, $V(p_{G_n}, p_{G'}) \leq C(\Theta, \Omega) W_1^\alpha(G_n, G')$, so $V(p_{G_n}, p_{G'}) \rightarrow 0$ as $W_1(G_n, G') \rightarrow 0$. Thus, $V(p_{G'}, p_{G_0}) = 0$ or equivalently $p_{G_0} = p_{G'}$ almost surely. From the first-order identifiability of $\{f(x|\theta, \Sigma), \theta \in \Theta, \Sigma \in \Omega\}$, it implies that $G' \equiv G_0$, which is a contradiction. This completes the proof.

2.6.2 The importance of boundedness conditions in the over-fitted setting

PROOF OF PROPOSITION 2.3.1 We choose $G_n = \sum_{i=1}^{k_0+1} p_i^n \delta_{(\theta_i^n, \Sigma_i^n)} \in \mathcal{O}_k(\Theta \times \Omega)$ such that $(\theta_i^n, \Sigma_i^n) = (\theta_i^0, \Sigma_i^0)$ for $i = 1, \dots, k_0$, $\theta_{k_0+1}^n = \theta_1^0$, $\Sigma_{k_0+1}^n = \Sigma_1^0 + \frac{\exp(n/r)}{n^\alpha} I_{d_2}$

where $\alpha = \frac{1}{2\beta}$. Additionally, $p_1^n = p_1^0 - \exp(-n)$, $p_i^n = p_i^0$ for all $2 \leq i \leq k_0$, and $p_{k_0+1}^n = \exp(-n)$. With this construction, we can check that $W_r^\beta(G, G_0) = d_2^{\beta/2}/\sqrt{n}$.

Now, as $h^2(p_{G_n}, p_{G_0}) \lesssim V(p_{G_n}, p_{G_0})$, we have

$$\begin{aligned} \exp\left(\frac{2}{W_r^\beta(G_n, G_0)}\right) h^2(p_G, p_{G_0}) &\lesssim \exp\left(-n + \frac{2\sqrt{n}}{d_2^{\beta/2}}\right) \times \\ &\int_{x \in \mathcal{X}} |f(x|\theta_1^0, \Sigma_{k_0+1}^n) - f(x|\theta_1^0, \Sigma_1^0)| dx, \end{aligned}$$

which converges to 0 as $n \rightarrow \infty$. The conclusion of our proposition is proved.

2.6.3 Characterization of strong identifiability

PROOF OF THEOREM 2.3.4 Here, we only present the proof for part (a) and part (b). The proofs for part (c) and (d) are somewhat similar and omitted.

(a) Assume that for given $k \geq 1$ and k different pairs $(\theta_1, \Sigma_1, m_1), \dots, (\theta_k, \Sigma_k, m_k)$, we can find $\alpha_j \in \mathbb{R}$, $\beta_j \in \mathbb{R}^d$, symmetric matrices $\gamma_j \in \mathbb{R}^{d \times d}$, and $\eta_j \in \mathbb{R}$, for $j = 1, \dots, k$ such that:

$$\begin{aligned} \sum_{j=1}^k \alpha_j f(x|\theta_j, \Sigma_j, m_j) + \beta_j^T \frac{\partial f}{\partial \theta}(x|\theta_j, \Sigma_j, m_j) + \text{tr}\left(\frac{\partial f}{\partial \Sigma}(x|\theta_j, \Sigma_j, m_j)^T \gamma_j\right) \\ + \eta_j \frac{\partial f}{\partial m}(x|\theta_j, \Sigma_j, m_j) = 0, \end{aligned}$$

Substituting the first derivatives of f to get

$$\begin{aligned} \sum_{j=1}^k \left\{ \alpha'_j + \left((\beta'_j)^T (x - \theta_j) + (x - \theta_j)^T \gamma'_j (x - \theta_j) \right) \times \right. \\ \left. \left[(x - \theta_j)^T \Sigma_j^{-1} (x - \theta_j) \right]^{m_j-1} + \eta'_j \log[(x - \theta_j)^T \Sigma_j^{-1} (x - \theta_j)] \right\} \times \\ \exp\left(-\left[(x - \theta_j)^T \Sigma_j^{-1} (x - \theta_j) \right]^{m_j}\right) = 0, \quad (2.11) \end{aligned}$$

where

$$\begin{aligned}\alpha'_j &= \frac{2\alpha_j m_j \Gamma(d/2) - m_j \Gamma(d/2) \text{tr}(\Sigma_j^{-1} \gamma_j) + 2\eta_j \Gamma(d/2) \left(1 - \frac{d}{2m_j} \psi\left(\frac{d}{2m_j}\right)\right)}{2\pi^{d/2} \Gamma(d/(2m_j)) |\Sigma_j|^{1/2}}, \\ \beta'_j &= \frac{2m_j^2 \Gamma(d/2)}{\pi^{d/2} \Gamma(d/(2m_j)) |\Sigma_j|^{1/2}} \Sigma_j^{-1} \beta_j, \quad \gamma'_j = \frac{m_j^2 \Gamma(d/2)}{\pi^{d/2} \Gamma(d/(2m_j)) |\Sigma_j|^{1/2}} \Sigma_j^{-1} \gamma_j \Sigma_j^{-1}, \text{ and} \\ \eta'_j &= \frac{-m_j \eta_j \Gamma(d/2)}{\pi^{d/2} \Gamma(d/(2m_j)) |\Sigma_j|^{1/2}}.\end{aligned}$$

Without loss of generality, assume $m_1 \leq m_2 \leq \dots \leq m_k$. Let $\bar{i} \in [1, k]$ be the maximum index such that $m_1 = m_{\bar{i}}$. As the tuples $(\theta_i, \Sigma_i, m_i)$ are distinct, so are the pairs $(\theta_1, \Sigma_1), \dots, (\theta_{\bar{i}}, \Sigma_{\bar{i}})$. In what follows, we denote $x = x_1 x'$ where x_1 is scalar and $x' \in \mathbb{R}^d$. Define

$$a_i = (x')^T \gamma'_i x', \quad b_i = [(\beta'_i)^T - 2\theta_i^T \gamma'_i] x', \quad c_i = \theta_i^T \gamma'_i \theta_i - (\beta'_i)^T \theta_i,$$

$$d_i = (x')^T \Sigma_i^{-1} x', \quad e_i = -2(x')^T \Sigma_i^{-1} \theta_i, \quad f_i = \theta_i^T \Sigma_i^{-1} \theta_i.$$

Borrowing a technique from [Yakowitz and Spragins \[1968\]](#), since $(\theta_1, \Sigma_1), \dots, (\theta_{\bar{i}}, \Sigma_{\bar{i}})$ are distinct, we have two possibilities:

Possibility 1 If Σ_j are the same for all $1 \leq j \leq \bar{i}$, then $\theta_1, \dots, \theta_{\bar{i}}$ are distinct. For any $i < j$, denote $\Delta_{ij} = \theta_i - \theta_j$. Now, if $x' \notin \bigcup_{1 \leq i < j \leq \bar{i}} \{u \in \mathbb{R}^d : u^T \Delta_{ij} = 0\}$, which is a finite union of hyperplanes, then $(x')^T \theta_1, \dots, (x')^T \theta_{\bar{i}}$ are distinct. Hence, if we choose $x' \in \mathbb{R}^d$ lying outside this union of hyperplanes, we will have $((x')^T \theta_1, (x')^T \Sigma_1 x'), \dots, ((x')^T \theta_{\bar{i}}, (x')^T \Sigma_{\bar{i}} x')$ are distinct.

Possibility 2 If Σ_j are not the same for all $1 \leq j \leq \bar{i}$, then we assume without loss of generality that $\Sigma_1, \dots, \Sigma_m$ are the only distinct matrices from $\Sigma_1, \dots, \Sigma_{\bar{i}}$, where $m \leq \bar{i}$. Denote $\delta_{ij} = \Sigma_i - \Sigma_j$ as $1 \leq i < j \leq m$, then as x' does not belong to $\bigcup_{1 \leq i < j \leq m} \{u \in \mathbb{R}^d : u^T \delta_{ij} u = 0\}$, we will have $(x')^T \Sigma_1 x', \dots, (x')^T \Sigma_m x'$ are distinct. Therefore, if x' does not belong to $\bigcup_{1 \leq i < j \leq m} \{u \in \mathbb{R}^d : u^T \delta_{ij} u = 0\}$, which is a finite

union of conics, then we have $((x')^T \theta_1, (x')^T \Sigma_1 x'), \dots, ((x')^T \theta_m, (x')^T \Sigma_m x')$ are distinct. Additionally, for any θ_j where $m+1 \leq j \leq \bar{i}$ that shares the same Σ_i where $1 \leq i \leq m$, using the argument in the first case, we can choose x' outside a finite hyperplane such that these $(x')^T \theta_j$ are again distinct. Hence, for x' lying outside a finite union of conics and hyperplanes, we have that $((x')^T \theta_1, (x')^T \Sigma_1 x'), \dots, ((x')^T \theta_{\bar{i}}, (x')^T \Sigma_{\bar{i}} x')$ are all different.

From these two cases, we can find a set D , which is a finite union of conics and hyperplanes, such that as $x' \notin D$, $((x')^T \theta_1, (x')^T \Sigma_1 x'), \dots, ((x')^T \theta_{\bar{i}}, (x')^T \Sigma_{\bar{i}} x')$ are distinct. Thus, (d_i, e_i) are different as $1 \leq i \leq \bar{i}$.

Choose $d_{i_1} = \min_{1 \leq i \leq \bar{i}} \{d_i\}$. Denote $J = \{1 \leq i \leq \bar{i} : d_i = d_{i_1}\}$. Choose $1 \leq i_2 \leq \bar{i}$ such that $e_{i_2} = \max_{i \in J} \{e_i\}$. Now, we define for all $1 \leq i \leq k$ that

$$A_i(x_1) = \alpha'_i + (a_i x_1^2 + b_i x_1 + c_i)(d_i x_1^2 + e_i x_1 + f_i)^{m_i-1} + \eta'_i \log(d_i x_1^2 + e_i x_1 + f_i).$$

Multiplying both sides of (2.11) with $\exp - (d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}}$, we get

$$A_{i_2}(x_1) + \sum_{j \neq i_2} A_j(x_1) \exp \left[(d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} \right] = 0. \quad (2.12)$$

Note that if $j \in J \setminus \{i_2\}$, $d_j = d_{i_2}$, $m_j = m_{i_2}$, and $e_j > e_{i_2}$. So,

$$(d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} \lesssim -x_1 \text{ as } x_1 \text{ is large enough.}$$

This implies that when $x_1 \rightarrow \infty$,

$$B_1(x_1) := \sum_{j \neq J \setminus \{i_2\}} A_j(x_1) \exp \left[(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} - (d_jx_1^2 + e_jx_1 + f_j)^{m_j} \right] \rightarrow 0.$$

On the other hand, if $j \notin J$ and $1 \leq j \leq \bar{i}$, then $d_j > d_{i_2}$ and $m_{i_2} = m_j$. So,

$$(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} - (d_jx_1^2 + e_jx_1 + f_j)^{m_j} \lesssim -x_1^{2m_{i_2}} \text{ as } x_1 \text{ is large.}$$

This implies that when $x_1 \rightarrow \infty$,

$$B_2(x_1) := \sum_{\substack{j \notin J, \\ 1 \leq j \leq \bar{i}}} A_j(x_1) \exp \left[(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} - (d_jx_1^2 + e_jx_1 + f_j)^{m_j} \right] \rightarrow 0.$$

Otherwise, if $j > \bar{i}$, then $m_j > m_{i_2}$. So,

$$(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} - (d_jx_1^2 + e_jx_1 + f_j)^{m_j} \lesssim -x_1^{2m_j}.$$

As a result,

$$B_3(x_1) := \sum_{j > \bar{i}} A_j(x_1) \exp \left[(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} - (d_jx_1^2 + e_jx_1 + f_j)^{m_j} \right] \rightarrow 0.$$

Now, by letting $x_1 \rightarrow \infty$,

$$\begin{aligned} \sum_{j \neq i_2} A_j(x_1) \exp \left[(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} - (d_jx_1^2 + e_jx_1 + f_j)^{m_j} \right] &= \\ A_1(x) + A_2(x) + A_3(x) &\rightarrow 0. \end{aligned} \quad (2.13)$$

Combining (2.12) and (2.13), we obtain that as $x_1 \rightarrow \infty$, $A_{i_2}(x_1) \rightarrow 0$. The only possibility for this result to happen is $a_{i_2} = b_{i_2} = \eta'_{i_2} = 0$. Or, equivalently, $(x')^T \gamma'_{i_2} x' = [(\beta'_i)^T - 2\theta_{i_2}^T \gamma'_{i_2}] x' = 0$. If $\gamma'_{i_2} \neq 0$, we can choose the element $x' \notin D$ lying outside the hyperplane $\{u \in \mathbb{R}^d : u^T \gamma'_{i_2} u = 0\}$. It means that $(x')^T \gamma'_{i_2} x' \neq 0$, which is a contradiction. Therefore, $\gamma'_{i_2} = 0$. It implies that $(\beta'_{i_2})^T x' = 0$. If $\beta'_{i_2} \neq 0$, we can choose $x' \notin D$ such that $(\beta'_{i_2})^T x' \neq 0$. Hence, $\beta'_{i_2} = 0$. With these results, $\alpha'_{i_2} = 0$. Overall, we obtain $\alpha'_{i_2} = \beta'_{i_2} = \gamma'_{i_2} = \eta'_{i_2} = 0$. Repeating the same argument to the remaining parameters $\alpha'_j, \beta'_j, \gamma'_j, \eta'_j$, we get $\alpha'_j = \beta'_j = \gamma'_j = \eta'_j = 0$ for $1 \leq j \leq k$. It is also equivalent that $\alpha_j = \beta_j = \gamma_j = \eta_j = 0$ for all $1 \leq j \leq k$. This concludes the proof of part (a) of our theorem.

(b) Consider that for given $k \geq 1$ and k different pairs $(\theta_1, \Sigma_1), \dots, (\theta_k, \Sigma_k)$, where $\theta_j \in \mathbb{R}^d$, $\Sigma_j \in S_d^{++}$ for all $1 \leq j \leq k$, we can find $\alpha_j \in \mathbb{R}$, $\beta_j \in \mathbb{R}^d$, and symmetric matrices $\gamma_j \in \mathbb{R}^{d \times d}$ such that:

$$\sum_{j=1}^k \alpha_j f(x|\theta_j, \Sigma_j) + \beta_j^T \frac{\partial f}{\partial \theta}(x|\theta_j, \Sigma_j) + \text{tr}\left(\frac{\partial f}{\partial \Sigma}(x|\theta_j, \Sigma_j)^T \gamma_j\right) = 0. \quad (2.14)$$

Multiplying both sides with $\exp(it^T x)$ and taking the integral in \mathbb{R}^d , after direct

calculations, the above equation can be rewritten as

$$\sum_{j=1}^k \left[\int_{\mathbb{R}^d} \left(\frac{\alpha'_j \exp(i(\Sigma_j^{1/2} t)^T x)}{(\nu + \|x\|^2)^{(\nu+d)/2}} + \frac{\exp(i(\Sigma_j^{1/2} t)^T x)(\beta'_j)^T x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} + \frac{\exp(i(\Sigma_j^{1/2} t)^T x)x^T M_j x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} \right) dx \right] \exp(it^T \theta_j) = 0, \quad (2.15)$$

where $\alpha'_j = \alpha_j - \frac{\text{tr}(\Sigma_j^{-1} \gamma_j)}{2}$, $\beta'_j = \frac{(\nu+d)}{2} \Sigma_j^{-1/2} \beta_j$, and $M_j = \frac{\nu+d}{2} \Sigma_j^{-1/2} \gamma_j \Sigma_j^{-1/2}$.

To simplify the left hand side of equation (2.15), it is sufficient to calculate the following quantities $A = \int_{\mathbb{R}^d} \frac{\exp(it^T x)}{(\nu + \|x\|^2)^{(\nu+d)/2}} dx$, $B = \int_{\mathbb{R}^d} \frac{\exp(it^T x)(\beta')^T x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} dx$, and $C = \int_{\mathbb{R}^d} \frac{\exp(it^T x)x^T M x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} dx$, where $\beta' \in \mathbb{R}^d$ and $M = (M_{ij}) \in \mathbb{R}^{d \times d}$.

In fact, by using an orthogonal transformation $x = O.z$, where $O \in \mathbb{R}^{d \times d}$ and its first column to be $(\frac{t_1}{\|t\|}, \dots, \frac{t_d}{\|t\|})^T$, we can verify that $\exp(it^T x) = \exp(i\|t\|z_1)$, $\|x\|^2 = \|z\|^2$, and $dx = |\det(O)|dz = dz$ and then we obtain the following results:

$$\begin{aligned} A &= \int_{\mathbb{R}^d} \frac{\exp(i\|t\|z_1)}{(\nu + \|z\|^2)^{(\nu+d)/2}} dz \\ &= \int_{\mathbb{R}} \exp(i\|t\|z_1) \int_{\mathbb{R}} \dots \int_{\mathbb{R}} \frac{1}{(\nu + \|z\|^2)^{(\nu+d)/2}} dz_d dz_{d-1} \dots dz_1 \\ &= C_1 A_1(\|t\|), \end{aligned}$$

where $C_1 = \prod_{j=2}^d \int_{\mathbb{R}} \frac{1}{(1+z^2)^{(\nu+j)/2}} dz$ and $A_1(t') = \int_{\mathbb{R}} \frac{\exp(i|t'|z)}{(v+z^2)^{(\nu+1)/2}} dz$ for any $t' \in \mathbb{R}$. Hence, for all $1 \leq j \leq k$

$$\int_{\mathbb{R}^d} \frac{\exp(i(\Sigma_j^{1/2} t)^T x)}{(\nu + \|x\|^2)^{(\nu+d)/2}} dx = C_1 A_1(\|\Sigma_j^{1/2} t\|). \quad (2.16)$$

Turning to B and C , by the same line of calculations we obtain

$$\begin{aligned} B &= \left(\sum_{j=1}^d O_{j1} \beta'_j \right) \int_{\mathbb{R}^d} \frac{\exp(it^t z_1) z_1}{(\nu + \|z\|^2)^{(\nu+d+2)/2}} dz \\ &= \left(\sum_{j=1}^d O_{j1} \beta'_j \right) C_2 A_2(\|t\|) \\ &= \frac{C_2(\beta')^T t A_2(\|t\|)}{\|t\|}. \end{aligned}$$

where $C_2 = \prod_{j=2}^d \int_{\mathbb{R}} \frac{1}{(1+z^2)^{(\nu+2+j)/2}} dz$ and $A_2(t') = \int_{\mathbb{R}} \frac{\exp(i|t'|z)z}{(\nu+z^2)^{(\nu+3)/2}} dz$ for any $t' \in \mathbb{R}$.

$$\begin{aligned} C &= C_3 \left(\sum_{j=1}^d M_{jj} \right) A_1(\|t\|) + \left(\sum_{jl} M_{jl} O_{j1} O_{l1} \right) (C_2 A_3(\|t\|) - C_3 A_1(\|t\|)) \\ &= C_3 \left(\sum_{j=1}^d M_{jj} \right) A_1(\|t\|) + \frac{1}{\|t\|^2} \left(\sum_{j,l} M_{jl} t_j t_l \right) (C_2 A_3(\|t\|) - C_3 A_1(\|t\|)). \end{aligned}$$

where we can define $C_3 = \int_{\mathbb{R}} \frac{z^2}{(1+z^2)^{(\nu+4)/2}} dz \prod_{j=3}^k \int_{\mathbb{R}} \frac{1}{(1+z^2)^{(\nu+2+j)/2}} dz$ and $A_3(t') = \int_{\mathbb{R}} \frac{\exp(i|t'|z)z^2}{(\nu+z^2)^{(\nu+3)/2}} dz$ for any $t' \in \mathbb{R}$. Thus, for all $1 \leq j \leq d$

$$\int_{\mathbb{R}^d} \frac{\exp(i(\Sigma_j^{1/2} t)^T x)(\beta'_j)^T x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} dx = \frac{C_2(\beta')^T \Sigma_j^{1/2} t A_2(\|\Sigma_j^{1/2} t\|)}{\|t\|}. \quad (2.17)$$

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{\exp(i(\Sigma_j^{1/2} t)^T x)x^T M_j x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} dx &= \frac{1}{\|\Sigma_j^{1/2} t\|^2} \left(\sum_{u,v} M_{uv}^j [\Sigma_j^{1/2} t]_u [\Sigma_j^{1/2} t]_v \right) \times \\ &\quad \times (C_2 A_3(\|\Sigma_j^{1/2} t\|) - C_3 A_1(\|\Sigma_j^{1/2} t\|)) + C_3 \left(\sum_{l=1}^d M_{ll}^j \right) A_1(\|\Sigma_j^{1/2} t\|), \end{aligned} \quad (2.18)$$

where M_{uv}^j indicates the element at u -th row and v -th column of M_j and $[\Sigma_j^{1/2} t]_u$

simply means the u -th component of $\Sigma_j^{1/2}t$.

As a consequence, by combining (2.16), (2.17), and (2.18), we can rewrite (2.15) as:

$$\begin{aligned} & \sum_{j=1}^k \left[\alpha'_j A_1(\|\Sigma_j^{1/2}t\|) + C_2 \frac{(\Sigma_j^{1/2}t)^T \beta'_j}{\|\Sigma_j^{1/2}t\|} A_2(\|\Sigma_j^{1/2}t\|) \right. + \\ & C_3 \left(\sum_{l=1}^d M_{ll}^j \right) A_1(\|\Sigma_j^{1/2}t\|) + \left(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2}t]_u [\Sigma_j^{1/2}t]_v}{\|\Sigma_j^{1/2}t\|^2} \right) (C_2 A_3(\|\Sigma_j^{1/2}t\|) - \\ & \left. C_3 A_1(\|\Sigma_j^{1/2}t\|)) \right] \exp(it^T \theta_j) = 0. \end{aligned}$$

Define $t = t_1 t'$, where $t_1 \in \mathbb{R}$ and $t' \in \mathbb{R}^d$. By using the same argument as in the case of the multivariate generalized Gaussian distribution, we can find D to be the finite union of conics and hyperplanes such that as $t' \notin D$, we have $((t')^T \theta_1, (t')^T \Sigma_1 t'), \dots ((t')^T \theta_k, (t')^T \Sigma_k t')$ are pairwise distinct. By denoting $\theta'_j = (t')^T \theta_j$, $\sigma_j = (t')^T \Sigma_j t'$, we can rewrite the above equation as:

$$\begin{aligned} & \sum_{j=1}^k \left[\alpha'_j A_1(\sigma_j |t_1|) + C_2 \frac{t_1 (\Sigma_j^{1/2} t')^T \beta'_j}{|t_1| \sigma_j} A_2(\sigma_j |t_1|) + C_3 \left(\sum_{l=1}^d M_{ll}^j \right) A_1(\sigma_j |t_1|) \right. + \\ & \left. \left(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} \right) (C_2 A_3(\sigma_j |t_1|) - C_3 A_1(\sigma_j |t_1|)) \right] \exp(i \theta'_j t_1) = 0. \end{aligned}$$

Since $A_2(\sigma_j |t_1|) = (i |t_1|) A_1(\sigma_j |t_1|)$, the above equation can be rewritten as:

$$\begin{aligned} & \sum_{j=1}^k \left[\left(\alpha'_j + C_3 \left(\sum_{l=1}^d M_{ll}^j \right) - C_3 \left(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} \right) \right) \right. \times \\ & \times A_1(\sigma_j |t_1|) + C_2 \left(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} \right) A_3(\sigma_j |t_1|) + \\ & \left. C_2 (i t_1) \frac{(\Sigma_j^{1/2} t')^T \beta'_j}{\sigma_j} A_1(\sigma_j |t_1|) \right] \exp(i \theta'_j t_1) = 0. \quad (2.19) \end{aligned}$$

As ν is an odd number, we assume $\nu = 2l - 1$. By using a classical result in complex

analysis, we obtain for any $m \in \mathbb{N}$ that

$$\int_{-\infty}^{+\infty} \frac{\exp(i|t_1|z)}{(z^2 + \nu)^m} dz = \frac{2\pi \exp(-|t_1|\sqrt{2l-1})}{(2\sqrt{2l-1})^{2m-1}} \left[\sum_{j=1}^m \binom{2m-1-j}{m-j} \frac{(2|t_1|\sqrt{2l-1})^{j-1}}{(j-1)!} \right].$$

It means that we can write $A_1(t_1) = C_4 \exp(-|t_1|\sqrt{2l-1}) \sum_{u=0}^{l-1} a_u |t_1|^u$, where $C_4 =$

$$\frac{2\pi}{(2\sqrt{2l-1})^{2l-1}}, a_u = \binom{2l-u-2}{l-u-1} \frac{(2\sqrt{2l-1})^u}{u!}.$$

Simultaneously, as $A_3(t_1) = A_1(t_1) - \nu \int_{\mathbb{R}} \frac{\exp(i|t_1|z)}{(\nu + z^2)^{(\nu+3)/2}} dz$, we can write

$$A_3(t_1) = C_4 \exp(-|t_1|\sqrt{2l-1}) \sum_{u=0}^l b_u |t_1|^u,$$

where $b_u = \left[\binom{2l-u-2}{l-u-1} - \frac{1}{4} \binom{2l-u}{l-u} \right] \frac{(2\sqrt{2l-1})^u}{u!}$ as $0 \leq u \leq l-1$, and $b_l = -\frac{1}{4} \frac{(2\sqrt{2l-1})^l}{l!}$. It is not hard to notice that $a_0, a_{l-1}, b_l \neq 0$.

Now, for all $t_1 \in \mathbb{R}$, equation (2.19) can be rewritten as:

$$\begin{aligned} \sum_{j=1}^k \left[\left(\alpha_j'' + \beta_j''(it_1) \right) \sum_{u=0}^{l-1} a_u \sigma_j^u |t_1|^u + \gamma_j'' \sum_{u=0}^l b_u \sigma_j^u |t_1|^u \right] &\times \\ \exp(it\theta_j' - \sigma_j \sqrt{2l-1} |t_1|) &= 0, \end{aligned}$$

where we have $\alpha_j'' = \alpha_j' + C_3 \left(\sum_{l=1}^d M_{ll}^j \right) - C_3 \left(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} \right)$, $\beta_j'' = C_2 \frac{(\Sigma_j^{1/2} t')^T \beta_j'}{\sigma_j}$,

and $\gamma_j'' = C_2 \left(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} \right)$. The above equation yields that for all $t_1 \geq 0$

$$\begin{aligned} \sum_{j=1}^k \left[\left(\alpha_j'' + \beta_j''(it_1) \right) \sum_{u=0}^{l-1} a_u \sigma_j^u t_1^u + \gamma_j'' \sum_{u=0}^l b_u \sigma_j^u t_1^u \right] &\times \\ \exp(it_1 \theta_j' - \sigma_j \sqrt{2l-1} t_1) &= 0. \end{aligned} \tag{2.20}$$

Using the Laplace transformation on both sides of (2.20) and denoting $c_j = \sigma_j \sqrt{2l-1} - i\theta'_j$ as $1 \leq j \leq k$, we obtain that as $\text{Re}(s) > \max_{1 \leq j \leq k} \{-\sigma_j \sqrt{2l-1}\}$

$$\begin{aligned} \sum_{j=1}^k \alpha''_j \sum_{u=0}^{l-1} \frac{u! a_u \sigma_j^u}{(s + c_j)^{u+1}} + i \beta''_j \sum_{u=1}^l \frac{u! a_{u-1} \sigma_j^{u-1}}{(s + c_j)^{u+1}} &+ \\ \gamma''_j \sum_{u=0}^l \frac{u! b_u \sigma_j^u}{(s + c_j)^{u+1}} &= 0. \end{aligned} \quad (2.21)$$

Without loss of generality, we assume that $\sigma_1 \leq \sigma_2 \leq \dots \leq \sigma_k$. It demonstrates that $-\sigma_1 \sqrt{2l-1} = \max_{1 \leq j \leq k} \{-\sigma_j \sqrt{2l-1}\}$. Denote $a_u^j = a_u \sigma_j^u$ and $b_u^j = b_u \sigma_j^u$ for all u . By multiplying both sides of (2.21) with $(s + c_1)^{l+1}$, as $\text{Re}(s) > -\sigma_1 \sqrt{2l-1}$ and $s \rightarrow -c_1$, we obtain $|i \beta''_1 l! a_{l-1}^1 + \gamma''_1 b_l l! b_l^1| = 0$ or equivalently $\beta''_1 = \gamma''_1 = 0$ since $a_{l-1}^1, b_l^1 \neq 0$. Likewise, multiply both sides of (2.21) with $(s + c_1)^l$ and using the same argument, as $s \rightarrow -c_1$, we obtain $\alpha''_1 = 0$. Overall, we obtain $\alpha''_1 = \beta''_1 = \gamma''_1 = 0$. Continue in this fashion until we get $\alpha''_j = \beta''_j = \gamma''_j = 0$ for all $1 \leq j \leq k$ or equivalently $\alpha_j = \beta_j = \gamma_j = 0$ for all $1 \leq j \leq k$.

As a consequence, for all $1 \leq j \leq k$, we have

$$\alpha'_j + C_3 \left(\sum_{l=1}^d M_{ll}^j \right) - C_3 \left(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} \right) = 0, \quad \frac{(\Sigma_j^{1/2} t')^T \beta'_j}{\sigma_j} = 0,$$

and $\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} = 0$. Since we have

$$\sum_{u,v} M_{uv}^j [\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v = (t')^T \Sigma_j^{1/2} M_j \Sigma_j^{1/2} t' = (t')^T \gamma_j t',$$

it is equivalent that

$$\alpha'_j + C_3 \left(\sum_{l=1}^d M_{ll}^j \right) = 0, \quad (t')^T \Sigma_j^{1/2} \beta'_j = 0, \quad \text{and} \quad (t')^T \gamma_j t' = 0.$$

By the same argument as that of part (a), we readily obtain that $\alpha'_j = 0$, $\beta'_j = 0 \in \mathbb{R}^d$, and $\gamma_j = 0 \in \mathbb{R}^{d \times d}$. From the formation of α'_j, β'_j , it follows that $\alpha_j = 0$, $\beta_j = 0 \in \mathbb{R}^d$, and $\gamma_j = 0 \in \mathbb{R}^{d \times d}$ for all $1 \leq j \leq k$. We achieve the conclusion of part (b) of our theorem.

CHAPTER III

Convergence rates of parameter estimation for some weakly identifiable finite mixtures

We establish minimax lower bounds and maximum likelihood convergence rates of parameter estimation for mean-covariance multivariate Gaussian mixtures, shape-rate Gamma mixtures, and some variants of finite mixture models, including the setting where the number of mixing components is bounded but unknown. These models belong to what we call "weakly identifiable" classes, which exhibit specific interactions among mixing parameters driven by the algebraic structures of the class of kernel densities and their partial derivatives. Accordingly both the minimax bounds and the maximum likelihood parameter estimation rates in these models, obtained under some compactness conditions on the parameter space, are shown to be typically much slower than the usual $n^{-1/2}$ or $n^{-1/4}$ rates of convergence.¹

3.1 Introduction

Location-scale Gaussian mixtures are one of the most widely utilized modeling tools in statistics. Shape-rate Gamma mixtures are also a useful modeling choice for non-negative valued data. Yet convergence behaviors of the parameters arising

¹This work has been published in [Ho and Nguyen, 2016a].

in these model classes remain largely open questions [Lindsay, 1995, McLachlan and Basford, 1988, DasGupta, 2008]. We seek to address these questions in this chapter.

For finite mixtures of Gaussians, some facts are known when only one type of parameter varies (such as the mean/location or the variance/scale but not both). Specifically, if the number of mixing components generating the data is given, then the optimal rate of parameter estimation is the standard $n^{-1/2}$, where n is the sample size. If the number of mixing components is unknown but bounded by a known constant, then the convergence rate $n^{-1/4}$ for estimating the mixing distribution is achieved by a procedure established by [Chen, 1995]. For multi-dimensional parameters, the $(\log n/n)^{1/4}$ rate of posterior concentration of the mixing distribution was established by [Nguyen, 2013], under Wasserstein distance W_2 . [Ho and Nguyen, 2016c] extended the results of Chen [1995] and Nguyen [2013] to a broader range of *strongly* identifiable models, which admit general rates for the mixing measure under maximum likelihood estimation (MLE): $(\log n/n)^{1/2}$ for exact-fitted mixtures under W_1 metric, and $(\log n/n)^{1/4}$ for over-fitted finite mixtures under W_2 metric.

Strong identifiability and related notions, as studied by Chen [1995], Nguyen [2013] and several others (e.g., [Liu and Shao, 2004, Rousseau and Mengersen, 2011]), refers to a linear independence condition on the class of kernel density functions and their first and second-order partial derivatives with respect to the parameters. It is fruitful to delineate this condition further: first-order identifiability requires linear independence of the density functions and their first-order derivatives; second-order identifiability requires linear independence of the density functions and their partial derivatives up to the second order [Ho and Nguyen, 2016c]. The classical identifiability condition — linear independence of the class of density functions — corresponds to zero-order identifiability. Gaussian mixtures with both the mean and covariance parameters varying are identifiable up to the first order, but *not* in the second-order. Gamma mixtures are not identifiable even in the first-order, despite being identifiable

in the classical sense. In each of these examples, the violation of such identifiability conditions is due to a specific interaction among different parameters being present in the model class. Such interactions are driven by specific algebraic structures of the class of kernel densities and their partial derivatives. They can be succinctly expressed by certain partial differential equations satisfied by the kernel density function.

We shall informally refer to those finite mixture models *weakly identifiable* if they fail either the first or second-order identifiability condition, but otherwise are identifiable in the classical sense. Most relevant existing works on the asymptotics of parameter estimation (e.g., [Chen \[1995\]](#), [Nguyen \[2013\]](#), [Ho and Nguyen \[2016c\]](#)) concern only the settings of strong identifiability, and thus quite inapplicable to weakly identifiable classes. In fact, for such model classes the standard rates of convergence $n^{-1/2}$ and $n^{-1/4}$ (modulo a logarithmic term) no longer hold in general — the rates that we establish in this chapter are non-standard, and new. For instance, we shall show that for a location-scale Gaussian mixture where the number of mixing components is unknown and bounded by a constant, a minimax lower bound and the MLE convergence rate for estimating the mixing measure depend on how much we potentially overfit the model: the estimation rate is $n^{-1/8}$ under the 4th order Wasserstein distance W_4 , if overfitting by one extra component; $n^{-1/12}$ under the 6th order Wasserstein distance W_6 if overfitting by two extra components. All these rates occur while the MLE convergence rate of the mixture density remains to be $n^{-1/2}$. Remarkably, for Gamma and some other mixtures, the minimax lower bound for estimating the mixing measure is shown to be worse than *any* polynomial rate of the form $n^{-1/r}$ even when the number of mixing components is known.

In the special case of overfitting location-scale Gaussian mixtures by one extra component, the poor convergence rate for parameter estimation has been noted before by several authors. Most notably, [Chen and Chen \[2003\]](#) established the convergence rate $n^{-1/8}$ of the mixing distribution under a hypothesis testing for homogeneity.

Kasahara and Shimotsu [[Kasahara and Shimotsu, 2014b](#)] also achieved the rate $n^{-1/8}$ of MLE of finite normal regression mixtures (overfitted by one more component) when parameters are reparameterized and mixing proportions are restricted to be bounded away from zero. We are not aware of existing work on Gamma mixtures.

3.1.1 Main results for Gaussian mixtures

Given an n -iid sample X_1, \dots, X_n generated according to a Gaussian mixture density $p_{G_0}(x) = \int f(x|\theta, \Sigma) G_0(d\theta, d\Sigma)$, where $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, \Sigma_i^0)}$ has $k_0 \geq 1$ distinct support points. The class of Gaussian densities is denoted by $\{f(x|\theta, \Sigma), \theta \in \Theta \subset \mathbb{R}^d, \Sigma \in \Omega \subset S_d^{++}\}$, where S_d^{++} indicates the set of all symmetric positive definite matrices on $\mathbb{R}^{d \times d}$ and $d \geq 1$. Throughout this chapter, Θ and Ω shall be restricted to be compact subsets where their precise formations are given in our main theorems. (We note that without these compactness conditions, the MLE of G_0 may not exist or be inconsistent.) Now, we shall fit a mixture of k Gaussian distributions using the n -sample, where $k \geq k_0 + 1$. Denote by $\mathcal{O}_k := \mathcal{O}_k(\Theta \times \Omega)$ the set of probability measures on $\Theta \times \Omega$ with at most k support points, $\mathcal{E}_{k_0} := \mathcal{E}_{k_0}(\Theta \times \Omega)$ the set of probability measures on $\Theta \times \Omega$ with exactly k_0 support points. In addition, given $c_0 \in [0, 1)$, define a subset of \mathcal{O}_k ,

$$\mathcal{O}_{k,c_0} := \left\{ G = \sum_{i=1}^{k^*} p_i \delta_{(\theta_i, \Sigma_i)} \in \mathcal{O}_k : p_i \geq c_0 \ \forall 1 \leq i \leq k^* \right\}.$$

Let \widehat{G}_n be an estimate of G_0 . We seek to derive the rate of convergence of \widehat{G}_n to G_0 under a number of settings. For evaluating the convergence of mixing measures, Wasserstein distances have proved to be a natural choice [[Nguyen, 2013, 2016](#)]. Given two discrete probability measures $G = \sum_{i=1}^k p_i \delta_{(\theta_i, \Sigma_i)}$ and $G' = \sum_{i=1}^{k'} p'_i \delta_{(\theta'_i, \Sigma'_i)}$ on $\Theta \times \Omega$, recall that the s -th ($s \geq 1$) order Wasserstein distance between G and G' takes the

form [Villani, 2009]:

$$W_s(G, G') = \left(\inf \sum_{i,j} q_{ij} (\|\theta_i - \theta'_j\| + \|\Sigma_i - \Sigma'_j\|)^s \right)^{1/s},$$

where the infimum is taken over all couplings \mathbf{q} between \mathbf{p} and \mathbf{p}' , i.e., $\mathbf{q} = (q_{ij})_{ij} \in [0, 1]^{k \times k'}$ such that $\sum_{i=1}^k q_{ij} = p'_j$ and $\sum_{j=1}^{k'} q_{ij} = p_i$ for any $i = 1, \dots, k$ and $j = 1, \dots, k'$. In addition, $\|\cdot\|$ denotes either the ℓ_2 norm for elements in \mathbb{R}^d or the entrywise ℓ_2 norm for matrices.

To see how a convergence rate in Wasserstein distance W_s is translated to that of the parameters, suppose that a sequence of mixing measures G_n tending to G_0 under W_s metric at a rate $\omega_n = o(1)$. If all G_n have the same number of atoms $k = k_0$ as that of G_0 , then the set of atoms of G_n converge to the k_0 atoms of G_0 , up to a permutation of the atoms, at the same rate ω_n under $\|\cdot\|$ metric. If G_n have varying $k_n \in [k_0, k]$ number of atoms, where k is a fixed upper bound, then a subsequence of G_n can be constructed so that each atom of G_0 is a limit point of a certain subset of atoms of G_n — the convergence to each such limit also happens at rate ω_n . Some atoms of G_n may have limit points that are not among G_0 's atoms — the total mass associated with those “redundant” atoms of G_n must vanish at the generally faster rate ω_n^s .

For over-fitted Gaussian mixtures with both mean and variance varying, a main result of this chapter is to show that the rate of convergence of the mixing measure is determined by the order of a set of polynomial equations, which we now describe precisely. Denote by $\bar{r} \geq 1$ the *minimum* value of $r \geq 1$ such that the following system of polynomial equations

$$\sum_{j=1}^{k-k_0+1} \sum_{n_1, n_2} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0 \quad \text{for each } \alpha = 1, \dots, r \tag{3.1}$$

does *not* have any non-trivial solution for the unknowns $(a_j, b_j, c_j)_{j=1}^{k-k_0+1}$. The ranges of n_1, n_2 in the second sum are all natural pairs satisfying $n_1 + 2n_2 = \alpha$. A solution is considered non-trivial if all of c_j s are non-zeros, while at least one of the a_j s is non-zero.

Theorem 3.1.1. (Gaussian mixtures) *Let $L, \gamma, \lambda < \bar{\lambda}$ be fixed positive numbers. Given $\Theta = [-a_n, a_n]^d$ where $a_n \leq L(\log n)^\gamma$, and Ω be a subset of S_d^{++} whose eigenvalues are bounded in an interval $[\underline{\lambda}, \bar{\lambda}]$.*

(a) (*Minimax lower bound*) *For any $r < 2\bar{r}$,*

$$\inf_{\widehat{G}_n \in \mathcal{O}_k} \sup_{G \in \mathcal{O}_k \setminus \mathcal{O}_{k_0}} E_{p_G} W_1(\widehat{G}_n, G) \geq c_1 n^{-1/r}.$$

Here, the infimum is taken over all sequences of estimates \widehat{G}_n ranging in \mathcal{O}_k , E_{p_G} denotes the expectation taken with respect to product measure with mixture density p_G^n , c_1 is a universal positive constant.

(b) (*Maximum likelihood estimation*) *Let $c_0 = 0$ if $k - k_0 = 1$ or 2, and $c_0 > 0$ otherwise. Assume that $G_0 \in \mathcal{O}_{k,c_0}$ and let \widehat{G}_n be the MLE ranging in \mathcal{O}_{k,c_0} . Then,*

$$\mathbb{P}(W_{\bar{r}}(\widehat{G}_n, G_0) > C(\log n/n)^{1/(2\bar{r})}) \lesssim \exp(-c \log n).$$

Here, probability \mathbb{P} is taken with respect to p_{G_0} . C, c are positive constants depending only on $d, L, \gamma, \lambda, \bar{\lambda}, c_0$ and G_0 .

Part (a) of Theorem 3.1.1 establishes a minimax lower bound for estimating mixing measure G under W_1 distance. Noting the general inequality $W_{\bar{r}} \geq W_1$, this lower bound obviously also holds for $W_{\bar{r}}$. In words, when the number of mixing components is unknown except that it lies in the interval $[k_0, k]$, then there is no method for estimating G at a rate better than $n^{-1/(2\bar{r})}$, uniformly for all $G \in \mathcal{O}_k \setminus \mathcal{O}_{k_0}$. The proof actually obtains something stronger: the lower bound holds uniformly for any fixed or

suitably shrinking W_1 neighborhood in \mathcal{O}_k of any $G_0 \in \mathcal{E}_{k_0}$. Part (b) of Theorem 3.1.1 establishes that, under the compactness of the parameter spaces Θ, Ω , the rate $n^{-1/(2\bar{r})}$ can be achieved, up to a logarithmic term $\log n$, by maximum likelihood estimation. We wish to emphasize that this is a pointwise convergence rate, i.e., constant C depends on G_0 . For a fixed G_0 , we do not know if the upper bound $n^{-1/(2\bar{r})}$ of the convergence rate for the MLE may still be improved without additional assumptions or not. As a consequence of part (a), the upper bound $n^{-1/(2\bar{r})}$ is sharp in the sense that it cannot be improved uniformly for any W_1 neighborhood for G_0 .

The link of the estimation rate for location-scale Gaussian mixtures to the solvability of the system of polynomial equations (4.24) established by the above theorem is rather striking, as it describes precisely the hardness of parameter estimation in over-fitted situations. Determining the solvability of a system of polynomial equations is a basic question in (computational) algebraic geometry. For system (4.24), there does not seem to be an obvious answer as to the general value of \bar{r} . Since the number of variables in this system is $3(k - k_0 + 1)$, one expects that \bar{r} keeps increasing as $k - k_0$ increases. Using a standard method of Groebner bases [Buchberger, 1965], we can show that for $k - k_0 = 1$ and 2, $\bar{r} = 4$ and 6, respectively. In addition if $k - k_0 \geq 3$, then $\bar{r} \geq 7$. Thus, the convergence rate of the mixing measure for Gaussian mixtures deteriorates rapidly as more extra components are included in the model. We expect, but do not have a proof, that the value \bar{r} in the rate $n^{-1/2\bar{r}}$ tends to infinity as the number of redundant Gaussian components increases to infinity. We note several recent results at the other end of the rate spectrum: when the number of mixing components is unbounded (infinite), the convergence rate of the mixing measure under W_2 is shown to be $(\log n)^{-1/2}$ for the location Gaussian mixtures [Caillerie et al., 2011, Nguyen, 2013]. This rate may also resonate with some classical results in the deconvolution literature (e.g. [Zhang, 1990, Fan, 1991]), but one should be reminded that these classical results are applicable to only location mixtures carry-

ing smooth mixing densities. Interestingly, although the convergence rate of mixing measures in over-fitted finite mixtures may be poor, if one is interested in mixing proportions only, it follows from the previous discussion of Wasserstein distance $W_{\bar{r}}$ that the rate $(n^{-1/(2\bar{r})})^{\bar{r}} = n^{-1/2}$ is still achieved by the MLE. This rate is also obtained by a Bayesian estimation procedure studied by [Rousseau and Mengersen, 2011].

3.1.2 Results for other weakly identifiable classes

We now briefly describe other model classes studied in this chapter. Gamma densities represent an interesting instance: the Gamma density $f(x|a,b)$ has two positive parameters, a for shape and b for rate. This family is not identifiable in the first order. Moreover, we will show that there are particular combinations of the true parameter values which prevent the Gamma class from enjoying strong convergence properties. One the other hand, by excluding the measure-zero set of pathological cases of true mixing measures, the Gamma density class in fact can be shown to be strongly identifiable in both orders. Thus, this class is *almost* strongly identifiable, using the terminology of [Allman et al., 2009]. The generic/pathological dichotomy in the convergence behavior within the Gamma class is quite interesting: in the generic case of true mixing measures, the mixing measure can be estimated at the standard rate (i.e., $n^{-1/2}$ under W_1 for exact-fitted and $n^{-1/4}$ under W_2 for over-fitted mixtures). The pathological cases are very unforgiving: even for exact-fitted mixtures, one can do no better than a logarithmic rate of convergence in a minimax sense.

Let some readers wonder whether this unusually slow rate for the exact-fitted mixture setting can happen only in the measurably negligible (pathological) cases, we also introduce a location-extension of the exponential distribution, the location-exponential class: $f(x|\theta, \sigma) := \frac{1}{\sigma} \exp(-\frac{x-\theta}{\sigma}) 1(x > \theta)$. We show that the minimax lower bound for estimating the mixing measure in an location-exponentials is no faster than a logarithmic rate, even when the number of mixing component is known.

Practical implications In theory, mixture models enjoy strong asymptotic properties as a black-box modeling device for density estimation, see [Genovese and Wasserman \[2000\]](#), [Ghosal and van der Vaart \[2001\]](#), [Rousseau \[2010\]](#), [Kruijer et al. \[2010\]](#) and the references therein. In practice, the parameters specific to each mixing components may carry useful information about the heterogeneity among the underlying (latent) subpopulations. Thus, understanding the statistical efficiency of parameter estimation in mixture modeling is also relevant from a practical standpoint. Problematic convergence behaviors exhibited by widely utilized models such as Gaussian mixtures may have long been observed in practice, but a concrete theory has been largely unavailable. The results established in this chapter present a cautionary tale about the limitation of Gaussian mixtures, when it comes to assessing the quality of parameter estimation, but only when the number of mixing components is unknown. Since a tendency in practice is to "over-fit" the mixture generously with many more extra mixing components, our theory warns against this because as we have shown, the convergence rate via standard methods such as MLE for subpopulation-specific parameters deteriorates rapidly with the number of redundant components. For Gamma and location-exponential distribution, our theory also paints wildly varied convergence behaviors within each model class and thus a similarly extreme caution. We hope that the theoretical results obtained may hint at practically useful ways for determining benign scenarios and imposing helpful constraints when the mixture models enjoy strong identifiability properties and favorable convergence rates, and for identifying pathological scenarios where the practitioners would do well by avoiding them.

Chapter organization Section 3.2 is devoted to the proof of the results for Gaussian mixture models. Section 3.3 investigates Gamma mixtures and a location extension of exponential distribution. The theoretical bounds are illustrated via simulations in Section 3.4. Remaining proofs are given in Section 3.5.

Notation In addition to Wasserstein distances for mixing measures, we also utilize several familiar notions of distance for mixture densities, with respect to Lebesgue measure. They are total variation distance $V(p_G, p_{G'}) = \frac{1}{2} \int |p_G(x) - p_{G'}(x)| d\mu(x)$ and Hellinger distance $h^2(p_G, p_{G'}) = \frac{1}{2} \int (\sqrt{p_G(x)} - \sqrt{p_{G'}(x)})^2 d\mu(x)$.

3.2 Proof of main results for Gaussian mixtures

This section is devoted to proving Theorem (3.1.1). This theorem addresses only over-fitted Gaussian mixtures, i.e., when the true number of mixing components is bounded but otherwise unknown. If the number of mixing Gaussian components is known, it was already shown that the rate of estimating the mixing measure G is the standard rate $n^{-1/2}$ under W_1 metric [Ho and Nguyen, 2016c]. This is due to the fact that the class of Gaussian densities with both mean and covariance parameters varying is identifiable in the first order. However, the Gaussian family is not identifiable in the second order — that is to say that the collection of Gaussian density functions and their partial derivatives up to the second order taken with respect to the mean and covariance parameters are *not* linearly independent. This can be seen by the following identity, which represents a partial differential equation satisfied by Gaussian density $f(x|\theta, \Sigma)$:

$$\frac{\partial^2 f}{\partial \theta^2}(x|\theta, \Sigma) = 2 \frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma). \quad (3.2)$$

This identity, also noted previously by Chen and Chen [2003], Kasahara and Shimotsu [2014b], will play a fundamental role in our proof of Theorem (3.1.1).

3.2.1 On the order \bar{r}

Before proceeding to the proof of the theorem, let us briefly discuss some properties of \bar{r} as defined in (4.24). This is a system of r polynomial equations with $3(k - k_0 + 1)$ unknowns. The condition $c_1, \dots, c_{k-k_0+1} \neq 0$ is important. In fact, if $c_1 = 0$, then by

choosing $a_1 \neq 0$, $a_i = 0$ for all $i = 2, \dots, k-k_0+1$ and $b_j = 0$ for all $j = 1, \dots, k-k_0+1$, we can check that system (4.24) is satisfied for all $\alpha \geq 1$. Therefore, without this condition, \bar{r} does not exist.

To illustrate the possible values of \bar{r} , let us consider the case $k = k_0 + 1$, and let $r = 3$. System (4.24) reduces to the equations:

$$\begin{aligned} c_1^2 a_1 + c_2^2 a_2 &= 0 \\ \frac{1}{2}(c_1^2 a_1^2 + c_2^2 a_2^2) + c_1^2 b_1 + c_2^2 b_2 &= 0 \\ \frac{1}{3!}(c_1^2 a_1^3 + c_2^2 a_2^3) + c_1^2 a_1 b_1 + c_2^2 a_2 b_2 &= 0. \end{aligned}$$

It is simple to see that a non-trivial solution exists, by choosing $c_2 = c_1 \neq 0$, $a_1 = 1$, $a_2 = -1$, $b_1 = b_2 = -1/2$. Hence, $\bar{r} \geq 4$. For $r = 4$, the system consists of the three equations given above, plus

$$\frac{1}{4!}(c_1^2 a_1^4 + c_2^2 a_2^4) + \frac{1}{2!}(c_1^2 a_1^2 b_1 + c_2^2 a_2^2 b_2) + \frac{1}{2!}(c_1^2 b_1^2 + c_2^2 b_2^2) = 0.$$

It will be shown in the sequel that this system has no non-trivial solution. Therefore for $k = k_0 + 1$, we have $\bar{r} = 4$.

Determining the exact value of \bar{r} in the general case appears quite challenging. For the specific value of $k - k_0$, one can find \bar{r} — there are well-developed methods in computational algebra for dealing with this type of polynomial equations, such as Groebner bases [Buchberger, 1965] and resultants [Sturmfels, 2002]. Using the Groebner bases method, we shall show in Section 3.5 that

Proposition 3.2.1. $\bar{r} = 4$ if $k = k_0 + 1$, $\bar{r} = 6$ if $k = k_0 + 2$. If $k \geq k_0 + 3$, then $\bar{r} \geq 7$.

3.2.2 Discussion of conditions in Theorem 3.1.1

The main conditions in the statement of Theorem 3.1.1 are concerned with compactness and boundedness of the mixture model's parameters, including the parameters of mixing components, and the parameters for mixing probabilities.

The parameters of mixing components lie in Ω and Θ . Compactness conditions for Ω and Θ are required for three reasons. First, the compactness of Ω is important in guaranteeing that the likelihood function is bounded. Indeed, if the smallest eigenvalue of the covariance parameter is not bounded below or the largest eigenvalue of the covariance parameter is not bounded above, the likelihood function will become unbounded [Day, 1969, Hathaway, 1985, Chen and Li, 2009]. Second, the compactness of Θ and Ω are also crucial in obtaining upper bounds of the (bracket) entropies that we need for Lemma 3.2.1. Such bounds yield convergence rate $n^{-1/2}$, up to logarithmic factor, for the convergence of mixture density p_G under Hellinger distance. Third, and most importantly, these compactness assumptions are required in establishing the lower bounds of Hellinger distance of mixture densities in terms of Wasserstein distance of mixing measures (cf. Proposition 3.2.2), thereby allowing us to translate the convergence rate of the mixture density into that of the corresponding mixing measure. Our proof technique hinges upon the compactness conditions. As pointed out by the referees, one may be able to relax somewhat the compactness assumptions by penalizing the likelihood function appropriately [Chen et al., 2008, Chen and Tan, 2009]. While the first two issues discussed above may still be addressed, the third issue will require a substantially new proof technique; moreover, the rate of convergence will be likely different.

It is required in part (b) of the theorem that \widehat{G}_n range in \mathcal{O}_{k,c_0} , where $c_0 > 0$ when $k - k_0 \geq 3$. This requirement is sufficient for establishing the bound in part (b) of Proposition 3.2.2. A consequence of this requirement is that it prevents the Fisher matrix at the masses from being degenerate [Chen and Li, 2009, Chen et al.,

2012, Kasahara and Shimotsu, 2014b]. As such, this condition is also crucial in obtaining the asymptotic distribution of parameter estimates. We note, however, that this requirement may not be necessary for the purpose of establishing rates of parameter estimation. In fact, when the Gaussian mixture is overfitted by at most two components, i.e., $1 \leq k - k_0 \leq 2$, it will be demonstrated by Proposition 3.2.3 that this requirement can be removed (by letting $c_0 = 0$) without affecting the conclusion of the theorem.

3.2.3 Sharp identifiability bounds

A central ingredient in the proof of Theorem 3.1.1 are sharp inequalities which relate the distance of two Gaussian mixture densities to a Wasserstein distance between corresponding mixing measures. Let $V(p_G, p_{G_0})$ denote the variational distance, and $h(p_G, p_{G_0})$ the Hellinger distance of p_G and p_{G_0} . The order \bar{r} enters the following bounds in an essential way:

Proposition 3.2.2. *Let \bar{r} be defined as above, and $G_0 \in \mathcal{E}_{k_0} \cap \mathcal{O}_{k_0, c_0}$ for some $c_0 > 0$.*

(a) *For any $1 \leq r < \bar{r}$, there holds:*

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_k} \left\{ h(p_G, p_{G_0}) / W_1^r(G, G_0) : W_1(G, G_0) \leq \epsilon \right\} = 0.$$

(b) *For any $G \in \mathcal{O}_{k, c_0}$ such that $W_{\bar{r}}(G, G_0)$ is sufficiently small, there holds:*

$$h(p_G, p_{G_0}) \geq V(p_G, p_{G_0}) \gtrsim W_{\bar{r}}^{\bar{r}}(G, G_0) \geq W_1^{\bar{r}}(G, G_0).$$

The proof of this proposition is deferred to Section 3.5. We make several remarks.

- (i) In part (a) the ratio h/W_1^r is set to ∞ if $W_1 = 0$. In part (b), the multiplying constant in \gtrsim bound depends only on G_0 .

(ii) Part (a) and part (b) together show that $W_{\bar{r}}^{\bar{r}}(G, G_0)$ is the sharp lower bound for the distance of mixture densities $V(p_G, p_{G_0})$. In particular, we cannot have $V \gtrsim W_1^r$ for any $r < \bar{r}$.

(iii) In part (b), G is restricted to a subset of \mathcal{O}_k , i.e., set \mathcal{O}_{k,c_0} , which places a lower bound constraint on the mixing probability mass. This restriction seems to be an artifact of our proof technique. It can be removed completely with some extra hard work, at least for the case $k - k_0 \leq 2$, as follows:

Proposition 3.2.3. *Let $k - k_0 = 1$ or 2 . Fix $G_0 \in \mathcal{E}_{k_0}$. For any $G \in \mathcal{O}_k$ such that $W_{\bar{r}}(G, G_0)$ is sufficiently small, we have $V(p_G, p_{G_0}) \gtrsim W_{\bar{r}}^{\bar{r}}(G, G_0)$.*

The proof of Proposition 3.2.3 is deferred to Section 3.5.3. Given the two propositions above, we can now complete the proof of Theorem 3.1.1.

3.2.4 Proof of Theorem 3.1.1

(a) The proof of this part follows from the same argument as that of Lemma 1 of Yu [1997] for establishing minimax lower bounds. Fix $r < \bar{r}$ and $G_0 \in \mathcal{E}_{k_0}$. Let $C_0 > 0$ be any fixed constant. According to part (a) of Proposition 3.2.2, for any sufficiently small $\epsilon > 0$, there exists $G'_0 \in \mathcal{O}_k$ such that $W_1(G_0, G'_0) = 2\epsilon$ and $h(p_{G_0}, p_{G'_0}) \leq C_0\epsilon^r$. Take any sequence of estimates \hat{G}_n ranging in \mathcal{O}_k , we have

$$2 \max_{G \in \{G_0, G'_0\}} E_{p_G} W_1(\hat{G}_n, G) \geq E_{p_{G_0}} W_1(\hat{G}_n, G_0) + E_{p_{G'_0}} W_1(\hat{G}_n, G'_0),$$

where $E_{p_{G_0}}$ (resp. $E_{p_{G'_0}}$) denotes the expectation taken with respect to the product measure with density $p_{G_0}^n$ ($p_{G'_0}^n$). By the triangle inequality, $W_1(\hat{G}_n, G_0) + W_1(\hat{G}_n, G'_0) \geq W_1(G_0, G'_0) = 2\epsilon$. Thus,

$$E_{p_{G_0}} W_1(\hat{G}_n, G_0) + E_{p_{G'_0}} W_1(\hat{G}_n, G'_0) \geq 2\epsilon \inf_{f_1, f_2} \left(E_{p_{G_0}} f_1 + E_{p_{G'_0}} f_2 \right),$$

where the infimum is taken over all measurable nonnegative functions f_1 and f_2 defined in terms of n arguments X_1, \dots, X_n , subject to the constraint that $f_1 + f_2 = 1$. From the definition of the variational distance, the infimum value in the above expression is equal to $(1 - V(p_{G_0}^n, p_{G'_0}^n))$. Hence,

$$\max_{G \in \{G_0, G'_0\}} E_{p_G} W_1(\hat{G}_n, G) \geq \epsilon \left(1 - V(p_{G_0}^n, p_{G'_0}^n)\right).$$

Now, due to the general relationship between variational distance and Hellinger distance, i.e., $V \leq h$, and by our construction that $h(p_{G_0}, p_{G'_0}) \leq C_0 \epsilon^r$, we have

$$\begin{aligned} V(p_{G_0}^n, p_{G'_0}^n) &\leq h(p_{G_0}^n, p_{G'_0}^n) \\ &= \sqrt{1 - (1 - h^2(p_{G_0}, p_{G'_0}))^n} \\ &\leq \sqrt{1 - (1 - C_0^2 \epsilon^{2r})^n}. \end{aligned}$$

As a result,

$$\max_{G \in \{G_0, G'_0\}} E_{p_G} W_1(\hat{G}_n, G) \geq \epsilon \left(1 - \sqrt{1 - (1 - C_0^2 \epsilon^{2r})^n}\right).$$

By choosing $\epsilon^{2r} = \frac{1}{C_0^2 n}$, the right hand side of the above inequality is bounded below by $c_1 \epsilon \asymp n^{-1/2r}$ for any $r < \bar{r}$ where c_1 is some positive universal constant. Noting that $G_0, G'_0 \in \mathcal{O}_k \setminus \mathcal{O}_{k_0-1}$, this concludes the proof for part (a).

(b) The proof follows from combining the result of part (b) of Proposition 3.2.2 with a standard result on convergence of density estimation via MLE, from [van de Geer, 2000]. To draw from the later, we first recall some additional standard notation from the empirical process theory literature (which after this proof will unfortunately not be needed for the rest of the chapter). Let $\Theta^* = \Theta \times \Omega$, $\mathcal{P}_k(\Theta^*) = \{p_G | G \in \mathcal{O}_k\}$. Let $N(\epsilon, \mathcal{P}_k(\Theta^*), \|\cdot\|_\infty)$ denote the covering number of the metric space $(\mathcal{P}_k(\Theta^*), \|\cdot\|_\infty)$, and $H_B(\epsilon, \mathcal{P}_k(\Theta^*), h)$ the bracketing entropy of $\mathcal{P}_k(\Theta^*)$ under Hellinger distance

metric h . Put $\bar{P}_k(\Theta^*) = \left\{ p_{\frac{G+G_0}{2}} : G \in \mathcal{O}_k \right\}$ and $\bar{\mathcal{P}}_k^{1/2}(\Theta^*) = \left\{ f^{1/2} | f \in \bar{\mathcal{P}}_k(\Theta^*) \right\}$. For any $\delta > 0$, denote the intersection of a Hellinger ball centered at p_{G_0} and $\bar{\mathcal{P}}_k^{1/2}(\Theta^*)$ as:

$$\bar{\mathcal{P}}_k^{1/2}(\Theta^*, \delta) = \left\{ f^{1/2} \in \bar{\mathcal{P}}_k^{1/2}(\Theta^*) | h(f, p_{G_0}) \leq \delta \right\}.$$

The size of this set is captured by the entropy integral:

$$\mathcal{J}_B(\delta, \bar{\mathcal{P}}_k^{1/2}(\Theta^*, \delta), \mu) = \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(u, \bar{\mathcal{P}}_k^{1/2}(\Theta^*, u), \mu) du \vee \delta,$$

where μ denotes Lebesgue measure. Since $\bar{\mathcal{P}}_k^{1/2}(\Theta^*, u) \subset \bar{\mathcal{P}}_k^{1/2}(\Theta^*)$, for any $u > 0$,

$$\begin{aligned} H_B(u, \bar{\mathcal{P}}_k^{1/2}(\Theta^*, u), L_2(\mu)) &\leq H_B(u, \bar{\mathcal{P}}_k^{1/2}(\Theta^*), L_2(\mu)) \\ &= H_B(u/\sqrt{2}, \bar{\mathcal{P}}_k(\Theta^*), h), \end{aligned} \quad (3.3)$$

where the identity is immediate from relationship between the Hellinger distance metric and $L_2(\mu)$.

Note that for any two mixing measures G_1, G_0 , $p_{(G_1+G_0)/2} = (p_{G_1} + p_{G_0})/2$. Note also the fact that for any probability densities f_0, f_1, f_2 defined on the same space, $h^2((f_1 + f_0)/2, (f_2 + f_0)/2) \leq h^2(f_1, f_2)/2$ (cf. Lemma 4.2 [van de Geer \[2000\]](#)). So, for any two mixing measures $G_1, G_2 \in \mathcal{O}_k$, we have

$$h^2(p_{\frac{G_1+G_0}{2}}, p_{\frac{G_2+G_0}{2}}) \leq h^2(p_{G_1}, p_{G_2})/2.$$

This inequality yields $H_B(u/\sqrt{2}, \bar{\mathcal{P}}_k(\Theta^*), h) \leq H_B(u, \bar{\mathcal{P}}_k(\Theta^*), h)$. Combining with Eq. (3.3) to obtain

$$H_B(u, \bar{\mathcal{P}}_k^{1/2}(\Theta^*, u), L_2(\mu)) \leq H_B(u, \bar{\mathcal{P}}_k(\Theta^*), h).$$

This inequality allows us to obtain an upper bound of the LHS in terms of a bound on the RHS. Specifically, we need the following

Lemma 3.2.1. *Suppose that $\Theta^* = [-a, a]^d \times \Omega$, where Ω is a subset of S_d^{++} whose eigenvalues are bounded in an interval $[\underline{\lambda}, \bar{\lambda}]$, $a \leq L(\log(1/\epsilon))^\gamma$, $\gamma \geq 1/2$, $L > 0$. Then for $0 < \epsilon < 1/2$,*

$$\log N(\epsilon, \mathcal{P}_k(\Theta^*), \|\cdot\|_\infty) \lesssim \log(1/\epsilon), \quad (3.4)$$

$$H_B(\epsilon, \mathcal{P}_k(\Theta^*), h) \lesssim \log(1/\epsilon). \quad (3.5)$$

The proof of this lemma is an extension of the arguments in [Ghosal and van der Vaart \[2001\]](#) to multivariate setting, and is deferred to Section 3.5.3. Now, we choose $L > 0$ and $\gamma_1 = \max \{1/2, \gamma\} \geq 1/2$ such that $a_n \leq L(\log(n))^{\gamma_1}$. From Lemma 3.2.1, as long as $0 < u < 1/2$, we have

$$H_B(u, \overline{\mathcal{P}}_k^{1/2}(\Theta^*, u), L_2(\mu)) \leq H_B(u, \mathcal{P}_k(\Theta^*), h) \lesssim \log(1/u). \quad (3.6)$$

Now, we state the result of Theorem 7.4 of [van de Geer \[2000\]](#) adapted to the notation used in our chapter

Theorem 3.2.1. *Take $\Psi(\delta) \geq \mathcal{J}_B(\delta, \overline{\mathcal{P}}_k^{1/2}(\Theta^*, \delta), \mu)$ in such a way that $\Psi(\delta)/\delta^2$ is a non-increasing function of δ . Then, for a universal constant c and for*

$$\sqrt{n}\delta_n^2 \geq c\Psi(\delta_n),$$

we have for all $\delta \geq \delta_n$

$$P(h(p_{\widehat{G}_n}, p_{G_0}) > \delta) \leq c \exp \left[-\frac{n\delta^2}{c^2} \right].$$

Based on the bracket entropy bound in (3.6), we can choose $\Psi(\delta) = \delta[\log(1/\delta)]^{1/2}$

for $\delta > 0$. Therefore, by choosing $\delta_n = O(\log n/n)^{1/2}$, we obtain $P(h(p_{\widehat{G}_n}, p_{G_0}) > \delta_n) \lesssim \exp(-c \log(n))$, where constant $c > 0$ depends only on $L, \gamma, \underline{\lambda}, \bar{\lambda}$. Combining this probability bound with part (b) of Proposition 3.2.2 concludes the proof.

3.3 Gamma mixtures and location extensions

The Gamma family of densities takes the form $f(x|a, b) := \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$ for $x > 0$, and 0 otherwise, where a, b are positive shape and rate parameters, respectively. The Gamma family is not identifiable in the first order when *both* shape and rate parameters vary—this is to say that the collection of Gamma density functions and their partial derivatives up to the first order taken with respect to the shape and rate parameters are *not* linearly independent. This can be seen by the following identity:

$$\frac{\partial f}{\partial b}(x|a, b) = \frac{a}{b} f(x|a, b) - \frac{a}{b} f(x|a + 1, b). \quad (3.7)$$

Examining the identity in the above display shows that the violation of linear independence of the collection of Gamma density functions and its derivatives is due to certain combinations of the Gamma parameter values. This suggests that outside of these value combinations the Gamma densities may well be identifiable in the first order and even the second order. This observation leads to a remarkable consequence for Gamma mixtures, which display wildly distinct behaviors in two disjoint categories of the parameter values, which we call “generic cases” and “pathological cases”.

Fix $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(a_i^0, b_i^0)} \in \mathcal{E}_{k_0} := \mathcal{E}_{k_0}(\Theta)$ where $k_0 \geq 2$ and $\Theta \subset \mathbb{R}_+^2$. Assume that $a_i^0 \geq 1$ for all $1 \leq i \leq k_0$. To delineate the structure underlying parameter values of G_0 , we define

(A.1) Generic cases: $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} \neq \{1, 0\}$ for all $1 \leq i, j \leq k_0$.

(A.2) Pathological cases: $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} = \{1, 0\}$ for some $1 \leq i, j \leq k_0$.

We have the following result under the exact-fitted setting of Gamma mixtures. Let $\widehat{G}_n \in \mathcal{E}_{k_0}$ denote the MLE estimate of G_0 .

Theorem 3.3.1. (Exact-fitted Gamma mixtures) *Given $\Theta = [\underline{a}, \bar{a}] \times [\underline{b}, \bar{b}]$ where $\underline{a} \geq 1, \bar{a}, \underline{b}, \bar{b}$ are given positive numbers.*

(a) Generic cases *If the support points of G_0 satisfy assumption (A.1), then*

$$\mathbb{P}(W_1(\widehat{G}_n, G_0) > \delta_n) \lesssim \exp(-c \log n),$$

where δ_n is sufficiently large multiple of $(\log n/n)^{1/2}$ and c is positive constant depending only on $\underline{a}, \bar{a}, \underline{b}, \bar{b}$.

(b) Pathological cases *For any $r \geq 2$,*

$$\inf_{\widehat{G}_n \in \mathcal{E}_{k_0}} \sup_{G \in \mathcal{E}_{k_0}} E_{p_G} W_r(\widehat{G}_n, G) \gtrsim n^{-1/r}.$$

While the result of part (a) may seem “obvious” due to the standard rate $(\log n/n)^{1/2}$, this should be put in the context of the minimax lower bound of part (b), which shows that one cannot estimate the Gamma parameters efficiently uniformly over a W_1 neighborhood of G_0 , when we do not know whether G_0 is pathological or not. As can be seen in the proof, the poor rate is due to the difficulty of distinguishing between the pathological and generic instances — no polynomial rate estimation method is possible.

Turning to the over-fitted Gamma mixture setting, as before let $G_0 \in \mathcal{E}_{k_0}$, while G varies in a larger subset of $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ for some given $k \geq k_0 + 1$. We have the following categories regarding the true G_0 :

$$(A.3) \text{ Generic cases: } \{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} \notin \left\{ \{1, 0\}, \{2, 0\} \right\} \text{ for all } 1 \leq i, j \leq k_0.$$

(A.4) Pathological cases: $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} \in \left\{ \{1, 0\}, \{2, 0\} \right\}$ for some $1 \leq i, j \leq k_0$.

Additionally, for any $c_0 > 0$ and $l \geq 1$, define the following constrained set of \mathcal{O}_l

$$\mathcal{O}_{l,c_0} = \left\{ G = \sum_{i=1}^{k'} p_i \delta_{(a_i, b_i)} \middle| \begin{array}{l} k' \leq k \text{ and } |a_i - a_j^0| \notin [1 - c_0, 1 + c_0] \\ \cup [2 - c_0, 2 + c_0] \forall (i, j) \end{array} \right\}.$$

Theorem 3.3.2. (Over-fitted Gamma mixtures) *Assume the same conditions on Θ as that of Theorem 3.3.1.*

(a) Generic cases If $G_0 \in \mathcal{O}_{k,c_0}$ and let $\hat{G}_n \in \mathcal{O}_{k,c_0}$ be the MLE estimation of G_0 , then $\mathbb{P}(W_2(\hat{G}_n, G_0) > \delta_n) \lesssim \exp(-c \log n)$, where δ_n is sufficiently large multiple of $(\log n/n)^{1/4}$ and c is positive constant depending only on $c_0, \underline{a}, \bar{a}, \underline{b}, \bar{b}$.

Moreover, the following minimax bound holds, for any $2 \leq r < 4$,

$$\inf_{\hat{G}_n \in \mathcal{O}_{k,c_0}} \sup_{G \in \mathcal{O}_{k,c_0} \setminus \mathcal{O}_{k_0-1}} E_{p_G} W_r(\hat{G}_n, G) \gtrsim n^{-1/r}.$$

(b) Pathological cases For any $r \geq 2$,

$$\inf_{\hat{G}_n \in \mathcal{O}_k} \sup_{G \in \mathcal{O}_k \setminus \mathcal{O}_{k_0-1}} E_{p_G} W_r(\hat{G}_n, G) \gtrsim n^{-1/r}.$$

Part (a) shows that in the over-fitted setting, if the true G_0 falls in the generic cases, then the standard MLE method restricted to a suitable subset of \mathcal{O}_k still yields the $(\log n/n)^{1/4}$ rate of convergence for the mixing measure. Outside of this category, however, one cannot hope to estimate G at any polynomial rate of convergence.

Not all is bad news for Gamma mixtures: since the pathological cases represent

a Lebesgue measure zero set, Gamma mixtures can be viewed as almost strongly identifiable with the strong convergence properties for the parameter estimation.

Exponential location extension Let the reader think that pathological cases are rare, we introduce a location extension of the exponential distribution, for which there is no such generic/pathological dichotomy. With this family, the convergence behavior of the mixing parameters is always slow, even when the number of mixing components is known. The class of location-exponential distribution $\{f(x|\theta, \sigma), \theta \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$ is defined as $f(x|\theta, \sigma) = \frac{1}{\sigma} \exp\left(-\frac{x-\theta}{\sigma}\right) \cdot 1_{\{x>\theta\}}$ for $x \in \mathbb{R}$. Direct calculation yields that

$$\frac{\partial f}{\partial \theta}(x|\theta, \sigma) = \frac{1}{\sigma} f(x|\theta, \sigma) \text{ when } x \neq \theta. \quad (3.8)$$

Since this identity holds in general, the linear independence of the kernel densities f and their partial derivatives is clearly violated regardless of the true values of G_0 . We shall state a result for the exact-fitted setting only. Let $\Theta = [-a, a]$ and $\Omega = [\underline{\sigma}, \bar{\sigma}]$ where $a, \underline{\sigma}, \bar{\sigma}$ are fixed positive constants.

Theorem 3.3.3. (Exact-fitted location-exponential mixtures) *For any $r \geq 2$,*

$$\inf_{\widehat{G}_n \in \mathcal{E}_{k_0}} \sup_{G \in \mathcal{E}_{k_0}} E_{p_G} W_1(\widehat{G}_n, G) \gtrsim n^{-1/r}.$$

This is quite a surprising bound, especially considering this is a finite mixture model with the known number of mixing components k_0 . Yet, one cannot hope to achieve a polynomial estimation rate uniformly over a neighborhood (in W_1) of any mixing measure G_0 . As in the pathological cases of Gamma mixtures, the poor convergence behavior of parameter estimation is due to the interaction of mixing parameters θ and σ , which is induced by the algebraic structures of f and its partial derivatives. As can be observed from the proof, the algebraic structure makes it

difficult to distinguish between mixing measures G carrying similar mixture densities.

3.4 Simulations

We illustrate via simulations the rich spectrum of convergence behaviors for weak identifiable classes. Both identifiability bounds $h \geq V \gtrsim W_r^r$, and the convergence behavior of the MLE are examined.

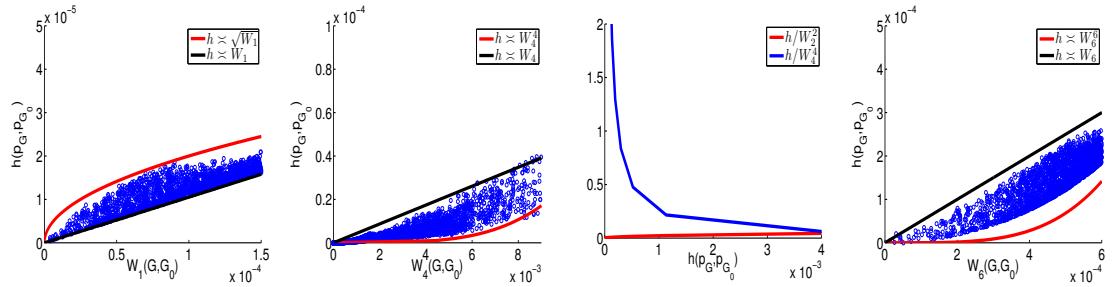


Figure 3.1: Location-scale Gaussian mixtures. From left to right: (1) Exact-fitted setting; (2) Over-fitted by one component; (3) Over-fitted by one component; (4) Over-fitted by two components.

Weak identifiability bounds We experiment with classes of Gaussian densities. The results for mixtures of location-scale Gaussian distributions are given in Figure 3.1. Simulation details are as follows. The true mixing measure G_0 has exactly $k_0 = 2$ support points with locations $\theta_1^0 = -2$, $\theta_2^0 = 4$, scales $\sigma_1^0 = 1$, $\sigma_2^0 = 2$, and $p_1^0 = 1/3, p_2^0 = 2/3$. 5000 random samples of discrete mixing measures $G \in \mathcal{E}_2$, 5000 samples of $G \in \mathcal{O}_3$ and another 5000 for $G \in \mathcal{O}_4$, where the support points are uniformly generated in $\Theta = [-10, 10]$ and $\Omega = [0.5, 5]$. Additionally, to illustrate the best lower bound W_4^4 when we overfit by one point, we also generate sequence G in accordance with the construction of sequence G in the proof of part (a) of Proposition 3.2.2. The ratios h/W_2^2 and h/W_4^4 are plotted in the third panel of Figure 3.1 to verify that $h \gtrsim W_4^4$ holds, but $h \gtrsim W_2^2$ does not. It can be observed that both the lower bounds and upper bounds are in agreement with the theorems established earlier.

Convergence rates of MLE First, we generate n -iid samples from a bivariate location-covariance Gaussian mixture with three components with an arbitrarily fixed choice of G_0 . The true parameters for the mixing measure G_0 are: $\theta_1^0 = (0, 3)$, $\theta_2^0 = (1, -4)$, $\theta_3^0 = (5, 2)$, $\Sigma_1^0 = \begin{pmatrix} 4.2824 & 1.7324 \\ 1.7324 & 0.81759 \end{pmatrix}$, $\Sigma_2^0 = \begin{pmatrix} 1.75 & -1.25 \\ -1.25 & 1.75 \end{pmatrix}$, $\Sigma_3^0 = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$, and $p_1^0 = 0.3$, $p_2^0 = 0.4$, $p_3^0 = 0.3$. MLE \hat{G}_n are obtained by the EM algorithm as we assume that the data come from a mixture of k Gaussians where $k \geq k_0 = 3$. See Figure 3.2 for a fixed choice of G_0 . Wasserstein distances between \hat{G}_n and G_0 are plotted against increasing sample size n . The error bars were obtained by running the experiment 7 times for each n . These simulation results match quite well with the established rates and highlight that convergence slows down rapidly as $k - k_0$ increases.

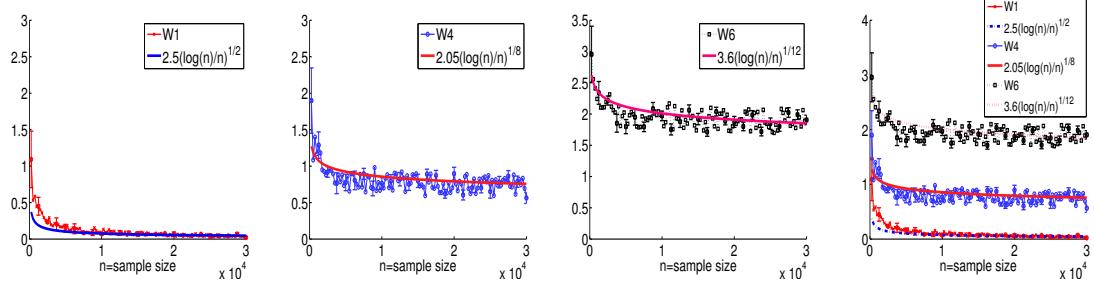


Figure 3.2: MLE rates for location-covariance mixtures of Gaussians. L to R: (1) Exact-fitted: $W_1 \asymp n^{-1/2}$. (2) Over-fitted by one: $W_4 \asymp n^{-1/8}$. (3) Over-fitted by two: $W_6 \asymp n^{-1/12}$.

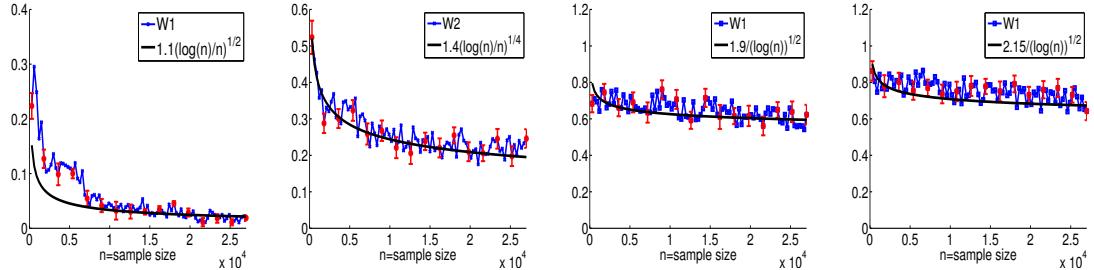


Figure 3.3: MLE rates for shape-rate mixtures of Gamma distributions. L to R: (1) Generic/Exact-fitted: $W_1(\hat{G}_n, G_0) \asymp n^{-1/2}$. (2) Generic/Over-fitted: $W_2 \asymp n^{-1/4}$. (3) Pathological/Exact-fitted: $W_1 \approx 1/(\log n)^{1/2}$. (4) Pathological/Over-fitted: $W_1 \approx 1/(\log n)^{1/2}$.

We turn to mixtures of Gamma distributions. For generic cases, we generate n -iid samples from a Gamma mixture model that has exactly two mixing components. The true parameters for the mixing measure G_0 are: $a_1^0 = 8$, $a_2^0 = 2$, $b_1^0 = 3$, $b_2^0 = 4$, $p_1^0 = 1/3$, $p_2^0 = 2/3$. For pathological cases, everything else remains the same, except for our choice of G_0 , for which we choose $a_1^0 = 8$, $a_2^0 = 7$, $b_1^0 = 3$, $b_2^0 = 3$, $p_1^0 = 1/3$, $p_2^0 = 2/3$.

It is remarkable to see the wild swing in behaviors within this same class. See Figure 3.3. Even for exact-fitted finite mixtures of Gamma, one can achieve very fast convergence rate of $n^{-1/2}$ in the generic case, or appear to be stagnant at a logarithmic rate if the true mixing measure G_0 belongs to the pathological category.

3.5 Proofs of other propositions and theorems

3.5.1 Proofs for over-fitted Gaussian mixtures

PROOF OF PROPOSITION 3.2.2 For the ease of exposition, we consider the setting of univariate location-scale Gaussian distributions, i.e., both θ and $\Sigma = \sigma^2$ are scalars. The proof for general $d \geq 1$ is similar and omitted. Put $v = \sigma^2$, so we write $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, v_i^0)}$.

Step 1 For any sequence $G_n \in \mathcal{O}_k$, since k is finite, there is some $k^* \in [k_0, k]$ such that there exists a subsequence of G_n having exactly k^* support points. Denote $G_n = \sum_{i=1}^{k^*} p_i^n \delta_{(\theta_i^n, v_i^n)}$ (here, without loss of generality, we replace the whole sequence by its subsequence). Now if $G_n \rightarrow G_0$ in W_r , there exists a subsequence of G_n such that each support point (θ_i^0, v_i^0) of G_0 is the limit of a subset of $s_i \geq 1$ support points of G_n . In general there may also a subset of support points of G_n whose limits are not among the support points of G_0 .

Note that with part (a), we shall construct one sequence of G_n to prove its conclusion. In our construction there are no constraints placed on p_i^n for all i . On the

other hand, regarding part (b), we shall impose the constraint that $p_i^n \geq c_0$ for all i . Under this constraint, all the limit points of support points of G_n will be only those of G_0 . To avoid notational cluttering, we replace the subsequence of G_n by the whole sequence $\{G_n\}$. By re-labeling the support points, G_n can be expressed by

$$G_n = \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(\theta_{ij}^n, v_{ij}^n)}, \quad (3.9)$$

where $(\theta_{ij}^n, v_{ij}^n) \rightarrow (\theta_i^0, v_i^0)$, $\sum_{l=1}^{s_i} p_{il}^n \rightarrow p_i^0$ for all $i = 1, \dots, k_0$ and $j = 1, \dots, s_i$, where s_1, \dots, s_{k_0} are some natural constants less than k . All G_n have exactly the same $k^* = \sum s_i \leq k$ number of support points. This is the representation for G_n that we shall utilize in the proof of both part (a) and part (b).

Step 2 For any $x \in \mathbb{R}$,

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n (f(x|\theta_{ij}^n, v_{ij}^n) - f(x|\theta_i^0, v_i^0)) + \\ &\quad \sum_{i=1}^{k_0} (p_i^n - p_i^0) f(x|\theta_i^0, v_i^0), \end{aligned}$$

where $p_i^n := \sum_{j=1}^{s_i} p_{ij}^n$. For any $r \geq 1$, integer $N \geq r$ and $x \in \mathbb{R}$, by means of Taylor expansion up to the order N , we obtain

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \sum_{|\alpha|=1}^N (\Delta \theta_{ij}^n)^{\alpha_1} (\Delta v_{ij}^n)^{\alpha_2} \frac{D^{|\alpha|} f(x|\theta_i^0, v_i^0)}{\alpha!} + \\ &\quad A_1(x) + R_1(x). \quad (3.10) \end{aligned}$$

Here, $\alpha = (\alpha_1, \alpha_2)$, $|\alpha| = \alpha_1 + \alpha_2$, $\alpha! = \alpha_1! \alpha_2!$, $\Delta \theta_{ij}^n = \theta_{ij}^n - \theta_i^0$, $\Delta v_{ij}^n = v_{ij}^n - v_i^0$. Additionally, $A_1(x) = \sum_{i=1}^{k_0} (p_i^n - p_i^0) f(x|\theta_i^0, v_i^0)$, and $R_1(x) = O\left(\sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n (|\Delta \theta_{ij}^n|^{N+\delta} + |\Delta v_{ij}^n|^{N+\delta})\right)$ for some positive constant $\delta > 0$.

Step 3 Enter the key identity (3.2): $\frac{\partial^2 f}{\partial \theta^2}(x|\theta, v) = 2\frac{\partial f}{\partial v}(x|\theta, v)$ for all x . This entails, for any natural orders n_1, n_2 , that $\frac{\partial^{n_1+n_2} f}{\partial \theta^{n_1} \partial v^{n_2}}(x|\theta, v) = \frac{1}{2^{n_2}} \frac{\partial^{n_1+2n_2} f}{\partial \theta^{n_1+2n_2}}(x|\theta, v)$. Thus, by converting all derivatives to those taken with respect to only θ , we may rewrite (3.10) as

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n \sum_{\alpha \geq 1} \sum_{n_1, n_2} \frac{(\Delta \theta_{ij}^n)^{n_1} (\Delta v_{ij}^n)^{n_2}}{2^{n_2} n_1! n_2!} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_i^0, v_i^0) \\ &\quad + A_1(x) + R_1(x) \\ &:= A_1(x) + B_1(x) + R_1(x), \end{aligned} \tag{3.11}$$

where n_1, n_2 in the sum satisfy $n_1 + 2n_2 = \alpha, n_1 + n_2 \leq N$.

Step 4 We proceed to proving part (a) of the proposition. From the definition of \bar{r} , by setting $r = \bar{r} - 1$, there exist non-trivial solutions $(c_i^*, a_i^*, b_i^*)_{i=1}^{k-k_0+1}$ for the system of equations (4.24). Construct a sequence of probability measures $G_n \in \mathcal{O}_k$ under the representation given by Eq. (3.9) as follows:

$$\theta_{1j}^n = \theta_1^0 + \frac{a_j^*}{n}, \quad v_{1j}^n = v_1^0 + \frac{2b_j^*}{n^2}, \quad p_{1j}^n = \frac{p_1^0 (c_j^*)^2}{\sum_{j=1}^{k-k_0+1} (c_j^*)^2}, \quad \text{for all } j = 1, \dots, k - k_0 + 1,$$

and $\theta_{i1}^n = \theta_i^0, v_{i1}^n = v_i^0, p_{i1}^n = p_i^0$ for all $i = 2, \dots, k_0$. (That is, we set $k^* = k$, $s_1 = k - k_0 + 1, s_i = 1$ for all $2 \leq i \leq k_0$). Note that b_j^* may be negative, but we are guaranteed that $v_{1j}^n > 0$ for sufficiently large n . It is easy to verify that $W_1(G_n, G_0) = \sum_{i=1}^{k-k_0+1} p_{1i}^n \left(\frac{|a_i^*|}{n} + \frac{2|b_i^*|}{n^2} \right) \asymp \frac{1}{n}$, because at least one of the a_i^* is non-zero.

Step 5 Select $N = \bar{r}$ in Eq. (3.11). By our construction of G_n , clearly $A_1(x) = 0$. Moreover,

$$\begin{aligned} B_1(x) &= \sum_{i=1}^{k-k_0+1} p_{1i}^n \sum_{\alpha=1}^{\bar{r}-1} \sum_{n_1, n_2} \frac{(\Delta \theta_{1i}^n)^{n_1} (\Delta v_{1i}^n)^{n_2}}{2^{n_2} n_1! n_2!} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x | \theta_1^0, v_1^0) \\ &+ \sum_{i=1}^{k-k_0+1} p_{1i}^n \sum_{\alpha=\bar{r}}^{\bar{r}} \sum_{n_1, n_2} \frac{(\Delta \theta_{1i}^n)^{n_1} (\Delta v_{1i}^n)^{n_2}}{2^{n_2} n_1! n_2!} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x | \theta_1^0, v_1^0) \\ &:= \sum_{\alpha=1}^{\bar{r}-1} B_{\alpha n} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x | \theta_1^0, v_1^0) + \sum_{\alpha \geq \bar{r}} C_{\alpha n} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x | \theta_1^0, v_1^0). \end{aligned}$$

In the above display, for each $\alpha \geq \bar{r}$, observe that $C_{\alpha n} = O(n^{-\alpha})$. Moreover, for each $1 \leq \alpha \leq \bar{r} - 1$,

$$B_{\alpha n} = \frac{1}{n^\alpha \sum_{i=1}^{k-k_0+1} (c_i^*)^2} \sum_{i=1}^{k-k_0+1} (c_i^*)^2 \sum_{n_1+2n_2=\alpha} \frac{(a_i^*)^{n_1} (b_i^*)^{n_2}}{n_1! n_2!} = 0,$$

because $(c_i^*, a_i^*, b_i^*)_{i=1}^{k-k_0+1}$ form a non-trivial solution to system (4.24).

Step 6 We arrive at an upper bound for the Hellinger distance of mixture densities.

$$\begin{aligned} h^2(p_{G_n}, p_{G_0}) &\leq \frac{1}{2p_1^0} \int_{\mathbb{R}} \frac{(p_{G_n}(x) - p_{G_0}(x))^2}{f(x | \theta_1^0, v_1^0)} dx \\ &\lesssim \int_{\mathbb{R}} \frac{\sum_{\alpha=\bar{r}}^{\bar{r}} C_{\alpha n}^2 \left(\frac{\partial^\alpha f}{\partial \theta^\alpha}(x | \theta_1^0, v_1^0) \right)^2 + R_1^2(x)}{f(x | \theta_1^0, v_1^0)} dx, \end{aligned}$$

For Gaussian densities, it can be verified that $\left(\frac{\partial^\alpha f}{\partial \theta^\alpha}(x | \theta_1^0, v_1^0) \right)^2 / f(x | \theta_1^0, v_1^0)$ is integrable for all $1 \leq \alpha \leq 2\bar{r}$. So, $h^2(p_{G_n}, p_{G_0}) \leq O(n^{-2\bar{r}}) + \int R_1^2(x) / f(x | \theta_1^0, v_1^0) dx$.

Turning to the Taylor remainder $R_1(x)$, note that

$$\begin{aligned} |R_1(x)| &\lesssim \sum_{i=1}^{k-k_0+1} \sum_{|\beta|=\bar{r}+1} \frac{(\bar{r}+1)}{\beta!} |\Delta\theta_{1i}^n|^{\beta_1} |\Delta v_{1i}^n|^{\beta_2} \times \\ &\quad \times \int_0^1 (1-t)^{\bar{r}} \left| \frac{\partial^{\bar{r}+1} f}{\partial\theta^{\beta_1} \partial v^{\beta_2}}(x|\theta_1^0 + t\Delta\theta_{1i}^n, v_1^0 + t\Delta v_{1i}^n) \right| dt. \end{aligned}$$

Now, $(\Delta\theta_{1i}^n)^{\beta_1} (\Delta v_{1i}^n)^{\beta_2} \asymp n^{-\beta_1-2\beta_2} = o(n^{-2\bar{r}})$. In addition, as n is sufficiently large, we have for all $|\beta| = \bar{r} + 1$ that

$$\sup_{t \in [0,1]} \int_{x \in \mathbb{R}} \left(\frac{\partial^{\bar{r}+1} f}{\partial\theta^{\beta_1} \partial v^{\beta_2}}(x|\theta_1^0 + t\Delta\theta_{1i}^n, v_1^0 + t\Delta v_{1i}^n) \right)^2 / f(x|\theta_1^0, v_1^0) dx < \infty.$$

It follows that $h(p_{G_n}, p_{G_0}) = O(n^{-\bar{r}})$. As noted above, $W_1(G_n, G_0) \asymp n^{-1}$, so the claim of part (a) is established.

Step 7 Turning to part (b) of Proposition 3.2.2, it suffices to show that

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_{k,c_0}} \left\{ \sup_{x \in \mathcal{X}} |p_G(x) - p_{G_0}(x)| / W_{\bar{r}}(G, G_0) : W_{\bar{r}}(G, G_0) \leq \epsilon \right\} > 0. \quad (3.12)$$

Then one can arrive at the proposition's claim by passing through an argument using Fatou's lemma (cf. proof of Theorem 1 of Nguyen [2013] or step 4 in the proof of Theorem 3.1 of Ho and Nguyen [2016c]). Suppose that (3.12) does not hold. Then we can find a sequence of probability measures $G_n \in \mathcal{O}_{k,c_0}$ that are represented by Eq. (3.9), such that $W_{\bar{r}}(G_n, G_0) \rightarrow 0$ and $\sup_x |p_{G_n}(x) - p_{G_0}(x)| / W_{\bar{r}}(G_n, G_0) \rightarrow 0$. Define

$$D_n := d(G_n, G_0) := \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n (|\Delta\theta_{ij}^n|^{\bar{r}} + |\Delta v_{ij}^n|^{\bar{r}}) + \sum_{i=1}^{k_0} |p_i^n - p_i^0|.$$

It is easy to see that $W_{\bar{r}}^{\bar{r}}(G_n, G_0) \lesssim D_n$, since D_n is the multiple of the $W_{\bar{r}}^{\bar{r}}$ cost of moving mass from G_n to G_0 by a (possibly) non-optimal coupling. So, for all $x \in \mathbb{R}$, $(p_{G_n}(x) - p_{G_0}(x))/D_n \rightarrow 0$. Combining this fact with (3.11), where $N = \bar{r}$, we obtain

$$(A_1(x) + B_1(x) + R_1(x))/D_n \rightarrow 0. \quad (3.13)$$

We have $R_1(x)/D_n = o(1)$ as $n \rightarrow \infty$.

Step 8 $A_1(x)/D_n$ and $B_1(x)/D_n$ are the linear combination of elements of $\frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta, v)$ where $\alpha = n_1 + 2n_2$ and $n_1 + n_2 \leq \bar{r}$. Note that the natural order α ranges in $[0, 2\bar{r}]$. Let $E_\alpha(\theta, v)$ denote the corresponding coefficient of $\frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta, v)$. Extracting from (3.11), for $\alpha = 0$, $E_0(\theta_i^0, v_i^0) = (p_i^n - p_i^0)/D_n$. For $\alpha \geq 1$,

$$E_\alpha(\theta_i^0, v_i^0) = \left[\sum_{j=1}^{s_i} p_{ij}^n \sum_{\substack{n_1+2n_2=\alpha \\ n_1+n_2 \leq \bar{r}}} \frac{(\Delta \theta_{ij}^n)^{n_1} (\Delta v_{ij}^n)^{n_2}}{2^{n_2} n_1! n_2!} \right] / D_n.$$

In the remainder of this proof step, we shall show that as $n \rightarrow \infty$, at least one of the coefficients $E_\alpha(\theta_i^0, v_i^0)$ must not vanish. Suppose this is not the case, i.e., $E_\alpha(\theta_i^0, v_i^0) \rightarrow 0$ for all $i = 1, \dots, k_0$ and $0 \leq \alpha \leq 2\bar{r}$ as $n \rightarrow \infty$. By taking the summation of all $|E_0(\theta_i^0, v_i^0)|$, we get $\sum_{i=1}^{k_0} |p_i^n - p_i^0|/D_n \rightarrow 0$ as $n \rightarrow \infty$. As a consequence, we obtain

$$\sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}^n (|\Delta \theta_{ij}^n|^{\bar{r}} + |\Delta v_{ij}^n|^{\bar{r}})/D_n \rightarrow 1 \text{ as } n \rightarrow \infty.$$

Hence, we can find an index $i^* \in \{1, 2, \dots, k_0\}$ such that as $n \rightarrow \infty$

$$\sum_{j=1}^{s_{i^*}} p_{i^*j}^n (|\Delta \theta_{i^*j}^n|^{\bar{r}} + |\Delta v_{i^*j}^n|^{\bar{r}})/D_n \not\rightarrow 0.$$

Without loss of generality, we assume that $i^* = 1$. Accordingly,

$$\begin{aligned} F_\alpha(\theta_1^0, v_1^0) &:= \frac{D_n E_\alpha(\theta_1^0, \sigma_1^0)}{\sum_{j=1}^{s_1} p_{1j}^n (|\Delta\theta_{1j}^n|^{\bar{r}} + |\Delta v_{1j}^n|^{\bar{r}})} \\ &= \frac{\sum_{j=1}^{s_1} p_{1j}^n \sum_{\substack{n_1+2n_2=\alpha \\ n_1+n_2 \leq \bar{r}}} \frac{(\Delta\theta_{1j}^n)^{n_1} (\Delta v_{1j}^n)^{n_2}}{2^{n_2} n_1! n_2!}}{\sum_{j=1}^{s_1} p_{1j}^n (|\Delta\theta_{1j}^n|^{\bar{r}} + |\Delta v_{1j}^n|^{\bar{r}})} \rightarrow 0. \end{aligned}$$

If $s_1 = 1$ then $F_1(\theta_1^0, v_1^0)$ and $F_{2\bar{r}}(\theta_1^0, v_1^0)$ yield

$$|\Delta\theta_{11}^n|^{\bar{r}} / (|\Delta\theta_{11}^n|^{\bar{r}} + |\Delta v_{11}^n|^{\bar{r}}), \quad |\Delta v_{11}^n|^{\bar{r}} / (|\Delta\theta_{11}^n|^{\bar{r}} + |\Delta v_{11}^n|^{\bar{r}}) \rightarrow 0,$$

which is a contradiction. As a consequence, $s_1 \geq 2$.

Denote $\bar{p}_n = \max_{1 \leq j \leq s_1} \{p_{1j}^n\}$, $\bar{M}_n = \max \left\{ |\Delta\theta_{11}^n|, \dots, |\Delta\theta_{1s_1}^n|, |\Delta v_{11}^n|^{1/2}, \dots, |\Delta v_{1s_1}^n|^{1/2} \right\}$. Since $0 < p_{1j}^n / \bar{p}_n \leq 1$ for all $1 \leq j \leq s_1$, by a subsequence argument, there exist $c_j^2 := \lim_{n \rightarrow \infty} p_{1j}^n / \bar{p}_n$ for all $j = 1, \dots, s_1$. Similarly, define $a_j := \lim_{n \rightarrow \infty} \Delta\theta_{1j}^n / \bar{M}_n$, and $2b_j := \lim_{n \rightarrow \infty} \Delta v_{1j}^n / \bar{M}_n^2$ for each $j = 1, \dots, s_1$. By the constraints of \mathcal{O}_{k,c_0} , $p_{1j}^n \geq c_0$, so all of c_j^2 differ from 0 and at least one of them equals to 1. Likewise, at least one element of $(a_j, b_j)_{j=1}^{s_1}$ equal to -1 or 1. Now, for each $\alpha = 1, \dots, \bar{r}$, divide both the numerator and denominator of $F_\alpha(\theta_1^0, v_1^0)$ by \bar{p}_n and then \bar{M}_n^α and let $n \rightarrow \infty$, we obtain the following system of polynomial equations

$$\sum_{j=1}^{s_1} \sum_{n_1+2n_2=\alpha} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0 \quad \text{for each } \alpha = 1, \dots, \bar{r}.$$

Since $s_1 \geq 2$, we get $\bar{r} \geq 4$. If $a_i = 0$ for all $1 \leq i \leq s_1$ then by choosing $\alpha = 4$, we obtain $\sum_{j=1}^{s_1} c_j^2 b_j^2 = 0$. However, it demonstrates that $b_i = 0$ for all $1 \leq i \leq s_1$ — a contradiction to the fact that at least one element of $(a_i, b_i)_{i=1}^{s_1}$ is different from 0.

Therefore, at least one element of $(a_i)_{i=1}^{s_1}$ is not equal to 0. Observe that $s_i \leq k - k_0 + 1$

(because the number of distinct atoms of G_n is $\sum_{i=1}^{k_0} s_i \leq k$ and all $s_i \geq 1$). Thus, the existence of non-trivial solutions for the system of equations given in the above display entails the existence of non-trivial solutions for system of equations (4.24). This contradicts with the definition of \bar{r} . Therefore, our hypothesis that all coefficients $E_\alpha(\theta_i^0, v_i^0)$ vanish does not hold — there must be at least one coefficient which does not converge to 0 as $n \rightarrow \infty$.

Step 9 Let m_n be the maximum of the absolute values of $E_\alpha(\theta_i^0, v_i^0)$ where $0 \leq \alpha \leq 2\bar{r}$, $1 \leq i \leq k_0$ and $d_n = 1/m_n$. Since $m_n \not\rightarrow 0$ as $n \rightarrow \infty$, d_n is uniformly bounded above for all n . As $d_n|E_\alpha(\theta_i^0, v_i^0)| \leq 1$, we have $d_n E_\alpha(\theta_i^0, v_i^0) \rightarrow \beta_{i\alpha}$ for all $0 \leq \alpha \leq 2\bar{r}$, $1 \leq i \leq k_0$ where at least one of $\beta_{i\alpha}$ differs from 0. Incorporating these limits to Eq.(3.13), we obtain that for all $x \in \mathbb{R}$,

$$(p_{G_n}(x) - p_{G_0}(x))/D_n \rightarrow \sum_{i=1}^{k_0} \sum_{\alpha=0}^{2\bar{r}} \beta_{i\alpha} \frac{\partial^\alpha f}{\partial \theta^\alpha}(x|\theta_i^0, v_i^0) = 0.$$

By direct calculation, we can rewrite the above equation as

$$\sum_{i=1}^{k_0} \left(\sum_{j=1}^{2\bar{r}+1} \gamma_{ij} (x - \theta_i^0)^{j-1} \right) \exp\left(-\frac{(x - \theta_i^0)^2}{2v_i^0}\right) = 0 \quad \text{for all } x \in \mathbb{R}, \quad (3.14)$$

where γ_{ij} for odd j are linear combinations of $\beta_{i(2l_1)}$, for $(j-1)/2 \leq l_1 \leq \bar{r}$, such that all of the coefficients are functions of v_i^0 differing from 0. For even j , γ_{ij} are linear combinations of $\beta_{i(2l_2+1)}$, for $j/2 \leq l_2 \leq \bar{r}$, such that all of the coefficients are functions of v_i^0 differing from 0. Now, without loss of generality, we assume that $v_1^0 \leq v_2^0 \leq \dots \leq v_{k_0}^0$. Denote $\bar{i} \in [1, k_0]$ to be the minimum index i such that $v_i^0 = v_{k_0}^0$. It implies that $v_{\bar{i}}^0 = v_{\bar{i}+1}^0 = \dots = v_{k_0}^0$. Therefore, θ_i^0 are pairwise different as $\bar{i} \leq i \leq k_0$. Now, let call $\underline{i} = \arg \max_{\bar{i} \leq i \leq k_0} \theta_i^0$. Multiply both sides of (3.14) with $\exp[(x - \theta_{\underline{i}}^0)^2/2v_{\underline{i}}^0]$

and let $x \rightarrow +\infty$, then we can check that

$$\sum_{j=1}^{2\bar{r}+1} \gamma_{ij}(x - \theta_i^0)^{j-1} \rightarrow 0,$$

which only happens when $\gamma_{ij} = 0$ for all $1 \leq j \leq 2\bar{r}+1$. Employing the same argument to the remained indices, we obtain $\gamma_{ij} = 0$ for all $i = 1, \dots, k_0, j = 1, \dots, 2\bar{r}+1$. This entails that $\beta_{i\alpha} = 0$ for all $i = 1, \dots, k_0, \alpha = 0, \dots, 2\bar{r}$ — a contradiction. Thus we achieve the conclusion of (3.12).

PROOF OF PROPOSITION 3.2.1 Our proof is based on Groebner bases method for determining solutions for a system of polynomial equations. (i) For the case $k - k_0 = 1$, the system (4.24) when $r = 4$ can be written as

$$c_1^2 a_1 + c_2^2 a_2 = 0 \quad (3.15)$$

$$\frac{1}{2}(c_1^2 a_1^2 + c_2^2 a_2^2) + c_1^2 b_1 + c_2^2 b_2 = 0 \quad (3.16)$$

$$\frac{1}{3!}(c_1^2 a_1^3 + c_2^2 a_2^3) + c_1^2 a_1 b_1 + c_2^2 a_2 b_2 = 0 \quad (3.17)$$

$$\frac{1}{4!}(c_1^2 a_1^4 + c_2^2 a_2^4) + \frac{1}{2!}(c_1^2 a_1^2 b_1 + c_2^2 a_2^2 b_2) + \frac{1}{2!}(c_1^2 b_1^2 + c_2^2 b_2^2) = 0 \quad (3.18)$$

Suppose that the above system has a non-trivial solution. If $c_1 a_1 = 0$, then equation (3.15) implies $c_2 a_2 = 0$. Since $c_1, c_2 \neq 0$, we have $a_1 = a_2 = 0$. This violates the constraint that one of a_1, a_2 is non-zero. Hence, $c_1 a_1, c_2 a_2 \neq 0$. Divide both sides of (3.15), (3.16), (3.17), (3.18) by $c_1^2 a_1, c_1^2 a_1^2, c_1^2 a_1^3, c_1^2 a_1^4$ respectively, we obtain the following

system of polynomial equations

$$1 + x^2a = 0$$

$$1 + x^2a^2 + 2(b + x^2c) = 0$$

$$1 + x^2a^3 + 6(b + x^2ac) = 0$$

$$1 + x^2a^4 + 12(b + x^2a^2c) + 12(b^2 + x^2c^2) = 0$$

where $x = c_2/c_1, a = a_2/a_1, b = b_1/a_1, c = b_2/a_1$. By taking the lexicographical order $a \succ b \succ c \succ x$, the Groebner basis of the above system contains $x^6 + 2x^4 + 2x^2 + 1 > 0$ for all $x \in \mathbb{R}$. Therefore, the above system of polynomial equations does not have real solutions. As a consequence, the original system of polynomial equations does not have non-trivial solution, which means that $\bar{r} \leq 4$. However, we have already shown that as $r = 3$, Eq.(4.24) has non-trivial solution. Therefore, $\bar{r} = 4$.

(ii) The case $k - k_0 = 2$. System (4.24) when $r = 6$ takes the form:

$$\sum_{i=1}^3 c_i^2 a_i = 0 \quad (3.19)$$

$$\frac{1}{2} \sum_{i=1}^3 c_i^2 a_i^2 + \sum_{i=1}^3 c_i^2 b_i = 0 \quad (3.20)$$

$$\frac{1}{6} \sum_{i=1}^3 c_i^2 a_i^3 + \frac{1}{2} \sum_{i=1}^3 c_i^2 a_i b_i = 0 \quad (3.21)$$

$$\frac{1}{24} \sum_{i=1}^3 c_i^2 a_i^4 + \frac{1}{2} \sum_{i=1}^3 c_i^2 a_i^2 b_i + \frac{1}{2} \sum_{i=1}^3 c_i^2 b_i^2 = 0 \quad (3.22)$$

$$\frac{1}{120} \sum_{i=1}^3 c_i^2 a_i^5 + \frac{1}{6} \sum_{i=1}^3 c_i^2 a_i^3 b_i + \frac{1}{2} \sum_{i=1}^3 c_i^2 a_i b_i^2 = 0 \quad (3.23)$$

$$\frac{1}{720} \sum_{i=1}^3 c_i^2 a_i^6 + \frac{1}{24} \sum_{i=1}^3 c_i^2 a_i^4 b_i + \frac{1}{4} \sum_{i=1}^3 c_i^2 a_i^2 b_i^2 + \frac{1}{6} \sum_{i=1}^3 c_i^2 b_i^3 = 0 \quad (3.24)$$

Non-trivial solution constraints require that $c_1, c_2, c_3 \neq 0$ and without loss of gener-

ality, $a_1 \neq 0$. Dividing both sides of the six equations above by $c_1^2 a_1, c_1^2 a_1^2, c_1^2 a_1^3, c_1^2 a_1^4, c_1^2 a_1^5, c_1^2 a_1^6$, respectively, we obtain

$$\begin{aligned} 1 + x^2 a + y^2 b &= 0 \\ \frac{1}{2}(1 + x^2 a^2 + y^2 b^2) + c + x^2 d + y^2 e &= 0 \\ \frac{1}{3}(1 + x^2 a^3 + y^2 b^3) + c + x^2 a d + y^2 b e &= 0 \\ \frac{1}{12}(1 + x^2 a^4 + y^2 b^4) + c + x^2 a^2 d + y^2 b^2 e + c^2 + x^2 d^2 + y^2 e^2 &= 0 \\ \frac{1}{60}(1 + x^2 a^5 + y^2 b^5) + \frac{1}{3}(c + x^2 a^3 d + y^2 b^3 e) + c^2 + x^2 a d^2 + y^2 b e^2 &= 0 \\ \frac{1}{360}(1 + x^2 a^6 + y^2 b^6) + \frac{1}{12}(c + x^2 a^4 d + y^2 b^4 e) + \frac{1}{2}(c^2 + x^2 a^3 d + y^2 b^3 e) \\ + \frac{1}{3}(c^3 + x^2 d^3 + y^2 e^3) &= 0 \end{aligned}$$

where $x = c_2/c_1, y = c_3/c_1, a = a_2/a_1, b = a_3/a_1, c = b_1/a_1^2, d = b_2/a_1^2, e = b_3/a_1^2$.

By taking the lexicographical order $a \succ b \succ c \succ d \succ x \succ y$, we can verify that the Groebner bases of the above system of polynomial equations contains a polynomial in terms of x^2, y^2 with all of the positive coefficient numbers, which cannot be 0 when $x, y \in \mathbb{R}$. Therefore, the original system of polynomial equations does not have a non-trivial solution. It follows that $\bar{r} \leq 6$.

When $r = 5$, we retain the first five equations in the system described in the above display. By choosing $x = y = 1$, under lexicographical order $a \succ b \succ c \succ d \succ e$, we can verify that the Groebner bases contains a polynomial of e with roots $e = \pm\sqrt{2}/3$ or $e = (-3 \pm \sqrt{2})/6$ while a, b, c, d can be uniquely determined by e . Thus, system of polynomial equations (4.24) has a non-trivial solution. It follows that $\bar{r} = 6$.

(iii) For the case $k - k_0 \geq 3$, we choose $c_1 = c_2 = \dots = c_{k-k_0+1} = 1, a_i = b_i = 0$ for all $4 \leq i \leq k - k_0 + 1$. Additionally, take $a_1 = a_2 = 1$. Now, by choosing $r = 6$ in system (4.24), we can check by Groebner bases that this system of polynomial equations has a non-trivial solution. As a result, $\bar{r} \geq 7$.

3.5.2 Mixture of Gamma distributions and location-exponential distributions

PROOF OF THEOREM 3.3.1 The proof of this theorem proceeds in the same manner as that of Theorem 3.1.1. Therefore, it suffices to prove the following.

Proposition 3.5.1. (Bounds for exact-fitted Gamma mixtures)

(a) (*Generic cases*) Assume that the support points of G_0 satisfy assumption (A.1).

Then for $G \in \mathcal{E}_{k_0}$ and $W_1(G, G_0)$ sufficiently small, we have

$$V(p_G, p_{G_0}) \gtrsim W_1(G, G_0).$$

(b) (*Pathological cases*) If the support points of G_0 satisfy assumption (A.2), then

for any $r \geq 1$

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{E}_{k_0}} \left\{ V(p_G, p_{G_0}) / W_r^r(G, G_0) : W_r(G, G_0) \leq \epsilon \right\} = 0.$$

Proof. (a) For the range of generic parameter values of G_0 , we shall show that the first-order identifiability still holds for Gamma mixtures, so that the conclusion can be drawn immediately from Theorem 3.1 of Ho and Nguyen [2016c]. It suffices to show that for any $\alpha_{ij} \in \mathbb{R}$ ($1 \leq i \leq 3, 1 \leq j \leq k_0$) such that for almost sure $x > 0$

$$\sum_{i=1}^{k_0} \alpha_{1i} f(x|a_i^0, b_i^0) + \alpha_{2i} \frac{\partial f}{\partial a}(x|a_i^0, b_i^0) + \alpha_{3i} \frac{\partial f}{\partial b}(x|a_i^0, b_i^0) = 0 \quad (3.25)$$

then $\alpha_{ij} = 0$ for all i, j . Equation (3.25) is rewritten as

$$\sum_{i=1}^{k_0} \left(\beta_{1i} x^{a_i^0 - 1} + \beta_{2i} (\log x) x^{a_i^0 - 1} + \beta_{3i} x^{a_i^0} \right) \exp(-b_i^0 x) = 0, \quad (3.26)$$

where $\beta_{1i} = \alpha_{1i} \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)} + \alpha_{2i} \frac{(b_i^0)^{a_i^0} (\log(b_i^0) - \psi(a_i^0))}{\Gamma(a_i^0)} + \alpha_{3i} \frac{a_i^0 (b_i^0)^{a_i^0 - 1}}{\Gamma(a_i^0)}$, $\beta_{2i} = \alpha_{2i} \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}$, and $\beta_{3i} = -\alpha_{3i} \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}$. Without loss of generality, we assume that $b_1^0 \leq b_2^0 \leq \dots \leq b_{k_0}^0$.

Denote \bar{i} to be the maximum index i such that $b_i^0 = b_1^0$. Then we have that $a_1^0, \dots, a_{\bar{i}}^0$ are pairwise different. Multiply both sides of (3.26) with $\exp(b_i^0 x)$ and let $x \rightarrow +\infty$, we obtain

$$\sum_{i=1}^{\bar{i}} \beta_{1i} x^{a_i^0 - 1} + \beta_{2i} (\log x) x^{a_i^0 - 1} + \beta_{3i} x^{a_i^0} \rightarrow 0.$$

Since $|a_i^0 - a_j^0| \neq 1$ and $a_i^0 \geq 1$ for all $1 \leq i, j \leq \bar{i}$, the above result implies that

$\beta_{1i} = \beta_{2i} = \beta_{3i} = 0$ for all $1 \leq i \leq \bar{i}$ or equivalently $\alpha_{1i} = \alpha_{2i} = \alpha_{3i}$ for all $1 \leq i \leq \bar{i}$.

Repeat the same argument for the remained indices, we obtain $\alpha_{1i} = \alpha_{2i} = \alpha_{3i} = 0$ for all $1 \leq i \leq k_0$. This concludes the proof.

(b) Without loss of generality, we assume that $\{|a_2^0 - a_1^0|, |b_2^0 - b_1^0|\} = \{1, 0\}$. In particular, $b_1^0 = b_2^0$ and assume $a_2^0 = a_1^0 - 1$. We construct the following sequence of measures $G_n = \sum_{i=1}^{k_0} p_i^n \delta_{(a_i^n, b_i^n)}$, where $a_i^n = a_i^0$ for all $1 \leq i \leq k_0$, $b_1^n = b_1^0, b_2^n = b_1^0 \left(1 + \frac{1}{a_2^0(np_2^0 - 1)}\right)$, $b_i^n = b_i^0$ for all $3 \leq i \leq k_0$, $p_1^n = p_1^0 + 1/n, p_2^n = p_2^0 - 1/n, p_i^n = p_i^0$ for all $3 \leq i \leq k_0$. We can check that $W_r^r(G_n, G_0) \asymp 1/n + (p_2^0 - 1/n)|b_2^n - b_1^0|^r \asymp n^{-1}$ as $n \rightarrow \infty$. For any natural order $r \geq 1$, by applying Taylor's expansion up to the $([r] + 1)$ th-order, we obtain:

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{k_0} p_i^n (f(x|a_i^n, b_i^n) - f(x|a_i^0, b_i^0)) + (p_i^n - p_i^0) f(x|a_i^0, b_i^0) \\ &= (p_1^n - p_1^0) f(x|a_1^0, b_1^0) + (p_2^n - p_2^0) f(x|a_2^0, b_2^0) + \\ &\quad \sum_{j=1}^{[r]+1} p_2^n \frac{(b_2^n - b_1^0)^j}{j!} \frac{\partial^j f}{\partial b^j}(x|a_2^0, b_2^0) + R_n(x). \end{aligned} \quad (3.27)$$

The Taylor expansion remainder $|R_n(x)| = O(p_2^n |b_2^n - b_1^0|^{[r]+1+\delta})$ for some $\delta > 0$ due to $a_2^0 \geq 1$. Therefore, $R_n(x) = o(W_r^r(G_n, G_0))$ as $n \rightarrow \infty$. For the choice of p_2^n, b_2^n , we

can check that as $j \geq 2$, $p_2^n(b_2^n - b_2^0)^j = o(W_r^r(G_n, G_0))$. Now, we can rewrite (3.27) as

$$p_{G_n}(x) - p_{G_0}(x) = A_n x^{a_2^0} \exp(-b_1^0 x) + B_n x^{a_2^0-1} \exp(-b_1^0 x) + \sum_{j=2}^{[r]+1} p_2^n \frac{(b_2^n - b_2^0)^j}{j!} \frac{\partial^j f}{\partial b^j}(x|a_2^0, b_2^0) + R_n(x),$$

where we have $A_n = \frac{(b_1^0)^{a_1^0}}{\Gamma(a_1^0)}(p_1^n - p_1^0) - \frac{(b_1^0)^{a_2^0}}{\Gamma(a_2^0)}p_2^n(b_2^n - b_1^0) = 0$ and similarly $B_n = \frac{(b_1)^{a_2^0}}{\Gamma(a_2^0)}(p_2^n - p_2^0) + \frac{a_2^0(b_1^0)^{a_2^0-1}}{\Gamma(a_2^0)}p_2^n(b_2^n - b_1^0) = 0$ for all n . Since $a_2^0 \geq 1$, $\left| \frac{\partial^j f}{\partial b^j}(x|a_2^0, b_2^0) \right|$ is bounded for all $2 \leq j \leq r+1$. It follows that $\sup_{x>0} |p_{G_n}(x) - p_{G_0}(x)| = O(n^{-2})$.

Observe that

$$\begin{aligned} V(p_{G_n}, p_{G_0}) &= 2 \int_{p_{G_n}(x) < p_{G_0}(x)} (p_{G_0}(x) - p_{G_n}(x)) d(x) \\ &\leq 2 \int_{x \in (0, a_2^0/b_1^0)} |p_{G_n}(x) - p_{G_0}(x)| dx. \end{aligned}$$

As a consequence $V(p_{G_n}, p_{G_0}) = O(n^{-1/2})$ so for any $r \geq 1$, $V(p_{G_n}, p_{G_0}) = o(W_r^r(G_n, G_0))$ as $n \rightarrow \infty$. \square

PROOF OF THEOREM 3.3.2 As in the proof of Theorem 3.3.1, it is sufficient to prove the following.

Proposition 3.5.2. (Bounds for over-fitted Gamma mixtures)

(a) (*Generic cases*) Assume that we have $G_0 \in \mathcal{O}_{k,c_0}$. Then, for $G \in \mathcal{O}_{k,c_0}$ and $W_2(G, G_0)$ sufficiently small, we obtain

$$V(p_G, p_{G_0}) \gtrsim W_2^2(G, G_0).$$

(b) (*Pathological cases*) Assume that the support points of G_0 satisfy assumption

(A.4), then for any $r \geq 1$,

$$\liminf_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_k} \left\{ V(p_G, p_{G_0}) / W_r^r(G, G_0) : W_r(G, G_0) \leq \epsilon \right\} = 0.$$

Proof. (a) As in step 7 in the proof of Proposition 3.2.2, it suffices to show that

$$\begin{aligned} \liminf_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_{k,c_0}} & \left\{ \sup_{x \in \mathcal{X}} |p_G(x) - p_{G_0}(x)| / W_2^2(G, G_0) : \right. \\ & \left. W_2(G, G_0) \leq \epsilon \right\} > 0. \end{aligned} \quad (3.28)$$

Suppose this does not hold, by repeating the arguments of step 1 of Proposition 3.2.2, there is a sequence $G_n = \sum_{i=1}^{k_0+m} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(a_{ij}^n, b_{ij}^n)} \rightarrow G_0 = \sum_{i=1}^{k_0+m} p_i^0 \delta_{(a_i^0, b_i^0)}$ such that $(a_{ij}^n, b_{ij}^n) \rightarrow (a_i^0, b_i^0)$ for all $1 \leq i \leq k_0 + m$ where (a_i^0, b_i^0) are limit points that lie outside the support points of G_0 as $k_0 + 1 \leq i \leq k_0 + m$. Additionally, $p_i^0 = 0$ as $k_0 + 1 \leq i \leq k_0 + m$. Invoke the Taylor expansion up to the second order and assume that all of the coefficients corresponding to the first and second derivatives with respect to the parameters go to 0. Use the same argument as that of step 8 in Proposition 3.2.2, by summing up all the coefficients of second derivative, we obtain the contradiction. Now, by proceeding in the same way as that of step 9 in Proposition 3.2.2, as we let $n \rightarrow \infty$, we have for almost every x ,

$$\begin{aligned} \frac{p_{G_n}(x) - p_{G_0}(x)}{d(G_n, G_0)} & \rightarrow \sum_{i=1}^{k_0+m} \left\{ \alpha_{1i} f(x|a_i^0, b_i^0) + \alpha_{2i} \frac{\partial f}{\partial a}(x|a_i^0, b_i^0) + \alpha_{3i} \frac{\partial f}{\partial b}(x|a_i^0, b_i^0) + \right. \\ & \left. \sum_{j=1}^{s_i} \alpha_{4ij}^2 \frac{\partial^2 f}{\partial a^2}(x|a_i^0, b_i^0) + \sum_{j=1}^{s_i} \alpha_{5ij}^2 \frac{\partial^2 f}{\partial b^2}(x|a_i^0, b_i^0) + 2 \sum_{j=1}^{s_i} \alpha_{4ij} \alpha_{5ij} \frac{\partial^2 f}{\partial a \partial b}(x|a_i^0, b_i^0) \right\} = 0, \end{aligned}$$

where at least one of $\alpha_{1i}, \alpha_{2i}, \alpha_{3i}, \sum_{j=1}^{s_i} \alpha_{4ij}^2, \sum_{j=1}^{s_i} \alpha_{5ij}^2, 2 \sum_{j=1}^{s_i} \alpha_{4ij} \alpha_{5ij}$ is non-zero. We can

rewrite the above equation as

$$\sum_{i=1}^{k_0+m} \left\{ \beta_{1i} x^{a_i^0-1} + \beta_{2i} x^{a_i^0} + \beta_{3i} x^{a_i^0+1} + \beta_{4i} (\log x) x^{a_i^0-1} + \beta_{5i} (\log x)^2 x^{a_i^0-1} + \beta_{6i} (\log x) x^{a_i^0} \right\} e^{-b_i^0 x} = 0, \quad (3.29)$$

where $\beta_{1i} = \alpha_{1i} \frac{b_i^0}{\Gamma(a_i^0)} + \beta_i^0 \frac{\partial}{\partial a} \left(\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)} \right) + \alpha_{3i} \frac{a_i^0 (b_i^0)^{a_i^0-1}}{\Gamma(a_i^0)} + \sum_{j=1}^{s_i} \alpha_{4ij}^2 \frac{\partial}{\partial a^2} \left(\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)} \right) + \sum_{j=1}^{s_i} \alpha_{5ij}^2 \frac{a_i^0 (a_i^0 - 1) (b_i^0)^{a_i^0-2}}{\Gamma(a_i^0)} + 2 \sum_{j=1}^{s_i} \alpha_{4ij} \alpha_{5ij} \frac{\partial}{\partial a} \left(\frac{a_i^0 (b_i^0)^{a_i^0-1}}{\Gamma(a_i^0)} \right), \quad \beta_{2i} = -\alpha_{3i} \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)} + 2 \sum_{j=1}^{s_i} \alpha_{5ij}^2 \frac{a_i^0 (b_i^0)^{a_i^0-1}}{\Gamma(a_i^0)} + 2 \sum_{j=1}^{s_i} \alpha_{4ij} \alpha_{5ij} \frac{\partial}{\partial a} \left(\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)} \right), \quad \beta_{3i} = \sum_{j=1}^{s_i} \alpha_{5ij}^2 \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}, \quad \beta_{4i} = \alpha_{2i} \frac{(b_i^0)}{\Gamma(a_i^0)} + 2 \sum_{j=1}^{s_i} \alpha_{4ij}^2 \frac{\partial}{\partial a} \left(\frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)} \right) + 2 \sum_{j=1}^{s_i} \alpha_{4ij} \alpha_{5ij} \frac{a_i^0 (b_i^0)^{a_i^0-1}}{\Gamma(a_i^0)}, \quad \beta_{5i} = \sum_{j=1}^{s_i} \alpha_{4ij}^2 \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)},$

and $\beta_{6i} = -2 \sum_{j=1}^{s_i} \alpha_{4ij} \alpha_{5ij} \frac{(b_i^0)^{a_i^0}}{\Gamma(a_i^0)}$. Using the same argument as that of the proof of part (a) of Proposition 3.5.1, by multiplying both sides of the above equation with $\exp(b_i^0 x)$ and let $x \rightarrow +\infty$, we obtain

$$\sum_{i=1}^{\bar{i}} \left\{ \beta_{1i} x^{a_i^0-1} + \beta_{2i} x^{a_i^0} + \beta_{3i} x^{a_i^0+1} + \beta_{4i} (\log x) x^{a_i^0-1} + \beta_{5i} (\log x)^2 x^{a_i^0-1} + \beta_{6i} (\log x) x^{a_i^0} \right\} \rightarrow 0.$$

By the constraints of \mathcal{O}_{k,c_0} , we have $|a_i^0 - a_j^0| \notin \{1, 2\}$ for all $1 \leq i, j \leq k_0 + m$. Therefore, this limit yields $\beta_{1i} = \beta_{2i} = \beta_{3i} = \beta_{4i} = \beta_{5i} = \beta_{6i} = 0$ for all $1 \leq i \leq \bar{i}$ or equivalently $\alpha_{1ij} = \alpha_{2ij} = \alpha_{3ij} = \alpha_{4ij} = \alpha_{5ij} = 0$ for all $1 \leq i \leq \bar{i}, 1 \leq j \leq s_i$. The same argument for remained indicies yields $\alpha_{1ij} = \alpha_{2ij} = \alpha_{3ij} = \alpha_{4ij} = \alpha_{5ij} = 0$ for all $1 \leq i \leq k_0 + m, 1 \leq j \leq s_i$, which leads to contradiction. This concludes the proof.

(b) If there exists (i, j) such that $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} \equiv \{1, 0\}$, then we can use the same way of construction as that of part (b) of Proposition 3.5.1. Now, the only

case of interest is when we have some (i, j) such that $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} \equiv \{2, 0\}$.

Without loss of generality, assume that $a_2^0 = a_1^0 - 2$. We construct the sequence

$$G_n = \sum_{i=1}^{k_0+1} p_i^n \delta_{(a_i^n, b_i^n)} \text{ as } a_1^n = a_1^0, a_2^n = a_3^n = a_2^0, a_i^n = a_{i-1}^0 \text{ for all } 4 \leq i \leq k_0 + 1,$$

$$b_1^n = b_1^0, b_2^n - b_1^n = b_1^0 - b_3^n = \frac{b_1^0}{a_2^0 n}, b_i^n = b_{i-1}^0 \text{ for all } 4 \leq i \leq k_0 + 1, p_1^n = p_1^0 - c_n,$$

$$p_2^n = \frac{p_2^0}{2} + \frac{1}{2} \left(c_n + \frac{1}{n} \right), p_3^n = \frac{p_2^0}{2} + \frac{1}{2} \left(c_n - \frac{1}{n} \right), p_i^n = p_{i-1}^0 \text{ for all } 4 \leq i \leq k_0 + 1 \text{ where}$$

$$c_n = \frac{(a_2^0 + 1)p_2^0}{(2n^2 - 1)a_2^0 - 1}. \text{ Now, we can check that for any } r \geq 1, W_r^r(G_n, G_0) \gtrsim c_n + \frac{1}{n^r}.$$

As $r \geq 2$, by means of Taylor expansions up to the $([r] + 1)$ -th order, we obtain

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= (p_1^n - p_1^0)f(x|a_1^0, b_1^0) + \left(\sum_{i=2}^3 p_i^n - p_2^0 \right) f(x|a_2^0, b_2^0) \\ &\quad + \sum_{j=1}^{[r]+1} \frac{\sum_{i=2}^3 p_i^n (b_i^n - b_i^0)^j}{j!} \frac{\partial^j f}{\partial b^j}(x|a_2^0, b_2^0) + R_n(x), \end{aligned} \quad (3.30)$$

where $R_n(x)$ is the remainder term and therefore $|R_n(x)|/W_r^r(G_n, G_0) \rightarrow 0$. We can check that as $j \geq 3$, $\sum_{i=2}^3 p_i^n (b_i^n - b_i^0)^j/W_r^r(G_n, G_0) \rightarrow 0$ as $n \rightarrow \infty$. Additionally, direct computation demonstrates that

$$\begin{aligned} (p_1^n - p_1^0)f(x|a_1^0, b_1^0) + \left(\sum_{i=2}^3 p_i^n - p_2^0 \right) f(x|a_2^0, b_2^0) &+ \\ \sum_{j=1}^2 \frac{\sum_{i=2}^3 p_i^n (b_i^n - b_i^0)^j}{j!} \frac{\partial^j f}{\partial b^j}(x|a_2^0, b_2^0) &= 0. \end{aligned}$$

The rest of the proof proceeds in the same way as that of Proposition 3.5.1 part (b). \square

PROOF OF THEOREM 3.3.3 It suffices to demonstrate the following bound:

Proposition 3.5.3. (Location-exponential mixtures) *For any $r \geq 1$,*

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{E}_{k_0}} \left\{ V(p_G, p_{G_0}) / W_1^r(G, G_0) : W_1(G, G_0) \leq \epsilon \right\} = 0.$$

Proof. Choose the sequence $G_n = \sum_{i=1}^{k_0} p_i^n \delta_{(\theta_i^n, \sigma_i^n)}$ such that $\sigma_i^n = \sigma_i^0$ for all $1 \leq i \leq k_0$, $(p_i^n, \theta_i^n) = (p_i^0, \theta_i^0)$ for all $3 \leq i \leq k_0$. The parameters $p_1^n, p_2^n, \theta_1^n, \theta_2^n$ are to be determined. With this construction of G_n , we obtain $W_1(G_n, G_0) \asymp |p_1^n - p_1^0| + |p_2^n - p_2^0| + p_1^0 |\theta_1^n - \theta_1^0| + p_2^0 |\theta_2^n - \theta_2^0|$. Now, for any $x \notin \{\theta_1^0, \theta_2^0\}$ and for any $r \geq 1$, taking the Taylor expansion with respect to θ up to the $([r] + 1)$ -th order, we obtain

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^2 p_i^0 (f(x|\theta_i^n, \sigma_i^0) - f(x|\theta_i^0, \sigma_i^0)) + (p_i^n - p_i^0) f(x|\theta_i^n, \sigma_i^0) \\ &= \sum_{i=1}^2 (p_i^n - p_i^0) f(x|\theta_i^n, \sigma_i^0) - p_i^0 \left[\sum_{j=1}^{[r]+1} \frac{(\theta_i^0 - \theta_i^n)^j}{j!} \frac{\partial^j f}{\partial \theta^j}(x|\theta_i^n, \sigma_i^0) \right] \\ &\quad + R(x) \\ &= \sum_{i=1}^2 \left[(p_i^n - p_i^0) - p_i^0 \sum_{j=1}^{[r]+1} \frac{(\theta_i^0 - \theta_i^n)^j}{j!(\sigma_i^0)^j} \right] f(x|\theta_i^n, \sigma_i^0) + R(x), \end{aligned}$$

where the last inequality is due to the identity (3.8) and $R(x)$ is the remainder of Taylor expansion. Note that

$$\sup_{x \notin \{\theta_1^0, \theta_2^0\}} |R(x)| / W_1^r(G_n, G_0) \leq \sum_{i=1}^2 O(|\theta_i^n - \theta_i^0|^{[r]+1+\delta}) / |\theta_i^n - \theta_i^0|^r \rightarrow 0.$$

Now, we choose $p_1^n = p_1^0 + 1/n$, $p_2^n = p_2^0 - 1/n$, which means $p_1^n + p_2^n = p_1^0 + p_2^0$ and $p_1^n \rightarrow p_1^0$, $p_2^n \rightarrow p_2^0$. As $p_i^0 / j!(\sigma_i^0)^j$ are fixed positive constants for all $1 \leq j \leq [r] + 1$. It is clear that there exists sequences θ_1^n and θ_2^n such that for both $i = 1$ and $i = 2$, $\theta_i^n - \theta_i^0 \rightarrow 0$, the identity $p_i^0 \sum_{j=1}^{[r]+1} \frac{(\theta_i^0 - \theta_i^n)^j}{j!(\sigma_i^0)^j} = p_i^n - p_i^0$ holds for all n (sufficiently large).

With these choices of $p_1^n, p_2^n, \theta_1^n, \theta_2^n$, we have

$$\sup_{x \notin \{\theta_1^0, \theta_2^0\}} |p_{G_n}(x) - p_{G_0}(x)|/W_1^r(G_n, G_0) = \sup_{x \notin \{\theta_1^0, \theta_2^0\}} |R(x)|/W_1^r(G_n, G_0) \rightarrow 0.$$

The rest of the proof proceeds in the same way as that of Prop. 3.5.1 part (b). \square

3.5.3 Proofs for remaining results

PROOF OF LEMMA 3.2.1 For any set \mathcal{M} , a set \mathcal{M}_ϵ is called an ϵ -net over \mathcal{M} if any element of \mathcal{M} is within ϵ distance of some metrics from an element of \mathcal{M}_ϵ . It is a known fact that we can choose an ϵ -net \mathcal{S}_1 over the k -dimensional simplex for the l_1 norm such that $|\mathcal{S}_1| \leq \left(\frac{5}{\epsilon}\right)^k$, where $|\cdot|$ denotes the cardinality of a set (e.g see Lemma A.4 of [Ghosal and van der Vaart \[2001\]](#)). Additionally, if we denote \mathcal{S}_2 to be $2d\epsilon$ -net of Ω under metric $\|\cdot\|$, then we can verify that $|\mathcal{S}_2| \leq \left(\frac{2d\bar{\lambda}}{\epsilon}\right)^{d(d+1)/2}$.

Now, denote \mathcal{S}_3 to be the set of all $p_G \in \mathcal{P}_k(\Theta^*)$ such that G is supported on $((\pm l_1\epsilon, \pm l_2\epsilon, \dots, \pm l_d\epsilon), \Sigma)$, where $\Sigma \in \mathcal{S}_2$, $0 \leq l_i \leq \frac{a}{\epsilon}$ for all $1 \leq i \leq d$, with weights come from \mathcal{S}_1 only. For each p_G in $\mathcal{P}_k(\Theta^*)$, we firstly move the support points of G to their closest support points in $((\pm l_1\epsilon, \pm l_2\epsilon, \dots, \pm l_d\epsilon), \Sigma)$ to form \tilde{G} and then we move the masses of \tilde{G} to their closest masses in \mathcal{S}_1 to form G^* . By means of triangle inequality, we obtain

$$\|p_G - p_{G^*}\|_\infty \leq \|p_G - p_{\tilde{G}}\|_\infty + \|p_{\tilde{G}} - p_{G^*}\|_\infty.$$

Due to the boundness of kernel density function $f(x|\theta, \Sigma)$, it is not hard to verify that $\|p_{\tilde{G}} - p_{G^*}\|_\infty \lesssim \epsilon$. Additionally, denote $G = \sum_{i=1}^{k_1} p_i \delta_{\{\theta_i, \Sigma_i\}}$ where $k_1 \leq k$. Then, $\tilde{G} = \sum_{i=1}^{k_1} p_i \delta_{\{\tilde{\theta}_i, \tilde{\Sigma}_i\}}$ where $(\tilde{\theta}_i, \tilde{\Sigma}_i)$ has the form $((\pm l_1\epsilon, \pm l_2\epsilon, \dots, \pm l_d\epsilon), \Sigma)$ and $\Sigma \in \mathcal{S}_2$.

By means of triangle inequality, we obtain

$$\|p_G - p_{\tilde{G}}\|_\infty \leq \sum_{i=1}^{k_1} p_i \left(\|f(x|\theta_i, \Sigma_i) - f(x|\tilde{\theta}_i, \Sigma_i)\|_\infty + \|f(x|\tilde{\theta}_i, \Sigma_i) - f(x|\tilde{\theta}_i, \tilde{\Sigma}_i)\|_\infty \right).$$

As $|\tilde{\Sigma}_i|$ is bounded for all $1 \leq i \leq k_1$, by means of mean value theorem, we achieve $\|f(x|\theta_i, \Sigma_i) - f(x|\tilde{\theta}_i, \Sigma_i)\|_\infty \lesssim \|\theta_i - \tilde{\theta}_i\| \lesssim \epsilon$. Similarly, by means of Taylor expansion up to the first order regarding $\Sigma_i, \tilde{\Sigma}_i$ and Cauchy-Schwarz's inequality, we have $\|f(x|\tilde{\theta}_i, \Sigma_i) - f(x|\tilde{\theta}_i, \tilde{\Sigma}_i)\|_\infty \lesssim \|\Sigma_i - \tilde{\Sigma}_i\| \lesssim \epsilon$. Therefore, $\|p_G - p_{G^*}\|_\infty \lesssim \epsilon$. As a consequence, the cardinality of \mathcal{S}_3 is bounded as

$$|\mathcal{S}_3| \leq \left(\frac{2d\bar{\lambda}}{\epsilon} \right)^{d(d+1)k/2} \times \left(\frac{2a}{\epsilon} \right)^{dk} \times \left(\frac{5}{\epsilon} \right)^k.$$

Hence, for some constants c_1 and c_2 ,

$$\log N(c_1\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq \log |\mathcal{S}_3| \lesssim \log(1/\epsilon),$$

which proves (3.4).

To establish (3.5), let $\eta \leq \epsilon$ to be chosen later. From the assumption, it also indicates that $a \lesssim (\log(1/\eta))^\gamma$. Denote f_1, f_2, \dots, f_N to be an η -net for $\|\cdot\|_\infty$ over $\mathcal{P}_k(\Theta^*)$. Notice that as $\Sigma \in \Omega$, $|\Sigma| \geq \underline{\lambda}^{2d}$ and as $\|x\| \geq 2\sqrt{da}$,

$$(x - \theta)^T \Sigma^{-1} (x - \theta) \geq \frac{\|x - \theta\|^2}{\lambda_d(\Sigma)} \geq \frac{\|x - \theta\|^2}{\bar{\lambda}^2} \geq \frac{(\|x\| - \|\theta\|)^2}{\bar{\lambda}^2} \geq \frac{\|x\|^2}{4\bar{\lambda}^2}.$$

Therefore, by defining

$$H(x) = \begin{cases} \frac{1}{(2\pi)^{d/2} \underline{\lambda}^d} \exp\left(-\frac{\|x\|^2}{8\bar{\lambda}^2}\right), & \text{if } \|x\| \geq 2\sqrt{da} \\ \frac{1}{(2\pi)^{d/2} \underline{\lambda}^d}, & \text{otherwise,} \end{cases}.$$

we obtain $H(x)$ is an envelope for $\mathcal{P}_k(\Theta^*)$. We construct the brackets $[p_i^L, p_i^U]$ as follows

$$p_i^L(x) = \max \{f_i(x) - \eta, 0\}, \quad p_i^U(x) = \min \{f_i(x) + \eta, H(x)\}.$$

It is clear that $\mathcal{P}_k(\Theta^*) \subset \bigcup_{i=1}^N [p_i^L, p_i^U]$. Additionally, $p_i^U(x) - p_i^L(x) \leq \min \{2\eta, H(x)\}$. As a consequence, for any $B \geq 2\sqrt{d}\eta$, we have

$$\int_{\mathbb{R}^d} (p_i^U(x) - p_i^L(x)) dx \leq \int_{\|x\| < B} 2\eta dx + \int_{\|x\| \geq B} H(x) dx.$$

By means of spherical coordinates, we obtain $\int_{\|x\| \geq B} H(x) dx \lesssim B^{d-1} \exp\left(-\frac{B^2}{8\bar{\lambda}^2}\right)$. Additionally, we also have

$$\int_{\|x\| \leq B} 2\eta dx \lesssim \eta \int_0^B R^{d-1} dR \lesssim \eta B^d.$$

By choosing $B = \max \left\{ 2\sqrt{d}L, \sqrt{8\bar{\lambda}} \right\} (\log(1/\eta))^\gamma$, then it is clear that

$$B^{d-1} \exp\left(-\frac{B^2}{8\bar{\lambda}^2}\right) \lesssim \eta (\log(1/\eta))^{(d-1)\gamma}, \quad \eta B^d \lesssim \eta (\log(1/\eta))^{d\gamma}.$$

Thus,

$$\int_{\mathbb{R}^d} (p_i^U(x) - p_i^L(x)) dx \lesssim \eta (\log(1/\eta))^{d\gamma}.$$

With this result and that of (3.4), they imply that for some positive constant c

$$H_B \left(c\eta (\log(1/\eta))^{d\gamma}, \mathcal{P}_k(\Theta^*), \|\cdot\|_1 \right) \leq N \lesssim \log(1/\eta).$$

By choosing $\epsilon = c\eta (\log(1/\eta))^{d\gamma}$, note that $\log(1/\eta) \sim \log(1/\epsilon)$. Therefore,

$$H_B(\epsilon, \mathcal{P}_k(\Theta^*), \|\cdot\|_1) \lesssim \log(1/\epsilon).$$

As we have $h^2 \leq \|\cdot\|_1$, the above result implies that $H_B(\sqrt{\epsilon}, \mathcal{P}_k(\Theta^*), h) \lesssim \log(1/\epsilon)$. Therefore,

$$H_B(\epsilon, \mathcal{P}_k(\Theta^*), h) \lesssim \log(1/\epsilon),$$

which proves (3.5).

PROOF OF PROPOSITION 3.2.3 We only consider the case $k - k_0 = 1$ (the proof for the case $k - k_0 = 2$ is rather similar, therefore it is omitted). Since $k - k_0 = 1$, from Proposition 3.2.1, we have $\bar{r} = 4$. As in the proof of Proposition 3.2.2, it suffices to show for $d = 1$ that

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_k} \left\{ \sup_{x \in \mathcal{X}} |p_G(x) - p_{G_0}(x)| / W_{\bar{r}}^{\bar{r}}(G, G_0) : W_{\bar{r}}(G, G_0) \leq \epsilon \right\} > 0. \quad (3.31)$$

Denote $v = \sigma^2$. Assume that the above result does not hold, i.e. we can find a sequence of $G_n = \sum_{i=1}^{k_0+m} \sum_{j=1}^{s_i} p_{ij}^n \delta_{(\theta_{ij}^n, v_{ij}^n)} \rightarrow G_0$ in $W_{\bar{r}}$ where $(p_{ij}^n, \theta_{ij}^n, v_{ij}^n) \rightarrow (p_i^0, \theta_i^0, v_i^0)$ for all $1 \leq i \leq k_0 + m$, $1 \leq j \leq s_i$ and $p_i^0 = 0$ as $k_0 + 1 \leq i \leq k_0 + m$. As $k - k_0 = 1$, we have $0 \leq m \leq 1$. Note that since we do not have the constraints on the masses of mixing measures G_n as those in part (b) of Proposition 3.2.2, there are some atoms of G_n that may converge to some limit points outside the set of atoms of G_0 . That is the reason why we define the possible additional atom $(p_{k_0+1}^0, \theta_{k_0+1}^0, v_{k_0+1}^0)$. Repeating the same arguments as the proof of Proposition 3.2.2 up to step 8 when we have the assumption that $E_\alpha(\theta_i^0, v_i^0) \rightarrow 0$ for all $1 \leq i \leq k_0 + m$ and $0 \leq \alpha \leq 2\bar{r}$ as $n \rightarrow \infty$.

Now, we can find an index $i^* \in \{1, 2, \dots, k_0 + m\}$ such that as $n \rightarrow \infty$

$$\sum_{j=1}^{s_{i^*}} p_{i^*j}^n (|\Delta\theta_{i^*j}^n|^{\bar{r}} + |\Delta v_{i^*j}^n|^{\bar{r}}) / D_n \not\rightarrow 0.$$

where $D_n = \sum_{i=1}^{k_0+m} \sum_{j=1}^{s_i} p_{ij}^n (|\Delta\theta_{ij}^n|^{\bar{r}} + |\Delta v_{ij}^n|^{\bar{r}}) + \sum_{i=1}^{k_0+m} |p_i^n - p_i^0|$. Since $E_{2\bar{r}}(\theta_{i^*}^0, v_{i^*}^0) \rightarrow 0$ for all $1 \leq i \leq k_0 + m$, it implies that $\sum_{j=1}^{s_{i^*}} p_{i^*j}^n |\Delta v_{i^*j}^n|^{\bar{r}} / D_n \rightarrow 0$. Therefore, we obtain

$$\sum_{j=1}^{s_{i^*}} p_{i^*j}^n |\Delta v_{i^*j}^n|^{\bar{r}} / \sum_{j=1}^{s_{i^*}} p_{i^*j}^n (|\Delta\theta_{i^*j}^n|^{\bar{r}} + |\Delta v_{i^*j}^n|^{\bar{r}}) \rightarrow 0.$$

It implies that

$$\sum_{j=1}^{s_{i^*}} p_{i^*j}^n |\Delta\theta_{i^*j}^n|^{\bar{r}} / \sum_{j=1}^{s_{i^*}} p_{i^*j}^n (|\Delta\theta_{i^*j}^n|^{\bar{r}} + |\Delta v_{i^*j}^n|^{\bar{r}}) \rightarrow 1.$$

As a consequence,

$$\begin{aligned} F'_\alpha(\theta_{i^*}^0, v_{i^*}^0) &= \frac{\sum_{j=1}^{s_{i^*}} p_{i^*j}^n (|\Delta\theta_{i^*j}^n|^{\bar{r}} + |\Delta v_{i^*j}^n|^{\bar{r}})}{\sum_{j=1}^{s_{i^*}} p_{i^*j}^n |\Delta\theta_{i^*j}^n|^{\bar{r}}} F_\alpha(\theta_{i^*}^0, v_{i^*}^0) \\ &= \frac{\sum_{j=1}^{s_{i^*}} p_{i^*j}^n \sum_{n_1, n_2} \frac{(\Delta\theta_{i^*j}^n)^{n_1} (\Delta v_{i^*j}^n)^{n_2}}{n_1! n_2!}}{\sum_{j=1}^{s_{i^*}} p_{i^*j}^n |\Delta\theta_{i^*j}^n|^4} \rightarrow 0, \end{aligned} \quad (3.32)$$

where $n_1 + 2n_2 = \alpha$ and $1 \leq \alpha \leq 4$. As $i^* \in \{1, 2, \dots, k_0 + m\}$, we have $i^* \in \{1, \dots, k_0\}$ or $i^* \in \{k_0 + 1, \dots, k_0 + m\}$. Firstly, we assume that $i^* \in \{1, \dots, k_0\}$. Without loss of generality, let $i^* = 1$. Since $s_1 \leq k - k_0 + 1 = 2$, there are two possibilities.

Case 1 If $s_1 = 1$, then $F'_1(\theta_1^0, v_1^0) = \Delta\theta_{11}^n / |\Delta\theta_{11}^n|^4 \not\rightarrow 0$, which is a contradiction.

Case 2 If $s_1 = 2$, without loss of generality, we assume that $p_{11}^n |\Delta\theta_{11}^n| \leq p_{12}^n |\Delta\theta_{12}^n|$ for infinitely many n , which we can assume to hold for all n (by choosing the subsequence). Since $p_{11}^n (\Delta\theta_{11}^n)^4 + p_{12}^n (\Delta\theta_{12}^n)^4 > 0$, we obtain $\theta_{12}^n \neq 0$ for all n . If $\Delta\theta_{11}^n = 0$ for infinitely many n , then $F'_1(\theta_1^0, v_1^0) = \Delta\theta_{12}^n / (\Delta\theta_{12}^n)^4 \not\rightarrow 0$, which is a contradiction. Therefore, we may assume $\theta_{11}^n \neq 0$ for all n . Let $a := \lim_{n \rightarrow \infty} p_{11}^n \Delta\theta_{11}^n / p_{12}^n \Delta\theta_{12}^n \in [-1, 1]$. Dividing both the numerator and denominator of $F'_1(\theta_1^0, v_1^0)$ by $p_{12}^n \Delta\theta_{12}^n$ and letting $n \rightarrow \infty$, we obtain $a = -1$. Consider the following scenarios regarding p_{11}^n / p_{12}^n :

(i) If $p_{11}^n / p_{12}^n \rightarrow \infty$, then $\Delta\theta_{11}^n / \Delta\theta_{12}^n \rightarrow 0$. Since $\Delta\theta_{11}^n, \Delta\theta_{12}^n \neq 0$, denote $\Delta v_{11}^n = k_1^n (\Delta\theta_{11}^n)^2$, $\Delta v_{12}^n = k_2^n (\Delta\theta_{12}^n)^2$ for all n . Now, by dividing the numerator and denominator of $F'_2(\theta_1^0, v_1^0), F'_3(\theta_1^0, v_1^0), F'_4(\theta_1^0, v_1^0)$ by $p_{12}^n (\Delta\theta_{12}^n)^2$, $p_{12}^n (\Delta\theta_{12}^n)^3$, and $p_{12}^n (\Delta\theta_{12}^n)^4$ respectively, we obtain

$$\begin{aligned} M_{n,1} &= \frac{1}{2} + k_2^n + k_1^n \frac{p_{11}^n (\Delta\theta_{11}^n)^2}{p_{12}^n (\Delta\theta_{12}^n)^2} \rightarrow 0, \\ M_{n,2} &= \frac{1}{3!} + k_2^n + k_1^n \frac{p_{11}^n (\Delta\theta_{11}^n)^3}{p_{12}^n (\Delta\theta_{12}^n)^3} \rightarrow 0, \\ M_{n,3} &= \frac{1}{4!} + \frac{k_2^n}{2} + \frac{(k_2^n)^2}{2} + \left(\frac{k_1^n}{2} + \frac{(k_1^n)^2}{2} \right) \frac{p_{11}^n (\Delta\theta_{11}^n)^4}{p_{12}^n (\Delta\theta_{12}^n)^4} \rightarrow 0. \end{aligned}$$

If $|k_1^n|, |k_2^n| \rightarrow \infty$ then $M_{n,3} > \frac{1}{4!}$ for sufficiently large n , which is a contradiction.

Therefore, at least one of $|k_1^n|, |k_2^n|$ does not converge to ∞ . If $|k_1^n| \rightarrow \infty$ and $|k_2^n| \not\rightarrow \infty$ then $M_{n,1}$ implies that $|k_1^n \frac{p_{11}^n (\Delta\theta_{11}^n)^2}{p_{12}^n (\Delta\theta_{12}^n)^2}| \not\rightarrow \infty$. Therefore, $|k_1^n \frac{p_{11}^n (\Delta\theta_{11}^n)^3}{p_{12}^n (\Delta\theta_{12}^n)^3}| \rightarrow 0$ as $\Delta\theta_{11}^n / \Delta\theta_{12}^n \rightarrow 0$ and $k_1^n \frac{(\Delta\theta_{11}^n)^2}{(\Delta\theta_{12}^n)^2} \rightarrow 0$ as $p_{11}^n / p_{12}^n \rightarrow \infty$. Combining these results with $M_{n,3}, M_{n,4}$, we get $k_2^n + \frac{1}{3!} \rightarrow 0$ and $\frac{1}{4!} + \frac{k_2^n}{2} + \frac{(k_2^n)^2}{2} \rightarrow 0$, which cannot happen. If $|k_1^n| \not\rightarrow \infty$, then $M_{n,1}$ and $M_{n,2}$ implies that $k_2^n + 1/2 \rightarrow 0$ and $k_2^n + 1/6 \rightarrow 0$, which cannot happen either. As a consequence, $p_{11}^n / p_{12}^n \not\rightarrow \infty$.

(ii) If $p_{11}^n / p_{12}^n \rightarrow 0$ then $p_{12}^n / p_{11}^n \rightarrow \infty$. Since $p_{11}^n \Delta\theta_{11}^n / p_{12}^n \Delta\theta_{12}^n \rightarrow -1$, we have $|\Delta\theta_{11}^n / \Delta\theta_{12}^n| \rightarrow \infty$ or equivalently $\Delta\theta_{12}^n / \Delta\theta_{11}^n \rightarrow 0$. From here, using the same argu-

ment as that above, we are also led to a contradiction. So, $p_{11}^n/p_{12}^n \not\rightarrow 0$.

(iii) If $p_{11}^n/p_{12}^n \rightarrow b \notin \{0, \infty\}$. It also means that $\Delta\theta_{11}^n/\Delta\theta_{12}^n \rightarrow -1/b$. Therefore, by dividing the numerator and denominator of $F_2'(\theta_1^0, v_1^0), F_3'(\theta_1^0, v_1^0), F_4'(\theta_1^0, v_1^0)$ by $p_{12}^n(\Delta\theta_{12}^n)^2, p_{12}^n(\Delta\theta_{12}^n)^3$, and $p_{12}^n(\Delta\theta_{12}^n)^4$ and let $n \rightarrow \infty$, we arrive at the scaling system of equations (4.24) when $r = 4$ for which we already know that non-trivial solution does not exist. Hence, the case $s_1 = 2$ cannot happen.

As a consequence, $i^* \notin \{1, \dots, k_0\}$. However, since $m \leq 1$, we have $i^* = k_0 + 1$. This implies that $s_{k_0+1} = 1$, which we already know from Case 1 that (3.32) cannot hold. Therefore, our hypothesis that all coefficients $E_\alpha(\theta_i^0, v_i^0)$ vanish does not hold — there must be at least one coefficient which does not converge to 0 as $n \rightarrow \infty$. Repeating the same argument as step 9 in the proof of Proposition 3.2.2, we achieve the conclusion of our result.

CHAPTER IV

Singularity structures and impacts on parameter estimation in finite mixtures of distributions

Singularities of a statistical model are the elements of the model's parameter space which make the corresponding Fisher information matrix degenerate. These are the points for which estimation techniques such as the maximum likelihood estimator and standard Bayesian procedures do not admit the root- n parametric rate of convergence. We propose a general framework for the identification of singularity structures of the parameter space of finite mixtures, and study the impacts of the singularity levels on minimax lower bounds and rates of convergence for the maximum likelihood estimator over a compact parameter space. Our study makes explicit the deep links between model singularities, parameter estimation convergence rates and minimax lower bounds, and the algebraic geometry of the parameter space for mixtures of continuous distributions. The theory is applied to establish concrete convergence rates of parameter estimation for finite mixture of skewnormal distributions. This rich and increasingly popular mixture model is shown to exhibit a remarkably complex range of asymptotic behaviors which have not been hitherto reported in the literature.¹

¹This chapter has been published in [Ho and Nguyen, 2016d].

4.1 Introduction

In the standard asymptotic theory of parametric estimation, a customary regularity assumption is the non-singularity of the Fisher information matrix defined by the statistical model (see, for example, [Lehmann and Casella \[1998\]](#) (pg. 124); or [van der Vaart \[1998\]](#), Sec. 5.5). This condition leads to the cherished root- n consistency, and in many cases the asymptotic normality of parameter estimates. When the non-singularity condition fails to hold, that is, when the true parameters represent a singular point in the statistical model, very little is known about the asymptotic behavior of their estimates.

The singularity situation might have been brushed aside as idiosyncratic by some parametric statistical modelers in the past. As complex and high-dimensional models are increasingly embraced by statisticians and practitioners alike, singularities are no longer a rarity — they start to take a highly visible place in modern statistics. For example, the many zeros present in a high-dimensional linear regression problem represent a type of singularities of the underlying model, points corresponding to rank-deficient Fisher information matrices [[Hastie et al., 2015](#)]. In another example, the zero skewness in the family of skewed distributions represents a singular point [[Chiogna, 2005](#)]. In both examples, singularity points are quite easy to spot out — it is the impacts of their presence on improved parameter estimation procedures and the asymptotic properties such procedures entail that are nontrivial matters occupying the best efforts of many researchers in the past decade. The textbooks by [Bühlmann and van de Geer \[2011\]](#), [Hastie et al. \[2015\]](#), for instance, address such issues for high-dimensional regression problems, while the recent papers by [Ley and Paindaveine \[2010\]](#), [Hallin and Ley \[2012, 2014\]](#) investigate statistical inference in the skewed families for distribution. By contrast, with finite mixture models — a popular and rich class of modeling tools for density estimation and heterogeneity

inference [Lindsay, 1995] and a subject of this chapter, the singularity phenomenon is not quite well understood, to the best our knowledge, except for specific instances.

One of the simplest instances is the singularity of Fisher information matrix in an (overfitted) finite mixture that includes a homogeneous distribution. Lee and Chesher [1986] analyzed a test of heterogeneity based on finite mixtures, addressing the challenge arising from the aforementioned singularity. Recent works on the related topic include Chen and Chen [2003], Kasahara and Shimotsu [2014b]. Rottnitzky et al. [2000] investigated likelihood-based parameter estimation in a somewhat general parametric modeling framework, subject to the constraint that the Fisher information matrix is one rank deficient. For overfitted finite mixtures, Chen [1995] showed that under a condition of strong identifiability, there are estimators which achieve the generic convergence rate $n^{-1/4}$ for parameter estimation. Recent works also established generic behaviors of estimation under somewhat broader settings of overfitted finite mixture models with both maximum likelihood estimation and Bayesian estimation [Rousseau and Mengersen, 2011, Nguyen, 2013, Ho and Nguyen, 2016c].

The family of mixture models is far too rich to submit to a uniform kind of behavior of parameter estimation. In fact, it was shown only recently that even classical models such as the location-scale Gaussian mixtures, and the shape-rate Gamma mixtures, do not admit such a generic rate of convergence for an estimation method such as MLE [Ho and Nguyen, 2016a]. For instance, singularities arise in the finite mixtures of Gamma distributions, even when the number of mixing components is known — this phenomenon results in an extremely slow convergence behavior for the model parameters lying in the vicinity of singular points, eventhough such parameters are (perfectly) identifiable. Finite mixtures of Gaussian distributions, though identifiable, exhibit both minimax lowerbounds and maximum likelihood estimation rates that are directly linked to the solvability of a system of real polynomial equations, rates

which deteriorate quickly with the increasing number of extra mixing components. The results obtained for such specific instances contain considerable insights about parameter estimation in finite mixture models, but they only touch upon the surface of a more general phenomenon. Indeed, as we shall see there is a much richer spectrum of asymptotic behavior in which regular (non-singular) mixtures, strongly identifiable mixtures, and weakly identifiable mixture models (such as the one studied by [Ho and Nguyen \[2016a\]](#)) occupy but a small spot.

Objectives and main results In this chapter we propose a theoretical framework for analyzing parameter estimation behavior in finite mixture models, addressing directly the situations where the non-singularity condition of the Fisher information matrix may not hold. Our approach is to take on a systematic investigation of the singularity structure of a compact and multi-dimensional parameter space of mixture models, and then study the impacts of the presence of singularities on parameter estimation. It is no longer sufficient to speak of the standard notion of Fisher information singularities. A more fundamental notion that we introduce is called *singularity level*, a natural or infinite value given to every element in the parameter space. Fisher information singularities simply correspond to points in the parameter space whose singularity level is non-zero. Within the set of Fisher information singularities the parameter space can be partitioned into disjoint subsets determined by different singularity levels. The singularity level describes in a precise manner the variation of the mixture likelihood with respect to changes in model parameters. This concept enables us to quantify the varying degrees of identifiability and singularity, some of which were implicitly exploited in previous works mentioned above.

The statistical implication of the singularity level is easy to describe: given an i.i.d. n -sample from a (true) mixture model, a parameter value of singularity level r admits $n^{-1/2(r+1)}$ minimax lower bound for any estimator tending to the true param-

eter(s), as well as the same maximum likelihood estimator's convergence rate (up to a logarithmic factor and under some conditions). Thus, singularity level 0 results in root- n convergence rate for parameter estimation. Fisher singularity corresponds to singularity level 1 or greater than 1, resulting in convergence rates $n^{-1/4}, n^{-1/6}, n^{-1/8}$ or so on. The detailed picture of the distribution of singularity levels, however, can be extremely complex to capture. Remarkably, there are examples of finite mixtures for which the compact parameter space can be partitioned into disjoint subsets whose singularity level ranges from 0 to 1 to 2, . . . , up to infinity. As a result, if we were to vary the true parameter values, we would encounter a phenomenon akin to that of “phase transition” on the statistical efficiency of parameter estimation occurring within the same model class.

Techniques A major component of our general framework is a procedure for characterizing subsets of points carrying the same singularity level. It will be shown that these points are in fact a subset of a real affine variety. A real affine variety is a set of solutions to a system of real polynomial equations. The polynomial equations can be derived explicitly by the kernel density functions that define a given mixture distribution. The study of the solutions of polynomial equations is a central subject of algebraic geometry [Sturmfels, 2002, Cox et al., 2007]. The connections between statistical models and algebraic geometry have been studied for discrete Markov random fields [Drton et al., 2009], as well as finite mixtures of categorical data [Allman et al., 2009]. For finite mixtures of continuous distributions, the link to algebraic geometry is distilled from a new source of algebraic structure, in addition to the presence of mixing measures: it is traced to the partial differential equations satisfied by the mixture model's kernel density function. For Gaussian mixtures, it is the relation captured by Eq. (4.3) for the Gaussian kernel. The partial differential equations can be nonlinear, with coefficients given by rational functions defined in terms of model

parameters. It is this relation that is primarily responsible for the complexity of the singularity structure. A quintessential example of such a relation is given by Eq. (4.2) for the skewnormal kernel densities.

Although our method for the analysis of singularity structure and the asymptotic theory for parameter estimation can be used to re-derive old and existing results such as those of [Chen \[1995\]](#), [Ho and Nguyen \[2016a\]](#), a substantial outcome is to establish new results on mixture models for which no asymptotic theory have hitherto been achieved. This leads us to a story of finite mixtures of skewnormal distributions. The skewnormal distribution was originally proposed in [Azzalini \[1986\]](#), [Azzalini and Valle \[1996\]](#), [Azzalini and Capitanio \[1999\]](#). The skewnormal generalizes normal (Gaussian) distribution, which is enhanced by the capability of handling asymmetric (skewed) data distributions. Due to its more realistic modeling capability for multi-modality and asymmetric components, skewnormal mixtures are increasingly adopted in recent years for model based inference of heterogeneity by many researchers [[Lin et al., 2007](#), [Arellano-Valle et al., 2008, 2009](#), [Lin, 2009](#), [Schnatter and Pyne, 2009](#), [Ghosal and Roy, 2011](#), [Lee and McLachlan, 2013](#), [Prates et al., 2013](#), [Canale and Scarpa, 2015](#), [Zeller et al., 2015](#)]. Due to its usefulness, a thorough understanding of the asymptotic behavior of parameter estimation for skewnormal mixtures is also of interest in its own right.

The singularity structure of the skewnormal mixtures is perhaps one of the more complex among the parametric mixture models that we have typically encountered in the literature. By comparison, strongly identifiable models admit the same singularity level (1, to be precise) for all parameter values residing in a compact space, resulting in $n^{-1/4}$ convergence rate for the MLE. Most mixture models whose kernel density function has only one type of parameter, such as location mixtures or scale mixtures, are in this category. Location-scale Gaussian mixtures are a step up in the complexity, in that all their parameter values carry the same singularity level, which depends only

on the number of extra mixing components. Yet this is not the picture of skewnormal mixtures. We will be able to identify subsets with singularity level 0, 1, 2, … all the way up to infinity. Even in the setting of mixtures with known number of mixing components, the singularity structure is remarkably complex. Thus, the results for skewnormal mixtures present an useful illustration for the full power of the general theory for finite mixtures of continuous distributions.

The source of complexity of skewnormal mixtures is the structure of the skewnormal kernel density. The evidence for the latter was already made clear by [Chiogna \[2005\]](#), [Ley and Paindaveine \[2010\]](#), [Hallin and Ley \[2012, 2014\]](#), whose works provided a thorough picture of the singularities for the class of skewnormal densities, and their impacts on the non-standard rates of convergence of MLE. Not only can we recover the results of [Hallin and Ley \[2012, 2014\]](#) in terms of rates of convergence, which correspond to a trivial “mixture” that has exactly one skewnormal component, an entirely new set of results are established for mixtures of two or more components. It is in this setting that new types of singularities arise out of the interactions between distinct skewnormal components. These interactions define the subset of singular points of a given level that can be characterized by a system of real polynomial equations. This algebraic geometric characterteration allows us to establish either the precise singularity level or an upper bound for the mixture model’s entire parameter space.

The plan for the remainder of our chapter is as follows. Section 4.2 lays out the notations and relevant concepts such as parameter spaces and the underlying geometries. Section 4.3 presents the general framework of analysis of singularity structure, and the impact on convergence rates of parameter estimation for singular points of a given singularity level. Section 4.4 and Section 4.5 illustrate the theory on the finite mixture of skewnormal distributions, by giving concrete minimax bounds and MLE convergence rates for this class of models for the first time. We conclude

with a discussion in Section 6.6. Further details of the proofs and some additional results are given in the Appendices.

4.2 Background

A finite mixture of continuous distributions admits density of the form $p_G(x) = \int f(x|\eta)dG(\eta)$ with respect to Lebesgue measure on an Euclidean space for x , where $f(x|\eta)$ denotes a probability density kernel, η is a multi-dimensional parameter taking values in a subset of an Euclidean space Θ , G denotes a discrete mixing distribution on Θ . The number of support points of G represents the number of mixing components in mixture model. Suppose that $G = \sum_{i=1}^k p_i \delta_{\eta_i}$, then $p_G(x) = \sum_{i=1}^k p_i f(x|\eta_i)$.

4.2.1 Parameter spaces and geometries

There are different kinds of parameter space and geometries that they carry which are relevant to our work. We proceed to describe them in the following.

Natural parameter space The customarily defined parameter space of the k -mixture of distributions is that of the mixing component parameters η_i , and mixing probabilities p_i . Throughout this chapter, it is assumed that $\eta_i \in \Theta$, which is a compact subset of \mathbb{R}^d for some $d \geq 1$, for $i = 1, \dots, k$. The mixing probability vector $\mathbf{p} = (p_1, \dots, p_k) \in \Delta^{k-1}$, the $(k-1)$ -probability simplex. To simplify the theory we will further assume (in Section 4.4) that all $p_i \geq c_0$ for some small positive constant c_0 . For the remainder of the chapter, we also use Ω to denote the compact subset of the Euclidean space for parameters $(\mathbf{p}, \boldsymbol{\eta})$.

Example 4.2.1. The skewnormal density kernel on the real line has three parameters $\eta = (\theta, \sigma, m) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}$, namely, the location, scale and skewness (shape)

parameters. It is given by, for $x \in \mathbb{R}$,

$$f(x|\theta, \sigma, m) := \frac{2}{\sigma} f\left(\frac{x-\theta}{\sigma}\right) \Phi(m(x-\theta)/\sigma),$$

where $f(x)$ is the standard normal density and $\Phi(x) = \int f(t)1(t \leq x) dt$. This generalizes the Gaussian density kernel, which corresponds to fixing $m = 0$. The parameter space for the k -mixture of skewnormals is therefore a subset of an Euclidean space for the mixing probabilities p_i and mixing component parameters $\eta_i = (\theta_i, v_i = \sigma_i^2, m_i) \in \mathbb{R}^3$. For each $i = 1, \dots, k$, θ_i, σ_i, m_i are restricted to reside in compact subsets $\Theta_1 \subset \mathbb{R}, \Theta_2 \subset \mathbb{R}_+, \Theta_3 \subset \mathbb{R}$ respectively, i.e., $\Theta = \Theta_1 \times \Theta_2 \times \Theta_3$.

Semialgebraic sets The singularity structure of the parameter space carries a different geometry. It will be described in terms of the zero sets (sets of solutions) of systems of real polynomial equations. The zero set of a system of real polynomial equations is called a (real) affine variety [Cox et al., 2007]. In fact, the sets which describe the singularity structure of finite mixtures are not affine varieties per se. We will see that they are the intersection between real affine varieties – the real-valued solutions of a finite collection of equations of the form $P(\mathbf{p}, \boldsymbol{\eta}) = 0$, and the set of parameter values satisfying $Q(\mathbf{p}, \boldsymbol{\eta}) > 0$, for some real polynomials P and Q . The intersection of these sets is also referred to as semialgebraic sets.

Example 4.2.2. Continuing on the example of skewnormal mixtures, we will see that first two types of singularities that arise from the mixture of skewnormals are solutions of the following polynomial equations

- (i) Type A: $P_1(\boldsymbol{\eta}) = \prod_{j=1}^k m_j$.
- (ii) Type B: $P_2(\boldsymbol{\eta}) = \prod_{1 \leq i \neq j \leq k} \left\{ (\theta_i - \theta_j)^2 + \left[\sigma_i^2(1 + m_j^2) - \sigma_j^2(1 + m_i^2) \right]^2 \right\}$.

These are just two among many more polynomials and types of singularities that we will be able to enumerate in the sequel. We quickly note that Type A refers to

the one (and only) kind of singularity intrinsic to the skewnormal kernel: $P_1 = 0$ if either one of the $m_j = 0$ — one of the skewnormal components is actually normal (symmetric). This type of singularity has received in-depth treatments by a number of authors [Chiogna, 2005, Ley and Paindaveine, 2010, Hallin and Ley, 2012, 2014]. On the other hand, Type B refers to something intrinsic to a mixture model, as it describes the relation of parameters of distinct mixing components i and j .

Space of mixing measures and transportation distance As described in the Introduction, a goal of this work is to turn the knowledge about the nature of singularities of parameter space Ω into the statistical efficiency of parameter estimation procedures. For this purpose, the convergence of parameters in a mixture model is most naturally analyzed in terms of the convergence in the space of mixing measures endowed by transportation distance (Wasserstein distance) metrics [Nguyen, 2013]. This is because the role played by parameters $\boldsymbol{p}, \boldsymbol{\eta}$ in the mixture model is via mixing measure G . It is mixing measure G that determines the mixture density p_G according to which the data are drawn from. Since the map $(\boldsymbol{p}, \boldsymbol{\eta}) \mapsto G(\boldsymbol{p}, \boldsymbol{\eta}) = G = \sum p_i \delta_{\boldsymbol{\eta}_i}$ is many-to-one, we shall treat a pair of parameter vectors $(\boldsymbol{p}, \boldsymbol{\eta}) = (p_1, \dots, p_k; \eta_1, \dots, \eta_k)$ and $(\boldsymbol{p}', \boldsymbol{\eta}') = (p'_1, \dots, p'_{k'}; \eta'_1, \dots, \eta'_{k'})$ to be equivalent if the corresponding mixing measures are equal, $G(\boldsymbol{p}, \boldsymbol{\eta}) = G(\boldsymbol{p}', \boldsymbol{\eta}')$.

For $r \geq 1$, the Wasserstein distance of order r between $G(\boldsymbol{p}, \boldsymbol{\eta})$ and $G(\boldsymbol{p}', \boldsymbol{\eta}')$ takes the form (cf. Villani [2003]),

$$W_r(G(\boldsymbol{p}, \boldsymbol{\eta}), G(\boldsymbol{p}', \boldsymbol{\eta}')) = \left(\inf \sum_{i,j} q_{ij} \|\eta_i - \eta'_j\|_r^r \right)^{1/r},$$

where $\|\cdot\|_r$ is the ℓ_r norm endowed by the natural parameter space, the infimum is taken over all couplings \boldsymbol{q} between \boldsymbol{p} and \boldsymbol{p}' , i.e., $\boldsymbol{q} = (q_{ij})_{ij} \in [0, 1]^{k \times k'}$ such that $\sum_{i=1}^{k'} q_{ij} = p_j$ and $\sum_{j=1}^k q_{ij} = p'_i$ for any $i = 1, \dots, k$ and $j = 1, \dots, k'$. (For the example of skewnormal mixtures, if $\boldsymbol{\eta} = (\theta, v, m)$ and $\boldsymbol{\eta}' = (\theta', v', m')$, then $\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|_r^r :=$

$$|\theta - \theta'|^r + |v - v'|^r + |m - m'|^r).$$

Suppose that a sequence of probability measures $G_n = \sum_i p_i^n \delta_{\eta_i^n}$ tending to G_0 under W_r metric at a rate $\omega_n = o(1)$. If all G_n have the same number of atoms $k_n = k_0$ as that of G_0 , then the set of atoms of G_n converge to the k_0 atoms of G_0 , up to a permutation of the atoms, at the same rate ω_n under $\|\cdot\|$. If G_n have the varying $k_n \in [k_0, k]$ number of atoms, where k is a fixed upper bound, then a subsequence of G_n can be constructed so that each atom of G_0 is a limit point of a certain subset of atoms of G_n — the convergence to each such limit also happens at rate ω_n . Some atoms of G_n may have limit points that are not among G_0 's atoms — the total mass associated with those “redundant” atoms of G_n must vanish at the generally faster rate ω_n^r .

4.2.2 Estimation settings

The impact of singularities on parameter estimation behavior is dependent on whether the mixture model is fitted with a known number of mixing components, or if only an upper bound on the number of mixing components is known. The former model fitting setting is called “e-mixtures” for short, while the latter “o-mixtures” (“e” for exact-fitted and “o” for over-fitted).

Specifically, given an i.i.d. n -sample X_1, X_2, \dots, X_n according to the mixture density $p_{G_0}(x) = \int f(x|\eta) G_0(d\eta)$, where $G_0 = G(\mathbf{p}^0, \boldsymbol{\eta}^0) = \sum_{i=1}^{k_0} p_i^0 \delta_{\eta_i^0}$ is unknown mixing measure with exactly k_0 distinct support points. We are interested in fitting a mixture of k mixing components, where $k \geq k_0$, using the n -sample X_1, \dots, X_n . In the e-mixture setting, $k = k_0$ is known, so an estimate G_n (such as the maximum likelihood estimate) is drawn from ambient space \mathcal{E}_{k_0} , the set of probability measures $G = G(\mathbf{p}, \boldsymbol{\eta})$ with exactly k_0 support points, where $(\mathbf{p}, \boldsymbol{\eta}) \in \Omega$. In the o-mixture setting, \hat{G}_n is drawn from ambient space \mathcal{O}_k , the set of probability measures $G = G(\mathbf{p}, \boldsymbol{\eta})$ with at most k support points, where $(\mathbf{p}, \boldsymbol{\eta}) \in \Omega$.

Assumption Throughout this chapter, several conditions on the kernel density $f(x|\eta)$ are assumed to hold. Firstly, the collection of kernel densities f as η varies is linearly independent. It follows that the mixture model is identifiable, i.e., $p_G(x) = p_{G_0}(x)$ for almost all x entails $G = G_0$. Secondly, we say $f(x|\eta)$ satisfies a uniform Lipschitz condition of order r , for some $r \geq 1$, if f as a function of η is differentiable up to order r , and that the partial derivatives with respect to η , namely $\partial^{|\kappa|} f / \partial \eta^\kappa$, for any $\kappa = (\kappa_1, \dots, \kappa_d) \in \mathbb{N}^d$ such that $|\kappa| := \kappa_1 + \dots + \kappa_d = r$ satisfy the following: for any $\gamma \in \mathbb{R}^d$,

$$\sum_{|\kappa|=r} \left| \left(\frac{\partial^{|\kappa|} f}{\partial \eta^\kappa}(x|\eta_1) - \frac{\partial^{|\kappa|} f}{\partial \eta^\kappa}(x|\eta_2) \right) \right| \gamma^\kappa \leq C \|\eta_1 - \eta_2\|_r^\delta \|\gamma\|_r^r$$

for some positive constants δ and C independent of x and $\eta_1, \eta_2 \in \Theta$. It is simple to verify that most kernel densities used in mixture modeling, including the skewnormal kernel, satisfy the uniform Lipschitz property over compact domain Θ , for any finite $r \geq 1$.

Notation We utilize several familiar notions of distance for mixture densities, with respect to Lebesgue measure μ . They include the total variation distance $V(p_G, p_{G_0}) = \frac{1}{2} \int |p_G(x) - p_{G_0}(x)| d\mu(x)$ and the Hellinger distance with formulation $h^2(p_G, p_{G_0}) = \frac{1}{2} \int \left(\sqrt{p_G(x)} - \sqrt{p_{G_0}(x)} \right)^2 d\mu(x)$.

4.3 Singularity structure in finite mixture models

4.3.1 Beyond Fisher information

Given a mixture model

$$\left\{ p_G(x) \middle| G = G(\boldsymbol{p}, \boldsymbol{\eta}) = \sum_{i=1}^k p_i \delta_{\eta_i}, (\boldsymbol{p}, \boldsymbol{\eta}) \in \Omega \right\}$$

from some given finite k and f a given kernel density (e.g., skewnormal), let l_G denote the score vector, that is, the column vector made of the partial derivatives of the log-likelihood function $\log p_G(x)$ with respect to each of the model parameters represented by $(\boldsymbol{p}, \boldsymbol{\eta})$. The Fisher information matrix is then given by $I(G) = \mathbb{E}(l_G l_G^\top)$, where the expectation is taken with respect to p_G . We say that the parameter vector $(\boldsymbol{p}, \boldsymbol{\eta})$ (and the corresponding mixing measure $G = G(\boldsymbol{p}, \boldsymbol{\eta})$) is a singular point in the parameter space (resp., ambient space of mixing measures), if $I(G)$ is degenerate. Otherwise, $(\boldsymbol{p}, \boldsymbol{\eta})$ (resp., G) is a non-singular point.

According to the standard asymptotic theory, if the true mixing measure G_0 is non-singular, *and* the number of mixing components $k_0 = k$ (that is, we are in the e-mixture setting), then basic estimators such as the MLE or Bayesian estimator yield the optimal root- n rate of convergence. Although the standard theory remains silent when $I(G_0)$ is degenerate, it is clear that the root- n rate may no longer hold. Moreover, there may be a richer range of behaviors for parameter estimation, requiring us to look into the deep structure of the zeros of $I(G_0)$. This will be our story for both settings of e-mixtures and o-mixtures. In fact, the Fisher information matrix $I(G_0)$ is no longer sufficient in assessing parameter estimation behaviors.

Example 4.3.1. To illustrate in the context of skewnormal mixtures, where parameter $\boldsymbol{\eta} = (\theta, v, m)$, observe that the mixture density structure allows the following characterization: $I(G)$ is degenerate if and only if the collection of partial derivatives

$$\left\{ \frac{\partial p_G(x)}{\partial p_j}, \frac{\partial p_G(x)}{\partial \eta_j} \right\} := \left\{ \frac{\partial p_G(x)}{\partial p_j}, \frac{\partial p_G(x)}{\partial \theta_j}, \frac{\partial p_G(x)}{\partial v_j}, \frac{\partial p_G(x)}{\partial m_j} \middle| j = 1, \dots, k \right\}$$

as functions of x are not linearly independent. This is equivalent to having that for some coefficients (α_{ij}) , $i = 1, \dots, 4$ and $j = 1, \dots, k$, not all of which are zeros, there

holds

$$\sum_{j=1}^k \alpha_{1j} f(x|\eta_j) + \alpha_{2j} \frac{\partial f}{\partial \theta}(x|\eta_j) + \alpha_{3j} \frac{\partial f}{\partial v}(x|\eta_j) + \alpha_{4j} \frac{\partial f}{\partial m}(x|\eta_j) = 0, \quad (4.1)$$

for almost all $x \in \mathbb{R}$. Lemma 4.4.1 later shows that the (Fisher information matrix's) singular points are the zeros of some polynomial equations.

We have seen that for the e-mixtures G is non-singular if the collection of density kernel functions $f(x|\eta)$ and their first partial derivatives with respect to each model parameter are linearly independent. This condition is also known as the first-order identifiability. For o-mixtures, the relevant notion is the second-order identifiability [Chen, 1995, Nguyen, 2013, Ho and Nguyen, 2016c]: the condition that the collection of kernel density functions $f(x|\eta)$, their first and second partial derivatives, are linearly independent. This condition fails to hold for skewnormal kernel densities, whose first and second partial derivatives are linked by the following nonlinear partial differential equations:

$$\begin{cases} \frac{\partial^2 f}{\partial \theta^2} - 2 \frac{\partial f}{\partial v} + \frac{m^3 + m}{v} \frac{\partial f}{\partial m} = 0, \\ 2m \frac{\partial f}{\partial m} + (m^2 + 1) \frac{\partial^2 f}{\partial m^2} + 2vm \frac{\partial^2 f}{\partial v \partial m} = 0. \end{cases} \quad (4.2)$$

The proof of these identities can be found in Lemma 4.8.1 in Appendix B. Note that if $m = 0$, the skewnormal kernel becomes normal kernel, which admits a (simpler) linear relationship:

$$\frac{\partial^2 f}{\partial \theta^2} = 2 \frac{\partial f}{\partial v}. \quad (4.3)$$

This relation plays a fundamental role in the analysis of finite mixtures of location-scale normal distributions [Ho and Nguyen, 2016a]. Compared to Gaussian density kernel, the nonlinear relationship exhibited by skewnormal density kernel results in a much richer behavior. Analyzing this requires the development of a more general

theory that we now embark on.

4.3.2 Behavior of likelihood in a Wasserstein neighborhood

Instead of dwelling on the Fisher information matrix, we shall employ a direct approach which studies the behavior of the likelihood function $p_G(x)$ as G varies in a Wasserstein neighborhood of a mixing measure $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\eta_i^0}$.

Fix $r \geq 1$, and consider a sequence of $G_n \in \mathcal{O}_k$, such that $W_r(G_0, G_n) \rightarrow 0$. Let $k_n \leq k$ be the number of distinct support points of G_n . Then each supporting atom η_i^0 as $i \in \{1, \dots, k_0\}$ of G_0 will have at least one atom of G_n that converges to. By relabelling the support points of G_n , we can express it as

$$G_n = \sum_{i=1}^{k_0} \sum_{j=1}^{s_i^n} p_{ij}^n \delta_{\eta_{ij}^n}, \quad (4.4)$$

where $\eta_{ij}^n \rightarrow \eta_i^0$ for all $i = 1, \dots, k_0$, $j = 1, \dots, s_i^n$. Additionally, $\sum_{i=1}^{k_0} s_i^n = k_n$. There exists a subsequence of G_n according to which k_n and all s_i^n are constant in n . (Note that for the setting of e-mixtures, the sequence of elements G_n is restricted to \mathcal{E}_{k_0} , so $k_n = k_0$ for all n . It follows that $s_i^n = 1$ for all $i = 1, \dots, k_0$. For o-mixtures, to simplify the presentation, we have omitted the cases where some G_n may have atoms that do not converge to the atoms of G_0). Thus, from here on we replace the sequence of G_n by this subsequence. To simplify the notation, n will be dropped from the superscript when the context is clear. In addition, we use the notation $\Delta\eta_{ij} := \eta_{ij} - \eta_i^0$ for $i = 1, \dots, k_0$, $j = 1, \dots, s_i$. Also, $p_{i.} := \sum_{j=1}^{s_i} p_{ij}$, and $\Delta p_{i.} := p_{i.} - p_i^0$, for $i = 1, \dots, k_0$. (For e-mixtures, since $s_i = 1$ for all i , the notation is simplified further: let $\Delta\eta_i := \Delta\eta_{i1} = \eta_i - \eta_i^0$, $\Delta p_i = \Delta p_{i.} = p_i - p_i^0$ for all $i = 1, \dots, k_0$.) The following lemma relates Wasserstein distance metric to a semipolynomial of degree r (a semipolynomial is a polynomial of the absolute value of some variables).

Lemma 4.3.1. Fix $r \geq 1$. For any element G represented by Eq. (4.4), define

$$D_r(G_0, G) := \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij} \|\Delta \eta_{ij}\|_r^r + \sum_{i=1}^{k_0} |\Delta p_{i\cdot}|.$$

We have that $W_r^r(G, G_0) \asymp D_r(G_0, G)$, as $W_r(G_0, G) \downarrow 0$.

To investigate the behavior of likelihood function $p_G(x)$ as G tends to G_0 in Wasserstein distance W_r , by representation (4.4), we can express

$$p_G(x) - p_{G_0}(x) = \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij} (f(x|\eta_{ij}) - f(x|\eta_i^0)) + \sum_{i=1}^{k_0} \Delta p_{i\cdot} f(x|\eta_i^0). \quad (4.5)$$

By Taylor expansion up to order r , we obtain

$$p_G(x) - p_{G_0}(x) = \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij} \sum_{|\kappa|=1}^r \frac{(\Delta \eta_{ij})^\kappa}{\kappa!} \frac{\partial^{|\kappa|} f}{\partial \eta^\kappa}(x|\eta_i^0) + \sum_{i=1}^{k_0} \Delta p_{i\cdot} f(x|\eta_i^0) + R_r(x), \quad (4.6)$$

where $R_r(x)$ is the Taylor remainder. Moreover, it can be verified that

$$\sup_x |R_r(x)/W_r^r(G, G_0)| \rightarrow 0$$

since f is uniform Lipschitz up to order r . We arrive at the following formulae, which measures the speed of change of the likelihood function as G varies in the Wasserstein neighborhood of G_0 :

$$\frac{p_G(x) - p_{G_0}(x)}{W_r^r(G, G_0)} = \sum_{|\kappa|=1}^r \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} \left(\frac{p_{ij} (\Delta \eta_{ij})^\kappa / \kappa!}{W_r^r(G_0, G)} \right) \frac{\partial^{|\kappa|} f}{\partial \eta^\kappa}(x|\eta_i^0) + \sum_{i=1}^{k_0} \frac{\Delta p_{i\cdot}}{W_r^r(G_0, G)} f(x|\eta_i^0) + o(1). \quad (4.7)$$

The right hand side of Eq. (4.7) is a linear combination of the partial derivatives of f evaluated at G_0 . In addition, by Lemma 4.3.1, the coefficients of this linear representation is asymptotically equivalent to the ratio of two semipolynomials.

Equation (4.7) highlights the distinct roles of model parameters and the kernel

density function in the formation of a mixture model's likelihood. The former appear only in the coefficients, while the latter provides the partial derivatives which appear as if basis functions for the linear combination. We wrote “as if”, because the partial derivatives of kernel f may not be linearly independent functions (recall the examples in Section 4.3.1). When a partial derivative of f can be represented as a linear combination of other partial derivatives, it can be eliminated from the expression in the right hand side. This reduction process may be repeatedly applied until all partial derivatives that remain are linearly independent functions. This motivates the following.

Definition 4.3.1. *The following representation is called r -minimal form of the mixture likelihood for a sequence of mixing measures G tending to G_0 in W_r metric:*

$$\frac{p_G(x) - p_{G_0}(x)}{W_r^r(G, G_0)} = \sum_{l=1}^{T_r} \left(\frac{\xi_l^{(r)}(G)}{W_r^r(G_0, G)} \right) H_l^{(r)}(x) + o(1), \quad (4.8)$$

which holds for all x , with the index l ranging from 1 to a finite T_r , if

- (1) $H_l^{(r)}(x)$ for all l are linearly independent functions of x , and
- (2) coefficients $\xi_l^{(r)}(G)$ are polynomials of the components of $\Delta\eta_{ij}$, and $\Delta p_i, p_{ij}$.

It is sufficient, but not necessary, to select functions $H_l^{(r)}$ from the collection of partial derivatives $\partial^{|\kappa|} f / \partial \eta^\kappa$ evaluated at particular atoms η_i^0 of G_0 , where $|\kappa| \leq r$, by adopting the elimination technique. The precise formulation of $\xi_l^{(r)}(G)$ and $H_l^{(r)}(x)$ will be determined explicitly by the specific G_0 . The r -minimal form for each G_0 is not unique, but they play a fundamental role in our notion of the singularity level of G_0 relative to a class of mixing distributions \mathcal{G} .

Definition 4.3.2. *Fix $r \geq 1$ and let \mathcal{G} be a class of discrete probability measures which has a bounded number of support points in Θ . We say G_0 is r -singular relative*

to \mathcal{G} , if G_0 admits a r -minimal form given by Eq. (4.8), according to which there exists a sequence of $G \in \mathcal{G}$ tending to G_0 under W_r such that

$$\xi_l^{(r)}(G)/W_r^r(G, G_0) \rightarrow 0 \text{ for all } l = 1, \dots, T_r.$$

We now verify that the r -singularity notion is well-defined, in that it does not depend on any specific choice of the r -minimal form. This invariant property is confirmed by part (a) of the following lemma. Part (b) establishes a crucial monotonic property.

Lemma 4.3.2. (a) (*Invariance*) *The existence of the sequence of G in the statement of Definition 4.3.2 holds for all r -minimal forms once it holds for at least one r -minimal form.*

(b) (*Monotonicity*) *If G_0 is r -singular for some $r > 1$, then G_0 is $(r-1)$ -singular.*

Proof. (a) The existence of the sequence of G described in the definition of a r -minimal form implies for that sequence, $(p_G(x) - p_{G_0}(x))/W_r^r(G, G_0) \rightarrow 0$ holds for any x . Now take any r -minimal form (4.8) given by the same sequence. Let $C(G) = \max_{l=1}^{T_r} \frac{\xi_l^{(r)}(G)}{W_r^r(G_0, G)}$. If $\liminf C(G) = 0$, we are done. If not, we have $\liminf C(G) > 0$.

It follows that

$$\sum_{l=1}^{T_r} \left(\frac{\xi_l^{(r)}(G)}{C(G)W_r^r(G, G_0)} \right) H_l^{(r)}(x) \rightarrow 0.$$

Moreover, all the coefficients in the above display are bounded from above by 1, one of which is in fact 1. There exists a subsequence of G by which these coefficients have limits, one of which is 1. This is also a contradiction due to the linear independence of functions $H_l^{(r)}(x)$.

(b) Let G be an element in the sequence that admits a r -minimal form such that $\xi_l^{(r)}(G)/W_r^r(G_0, G) \rightarrow 0$ for all $l = 1, \dots, T_r$. It suffices to assume that the basis functions $H_l^{(r)}$ are selected from the collection of partial derivatives of f . We will show that the same

sequence of G and the elimination procedure for the r -minimal form can be used to construct a $r - 1$ -minimal form by which

$$\xi_l^{(r-1)}(G)/W_{r-1}^{r-1}(G_0, G) \rightarrow 0$$

for all $l = 1, \dots, T_{r-1}$. There are two possibilities to consider.

First, suppose that each of the r -th partial derivatives of density kernel f (i.e., $\partial^\kappa f / \partial \eta^\kappa$, where $|\kappa| = r$) is not in the linear span of the collection of partial derivatives of f at order $r - 1$ or less. Then, for each $l = 1, \dots, T_{r-1}$, $\xi_l^{(r-1)}(G) = \xi_{l'}^{(r)}(G)$ for some $l' \in [1, T_r]$. Since $W_{r-1}^{r-1}(G, G_0) \gtrsim W_r^r(G, G_0)$, due to the fact that the support points of G and G_0 are in a bounded set, we have that

$$\xi_l^{(r-1)}(G)/W_{r-1}^{r-1}(G_0, G) \lesssim \xi_{l'}^{(r)}(G)/W_r^r(G_0, G)$$

which vanishes by the hypothesis.

Second, suppose that some of the r -th partial derivatives, say, $\partial^{|\kappa|} f / \partial \eta^\kappa$ where $|\kappa| = r$, can be eliminated because they can be represented by a linear combination of a subset of other partial derivatives $H_l^{(r-1)}$ (in addition to possibly a subset of other partial derivatives $H_l^{(r)}$) with corresponding finite coefficients $\alpha_{\kappa, i, l}$. It follows that for each $l = 1, \dots, T_{r-1}$, the coefficient $\xi_l^{(r-1)}(G)$ that defines the $r - 1$ -minimal form is transformed into a coefficient in the r -minimal form by

$$\xi_{l'}^{(r)}(G) := \xi_l^{(r-1)}(G) + \sum_{\kappa; |\kappa|=r} \sum_{i=1}^{k_0} \alpha_{\kappa, i, l} \sum_{j=1}^{s_i} p_{ij} (\Delta \eta_{ij})^\kappa / \kappa!.$$

Since $\xi_{l'}^{(r)}(G)/W_r^r(G, G_0)$ tends to 0, so does $\xi_{l'}^{(r)}(G)/W_{r-1}^{r-1}(G, G_0)$. By Lemma 4.3.1 for each κ such that $|\kappa| = r$, we have

$$\sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij} (\Delta \eta_{ij})^\kappa / \kappa! = o(D_{r-1}(G_0, G)) = o(W_{r-1}^{r-1}(G, G_0)).$$

It follows that $\xi_l^{(r-1)}(G)/W_{r-1}^{r-1}(G, G_0)$ tends to 0, for each $l = 1, \dots, T_{r-1}$. This completes the proof. \square

The monotonicity of r -singularity naturally leads to the notion of singularity level of a mixing measure G_0 (and the corresponding parameters) relative to an ambient space \mathcal{G} .

Definition 4.3.3. *The singularity level of G_0 relative to a given class \mathcal{G} , denoted by $\ell(G_0|\mathcal{G})$, is*

0, if G_0 is not r -singular for any $r \geq 1$;

∞ , if G_0 is r -singular for all $r \geq 1$;

otherwise, the largest natural number $r \geq 1$ for which G_0 is r -singular.

The role of the ambient space \mathcal{G} is critical in determining the singularity level of $G_0 \in \mathcal{G}$. Clearly, if $\mathcal{G} \subset \mathcal{G}'$ are both subsets of probability measures that contain G_0 , r -singularity relative to \mathcal{G} entails r -singularity relative to \mathcal{G}' . This means $\ell(G_0|\mathcal{G}) \leq \ell(G_0|\mathcal{G}')$. Let us look at the following examples.

- Take $\mathcal{G} = \mathcal{E}_{k_0}$, i.e., the setting of e-mixtures. It is easy to verify that if the collection of $\{\partial^\kappa f / \partial \eta_j^\kappa(x|\eta_j) | j = 1, \dots, k_0; |\kappa| \leq 1\}$ evaluated at G_0 is linearly independent, then G_0 is not 1-singular relative to \mathcal{E}_{k_0} . It follows that $\ell(G_0|\mathcal{G}) = 0$.
- On the other hand, if $\mathcal{G} = \mathcal{O}_k$ for any $k > k_0$, i.e., the setting of o-mixtures. Then it can be shown that G_0 is always 1-singular relative to \mathcal{O}_k for any $k > k_0$. Thus, $\ell(G_0|\mathcal{G}) \geq 1$. Now, if the collection of $\{\partial^\kappa f / \partial \eta_j^\kappa(x|\eta_j) | j = 1, \dots, k_0; |\kappa| \leq 2\}$ evaluated at G_0 is linearly independent, then it can be observed that G_0 is not 2-singular relative to \mathcal{O}_k . Thus, $\ell(G_0|\mathcal{G}) = 1$.

In fact, the conditions described in the two examples above are referred to as strong identifiability conditions studied by Chen [1995], Nguyen [2013], Ho and Nguyen [2016c]. Our concept of singularity level generalizes such strong identifiability conditions, by allowing us to consider situations where such conditions fail to hold. This is when $\ell(G_0|\mathcal{G}) = 2, 3, \dots, \infty$. The significance of this concept can be appreciated by the following theorem.

Theorem 4.3.1. *Let \mathcal{G} be a class of probability measures on Θ that have a bounded number of support points, and fix $G_0 \in \mathcal{G}$. Suppose that $\ell(G_0|\mathcal{G}) = r$, for some $0 \leq r \leq \infty$.*

(i) *If $r < \infty$, then $\inf_{G \in \mathcal{G}} \frac{\|p_G - p_{G_0}\|_\infty}{W_s^s(G, G_0)} > 0$ for any $s \geq r + 1$.*

(ii) *If $r < \infty$, then $\inf_{G \in \mathcal{G}} \frac{V(p_G, p_{G_0})}{W_s^s(G, G_0)} > 0$ for any $s \geq r + 1$.*

(iii) *If $1 \leq r < \infty$ and in addition,*

(a) *f is $(r + 1)$ -order differentiable with respect to η and for some constant $c_0 > 0$,*

$$\sup_{\|\eta - \eta'\| \leq c_0} \int_{x \in \mathcal{X}} \left(\frac{\partial^{r+1} f}{\partial \eta^\alpha}(x|\eta) \right)^2 / f(x|\eta') dx < \infty \quad (4.9)$$

for any $|\alpha| = r + 1$.

(b) *There is a sequence $G \in \mathcal{G}$ tending to G_0 in Wasserstein metric W_r and the coefficients of the r -minimal form $\xi_l^{(r)}(G) = 0$ for all $l = 1, \dots, T_r$.*

Then, for any $1 \leq s < r + 1$,

$$\liminf_{G \in \mathcal{G}: W_1(G, G_0) \rightarrow 0} \frac{h(p_G, p_{G_0})}{W_1^s(G, G_0)} = 0.$$

(iv) If $r = \infty$ and the conditions (a), (b) in part (iii) hold for any $l \in \mathbb{N}$ (here, the parameter r in these conditions is replaced by l), then the conclusion of part (iii) holds for any $s \geq 1$.

We make a few remarks.

- Part (i) and part (ii) show how the singularity level of G_0 relative to an ambient space \mathcal{G} may be used to translate the convergence of mixture densities (under the sup-norm and the total variation distance) into the convergence of mixing measures under a Wasserstein metric. Part (iii) shows a sufficient condition under which the power $r + 1$ in the bounds from part (i) and (ii) is in fact tight.
- In part (iii) the condition regarding the integrand of the partial derivative of f (cf. Eq. (4.9)) can be easily checked to be satisfied by many kernels, such as Gaussian kernel, Gamma kernel, Student t's kernel, and skewnormal kernel.
- Condition (b) regarding the sequence of G appears somewhat opaque in general, but it will be seen in specific examples for skewnormal mixtures in the sequel. It is sufficient, but not necessary, for verifying the r -singularity of G_0 to construct the sequence of G so that $\xi_l^{(r)}(G) = 0$ for all $1 \leq l \leq T_r$, provided such a sequence exists. This requires finding an appropriate parameterization of a sequence of G tending toward G_0 that satisfy a number of polynomial equations defined in terms of the parameter perturbations.

Proof. Here, we provide the proof for part (i) and (ii) of the theorem. The proof for part (iii) and (iv) is deferred to the Appendix.

(i) It suffices to prove the first inequality for $s = r+1$. Firstly, we will demonstrate that

$$\liminf_{G \in \mathcal{G}: W_s(G, G_0) \rightarrow 0} \|p_G - p_{G_0}\|_\infty / W_s^s(G, G_0) > 0.$$

If this is not true, then there exists a sequence of G such that $W_s(G, G_0) \rightarrow 0$, and for any x , $(p_G(x) - p_{G_0}(x))/W_s^s(G, G_0) \rightarrow 0$. Take any s -minimal form for this ratio, we have

$$\frac{p_G(x) - p_{G_0}(x)}{W_s^s(G, G_0)} = \sum_{l=1}^{T_s} \left(\frac{\xi_l^{(s)}(G)}{W_s^s(G, G_0)} \right) H_l^{(s)}(x) + o(1) \rightarrow 0.$$

For each G in the sequence, let $C(G) = \max_l \frac{\xi_l^{(s)}(G)}{W_s^s(G_0, G)}$. If $\liminf C(G) = 0$, then this means G_0 is s -singular, so $\ell(G_0|\mathcal{G}) \geq s$. This violates the given assumption. So we have $\liminf C(G) > 0$. It follows that

$$\sum_{l=1}^{T_s} \left(\frac{\xi_l^{(s)}(G)}{C(G)W_s^s(G, G_0)} \right) H_l^{(s)}(x) \rightarrow 0.$$

Moreover, all coefficients in the above display are bounded from above by 1, one of which is in fact 1. There exists a subsequence of G by which these coefficients have a limit, one of which is 1. This is also a contradiction due to the linear independence of functions $H_l^{(s)}$.

Therefore, we can find a positive number ϵ_0 such that $\|p_G - p_{G_0}\|_\infty \gtrsim W_s^s(G, G_0)$ for any $W_s(G, G_0) \leq \epsilon_0$. Now, to obtain the conclusion of part (i), it suffices to demonstrate that

$$\inf_{G \in \mathcal{G}: W_s(G, G_0) > \epsilon_0} \|p_G - p_{G_0}\|_\infty / W_s^s(G, G_0) > 0.$$

If this is not the case, there is a sequence G' such that $W_s(G', G_0) > \epsilon_0$ and $\|p_{G'} - p_{G_0}\|_\infty / W_s^s(G', G_0) \rightarrow 0$. Since Θ is compact and \mathcal{G} contains only probability measures with bounded number of support points in Θ , we can find $G^* \in \mathcal{G}$ such that $W_s(G', G^*) \rightarrow 0$ and $W_s(G^*, G_0) \geq \epsilon_0$. As $W_s(G', G_0) \rightarrow W_s(G^*, G_0) > 0$, we have $\|p_{G'} - p_{G_0}\|_\infty \rightarrow 0$. Now, due to the first order uniform Lipschitz condition of f , we obtain $p_{G'}(x) \rightarrow p_{G^*}(x)$ for all $x \in \mathcal{X}$. Thus, $p_{G^*}(x) = p_{G_0}(x)$ for almost all $x \in \mathcal{X}$, which entails that $G^* = G_0$,

a contradiction. This completes the proof.

(ii) Turning to the second inequality, we also firstly demonstrate that

$$\liminf_{G \in \mathcal{G}: W_s(G, G_0) \rightarrow 0} V(p_G, p_{G_0}) / W_s^s(G, G_0) > 0.$$

If it is not true, then we have a sequence of G such that $W_s(G, G_0) \rightarrow 0$ and $V(p_G, p_{G_0}) / W_s^s(G, G_0) \rightarrow 0$. By Fatou's lemma

$$0 = \liminf \frac{V(p_G, p_{G_0})}{C(G)W_s^s(G, G_0)} \geq \int \liminf_G \left| \frac{\xi_l^{(s)}(G)}{C(G)W_s^s(G, G_0)} H_l^{(s)}(x) \right| dx.$$

The integrand must be zero for almost all x , leading to a contradiction as before. Hence, to obtain the conclusion of part (ii), we only need to show that

$$\inf_{G \in \mathcal{G}: W_s(G, G_0) > \epsilon_0} V(p_G, p_{G_0}) / W_s^s(G, G_0) > 0.$$

where $\epsilon_0 > 0$ such that $V(p_G, p_{G_0}) \gtrsim W_s^s(G, G_0)$ for any $W_s(G, G_0) \leq \epsilon_0$. If it is not true, by using the same argument as that of part (i), there is a sequence of G' such that $W_s(G', G^*) \rightarrow 0$, $V(p_{G'}, p_{G_0}) \rightarrow 0$, while $W_s(G^*, G_0) \geq \epsilon_0$ and $p_{G'}(x) \rightarrow p_{G^*}(x)$ for all $x \in \mathcal{X}$. By Fatou's lemma,

$$0 = \liminf V(p_{G'}, p_{G_0}) \geq \int \liminf |p_{G'}(x) - p_{G_0}(x)| dx = V(p_{G^*}, p_{G_0}),$$

which leads to $G^* = G_0$, a contradiction. We obtain the conclusion of this part. \square

We are ready to state the impact of the singularity level of mixing measure G_0 relative to an ambient space \mathcal{G} on the rate of convergence for an estimate of G_0 , where $\mathcal{G} = \mathcal{E}_{k_0}$ in e-mixtures, and $\mathcal{G} = \mathcal{O}_k$ in o-mixtures. Let \mathcal{G} be structured into a sieve of subsets defined by the maximum singularity level relative to \mathcal{G} .

$$\mathcal{G} = \bigcup_{r=1}^{\infty} \mathcal{G}_r, \text{ where } \mathcal{G}_r := \left\{ G \in \mathcal{G} \mid \ell(G|\mathcal{G}) \leq r \right\}, \quad r = 0, 1, \dots, \infty.$$

The first part of the following theorem gives a minimax lower bound for the estimation of the mixing measure G_0 , given that the singularity level of G_0 is known up to a constant $r \geq 1$. The second part gives a quick result on the convergence rate of a point estimate such as the MLE.

Theorem 4.3.2. (a) Fix $r \geq 1$. Assume that for any $G_0 \in \mathcal{G}_r$, the conclusion of part (iii) of Theorem 4.3.1 holds for \mathcal{G}_r (i.e., \mathcal{G} is replaced by \mathcal{G}_r in that theorem). Then, for any $s \in [1, r+1)$ there holds

$$\inf_{\widehat{G}_n \in \mathcal{G}_r} \sup_{G_0 \in \mathcal{G}_r} E_{p_{G_0}} W_s(\widehat{G}_n, G_0) \gtrsim n^{-1/2s}.$$

Here, the infimum is taken over all sequences of estimates $\widehat{G}_n \in \mathcal{G}_r$ and $E_{p_{G_0}}$ denotes the expectation taken with respect to product measure with mixture density $p_{G_0}^n$.

(b) Let $G_0 \in \mathcal{G}_r$ for some fixed $r \geq 1$. Let $\widehat{G}_n \in \mathcal{G}_r$ be a point estimate for G_0 , which is obtained from an n -sample of i.i.d. observations drawn from p_{G_0} . As long as $h(p_{\widehat{G}_n}, p_{G_0}) = O_P(n^{-1/2})$, we have

$$W_{r+1}(\widehat{G}_n, G_0) = O_P(n^{-1/2(r+1)}).$$

Proof. Part (a) of this theorem is a consequence of the conclusion of Theorem 4.3.1, part (iii). The proof of this fact is quite standard, and similar to that of Theorem 1.1. of [Ho and Nguyen, 2016a] and is omitted. Part (b) follows immediately from part (ii) of Theorem 4.3.1, as we have $h(p_{\widehat{G}_n}, p_{G_0}) \geq V(p_{\widehat{G}_n}, p_{G_0}) \gtrsim W_{r+1}^{r+1}(\widehat{G}_n, G_0)$. \square

We conclude this section with some comments. It is well-known that many density estimation methods, such as MLE and Bayesian estimation applied to a compact parameter space for parametric mixture models, guarantee a root- n rate (up to a logarithmic term) of convergence under Hellinger distance metric on the density functions (cf. [van de Geer, 2000, Ghosal and van der Vaart, 2001, DasGupta, 2008]). It follows that, as far as we are concerned, the remaining work in establishing the convergence behavior of parameter estimation (as opposed to density estimation) lies in the calculation of the singularity levels, i.e., the identification of sets \mathcal{G}_r . For skewnormal mixtures, such calculations will be carried out in Section 4.4 and Section 4.5. For the settings of G_0 where we are able to obtain the exact singularity levels, we can also construct the sequence of G required by part (iii) of Theorem 4.3.1. Whenever the exact singularity level is obtained, we automatically obtain a minimax lower bound and a matching upper bound for MLE convergence rate under a Wasserstein distance metric, thanks to the above theorem. In some cases, however, the singularity level of G_0 may be not determined exactly, but only an upper bound is given. In such cases, only an upper bound to the convergence rate of the MLE can be obtained, while minimax lower bounds may be unknown.

4.3.3 Construction of r -minimal forms

As we mentioned above, a simple way of constructing an r -minimal form is to select a subset of partial derivatives of f taken up to order r such that all these functions are linearly independent. A simple procedure is to start from the smallest order $r = 1$ and then move up to $r = 2, 3, \dots$ and so on. For each r , assume that we have obtained a linearly independent subset of partial derivatives up to order $r - 1$. Now, going over the ordered list of r -th partial derivatives: $\{\partial^{|\kappa|}f/\partial\eta^\kappa | \kappa \in \mathbb{N}^d, |\kappa| = r\}$. For each κ such that $|\kappa| = r$, if the partial derivative of f of order κ can be expressed as a linear combination of other partial derivatives already selected, then this one is eliminated.

The process goes on until we exhaust the list of the partial derivatives.

Example 4.3.2. Continuing from Example 4.3.1, suppose that G_0 satisfies Eq. (4.1).

According to the proof of Lemma 4.4.1, we can choose $\alpha_{4k} \neq 0$, so the partial derivative may be eliminated via the reduction:

$$\frac{\partial f(x|\eta_k^0)}{\partial m} = -\sum_{j=1}^k \frac{\alpha_{1j}}{\alpha_{4k}} f(x|\eta_j^0) + \frac{\alpha_{2j}}{\alpha_{4k}} \frac{\partial f(x|\eta_j^0)}{\partial \theta} + \frac{\alpha_{3j}}{\alpha_{4k}} \frac{\partial f(x|\eta_j^0)}{\partial v} - \sum_{j=1}^{k-1} \frac{\alpha_{4j}}{\alpha_{4k}} \frac{\partial f(x|\eta_j^0)}{\partial m}$$

Note that this elimination step is valid only for a subset of $G_0 = G(\mathbf{p}^0, \boldsymbol{\eta}^0)$ for which Eq. (4.1) holds. That is, only if $P_1(\boldsymbol{\eta}^0) = 0$ or $P_2(\boldsymbol{\eta}^0) = 0$.

Example 4.3.3. If $f(x|\eta) = f(x|\theta, v, m)$ where $m = 0$, the skewnormal kernel becomes the Gaussian kernel. Due to (4.3), all partial derivatives with respect to both θ and v can be eliminated via the following reduction: for any $\kappa_1, \kappa_2 \in \mathbb{N}$, for any $j = 1, \dots, k_0$,

$$\frac{\partial^{\kappa_1+\kappa_2} f(x|\eta_j^0)}{\partial \theta^{\kappa_1} v^{\kappa_2}} = \frac{1}{2^{\kappa_2}} \frac{\partial^{\kappa_1+2\kappa_2} f(x|\eta_j^0)}{\partial \theta^{\kappa_1+2\kappa_2}}.$$

Thus, this elimination is valid for all parameter values $(\mathbf{p}^0, \boldsymbol{\eta}^0)$, and r -minimal forms for all orders.

Example 4.3.4. For the skewnormal kernel density $f(x|\eta) = f(x|\theta, v, m)$, Eq. (4.2) yields the following reductions: for any $j = 1, \dots, k_0$, any $\eta = (\theta, v, m) = \eta_j^0 = (\theta_j^0, v_j^0, m_j^0)$ such that $m \neq 0$

$$\frac{\partial^2 f}{\partial \theta^2} = 2 \frac{\partial f}{\partial v} - \frac{m^3 + m}{v} \frac{\partial f}{\partial m}, \quad (4.10)$$

$$\frac{\partial^2 f}{\partial v \partial m} = -\frac{1}{v} \frac{\partial f}{\partial m} - \frac{m^2 + 1}{2vm} \frac{\partial^2 f}{\partial m^2}. \quad (4.11)$$

This results in a ripple effect on subsequent eliminations at higher orders. For examples, partial derivatives up to the third order of f evaluated at $\eta = \eta_j^0 = (\theta_j^0, v_j^0, m_j^0)$

for any $j = 1, \dots, k_0$ where $m_j^0 \neq 0$ can be expressed as follows:

$$\begin{aligned}
\frac{\partial^3 f}{\partial \theta^3} &= 2 \frac{\partial^2 f}{\partial \theta \partial v} - \frac{m^3 + m}{v} \frac{\partial^2 f}{\partial \theta \partial m}, \\
\frac{\partial^3 f}{\partial \theta^2 \partial v} &= 2 \frac{\partial^2 f}{\partial v^2} + \frac{m^3 + m}{v^2} \frac{\partial f}{\partial m} - \frac{m^3 + m}{v} \frac{\partial^2 f}{\partial v \partial m}, \\
\frac{\partial^3 f}{\partial \theta^2 \partial m} &= 2 \frac{\partial^2 f}{\partial v \partial m} - \frac{3m^2 + 1}{v} \frac{\partial f}{\partial m} - \frac{m^3 + m}{v} \frac{\partial^2 f}{\partial m^2}, \\
\frac{\partial^3 f}{\partial v \partial m^2} &= -\frac{m^2 + 1}{2vm} \frac{\partial^3 f}{\partial m^3} - \frac{3m^2 - 1}{2vm^2} \frac{\partial^2 f}{\partial m^2}, \\
\frac{\partial^3 f}{\partial v^2 \partial m} &= -\frac{2}{v} \frac{\partial^2 f}{\partial v \partial m} - \frac{m^2 + 1}{2vm} \frac{\partial^3 f}{\partial v \partial m^2} \\
&= \frac{(m^2 + 1)^2}{4v^2 m^2} \frac{\partial^3 f}{\partial m^3} + \frac{(m^2 + 1)(7m^2 - 1)}{4m^3 v^2} \frac{\partial^2 f}{\partial m^2} + \frac{2}{v^2} \frac{\partial f}{\partial m}, \\
\frac{\partial^3 f}{\partial \theta \partial v \partial m} &= -\frac{m^2 + 1}{2vm} \frac{\partial^3 f}{\partial \theta \partial m^2} - \frac{1}{v} \frac{\partial^2 f}{\partial \theta \partial m}.
\end{aligned} \tag{4.12}$$

All three examples above demonstrate how the dependence among partial derivatives of kernel density f , among different orders κ , and among those evaluated at different component i , has a deep impact on the representation of r -minimal forms.

In general, the r -minimal form (4.8) may be expressed somewhat more explicitly as follows

$$\frac{p_G(x) - p_{G_0}(x)}{W_r^r(G, G_0)} = \sum_{(i, \kappa) \in \mathcal{I}, \mathcal{K}} \frac{\xi_{i, \kappa}^{(r)}(G)}{W_r^r(G_0, G)} H_{i, \kappa}^{(r)}(x|G_0) + \sum_{i=1}^{k_0} \frac{\zeta_i^{(r)}(G)}{W_r^r(G_0, G)} f(x|\eta_i^0) + o(1).$$

where $\mathcal{I} \subset \{1, \dots, k_0\}$ and $\mathcal{K} \subset \mathbb{N}^d$ of elements κ such that $|\kappa| \leq r$. It is emphasized that the sets \mathcal{I} and \mathcal{K} are specific to a particular r -minimal form under consideration. $H_{i, \kappa}^{(r)}$ are a collection of linearly independent partial derivatives of f that are also independent of all functions $f(x|\eta_i^0)$. $H_{i, \kappa}^{(r)}$ are taken from the collection of partial derivatives with order at most r . We also observe that $\xi_{i, \kappa}^{(r)}$ and $\zeta_i^{(r)}$ take the following

polynomial forms:

$$\xi_{i,\kappa}^{(r)}(G) = \sum_{j=1}^{s_i} \frac{p_{ij}(\Delta\eta_{ij})^\kappa}{\kappa!} + \sum_{i',\kappa'} \beta_{i,\kappa,i',\kappa'}(G_0) \sum_{j=1}^{s_{i'}} \frac{p_{ij}(\Delta\eta_{ij})^{\kappa'}}{\kappa'!}, \quad (4.13)$$

$$\zeta_i^{(r)}(G) = \Delta p_i + \sum_{i',\kappa'} \gamma_{i,\kappa,i',\kappa'}(G_0) \sum_{j=1}^{s_{i'}} \frac{p_{ij}(\Delta\eta_{ij})^{\kappa'}}{\kappa'!}. \quad (4.14)$$

In the right hand side of each of the last two equations, i' is taken from a subset of $\{1, \dots, k_0\}$ and κ' is from a subset of \mathbb{N}^d such that $|\kappa| \leq |\kappa'| \leq r$. The actual detail of these subsets depend on the specific elimination scheme leading to the r -minimal form. Likewise, the non-zero coefficients $\beta_{i,\kappa,i',\kappa'}(G_0)$ and $\gamma_{i,\kappa,i',\kappa'}(G_0)$ arise from the specific elimination scheme. We include argument G_0 in these coefficients to highlight the fact that they may be dependent on the atoms of G_0 (cf. Example 4.3.2 and 4.3.4).

By the definition of r -singularity for any $r \geq 1$, G_0 is r -singular relative to \mathcal{G} if there exists a sequence of G tending to G_0 in the ambient space \mathcal{G} such that the sequences of semipolynomial fractions, namely, $\xi_{i,\kappa}^{(r)}(G)/W_r^r(G, G_0)$ and $\zeta_i^{(r)}(G)/W_r^r(G, G_0)$ (whose numerators are given by Eq. (4.13) and Eq. (4.14)), must vanish. As a consequence, the question of r -singularity for a given element G_0 is determined by the limiting behavior of a finite collection of infinite sequences of semipolynomial fractions.

4.3.4 Polynomial limits of r -minimal form coefficients

It is worth noting that the limiting behavior of semipolynomial fractions described above is independent of a particular choice of the r -minimal form, in a sense that we now explain. In part (a) of Lemma 4.3.2, we established an invariance property of the r -singularity, which does not depend on a specific form of the r -minimal form. Let us restrict the basis functions to be members of the collection of all partial derivatives of f up to order r . In the proof of part (b) of Lemma 4.3.2 it was shown that the coefficients $\xi_l^{(r)}(G)$ have to be elements of a set of polynomials of $\Delta\eta_{ij}$, Δp_i , and p_{ij} ,

which are closed under linear combinations of its elements. Let us denote this set by $\mathcal{P}(G, G_0)$, which is invariant with respect to any specific choice of the basis functions (from the collection of partial derivatives) for the r -minimal form. Moreover, G_0 is r -singular if and only if a sequence of G tending to G_0 in W_r can be constructed such that for any element $\xi_l^{(r)}(G) \in \mathcal{P}(G, G_0)$, we have $\xi_l^{(r)}(G)/W_r^r(G, G_0) \rightarrow 0$. Equivalently,

$$\xi_l^{(r)}(G)/D_r(G, G_0) \rightarrow 0 \text{ for all } \xi_l^r(G) \in \mathcal{P}(G, G_0). \quad (4.15)$$

Extracting the limits of a single multivariate semipolynomial fraction (a.k.a. rational semipolynomial functions) is quite challenging in general, due to the interaction among multiple variables involved [Xiao and Zeng, 2014]. Analyzing the limits of not one but a collection of multivariate rational semipolynomials is considerably more difficult. To obtain meaningful and concrete results, we need to consider specific systems of multivariate rational semipolynomials that arise from the r -minimal form.

In the remainder of this chapter we will proceed to do just that. By working with a specific choice of kernel density f (the skewnormal), it will be shown that under the compactness of the parameter spaces, one can extract a subset of limits from the system of rational semipolynomials $\xi_l^{(r)}(G)/D_r(G, G_0)$. These limits are expressed as a system of polynomials admitting non-trivial solutions. For a given $r \geq 1$, if the extracted system of polynomial limits does not contain admissible solutions, then it means that there does not exist any sequence of mixing measures G for which a valid r -minimal form can be constructed, because (4.15) is not fulfilled. This would entail the upper bound $\ell(G_0|\mathcal{G}) < r$. On the other hand, if the extracted system of polynomial limits does contain at least one admissible solution, this is a hint that the r -singularity level of G_0 relative to the ambient space G *might* hold. Whether this is actually the case or not requires an explicit construction of a sequence of $G \in \mathcal{G}$ (often building upon the admissible solutions of the polynomial limits) and then the verification that condition (4.15) indeed holds. For the verification purpose,

it is sufficient (and simpler) to work with a specific choice of r -minimal form, as Definition 4.3.2 allows.

The foregoing description, along with the presentation in the previous subsection on the construction of r -minimal forms, provides the outline of a general procedure which links the determination of the singularity level to the solvability of a system of polynomial limits. This procedure will be illustrated in the next sections for the remarkable world of mixtures of skewnormal distributions.

4.4 O-mixtures of skewnormal distributions

In this section, we focus on the o-mixture setting of skewnormal distributions. To avoid a heavy dose of technicality, we study the singularity level of $G_0 \in \mathcal{E}_{k_0}$ relative to ambient space \mathcal{O}_{k,c_0} for some $k > k_0$ and small constant $c_0 > 0$, where $\mathcal{O}_{k,c_0} \subset \mathcal{O}_k$ contains only (discrete) probability measures whose point masses are bounded from below by c_0 . Moreover, we will analyze the singularity level of $G_0 \in \mathcal{S}_0$, a subset to be defined shortly by (4.16). This case is interesting in that it illustrates the full power of the general method of analysis that was described in Section 4.3 in a concrete fashion. Due to the complex nature and space constraints, we will not report any result on the case where G_0 is in the complement of \mathcal{S}_0 .² Instead, in Section 4.5 we study the singularity level of G_0 relative to the smaller ambient space \mathcal{E}_{k_0} (that is, e-mixture setting), for which a more complete picture of the singularity structure is achieved.

Lemma 4.4.1. *For skewnormal density kernel $f(x|\boldsymbol{\eta})$, the collection of $\{\partial^\kappa f / \partial \eta^\kappa(x|\eta_j) | j = 1, \dots, k_0; 0 \leq |\kappa| \leq 1\}$ is not linearly independent if and only if $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)$ are the zeros of either polynomial P_1 or P_2 , which are defined as follows:*

$$\text{Type A: } P_1(\boldsymbol{\eta}) = \prod_{j=1}^{k_0} m_j.$$

²Interested readers may consult Section 6 in technical report [Ho and Nguyen, 2016d] for such results.

$$Type\ B: P_2(\boldsymbol{\eta}) = \prod_{1 \leq i \neq j \leq k_0} \left\{ (\theta_i - \theta_j)^2 + \left[\sigma_i^2(1 + m_j^2) - \sigma_j^2(1 + m_i^2) \right]^2 \right\}.$$

This lemma leads us to consider

$$\mathcal{S}_0 = \left\{ G = G(\mathbf{p}, \boldsymbol{\eta}) \mid (\mathbf{p}, \boldsymbol{\eta}) \in \Omega, P_1(\boldsymbol{\eta}) \neq 0, P_2(\boldsymbol{\eta}) \neq 0 \right\}. \quad (4.16)$$

In the o-mixture setting, we will see that $\ell(G_0 | \mathcal{O}_{k,c_0})$ may grow with $k - k_0$, the number of extra mixing components. The main exercise is to arrive at a suitable r -minimal form, for which the vanishing behavior of its coefficients can be analyzed. Section 4.3.3 describes a general strategy for the construction of r -minimal form based on the partial derivatives of the density kernel f with respect to the parameters $\eta = (\theta, v, m)$ up to order r .

For skewnormal kernel density f , the following lemma provides an explicit form for reducing a partial derivative of f to other partial derivatives of lower orders.

Lemma 4.4.2. *For any $r \geq 1$, denote*

$$A_1^r = \{(\alpha_1, \alpha_2, \alpha_3) : \alpha_1 \leq 1, \alpha_3 = 0, \text{ and } |\alpha| \leq r\}.$$

$$A_2^r = \{(\alpha_1, \alpha_2, \alpha_3) : \alpha_1 \leq 1, \alpha_2 = 0, \alpha_3 \geq 1, \text{ and } |\alpha| \leq r\}.$$

$$\mathcal{F}_r = A_1^r \cup A_2^r.$$

Let $f(x|\eta) = f(x|\theta, v, m)$ denote the skewnormal kernel. Then, for any $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{N}^3$ and $m \neq 0$, there holds

$$\frac{\partial^{|\alpha|} f}{\partial \theta^{\alpha_1} \partial v^{\alpha_2} \partial m^{\alpha_3}} = \sum_{\kappa \in \mathcal{F}_{|\alpha|}} \frac{P_{\alpha_1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2} \partial m^{\kappa_3}},$$

where, $P_{\alpha_1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)$, $H_{\alpha_1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)$, and $Q_{\alpha_1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)$ are polynomials in terms of m, v respectively.

Next, we show that the partial derivatives on the RHS of the above identity are

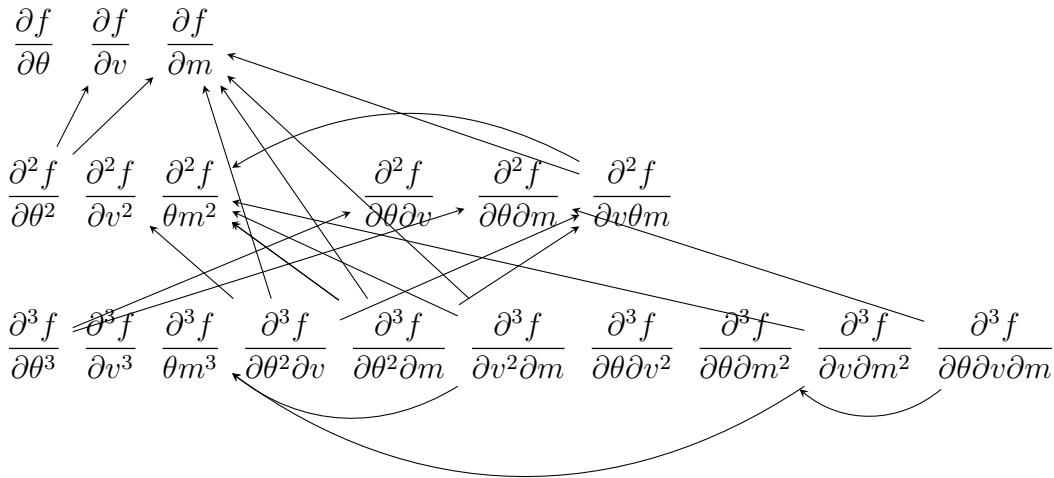


Figure 4.1: The illustration of the elimination steps from a complete collection of derivatives of f up to the order 3 to a reduced system of linearly independent partial derivatives, cf. Lemma 4.4.3. The circled derivatives are eliminated from the partial derivatives present in the 3-minimal form. $A \rightarrow B$ means that B is involved in the representation of A under the reduction.

in fact linearly independent, under additional assumptions on G_0 .

Lemma 4.4.3. *Recall the notation from Lemma 4.4.2. If $G_0 \in \mathcal{S}_0$, then for any $r \geq 1$, the collection of partial derivatives of the skewnormal density kernel $f(x|\eta)$, namely*

$$\left\{ \frac{\partial^{|\kappa|} f(x|\eta)}{\theta^{\kappa_1} v^{\kappa_2} m^{\kappa_3}} \middle| \kappa = (\kappa_1, \kappa_2, \kappa_3) \in \mathcal{F}_r, \eta = \eta_1^0, \dots, \eta_{k_0}^0 \right\}$$

is linearly independent.

Figure 4.1 gives an illustration of Lemma 4.4.3 when $r = 3$. Armed with the foregoing lemmas we can easily obtain a suitable r -minimal form for the mixture densities of skewnormals.

4.4.1 Special cases

To illustrate our techniques and results, consider a special case in which G_0 has exactly one atom, and $k = k_0 + 1 = 2$. The general results will be presented in Section 4.4.2.

G_0 is 1-singular $G_0 \in \mathcal{S}_0$ implies that all first order derivatives of f are linearly independent. Hence, from Eq. (4.8), the 1-minimal form takes the form:

$$\begin{aligned} \frac{p_G(x) - p_{G_0}(x)}{W_1(G, G_0)} &\asymp \frac{1}{W_1(G, G_0)} \left(\Delta p_{1.} f(x|\eta_1^0) + \sum_{i=1}^2 p_{1i} \Delta \theta_{1i} \frac{\partial f}{\partial \theta}(x|\eta_1^0) \right. \\ &\quad \left. + \sum_{i=1}^2 p_{1i} \Delta v_{1i} \frac{\partial f}{\partial v}(x|\eta_1^0) + \sum_{i=1}^2 p_{1i} \Delta m_{1i} \frac{\partial f}{\partial m}(x|\eta_1^0) \right) + o(1). \end{aligned} \quad (4.17)$$

Since $k = 2$ and $k_0 = 1$, we have $\Delta p_{1.} = 0$. A sequence of G can be chosen so that $\sum_{i=1}^2 p_{1i} \Delta \theta_{1i} = 0$, $\sum_{i=1}^2 p_{1i} \Delta v_{1i} = 0$, $\sum_{i=1}^2 p_{1i} \Delta m_{1i} = 0$. Clearly, all of the coefficients in (4.8) are 0. Hence G_0 is 1-singular relative to \mathcal{O}_{2,c_0} .

G_0 is 2-singular Using the method of elimination described in Example 4.3.4 we obtain the following 2-minimal form:

$$\frac{1}{W_2^2(G, G_0)} \left(\sum_{\kappa \in \mathcal{F}_2} \xi_{\kappa_1, \kappa_2, \kappa_3}^{(2)} \frac{\partial^{|\alpha|} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2} \partial m^{\kappa_3}}(x|\eta_1^0) \right) + o(1), \quad (4.18)$$

where $\xi_{\kappa_1, \kappa_2, \kappa_3}^{(2)}$ are given by

$$\begin{aligned} \xi_{1,0,0}^{(2)} &= \sum_{i=1}^2 p_{1i} \Delta \theta_{1i}, \quad \xi_{0,1,0}^{(2)} = \sum_{i=1}^2 p_{1i} \Delta v_{1i} + \sum_{i=1}^2 p_{1i} (\Delta \theta_{1i})^2, \\ \xi_{0,0,1}^{(2)} &= -\frac{(m_1^0)^3 + m_1^0}{2v_1^0} \sum_{i=1}^2 p_{1i} (\Delta \theta_{1i})^2 - \frac{1}{v_1^0} \sum_{i=1}^2 p_{1i} \Delta v_{1i} \Delta m_{1i} + \sum_{i=1}^2 p_{1i} \Delta m_{1i}, \\ \xi_{0,2,0}^{(2)} &= \sum_{i=1}^2 p_{1i} (\Delta v_{1i})^2, \quad \xi_{0,0,2}^{(2)} = -\frac{(m_1^0)^2 + 1}{2v_1^0 m_1^0} \sum_{i=1}^2 p_{1i} \Delta v_{1i} \Delta m_{1i} + \sum_{i=1}^2 p_{1i} \Delta (m_{1i})^2, \\ \xi_{1,1,0}^{(2)} &= \sum_{i=1}^2 p_{1i} \Delta \theta_{1i} \Delta v_{1i}, \quad \xi_{1,0,1}^{(2)} = \sum_{i=1}^2 p_{1i} \Delta \theta_{1i} \Delta m_{1i}. \end{aligned}$$

Note in particular the formulas for $\xi_{0,1,0}^{(2)}$, $\xi_{0,0,1}^{(2)}$ and $\xi_{0,0,2}^{(2)}$ are the results of reduction equation (4.10). It remains to construct a sequence of G tending to G_0 so that

$\xi_\kappa^{(2)}/W_2^2(G, G_0)$ vanish for all $\kappa = (\kappa_1, \kappa_2, \kappa_3) \in \mathcal{F}_2$. Define

$$\overline{M} = \max \{|\Delta\theta_{11}|, |\Delta\theta_{12}|, |\Delta v_{11}|^{1/2}, |\Delta v_{12}|^{1/2}, |\Delta m_{11}|^{1/2}, |\Delta m_{12}|^{1/2}\}.$$

Then, it can be observed that $W_2^2(G, G_0) \gtrsim \overline{M}^2$ and $\xi_{\kappa_1, \kappa_2, \kappa_3}^{(2)} = O(\overline{M}^{\kappa_1+2\kappa_2+2\kappa_3})$. So, for any $\kappa \in \mathcal{F}_2$ such that $\kappa_1 + 2\kappa_2 + 2\kappa_3 \geq 3$, as $\xi_{\kappa_1, \kappa_2, \kappa_3}^{(2)} = O(\overline{M}^s)$ where $s \geq 3$, it implies that $\xi_{\kappa_1, \kappa_2, \kappa_3}^{(2)}/W_2^2(G, G_0) \rightarrow 0$. So we only need to consider the coefficients where $\kappa_1 + 2\kappa_2 + 2\kappa_3 \leq 2$ and $\kappa_1 \leq 1$. They are $\xi_{1,0,0}^{(2)}/W_2^2(G, G_0)$, $\xi_{0,1,0}^{(2)}/W_2^2(G, G_0)$, and $\xi_{0,0,1}^{(2)}/W_2^2(G, G_0)$. Now, by dividing both the numerator and denominator of each of these coefficients by \overline{M} , \overline{M}^2 , and \overline{M}^2 , respectively, we extract the following system of polynomial limits:

$$\begin{aligned} d_1^2 a_1 + d_2^2 a_2 &= 0, \\ d_1^2 a_1^2 + d_2^2 a_2^2 + d_1^2 b_1 + d_2^2 b_2 &= 0, \\ -\frac{(m_1^0)^3 + m_1^0}{2v_1^0} (d_1^2 a_1^2 + d_2^2 a_2^2) + d_1^2 c_1 + d_2^2 c_2 &= 0, \end{aligned} \quad (4.19)$$

where $\Delta\theta_{1i}/\overline{M} \rightarrow a_i$, $\Delta v_{1i}/\overline{M}^2 \rightarrow b_i$, $\Delta m_{1i}/\overline{M}^2 \rightarrow c_i$, $p_{1i} \rightarrow d_i^2$ for all $1 \leq i \leq 2$.

One solution to the above system of polynomial equations is $d_1 = d_2$, $a_1 = -a_2$, $b_1 = b_2 = a_1^2/2$, $c_1 = c_2 = (-(m_1^0)^3 + m_1^0)/2v_1^0$. It follows that if we choose the sequence of G so that $p_{11} = p_{12} = 1/2$, $\Delta\theta_{11} = -\Delta\theta_{12}$, $\Delta v_{11} = \Delta v_{12} = (\Delta\theta_{11})^2/2$, and $\Delta m_{11} = \Delta m_{12} = (\Delta\theta_{11})^2(-(m_1^0)^3 + m_1^0)/2v_1^0$, then all coefficients of the 2-minimal form vanish. Hence, G_0 is 2-singular relative to \mathcal{O}_{2,c_0} .

G_0 is 3-singular The proof for this is similar. A 3-minimal form can be obtained by applying the reductions (4.12), which eliminate all third order partial derivatives in terms of lower order ones that are in fact linearly independent by the condition that $G_0 \in \mathcal{S}_0$. As in the foregoing paragraphs, we can obtain a system of polynomials that turn out to share the same solution as the one described. This leads to the same

choice of sequence for G according to which all coefficients of the 3-minimal form vanish. Thus, G_0 is 3-singular relatively to \mathcal{O}_{k,c_0} .

G_0 is not 4-singular Applying the same approach to obtain a 4-minimal form and their rational semipolynomial coefficients, from which we extract the following system of real polynomial limits:

$$\begin{aligned}
d_1^2 a_1 + d_2^2 a_2 &= 0, \\
d_1^2 a_1^2 + d_2^2 a_2^2 + d_1^2 b_1 + d_2^2 b_2 &= 0, \\
-\frac{(m_1^0)^3 + m_1^0}{2v_1^0} (d_1^2 a_1^2 + d_2^2 a_2^2) + d_1^2 c_1 + d_2^2 c_2 &= 0, \\
\frac{1}{3} (d_1^2 a_1^3 + d_2^2 a_2^3) + d_1^2 a_1 b_1 + d_2^2 a_2 b_2 &= 0, \\
-\frac{(m_1^0)^3 + m_1^0}{6v_1^0} (d_1^2 a_1^3 + d_2^2 a_2^3) + d_1^2 a_1 c_1 + d_2^2 a_2 c_2 &= 0, \\
\frac{1}{6} (d_1^2 a_1^4 + d_2^2 a_2^4) + d_1^2 a_1^2 b_1 + d_2^2 a_2^2 b_2 + \frac{1}{2} (d_1^2 b_1^2 + d_2^2 b_2^2) &= 0, \\
\frac{((m_1^0)^3 + m_1^0)^2}{12(v_1^0)^2} (d_1^2 a_1^4 + d_2^2 a_2^4) - \frac{(m_1^0)^3 + m_1^0}{v_1^0} (d_1^2 a_1^2 c_1 + d_2^2 a_2^2 c_2) - \\
\frac{(m_1^0)^2 + 1}{v_1^0 m_1^0} (d_1^2 b_1 c_1 + d_2^2 b_2 c_2) + d_1^2 c_1^2 + d_2^2 c_2^2 &= 0, \tag{4.20}
\end{aligned}$$

such that at least one among $a_1, a_2, b_1, b_2, c_1, c_2$ is non-zero and $d_1, d_2 \neq 0$. At the first glance, the behavior of this system may be dependent on the specific value of v_1^0, m_1^0 . However, if we remove the third, fifth and eighth equations, we obtain a system of real polynomials that does not depend on the specific value of G_0 . In fact, it can be verified that this system does not admit any non-trivial real solution. Thus, there does not exist any sequence of $G \in \mathcal{O}_{2,c_0}$ according to which all coefficients of the 4-minimal form vanish. So, G_0 is *not* 4-singular. We conclude that $\ell(G_0|\mathcal{O}_{2,c_0}) = 3$.

We end this illustrative exercise with a remark. The fact that there exists a subset of the limiting polynomials of the coefficients of r -minimal forms that do not depend on specific value of G_0 is very useful, because it allows us to provide an upper

bound on the singularity level the holds uniformly for all $G_0 \in \mathcal{S}_0$. It is interesting to note that this subset of polynomials also arises from the same analysis applied to the Gaussian kernels studied by [Ho and Nguyen, 2016a]. This observation can be partially explained by the fact that Gaussian kernels are a special case of skewnormal kernels with zero skewness. A highly nontrivial consequence from this observation is that the singularity level in a skewnormal mixture is always bounded from above than the singularity level in a Gaussian mixture. Thanks to Theorem 4.3.2 we arrive at a somewhat surprising conclusion that the MLE and minimax bounds for parameter estimation in skewnormal o-mixtures are generally *faster* than that of Gaussian o-mixtures. Now we are ready for results for the general setting of $G_0 \in \mathcal{S}_0$, which also articulates this remark more precisely.

4.4.2 General results

In this section we shall present results on $\ell(G_0 | \mathcal{O}_{k,c_0})$ for the general case $k > k_0$. To do so, we shall define the system of the limiting polynomials that characterizes the singularity level of G_0 .

Recall the notation introduced by the statement of Lemma 4.4.2. For given $r \geq 1$, for each $i = 1, \dots, k_0$, the system is given by the equations of real unknowns $(a_j, b_j, c_j, d_j)_{j=1}^{k-k_0+1}$:

$$\left\{ \sum_{j=1}^{k-k_0+1} \sum_{\alpha} \frac{P_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_i^0)}{H_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_i^0) Q_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(v_i^0)} \frac{d_j^2 a_j^{\alpha_1} b_j^{\alpha_2} c_j^{\alpha_3}}{\alpha_1! \alpha_2! \alpha_3!} = 0 \middle| \beta \in \mathcal{F}_r \cap \{\beta_1 + 2\beta_2 + 2\beta_3 \leq r\} \right\} \quad (4.21)$$

where the range of $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{N}^3$ in the above sum satisfies $\alpha_1 + 2\alpha_2 + 2\alpha_3 = \beta_1 + 2\beta_2 + 2\beta_3$.

Note that the above system of polynomial equations is the general version of the systems of polynomial equations described in the previous section. There are

$2r - 1$ equations in the above system of $4(k - k_0 + 1)$ unknowns. A solution of (4.21) is considered *non-trivial* if all of d_j are non-zeros while at least one among $a_1, \dots, a_i, b_1, \dots, b_i, c_1, \dots, c_i$ is non-zero. We say that system (4.21) is unsolvable if it does not have any non-trivial (or admissible) solution. The main result of this section is the following.

Theorem 4.4.1. *For each $i = 1, \dots, k_0$, let $\rho(v_i^0, m_i^0, k - k_0)$ be the minimum r for which system of polynomial equations (4.21) is unsolvable. Define*

$$R(G_0, k) = \max_{1 \leq i \leq k_0} \rho(v_i^0, m_i^0, k - k_0). \quad (4.22)$$

If $G_0 \in \mathcal{S}_0$, then $\ell(G_0 | \mathcal{O}_{k, c_0}) \leq R(G_0, k) - 1$.

Remark: We make the following comments regarding the results of Theorem 4.4.1.

- (i) If $k - k_0 = 1$, we can obtain $R(G_0, k) = 4$ from the examples given in Section 4.4.1 (although in the examples we only worked out the case that $k_0 = 1$, for general $k_0 \geq 1$ the techniques are the same). Since G_0 is in fact 3-singular, the bound is tight.
- (ii) In order to determine $R(G_0, k)$, we need to find the value of $\rho(v_i^0, m_i^0, k - k_0)$ for all $1 \leq i \leq k_0$. One may ask whether the value of $\rho(v_i^0, m_i^0, k - k_0)$ depends on the specific values of v_i^0, m_i^0 . The structure of $\rho(v_i^0, m_i^0, k - k_0)$ will be looked at in more detail in the next subsection.

Proof. The strategy is clear: First, we shall obtain a valid r -minimal form for G_0 , cf. Eq. (4.8). This requires a method for obtaining linearly independent basis functions $H_l(x)$ out of the partial derivatives of kernel density f . Second, we obtain the polynomial limits of collection of coefficients of the r -minimal form. Third, we obtain bounds on r according to which this system of limiting polynomials does not admit non-trivial real solutions. This provides upper bounds on the singularity level of G_0 .

Step 1: Construction of r -minimal form It follows from Lemma 4.4.2 and Lemma 4.4.3 that a r -th minimal form for G_0 can be obtained as

$$\frac{p_G(x) - p_{G_0}(x)}{W_r^r(G, G_0)} \asymp \frac{A_1(x) + B_1(x)}{W_r^r(G, G_0)},$$

where $A_1(x)$ and $B_1(x)$ are given as follows

$$\begin{aligned} A_1(x) &= \sum_{i=1}^{k_0} \sum_{\beta \in \mathcal{F}_r} \left(\sum_{j=1}^{s_i} \sum_{|\alpha| \leq r} \frac{P_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_i^0)}{H_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_i^0) Q_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(v_i^0)} \frac{p_{ij}(\Delta \theta_{ij})^{\alpha_1} (\Delta v_{ij})^{\alpha_2} (\Delta m_{ij})^{\alpha_3}}{\alpha_1! \alpha_2! \alpha_3!} \right) \\ &\quad \times \frac{\partial^{|\beta|} f}{\partial \theta^{\beta_1} \partial v^{\beta_2} \partial m^{\beta_3}}(x | \theta_i^0, \sigma_i^0, m_i^0), \\ B_1(x) &= \sum_{i=1}^{k_0} \Delta p_i.f(x | \theta_i^0, \sigma_i^0, m_i^0). \end{aligned}$$

Suppose that there exists a sequence of G tending to G_0 under W_r such that all the coefficients of $A_1(x)/W_r^r(G, G_0)$ and $B_1(x)/W_r^r(G, G_0)$ vanish. Then for all $1 \leq i \leq k_0$, we obtain that $\Delta p_i./W_r^r(G, G_0) \rightarrow 0$ and

$$E_{\beta_1, \beta_2, \beta_3}(\theta_i^0, v_i^0, m_i^0) := \frac{\sum_{j=1}^{s_i} \sum_{|\alpha| \leq r} \frac{P_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_i^0)}{H_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_i^0) Q_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(v_i^0)} \frac{p_{ij}(\Delta \theta_{ij})^{\alpha_1} (\Delta v_{ij})^{\alpha_2} (\Delta m_{ij})^{\alpha_3}}{\alpha_1! \alpha_2! \alpha_3!}}{W_r^r(G, G_0)} \rightarrow 0,$$

as $\beta \in \mathcal{F}_r$. According to Lemma 4.3.1,

$$W_r^r(G, G_0) \asymp D_r(G_0, G).$$

So, $\sum_{i=1}^{k_0} |\Delta p_i.|/D_r(G_0, G) \rightarrow 0$. It follows that

$$\sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij}(|\Delta \theta_{ij}|^r + |\Delta v_{ij}|^r + |\Delta m_{ij}|^r)/D_r(G_0, G) \rightarrow 1.$$

This means there exists some index $i^* \in \{1, \dots, k_0\}$ such that

$$\sum_{j=1}^{s_{i^*}} p_{i^*j} (|\Delta\theta_{i^*j}|^r + |\Delta v_{i^*j}|^r + |\Delta m_{i^*j}|^r) / D_r(G_0, G) \not\rightarrow 0.$$

By multiplying the inverse of the above term with $E_{\beta_1, \beta_2, \beta_3}(\theta_{i^*}^0, v_{i^*}^0, m_{i^*}^0)$ as $\beta \in \mathcal{F}_r$ and using the fact that $W_r(G, G_0) \asymp D_r(G_0, G)$, we obtain

$$F_{\beta_1, \beta_2, \beta_3}(\theta_{i^*}^0, v_{i^*}^0, m_{i^*}^0) := \frac{\sum_{j=1}^{s_1} \sum_{|\alpha| \leq r} \frac{P_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_{i^*}^0)}{H_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_{i^*}^0) Q_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(v_{i^*}^0)} \frac{p_{i^*j} (\Delta\theta_{i^*j})^{\alpha_1} (\Delta v_{i^*j})^{\alpha_2} (\Delta m_{i^*j})^{\alpha_3}}{\alpha_1! \alpha_2! \alpha_3!}}{\sum_{j=1}^{s_1} p_{i^*j} (|\Delta\theta_{i^*j}|^r + |\Delta v_{i^*j}|^r + |\Delta m_{i^*j}|^r)} \rightarrow 0,$$

Step 2: Greedy extraction of polynomial limits We proceed to extract polynomial limits of all $F_{\beta_1, \beta_2, \beta_3}(\theta_{i^*}^0, v_{i^*}^0, m_{i^*}^0)$. This technique has been demonstrated in Section 4.4.1 for specific cases. Note that the numerators of the $F_{\beta_1, \beta_2, \beta_3}(\theta_{i^*}^0, v_{i^*}^0, m_{i^*}^0)$ are inhomogeneous polynomials in general. Let

$$\overline{M}_g = \max \left\{ |\Delta\theta_{i^*1}|, \dots, |\Delta\theta_{i^*s_{i^*}}|, |\Delta v_{i^*1}|^{1/2}, \dots, |\Delta v_{i^*s_{i^*}}|^{1/2}, |\Delta m_{i^*1}|^{1/2}, \dots, |\Delta m_{i^*s_{i^*}}|^{1/2} \right\}.$$

Denote the limits for the relevant subsequences, which exist due to the boundedness: $\Delta\theta_{i^*j}/\overline{M}_g \rightarrow a_j$, $\Delta v_{i^*j}/\overline{M}_g^2 \rightarrow b_j$, and $\Delta m_{i^*j}/\overline{M}_g^2 \rightarrow c_j$, and $p_{i^*j} \rightarrow d_j^2$ for each $j = 1, \dots, s_{i^*}$. Here, at least one element of $(a_j, b_j, c_j)_{j=1}^{s_{i^*}}$ equals to -1 or 1. For any index vector $\beta = (\beta_1, \beta_2, \beta_3)$ such that $\beta \in \mathcal{F}_r$, the lowest order of \overline{M}_g in the numerator of $F_{\beta_1, \beta_2, \beta_3}(\theta_{i^*}^0, v_{i^*}^0, m_{i^*}^0)$ is $\overline{M}_g^{\beta_1+2\beta_2+2\beta_3}$. Since $\sum_{j=1}^{s_1} p_{i^*j} (|\Delta\theta_{i^*j}|^r + |\Delta v_{i^*j}|^r + |\Delta m_{i^*j}|^r) \asymp \overline{M}_g^r$, it is clear that $F_{\beta_1, \beta_2, \beta_3}(\theta_{i^*}^0, v_{i^*}^0, m_{i^*}^0)$ vanishes as long as $\beta_1 + 2\beta_2 + 2\beta_3 \geq r + 1$. Thus, we only need to concern with $F_{\beta_1, \beta_2, \beta_3}(\theta_{i^*}^0, v_{i^*}^0, m_{i^*}^0)$ when $\beta \in \mathcal{F}_r$ and $\beta_1 + 2\beta_2 + 2\beta_3 \leq r$.

For any $\beta = (\beta_1, \beta_2, \beta_3)$ such that $\beta \in \mathcal{F}_r$ and $\beta_1 + 2\beta_2 + 2\beta_3 \leq r$, by dividing the numerator and denominator of $F_{\beta_1, \beta_2, \beta_3}(\theta_{i^*}^0, v_{i^*}^0, m_{i^*}^0)$ by $\overline{M}_g^{\beta_1+2\beta_2+2\beta_3}$ (i.e the lowest order of \overline{M}_g in the numerator of $F_{\beta_1, \beta_2, \beta_3}(\theta_{i^*}^0, v_{i^*}^0, m_{i^*}^0)$), we obtain the following system of equations

$$\sum_{j=1}^{s_{i^*}} \sum_{\alpha} \frac{P_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_{i^*}^0)}{H_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m_{i^*}^0) Q_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(v_{i^*}^0)} \frac{d_j^2 a_j^{\alpha_1} b_j^{\alpha_2} c_j^{\alpha_3}}{\alpha_1! \alpha_2! \alpha_3!} = 0, \quad (4.23)$$

where the range of $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ in the above sum satisfies $\alpha_1 + 2\alpha_2 + 2\alpha_3 = \beta_1 + 2\beta_2 + 2\beta_3$. The above system of polynomial equations is the general version of the systems of polynomial equations (4.19) and (4.20) that we considered in Section 4.4.1. Now, one of the elements of a_j s, b_j s, c_j s is non-zero. Since $G \in \mathcal{O}_{k, c_0}$ and $\sum_{j=1}^{s_{i^*}} p_{i^* j} \rightarrow p_{i^*}^0$, we have the constraints $d_j^2 > 0$ and $\sum_{j=1}^{s_{i^*}} d_j^2 = p_{i^*}^0$. However, we can remove the constraint on the summation of d_j^2 by putting $d_j^2 = p_{i^*}^0 (d'_j)^2 / \sum_{j=1}^{s_{i^*}} (d'_j)^2$ where we the only constraint on d'_j s is $d'_j \neq 0$ for all $1 \leq j \leq s_{i^*}$. As a consequence, when we talk about system of polynomial equations (4.23), we can consider only the constraint $d_j^2 \neq 0$ for any $1 \leq j \leq s_{i^*}$. By Definition 4.3.2, G_0 is not r -singular relative to \mathcal{O}_{k, c_0} as long as the system (4.23) does not admit any non-trivial solution for the unknowns $(a_j, b_j, c_j, d_j)_{j=1}^{s_{i^*}}$.

Step 3: Deriving an upper bound There are two distinct features of system of polynomial equations (4.23). First, i^* varies in $\{1, 2, \dots, k_0\}$ as $G \in \mathcal{O}_{k, c_0}$ tends to G_0 . Second, the value of s_{i^*} of the subsequence of G is subject to the constraint that $s_{i^*} \leq k - k_0 + 1$. (This constraint arises due to number of distinct atoms of G , $\sum_{j=1}^{k_0} s_j \leq k' \leq k$ and all $s_j \geq 1$ for all $1 \leq j \leq k_0$). It follows from these two observations that the system (4.23) admits a non-trivial solution only if the system (4.21) also admits a non-trivial solution. This cannot be the case if $r \geq R(G_0, k)$, by the definition given in Eq. (4.22). This concludes our proof.

□

4.4.3 Properties of the system of limiting polynomial equations

The goal of this subsection is the present additional results on the structure of function $\rho(v, m, k - k_0)$, which is a fundamental quantity in Theorem 4.4.1 (Here, v_i^0, m_i^0 are replaced by v, m). It is difficult to obtain explicit values for $\rho(v, m, k - k_0)$ in general. Nonetheless, we can obtain a nontrivial upper bound for ρ . Now, let $\Xi_1 := \{(v, m) \in \Theta_2 \times \Theta_3 : m \neq 0\}$. Recall that $\rho(v, m, l)$, where $l = k - k_0 \geq 1$, is the minimum value according to which system (4.21) does not admit non-trivial real-solution.

Proposition 4.4.1. *Let $\bar{r}(l)$ the minimal value of $s > 0$ such that the following system of polynomial equations*

$$\sum_{j=1}^l \sum_{\substack{n_1+2n_2=\alpha \\ n_1, n_2 \geq 0}} \frac{x_j^2 y_j^{n_1} z_j^{n_2}}{n_1! n_2!} = 0 \quad \text{for each } \alpha = 1, \dots, s \quad (4.24)$$

does not have any solution for $(x_1, \dots, x_l, y_1, \dots, y_l, z_1, \dots, z_l)$ such that x_1, \dots, x_l are non-zeros, and at least one of y_1, \dots, y_l is non-zero. For all $l = 1, 2, \dots$, there holds

$$\sup_{(v,m) \in \Xi_1} \rho(v, m, l) \leq \bar{r}(l).$$

Remarks (i) The proof of this proposition is given in Appendix B, which proceeds by verifying that system (4.24) forms a subset of equations that define system (4.21). Combining with the statement of Theorem 4.4.1, we obtain

$$\ell(G_0 | \mathcal{O}_{k,c_0}) \leq \bar{r}(l) - 1.$$

(ii) A remarkable fact is that $\bar{r}(l)$ is nothing but the singularity level of G_0 relative to \mathcal{O}_{k,c_0} in the context of location-scale Gaussian mixture. This statement can be proved

directly using the same method of proof described in the previous section for the skewnormal mixtures. The proof for the Gaussian mixture is much simpler, because the r -minimal form for Gaussian mixtures can be obtained via the relatively simpler elimination steps given by Example 4.3.3. The fact that the coefficients involved in this elimination are constant with respect to the model parameters is the fundamental reason why the singularity level of G_0 for the Gaussian mixtures is uniform over the entire space of parameters. See also Theorem 1.1 of Ho and Nguyen [2016a].

(iii) Combining the above remark with the results established by Theorem 4.3.2 leads us to conclude this: it is statistically more efficient to estimate location-scale-shape parameters of skewnormal o-mixtures than to estimate location-scale parameters of Gaussian o-mixtures that carry the same number of extra mixing components.

Dependence of ρ on (v, m) To understand the role of parameter value (v, m) on singularity levels, we shall construct a partition of the parameter space for (v, m) based on the value of function ρ . For each $l, r \geq 1$, define an “inverse” function

$$\rho_l^{-1}(r) = \{(v, m) \in \Xi_1 : \rho(v, m, l) = r\}.$$

Additionally, take

$$\varrho(l) = \min \{r : \rho_l^{-1}(r) \neq \emptyset\}, \quad \bar{\rho}(l) = \max \{r : \rho_l^{-1}(r) \neq \emptyset\}.$$

It follows from Proposition 4.4.1 that $\bar{\rho}(l) \leq \bar{r}(l)$. In addition, $\rho_l^{-1}(r)$ are mutually disjoint for different values of r . So, for each fixed $l \geq 1$,

$$\Xi_1 = \bigcup_{r=\varrho(l)}^{\bar{\rho}(l)} \rho_l^{-1}(r).$$

Proposition 4.4.2. *For each $l \geq 1, r \geq 1$, $\rho_l^{-1}(r)$ is a semialgebraic set.*

Proof. For each $r \geq 1$, let \mathbb{A}_r be the collection of all $(v, m) \in \Xi_1$ such that the system of polynomial equations (4.21) contains admissible solutions. Furthermore, \mathbb{B}_r denotes the collection of all solutions $(v, m, \{a_i\}_{i=1}^l, \{b_i\}_{i=1}^l, \{c_i\}_{i=1}^l, \{d_i\}_{i=1}^l)$ of system of polynomial equations (4.21), i.e., we treat v, m as two additional unknowns of the system. Since $P_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m)$, $H_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m)$, and $Q_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(v)$ are polynomial functions of m, v for all α, β , by definition \mathbb{B}_r is a semialgebraic set for all $r \geq 1$. By Tarski-Seidenberg theorem [Basu et al., 2006], since \mathbb{A}_r is the projection of \mathbb{B}_r from dimension $(4l + 2)$ to dimension 2, \mathbb{A}_r is a semialgebraic set for all $r \geq 1$. It follows that \mathbb{A}_r^c is semialgebraic for all $r \geq 1$. Since $\rho_l^{-1}(r) = \mathbb{A}_r^c \cap \mathbb{A}_{r-1}$ for all $r \geq 1$, the conclusion of the proposition follows. \square

The following result gives us some exact values of $\varrho(l)$ and $\bar{\varrho}(l)$ in specific cases.

Proposition 4.4.3. (a) If $l = k - k_0 = 1$, then $\varrho(l) = \bar{\varrho}(l) = 4$.

(b) If $l = k - k_0 = 2$, then $\varrho(l) = 5$ and $\bar{\varrho}(l) = 6$. Thus, Ξ_1 is partitioned into two subsets, both of which are non-empty because $\{(1, -2), (1, 2)\} \subset \rho_l^{-1}(5)$, and $(1, \frac{1}{10}) \in \rho_l^{-1}(6)$.

From the definition of $R(G_0, k)$, we can write

$$R(G_0, k) = \max \left\{ r \left| \text{there is } i = 1, \dots, k_0 \text{ such that } (v_i^0, m_i^0) \in \rho_{k-k_0}^{-1}(r) \right. \right\}.$$

According to the Proposition 4.4.3, if $k - k_0 = 1$, we have $R(G_0, k) = 4$ (see also our earlier remark). If $k - k_0 = 2$, we may have either $R(G_0, k) = 5$ or 6, depending on the value of parameters (v, m) that provide the support for G_0 .

We end this subsection by noting that we have just provided specific examples in which $R(G_0, k) - 1$ may vary with the actual parameter values that define G_0 . Although this is an upper bound of the singularity level, we have *not* actually proved that the singularity level of G_0 may generally vary with its parameter values. We will

be able to do so in the sequel, when we work with the e-mixture setting.

4.5 E-mixtures of skewnormal distributions

E-mixtures are the setting in which the number of mixing components is known $k = k_0$. In this section, we study the singularity structure of mixing measure G_0 relative to the ambient space \mathcal{E}_{k_0} , where k_0 is the number of supporting atoms for G_0 .

Recall from the previous section the definition of \mathcal{S}_0 , the subset $\mathcal{S}_0 \subset \mathcal{E}_{k_0}$ of measure $G_0 = G_0(\boldsymbol{p}^0, \boldsymbol{\eta}^0)$ such that $(\boldsymbol{p}^0, \boldsymbol{\eta}^0)$ satisfy $P_1(\boldsymbol{\eta}^0)P_2(\boldsymbol{\eta}^0) \neq 0$. P_1 and P_2 are polynomials given in the statement of Lemma 4.4.1. It is simple to verify that for any $G_0 \in \mathcal{S}_0$, as a consequence of this lemma, the Fisher information matrix $I(G_0)$ is non-singular. It follows that

Theorem 4.5.1. *If $G_0 \in \mathcal{S}_0$, then $\ell(G_0 | \mathcal{E}_{k_0}) = 0$.*

We turn our attention to the singularity structure of set $\mathcal{E}_{k_0} \setminus \mathcal{S}_0$. For any $G_0 \in \mathcal{E}_{k_0} \setminus \mathcal{S}_0$, the parameters of G_0 satisfy $P_1(\boldsymbol{\eta}^0)P_2(\boldsymbol{\eta}^0) = 0$. Accordingly, for each pair of $(i, j) = 1, \dots, k_0$ the two components indexed by i and j are said to be **homologous** if

$$(\theta_i^0 - \theta_j^0)^2 + [v_i^0(1 + (m_i^0)^2) - v_j^0(1 + (m_j^0)^2)]^2 = 0.$$

Moreover, for each $1 \leq i \leq k_0$, let I_i denote the set of all components homologous to (component) i . By definition, it is clear that if i and j are homologous, $I_i \equiv I_j$. Therefore, these homologous sets form equivalence classes. From here on, when we say a homologous set I , we implicitly mean that it is the representation of the equivalent classes.

Now, the homologous set consists of the indices of skewnormal components that share the same location and a rescaled version of the scale parameter. A non-empty homologous set I is said to be **conformant** if for any $i \neq j \in I$, $m_i^0 m_j^0 > 0$. A non-empty homologous set I is said to be **nonconformant** if we can find two indices

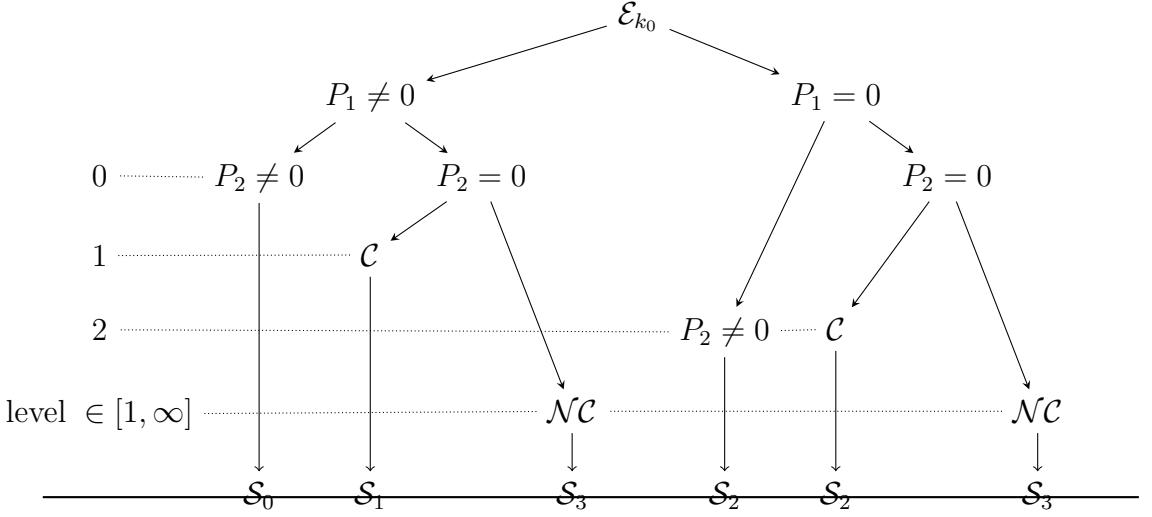


Figure 4.2: The singularity level of G_0 relative to \mathcal{E}_{k_0} is determined by partition based on zeros of polynomials P_1, P_2 into subsets $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \mathcal{S}_3$. Here, "NC" stands for nonconformant.

$i, j \in I$ such that $m_i^0 m_j^0 < 0$. Additionally, G_0 is said to be **conformant** if all the homologous sets are conformant or **nonconformant** (NC) if at least one homologous set is nonconformant. Now, we define a partition of $\mathcal{E}_{k_0} \setminus \mathcal{S}_0$ as follows $\mathcal{E}_{k_0} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2 \cup \mathcal{S}_3$, where

$$\begin{cases} \mathcal{S}_1 = \{G = G(\boldsymbol{p}, \boldsymbol{\eta}) \in \mathcal{E}_{k_0} \mid P_1(\boldsymbol{\eta}) \neq 0, P_2(\boldsymbol{\eta}) = 0, G \text{ is conformant}\} \\ \mathcal{S}_2 = \{G = G(\boldsymbol{p}, \boldsymbol{\eta}) \in \mathcal{E}_{k_0} \mid P_1(\boldsymbol{\eta}) = 0, \text{ if } P_2(\boldsymbol{\eta}) = 0 \text{ then } G \text{ is conformant}\} \\ \mathcal{S}_3 = \{G = G(\boldsymbol{p}, \boldsymbol{\eta}) \in \mathcal{E}_{k_0} \mid P_2(\boldsymbol{\eta}) = 0 \text{ and } G \text{ is nonconformant}\}. \end{cases}$$

Figure 4.2 summarizes singularity levels of elements residing in \mathcal{E}_{k_0} , except for \mathcal{S}_3 .

4.5.1 Singularity level of $G_0 \in \mathcal{S}_1 \cup \mathcal{S}_2$

The main results of this subsection are the following two theorems

Theorem 4.5.2. *If $G_0 \in \mathcal{S}_1$, then $\ell(G_0 | \mathcal{E}_{k_0}) = 1$.*

Theorem 4.5.3. *If $G_0 \in \mathcal{S}_2$, then $\ell(G_0 | \mathcal{E}_{k_0}) = 2$.*

The complete proofs of both theorems are given in in Appendix. In the following, we shall present the proof for a simple setting of $G_0 \in \mathcal{S}_1$, which illustrates the complete proofs, and also helps to explain why the partition of according to \mathcal{S}_1 , i.e., the notion of conformant, arises in the first place.

Proof. (of a simplified setting) The simplified setting is that all components of G_0 are homologous to one another. By definition all components of G_0 are non-Gaussian (because $P_1(\boldsymbol{\eta}^0) \neq 0$). Thus, we have $\theta_1^0 = \dots = \theta_{k_0}^0$ and $\frac{v_1^0}{1 + (m_1^0)^2} = \dots = \frac{v_{k_0}^0}{1 + (m_{k_0}^0)^2}$. Additionally, $m_i^0 \neq 0$ for all $1 \leq i \leq k_0$. Since G_0 is conformant, m_i^0 share the same sign for all $1 \leq i \leq k_0$. Without loss of generality, we assume $m_i^0 > 0$. We need to show that G_0 is 1-singular, but not 2-singular.

G_0 is 1-singular Given constraints on the parameters of G_0 , it is simple to arrive at the following 1-minimal form (cf. Eq. (4.8)):

$$\begin{aligned} & \frac{1}{W_1(G, G_0)} \left\{ \sum_{i=1}^{k_0} \left[\beta_{1i}^{(1)} + \beta_{2i}^{(1)}(x - \theta_1^0) + \beta_{3i}^{(1)}(x - \theta_1^0)^2 \right] f \left(\frac{x - \theta_1^0}{\sigma_i^0} \right) \Phi \left(\frac{m_i^0(x - \theta_1^0)}{\sigma_i^0} \right) \right. \\ & \left. + \left[\gamma_1^{(1)} + \gamma_2^{(1)}(x - \theta_1^0) \right] \exp \left(-\frac{(m_1^0)^2 + 1}{2v_1^0}(x - \theta_1^0)^2 \right) \right\} + o(1), \end{aligned} \quad (4.25)$$

where coefficients $\beta_{1i}^{(1)}, \beta_{2i}^{(1)}, \beta_{3i}^{(1)}, \gamma_1^{(1)}, \gamma_2^{(1)}$ are the polynomials of $\Delta\theta_j, \Delta v_j, \Delta m_j$, and Δp_j :

$$\begin{aligned} \beta_{1i}^{(1)} &= \frac{2\Delta p_i}{\sigma_i^0} - \frac{p_i \Delta v_i}{(\sigma_i^0)^3}, \quad \beta_{2i}^{(1)} = \frac{2p_i \Delta \theta_i}{(\sigma_i^0)^3}, \quad \beta_{3i}^{(1)} = \frac{p_i \Delta v_i}{(\sigma_i^0)^5}, \\ \gamma_1^{(1)} &= \sum_{j=1}^{k_0} -\frac{p_j m_j^0 \Delta \theta_j}{\pi(\sigma_j^0)^2}, \quad \gamma_2^{(1)} = \sum_{j=1}^{k_0} -\frac{p_j m_j^0 \Delta v_j}{2\pi(\sigma_j^0)^4} + \frac{p_j \Delta m_j}{\pi(\sigma_j^0)^2}. \end{aligned}$$

Note that, the conditions $m_i^0 \neq 0$ for all $1 \leq i \leq k_0$ allow us to have the following terms $f \left(\frac{x - \theta_1^0}{\sigma_i^0} \right) \Phi \left(\frac{m_i^0(x - \theta_1^0)}{\sigma_i^0} \right)$ and $\exp \left(-\frac{(m_1^0)^2 + 1}{2v_1^0}(x - \theta_1^0)^2 \right)$ are linearly independent. It is clear that if a sequence of G (represented by Eq. (4.4)) is chosen

such that $\Delta\theta_i = \Delta v_i = \Delta p_i = 0$ for all $1 \leq i \leq k_0$, and $\sum_{i=1}^{k_0} p_i \Delta m_i / v_i^0 = 0$, then we obtain $\beta_{1i}^{(1)}/W_1(G, G_0) = \beta_{2i}^{(1)}/W_1(G, G_0) = \beta_{3i}^{(1)}/W_1(G, G_0) = \gamma_1^{(1)}/W_1(G, G_0) = \gamma_2^{(1)}/W_1(G, G_0) = 0$. Hence, G_0 is 1-singular relative to \mathcal{E}_{k_0} .

G_0 is not 2-singular Indeed, suppose that this is not true. Then from Definition 4.3.2, for any sequence of G that tends to G_0 under W_2 , all coefficients of the 2-minimal form must vanish. A 2-minimal form is given as follows:

$$\begin{aligned} & \frac{1}{W_2^2(G, G_0)} \left[\sum_{i=1}^{k_0} \left(\sum_{j=1}^5 \beta_{ji}^{(2)} (x - \theta_1^0)^{j-1} \right) f \left(\frac{x - \theta_1^0}{\sigma_i^0} \right) \Phi \left(\frac{m_i^0(x - \theta_1^0)}{\sigma_i^0} \right) \right. \\ & \quad \left. + \left(\sum_{j=1}^4 \gamma_j^{(2)} (x - \theta_1^0)^{j-1} \right) \exp \left(-\frac{(m_1^0)^2 + 1}{2v_1^0} (x - \theta_1^0)^2 \right) \right] + o(1), \end{aligned} \quad (4.26)$$

where $\beta_{ji}^{(2)}, \gamma_j^{(2)}$ are polynomials of $\Delta\theta_l, \Delta v_l, \Delta m_l$, and Δp_l for $l = 1, \dots, k_0$:

$$\begin{aligned} \beta_{1i}^{(2)} &= \frac{2\Delta p_i}{\sigma_i^0} - \frac{p_i \Delta v_i}{(\sigma_i^0)^3} - \frac{p_i (\Delta\theta_i)^2}{(\sigma_i^0)^3} + \frac{3p_i (\Delta v_i)^2}{4(\sigma_i^0)^5}, \quad \beta_{2i}^{(2)} = \frac{2p_i \Delta\theta_i}{(\sigma_i^0)^3} - \frac{6p_i \Delta\theta_i \Delta v_i}{(\sigma_i^0)^5}, \\ \beta_{3i}^{(2)} &= \frac{p_i \Delta v_i}{(\sigma_i^0)^5} + \frac{p_i (\Delta\theta_i)^2}{(\sigma_i^0)^5} - \frac{3p_i (\Delta v_i)^2}{2(\sigma_i^0)^7}, \quad \beta_{4i}^{(2)} = \frac{2p_i \Delta\theta_i \Delta v_i}{(\sigma_i^0)^7}, \quad \beta_{5i}^{(2)} = \frac{p_i (\Delta v_i)^2}{4(\sigma_i^0)^9}, \\ \gamma_1^{(2)} &= \sum_{j=1}^{k_0} -\frac{p_j m_j^0 \Delta\theta_j}{\pi(\sigma_j^0)^2} + \frac{2p_j m_j^0 (\Delta\theta_j)(\Delta v_j)}{\pi(\sigma_j^0)^4} - \frac{2p_j \Delta\theta_j \Delta m_j}{\pi(\sigma_j^0)^2}, \\ \gamma_2^{(2)} &= \sum_{j=1}^{k_0} -\frac{p_j m_j^0 \Delta v_j}{2\pi(\sigma_j^0)^4} - \frac{p_j ((m_j^0)^3 + 2m_j^0)(\Delta\theta_j)^2}{2\pi(\sigma_j^0)^4} + \frac{p_j \Delta m_j}{\pi(\sigma_j^0)^2} + \frac{5p_j m_j^0 (\Delta v_j)^2}{8\pi(\sigma_j^0)^6} - \frac{p_j \Delta v_j \Delta m_j}{\pi(\sigma_j^0)^4}, \\ \gamma_3^{(2)} &= \sum_{j=1}^{k_0} \frac{p_j (2(m_j^0)^2 + 2)\Delta\theta_j \Delta m_j}{\pi(\sigma_j^0)^4} - \frac{p_j ((m_j^0)^3 + 2m_j^0)\Delta\theta_j \Delta v_j}{2\pi(\sigma_j^0)^6}, \\ \gamma_4^{(2)} &= \sum_{j=1}^{k_0} -\frac{p_j ((m_j^0)^3 + 2m_j^0)(\Delta v_j)^2}{8\pi(\sigma_j^0)^8} - \frac{p_j m_j^0 (\Delta m_j)^2}{2\pi(\sigma_j^0)^4} + \frac{p_j ((m_j^0)^2 + 1)\Delta v_j \Delta m_j}{\pi(\sigma_j^0)^6}. \end{aligned}$$

Now, $\beta_{ji}^{(2)}/W_2^2(G, G_0) \rightarrow 0$ leads to $\Delta p_i/W_2^2(G, G_0), \Delta\theta_i/W_2^2(G, G_0), \Delta v_i/W_2^2(G, G_0) \rightarrow 0$ for all $1 \leq i \leq k_0$ (The rigorous argument for that result is in Step 1.1 of the full

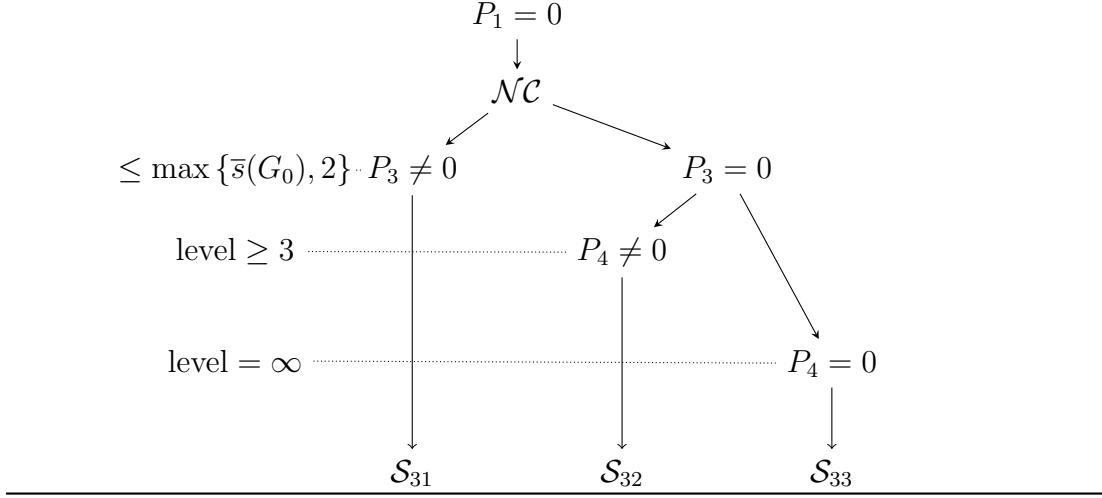


Figure 4.3: The level of singularity structure of $G_0 \in \mathcal{S}_3$ when $P_1(\boldsymbol{\eta}^0) = 0$. Here, "NC" stands for nonconformant. The term $\bar{s}(G_0)$ is defined in (4.37).

proof of this theorem in Appendix B). Combining with Lemma 4.3.1, we obtain

$$\sum_{i=1}^{k_0} p_i |\Delta m_i|^2 / W_2^2(G, G_0) \not\rightarrow 0. \quad (4.27)$$

Additionally, the vanishing of coefficients $\gamma_j^{(2)} / W_2^2(G, G_0)$ for $1 \leq j \leq 4$ entails

$$\begin{aligned} & \left(\sum_{i=1}^{k_0} p_i \Delta m_i / v_i^0 \right) / W_2^2(G, G_0) \rightarrow 0, \\ & \left(\sum_{i=1}^{k_0} p_i m_i^0 (\Delta m_i)^2 / (v_i^0)^2 \right) / W_2^2(G, G_0) \rightarrow 0. \end{aligned} \quad (4.28)$$

Combining (4.27) and (4.28), it follows that

$$\left(\sum_{i=1}^{k_0} p_i m_i^0 (\Delta m_i)^2 / (v_i^0)^2 \right) / \sum_{i=1}^{k_0} p_i |\Delta m_i|^2 \rightarrow 0,$$

which is a contradiction due to $m_i^0 > 0$ for all $1 \leq i \leq k_0$. Hence, G_0 is not 2-singular relative to \mathcal{E}_{k_0} . We conclude that $\ell(G_0 | \mathcal{E}_{k_0}) = 1$. \square

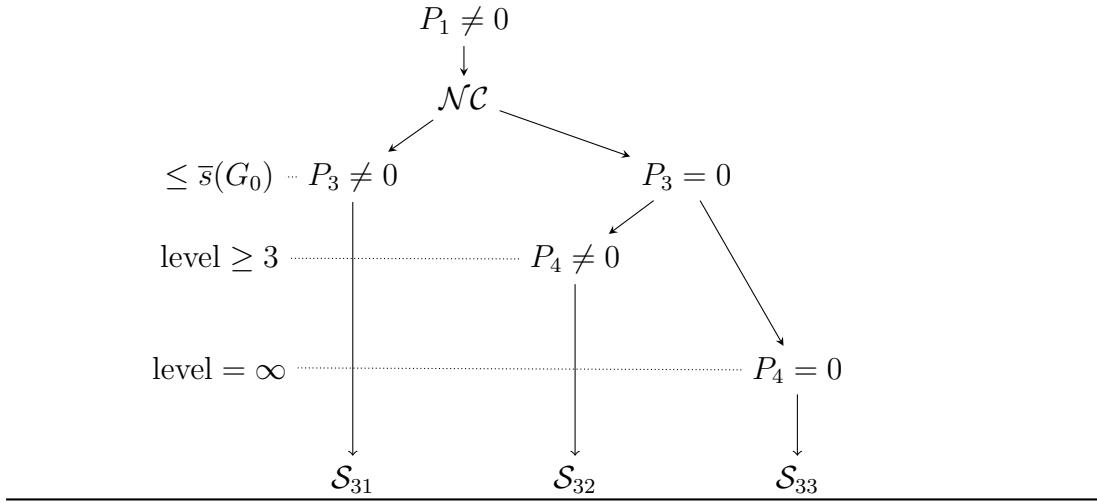


Figure 4.4: The level of singularity structure of $G_0 \in \mathcal{S}_3$ when $P_1(\boldsymbol{\eta}^0) \neq 0$. Here, "NC" stands for nonconformant. The term $\bar{s}(G_0)$ is defined in (4.37).

4.5.2 Singularity levels of $G_0 \in \mathcal{S}_3$: a summary

The singularity structure of \mathcal{S}_3 is much more complex than those of previous settings of G_0 . \mathcal{S}_3 does not admit an uniform level of singularity for all its elements — it needs to be partitioned into many subsets via intersections with additional semialgebraic sets of the parameter space. In addition, we can establish the existence of subsets that correspond to the infinite level of singularity. In most cases when the singularity level is finite, we may only be able to provide some bounds rather than an exact values. As in o-mixtures, the unifying theme of such bounds is their connection to the solvability of a system of real polynomial equations.

If $G_0 = G_0(\mathbf{p}^0, \boldsymbol{\eta}^0) \in \mathcal{S}_3$, then its corresponding parameters satisfy $P_2(\boldsymbol{\eta}^0) = 0$, i.e., there is at least one homologous set of G_0 . Moreover, at least one such homologous set is nonconformant. For any $G_0 \in \mathcal{S}_3$, let I_1, \dots, I_t be all nonconformant homologous sets of G_0 . The singularity structures of \mathcal{S}_3 arise from the zeros of the following polynomials:

- Type C(1): $P_3(\mathbf{p}^0, \boldsymbol{\eta}^0) := \prod_{i=1}^t \left(\prod_{S \subseteq I_i, |S| \geq 2} \left(\sum_{j \in S} p_j^0 \prod_{l \neq j} m_l^0 \right) \right)$.

- Type C(2): $P_4(\mathbf{p}^0, \boldsymbol{\eta}^0) := \prod_{1 \leq i \neq j \leq k_0} \left[u_{ij}^2 + (m_i^0 \sigma_j^0 + m_j^0 \sigma_i^0)^2 + (p_i^0 \sigma_j^0 - p_j^0 \sigma_i^0)^2 \right]$,
where $u_{ij}^2 = (\theta_i^0 - \theta_j^0)^2 + (v_i^0(1 + (m_j^0)^2) - v_j^0(1 + (m_i^0)^2))^2$.

Type C singularities, including both C(1) and C(2), are distinguished from Type A and Type B singularities by the fact that the Type C polynomials are defined by not only component parameters $\boldsymbol{\eta}^0$, but also mixing probability parameters \mathbf{p}^0 . Note that C(1) singularity implies that there is some homologous set I_i of G_0 such that $\prod_{S \subseteq I_i, |S| \geq 2} \left(\sum_{j \in S} p_j^0 \prod_{l \neq j} m_l^0 \right) = 0$. A homologous set of G_0 having the above property is said to contain type C(1) singularity locally. Similarly, type C(2) singularity implies that there is some pair $1 \leq i \neq j \leq k_0$ such that $u_{ij}^2 + (m_i^0 \sigma_j^0 + m_j^0 \sigma_i^0)^2 + (p_i^0 \sigma_j^0 - p_j^0 \sigma_i^0)^2 = 0$. A homologous set of G_0 having this pair is said to contain type C(2) singularity locally. It can be easily checked that a homologous set containing type C(2) singularity must also contain type C(1) singularity, since $P_4(\mathbf{p}^0, \boldsymbol{\eta}^0) = 0$ entails $P_3(\mathbf{p}^0, \boldsymbol{\eta}^0) = 0$. Now, we define the following partition of \mathcal{S}_3 according to the definition of type C(1) and C(2) singularity: $\mathcal{S}_3 = \mathcal{S}_{31} \cup \mathcal{S}_{32} \cup \mathcal{S}_{33}$, where

$$\begin{cases} \mathcal{S}_{31} = \{G = G(\mathbf{p}, \boldsymbol{\eta}) \in \mathcal{S}_3 \mid P_3(\mathbf{p}, \boldsymbol{\eta}) \neq 0\} \\ \mathcal{S}_{32} = \{G = G(\mathbf{p}, \boldsymbol{\eta}) \in \mathcal{S}_3 \mid P_3(\mathbf{p}, \boldsymbol{\eta}) = 0, P_4(\mathbf{p}, \boldsymbol{\eta}) \neq 0\} \\ \mathcal{S}_{33} = \{G = G(\mathbf{p}, \boldsymbol{\eta}) \in \mathcal{S}_3 \mid P_3(\mathbf{p}, \boldsymbol{\eta}) = 0, P_4(\mathbf{p}, \boldsymbol{\eta}) = 0\}. \end{cases}$$

Due to the highly technical nature of our analysis of the singularity structure of \mathcal{S}_3 , we defer such details to Section 4.7.1 in Appendix A. Here, we only provide a summary of such results. Figure 4.3 and 4.4 provide additional illustrations. Specifically, when $G_0 \in \mathcal{S}_{31}$, it is shown that $\ell(G_0 | \mathcal{E}_{k_0}) \leq \max\{2, \bar{s}(G_0)\}$, where $\bar{s}(G_0)$ is defined by a system of polynomial equations that we obtain via a method of greedy extraction of polynomial limits, see Section 4.7.1.1. In some specific cases, the precise singularity level of $G_0 \in \mathcal{S}_{31}$ will be given. If $G_0 \in \mathcal{S}_{32}$, we need a more sophisticated method of extraction for polynomial limits; our technique is illustrated on a specific example

of G_0 in Section 4.7.1.2. Finally, if $G_0 \in \mathcal{S}_{33}$, it is shown that $\ell(G_0|\mathcal{E}_{k_0}) = \infty$ in Section 4.7.1.3.

4.6 Discussion and concluding remarks

Understanding the behavior of parameter estimates of mixture models is useful because the mixing parameters represent explicitly the heterogeneity of the underlying data population that mixture models are most suitable for. In this chapter, a general framework for the identification of singularity structure arising from finite mixture models is proposed. It is shown that the singularity levels of the model's parameter space directly determine minimax lower bounds and maximum likelihood estimation convergence rates, under conditions on the compactness of the parameter space.

The systematic identification of singularity levels and the implications on parameter estimation is a crucial step toward the development of more efficient model-based inference procedures. It is our view that such procedures must account for the presence of singular points residing in the parameter space of the model. As a matter of fact, there are quite a few examples of such efforts applied to specific statistical models, even if the picture of the singularity structure associating with those models might not have been discussed explicitly. This raises a question of whether or not it is possible to extend and generalize such techniques in order to address the presence of singularities in a direct fashion. We give several examples:

- (1) For overfitted mixture models, methods based on likelihood-based penalization techniques were shown to be quite effective (e.g., [Toussile and Gassiat, 2009, Chen, 2016]). Our work shows that parameter values residing in the vicinity of regions of high singularity levels should be hard to estimate efficiently. Can a penalization technique be generalized to regularize the estimates toward subsets containing singularity points of smaller levels?

- (2) Suitable choices of Bayesian prior have been proposed to induce favorable posterior contraction behavior for overfitted finite mixtures [Rousseau and Mengersen, 2011]. Can we develop an appropriate prior for the mixture model parameters, given our knowledge of singular points residing in the parameter space?
- (3) Reparametrization is an effective technique that can be employed to combat singularities present in the class of skewed distributions [Hallin and Ley, 2014]. It would be interesting to study if such reparameterization technique can be systematically developed for the mixture models as well.

We also expect that the theory of singularity structures carries important consequences on the computational complexity of parameter estimation procedures, including both optimization and sampling based methods. The non-uniform nature of the singularity levels reveals a complex structure of the likelihood function: regions in parameter space that carry low singularity levels may observe a relatively high curvature of the likelihood surface, while high singularity levels imply a “flatter” likelihood surface along a certain subspace of the parameters. Such a subspace is manifested by our construction of sequences of mixing measures that attest to the condition of r -singularity. It is of interest to exploit the explicit knowledge of singularity levels obtained for a given mixture model class, so as to improve upon the computational efficiency of the optimization and sampling procedures that operate on the model’s parameter space.

4.7 Appendix A

This Appendix contains additional results on the singularity structure of e-mixtures of skewnormal distributions.

4.7.1 Singularity structure of \mathcal{S}_3 : detailed analysis

To develop intuition and obtain bounds for singularity level for $G_0 \in \mathcal{S}_3$, we start by considering a simple case similar to the exposition of subsection 4.5.1. That is, G_0 has only one homologous set of size k_0 . $G_0 \in \mathcal{S}_3$ means that m_i^0 do not share the same signs for all $i = 1, \dots, k_0$. To investigate the singularity level for G_0 , we first obtain an r -minimal form, for $r \geq 2$, of $(p_G(x) - p_{G_0}(x))/W_r^r(G, G_0)$ by

$$\frac{1}{W_r^r(G, G_0)} \left[\sum_{i=1}^{k_0} \left(\sum_{j=1}^{2r+1} \beta_{ji}^{(r)} (x - \theta_1^0)^{j-1} \right) f\left(\frac{x - \theta_1^0}{\sigma_i^0}\right) \Phi\left(\frac{m_i^0(x - \theta_1^0)}{\sigma_i^0}\right) \right. \\ \left. + \left(\sum_{j=1}^{2r} \gamma_j^{(r)} (x - \theta_1^0)^{j-1} \right) \exp\left(-\frac{(m_1^0)^2 + 1}{2v_1^0} (x - \theta_1^0)^2\right) \right] + o(1), \quad (4.29)$$

where $\beta_{ji}^{(r)}, \gamma_j^{(r)}$ are polynomials of $\Delta\theta_l, \Delta v_l, \Delta m_l$, and Δp_l as $1 \leq i, l \leq k_0$ and $1 \leq j \leq 2r+1$. For concrete formulas of $\beta_{ji}^{(r)}, \gamma_j^{(r)}$, we note that for any $\alpha = (\alpha_1, \alpha_2, \alpha_3)$ such that $|\alpha| \leq r$, there holds

$$\frac{\partial^{|\alpha|} f}{\partial \theta^{\alpha_1} \partial v^{\alpha_2} \partial m^{\alpha_3}} = \left(\sum_{i=1}^{2r} \frac{U_i^{\alpha_1, \alpha_2, \alpha_3}(m)}{V_i^{\alpha_1, \alpha_2, \alpha_3}(v)} (x - \theta)^{i-1} \right) f\left(\frac{x - \theta}{\sigma}\right) f\left(\frac{m(x - \theta)}{\sigma}\right) + \\ \frac{1}{\sigma} \left(\sum_{i=1}^{2r+1} \frac{L_i^{\alpha_1, \alpha_2, \alpha_3}}{N_i^{\alpha_1, \alpha_2, \alpha_3}(v)} (x - \theta)^{i-1} \right) f\left(\frac{x - \theta}{\sigma}\right) \Phi\left(\frac{m(x - \theta)}{\sigma}\right).$$

In the above display $U_i^{\alpha_1, \alpha_2, \alpha_3}(m), V_i^{\alpha_1, \alpha_2, \alpha_3}(v), N_i^{\alpha_1, \alpha_2, \alpha_3}(v)$ are polynomials in terms of m, v and $L_i^{\alpha_1, \alpha_2, \alpha_3}$ are some constant numbers. As $\alpha_3 \geq 1$, we can further check

that $L_i^{\alpha_1, \alpha_2, \alpha_3} = 0$ for all $1 \leq i \leq 2r$ and α_1, α_2 such that $|\alpha| \leq r$. It follows that

$$\begin{aligned}\beta_{ji}^{(r)} &= \frac{2\Delta p_i}{\sigma_j^0} 1_{\{j=1\}} + \frac{1}{\sigma_i^0} \sum_{|\alpha| \leq r} \frac{L_j^{\alpha_1, \alpha_2, \alpha_3}}{N_j^{\alpha_1, \alpha_2, \alpha_3}(v_i^0)} \frac{p_i(\Delta\theta_i)^{\alpha_1} (\Delta v_i)^{\alpha_2} (\Delta m_i)^{\alpha_3}}{\alpha_1! \alpha_2! \alpha_3!}, \\ \gamma_j^{(r)} &= \sum_{i=1}^{k_0} \sum_{|\alpha| \leq r} \frac{U_j^{\alpha_1, \alpha_2, \alpha_3}(m_i^0)}{V_j^{\alpha_1, \alpha_2, \alpha_3}(v_i^0)} \frac{p_i(\Delta\theta_i)^{\alpha_1} (\Delta v_i)^{\alpha_2} (\Delta m_i)^{\alpha_3}}{\alpha_1! \alpha_2! \alpha_3!},\end{aligned}$$

where $1 \leq i \leq k_0$ and $1 \leq j \leq 2r+1$. Since $L_j^{\alpha_1, \alpha_2, \alpha_3} = 0$ as $\alpha_3 \geq 1$, we further obtain that

$$\beta_{ji}^{(r)} = \frac{2\Delta p_i}{\sigma_j^0} 1_{\{j=1\}} + \frac{1}{\sigma_i^0} \sum_{\alpha_1 + \alpha_2 \leq r} \frac{L_j^{\alpha_1, \alpha_2, 0}}{N_j^{\alpha_1, \alpha_2, 0}(v_i^0)} \frac{p_i(\Delta\theta_i)^{\alpha_1} (\Delta v_i)^{\alpha_2}}{\alpha_1! \alpha_2!}.$$

Therefore, $\beta_{ji}^{(r)}$ are polynomials of $\Delta p_i, \Delta\theta_i, \Delta v_i$, while $\gamma_j^{(r)}$ are polynomials in terms of $\Delta\theta_i, \Delta v_i, \Delta m_i$, for $1 \leq i \leq k_0, 1 \leq j \leq 2r+1$.

Suppose that there is a sequence of G tending to G_0 (in W_r distance) such that all coefficients of its r -minimal form in (4.29) vanish. It can be checked that $\beta_{ji}^{(r)}/W_r^r(G, G_0) \rightarrow 0$ for all even $j \in [1, 2r+1]$ entails that $\Delta\theta_i/W_r^r(G, G_0) \rightarrow 0$ for all $1 \leq i \leq k_0$. Similarly, $\beta_{ji}^{(r)}/W_r^r(G, G_0) \rightarrow 0$ for all odd $j \in [3, 2r+1]$ entails that $\Delta v_i/W_r^r(G, G_0) \rightarrow 0$ for all $1 \leq i \leq k_0$. So, as $\beta_{1i}^{(r)}/W_r^r(G, G_0) \rightarrow 0$, we obtain $\Delta p_i/W_r^r(G, G_0) \rightarrow 0$. It follows that, as $\beta_{ji}^{(r)}/W_r^r(G, G_0) \rightarrow 0$ for all $1 \leq j \leq 2r+1$, we must have $\Delta p_i/W_r^r(G, G_0), \Delta\theta_i/W_r^r(G, G_0), \Delta v_i/W_r^r(G, G_0) \rightarrow 0$ for all $1 \leq i \leq k_0$. These results imply that

$$\frac{\sum_{i=1}^{k_0} |\Delta p_i| + p_i(|\Delta\theta_i|^r + |\Delta v_i|^r)}{W_r^r(G, G_0)} \rightarrow 0.$$

If $\Delta m_i = 0$ for all $1 \leq i \leq k_0$, then by means of Lemma 4.3.1,

$$\sum_{i=1}^{k_0} |\Delta p_i| + p_i(|\Delta\theta_i|^r + |\Delta v_i|^r) = D_r(G_0, G) \asymp W_r^r(G, G_0),$$

which contradicts with the above limit. Therefore, we must have $\max_{1 \leq i \leq k_0} |\Delta m_i| > 0$.

Turning to $\gamma_l^{(r)}$ and the fact that

$$\Delta p_i/W_r^r(G, G_0), \Delta \theta_i/W_r^r(G, G_0), \Delta v_i/W_r^r(G, G_0) \rightarrow 0,$$

if $\gamma_l^{(r)}/W_r^r(G, G_0) \rightarrow 0$ as $1 \leq l \leq 2r$, we also have that

$$\left(\sum_{i=1}^{k_0} \sum_{\alpha_3 \leq r} \frac{U_l^{0,0,\alpha_3}(m_i^0)}{V_l^{0,0,\alpha_3}(v_i^0)} \frac{p_i(\Delta m_i)^{\alpha_3}}{\alpha_3!} \right) / W_r^r(G, G_0) \rightarrow 0.$$

We can verify that as $1 \leq l \leq 2r$ is odd, $U_j^{0,0,\alpha_3}(m_i^0) = 0$ for all $\alpha_3 \leq r$ and $1 \leq i \leq k_0$.

Additionally, as $1 \leq l \leq 2r$ is even, the above system of limits becomes

$$\left(\sum_{i_1-i_2=l/2} \frac{q_{i_1,i_2}}{i_1!} \sum_{i=1}^{k_0} \frac{p_i(m_i^0)^{i_1-2i_2-1} (\Delta m_i)^{i_1}}{(\sigma_i^0)^l} \right) / W_r^r(G_0, G) \rightarrow 0, \quad (4.30)$$

where $1 \leq i_1 \leq r$, $i_2 \leq (i_1 - 1)/2$ as i_1 is odds or $i_2 \leq i_1/2 - 1$ as i_1 is even. Here, $q_{i,j}$ are the integer coefficients that appear in the high order derivatives of $f(x|\theta, \sigma, m)$ with respect to m :

$$\frac{\partial^{s+1} f}{\partial m^{s+1}} = \left[\sum_{j=0}^{(s-1)/2} \frac{q_{(s+1),j} m^{s-2j}}{\sigma^{2s+2-2j}} (x-\theta)^{2s-2j+1} \right] f\left(\frac{x-\theta}{\sigma}\right) f\left(\frac{m(x-\theta)}{\sigma}\right)$$

when s is an odd number and

$$\frac{\partial^{s+1} f}{\partial m^{s+1}} = \left[\sum_{j=0}^{s/2} \frac{q_{(s+1),j} m^{s-2j}}{\sigma^{2s+2-2j}} (x-\theta)^{2s-2j+1} \right] f\left(\frac{x-\theta}{\sigma}\right) f\left(\frac{m(x-\theta)}{\sigma}\right)$$

when s is an even number. For instance, when $s = 0$, we have $q_{1,0} = 2$ and when $s = 1$, we have $q_{2,0} = -2$.

Summarizing, in order for all the coefficients in the r -minimal form (4.29) to vanish, i.e we have $\beta_{ji}^{(r)}, \gamma_l^{(r)}/W_r^r(G, G_0) \rightarrow 0$, the system of limits (4.30) is the key

factor to determine the singularity structure of $G_0 \in \mathcal{S}_3$. We are going to explore the structure of this system of limits under the specific settings of $G_0 \in \mathcal{S}_3$.

4.7.1.1 Singularity level of $G_0 \in \mathcal{S}_{31}$

Recall from the above argument that, as we have $\beta_{ji}^{(r)}/W_r^r(G, G_0)$ when $1 \leq i, l \leq k_0$ and $1 \leq j \leq 2r + 1$, we obtain $\Delta\theta_i, \Delta v_i, \Delta p_i/W_r^r(G, G_0) \rightarrow 0$ for all $1 \leq i \leq k_0$. Combining with Lemma 4.3.1, it follows that

$$\sum_{i=1}^{k_0} p_i |\Delta m_i|^r / W_r^r(G_1, G) \not\rightarrow 0. \quad (4.31)$$

Since we have $\max_{1 \leq i \leq k_0} |\Delta m_i| > 0$, a combination of (4.30) and (4.31) leads to

$$\left(\sum_{i_1-i_2=l/2} \frac{q_{i_1,i_2}}{i_1!} \sum_{i=1}^{k_0} \frac{p_i (m_i^0)^{i_1-2i_2-1} (\Delta m_i)^{i_1}}{(\sigma_i^0)^l} \right) / \sum_{i=1}^{k_0} p_i |\Delta m_i|^r \rightarrow 0, \quad (4.32)$$

for any even l such that $1 \leq l \leq 2r$. Let $q_i = p_i/\sigma_i^0$, $t_i^0 = m_i^0/\sigma_i^0$, and $\Delta t_i = \Delta m_i/\sigma_i^0$ for all $1 \leq i \leq k_0$, then the above limits can be rewritten as

$$\left(\sum_{i=1}^{k_0} \sum_{i_1-i_2=l/2} \frac{q_{i_1,i_2}}{i_1!} q_i (t_i^0)^{i_1-2i_2-1} (\Delta t_i)^{i_1} \right) / \sum_{i=1}^{k_0} q_i |\Delta t_i|^r \rightarrow 0, \quad (4.33)$$

where in the summation of the above display, $1 \leq i_1 \leq r$, $i_2 \leq (i_1 - 1)/2$ as i_1 is odd, or $i_2 \leq i_1/2 - 1$ as i_1 is even and l is an even number ranging from 2 to $2r$. These are the limits of the ratio of two semipolynomial functions. The existence of these limits will be shown to entail the existence of zeros of a system of polynomial equations.

Greedy extraction of limiting polynomials As explained in the main text, it is generally difficult to obtain all polynomial limits of the system of rational semipolynomial functions given by (4.33). However, it is possible to obtain a subset of polynomial limits via a greedy method of extraction. We shall demonstrate this technique for the

specific case $r = 3$, and then present a general result, not unlike what we have done in subsections 4.4.1 and 4.4.2 for o-mixtures. For $r = 3$, we only have three possible choices of l in (4.33), which are $l = 2, 4$ and 6 . As $l = 2$, we have $(i_1, i_2) = (1, 0)$. As $l = 4$, we obtain $(i_1, i_2) \in \{(2, 0), (3, 1)\}$. Finally, as $l = 6$, we get $(i_1, i_2) = (3, 0)$. Here, we can compute that $q_{1,0} = 2, q_{2,0} = -2, q_{3,1} = -2, q_{3,0} = 2$. Therefore, as $r = 3$, the system of limits (4.33) becomes

$$\begin{aligned} & \left(\sum_{i=1}^{k_0} q_i \Delta t_i \right) / \sum_{i=1}^{k_0} q_i |\Delta t_i|^3 \rightarrow 0, \\ & \left(\sum_{i=1}^{k_0} q_i t_i^0 (\Delta t_i)^2 + \frac{1}{3} q_i (\Delta t_i)^3 \right) / \sum_{i=1}^{k_0} q_i |\Delta t_i|^3 \rightarrow 0, \\ & \left(\sum_{i=1}^{k_0} q_i (t_i^0)^2 (\Delta t_i)^3 \right) / \sum_{i=1}^{k_0} q_i |\Delta t_i|^3 \rightarrow 0. \end{aligned} \quad (4.34)$$

Denote $|\Delta t_{k_0}| := \max_{1 \leq i \leq k_0} \{|\Delta t_i|\}$. In each of the limiting expressions in the above display, we shall divide both the numerator and denominator of the left hand side by $|\Delta t_{k_0}|^\alpha$, where α is the smallest degree that appears in one of the monomials in the numerator. Since $|\Delta t_i|/|\Delta t_{k_0}|$ is bounded, there exist a subsequence according to which $\Delta t_i/|\Delta t_{k_0}|$ tends to a constant, say k_i , for each $i = 1, \dots, k_0$. Note that at least one of the k_i is non-zero. Moreover, we obtain the following equations in the limit

$$\sum_{i=1}^{k_0} q_i^0 k_i = 0, \quad \sum_{i=1}^{k_0} q_i^0 t_i^0 (k_i)^2 = 0, \quad \sum_{i=1}^{k_0} q_i^0 (t_i^0)^2 (k_i)^3 = 0.$$

Since $q_i^0 = p_i^0/\sigma_i^0$, $t_i^0 = m_i^0/\sigma_i^0$ for all $1 \leq i \leq k_0$, by rescaling k_i , the above system of polynomial equations can be rewritten as

$$\sum_{i=1}^{k_0} p_i^0 k_i = 0, \quad \sum_{i=1}^{k_0} p_i^0 m_i^0 (k_i)^2 = 0, \quad \sum_{i=1}^{k_0} p_i^0 (m_i^0)^2 (k_i)^3 = 0.$$

Now we shall apply the greedy extraction technique to the general system (4.33).

This involves dividing both the numerator and the denominator of the left hand side in each equation of the system by $(\Delta t_{k_0})^{l/2}$ for any $2 \leq l \leq 2r$ and l is even. This leads to the existence of solution for the following system of polynomial equations

$$\sum_{i=1}^{k_0} p_i^0 (m_i^0)^{l/2-1} k_i^{l/2} = 0, \quad (4.35)$$

where the index l is even and $2 \leq l \leq 2r$. In this system, at least one of k_i is non-zero.

At this point, by a contrapositive argument we immediately deduces that if system of polynomial equations (4.35) does *not* have a valid solution for the k_i , one of which must be non-zero, then G_0 is *not* r -singular relative to \mathcal{E}_{k_0} . It follows that $\ell(G_0 | \mathcal{E}_{k_0}) \leq r-1$. This connection motivates a deeper investigation into the behavior of the system of real polynomial equations (4.35).

Behavior of system of limiting polynomial equations We proceed to study the solvability of the system of polynomial equations like (4.35). Consider two parameter sequences $\mathbf{a} = \{a_i\}_{i=1}^{k_0}$, $\mathbf{b} = \{b_i\}_{i=1}^{k_0}$ such that $a_i > 0, b_i \neq 0$ for all $1 \leq i \leq l$ and b_i are pairwise different. Additionally, there exists two indices $1 \leq i_1 \neq j_1 \leq l$ such that $b_{i_1} b_{j_1} < 0$. We can think of a_i as taking the role of p_i^0 and b_i the role of m_i^0 .

Define $\bar{s}(k_0, \mathbf{a}, \mathbf{b})$ to be the *minimum* value of $s \geq 1$ such that the following system of polynomial equations

$$\sum_{i=1}^{k_0} a_i b_i^u c_i^{u+1} = 0, \text{ for } u = 0, 1, \dots, s \quad (4.36)$$

does not admit any *non-trivial* solution, by which we require that at least one of c_i is non-zero. For example, if $s = 2$, and $k_0 = 2$, the above system of polynomial equations is

$$a_1 c_1 + a_2 c_2 = 0, \quad a_1 b_1 c_1^2 + a_2 b_2 c_2^2 = 0, \quad a_1 b_1^2 c_1^3 + a_2 b_2^2 c_2^3 = 0.$$

In general, it is difficult to determine the exact value of $\bar{s}(k_0, \mathbf{a}, \mathbf{b})$ since it depends on the specific values of parameter sequences \mathbf{a} and \mathbf{b} . However, it is possible to obtain some nontrivial bounds:

Proposition 4.7.1. *Let $k_0 \geq 2$.*

- (a) *If for any subset I of $\{1, 2, \dots, k_0\}$ we have $\sum_{i \in I} a_i \prod_{j \in I \setminus \{i\}} b_j \neq 0$, then $\bar{s}(k_0, \mathbf{a}, \mathbf{b}) \leq k_0 - 1$.*
- (b) *If there is a subset I of $\{1, 2, \dots, k_0\}$ such that $\sum_{i \in I} a_i \prod_{j \in I \setminus \{i\}} b_j = 0$, then $\bar{s}(k_0, \mathbf{a}, \mathbf{b}) = \infty$.*
- (c) *Under the same condition as that of part (a):*

If $k_0 = 2$, then $\bar{s}(k_0, \mathbf{a}, \mathbf{b}) = 1$.

If $k_0 = 3$, and $\sum_{i=1}^{k_0} a_i \prod_{j \neq i, j \leq k_0} b_j > 0$, then $\bar{s}(k_0, \mathbf{a}, \mathbf{b}) = 1$. Otherwise, $\bar{s}(k_0, \mathbf{a}, \mathbf{b}) = 2$.

Remarks (i) Applying part (a) of this proposition to system (4.35), since $G_0 \in \mathcal{S}_{31}$, i.e $P_3(\mathbf{p}^0, \boldsymbol{\eta}^0) = \sum_{i=1}^{k_0} p_i^0 \prod_{j \neq i} m_j^0 \neq 0$, G_0 is not $\bar{s}(k_0, \{p_i^0\}_{i=1}^{k_0}, \{m_i^0\}_{i=1}^{k_0}) + 1$ -singular relative to \mathcal{E}_{k_0} . Therefore, the singularity level of G_0 is at most $\bar{s}(k_0, \{p_i^0\}_{i=1}^{k_0}, \{m_i^0\}_{i=1}^{k_0})$.
(ii) Part (a) provides a mild condition of parameter sequences \mathbf{a}, \mathbf{b} under which a nontrivial finite upper bound can be obtained. A closer investigation of the proof establishes that this bound is tight, i.e., there exists (\mathbf{a}, \mathbf{b}) such that $\bar{s}(k_0, \mathbf{a}, \mathbf{b}) = k_0 - 1$ holds. This motivates the definition of \mathcal{S}_{31} . (iii) Part (b) suggests the possibility of infinite level of singularity, even as k_0 is fixed. We will show that this happens when $G_0 \in \mathcal{S}_{33}$. (iv) Part (c) suggests that the singularity levels of G_0 may be different for different values of $(\mathbf{p}^0, \boldsymbol{\eta}^0)$ for the same k_0 .

General bounds for singularity level of $G_0 \in \mathcal{S}_{31}$ So far, we assume that G_0 has exactly one homologous set without C(1) singularity of size k_0 . Now, we

suppose that G_0 has more than one nonconformant homologous set without C(1) singularity of components, and that there are no Gaussian components (i.e., $P_1(\boldsymbol{\eta}^0) = \prod_{j=1}^{k_0} m_j^0 \neq 0$). It can be observed that the singularity level of G_0 can be bounded in terms of a number of system of polynomial equations of the same form as Eq. (4.35), which are applied to *disjoint* subsets of nonconformant homologous components. The application to each subset yields a corresponding system of polynomial limits like (4.33). If none of such systems admit non-trivial solutions, then we are absolutely certain that their corresponding systems of limiting equations cannot hold. As a consequence, we obtain that $\ell(G_0|\mathcal{E}_{k_0}) \leq \bar{s}(G_0)$, where

$$\bar{s}(G_0) := \max_I \bar{s}(|I|, \{p_i^0\}_{i \in I}, \{m_i^0\}_{i \in I}), \quad (4.37)$$

where the maximum is taken over all nonconformant homologous subsets I of components of G_0 .

If, on the other hand, G_0 has one or more Gaussian components, in addition to having some nonconformant homologous subsets, then by combining the argument presented in Section 4.5.1 with the foregoing argument, we deduce that the singularity level of G_0 is at most $\max\{2, \bar{s}(G_0)\}$. Summarizing, we have the following theorem regarding the upper bound of singularity levels of $G_0 \in \mathcal{S}_{31}$ whose rigorous proof is deferred to Appendix B.

Theorem 4.7.1. *Suppose that $G_0 \in \mathcal{S}_{31}$.*

- (a) *If $P_1(\boldsymbol{\eta}^0) \neq 0$, then $\ell(G_0|\mathcal{E}_{k_0}) \leq \bar{s}(G_0) \leq k^* - 1 \leq k_0 - 1$.*
- (b) *If $P_1(\boldsymbol{\eta}^0) = 0$, then $\ell(G_0|\mathcal{E}_{k_0}) \leq \max\{2, \bar{s}(G_0)\} \leq \max\{2, k^* - 1\} \leq \max\{2, k_0 - 1\}$.*

where k^ is the maximum length among all nonconformant homologous sets without C(1) singularity of G_0 .*

Exact calculations in special cases Since our proof method was to extract only an (incomplete) subset of polynomial limits, we could only speak of upper bounds of the singularity level, not lower bounds in general. For some special cases of $G_0 \in \mathcal{S}_{31}$, with extra work we can determine the exact singularity level of G_0 . This is based on the specific value of k^* , which is defined to be the maximum length among all nonconformant homologous sets without C(1) singularity of G_0 in Theorem 4.7.1:

Proposition 4.7.2. (Exact singularity level) *Assume that $G_0 \in \mathcal{S}_{31}$ and $P_1(\boldsymbol{\eta}^0) \neq 0$.*

(a) *If $k^* = 2$, then $\ell(G_0|\mathcal{E}_{k_0}) = 1$.*

(b) *Let $k^* = 3$. In addition, if all homologous sets I of G_0 such that $|I| = k^*$ satisfy*

$$\sum_{i \in I} p_i^0 \prod_{j \in I \setminus \{i\}} m_j^0 > 0, \text{ then } \ell(G_0|\mathcal{E}_{k_0}) = 1. \text{ Otherwise, } \ell(G_0|\mathcal{E}_{k_0}) = 2.$$

4.7.1.2 Singularity structure of \mathcal{S}_{32}

For the simplicity of the argument in this section, we go back to the simple setting of G_0 , i.e., G_0 has only one homologous set of size k_0 . Since $G_0 \in \mathcal{S}_{32}$, we have $P_3(\mathbf{p}^0, \boldsymbol{\eta}^0) = \sum_{i=1}^{k_0} p_i^0 \prod_{j \neq i} m_j^0 = 0$. This entails that $\bar{s}(k_0, \{p_i^0\}, \{m_i^0\}) = \infty$ according to part (b) of Proposition 4.7.1. As a result, $\bar{s}(G_0) = \infty$, i.e., the upper bound given by Theorem 4.7.1, that is, $\ell(G_0|\mathcal{E}_{k_0}) \leq \bar{s}(G_0)$, is no longer meaningful for \mathcal{S}_{32} . This does not necessarily imply that the singularity level for $G_0 \in \mathcal{S}_{32}$ is infinite. It simply means that the system of polynomial equations in (4.35) will not lead to any contradiction for any order r . In fact, these equations described by (4.35) are no longer sufficient to express the polynomial limits of the system (4.32). The issue is that our greedy extraction of polynomial limits for the system (4.32) treats each equation of the system separately. For instance, in system (4.34), a special case of system (4.32) when $r = 3$, we do not consider the interaction between two summations $\sum_{i=1}^{k_0} q_i t_i^0 (\Delta t_i)^2$ and $\sum_{i=1}^{k_0} \frac{1}{3} q_i (\Delta t_i)^3$ in the numerator of the second limit. As a result, the

limiting polynomials obtained are dependent only on the lowest order monomial terms that appear in the numerator of each of the r -minimal form's coefficients.

To go further with \mathcal{S}_{32} , we introduce a more sophisticated technique for the polynomial limit extraction, which seeks to partially account for the interactions among different summations in the numerators of all the limits in system (4.32). This can be achieved by keeping not only the lowest order monomial in the numerator of the r -form's coefficient, but also the second lowest order monomials. As a result, we can extract a larger set of polynomial limits than (4.35). This would allow us to obtain a tighter bound of the singularity level for elements of \mathcal{S}_{32} . Although our extraction technique is general, the system of limiting polynomials that can be extracted is difficult to express explicitly for large values of k_0 . For this reason in the following we shall illustrate this technique of polynomial limit extraction on a specific case of $k_0 = 2$.

Proposition 4.7.3. *Assume that $G_0 \in \mathcal{S}_{32}$ and G_0 has only one homologous set of size k_0 . Then as $k_0 = 2$, we have $\ell(G_0|\mathcal{E}_{k_0}) = 3$.*

Remark: (i) The assumption that G_0 has only one homologous set is just for the convenience of the argument. The conclusion of this proposition still holds when $G_0 \in \mathcal{S}_{32}$ has multiple homologous sets and the maximum length of homologous sets with C(1) singularity is 2. (ii) By using the same technique, we can demonstrate that $\ell(G_0|\mathcal{E}_{k_0}) = k_0 + 1$ when $k_0 \leq 5$ and $G_0 \in \mathcal{S}_{32}$ has only one homologous set of size k_0 . We conjecture that this result also holds for general k_0 .

Proof. The proof proceeds in two main steps

Step 1: We will demonstrate that G_0 is 3-singular relative to \mathcal{E}_{k_0} . As $r = 3$, the system (4.32) consists of the following limiting equations, as $q_i \rightarrow q_i^0 > 0$ and $\Delta t_i \rightarrow 0$

for all $i = 1, 2$,

$$\begin{aligned} \sum_{i=1}^2 q_i \Delta t_i / \sum_{i=1}^2 q_i |\Delta t_i|^3 &\rightarrow 0, \\ \left(\sum_{i=1}^2 q_i t_i^0 (\Delta t_i)^2 + \frac{1}{3} q_i (\Delta t_i)^3 \right) / \sum_{i=1}^2 q_i |\Delta t_i|^3 &\rightarrow 0, \\ \left(\sum_{i=1}^2 q_i (t_i^0)^2 (\Delta t_i)^3 \right) / \sum_{i=1}^2 q_i |\Delta t_i|^3 &\rightarrow 0, \end{aligned}$$

where $q_i = p_i/\sigma_i^0$, $q_i^0 = p_i^0/\sigma_i^0$, $t_i^0 = m_i^0/\sigma_i^0$, and $\Delta t_i = \Delta m_i/\sigma_i^0$ for all $i = 1, 2$. The condition of C(1) singularity means $P_3(\mathbf{p}^0, \boldsymbol{\eta}^0) = 0$. That is $p_1^0 m_2^0 + p_2^0 m_1^0 = 0$. So, $q_1^0 t_2^0 + q_2^0 t_1^0 = 0$. By choosing $\Delta t_2 = 1/n$, $\Delta t_1 = \frac{1}{n} \left(-\frac{q_2}{q_1} + \frac{1}{n^4} \right)$ where $q_1 = q_1^0 + 1/n$ and $q_2 = -q_1 t_2^0 / t_1^0 + 1/n^2$, we can check that all of the above limits are satisfied. Hence, G_0 is 3-singular relative to \mathcal{E}_{k_0} .

Step 2: It remains to show that G_0 is *not* 4-singular relative to \mathcal{E}_{k_0} , and hence, G_0 's singularity level is 3. Let $r = 4$, the system (4.32) consists of the following limiting equations

$$\begin{aligned} \sum_{i=1}^2 q_i^n \Delta t_i^n / \sum_{i=1}^2 q_i^n |\Delta t_i^n|^4 &\rightarrow 0, \\ \left(\sum_{i=1}^2 q_i t_i^0 (\Delta t_i)^2 + \frac{1}{3} q_i (\Delta t_i)^3 \right) / \sum_{i=1}^2 q_i |\Delta t_i|^4 &\rightarrow 0, \\ \left(\sum_{i=1}^2 \frac{1}{3} q_i (t_i^0)^2 (\Delta t_i)^3 + \frac{1}{4} q_i t_i^0 (\Delta t_i)^4 \right) / \sum_{i=1}^2 q_i |\Delta t_i|^4 &\rightarrow 0, \\ \sum_{i=1}^2 q_i (t_i^0)^3 (\Delta t_i)^4 / \sum_{i=1}^2 q_i |\Delta t_i|^4 &\rightarrow 0. \end{aligned}$$

In order to account for the second-lowest order monomials of the numerator in each of the equations, we raise the order of the denominator in each equation to the former.

That is,

$$\begin{aligned}
K_1 &:= \sum_{i=1}^2 q_i \Delta t_i / \sum_{i=1}^2 q_i |\Delta t_i|^2 \rightarrow 0, \\
K_2 &:= \left(\sum_{i=1}^2 q_i t_i^0 (\Delta t_i)^2 + \frac{1}{3} q_i (\Delta t_i^n)^3 \right) / \sum_{i=1}^2 q_i |\Delta t_i|^3 \rightarrow 0, \\
K_3 &:= \left(\sum_{i=1}^2 \frac{1}{3} q_i (t_i^0)^2 (\Delta t_i)^3 + \frac{1}{4} q_i t_i^0 (\Delta t_i)^4 \right) / \sum_{i=1}^2 q_i |\Delta t_i|^4 \rightarrow 0, \\
K_4 &:= \sum_{i=1}^2 q_i (t_i^0)^3 (\Delta t_i)^4 / \sum_{i=1}^2 q_i |\Delta t_i|^4 \rightarrow 0.
\end{aligned}$$

We assume without loss of generality that $|\Delta t_2|$ is the maximum between $|\Delta t_1|$ and $|\Delta t_2|$. Denote $\Delta t_1 = k_1 \Delta t_2$ where $k_1 \in [-1, 1]$ and $k_1 \rightarrow k'_1$. The vanishing of K_1 yields $q_1^0 k'_1 + q_2^0 = 0$. So, $k'_1 = -q_2^0/q_1^0 = t_2^0/t_1^0$.

Divide both the numerator and denominator of K_1 by $(\Delta t_2)^2$, we obtain $(q_1 k_1 + q_2)/\Delta t_2 \rightarrow 0$. Write $u = k_1 + q_2/q_1$, then $q_1 u / \Delta t_2 \rightarrow 0$, which implies that $u/\Delta t_2 \rightarrow 0$.

Next, divide both the numerator and denominator of K_2 by $(\Delta t_2)^3$, we obtain

$$\left(\sum_{i=1}^2 q_i t_i^0 (\Delta t_i)^2 + \frac{1}{3} q_i (\Delta t_i)^3 \right) / (\Delta t_2)^3 \rightarrow 0.$$

Plug in the formula of k_1 and the fact that $u/\Delta t_2 \rightarrow 0$, it follows that

$$\left(q_1 t_1^0 \left(\frac{q_2}{q_1} \right)^2 + q_2 t_2^0 \right) / (\Delta t_2) \rightarrow -\frac{1}{3} (q_1^0 (k'_1)^3 + q_2^0).$$

Thus, we get $P_1 := (t_1^0 q_2 + t_2^0 q_1) / \Delta t_2 \rightarrow -\frac{q_1^0}{3q_2^0} (q_1^0 (k'_1)^3 + q_2^0)$. It is simple to verify that this limit is non-zero, otherwise we would have $q_1^0 = q_2^0$, which violates the definition that G_0 does not have C(2) singularity, i.e $G_0 \in \mathcal{S}_{32}$.

Continuing, divide both the numerator and denominator of K_3 by $(\Delta t_2)^4$, and with the same argument, we obtain $P_2 := (t_1^0 q_2 - t_2^0 q_1)(t_1^0 q_2 + t_2^0 q_1) / \Delta t_2 \rightarrow -\frac{3(q_1^0)^2}{4q_2^0} (q_1^0 t_1^0 (k'_1)^4 +$

$$q_2^0 t_2^0).$$

By dividing P_2 by P_1 and let it to vanish, we can extract the following polynomial in the limit:

$$4(q_1^0(k'_1)^3 + q_2^0)(t_1^0 q_2^0 - t_2^0 q_1^0) = 9q_1^0(q_1^0 t_1^0(k'_1)^4 + q_2^0 t_2^0).$$

By plugging in $k'_1 = -q_2^0/q_1^0$ and $t_1^0 q_2^0 + t_2^0 q_1^0 = 0$, we can deduce that $q_1^0 = q_2^0$, which is a contradiction. Thus, we conclude that G_0 is not 4-singular relative to \mathcal{E}_{k_0} . \square

4.7.1.3 Singularity level of $G_0 \in \mathcal{S}_{33}$

As we can see from the proof of Proposition 4.7.2, the condition of without C(2) singularity plays a major role in guaranteeing that $G_0 \in \mathcal{S}_{32}$ is not 4-singular relative to \mathcal{E}_{k_0} when G_0 has only one homologous set of $k_0 = 2$. Therefore, for elements G_0 in \mathcal{S}_{33} , we expect the singularity level of G_0 may be very large. In fact, we can show that

Theorem 4.7.2. *If $G_0 \in \mathcal{S}_{33}$, then $\ell(G_0 | \mathcal{E}_{k_0}) = \infty$.*

Proof. Here, we present the proof for $k_0 = 2$. For general values of k_0 , the proof is similar and deferred to Appendix B. For $k_0 = 2$, the condition that $G_0 \in \mathcal{S}_{33}$ entails $P_4(\mathbf{p}^0, \boldsymbol{\eta}^0) = 0$, i.e $p_1^0/\sigma_1^0 = p_2^0/\sigma_2^0$ and $m_1^0/\sigma_1^0 = -m_2^0/\sigma_2^0$. By choosing $\Delta m_1/\sigma_1^0 = -\Delta m_2/\sigma_2^0$, $p_1 = p_2 = p_1^0 = p_2^0$, we can check that

$$\sum_{i=1}^2 \frac{p_i(m_i^0)^u (\Delta m_j)^v}{(\sigma_i^0)^{u+v+1}} = 0,$$

for all odd numbers $u \in [1, v]$ when v is even number, or for all even numbers $u \in [0, v]$ when v is odd number.

Take order $r \geq 1$ to be an arbitrary natural number. Incorporating the identity in the previous display into (4.29) and (4.30), we obtain the vanishing of all $\gamma_l^{(r)}/W_r^r(G_1, G)$ for all $1 \leq l \leq 2r$ and l is even. If we choose $\Delta\theta_i = \Delta v_i = 0$ for all

$1 \leq i \leq 2$, we also have the coefficients $\beta_{ji}^{(r)}/W_r^r(G, G_0) = 0$ for all $1 \leq i \leq 2$ and $1 \leq j \leq 2r + 1$. Additionally, we also have $\gamma_l^{(r)}/W_r^r(G, G_0) = 0$ for all $1 \leq l \leq 2r$ and l is odd. Hence, G_0 is r -singular relative to \mathcal{E}_{k_0} for any $r \geq 1$. As a consequence, $\ell(G_0|\mathcal{E}_{k_0}) = \infty$. \square

4.8 Appendix B

This Appendix contains the remaining proofs of the results presented in this chapter.

4.8.1 Proofs for Section 3

PROOF OF THEOREM 4.3.1 Since the proofs for part (iii) and (iv) are similar, we only provide the proof for part (iii). The proof of this part is the generalization of that of part (c) in Theorem 3.2 in [Ho and Nguyen, 2016c]. By means of Taylor expansion up to r -th order, we have

$$\begin{aligned} h^2(p_G, p_{G_0}) &< \int_{x \in \mathcal{X}} \frac{(p_G(x) - p_{G_0}(x))^2}{p_{G_0}(x)} dx = \int_{x \in \mathcal{X}} \frac{\left(\sum_{l=1}^{T_r} \xi_l^{(r)}(G) H_l^{(r)}(x) + R_r(x) \right)^2}{p_{G_0}(x)} dx \\ &= \int_{x \in \mathcal{X}} \frac{R_r^2(x)}{p_{G_0}(x)} dx. \end{aligned}$$

Here, $R_r(x)$ has the following form

$$R_r(x) = \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} \sum_{|\alpha|=r+1} \frac{r+1}{\alpha!} (\Delta \eta_{ij})^\alpha \int_0^1 (1-t)^r \frac{\partial^{r+1} f}{\partial \eta^\alpha} (x|\eta_i^0 + t\Delta \eta_{ij}) dt.$$

Hence, as $p_{G_0}(x) > p_i^0 f(x|\eta_i^0)$ for all $1 \leq i \leq k_0$, for any $s < r + 1$, we have

$$\begin{aligned} \frac{h^2(p_G, p_{G_0})}{W_1^{2s}(G, G_0)} &< \int_{x \in \mathcal{X}} \frac{R_r^2(x)}{W_1^{2s}(G, G_0)p_{G_0}(x)} dx \\ &< \sum_{i=1}^{k_0} \int_{x \in \mathcal{X}} \frac{\left(\sum_{j=1}^{s_i} \sum_{|\alpha|=r+1} \frac{r+1}{\alpha!} (\Delta \eta_{ij})^\alpha \int_0^1 (1-t)^r \frac{\partial^{r+1} f}{\partial \eta^\alpha}(x|\eta_i^0 + t\Delta \eta_{ij}) dt \right)^2}{W_1^{2s}(G, G_0)p_i^0 f(x|\eta_i^0)} dx \\ &\lesssim \sum_{i=1}^{k_0} \int_{x \in \mathcal{X}} \frac{\sum_{j=1}^{s_i} \sum_{|\alpha|=r+1} \left(\frac{r+1}{\alpha!} (\Delta \eta_{ij})^\alpha \int_0^1 (1-t)^r \frac{\partial^{r+1} f}{\partial \eta^\alpha}(x|\eta_i^0 + t\Delta \eta_{ij}) dt \right)^2}{W_1^{2s}(G, G_0)p_i^0 f(x|\eta_i^0)} dx, \end{aligned}$$

where the last inequality comes from Cauchy-Schwarz's inequality. Now, for any $s < r + 1$, by utilizing Lemma 4.3.1, we obtain

$$\frac{|(\Delta \eta_{ij})^\alpha|}{W_1^s(G, G_0)} \asymp \frac{|(\Delta \eta_{ij})^\alpha|}{D_1^s(G_0, G)} < \frac{|(\Delta \eta_{ij})^\alpha|}{\|\Delta \eta_{ij}\|^s} \rightarrow 0, \quad (4.38)$$

for any $|\alpha| = r + 1$. According to the hypothesis, as $\Delta \eta_{ij} < c_0$, we have

$$\begin{aligned} \int_{x \in \mathcal{X}} \frac{\left(\int_0^1 (1-t)^r \frac{\partial^{r+1} f}{\partial \eta^\alpha}(x|\eta_i^0 + t\Delta \eta_{ij}) dt \right)^2}{p_i^0 f(x|\eta_i^0)} dx &< \int_{x \in \mathcal{X}} \frac{\left(\frac{\partial^{r+1} f}{\partial \eta^\alpha}(x|\eta_i^0 + t\Delta \eta_{ij}) \right)^2}{p_i^0 f(x|\eta_i^0)} dx \\ &< \infty. \end{aligned} \quad (4.39)$$

Combining (4.38) and (4.39), we achieve $h(p_G, p_{G_0})/W_1^s(G, G_0) \rightarrow 0$, which yields the conclusion of this part.

4.8.2 Proofs for Section 4

PROOF OF LEMMA 4.4.1 For any $k_0 \geq 1$ and k_0 different pairs $\eta_1 = (\theta_1, \sigma_1, m_1), \dots, \eta_{k_0} = (\theta_{k_0}, \sigma_{k_0}, m_{k_0})$, let $\alpha_{ij} \in \mathbb{R}$ for $i = 1, \dots, 4$, $j = 1, \dots, k_0$ such that for al-

most all $x \in \mathbb{R}$

$$\sum_{j=1}^{k_0} \alpha_{1j} f(x|\eta_j) + \alpha_{2j} \frac{\partial f}{\partial \theta}(x|\eta_j) + \alpha_{3j} \frac{\partial f}{\partial \sigma^2}(x|\eta_j) \alpha_{4j} \frac{\partial f}{\partial m}(x|\eta_j) = 0.$$

We can rewrite the above equation as

$$\sum_{j=1}^{k_0} \left\{ [\beta_{1j} + \beta_{2j}(x - \theta_j) + \beta_{3j}(x - \theta_j)^2] \Phi \left(\frac{m_j(x - \theta_j)}{\sigma_j} \right) \exp \left(-\frac{(x - \theta_j)^2}{2\sigma_j^2} \right) + (\gamma_{1j} + \gamma_{2j}(x - \theta_j)) f \left(\frac{m_j(x - \theta_j)}{\sigma_j} \right) \exp \left(-\frac{(x - \theta_j)^2}{2\sigma_j^2} \right) \right\} = 0, \quad (4.40)$$

where $\beta_{1j} = \frac{2\alpha_{1j}}{\sqrt{2\pi}\sigma_j} - \frac{\alpha_{3j}}{\sqrt{2\pi}\sigma_j^3}$, $\beta_{2j} = \frac{2\alpha_{2j}}{\sqrt{2\pi}\sigma_j^3}$, $\beta_{3j} = \frac{\alpha_{3j}}{\sqrt{2\pi}\sigma_j^5}$, $\gamma_{1j} = -\frac{2\alpha_{2j}m_j}{\sqrt{2\pi}\sigma_j^2}$, and $\gamma_{2j} = -\frac{\alpha_{3j}m_j}{\sqrt{2\pi}\sigma_j^4} + \frac{2\alpha_{4j}}{\sqrt{2\pi}\sigma_j^2}$ for all $j = 1, \dots, k_0$.

"Only if" direction: Assume by contrary that the conclusion does not hold, i.e., both type A and type B conditions do not hold. Denote $\sigma_{j+k_0} = \frac{\sigma_j^2}{1+m_j^2}$ for all $1 \leq j \leq k_0$. For the simplicity of the argument, we assume that σ_i are pairwise different and $\frac{\sigma_i^2}{1+m_i^2} \notin \{\sigma_j^2 : 1 \leq j \leq k_0\}$ for all $1 \leq i \leq k_0$. The argument for the other cases is similar. Now, σ_j are pairwise different as $1 \leq j \leq 2k_0$. The equation (4.40) can be rewritten as

$$\sum_{j=1}^{2k_0} \left\{ [\beta_{1j} + \beta_{2j}(x - \theta_j) + \beta_{3j}(x - \theta_j)^2] \times \Phi \left(\frac{m_j(x - \theta_j)}{\sigma_j} \right) \exp \left(-\frac{(x - \theta_j)^2}{2\sigma_j^2} \right) \right\} = 0, \quad (4.41)$$

where $m_j = 0$, $\theta_{j+k_0} = \theta_j$, $\beta_{1(j+k_0)} = \frac{2\gamma_{1j}}{\sqrt{2\pi}}$, $\beta_{2(j+k_0)} = \frac{2\gamma_{2j}}{\sqrt{2\pi}}$, $\beta_{3j} = 0$ as $k_0 + 1 \leq j \leq 2k_0$. Denote $\bar{i} = \arg \max_{1 \leq i \leq 2k_0} \{\sigma_i\}$. Multiply both sides of (4.41) with the term $\exp \left(\frac{(x - \theta_{\bar{i}})^2}{2\sigma_{\bar{i}}^2} \right) / \Phi \left(\frac{m_{\bar{i}}(x - \theta_{\bar{i}})}{\sigma_{\bar{i}}} \right)$ and let $x \rightarrow +\infty$ if $m_{\bar{i}} \geq 0$ or let $x \rightarrow -\infty$ if $m_{\bar{i}} < 0$ on both sides of the new equation, we obtain $\beta_{1\bar{i}} + \beta_{2\bar{i}}(x - \theta_{\bar{i}}) + \beta_{3\bar{i}}(x - \theta_{\bar{i}})^2 \rightarrow 0$.

It implies that $\beta_{1\bar{i}} = \beta_{2\bar{i}} = \beta_{3\bar{i}} = 0$. Repeatedly apply the same argument to the remaining σ_i until we obtain $\beta_{1i} = \beta_{2i} = \beta_{3i} = 0$ for all $1 \leq i \leq 2k_0$. It is equivalent to $\alpha_{1i} = \alpha_{2i} = \alpha_{3i} = \alpha_{4i} = 0$ for all $1 \leq i \leq k_0$, which is a contradiction.

"If" direction: There are two possible scenarios.

Type A singularity There exists some $m_j = 0$ as $1 \leq j \leq k_0$. In this case, we assume that $m_1 = 0$. If we choose $\alpha_{1j} = \alpha_{2j} = \alpha_{3j} = \alpha_{4j} = 0$ for all $2 \leq j \leq k_0$, then equation (4.40) can be rewritten as

$$\frac{\beta_{11}}{2} + \frac{\gamma_{11}}{\sqrt{2\pi}} + \left(\frac{\beta_{21}}{2} + \frac{\gamma_{21}}{\sqrt{2\pi}} \right) (x - \theta_1) + \frac{\beta_{31}}{2} (x - \theta_1)^2 = 0.$$

By choosing $\alpha_{31} = 0$, $\alpha_{11} = \frac{\alpha_{21}m_1}{\sqrt{2\pi}\sigma_1}$, $\alpha_{21} = -\frac{\alpha_{41}\sigma_1}{\sqrt{2\pi}}$, the above equation always equal to 0. Since $\alpha_{11}, \alpha_{21}, \alpha_{41}$ are not necessarily zero, the first-order identifiability (i.e., linear independence condition) is violated.

Type B singularity There exists indices $1 \leq i \neq j \leq k_0$ such that $\left(\frac{\sigma_i^2}{1+m_i^2}, \theta_i \right) = \left(\frac{\sigma_j^2}{1+m_j^2}, \theta_j \right)$. Without loss of generality, we assume that $i = 1, j = 2$. If we choose $\alpha_{1j} = \alpha_{2j} = \alpha_{3j} = \alpha_{4j} = 0$ for all $3 \leq j \leq k_0$, then equation in (4.40) can be rewritten as

$$\sum_{j=1}^2 \left\{ [\beta_{1j} + \beta_{2j}(x - \theta_j) + \beta_{3j}(x - \theta_j)^2] \Phi \left(\frac{m_j(x - \theta_j)}{\sigma_j} \right) \exp \left(-\frac{(x - \theta_j)^2}{2\sigma_j^2} \right) \right\} + \frac{1}{\sqrt{2\pi}} \left(\sum_{j=1}^2 \gamma_{1j} + \sum_{j=1}^2 \gamma_{2j}(x - \theta_1) \right) \exp \left(-\frac{(m_1^2 + 1)(x - \theta_1)^2}{2\sigma_1^2} \right) = 0.$$

Now, we choose $\alpha_{1j} = \alpha_{2j} = \alpha_{3j} = 0$ for all $1 \leq j \leq 2$, $\frac{\alpha_{41}}{\sigma_1^2} + \frac{\alpha_{42}}{\sigma_2^2} = 0$ then the above equation always hold. Since α_{41} and α_{42} need not be zero, the first-order identifiability condition is violated. This concludes the proof.

PROOF OF LEMMA 4.4.2 The proof proceeds via induction on $|\alpha|$. As $|\alpha| \leq 2$, we can easily check the conclusion of the lemma. Assume that the conclusion holds for any $|\alpha| \leq k - 1$. We shall demonstrate that it also holds for $|\alpha| = k$. Indeed, there are two settings:

Case 1: $\alpha_1 = k$ Under this setting, $\alpha_2 = \alpha_3 = 0$. From the induction hypothesis,

$$\begin{aligned}
\frac{\partial^{|\alpha|} f}{\partial \theta^{\alpha_1} \partial v^{\alpha_2} \partial m^{\alpha_3}} &= \frac{\partial}{\partial \theta} \left(\frac{\partial^{|\alpha|-1} f}{\partial \theta^{\alpha_1-1} \partial v^{\alpha_2} \partial m^{\alpha_3}} \right) \\
&= \frac{\partial}{\partial \theta} \left(\sum_{\kappa \in \mathcal{F}_{|\alpha|-1}} \frac{P_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2} \partial m^{\kappa_3}} \right) \\
&= \sum_{\kappa \in \mathcal{F}_{|\alpha|-1}} \frac{P_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|+1} f}{\partial \theta^{\kappa_1+1} \partial v^{\kappa_2} \partial m^{\kappa_3}}, \\
&= \sum_{\kappa \in \mathcal{F}_{k-1}: \kappa_1=0} \frac{P_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|+1} f}{\partial \theta^{\kappa_1+1} \partial v^{\kappa_2} \partial m^{\kappa_3}} \\
&\quad + \sum_{\kappa \in \mathcal{F}_{k-1}: \kappa_1=1} \frac{P_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1-1, \alpha_2, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|+1} f}{\partial \theta^{\kappa_1+1} \partial v^{\kappa_2} \partial m^{\kappa_3}} \tag{4.42}
\end{aligned}$$

where the second equality is due to the application of the hypothesis for $\alpha_1 - 1 + \alpha_2 + \alpha_3 = k - 1$. For any $\kappa \in \mathcal{F}_{k-1}$ such that $\kappa_1 = 1$,

$$\begin{aligned}
\frac{\partial^{|\kappa|+1} f}{\partial \theta^{\kappa_1+1} \partial v^{\kappa_2} \partial m^{\kappa_3}} &= \frac{\partial^{|\kappa|-1} f}{\partial v^{\kappa_2} \partial m^{\kappa_3}} \left(2 \frac{\partial f}{\partial v} - \frac{m^3 + m}{v} \frac{\partial f}{\partial m} \right) \\
&= 2 \frac{\partial^{|\kappa|} f}{\partial v^{\kappa_2+1} \partial m^{\kappa_3}} - \frac{\partial^{|\kappa|-1} f}{\partial v^{\kappa_2} \partial m^{\kappa_3}} \left(\frac{m^3 + m}{v} \frac{\partial f}{\partial m} \right). \tag{4.43}
\end{aligned}$$

From the inductive hypothesis, since $|\kappa| = \kappa_2 + \kappa_3 + 1 \leq k - 1$,

$$\frac{\partial^{|\kappa|} f}{\partial v^{\kappa_2+1} \partial m^{\kappa_3}} = \sum_{\kappa' \in \mathcal{F}_{|\kappa|}} \frac{P_{0, \kappa_2+1, \kappa_3}^{\kappa'_1, \kappa'_2, \kappa'_3}(m)}{H_{0, \kappa_2+1, \kappa_3}^{\kappa'_1, \kappa'_2, \kappa'_3}(m) Q_{0, \kappa_2+1, \kappa_3}^{\kappa'_1, \kappa'_2, \kappa'_3}(v)} \frac{\partial^{|\kappa'|} f}{\partial \theta^{\kappa'_1} \partial v^{\kappa'_2} \partial m^{\kappa'_3}}. \tag{4.44}$$

In addition,

$$\frac{\partial^{|\kappa|-1} f}{\partial v^{\kappa_2} \partial m^{\kappa_3}} \left(\frac{m^3 + m}{v} \frac{\partial f}{\partial m} \right) = \sum_{\beta: |\beta| \leq |\kappa|, \beta_1 \leq \kappa_2, \beta_2 \leq \kappa_3+1} \frac{A_{\beta_1, \beta_2}(m)}{B_{\beta_1, \beta_2}(v)} \frac{\partial^{|\beta|} f}{\partial v^{\beta_1} \partial m^{\beta_2}}. \quad (4.45)$$

Since $|\beta| \leq |\kappa| \leq k-1$, from the hypothesis,

$$\frac{\partial^{|\beta|} f}{\partial v^{\beta_1} \partial m^{\beta_2}} = \sum_{\kappa'' \in \mathcal{F}_{|\beta|}} \frac{P_{0, \beta_1, \beta_2}^{\kappa''_1, \kappa''_2, \kappa''_3}(m)}{H_{0, \beta_1, \beta_2}^{\kappa''_1, \kappa''_2, \kappa''_3}(m) Q_{0, \beta_1, \beta_2}^{\kappa''_1, \kappa''_2, \kappa''_3}(v)} \frac{\partial^{|\kappa''|} f}{\partial \theta^{\kappa''_1} \partial v^{\kappa''_2} \partial m^{\kappa''_3}}. \quad (4.46)$$

Combining equations (4.42), (4.43), (4.44), (4.45), and (4.46), we arrive at the conclusion of the lemma.

Case 2: $\alpha_1 \leq k-1$ Under this setting, assume without loss of generality that $\alpha_2 \geq 1$.

$$\begin{aligned} \frac{\partial^{|\alpha|} f}{\partial \theta^{\alpha_1} \partial v^{\alpha_2} \partial m^{\alpha_3}} &= \frac{\partial}{\partial v} \left(\frac{\partial^{|\alpha|-1} f}{\partial \theta^{\alpha_1} \partial v^{\alpha_2-1} \partial m^{\alpha_3}} \right) \\ &= \frac{\partial}{\partial v} \left(\sum_{\kappa \in \mathcal{F}_{|\alpha|-1}} \frac{P_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2} \partial m^{\kappa_3}} \right) \\ &= \sum_{\kappa \in \mathcal{F}_{|\alpha|-1}} \frac{\partial}{\partial v} \left(\frac{P_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \right) \frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2} \partial m^{\kappa_3}} \\ &\quad + \sum_{\kappa \in \mathcal{F}_{|\alpha|-1}} \frac{P_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|+1} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2+1} \partial m^{\kappa_3}}. \end{aligned} \quad (4.47)$$

Denote $A := \sum_{\kappa \in \mathcal{F}_{|\alpha|-1}} \frac{P_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|+1} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2+1} \partial m^{\kappa_3}}$, we further have that

$$\begin{aligned} A &= \sum_{\kappa \in \mathcal{F}_{|\alpha|-1}: \kappa_3=0} \frac{P_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|+1} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2+1} \partial m^{\kappa_3}} \\ &\quad + \sum_{\kappa \in \mathcal{F}_{|\alpha|-1}: \kappa_2=0, \kappa_3 \geq 1} \frac{P_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m)}{H_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(m) Q_{\alpha_1, \alpha_2-1, \alpha_3}^{\kappa_1, \kappa_2, \kappa_3}(v)} \frac{\partial^{|\kappa|+1} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2+1} \partial m^{\kappa_3}}. \end{aligned} \quad (4.48)$$

Since $m \neq 0$, for any $\kappa \in \mathcal{F}_{|\alpha|-1}$ such that $\kappa_2 = 0$ and $\kappa_3 \geq 1$, we have

$$\begin{aligned} \frac{\partial^{|\kappa|+1} f}{\partial \theta^{\kappa_1} \partial v^{\kappa_2+1} \partial m^{\kappa_3}} &= \frac{\partial^{|\kappa|-1} f}{\partial \theta^{\kappa_1} \partial m^{\kappa_3-1}} \left(-\frac{1}{v} \frac{\partial f}{\partial m} - \frac{m^2 + 1}{2mv} \frac{\partial^2 f}{\partial m^2} \right) \\ &= -\frac{1}{v} \frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa_1} \partial m^{\kappa_3}} - \frac{\partial^{|\kappa|-1} f}{\partial \theta^{\kappa_1} \partial m^{\kappa_3-1}} \left(\frac{m^2 + 1}{2mv} \frac{\partial^2 f}{\partial m^2} \right). \end{aligned} \quad (4.49)$$

Since $|\kappa| = \kappa_1 + \kappa_3 \leq k-1$ and $\kappa_1 \leq 1$, we have $(\kappa_1, 0, \kappa_3) \in \mathcal{F}_k$. Additionally, we can represent

$$\frac{\partial^{|\kappa|-1} f}{\partial \theta^{\kappa_1} \partial m^{\kappa_3-1}} \left(\frac{m^2 + 1}{2mv} \frac{\partial^2 f}{\partial m^2} \right) = \sum_{1 \leq \tau \leq \kappa_3+1} \frac{A'_\tau(m)}{B'_\tau(m)C'_\tau(v)} \frac{\partial^{\kappa_1+\tau} f}{\partial \theta^{\kappa_1} \partial m^\tau},$$

where $A'_\tau(m)$, $B'_\tau(m)$, $C'_\tau(v)$ are some polynomials of m and v . Since $\kappa_1 + \tau \leq \kappa_1 + \kappa_3 + 1 \leq k$ and $\kappa_1 \leq 1$, we have $(\kappa_1, 0, \tau) \in \mathcal{F}_k$. Combining these results with equations (4.47), (4.48), and (4.49), we achieve the conclusion of the lemma.

PROOF OF LEMMA 4.4.3 The proof of this lemma proceeds by induction on r . If $r = 1$,

$$\left\{ \frac{\partial^{|\alpha|} f}{\theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}} : (\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_r \right\} = \left\{ \frac{\partial f}{\partial \theta}, \frac{\partial f}{\partial v}, \frac{\partial f}{\partial m} \right\},$$

which are linearly independent with respect to $G_0 \in \mathcal{S}_0$ due to the conclusion of Lemma 4.4.1. Assume that the conclusion of the lemma holds up to r . We will demonstrate that it continues to hold for $r+1$. In fact,

$$\begin{aligned} \left\{ \frac{\partial^{|\alpha|} f}{\theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}} : (\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_{r+1} \right\} &= \left\{ \frac{\partial^{|\alpha|} f}{\theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}} : (\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_r \right\} \cup \\ &\quad \left\{ \frac{\partial^{r+1} f}{\partial \theta \partial v^r}, \frac{\partial^{r+1} f}{\partial v^{r+1}}, \frac{\partial^{r+1} f}{\partial \theta \partial m^r}, \frac{\partial^{r+1} f}{\partial m^{r+1}} \right\}. \end{aligned} \quad (4.50)$$

Assume that there are coefficients $\beta_{\alpha_1, \alpha_2, \alpha_3}^{(i)}$ where $1 \leq i \leq k_0$ and $(\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_{r+1}$ such that for all x

$$\sum_{i=1}^{k_0} \sum_{(\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_{r+1}} \beta_{\alpha_1, \alpha_2, \alpha_3}^{(i)} \frac{\partial^{|\alpha|} f}{\theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}}(x | \eta_i^0) = 0.$$

Using the fact from (4.50), we rewrite the above equation as

$$\begin{aligned} & \sum_{i=1}^{k_0} \sum_{(\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_r} \beta_{\alpha_1, \alpha_2, \alpha_3}^{(i)} \frac{\partial^{|\alpha|} f}{\theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}}(x | \eta_i^0) + \beta_{1,r,0}^{(i)} \frac{\partial^{r+1} f}{\partial \theta \partial v^r}(x | \eta_i^0) + \\ & \beta_{0,r+1,0}^{(i)} \frac{\partial^{r+1} f}{\partial v^{r+1}}((x | \eta_i^0) + \beta_{1,0,r}^{(i)} \frac{\partial^{r+1} f}{\partial \theta \partial m^r}(x | \eta_i^0) + \beta_{0,0,r+1}^{(i)} \frac{\partial^{r+1} f}{\partial m^{r+1}}(x | \eta_i^0)) = 0. \end{aligned} \quad (4.51)$$

Equation (4.51) can be rewritten as

$$\begin{aligned} & \sum_{i=1}^{k_0} \left(\sum_{j=1}^{2r+3} \gamma_{j,i}^{(r+1)} (x - \theta_i^0)^{j-1} \right) f \left(\frac{x - \theta_i^0}{\sigma_i^0} \right) \Phi \left(\frac{m_i^0 (x - \theta_i^0)}{\sigma_i^0} \right) \\ & + \sum_{i=1}^{k_0} \left(\sum_{j=1}^{2r+2} \tau_{j,i}^{(r+1)} (x - \theta_i^0)^{j-1} \right) \exp \left(-\frac{(m_i^0)^2 + 1}{2v_i^0} (x - \theta_i^0)^2 \right) = 0, \end{aligned}$$

where $\gamma_{j,i}^{(r+1)}$ are a combination of $\beta_{\alpha_1, \alpha_2, \alpha_3}^{(i)}$ when $(\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_{r+1}$ and $\alpha_3 = 0$. Additionally, $\tau_{j,i}^{(r+1)}$ are a combination of $\beta_{\alpha_1, \alpha_2, \alpha_3}^{(i)}$ when $(\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_{r+1}$. Due to the fact that there are no type A or type B singularities in $\{\eta_1^0, \dots, \eta_{k_0}^0\}$, by using the same argument as that of the proof of Lemma 4.4.1, we obtain that $\gamma_{j,i}^{(r+1)} = 0$ for all $1 \leq i \leq k_0$, $1 \leq j \leq 2r+3$ and $\tau_{j,i}^{(r+1)} = 0$ for all $1 \leq i \leq k_0$, $1 \leq j \leq 2r+2$. It can be checked that $\gamma_{2r+3,i}^{(r+1)} = 0$ implies $\beta_{0,r+1,0}^{(i)} = 0$ while $\gamma_{2r+2,i}^{(r+1)} = 0$ implies $\beta_{1,r,0}^{(i)} = 0$ for all $1 \leq i \leq k_0$. Similarly, $\tau_{2r+2,i}^{(r+1)} = 0$ implies $\beta_{0,0,r+1}^{(i)} = 0$ while $\tau_{2r+1,i}^{(r+1)} = 0$ implies $\beta_{1,0,r}^{(i)} = 0$ for all $1 \leq i \leq k_0$. As a consequence, Eq. (4.51) is reduced to

$$\sum_{i=1}^{k_0} \sum_{(\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_r} \beta_{\alpha_1, \alpha_2, \alpha_3}^{(i)} \frac{\partial^{|\alpha|} f}{\theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}}(x | \eta_i^0) = 0. \quad (4.52)$$

According to the hypothesis with r , we obtain that $\beta_{\alpha_1, \alpha_2, \alpha_3}^{(i)} = 0$ for all $1 \leq i \leq k_0$, $(\alpha_1, \alpha_2, \alpha_3) \in \mathcal{F}_r$. This concludes our proof.

PROOF OF PROPOSITION 4.4.1 From the formation of system of polynomial equations (4.21), if we choose $\beta_3 = 0$ (i.e., we only reduce to derivatives with respect to the location and scale parameter), then we have $P_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m)/H_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m)Q_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(v) = 2^{\alpha_2}$ when $\alpha_3 = 0$ and $P_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m)/H_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(m)Q_{\alpha_1, \alpha_2, \alpha_3}^{\beta_1, \beta_2, \beta_3}(v) = 0$ as $\alpha_3 \geq 1$ for any v, m and $\alpha_1 + 2\alpha_2 + 2\alpha_3 = \beta_1 + 2\beta_2 + 2\beta_3$. This shows that the system of polynomial equations (4.21) contains the following system of equations

$$\sum_{j=1}^l \sum_{\alpha_1+2\alpha_2=\beta_1+2\beta_2} \frac{2^{\alpha_2} d_j^2 a_j^{\alpha_1} b_j^{\alpha_2}}{\alpha_1! \alpha_2!} = 0, \quad (4.53)$$

where $\beta_1 + 2\beta_2 \leq r$ and $\beta_1 \leq 1$. This is precisely the system of polynomial equations (4.24) if we replace d_j by x_j , a_j by y_j , $2b_j$ by z_j , α_1, α_2 by n_1, n_2 . Now, if we choose $r > \bar{r}(l)$, the system of polynomial equations (4.53) has only trivial solution $a_j = b_j = 0$ for all $1 \leq j \leq l$. Substitute these results back to system of polynomial equations (4.21), we also obtain $c_j = 0$ for all $1 \leq j \leq l$, which is a contradiction. This completes our proof.

PROOF OF PROPOSITION 4.4.3 The proof of part (a) is straightforward from the discussion in Section 4.4.1. For the proof for part (b), we will present an explicit form for the system of polynomial equations to illustrate the variability of $\rho(l)$ and $\bar{\rho}(l)$ based on the values of (m, v) .

(b) As $l = 2$ and $r = 6$, the system of polynomial equations (4.21) can be rewritten

as

$$\begin{aligned}
& \sum_{i=1}^3 d_i^2 a_i = 0, \quad \sum_{i=1}^3 d_i^2 a_i^2 + d_i^2 b_i = 0, \quad \sum_{i=1}^3 -(m^3 + m) d_i^2 a_i^2 + 2v d_i^2 c_i = 0, \\
& \sum_{i=1}^3 \frac{1}{3} d_i^2 a_i^3 + d_i^2 a_i b_i = 0, \quad \sum_{i=1}^3 -(m^3 + m) d_i^2 a_i^3 + 6v d_i^2 a_i c_i = 0, \\
& \sum_{i=1}^3 \frac{(m^3 + m)^2}{12v^2} d_i^2 a_i^4 - \frac{m^3 + m}{v} d_i^2 a_i^2 c_i - \frac{m^2 + 1}{vm} d_i^2 b_i c_i + d_i^2 c_i^2 = 0, \\
& \sum_{i=1}^3 \frac{1}{6} d_i^2 a_i^4 + d_i^2 a_i^2 b_i + \frac{1}{2} d_i^2 b_i^2 = 0, \quad \sum_{i=1}^3 \frac{1}{30} d_i^2 a_i^5 + \frac{1}{3} d_i^2 a_i^3 b_i + \frac{1}{2} d_i^2 a_i b_i^2 = 0, \\
& \sum_{i=1}^3 \frac{(m^3 + m)^2}{120v^2} d_i^2 a_i^5 - \frac{(m^3 + m)}{6v} d_i^2 a_i^3 c_i - \frac{m^2 + 1}{2vm} d_i^2 a_i b_i c_i + \frac{1}{2} d_i^2 a_i c_i^2 = 0, \\
& \sum_{i=1}^3 \frac{1}{90} d_i^2 a_i^6 + \frac{1}{12} d_i^2 a_i^4 b_i + \frac{1}{2} d_i^2 a_i^2 b_i^2 + \frac{1}{6} d_i^2 b_i^3 = 0, \\
& \sum_{i=1}^3 \frac{(m^3 + m)^3}{720v^3} d_i^2 a_i^6 + \frac{(m^3 + m)^2}{24v^2} d_i^2 a_i^4 c_i + \frac{m^3 + m}{4v} d_i^2 a_i^2 c_i^2 + \\
& \quad \frac{(m^2 + 1)^2}{8v^2 m^2} d_i^2 b_i^2 c_i - \frac{m^2 + 1}{4mv} d_i^2 b_i c_i^2 + \frac{1}{6} d_i^2 c_i^3 = 0. \quad (4.54)
\end{aligned}$$

When $r = 4$, the system of polynomial equations (4.21) contains the first 7 equations in the system of polynomial equations (4.54). Now, m and v are considered as two additional variables in the above system of polynomial equations. Hence, there are 13 variables with only 7 equations. If we choose $d_1 = d_2 = d_3$ and take the lexicographical ordering $a_1 \succ a_2 \succ a_3 \succ b_1 \succ b_2 \succ b_3 \succ c_1 \succ c_2 \succ c_3 \succ m \succ v$, the Grobener bases (cf. [Buchberger \[1965\]](#)) of the above system of polynomial equations will return a non-trivial solution (due to the complexity of the roots, we will not present them here). As a consequence, $\rho(l) \geq 5$ under the case $l = 2$.

For $l = 2$ and $r = 5$, the system of polynomial equations (4.21) retains the first 9 equations in system (4.54). It can be checked that if we choose $m = \pm 2, v = 1$, then the system of polynomial equations when $r = 5$ does not have any non-trivial solution (note that, we also use the same lexicographical order as that being used in

the case $r = 4$). So, $\rho(l) = 5$. However, we can check that the value of $m = \frac{1}{10}$ (close to 0 in general) and $v = 1$ will lead the system of polynomial equations (4.54) to not having any non-trivial solution. Thus, $\bar{\rho}(l) = 6$. This concludes the proof or part (b) of the proposition.

4.8.3 Proofs for Section 5

FULL PROOF OF THEOREM 4.5.2 Here, we shall complete the proof of Theorem 4.5.2, which is the generalization of the argument in Section 4.5.1 for a special case for G_0 . Note that, the idea of this generalization is also used to the other settings of $G_0 \notin \mathcal{S}_1$. Now, we consider the possible existence of generic components in G_0 , i.e., there are no homologous sets or symmetry components. Let $u_1 = 1 < u_2 < \dots < u_{\bar{i}_1} \in [1, k_0+1]$ such that $(\frac{v_j^0}{1 + (m_j^0)^2}, \theta_j^0) = (\frac{v_l^0}{1 + (m_l^0)^2}, \theta_l^0)$ and $m_j^0 m_l^0 > 0$ for all $u_i \leq j, l \leq u_{i+1}-1, 1 \leq i \leq \bar{i}_1-1$. The constraint $m_j^0 m_l^0 > 0$ is due to the conformant property of the homologous sets of G_0 . By definition, we have $|I_{u_i}| = u_{i+1} - u_i$ for all $1 \leq i \leq \bar{i}_1 - 1$ where I_{u_i} denotes the set of all components homologous to component u_i .

To show that G_0 is 1-singular, we construct a sequence of $G \in \mathcal{E}_{k_0}$ such that $(p_i, \theta_i, v_i, m_i) = (p_i^0, \theta_i^0, v_i^0, m_i^0)$ for all $u_2 \leq i \leq k_0$, i.e., all the components of G and G_0 are identical from index u_2 up to k_0 . Hence, in the construction of the components from index u_1 to $u_2 - 1$ of G we consider only the homologous set I_{u_1} of G_0 . Utilizing the argument from the special case proof of Theorem 4.5.2 in Section 4.5.1, the construction of the sequence of G is specified by $\Delta\theta_i = \Delta v_i = \Delta p_i = 0$ and $\sum_{i=u_1}^{u_2-1} p_i \Delta m_i / v_i^0 = 0$. Thus G_0 is 1-singular. It remains to demonstrate that $G_0 \in \mathcal{S}_1$ is not 2-singular relative to \mathcal{E}_{k_0} .

Indeed, consider any sequence $G \in \mathcal{E}_{k_0} \rightarrow G_0$ under W_2 distance. Since $W_2^2(G, G_0) \asymp$

$D_2(G_0, G)$ (cf. Lemma 4.3.1), we have the 2-minimal form for the sequence G as

$$\frac{p_G(x) - p_{G_0}(x)}{W_2^2(G, G_0)} \asymp \frac{A_1(x) + A_2(x)}{D_2(G_0, G)},$$

where $A_1(x)/D_2(G_0, G)$ and $A_2(x)/D_2(G_0, G)$ are linear combinations of the elements of the forms $\frac{\partial^{|\alpha|} f}{\theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}}(x|\eta_i^0)$ for any $1 \leq i \leq k_0$ and $0 \leq |\alpha| \leq 2$. In $A_1(x)/D_2(G_0, G)$, the indices of the components range from 1 to $s_{\bar{i}_1} - 1$ while in $A_2(x)/D_2(G_0, G)$, the indices of the components range from $u_{\bar{i}_1}$ to k_0 . It is convenient to think of the term $A_1(x)/D_2(G_0, G)$ as the linear combination of homologous components, and $A_2(x)/D_2(G_0, G)$ as the linear combination of generic components, i.e., no Gaussian nor homologous components.

Regarding $A_2(x)/D_2(G_0, G)$, since we have the system of partial differential equations in (4.2), the collection of functions in $\left\{ \frac{\partial^{|\alpha|} f}{\partial \theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}}(x|\eta_i^0) : |\alpha| \leq 2, 1 \leq i \leq k_0 \right\}$ are not linearly independent. Employing the same strategy described in Section 4.4, we obtain a reduced system of linearly independent partial derivatives in Lemma 4.4.3. This is the set $\left\{ \frac{\partial^{|\alpha|} f}{\partial \theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}}(x|\eta_i^0) : \alpha \in \mathcal{F}_2, 1 \leq i \leq k_0 \right\}$. Let $\lambda_{\alpha_1 \alpha_2 \alpha_3}^{(2)}(\eta_i^0)/D_2(G_0, G)$ be the coefficient of the terms $\frac{\partial^{|\alpha|} f}{\theta^{\alpha_1} v^{\alpha_2} m^{\alpha_3}}(x|\eta_i^0)$ for any $s_{\bar{i}_1} \leq i \leq k_0$ and $\alpha \in \mathcal{F}_2$. The formulae for $\lambda_{\alpha_1, \alpha_2, \alpha_3}^{(2)}$ will be given later in Case 2.

Regarding $A_1(x)/D_2(G_0, G)$, by exploiting the fact that $(\frac{v_j^0}{1 + (m_j^0)^2}, \theta_j^0) = (\frac{v_l^0}{1 + (m_l^0)^2}, \theta_l^0)$ for all $u_i \leq j, l \leq u_{i+1} - 1, 1 \leq i \leq \bar{i}_1 - 1$, the term $A_1(x)/D_2(G_0, G)$ can be written as

$$\begin{aligned} \frac{A_1(x)}{D_2(G_0, G)} &= \frac{1}{D_2(G_0, G)} \left(\sum_{l=1}^{\bar{i}_1-1} \left\{ \sum_{i=u_l}^{u_{l+1}-1} \left[\sum_{j=1}^5 \beta_{jl}^{(2)} (x - \theta_{u_l}^0)^{j-1} \right] f\left(\frac{x - \theta_{u_l}^0}{\sigma_i^0}\right) \times \right. \right. \\ &\quad \left. \left. \Phi\left(\frac{m_i^0(x - \theta_{u_l}^0)}{\sigma_i^0}\right) \right\} + \left[\sum_{j=1}^4 \gamma_{jl}^{(2)} (x - \theta_{u_l}^0)^{j-1} \right] \exp\left(-\frac{(m_{u_l}^0)^2 + 1}{2v_{u_l}^0} (x - \theta_{u_l}^0)^2\right) \right), \end{aligned}$$

where $f(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$. (This form is a general version of Eq. (4.26) in Section

(4.5.1) when $\bar{i}_1 = 2, u_1 = 1, u_2 = k_0 + 1$). The detailed formulas of $\beta_{jil}^{(2)}$ and $\gamma_{jl}^{(2)}$ for $1 \leq l \leq \bar{i}_1 - 1, u_l \leq i \leq u_{l+1} - 1$, and $1 \leq j \leq 5$ are thus similar to that of (4.26). Here, we rewrite their general fomulations for the transparency of subsequent arguments:

$$\begin{aligned}
\beta_{1il}^{(2)} &= \frac{2\Delta p_i}{\sigma_i^0} - \frac{p_i\Delta v_i}{(\sigma_i^0)^3} - \frac{p_i(\Delta\theta_i)^2}{(\sigma_i^0)^3} + \frac{3p_i(\Delta v_i)^2}{4(\sigma_i^0)^5}, \quad \beta_{2il}^{(2)} = \frac{2p_i\Delta\theta_i}{(\sigma_i^0)^3} - \frac{6p_i\Delta\theta_i\Delta v_i}{(\sigma_i^0)^5}, \\
\beta_{3il}^{(2)} &= \frac{p_i\Delta v_i}{(\sigma_i^0)^5} + \frac{p_i(\Delta\theta_i)^2}{(\sigma_i^0)^5} - \frac{3p_i(\Delta v_i)^2}{2(\sigma_i^0)^7}, \quad \beta_{4il}^{(2)} = \frac{2p_i\Delta\theta_i\Delta v_i}{(\sigma_i^0)^7}, \quad \beta_{5il}^{(2)} = \frac{p_i(\Delta v_i)^2}{4(\sigma_i^0)^9}, \\
\gamma_{1l}^{(2)} &= \sum_{j=u_l}^{u_{l+1}-1} -\frac{p_j m_j^0 \Delta\theta_j}{\pi(\sigma_j^0)^2} + \frac{2p_j m_j^0 \Delta\theta_j \Delta v_j}{\pi(\sigma_j^0)^4} - \frac{2p_j \Delta\theta_j \Delta m_j}{\pi(\sigma_j^0)^2}, \\
\gamma_{2l}^{(2)} &= \sum_{j=s_l}^{s_{l+1}-1} -\frac{p_j m_j^0 \Delta v_j}{2\pi(\sigma_j^0)^4} - \frac{p_j((m_j^0)^3 + 2m_j^0)(\Delta\theta_j)^2}{2\pi(\sigma_j^0)^4} + \frac{p_j \Delta m_j}{\pi(\sigma_j^0)^2} \\
&\quad + \frac{5p_j m_j^0 (\Delta v_j)^2}{8\pi(\sigma_j^0)^6} - \frac{p_j \Delta m_j \Delta v_j}{\pi(\sigma_j^0)^4}, \\
\gamma_{3l}^{(2)} &= \sum_{j=s_l}^{s_{l+1}-1} \frac{p_j(2(m_j^0)^2 + 2)\Delta m_j \Delta\theta_j}{\pi(\sigma_j^0)^4} - \frac{p_j((m_j^0)^3 + 2m_j^0)\Delta\theta_j \Delta v_j}{2\pi(\sigma_j^0)^6}, \\
\gamma_{4l}^{(2)} &= \sum_{j=s_l}^{s_{l+1}-1} -\frac{p_j((m_j^0)^3 + 2m_j^0)(\Delta v_j)^2}{8\pi(\sigma_j^0)^8} - \frac{p_j m_j^0 (\Delta m_j)^2}{2\pi(\sigma_j^0)^4} \\
&\quad + \frac{p_j((m_j^0)^2 + 1)\Delta m_j \Delta v_j}{\pi(\sigma_j^0)^6},
\end{aligned}$$

where $1 \leq l \leq \bar{i}_1 - 1$ and $u_l \leq i \leq u_{l+1} - 1$. Now, suppose that all the coefficients of $A_1(x)/D_2(G_0, G)$ and $A_2(x)/D_2(G_0, G)$ go to 0. It implies that $\gamma_{jl}^{(2)}/D_2(G_0, G)$ ($1 \leq j \leq 4, 1 \leq l \leq \bar{i}_1 - 1$), $\beta_{jil}^{(2)}/D_2(G_0, G)$ ($1 \leq j \leq 5, u_l \leq i \leq u_{l+1} - 1, 1 \leq l \leq \bar{i}_1 - 1$), and $\lambda_{\alpha_1\alpha_2\alpha_3}^{(2)}(\eta_i^0)/D_2(G_0, G)$ (for all $|\alpha| \leq 2$) go to 0. From the formation of $D_2(G_0, G)$, we can find at least one index $1 \leq i^* \leq k_0$ such that $(|\Delta p_{i^*}| + p_{i^*}(|\Delta\theta_{i^*}|^2 + |\Delta v_{i^*}|^2 + |\Delta m_{i^*}|^2))/D_2(G_0, G) \not\rightarrow 0$. Let

$$\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) = |\Delta p_{i^*}| + p_{i^*}(|\Delta\theta_{i^*}|^2 + |\Delta v_{i^*}|^2 + |\Delta m_{i^*}|^2).$$

Now, there are two possible cases for i^* :

Case 1 $u_1 \leq i^* \leq u_{\bar{i}_1} - 1$. Without loss of generality, we assume that $u_1 \leq i^* \leq u_2 - 1$. Denote

$$d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) = \sum_{j=u_1}^{u_2-1} |\Delta p_j| + p_j(|\Delta \theta_j|^2 + |\Delta v_j|^2 + |\Delta m_j|^2).$$

Since $\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*})/D_2(G_0, G) \not\rightarrow 0$, we have

$$d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*})/D_2(G_0, G) \not\rightarrow 0.$$

Therefore, for $1 \leq j \leq 5$ and $u_1 \leq i \leq u_2 - 1$, $D_j := \frac{\alpha_{ji1}}{d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*})} \rightarrow 0$. Now, our argument for this case is organized further into two steps:

Step 1.1 From the vanishes of D_2 and D_4 , we obtain $p_i \Delta \theta_i / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0$ for all $u_1 \leq i \leq u_2 - 1$. Combining this result with $D_1 \rightarrow 0$ and $D_5 \rightarrow 0$, we achieve for all $u_1 \leq i \leq u_2 - 1$ that

$$\Delta p_i / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}), p_i \Delta v_i / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0.$$

Therefore, for all $u_1 \leq i \leq u_2 - 1$,

$$p_i (\Delta \theta_i)^2 / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}), p_i (v_i)^2 / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0.$$

These results eventually show that

$$U := \left(\sum_{j=u_1}^{u_2-1} p_j (\Delta m_j)^2 \right) / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \not\rightarrow 0.$$

Step 1.2 Since $p_i \Delta \theta_i / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*})$, $p_i \Delta v_i / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0$, by using the result that $\gamma_{41}^{(2)} / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0$, we have

$$V := \left[\sum_{j=u_1}^{u_2-1} \frac{p_j m_j^0 (\Delta m_j)^2}{(\sigma_j^0)^4} \right] / d(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0.$$

As $U \not\rightarrow 0$, we obtain

$$V/U = \left[\sum_{j=u_1}^{u_2-1} \frac{p_j^n m_j^0 (\Delta m_j)^2}{(\sigma_j^0)^4} \right] / \sum_{j=u_1}^{u_2-1} p_j (\Delta m_j)^2 \rightarrow 0. \quad (4.55)$$

Since $m_i^0 m_j^0 > 0$ for all $u_1 \leq i, j \leq u_2 - 1$, without loss of generality we assume that $m_j^0 > 0$ for all $s_1 \leq j \leq s_2 - 1$. However, it implies that

$$\left[\sum_{j=u_1}^{u_2-1} \frac{p_j m_j^0 (\Delta m_j)^2}{(\sigma_j^0)^4} \right] / \sum_{j=u_1}^{u_2-1} p_j (\Delta m_j)^2 \geq m_{\min} \sum_{j=u_1}^{u_2-1} p_j (\Delta m_j)^2 / \sum_{j=u_1}^{u_2-1} p_j (\Delta m_j)^2, \quad (4.56)$$

where $m_{\min} := \min_{u_1 \leq j \leq u_2-1} \left\{ \frac{m_j^0}{(\sigma_j^0)^4} \right\}$. Combining with (4.55), $m_{\min} = 0$ — a contradiction. In sum, Case 1 cannot happen.

Case 2 $u_{\bar{i}_1} \leq i^* \leq k_0$. We can write down the formation of $A_2(x)/D_2(G_0, G)$ as follows

$$\frac{A_2(x)}{D_2(G_0, G)} = \frac{1}{D_2(G_0, G)} \left(\sum_{i=u_{\bar{i}_1}}^{k_0} \sum_{\alpha \in \mathcal{F}_2} \lambda_{\alpha_1, \alpha_2, \alpha_3}^{(2)}(\eta_i^0) \frac{\partial^{|\alpha|} f}{\partial \theta^{\alpha_1} \partial v^{\alpha_2} \partial m^{\alpha_3}}(x | \eta_i^0) \right),$$

where $\lambda_{\alpha_1, \alpha_2, \alpha_3}^{(2)}(\eta_i^0)$ are given by

$$\begin{aligned}\lambda_{0,0,0}^{(2)}(\eta_i^0) &= \Delta p_i, \quad \lambda_{1,0,0}^{(2)}(\eta_i^0) = p_i \Delta \theta_i, \quad \lambda_{0,1,0}^{(2)}(\eta_i^0) = p_i \Delta v_i + p_i (\Delta \theta_i)^2, \\ \lambda_{0,0,1}^{(2)}(\eta_i^0) &= -\frac{(m_1^0)^3 + m_1^0}{2v_1^0} p_i (\Delta \theta_{1i})^2 - \frac{1}{v_1^0} p_i \Delta v_i \Delta m_i + p_i \Delta m_i, \\ \lambda_{0,2,0}^{(2)}(\eta_i^0) &= p_i (\Delta v_i)^2, \quad \lambda_{0,0,2}^{(2)}(\eta_i^0) = -\frac{(m_1^0)^2 + 1}{2v_1^0 m_1^0} p_i \Delta v_i \Delta m_i + p_i \Delta (m_i)^2, \\ \lambda_{1,1,0}^{(2)}(\eta_i^0) &= p_i \Delta \theta_i \Delta v_i, \quad \lambda_{1,0,1}^{(2)}(\eta_i^0) = p_i \Delta \theta_i \Delta m_i.\end{aligned}$$

From the assumption with the coefficients of $A_2(x)/D_2(G_0, G)$, we have

$$\lambda_{\alpha_1, \alpha_2, \alpha_3}^{(2)}(\eta_i^0)/D_2(G_0, G) \rightarrow 0$$

for any $u_{\bar{i}_1} \leq i \leq k_0$. From the hypothesis with i^* , $\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*})/D_2(G_0, G) \not\rightarrow 0$.

Therefore, it leads to $\lambda_{\alpha_1, \alpha_2, \alpha_3}^{(2)}(\eta_{i^*}^0)/\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*})$ for any $u_{\bar{i}_1} \leq i \leq k_0$ and $\alpha \in \mathcal{F}_2$.

Now, since $\lambda_{1,0,0}^{(2)}(\eta_{i^*}^0)/\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0$, we obtain $\Delta \theta_{i^*}/\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0$. Combining this result with $\lambda_{1,0,0}^{(2)}(\eta_{i^*}^0)/\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0$, we have

$$\Delta v_{i^*}/\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0.$$

Furthermore, as $\lambda_{0,0,1}^{(2)}(\eta_{i^*}^0)/\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0$, we get $\Delta m_{i^*}/\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0$. Hence, since $\lambda_{0,0,0}^{(2)}(\eta_{i^*}^0)/\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*}) \rightarrow 0$, we ultimately obtain

$$1 = \frac{|\Delta p_{i^*}| + p_{i^*}(|\Delta \theta_{i^*}|^2 + |\Delta v_{i^*}|^2 + |\Delta m_{i^*}|^2)}{\tau(p_{i^*}, \theta_{i^*}, v_{i^*}, m_{i^*})} \rightarrow 0,$$

which is a contradiction. As a consequence, Case 2 cannot happen.

Summarizing, not all the coefficients $\gamma_{jl}^{(2)}/D_2(G_0, G)$ ($1 \leq j \leq 4$, $1 \leq l \leq \bar{i}_1 - 1$), $\beta_{jil}^{(2)}/D_2(G_0, G)$ ($1 \leq j \leq 5$, $u_l \leq i \leq u_{l+1} - 1$, $1 \leq l \leq \bar{i}_1 - 1$), $\lambda_{\alpha_1 \alpha_2 \alpha_3}^{(2)}(\eta_i^0)/D_2(G_0, G)$ (for all $\alpha \in \mathcal{F}_2$) go to 0. From Definition 4.3.2, G_0 is not 2-singular relative to \mathcal{E}_{k_0} .

This concludes our proof.

FULL PROOF OF THEOREM 4.5.3 We divide the proof of this theorem into two main steps.

Step 1: To illustrate our calculations, we consider at first a simple setting of $G_0 \in \mathcal{S}_2$ in which $m_1^0, m_2^0, \dots, m_{k_0}^0 = 0$, leaving out the possible setting of conformant homologous sets and generic components. A complete proof for all possible settings of $G_0 \in \mathcal{S}_2$ will be given in Step 2.

G_0 is 2-singular To establish this, we look at 2-minimal form for $\frac{p_G(x) - p_{G_0}(x)}{W_2^2(G, G_0)}$, which is asymptotically equal to

$$\frac{1}{W_2^2(G, G_0)} \left[\sum_{i=1}^{k_0} \left(\sum_{j=1}^5 \zeta_{li}^{(2)} (x - \theta_i^0)^{j-1} \right) f \left(\frac{x - \theta_i^0}{\sigma_i^0} \right) \right], \quad (4.57)$$

where $\zeta_{li}^{(2)}$ are the polynomials in terms of $\Delta\theta_j$, Δv_j , Δm_j , and Δp_j as $1 \leq i, j \leq k_0$ and $1 \leq l \leq 5$. To make all the coefficients vanish, it suffices to have $(\Delta v_i)^2 / W_2^2(G, G_0) \rightarrow 0$ and

$$\begin{aligned} & \left[-\frac{p_i \Delta v_i}{2(\sigma_i^0)^3} - \frac{p_i (\Delta \theta_i)^2}{2(\sigma_i^0)^3} + \frac{3p_i (\Delta v_i)^2}{8(\sigma_i^0)^5} - \frac{2p_i \Delta \theta_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^2} + \frac{\Delta p_i}{\sigma_i^0} \right] / W_2^2(G, G_0) \rightarrow 0, \\ & \left[\frac{\Delta \theta_i}{(\sigma_i^0)^3} + \frac{2\Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^2} - \frac{3\Delta \theta_i \Delta v_i}{2(\sigma_i^0)^5} - \frac{2\Delta v_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^4} \right] / W_2^2(G, G_0) \rightarrow 0, \\ & \left[\frac{\Delta v_i}{2(\sigma_i^0)^5} + \frac{(\Delta \theta_i)^2}{2(\sigma_i^0)^5} + \frac{2\Delta \theta_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^4} \right] / W_2^2(G, G_0) \rightarrow 0, \\ & \left[\frac{\Delta \theta_i \Delta v_i}{2(\sigma_i^0)^7} + \frac{\Delta v_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^6} \right] / W_2^2(G, G_0) \rightarrow 0. \end{aligned} \quad (4.58)$$

This can be achieved by choosing a sequence of $G \rightarrow G_0$ in W_2 such that $\Delta \theta_i = \Delta v_i = \Delta m_i = \Delta p_i = 0$ for all $2 \leq i \leq k_0$; only for component 1 do we set $\Delta \theta_1 = -2\Delta m_1 \sigma_1^0 / \sqrt{2\pi}$ and $\Delta v_1 = (\Delta \theta_1)^2 / 2$. It follows that G_0 is 2-singular relative to \mathcal{E}_{k_0} .

G_0 is not 3-singular The 3-minimal form of $(p_G(x) - p_{G_0}(x))/W_3^3(G, G_0)$ is asymptotically equal to

$$\frac{1}{W_3^3(G, G_0)} \left[\sum_{i=1}^{k_0} \left(\sum_{j=1}^7 \zeta_{ji}^{(3)} (x - \theta_i^0)^{j-1} \right) f \left(\frac{x - \theta_i^0}{\sigma_i^0} \right) \right], \quad (4.59)$$

where $\zeta_{li}^{(3)}$ are the polynomials in terms of $\Delta\theta_j$, Δv_j , Δm_j , and Δp_j as $1 \leq i, j \leq k_0$ and $1 \leq l \leq 7$. Suppose that there exists a sequence $G \rightarrow G_0$ under W_3 such that all the coefficients of the 3-minimal form vanish. For any $1 \leq i \leq k_0$, it follows after some calculations that

$$\begin{aligned} C_1^{(i)} &:= \left[-\frac{p_i \Delta v_i}{2(\sigma_i^0)^3} - \frac{p_i (\Delta \theta_i)^2}{2(\sigma_i^0)^3} + \frac{3p_i (\Delta v_i)^2}{8(\sigma_i^0)^5} - \frac{2p_i \Delta \theta_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^2} + \frac{3p_i (\Delta \theta_i)^2 \Delta v_i}{4(\sigma_i^0)^5} + \right. \\ &\quad \left. \frac{2p_i \Delta \theta_i \Delta v_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^4} + \frac{\Delta p_i}{\sigma_i^0} \right] / W_3^3(G, G_0) \rightarrow 0, \\ C_2^{(i)} &:= \left[\frac{p_i \Delta \theta_i}{(\sigma_i^0)^3} + \frac{2p_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^2} - \frac{3p_i \Delta \theta_i \Delta v_i}{2(\sigma_i^0)^5} - \frac{2p_i \Delta v_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^4} - \frac{p_i (\Delta \theta_i)^3}{2(\sigma_i^0)^5} - \right. \\ &\quad \left. \frac{3p_i (\Delta \theta_i)^2 \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^4} + \frac{15p_i \Delta \theta_i (\Delta v_i)^2}{8(\sigma_i^0)^7} + \frac{2p_i (\Delta v_i)^2 \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^6} \right] / W_3^3(G, G_0) \rightarrow 0, \\ C_3^{(i)} &:= \left[\frac{p_i \Delta v_i}{2(\sigma_i^0)^5} + \frac{p_i (\Delta \theta_i)^2}{2(\sigma_i^0)^5} - \frac{3p_i (\Delta v_i)^2}{4(\sigma_i^0)^7} + \frac{2p_i \Delta \theta_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^4} - \frac{3p_i (\Delta \theta_i)^2 \Delta v_i}{2(\sigma_i^0)^7} - \right. \\ &\quad \left. \frac{5\Delta \theta_i \Delta v_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^6} \right] / W_3^3(G, G_0) \rightarrow 0, \\ C_4^{(i)} &:= \left[\frac{p_i \Delta \theta_i \Delta v_i}{2(\sigma_i^0)^7} + \frac{p_i \Delta v_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^6} + \frac{p_i (\Delta \theta_i)^3}{6(\sigma_i^0)^7} - \frac{p_i (\Delta m_i)^3}{3\sqrt{2\pi}(\sigma_i^0)^4} + \right. \\ &\quad \left. \frac{p_i (\Delta \theta_i)^2 \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^6} - \frac{5p_i \Delta \theta_i (\Delta v_i)^2}{4(\sigma_i^0)^9} - \frac{2p_i (\Delta v_i)^2 \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^8} \right] / W_3^3(G, G_0) \rightarrow 0, \\ C_5^{(i)} &:= \left[\frac{p_i (\Delta v_i)^2}{8(\sigma_i^0)^9} - \frac{5p_i (\Delta v_i)^3}{16(\sigma_i^0)^{11}} + \frac{p_i (\Delta \theta_i)^2 \Delta v_i}{4(\sigma_i^0)^9} + \frac{p_i \Delta \theta_i \Delta v_i \Delta m_i}{\sqrt{2\pi}(\sigma_i^0)^8} \right] / W_3^3(G, G_0) \rightarrow 0, \\ C_6^{(i)} &:= \left[\frac{p_i \Delta \theta_i (\Delta v_i)^2}{8(\sigma_i^0)^{11}} + \frac{p_i (\Delta v_i)^2 \Delta m_i}{4\sqrt{2\pi}(\sigma_i^0)^{10}} \right] / W_3^3(G, G_0) \rightarrow 0, \\ C_7^{(i)} &:= p_i (\Delta v_i)^3 / 48(\sigma_i^0)^3 W_3^3(G, G_0) \rightarrow 0. \end{aligned} \quad (4.60)$$

Since the system of limits in (4.60) holds for any $1 \leq i \leq k_0$, to further simplify the argument without loss of generality, we consider $k_0 = 1$. Under that scenario, we can rewrite $W_3^3(G, G_0) = p_1(|\Delta\theta_1|^3 + |\Delta v_1|^3 + |\Delta m_1|^3)$ where $p_1 = 1$. Additionally, for the simplicity of the presentation, we denote $C_i := C_i^{(1)}$ for any $1 \leq i \leq 7$. Now, our argument is organized into the following key steps

Step 1.1: We will argue that $\Delta\theta_1, \Delta v_1, \Delta m_1 \neq 0$. If $\Delta\theta_1 = 0$, by combining the vanishing of C_5 and C_7 , we achieve $(\Delta v_1)^2/W_3^3(G, G_0) \rightarrow 0$. Combining this result with $C_3 \rightarrow 0$, we obtain $\Delta v_1/W_3^3(G, G_0) \rightarrow 0$. Combining the previous results with $C_4 \rightarrow 0$ eventually yields that $(\Delta m_1)^3/W_3^3(G, G_0) \rightarrow 0$. Hence, $1 = p_1(|\Delta v_1|^3 + |\Delta m_1|^3)/W_3^3(G, G_0) \rightarrow 0$, which is a contradiction.

If $\Delta v_1 = 0$, then $C_1 + \Delta\theta_1 C_2 \rightarrow 0$ implies that $(\Delta\theta_1)^2/W_3^3(G, G_0) \rightarrow 0$. Combining this result with $C_4 \rightarrow 0$, we achieve $(\Delta m_1)^3/W_3^3(G, G_0) \rightarrow 0$, which also leads to a contradiction.

If $\Delta m_1 = 0$, then $C_6 \rightarrow 0$ leads to $(\Delta\theta_1)(\Delta v_1)^2/W_3^3(G, G_0) \rightarrow 0$. Combine this result with $C_4 \rightarrow 0$ leads to

$$\left[\frac{(\Delta\theta_1)(\Delta v_1)}{2(\sigma_1)^7} + \frac{(\Delta\theta_1)^3}{6(\sigma_1)^7} \right] / W_3^3(G, G_0) \rightarrow 0. \quad (4.61)$$

The combination of the above result and $C_3 \rightarrow 0$ implies that $\Delta v_1/W_3^3(G, G_0) \rightarrow 0$. Combine the former result with (4.61), we obtain $(\Delta\theta_1)^3/W_3^3(G, G_0) \rightarrow 0$, which is also a contradiction. Overall, we obtain the conclusion of this step.

Step 1.2: If $|\Delta v_1|$ is the maximum among $|\Delta\theta_1|$, $|\Delta v_1|$, and $|\Delta m_1|$. Then from $C_7 \rightarrow 0$, we obtain $|\Delta v_1|^3/(|\Delta\theta_1|^3 + |\Delta v_1|^3 + |\Delta m_1|^3) \rightarrow 0$, which is a contradiction.

Step 1.3: If $|\Delta\theta_1|$ is the maximum among $|\Delta\theta_1|$, $|\Delta v_1|$, and $|\Delta m_1|$. Denote $\Delta v_1/\Delta\theta_1 \rightarrow k_1$ and $\Delta m_1/\Delta\theta_1 \rightarrow k_2$. From C_7 , we obtain $k_1 = 0$. As $C_2 \rightarrow 0$, we obtain

$$\left[-\Delta\theta_1/(\sigma_1^0)^3 + 2\Delta m_1/\sqrt{2\pi}(\sigma_1^0)^2 \right] / (|\Delta\theta_1| + |\Delta v_1| + |\Delta m_1|) \rightarrow 0.$$

By diving both the numerator and denominator of this ratio by $\Delta\theta_1$, we quickly obtain the equation $1/(\sigma_1^0)^3 + 2k_2/\sqrt{2\pi}(\sigma_1^0)^2 = 0$, which yields the solution $k_2 = -\sqrt{\pi}/\sqrt{2}\sigma_1^0$.

Now, $C_5 \rightarrow 0$ yields that $(\Delta v_1)^2/(|\Delta\theta_1|^3 + |\Delta v_1|^3 + |\Delta m_1|^3) \rightarrow 0$. Applying this result to $C_3 \rightarrow 0$ and $C_4 \rightarrow 0$, we have $M_1, M_2 \rightarrow 0$ where the formations of M_1, M_2 are as follows:

$$\begin{aligned} M_1 &:= \left(\frac{\Delta v_1}{2(\sigma_1^0)^5} + \frac{(\Delta\theta_1)^2}{2(\sigma_1^0)^5} + \frac{2(\Delta\theta_1)(\Delta m_1)}{\sqrt{2\pi}(\sigma_1^0)^4} \right) / (|\Delta\theta_1|^3 + |\Delta v_1|^3 + |\Delta m_1|^3), \\ M_2 &:= \left(\frac{(\Delta\theta_1)(\Delta v_1)}{2(\sigma_1^0)^7} + \frac{(\Delta v_1)(\Delta m_1)}{\sqrt{2\pi}(\sigma_1^0)^6} + \frac{(\Delta\theta_1)^3}{6(\sigma_1^0)^7} - \frac{(\Delta m_1)^3}{3\sqrt{2\pi}(\sigma_1^0)^4} + \right. \\ &\quad \left. + \frac{(\Delta\theta_1)^2(\Delta m_1)}{\sqrt{2\pi}(\sigma_1^0)^6} \right) / (|\Delta\theta_1|^3 + |\Delta v_1|^3 + |\Delta m_1|^3). \end{aligned}$$

Now, $\left(\frac{\Delta\theta_1}{(\sigma_1^0)^2} + \frac{2\Delta m_1}{\sqrt{2\pi}\sigma_1^0} \right) M_1 - M_2$ yields that

$$\left[\frac{(\Delta m_1)^3}{3\sqrt{2\pi}} + \frac{2(\Delta\theta_1)(\Delta m_1)^2}{\pi\sigma_1^0} + \frac{2(\Delta\theta_1)^2(\Delta m_1)}{\sqrt{2\pi}(\sigma_1^0)^2} + \frac{(\Delta\theta_1)^3}{3(\sigma_1^0)^3} \right] / (|\Delta\theta_1|^3 + |\Delta v_1|^3 + |\Delta m_1|^3) \rightarrow 0.$$

By dividing both the numerator and denominator of this term by $(\Delta\theta_1)^3$, we obtain the equation $\frac{k_2^3}{3\sqrt{2\pi}} + \frac{2k_2^2}{\pi\sigma_1^0} + \frac{2k_2}{\sqrt{2\pi}(\sigma_1^0)^2} + \frac{1}{3(\sigma_1^0)^3} = 0$. Since $k_2 = -\frac{\sqrt{\pi}}{\sqrt{2}\sigma_1^0}$, this equation yields $\pi/6 - 1/3 = 0$, which is a contradiction. Therefore, this step cannot hold.

Step 1.4: If $|\Delta m_1|$ is the maximum among $|\Delta\theta_1|$, $|\Delta v_1|$, and $|\Delta m_1|$. The argument in this step is similar to that of Step 1.3. In fact, by denoting $\Delta\theta_1/\Delta m_1 \rightarrow k_3$ and $\Delta v_1/\Delta m_1 \rightarrow k_4$ then we also achieve $k_4 = 0$ and $k_3 = -\frac{\sqrt{2}}{\sqrt{\pi}\sigma_1^0}$ (by $C_2 \rightarrow 0$). Now by using the limits $C_3, C_4 \rightarrow 0$ as that of Step 1.3 and after some calculations, we

obtain the equation $\frac{k_3^3}{3(\sigma_1^0)^3} + \frac{2k_3^2}{\sqrt{2\pi}(\sigma_1)^2} + \frac{2k_3}{\pi\sigma_1^0} + \frac{1}{3\sqrt{2\pi}} = 0$, which also does not admit $k_3 = -\frac{\sqrt{2}}{\sqrt{\pi}\sigma_1^0}$ as a solution — a contradiction.

In sum, we have shown under that simple setting of $G_0 \in \mathcal{S}_2$, it is 2-singular, but not 3-singular relative to \mathcal{E}_{k_0} . Therefore, $\ell(G_0|\mathcal{E}_{k_0}) = 2$.

Step 2: Now, we address the general setting of $G_0 \in \mathcal{S}_2$, which accounts for the possible presence of both generic components and conformant homologous sets. Without loss of generality, we assume that $m_1^0, m_2^0, \dots, m_{\bar{i}_2}^0 = 0$ where $1 \leq \bar{i}_2 \leq k_0$ denotes the largest index i such that $m_i^0 = 0$. The remaining components are either conformant homologous sets or generic components. Using the exact same construction as that of Step 1, we establish easily that G_0 is 2-singular relative to \mathcal{E}_{k_0} . It remains to show that G_0 is not 3-singular relative to \mathcal{E}_{k_0} .

Consider the 3-minimal form for any sequence $G \in \mathcal{E}_{k_0} \rightarrow G_0$ under W_3 distance. Since $W_3^3(G, G_0) \asymp D_3(G_0, G)$ (cf. Lemma 4.3.1), we have

$$\frac{p_G(x) - p_{G_0}(x)}{W_3^3(G, G_0)} \asymp \frac{A'_1(x) + A'_2(x)}{D_3(G_0, G)},$$

where $A'_1(x)/D_3(G_0, G)$ is the linear combination of Gaussian components, i.e., the indices of components range from 1 to \bar{i}_2 , while $A'_2(x)/D_3(G_0, G)$ is the linear combination of conformant homologous components and generic components.

Suppose that all the coefficients of $A'_1(x)/D_3(G_0, G), A'_2(x)/D_3(G_0, G)$ go to 0. Similar to the argument in the proof of Theorem 4.5.2, observe that there is some index $\underline{i} \in [1, k_0]$ such that $(|\Delta p_{\underline{i}}| + p_{\underline{i}}(|\Delta \theta_{\underline{i}}|^3 + |\Delta v_{\underline{i}}|^3 + |\Delta m_{\underline{i}}|^3))/D_3(G_0, G) \not\rightarrow 0$. There are two possible cases regarding \underline{i} .

Case 2.1 $\underline{i} \in [1, \bar{i}_2]$. Applying a similar argument as that from Step 1 of this proof where we have only Gaussian components, we conclude that not all of the coefficients of $A'_1(x)/D_3(G_0, G)$ vanish, which is a contradiction. Therefore, Case 2.1 cannot

happen.

Case 2.2 $i \in [\bar{i}_2 + 1, k_0]$. Define

$$D_{r,new}(G_0, G) = \sum_{i=\bar{i}_2+1}^{k_0} (|\Delta p_i| + p_i(|\Delta \theta_i|^r + |\Delta v_i|^r + |\Delta m_i|^r)),$$

for any $r \in \{2, 3\}$. The idea of $D_{r,new}(G_0, G)$ is that we truncate the value of $D_r(G_0, G)$ from the index 1 to \bar{i}_2 , i.e., all the indices correspond to Gaussian components.

It is clear that $D_{3,new}(G_0, G) \lesssim D_{2,new}(G_0, G)$. Since $D_{3,new}(G_0, G)/D_3(G_0, G) \not\rightarrow 0$, we have $D_{2,new}(G_0, G)/D_3(G_0, G) \not\rightarrow 0$. By multiplying all the coefficients of $A'_2(x)/D_3(G_0, G)$ with $D_{2,new}(G_0, G)/D_3(G_0, G)$, we eventually obtain all the coefficients of $A'_2(x)/D_{2,new}(G_0, G)$ go to 0. However, by utilizing the same argument as in the proof of Theorem 4.5.2, we reach to the conclusion that the second order Taylor expansion is sufficient to have all the coefficients of $A'_2(x)/D_{2,new}(G_0, G)$ not vanish. Thus, not all the coefficients of $A'_2(x)/D_3(G_0, G)$ go to 0, which is a contradiction. As a consequence, Case 2.2 also cannot happen.

In sum, under no circumstance can all the coefficients of $A'_1(x)/D_3(G_0, G)$ and $A'_2(x)/D_3(G_0, G)$ be made to vanish. Hence, $G_0 \in \mathcal{S}_2$ is not 3-singular relative to \mathcal{E}_{k_0} , which concludes the proof.

4.8.4 Proofs for Section 4.7

PROOF OF PROPOSITION 4.7.1 (a) The proof proceeds by induction on l . When $l = 1$, the conclusion clearly holds. Assume that that conclusion of the proposition holds for $l - 1$. We will demonstrate that it also holds for l . Denote $y_i = a_i c_i$ and $z_i = b_i c_i$ for all $1 \leq i \leq l + 1$. Then, we can rewrite system of polynomial equations (4.36) as follows: $\sum_{i=1}^{l+1} z_i^u y_i = 0$ for any $0 \leq u \leq l$. If there exists some $1 \leq i_1 \leq l + 1$ such that $c_{i_1} = 0$, then we go back to the case $l - 1$, which

we have already known from the hypothesis that we do not have non-trivial solution.

Therefore, we assume that $c_i \neq 0$ for all $1 \leq i \leq l+1$, which implies that $y_i \neq 0$ for all $1 \leq i \leq l+1$. Now, the system of equations has the form of Vardermonde matrix,

which is $\begin{bmatrix} 1 & 1 & \dots & 1 \\ z_1 & z_2 & \dots & z_{l+1} \\ \vdots & \vdots & \ddots & \vdots \\ z_1^s & z_2^s & \dots & z_{l+1}^s \end{bmatrix}$. By suitable linear transformations, we can rewrite the

original system of equations as the following equivalent equations $\prod_{j \neq i} (z_j - z_i) y_i = 0$ for all $1 \leq i \leq l+1$. Since $y_i \neq 0$ for all $1 \leq i \leq l+1$, we obtain $\prod_{j \neq i} (z_j - z_i) = 0$ for all $1 \leq i \leq l+1$. As a consequence, there exists a partition J_1, J_2, \dots, J_s of $\{1, 2, \dots, l+1\}$ for some $1 \leq s \leq [l/2]$ such that if $i_2, i_3 \in J_u$ for $1 \leq u \leq s$, we have $z_{i_2} = z_{i_3}$ and for any $1 \leq i \neq j \leq s$, any two elements $z_{i_4} \in J_i, z_{j_4} \in J_j$ are different.

Choose any $j_i \in J_i$ for all $1 \leq i \leq s$. It is clear that the system of equations can be rewritten as $\sum_{i=1}^s z_{j_i}^u \sum_{j \in J_i} y_j = 0$ for all $0 \leq u \leq l+1$. If $s \geq 2$, it indicates that $|J_i| \leq l$ for all $1 \leq i \leq s$. Now, if we have some $1 \leq i_4 \leq s$ such that $\sum_{j \in J_{i_4}} y_j = 0$ then we obtain $\sum_{j \in J_{i_4}} a_j c_j = 0$. Since $z_{i_1} = z_{i_2}$ for any $i_1, i_2 \in J_{i_4}$, this equation can be rewritten as $\sum_{j \in J_{i_4}} a_j \prod_{v \neq j} b_v = 0$, which is a contradiction to the assumption of part (a) of the proposition. Therefore, $\sum_{j \in J_i} y_j \neq 0$ for all $1 \leq i \leq s$. However, by using the same argument as before, again by linear transformation, we can rewrite the new system of polynomial equations as $\sum_{j \in J_i} y_j \prod_{v \neq i} (z_{j_u} - z_{j_i}) = 0$ for all $1 \leq i \leq s$. This implies that there should be some $1 \leq u_1 \neq u_2 \leq s$ such that $z_{j_{u_1}} = z_{j_{u_2}}$, which is a contradiction.

As a consequence, we have $s = 1$, i.e., $|J_1| = l+1$. Hence, $b_1 c_1 = b_2 c_2 = \dots = b_{l+1} c_{l+1}$. Combining this fact with the equation $\sum_{i=1}^{l+1} a_i c_i = 0$, we obtain $\sum_{i=1}^{l+1} a_i \prod_{j \neq i} b_j = 0$, which is a contradiction to the assumption of the proposition. This concludes the proof.

(b) We choose $c_i = 0$ for all $i \notin I \subset \{1, \dots, l\}$. The system of polynomial equations

(4.36) becomes $\sum_{i \in s} a_i b_i^u c_i^{u+1} = 0$ for all $u \geq 0$. Notice that by choosing $b_i c_i = b_j c_j$ for all $i, j \in I$, we have $\sum_{i \in I} a_i b_i^u c_i^{u+1} = b_j c_j \sum_{i \in I} a_i c_i = 0$ for some $j \in I$ and for all $u \geq 1$ as long as $\sum_{i \in I} a_i c_i = 0$. Combining all the conditions, we obtain $\sum_{i \in J} a_i \prod_{j \neq i} b_j = 0$, which completes the proof.

(c) The result for the case $l = 1$ is obvious. For the case $l = 2$, after replacing c_3 in terms of c_1, c_2 , we obtain the following quadratic equation $(a_1 a_3 b_1 + a_1^2 b_3) c_1^2 + 2a_1 a_2 b_3 c_1 c_2 + (a_2 a_3 b_2 + a_2^2 b_3) c_2^2 = 0$. Note that, $c_1, c_2 \neq 0$ due to the assumption of part (c). Therefore, we does not have solution of this quadratic equation when $a_1^2 a_2^2 b_3^2 < (a_1 a_3 b_1 + a_1^2 b_3)(a_2 a_3 b_2 + a_2^2 b_3)$. It is equivalent to $\sum_{i=1}^3 a_i \prod_{j \neq i} b_j > 0$, which confirms our hypothesis. We are done.

FULL PROOF OF THEOREM 4.7.1 Here, we only provide the proof for part (b) as the proof for part (a) is similar. This is a generalization of the argument in Section 4.7.1. Under this situation, apart from the nonconformant homologous sets without C(1) singularity, we also have for G_0 the presence of Gaussian components components and possibly some conformant homologous sets, in addition to some generic components.

Let $u_1 = 1 < u_2 < \dots < u_{\bar{i}_3} \in [1, k_0+1]$ such that $(\frac{v_j^0}{1 + (m_j^0)^2}, \theta_j^0) = (\frac{v_l^0}{1 + (m_l^0)^2}, \theta_l^0)$ for all $u_i \leq j, l \leq u_{i+1} - 1, 1 \leq i \leq \bar{i}_3 - 1$, i.e., all the nonconformant homologous components without type C(1) singularity are from index 1 to $u_{\bar{i}_3}$. The remaining components are either Gaussian ones or conformant homologous sets or generic ones. It follows that $|I_{u_i}| = u_{i+1} - u_i$ for all $1 \leq i \leq \bar{i}_3 - 1$ and all I_{u_i} are nonconformant homologous sets without C(1) singularity.

Consider the \bar{r} -th minimal form for any sequence $G \in \mathcal{E}_{k_0} \rightarrow G_0$ under $W_{\bar{r}}$ distance where $\bar{r} = \max \left\{ 3, \bar{s}(G_0) + 1 \right\}$. Since $W_{\bar{r}}^{\bar{r}}(G, G_0) \asymp D_{\bar{r}}(G_0, G)$ (cf. Lemma 4.3.1), we

have

$$\frac{p_G(x) - p_{G_0}(x)}{W_{\bar{r}}^{\bar{r}}(G, G_0)} \asymp \frac{B_1(x) + B_2(x)}{D_{\bar{r}}(G_0, G)},$$

where $B_1(x)/D_{\bar{r}}(G_0, G)$ is the linear combination of nonconformant homologous components, i.e., the indices of components range from 1 to \bar{i}_3 while $B_2(x)/D_{\bar{r}}(G_0, G)$ is the linear combination of conformant homologous components, Gaussian components, and generic components.

Now, suppose that all the coefficients of $B_1(x)/D_{\bar{r}}(G_0, G), B_2(x)/D_{\bar{r}}(G_0, G)$ go to 0. Similar to the argument employed in the proof of Theorem 4.5.2, there is some index $\underline{i} \in [1, k_0]$ such that $(|\Delta p_{\underline{i}}| + p_{\underline{i}}(|\Delta \theta_{\underline{i}}|^{\bar{r}} + |\Delta v_{\underline{i}}|^{\bar{r}} + |\Delta m_{\underline{i}}|^{\bar{r}}))/D_{\bar{r}}(G_0, G) \not\rightarrow 0$. Now, there are two possible scenarios regarding \underline{i}

Case 1.1 $\underline{i} \in [1, u_{\bar{i}_3} - 1]$. Under that case, we can check that

$$\begin{aligned} \frac{B_1(x)}{D_{\bar{r}}(G_0, G)} &= \frac{1}{D_{\bar{r}}(G_0, G)} \left(\sum_{l=1}^{\bar{i}_3-1} \left\{ \sum_{i=u_l}^{u_{l+1}-1} \left[\sum_{j=1}^{2\bar{r}+1} \beta_{jil}^{(\bar{r})} (x - \theta_{u_l}^0)^{j-1} \right] f\left(\frac{x - \theta_{u_l}^0}{\sigma_i^0}\right) \times \right. \right. \\ &\quad \left. \left. \Phi\left(\frac{m_i^0(x - \theta_{u_l}^0)}{\sigma_i^0}\right) \right\} + \left[\sum_{j=1}^{2\bar{r}} \gamma_{jl}^{(\bar{r})} (x - \theta_{u_l}^0)^{j-1} \right] \exp\left(-\frac{(m_{u_l}^0)^2 + 1}{2v_{u_l}^0} (x - \theta_{u_l}^0)^2\right) \right). \end{aligned}$$

This representation of $B_1(x)/D_{\bar{r}}(G_0, G)$ is the general formulation of the equation (4.26) in Section (4.5.1) where $\bar{i}_3 = 2, u_1 = 1, u_2 = k_0 + 1$, and $\bar{r} = r$. Since $\underline{i} \in [1, u_{\bar{i}_3} - 1]$, there exists some index $l^* \in [1, \bar{i}_3 - 1]$ such that $\underline{i} \in [u_{l^*}, u_{l^*+1} - 1]$. By means of the same argument as that of Section 4.7.1 for $\beta_{jil}^{(\bar{r})}/D_{\bar{r}}(G_0, G) \rightarrow 0$ and $\gamma_{jl}^{(\bar{r})}/D_{\bar{r}}(G_0, G) \rightarrow 0$, we can extract the following system of polynomial limits:

$$\sum_{i=u_{l^*}}^{u_{l^*+1}-1} p_i^0 (m_i^0)^{l/2-1} (k_i)^{l/2} = 0,$$

where at least one of k_i differs from 0. Here, l is any even number such that $2 \leq l \leq 2\bar{r}$. From the formulation of $\bar{s}(G_0)$, since $\bar{r} \geq \bar{s}(G_0) + 1 \geq \bar{s}(|I_{u_l^*}|, \{p_i^0\}_{i \in I_{u_l^*}}, \{m_i^0\}_{i \in I_{u_l^*}}) + 1$, we can guarantee that the above system of polynomial equations does not have any non-trivial solution, which is a contradiction. Therefore, Case 1.1 cannot happen.

Case 1.2 $i \in [u_{\bar{i}_3}, k_0]$. Using the same argument as that in the proof of Theorem 4.5.3, the third order Taylor expansion is sufficient so that not all the coefficients of $B_2(x)/D_{3,new}(G_0, G)$ go to 0 where

$$D_{3,new}(G_0, G) = \sum_{i=u_{\bar{i}_3}}^{k_0} (|\Delta p_i| + p_i(|\Delta \theta_i|^3 + |\Delta v_i|^3 + |\Delta m_i|^3)).$$

Since $\bar{r} \geq 3$, we have $D_{3,new}(G_0, G)/D_{\bar{r}}(G_0, G) \not\rightarrow 0$. As all the coefficients of $B_2(x)/D_{\bar{r}}(G_0, G)$ vanish, it leads to all the coefficients of $B_2(x)/D_{3,new}(G_0, G)$ go to 0, which is a contradiction. Thus, Case 1.2 cannot happen.

In sum, for any sequence of G tending to G_0 in $W_{\bar{r}}$, not all the coefficients of $B_1(x)/D_{\bar{r}}(G_0, G)$ and $B_2(x)/D_{\bar{r}}(G_0, G)$ go to 0. By Definition 4.3.2, we conclude that $G_0 \in \mathcal{S}_2$ is not \bar{r} -singular relative to \mathcal{E}_{k_0} . As a consequence, $\ell(G_0|\mathcal{E}_{k_0}) \leq \bar{r} - 1 = \max \left\{ 2, \bar{s}(G_0) \right\}$.

PROOF OF PROPOSITION 4.7.2 Here, we utilize the same assumption on G_0 as that in the proof of Theorem 4.7.1, i.e., all the nonconformant homologous sets without C(1) singularity are from index 1 to $u_{\bar{i}_3}$. We also rearrange the components of G_0 such that the first nonconformant homologous set without C(1) singularity I_{u_1} has exactly k^* elements, i.e., $u_2 - u_1 = k^*$. As $u_1 = 1$, we have $u_2 = k^* + 1$.

(a) We will demonstrate that G_0 is 1-singular relative to \mathcal{E}_{k_0} . Indeed, the sequence of G is constructed as follows: $p_i = p_i^0, \theta_i = \theta_i^0, v_i = v_i^0$ for all $u_2 = k^* + 1 \leq i \leq k_0$, i.e., we match all the components of G and G_0 except the first k^* components of G_0 . Now, by proceeding in the same way as described in Section 4.7.1 up to Eq. (4.33), to

verify that G_0 is indeed 1-singular, the choice of the first k^* components of G needs to satisfy

$$\sum_{i=u_1}^{u_2-1} q_i \Delta t_i / \sum_{i=u_1}^{u_2-1} q_i |\Delta t_i| \rightarrow 0,$$

where $q_i = p_i/\sigma_i^0$ and $\Delta t_i = \Delta m_i/\sigma_i^0$ as $u_1 \leq i \leq u_2 - 1$. A simple choice is to take the first k^* components of G by $\sum_{i=u_1}^{u_2-1} q_i \Delta t_i = q_1 \Delta t_1 + q_2 \Delta t_2 = 0$, which is always possible. We conclude that G_0 is 1-singular relative to \mathcal{E}_{k_0} . Since $\bar{s}(G_0) = 1$ as $k^* = 2$, by combining with the upper bound of Theorem 4.7.1, we have $\ell(G_0|\mathcal{E}_{k_0}) = 1$.

(b) There are two cases to consider in this part

Case 1: All the homologous sets I of G_0 such that $|I| = k^*$ satisfy $\sum_{i \in I} p_i^0 \prod_{j \in I \setminus \{i\}} m_j^0 > 0$. To demonstrate that G_0 is 1-singular relative to \mathcal{E}_{k_0} , we utilize the same construction of G as that of part (a), i.e., $p_i = p_i^0, \theta_i = \theta_i^0, v_i = v_i^0$ for all $u_2 = k^* + 1 \leq i \leq k_0$ and $\sum_{i=u_1}^{u_2-1} q_i \Delta t_i = 0$. Next, we will show that G_0 is not 2-singular relative to \mathcal{E}_{k_0} . Using the same argument as that of the proof of Theorem 4.7.1, we obtain the following system of limiting rational polynomial functions:

$$\begin{aligned} & \sum_{i=u_{l^*}}^{u_{l^*+1}-1} q_i \Delta t_i / \sum_{i=u_{l^*}}^{u_{l^*+1}-1} q_i |\Delta t_i|^2 \rightarrow 0, \\ & \sum_{i=u_{l^*}}^{u_{l^*+1}-1} q_i t_i^0 (\Delta t_i)^2 / \sum_{i=u_{l^*}}^{u_{l^*+1}-1} q_i |\Delta t_i|^2 \rightarrow 0, \end{aligned}$$

where l^* is some index in $[1, \bar{i}_3 - 1]$ and $q_i = p_i/\sigma_i^0, \Delta t_i = \Delta m_i/\sigma_i^0, t_i^0 = m_i^0/\sigma_i^0$ for all $u_{l^*} \leq i \leq u_{l^*+1} - 1$. By employing the greedy extraction technique being described in Section 4.7.1.1, we obtain the following system of polynomial equations:

$$\sum_{i=u_{l^*}}^{u_{l^*+1}-1} p_i^0 c_i = 0, \quad \sum_{i=u_{l^*}}^{u_{l^*+1}-1} p_i^0 m_i^0 c_i^2 = 0,$$

where at least one of c_i differs from 0. Now, we have two possible scenarios:

Case 1.1: $|I_{u_l^*}| = u_{l^*+1} - u_{l^*} = 2$. Then, by solving the above system of equations, we obtain $\sum_{i \in I_{u_l^*}} p_i^0 \prod_{j \in I_{u_l^*} \setminus \{i\}} m_j^0 = 0$, which means $I_{u_l^*}$ is nonconformant homologous set with C(1) singularity of G_0 — a contradiction to the fact that $G_0 \in \mathcal{S}_{31}$.

Case 1.2: $|I_{u_l^*}| = u_{l^*+1} - u_{l^*} = k^* = 3$. Then, by solving the above system of equations, we obtain $\sum_{i \in I_{u_l^*}} p_i^0 \prod_{j \in I_{u_l^*} \setminus \{i\}} m_j^0 < 0$ — a contradiction to the assumption of Case 1.

Thus, G_0 is not 2-singular relative to \mathcal{E}_{k_0} . As a consequence, $\ell(G_0 | \mathcal{E}_{k_0}) = 1$ under Case 1.

Case 2: There exists at least one nonconformant homologous set I of G_0 such that $|I| = k^*$ satisfies $\sum_{i \in I} p_i^0 \prod_{j \in I \setminus \{i\}} m_j^0 < 0$. Without loss of generality, we assume the homologous set I_{u_1} of G_0 to have the property $\sum_{i \in I_{u_1}} p_i^0 \prod_{j \in I_{u_1} \setminus \{i\}} m_j^0 < 0$. We will show that G_0 is 2-singular relative to \mathcal{E}_{k_0} . In fact, we construct the sequence of G by letting $p_i = p_i^0, \theta_i = \theta_i^0, v_i = v_i^0$ for all $u_2 = k^* + 1 \leq i \leq k_0$. In order for G_0 to be 2-singular, it is sufficient that

$$\begin{aligned} \sum_{i=u_1}^{u_2-1} q_i \Delta t_i / \sum_{i=u_1}^{u_2-1} q_i |\Delta t_i|^2 &\rightarrow 0, \\ \sum_{i=u_1}^{u_2-1} q_i t_i^0 (\Delta t_i)^2 / \sum_{i=u_1}^{u_2-1} q_i |\Delta t_i|^2 &\rightarrow 0. \end{aligned}$$

The simple solution to the above system of limits is $\sum_{i=u_1}^{u_2-1} q_i \Delta t_i = 0$ and $\sum_{i=u_1}^{u_2-1} q_i t_i^0 (\Delta t_i)^2 = 0$. One solution to these two equations is $p_i = p_i^0$ and $\Delta m_i = (\sigma_i^0)^2 d_i / n$ for all $u_1 \leq i \leq u_2 - 1$ where d_1, d_2, d_3 satisfy

$$\sum_{i=u_1}^{u_2-1} p_i^0 d_i = 0, \quad \sum_{i=u_1}^{u_2-1} p_i^0 m_i^0 d_i^2 = 0,$$

which is guaranteed to have non-trivial solution as $\sum_{i \in I_{u_1}} p_i^0 \prod_{j \in I_{u_1} \setminus \{i\}} m_j^0 < 0$. Therefore, G_0 is 2-singular relative to \mathcal{E}_{k_0} . Since $\bar{s}(G_0) = 2$ as $k^* = 3$, combining with the upper bound of Theorem 4.7.1, we obtain $\ell(G_0 | \mathcal{E}_{k_0}) = 2$ under Case 2. This concludes our proof.

FULL PROOF OF THEOREM 4.7.2 Here, we shall provide the complete proof of Theorem 4.7.2, which is also the generalization of the argument in Section 4.7.1.3. Indeed, without loss of generality, we assume that $(p_1^0/\sigma_1^0, m_1^0/\sigma_1^0) = (p_2^0/\sigma_2^0, -m_2^0/\sigma_2^0)$. Next, we proceed to choosing a sequence of $G \in \mathcal{E}_{k_0}$ as follows: $p_i = p_i^0, \theta_i = \theta_i^0, v_i = v_i^0$ for all $1 \leq i \leq k_0$, and $m_1 = m_1^0 + 1/n, m_2 = m_2^0 - \sigma_2^0/n\sigma_1^0, m_i = m_i^0$ for all $3 \leq i \leq k_0$. The choice of m_1, m_2 is taken to guarantee that $\Delta m_1/\sigma_1^0 + \Delta m_2/\sigma_2^0 = 0$ as we have discussed in Section 4.7.1.3. Then, we can check that $\sum_{j=1}^2 p_j(m_j^0)^u (\Delta m_j)^v / (\sigma_j^0)^{u+v+1} = 0$ for all odd numbers $u \leq v$ when v is even number or for all even numbers $0 \leq u \leq v$ when v is odd number. From here, the completion of the proof follows in the same way as that of the special case previously described.

4.8.5 Proofs for auxiliary results

Lemma 4.8.1. *Let $\{f(x|\theta, \sigma, m), \theta \in \Theta_1, \sigma \in \Theta_2, m \in \Theta_3\}$ be a class of skew normal distribution. Denote $v := \sigma^2$, then*

$$\begin{cases} \frac{\partial^2 f}{\partial \theta^2}(x|\theta, \sigma, m) - 2\frac{\partial f}{\partial v}(x|\theta, \sigma, m) + \frac{m^3 + m}{v} \frac{\partial f}{\partial m}(x|\theta, \sigma, m) = 0, \\ 2m \frac{\partial f}{\partial m}(x|\theta, \sigma, m) + (m^2 + 1) \frac{\partial^2 f}{\partial m^2}(x|\theta, \sigma, m) + 2vm \frac{\partial^2 f}{\partial v \partial m}(x|\theta, \sigma, m) = 0. \end{cases}$$

Proof. Direct calculation yields

$$\begin{aligned}
\frac{\partial^2 f}{\partial \theta^2}(x|\theta, \sigma, m) &= \left\{ \left(-\frac{2}{\sqrt{2\pi}\sigma^3} + \frac{2(x-\theta)^2}{\sqrt{2\pi}\sigma^5} \right) \Phi\left(\frac{m(x-\theta)}{\sigma}\right) - \right. \\
&\quad \left. \frac{2m(m^2+2)(x-\theta)}{\sqrt{2\pi}\sigma^4} f\left(\frac{m(x-\theta)}{\sigma}\right) \right\} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right), \\
\frac{\partial f}{\partial v}(x|\theta, \sigma, m) &= \left\{ \left(-\frac{1}{\sqrt{2\pi}\sigma^3} + \frac{(x-\theta)^2}{\sqrt{2\pi}\sigma^5} \right) \Phi\left(\frac{m(x-\theta)}{\sigma}\right) - \right. \\
&\quad \left. \frac{m(x-\theta)}{\sqrt{2\pi}\sigma^4} f\left(\frac{m(x-\theta)}{\sigma}\right) \right\} \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right), \\
\frac{\partial f}{\partial m}(x|\theta, \sigma, m) &= \frac{2(x-\theta)}{\sqrt{2\pi}\sigma^2} f\left(\frac{m(x-\theta)}{\sigma}\right) \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right), \\
\frac{\partial^2 f}{\partial m^2}(x|\theta, \sigma, m) &= \frac{-2m(x-\theta)^3}{\sqrt{2\pi}\sigma^4} f\left(\frac{m(x-\theta)}{\sigma}\right) \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right), \\
\frac{\partial^2 f}{\partial v \partial m}(x|\theta, \sigma, m) &= \left(-\frac{2(x-\theta)}{\sqrt{2\pi}\sigma^4} + \frac{(m^2+1)(x-\theta)^3}{\sqrt{2\pi}\sigma^6} \right) \exp\left(-\frac{(x-\theta)^2}{2\sigma^2}\right).
\end{aligned}$$

From these equations, we can easily verify the conclusion of our lemma. \square

CHAPTER V

Robust estimation of mixing measures in finite mixture models

In finite mixture models, apart from underlying mixing measures, true kernel density functions of each subpopulation in the data are, in many scenarios, unknown. Perhaps the most popular approach is to choose some kernel functions that we empirically believe our data are generated from and use these kernels to fit our models. Nevertheless, as long as the chosen and the true kernels are different, statistical inference of mixing measures under this setting may be highly unstable. To overcome this challenge, we propose simple yet efficient robust estimators of the mixing measures in these models, which are inspired by the combination of minimum Hellinger distance estimators, model selection criteria, and the superefficiency phenomenon. We demonstrate that our estimators consistently recover the true number of components and achieve the optimal convergence rates of parameter estimates under both the well- and mis-specified kernel settings for any fixed bandwidth. These desirable asymptotic properties are illustrated via careful simulation studies with both synthetic and real data.

5.1 Introduction

Finite mixture models have become a popular model tool for making inference about the heterogeneity in data, starting, at least, with the classical work of Pearson [1894] on biometrical ratios on crabs. They have been used in various domains arising from biological, physical, and social sciences. For a comprehensive introduction of statistical inference in mixture models, we refer the readers to the books of McLachlan and Basford [1988], Lindsay [1995], Peel and McLachlan [2000].

In finite mixture models, we have our data $X_1, X_2, \dots, X_n \in \mathcal{X} \subset \mathbb{R}^d$ ($d \geq 1$) to be i.i.d observations from a finite mixture density function

$$p_{G_0^{f_0}}(x) = \int f_0(x|\theta) dG_0(\theta) = \sum_{i=1}^{k_0} p_i^0 f_0(x|\theta_i^0),$$

where $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ is a true but unknown mixing measure with exactly $k_0 < \infty$ non-zero components and $\left\{ f_0(x|\theta), \theta \in \Theta \subset \mathbb{R}^{d_1} \right\}$ is a true family of density functions, possibly partially unknown where $d_1 \geq 1$. There are essentially three principal challenges to the models that have attracted a great deal of attention from various researchers. They include estimating the true number of components k_0 , understanding the behavior of parameter estimations, i.e., the atoms and weights of true mixing measures G_0 , and determining the underlying kernel density f_0 of each subpopulation in the data. The first topic has been an intense area of research recently, see for example [Roeder, 1994, Escobar and West, 1995, Dacunha-Castelle and Gassiat, 1997, Richardson and Green, 1997, Dacunha-Castelle and Gassiat, 1999, Keribin, 2000, James et al., 2001, Chen et al., 2012, J.Chen and Khalili, 2012, Kasahara and Shimotsu, 2014a]. However, the second and third topic have receive much less attention due to their great mathematical difficulty. When the kernel density function f_0 is assumed to be known and k_0 is bounded by some fixed positive integer number, there

have been substantial advances in the understanding of parameter estimations of G_0 . More specifically, when k_0 is known, i.e., the exact-fitted setting, Ho and Nguyen [2016c] introduced a stronger version of classical parameter identifiability condition, which is first order identifiability, to guarantee the standard convergence rate $n^{-1/2}$ of parameter estimations. When k_0 is unknown and bounded, i.e., the over-fitted setting, Chen [1995], Nguyen [2013], Ho and Nguyen [2016c] utilized the notion of second order identifiability to establish the convergence rate $n^{-1/4}$ of parameter estimations. Recently, Ho and Nguyen [2016a,b] introduced the "singularity level" notion of the Fisher information matrix to characterize the convergence rates of parameter estimations when either the first or the second order identifiability condition fails to hold. When the kernel density function f_0 is unknown, there have been some work utilizing the semiparametric approaches [Bordes et al., 2006, Hunter et al., 2007]. The high level idea of these work is that we estimate f_0 from some classes of functions with infinite dimension and achieve the estimations of mixing measure G_0 accordingly. However, it is very difficult to establish a guarantee for the identifiability of the parameters, even when the parameter space is simple [Hunter et al., 2007]. Therefore, semiparametric approaches for estimating G_0 are usually not reliable.

Perhaps, the most common approach to avoid the identifiability issue of f_0 is to choose some kernel function f that we tactically believe the data are generated from, and utilize that kernel to fit the model to obtain an estimate of the mixing measure G_0 . In view of its simplicity and prevalence, this is also the approach that we consider in this chapter. However, it is likely that the chosen kernel f and the true kernel f_0 are different, i.e., we are under a misspecified kernel setting. Hence, the estimation of mixing measure G_0 under this approach may be highly unstable. The robustness question is unavoidable. Our principal goal in this chapter therefore, is the construction of robust estimators of G_0 where the estimation of both the number of components and the values of their parameters are of interest. Moreover, these estimators should

achieve the best possible convergence rates under various assumptions on both f and f_0 . When the true number of components k_0 is known, various robust methods have been proposed in the literature, see for example [Woodward et al., 1984, Donoho and Liu, 1988, Cutler and Cordero-Brana, 1996]. However, there is scarce work for robust estimators when the true number of components k_0 is unknown. Recently, Woo and Sriram [2006] proposed a robust estimator of the number of components based on the idea of minimum Hellinger distance estimator [Beran, 1977, Lindsay, 1994, Lin and He, 2006, Karunamuni and Wu, 2009]. However, their work faces certain limitations. Firstly, their estimator relied greatly upon the choice of bandwidth. In particular, in order to achieve the consistency of the number of components under the well-specified kernel setting, i.e., when $\{f\} = \{f_0\}$, the bandwidth should vanish to 0 sufficiently slowly (cf. Theorem 3.1 in Woo and Sriram [2006]). Secondly, the behaviors of parameter estimates from their estimators are hard to interpret due to the subtle choice of bandwidth. Last but not least, they also argued that their method achieved the robust estimation of the number of components under the misspecified kernel setting, i.e., when f and f_0 are different. Not only does their statement lack theoretical guarantee, their argument turns out to be also erroneous (see Section 5.3 in Woo and Sriram [2006]). More specifically, they considered the chosen kernel f to be Gaussian kernel and the true kernel f_0 to be Student's kernel with some fixed degree of freedom. The parameter space Θ consists of mean and scale parameter while the number of components $k_0 = 2$. They demonstrated that their estimator still maintained the correct number of components of G_0 , i.e., $k_0 = 2$, under that setting of f and f_0 . Unfortunately, their argument is not clear as their estimator should maintain the number of components of some mixing measure G_* which minimizes the appropriate Hellinger distance to the true model. Of course, establishing the consistency of their parameter estimators under the misspecified kernel setting is also a non-trivial problem.

Inspired by the idea of minimum Hellinger distance estimator, we propose flexible robust estimators of the mixing measure G_0 that address all the limitations of the estimator in [Woo and Sriram, 2006]. Not only our estimators are computationally feasible and robust but they also possess various desirable properties, such as the flexible choice of bandwidth, the consistency of the number of components, and the best possible convergence rates of the parameters. In particular, our main contributions in this chapter can be summarized as follows

- (i) We treat the well-specified kernel setting, i.e., $\{f\} = \{f_0\}$, and misspecified kernel setting, i.e., $\{f\} \neq \{f_0\}$, separately. Under both settings, we achieve the consistency of our estimators regarding the true number of components for any fixed bandwidth. Furthermore, when the bandwidth vanishes to 0 at appropriate rate, the consistency of estimating the true number of components is also guaranteed.
- (ii) For any fixed bandwidth, when f_0 is identifiable in the first order the optimal convergence rates $n^{-1/2}$ of parameter estimates are established under the well-specified kernel setting. Additionally, when f_0 is not identifiable in the first order, we demonstrate that our estimators still achieve the best possible convergence rates of parameter estimates.
- (iii) Under the misspecified kernel setting, we demonstrate that our estimators converge to some mixing measure G_* that is close to the true model under the Hellinger metric for any fixed bandwidth. When f is first order identifiable and G_* has finite number of components, the optimal convergence rates $n^{-1/2}$ are also established under mild conditions of both f and f_0 . Moreover, when G_* has infinite number of components, some analyses about the consistency of our estimators are also discussed.

Finally, our argument, so far, has mostly focused on the setting when the true mixing

measure G_0 is fixed with the sample size n . However, we note in passing that in a proper asymptotic model, G_0 may also vary with n and converge to some distribution in the limit. Under the well-specified kernel setting, we demonstrate that our estimators also achieve the minimax convergence rate of estimating G_0 under certain condition on the identifiability of kernel density function f_0 .

Chapter organization: The rest of the chapter is organized as follows. Section 6.2 provides preliminary backgrounds and facts. Section 5.3 presents an algorithm to construct a robust estimator of mixing measure based on model selection perspective. Theoretical results regarding that estimator are treated separately under both the well- and misspecified kernel setting. Section 5.4 introduces another algorithm to construct a robust estimator of mixing measure based on the idea of superefficiency. Section 5.5 addresses the performance of estimators developed in Section 5.3 and Section 5.4 under non-standard setting of kernel density function and true mixing measure. The theoretical results are illustrated via careful simulation studies with both synthetic and real data in Section 5.6. Discussions regarding possible future work are presented in Section 6.6 while self-contained proofs of key results are given in Section 5.8 and proofs of the remaining results are given in the Appendices.

Notation: Given two densities p, q (with respect to the Lebesgue measure μ), the total variation distance is given by $TV(p, q) = \frac{1}{2} \int |p(x) - q(x)| d\mu(x)$. Additionally, the square of Hellinger distance is given by $h^2(p, q) = \frac{1}{2} \int (\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x)$.

For any $\kappa = (\kappa_1, \dots, \kappa_{d_1}) \in \mathbb{N}^{d_1}$, we denote $\frac{\partial^{|\kappa|} f}{\partial \theta^\kappa}(x|\theta) = \frac{\partial^{|\kappa|} f}{\partial \theta_1^{\kappa_1} \dots \partial \theta_{d_1}^{\kappa_{d_1}}}(x|\theta)$ where $\theta = (\theta_1, \dots, \theta_{d_1})$. Additionally, the expression $a_n \gtrsim b_n$ will be used to denote the inequality up to a constant multiple where the value of the constant is independent of n and fixed within our setting. We also denote $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold.

5.2 Background

Throughout the chapter, we assume that the parameter space Θ is a compact subset of \mathbb{R}^{d_1} . For any kernel density function f and mixing measure G , we define $p_{G^f}(x) := \int f(x|\theta)dG(\theta)$. Additionally, we denote $\mathcal{E}_{k_0} := \mathcal{E}_{k_0}(\Theta)$ the space of discrete mixing measures with exactly k_0 distinct support points on Θ and $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ the space of discrete mixing measures with at most k distinct support points on Θ . Additionally, denote $\mathcal{G} := \mathcal{G}(\Theta) = \bigcup_{k \in \mathbb{N}_+} \mathcal{E}_k$ the set of all discrete measures with finite supports on Θ . Finally, $\bar{\mathcal{G}}$ denotes the space of all discrete measures (including those with countably infinite supports) on Θ .

As described in the introduction, a goal of our paper is to construct robust estimators that maintain the consistency of the number of components and the best possible convergence rates of parameter estimations. As in [Nguyen \[2013\]](#), our toolkit for analyzing the identifiability and convergence of parameters in a mixture model is based on the Wasserstein distances, which can be defined as the optimal cost of moving masses transforming one probability measure to another [[Villani, 2009](#)]. In particular, consider a mixing measure $G = \sum_{i=1}^k p_i \delta_{\theta_i}$, where $\mathbf{p} = (p_1, p_2, \dots, p_k)$ denotes the proportion vector. Likewise, let $G' = \sum_{i=1}^{k'} p'_i \delta_{\theta'_i}$. A coupling between \mathbf{p} and \mathbf{p}' is a joint distribution \mathbf{q} on $[1, \dots, k] \times [1, \dots, k']$, which is expressed as a matrix $\mathbf{q} = (q_{ij})_{1 \leq i \leq k, 1 \leq j \leq k'} \in [0, 1]^{k \times k'}$ with margins $\sum_{m=1}^k q_{mj} = p'_j$ and $\sum_{m=1}^{k'} q_{im} = p_i$ for any $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, k'$. We use $\mathcal{Q}(\mathbf{p}, \mathbf{p}')$ to denote the space of all such couplings. For any $r \geq 1$, the r -th order Wasserstein distance between G and G' is given by

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

where $\|\cdot\|$ denotes the l_2 norm for elements in \mathbb{R}^{d_1} . It is simple to argue that if a

sequence of probability measures $G_n \in \mathcal{E}_{k_0}$ converges to $G_0 \in \mathcal{E}_{k_0}$ under the W_r metric at a rate $\omega_n = o(1)$ then the set of atoms of G_n converges to the k_0 atoms of G_0 , up to a permutation of the atoms, at the same rate ω_n .

We recall now the following key definitions that are used to analyze the behavior of mixing measures in finite mixture models (cf. [Heinrich and Kahn, 2016+, Ho and Nguyen, 2016b]). We start with

Definition 5.2.1. *We say the family of densities $\{f(x|\theta), \theta \in \Theta\}$ is uniformly Lipschitz up to the order r , for some $r \geq 1$, if f as a function of θ is differentiable up to the order r and its partial derivatives with respect to θ satisfy the following inequality*

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

for any $\gamma \in \mathbb{R}^{d_1}$ and for some positive constant δ and C independent of x and $\theta_1, \theta_2 \in \Theta$. Here, $\gamma^\kappa = \prod_{i=1}^{d_1} \gamma_i^{\kappa_i}$ where $\kappa = (\kappa_1, \dots, \kappa_{d_1})$.

We can verify that many popular classes of density functions, including Gaussian, Student's t, and skew normal family, satisfy the uniform Lipschitz condition up to any order $r \geq 1$. Now, we have the following stronger notion of identifiability

Definition 5.2.2. *For any $r \geq 1$, we say that the family $\{f(x|\theta), \theta \in \Theta\}$ is identifiable in the r -th order if $f(x|\theta)$ is differentiable up to the r -th order in θ and the following holds*

A1. *For any $k \geq 1$, given k different elements $\theta_1, \dots, \theta_k \in \Theta$. If we have $\alpha_\kappa^{(i)}$ for $1 \leq i \leq k$, $\kappa \in \mathbb{N}^{d_1}$ and $|\kappa| \leq r$ such that for almost all x*

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

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

Rationale of the first order identifiability: Throughout the chapter, we denote $I(G, f) := E(l_G l_G^T)$ the Fisher information matrix of the kernel density f at the probability measure G . Here, $l_G := \frac{\partial}{\partial G} \log p_{G^f}(x)$ is the score function, where $\frac{\partial}{\partial G}$ means the derivatives with respect to all the components and masses of G . The first order identifiability of f is an equivalent way to say that the Fisher information matrix $I(G, f)$ is non-singular for any G . Now, under the first order identifiability and the first order uniform Lipschitz condition on f , a careful investigation of Theorem 3.1 and Corollary 3.1 in [Ho and Nguyen \[2016c\]](#) yields the following result

Proposition 5.2.1. *Suppose that the density family f is identifiable in the first order and uniformly Lipschitz up to the first order. Then, there is a positive constant C_0 depending on G_0 such that as long as $G \in \mathcal{O}_{k_0}$ we have*

$$h(p_{G^f}, p_{G_0^f}) \geq C_0 W_1(G, G_0).$$

Note that, the result of Proposition 5.2.1 is slightly stronger than that of Theorem 3.1 and Corollary 3.1 in [Ho and Nguyen \[2016c\]](#) as it holds for any $G \in \mathcal{O}_{k_0}$ instead of only for any $G \in \mathcal{E}_{k_0}$ as in these later results. The first order identifiability property of kernel density function f implies that any estimation method that yields the convergence rate $n^{-1/2}$ for $p_{G_0^f}$ under the Hellinger distance, the induced rate of convergence for the mixing measure G_0 is $n^{-1/2}$ under W_1 distance.

5.3 Minimum Hellinger distance estimator with non-singular Fisher information matrix

Throughout this section, we assume that two density families $\{f_0(x|\theta), \theta \in \Theta\}$ and $\{f(x|\theta), \theta \in \Theta\}$ are identifiable in the first order and admit uniform Lipschitz condition up to the first order. Now, let K be any fixed multivariate density function

and $K_\sigma(x) = \frac{1}{\sigma^d} K\left(\frac{x}{\sigma}\right)$ for any $\sigma > 0$. We define

$$f * K_\sigma(x|\theta) := \int f(x-y|\theta) K_\sigma(y) dy$$

for any $\theta \in \Theta$. The notation $f * K_\sigma$ can be thought as the convolution of the density family $\{f(x|\theta)\}$ with kernel function K_σ . From that definition, we have

$$p_{G^f} * K_\sigma = \sum_{i=1}^k p_i f * K_\sigma(x|\theta_i) = \sum_{i=1}^k p_i \int f(x-y|\theta_i) K_\sigma(y) dy$$

for any discrete measure $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ in $\bar{\mathcal{G}}$. Now, our approach to define a robust estimator of G_0 is inspired by the minimum Hellinger distance estimator [Beran, 1977] and the model selection criteria. Indeed, we have the following algorithm

Algorithm 1: Let $C_n n^{-1/2} \rightarrow 0$ and $C_n n^{1/2} \rightarrow \infty$ as $n \rightarrow \infty$.

- Step 1: Determine $\hat{G}_{n,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G^f} * K_\sigma, P_n * K_\sigma)$ for any $m \geq 1$.
- Step 2: Choose

$$\hat{m}_n = \inf \left\{ m \geq 1 : h(p_{\hat{G}_{n,m}^f} * K_\sigma, P_n * K_\sigma) \leq h(p_{\hat{G}_{n,m+1}^f} * K_\sigma, P_n * K_\sigma) + C_n n^{-1/2} \right\},$$

- Step 3: Let $\hat{G}_n = \hat{G}_{n,\hat{m}_n}$ for each n .

Note that, the choice of C_n is to guarantee that \hat{m}_n is finite. Additionally, it can be chosen based on certain model selection criterion. For instance, if we use BIC, then $C_n = \sqrt{(d_1 + 1)\log n / 2}$. The above algorithm is rather similar to the algorithm considered in Woo and Sriram [2006] except the fact that we take the convolution of p_{G^f} with K_σ in both Step 1 and Step 2. In fact, with the adaptation of notations as those in our paper, the algorithm in Woo and Sriram [2006] is as follows

Woo-Sriram (WS) Algorithm:

- Step 1: Determine $\bar{G}_{n,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G^f}, P_n * K_\sigma)$ for any $n, m \geq 1$.
- Step 2: Choose

$$\bar{m}_n = \inf \left\{ m \geq 1 : h(p_{\bar{G}_{n,m}^f}, P_n * K_\sigma) \leq h(p_{\bar{G}_{n,m+1}^f}, P_n * K_\sigma) + C'_n n^{-1/2} \right\},$$

where $C'_n n^{-1/2} \rightarrow 0$.

- Step 3: Let $\bar{G}_n = \bar{G}_{n,\bar{m}_n}$ for each n .

The convolution trick in Algorithm 1 was also considered in [James et al. \[2001\]](#) to construct the consistent estimation of mixture complexity. However, their work was based on the Kullback-Leibler (KL) divergence rather than the Hellinger distance. Under the misspecified kernel setting, i.e., $\{f\} \neq \{f_0\}$, the estimations of mixing measures G_0 from KL divergence may be unstable. Additionally, they only worked with true kernel function f_0 to be Gaussian, while in many applications, it is not realistic to expect that f_0 is Gaussian.

To demonstrate the advantages of our proposed estimator \hat{G}_n over Woo-Sriram's estimator \bar{G}_n , we will provide a careful theoretical study of both these estimators in the chapter. For readers' convenience, we provide now a summary of our later analyses of the convergence behaviors of \hat{G}_n and \bar{G}_n . Under the well-specified setting, i.e., $\{f\} = \{f_0\}$, the convolution trick in Algorithm 1 is crucial to guarantee the optimal rate $n^{-1/2}$ of \hat{G}_n to G_0 for any fixed bandwidth $\sigma > 0$. It comes from the fact that $P_n * K_\sigma(x)$ is the unbiased estimator of $p_{G_0} * K_\sigma(x)$ for all $x \in \mathcal{X}$. Hence, under suitable conditions of f_0 we can guarantee that $h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma) = O_p(n^{-1/2})$ when the bandwidth σ is fixed. However, it is not the case for WS Algorithm. Indeed, we demonstrate later in Section 5.3.3 that for any fixed bandwidth $\sigma > 0$, \bar{G}_n converges to \bar{G}_0 where $\bar{G}_0 = \arg \min_{G \in \bar{\mathcal{G}}} h(p_{G^{f_0}}, p_{G_0^{f_0}} * K_\sigma)$ under certain conditions of true kernel

f_0 , K and \bar{G}_0 (cf. Theorem 5.3.3). Unfortunately, \bar{G}_0 may be very different from G_0 even if they may have the same number of components. Therefore, even though we may recover the true number of components with WS algorithm, we hardly can obtain good estimates of parameter values. It shows that Algorithm 1 is more appealing than WS Algorithm under the well-specified kernel setting with fixed bandwidth σ .

When the bandwidth σ is allowed to vanish to 0 as $n \rightarrow \infty$, with the additional condition $n\sigma^d \rightarrow 0$, we are able to guarantee that $\hat{m}_n \rightarrow k_0$ almost surely (cf. Proposition 5.3.1). This result is also consistent with the result $\bar{m}_n \rightarrow k_0$ almost surely from Theorem 1 in [Woo and Sriram, 2006]. Moreover, under these conditions of bandwidth σ , both the estimators \hat{G}_n and \bar{G}_n converge to G_0 as $n \rightarrow \infty$. However, we can not establish the exact convergence rate $n^{-1/2}$ of \hat{G}_n to G_0 but only $n^{-1/2}$ up to some logarithmic factor under some choices of the bandwidth σ . It is due to the fact that our current technique is based on the evaluation of the term $h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$, which does not converge to 0 at the rate $n^{-1/2}$ when $\sigma \rightarrow 0$. The situation is similar for the convergence rates of \bar{G}_n as we also need to rely on studying $h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$.

Under the misspecified kernel setting, i.e., when chosen kernel f may be different from true kernel f_0 , the convolution trick continues to be useful for studying the convergence rate of \hat{G}_n to some G_* where $G_* = \arg \min_{G \in \bar{\mathcal{G}}} h(p_{G^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$. In fact, as G_* has finite number of components, under certain conditions on f , f_0 , and K , we are able to establish the convergence rate $n^{-1/2}$ of \hat{G}_n to G_* (cf. Theorem 5.3.2) for any fixed bandwidth $\sigma > 0$. When the number of components of G_* is infinite, we also report the consistency of the number of components of \hat{G}_n (cf. Proposition 5.3.2) under certain conditions of f and K . However, the convergence rates of \hat{G}_n to G_* are very complicated to establish. Therefore, we leave that scenario for future work.

5.3.1 Well-specified kernel setting

In this section, we consider the setting that f_0 is known, i.e., $\{f\} = \{f_0\}$. As we have seen from the discussion in Section 6.2, the first order identifiability condition plays an important role to obtain the convergence rate $n^{-1/2}$ of parameter estimations under mixture models. As Algorithm 1 relies on studying the variation around kernel function $f_0 * K_\sigma$ in the limit, we would like to guarantee that $f_0 * K_\sigma$ is identifiable in the first order for any $\sigma > 0$. Fortunately, we have a mild condition of K such that the first order identifiability of $f_0 * K_\sigma$ is maintained

Lemma 5.3.1. *Assume that $\widehat{K}(t) \neq 0$ for almost all $t \in \mathbb{R}^d$ where $\widehat{K}(t)$ is the Fourier transform of kernel function K . Then, as long as f_0 is identifiable in the first order, we obtain that $f_0 * K_\sigma$ is identifiable in the first order for any $\sigma > 0$.*

The assumption $\widehat{K}(t) \neq 0$ is very mild. Indeed, some popular choices of K to satisfy the condition of Lemma 5.3.1 are the Gaussian and Student's t kernel. Inspired by the result of Lemma 5.3.1, we have the following result establishing the convergence rates of $\widehat{G}_{n,\sigma}$ to G_0 under W_1 distance for any fixed bandwidth $\sigma > 0$

Theorem 5.3.1. *Let $\sigma > 0$ be given.*

(i) *If $f_0 * K_\sigma$ is identifiable, then $\widehat{m}_n \rightarrow k_0$ almost surely.*

(ii) *Assume further the following conditions*

(P.1) *The kernel K is chosen such that $f_0 * K_\sigma$ is also identifiable in the first order and admits a uniform Lipschitz property up to the first order.*

$$(P.2) \quad \Psi(G_0, \sigma) := \int \frac{g(x|G_0, \sigma)}{p_{G_0^{f_0}} * K_\sigma(x)} dx < \infty \text{ where } g(x|G_0, \sigma) := \int K_\sigma^2(x - y) p_{G_0^{f_0}}(y) dy.$$

Then, we have

$$W_1(\widehat{G}_n, G_0) = O_p\left(\sqrt{\frac{\Psi(G_0, \sigma)}{C_1^2(\sigma)} n^{-1/2}}\right)$$

where $C_1(\sigma) := \inf_{G \in \mathcal{O}_{k_0}} \frac{h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)}{W_1(G, G_0)}$.

Remarks:

- (i) Condition (P.1) is satisfied by many kernels K as the consequence of Lemma 5.3.1. By assumption (P.1) and Proposition 5.2.1, we obtain the following bound

$$h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \gtrsim W_1(G, G_0)$$

for any $G \in \mathcal{O}_{k_0}$, i.e., $C_1(\sigma) > 0$.

- (ii) Condition (P.2) is mild. One easy example for such setting is when f_0 and K are both Gaussian kernels. In fact, when f_0 and K are standard univariate Gaussian kernels, we achieve

$$\begin{aligned} \Psi(G_0, \sigma) &= \sum_{i=1}^{k_0} \int \frac{p_i^0 \int K_\sigma^2(x-y) f_0(y|\theta_i^0, \sigma_i^0) dy}{p_{G_0^{f_0}} * K_\sigma(x)} dx \\ &< \sum_{i=1}^{k_0} \int \frac{\int K_\sigma^2(x-y) f_0(y|\theta_i^0, \sigma_i^0) dy}{f_0 * K_\sigma(x|\theta_i^0)} dx \\ &\propto \sum_{i=1}^{k_0} ((\sigma_i^0)^2 + \sigma^2) / \sigma^2 < \infty. \end{aligned}$$

Another specific example is when f_0 and K are both Cauchy kernels or generally Student's t kernels with odd degree of freedom. However, assumption (P.2) may fail when K has much shorter tails than f_0 . For example if f_0 is Laplacian kernel and K is Gaussian kernel, then $\Psi(G_0, \sigma) = \infty$.

Comment on \hat{G}_n as $\sigma \rightarrow 0$: To avoid the ambiguity, we now denote $\{\sigma_n\}$ as the sequence of varied bandwidths σ . The following result shows the consistency of \hat{m}_n under specific conditions on $\sigma_n \rightarrow 0$

Proposition 5.3.1. *Given a sequence of bandwidths $\{\sigma_n\}$ such that $\sigma_n \rightarrow 0$ and $n\sigma_n^d \rightarrow \infty$ as $n \rightarrow \infty$. If f_0 is identifiable, then $\hat{m}_n \rightarrow k_0$ almost surely.*

Our result shows that if σ_n is small enough, the parametric $n^{1/2}$ rate of convergence of \hat{G}_n to G_0 is achieved. It would be more elegant to argue that this rate is achieved for some sequence $\sigma_n \rightarrow 0$. However, this cannot be done with the technique employed in the proof of Theorem 5.3.1. In particular, even though we still can guarantee that $\lim_{\sigma_n \rightarrow 0} C_1(\sigma_n) > 0$ (cf. Lemma 5.10.1 in Appendix B), the difficulty is that $\Psi(G_0, \sigma_n) = O(\sigma_n^{-\beta(d)})$ for some $\beta(d) > 0$ depending on d as $\sigma_n \rightarrow 0$. As a consequence, whatever the sequence of bandwidths $\sigma_n \rightarrow 0$ we choose, we will be only able to obtain the convergence rate $n^{-1/2}$ up to the logarithmic term of \hat{G}_n to G_0 . It can be thought as the limitation of the elegant technique employed in Theorem 5.3.1. We leave the exact convergence rate $n^{-1/2}$ of \hat{G}_n to G_0 under the setting $\sigma_n \rightarrow 0$ for the future work.

5.3.2 Misspecified kernel setting

In the previous section, we assume the well-specified kernel setting, i.e., $\{f\} = \{f_0\}$, and achieve the standard convergence rate $n^{-1/2}$ of \hat{G}_n to G_0 under mild conditions on f_0 and K for any fixed bandwidth $\sigma > 0$. However, the well-specified kernel assumption is often violated in practice, i.e., the chosen kernel f may be different from the true kernel f_0 . Motivated by this challenge, in this section we consider the setting when $\{f\} \neq \{f_0\}$ and demonstrate that the convergence rates of \hat{G}_n are still desirable under certain assumptions on f , f_0 , and K . Due to the complex nature of misspecified kernel setting, we will only study the behavior of \hat{G}_n when the bandwidth $\sigma > 0$ is fixed in this section. Now, for any $\sigma > 0$ assume that we can find

$$G_* = \arg \min_{G \in \bar{\mathcal{G}}} h(p_{G^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma).$$

With the above formulation, G_* is a discrete mixing measure that minimizes that Hellinger distance between $p_{G^f} * K_\sigma$ and $p_{G_0^{f0}} * K_\sigma$. As G_* may be not unique, we denote

$$\mathcal{M} = \left\{ G_* \in \bar{\mathcal{G}} : G_* = \arg \min_{G \in \bar{\mathcal{G}}} h(p_{G^f} * K_\sigma, p_{G_0^{f0}} * K_\sigma) \right\}.$$

That is, \mathcal{M} is the collection of G_* being the discrete mixing measure that minimizes the Hellinger distance between $p_{G^f} * K_\sigma$ and $p_{G_0^{f0}} * K_\sigma$. We start with the following key property of elements G_* in \mathcal{M}

Lemma 5.3.2. *For any $G \in \bar{\mathcal{G}}$ and $G_* \in \mathcal{M}$, it holds*

$$\int p_{G^f} * K_\sigma(x) \sqrt{\frac{p_{G_0^{f0}} * K_\sigma(x)}{p_{G_*^f} * K_\sigma(x)}} dx \leq \int \sqrt{p_{G_*^f} * K_\sigma(x)} \sqrt{p_{G_0^{f0}} * K_\sigma(x)} dx \quad (5.1)$$

With the above result, we have the following important property of \mathcal{M}

Lemma 5.3.3. *For any two elements $G_{1,*}, G_{2,*} \in \mathcal{M}$, we obtain $p_{G_{1,*}^f} * K_\sigma(x) = p_{G_{2,*}^f} * K_\sigma(x)$ for almost surely $x \in \mathcal{X}$.*

Now, we consider the partition of \mathcal{M} to the union of $\mathcal{M}_k = \{G_* \in \mathcal{M} : G_* \text{ has } k \text{ elements}\}$ where $k \in [1, \infty]$. Let $k_* := k_*(\mathcal{M})$ be the minimum number $k \in [1, \infty]$ such that \mathcal{M}_k is non-empty. Now we divide our argument into two distinct settings of k_* : k_* is finite and k_* is infinite.

5.3.2.1 Finite k_* :

By Lemma 5.3.3, \mathcal{M}_{k_*} will have exactly one element G_* as long as $f * K_\sigma$ is identifiable. Furthermore, \mathcal{M}_k is empty for all $k_* < k < \infty$. However, it is possible that \mathcal{M}_∞ still contains various elements. Fortunately, due to the parsimonious nature of Algorithm 1 and the result of Theorem 5.3.2, we will be able to demonstrate that \widehat{G}_n

still converges to the unique element $G_* \in \mathcal{M}_{k_*}$ at the optimal rate $n^{-1/2}$ regardless of the behavior of \mathcal{M}_∞ .

For the simplicity of our later argument under that setting of k_* , we denote G_* the unique element in \mathcal{M}_{k_*} . One simple example for $k_* < \infty$ is when f is location-scale family and f_0 is finite mixture of f . In particular, $f_0(x|\eta, \tau) = \frac{1}{\tau} f_0((x - \eta)/\tau)$ where η and τ are location and scale parameters respectively. Additionally, $f_0(x) = \sum_{i=1}^m p_i^* f(x|\eta_i^*, \tau_i^*)$ for some fixed positive integer m and fixed pairwise distinct components $(p_i^*, \eta_i^*, \tau_i^*)$ where $1 \leq i \leq m$. Under that setting, we can check that $k_* \leq mk_0$ and $p_{G_*^f}(x) = p_{G_0^{f_0}}(x)$ almost surely. The explicit formulation of G_* , therefore, can be found from the combinations of G_0 and $(p_i^*, \eta_i^*, \tau_i^*)$ where $1 \leq i \leq m$. From inequality (5.1) in Lemma 5.3.2, we have the following well-defined modification of Hellinger distance

Definition 5.3.1. *Given $\sigma > 0$. For any two mixing measures $G_1, G_2 \in \bar{\mathcal{G}}$, we define the distance $h^*(p_{G_1^f} * K_\sigma, p_{G_2^f} * K_\sigma)$ by*

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

The notatable feature of h^* is the involvement of term $\sqrt{p_{G_0^{f_0}} * K_\sigma(x)/p_{G_*^f} * K_\sigma(x)}$ in its formulation, which makes it slightly different from the traditional Hellinger distance. As long as $\{f\} \equiv \{f_0\}$, we obtain $h^*(p_{G_1^f} * K_\sigma, p_{G_2^f} * K_\sigma) \equiv h(p_{G_1^f} * K_\sigma, p_{G_2^f} * K_\sigma)$ for any $G_1, G_2 \in \bar{\mathcal{G}}$, i.e., the traditional Hellinger distance is a special case of h^* under the well-specified kernel setting. The modified version of Hellinger distance h^* is particularly useful for establishing the convergence rates of \hat{G}_n to G_* for any fixed $\sigma > 0$.

Note that, in the context of the well-specified kernel setting in Section 5.3.1 the key step we utilized to obtain the convergence rate $n^{-1/2}$ of \hat{G}_n to G_0 was based on the lower bound of the Hellinger distance and the first order Wasserstein distance in

inequality (5.1). With the modified Hellinger distance h^* , it turns out that we have the similar kind of lower bound as long as $k_* < \infty$.

Lemma 5.3.4. *Assume that $f * K_\sigma$ is identifiable in the first order and admits Lipschitz property up to the first order. If $k_* < \infty$, then for any $G \in \mathcal{O}_{k_*}$ there holds*

$$h^*(p_{G^f} * K_\sigma, p_{G_*^f} * K_\sigma) \gtrsim W_1(G, G_*).$$

Equipped with the above inequality, we have the following result regarding the convergence rate of \widehat{G}_n to G_* under the setting $k_* < \infty$

Theorem 5.3.2. *Assume $k_* < \infty$ for some $\sigma > 0$.*

(i) *If $f * K_\sigma$ is identifiable, then $\widehat{m}_n \rightarrow k_*$ almost surely.*

(ii) *Assume further that condition (P.2) in Theorem 5.3.1 holds, i.e., $\Psi(G_0, \sigma) < \infty$ and the following conditions hold:*

(M.1) *The kernel K is chosen such that $f * K_\sigma$ is identifiable in the first order and admits a uniform Lipschitz property up to the first order.*

(M.2) $\sup_{\theta \in \Theta} \int \sqrt{f * K_\sigma(x|\theta)} dx \leq M_1(\sigma)$ for some positive constant $M_1(\sigma)$.

(M.3) $\sup_{\theta \in \Theta} \left\| \frac{\partial f * K_\sigma}{\partial \theta}(x|\theta) / (f * K_\sigma(x|\theta))^{3/4} \right\|_\infty \leq M_2(\sigma)$ for some positive constant $M_2(\sigma)$.

Then, we have

$$W_1(\widehat{G}_n, G_*) = O_p \left(\sqrt{\frac{M^2(\sigma)\Psi(G_0, \sigma)}{C_{*,1}^4(\sigma)}} n^{-1/2} \right)$$

where $C_{*,1}(\sigma) := \inf_{G \in \mathcal{O}_{k_*}} \frac{h^*(p_{G^f} * K_\sigma, p_{G_*^f} * K_\sigma)}{W_1(G, G_*)}$ and $M(\sigma)$ is some positive constant.

Remarks:

- (i) As being mentioned in Lemma 5.3.4, condition (M.1) guarantees that $C_{*,1}(\sigma) > 0$.
- (ii) Conditions (M.2) and (M.3) are mild. An easy example is when f is Gaussian kernel and K is standard Gaussian kernel.
- (iii) When f_0 is indeed a finite mixture of f while both of them are location-scale kernels, a close investigation of the proof of this theorem reveals that we can relax condition (M.2) and (M.3) for the conclusion of this theorem to hold.

5.3.2.2 Infinite k_* :

So far, we have assumed that k_* has finite number of support points and achieve the cherished convergence rate $n^{-1/2}$ of \widehat{G}_n to unique element $G_* \in \mathcal{M}_{k_*}$ under certain conditions on f, f_0 , and K . It is due to the fact that $\widehat{m}_n \rightarrow k_* < \infty$ almost surely, which is eventually a consequence of the identifiability of kernel density function $f * K_\sigma$. However, for the setting $k_* = \infty$, to establish the consistency of \widehat{m}_n , we need to resort to a slightly stronger version of identifiability, which is finitely identifiable condition. We adapt Definition 3 in [Nguyen \[2013\]](#):

Definition 5.3.2. *The family $\{f(x|\theta), \theta \in \Theta\}$ is finitely identifiable if for any $G_1 \in \mathcal{G}$ and $G_2 \in \overline{\mathcal{G}}$, $|p_{G_1^f}(x) - p_{G_2^f}(x)| = 0$ for almost surely $x \in \mathcal{X}$ implies that $G_1 \equiv G_2$.*

An example of finite identifiability is when f is Gaussian kernel with both location and variance parameter. Now, a close investigation of Step 1 in the proof of Theorem 5.3.2 quickly yields the following result

Proposition 5.3.2. *Given $\sigma > 0$ such that $f * K_\sigma$ is finitely identifiable. If $k_* = \infty$, we achieve $\widehat{m}_n \rightarrow \infty$ almost surely.*

Even though we achieve the consistency result of \hat{m}_n when $k_* = \infty$, the convergence rate of \hat{G}_n to G_* still remains an elusive problem. However, an important insight from Proposition 5.3.2 indicates that the convergence rate of \hat{G}_n to some element $G_* \in \mathcal{M}_\infty$ may be much slower than $n^{-1/2}$ when $k_* = \infty$. It is due to the fact that both \hat{G}_n and $G_* \in \mathcal{M}_\infty$ have unbounded numbers of components in which the kind of bound in Lemma 5.3.4 is no longer sufficient. We leave the detail analyses of \hat{G}_n under that setting of k_* for the future work.

5.3.3 Analysis of WS Algorithm

So far, we have focused on studying the behaviors of \hat{G}_n in Algorithm 1, i.e., we established the consistency of the number of components of \hat{G}_n as well as the convergence rates of parameter estimates of \hat{G}_n under various settings of f and f_0 when the bandwidth σ is fixed. As we mentioned at the beginning of Section 3, we also would like to demonstrate the flexibilities and advantages of our estimator \hat{G}_n over Woo-Sriram's estimator \bar{G}_n in WS algorithm. As a consequence, in this section we also provide a careful analysis for the estimators \bar{G}_n from WS Algorithm under the fixed bandwidth setting. For the simplicity of our argument, we only consider the well-specified kernel setting, i.e., $\{f\} = \{f_0\}$. Rememeber that f_0 is identifiable in the first order and has uniform Lipschitz up to the first order. Assume now we can find

$$\bar{G}_0 = \arg \min_{G \in \bar{\mathcal{G}}} h(p_{G^{f_0}}, p_{G_0^{f_0}} * K_\sigma),$$

i.e., \bar{G}_0 is the discrete mixing measure that minimizes that Hellinger distance between $p_{G^{f_0}}$ and $p_{G_0^{f_0}} * K_\sigma$. Similar to Lemma 5.3.2, we also have the following property

characterizing \bar{G}_0 :

$$\int p_{G^f}(x) \sqrt{\frac{p_{G_0^{f_0}} * K_\sigma(x)}{p_{\bar{G}_0^{f_0}}(x)}} dx \leq \int \sqrt{p_{\bar{G}_0^{f_0}}(x)} \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} dx \quad (5.2)$$

for any $G \in \bar{\mathcal{G}}$. As being argued in Section 5.3.2, \bar{G}_0 may either have infinite number of components or unique; however, for the sake of simplicity, we assume in this section that there exists \bar{G}_0 having finite number of components. Using the same argument as in the proof of Lemma 5.3.3, we can treat \bar{G}_0 as unique mixing measure with finite number of components.

We denote k_0 the number of components of \bar{G}_0 . Fortunately, the form of \bar{G}_0 can be determined explicitly under various settings of f_0 and K . For instance, assume that f_0 are either univariate Gaussian kernels or Cauchy kernels with parameters $\theta = (\eta, \tau)$ where η and τ are location and variance parameter respectively and K are either standard univariate Gaussian kernels or Cauchy kernels respectively. Then, a simple calculation shows that $\bar{k}_0 = k_0$ and $\bar{G}_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, \bar{\tau}_i^0)}$ where $\bar{\tau}_i^0 = \sqrt{(\tau_i^0)^2 + \sigma^2}$ for any $1 \leq i \leq k_0$ and $\sigma > 0$.

From inequality (5.2), we have the following well-defined modification of Hellinger distance

Definition 5.3.3. Given $\sigma > 0$. For any two mixing measures $G_1, G_2 \in \bar{\mathcal{G}}$, we define the distance between $p_{G_1^{f_0}}$ and $p_{G_2^{f_0}}$ as

$$\left(\bar{h}(p_{G_1^{f_0}}, p_{G_2^{f_0}}) \right)^2 = \frac{1}{2} \int \left(\sqrt{p_{G_1^{f_0}}(x)} - \sqrt{p_{G_2^{f_0}}(x)} \right)^2 \sqrt{\frac{p_{G_0^{f_0}} * K_\sigma(x)}{p_{\bar{G}_0^{f_0}}(x)}} dx$$

As f_0 is identifiable in the first order and admits uniform Lipschitz condition up to the first order, we obtain that

$$\bar{h}(p_{G^{f_0}}, p_{\bar{G}_0^{f_0}}) \gtrsim W_1(G, \bar{G}_0) \quad (5.3)$$

for any $G \in \mathcal{O}_{\bar{k}_0}$. The proof of this bound is omitted since it is similar to that of Lemma 5.3.4. The above inequality leads to the following result concerning the convergence rate of \bar{G}_n to \bar{G}_0 when $\bar{k}_0 < \infty$

Theorem 5.3.3. *Given $\sigma > 0$. Assume that $\bar{k}_0 < \infty$.*

- (i) *If $f_0 * K_\sigma$ is identifiable, then $\bar{m}_n \rightarrow \bar{k}_0$ almost surely.*
- (ii) *Assume further that condition (P.2) in Theorem 5.3.1 holds, i.e., $\Psi(G_0, \sigma) < \infty$. Additionally, we have the following conditions*

$$(S.1) \quad \sup_{\theta \in \Theta} \int \sqrt{f_0(x|\theta)} dx \leq \bar{M}_1 \text{ for some positive constant } \bar{M}_1.$$

$$(S.2) \quad \sup_{\theta \in \Theta} \left\| \frac{\partial f_0}{\partial \theta}(x|\theta) / (f_0(x|\theta))^{3/4} \right\|_\infty \leq \bar{M}_2 \text{ for some positive constant } \bar{M}_2.$$

Then, we obtain

$$W_1(\bar{G}_n, \bar{G}_0) = O_p \left(\sqrt{\frac{[\bar{M}(\sigma)]^2 \Psi(G_0, \sigma)}{[\bar{C}(\sigma)]^4}} n^{-1/2} \right)$$

where $\bar{C}(\sigma) := \inf_{G \in \mathcal{O}_{\bar{k}_0}} \frac{\bar{h}(p_{G^{f_0}}, p_{\bar{G}_0^{f_0}})}{W_1(G, \bar{G}_0)}$ and $\bar{M}(\sigma)$ is some positive constant.

Condition (S.1) and (S.2) are indeed very mild as it is satisfied by many popular kernels, such as Gaussian, Laplacian, and Student's t with degree of freedom greater than 3. As being indicated in Theorem 5.3.3, the estimators \bar{G}_n from WS algorithm will not converge to the true mixing measure G_0 for any fixed bandwidth σ . It demonstrates that Algorithm 1 is more appealing than WS algorithm under this setting. For the setting when the bandwidth is allowed to vanish to 0, our current approach in the proof of Theorem 5.3.3 is not sufficient to determine whether it is possible to achieve the convergence rate $n^{-1/2}$ of WS's estimator \bar{G}_n to G_0 . Similar to the remark after Proposition 5.3.1, one of the main difficulties with our current technique is that the term $\Psi(G_0, \sigma) = O(\sigma^{-\beta(d)})$ for some $\beta(d) > 0$ depending on d . Another difficulty

is associated with the term $\overline{M}(\sigma)$ as it may also converge to 0 as $\sigma \rightarrow 0$. We leave this intriguing question for the future work.

5.4 Different approach with minimum Hellinger distance estimator

From the previous section, we develop a robust estimator of mixing measure G_0 based on the idea of minimum Hellinger distance estimator and model selection criteria. That estimator is shown to possess various desirable properties, including the consistency of number of components \hat{m}_n and the optimal convergence rates of \hat{G}_n . In this section, we take a rather different approach of constructing such robust estimator. In fact, we have the following algorithm

Algorithm 2:

- Step 1: Determine $\hat{G}_{n,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G^f} * K_\sigma, P_n * K_\sigma)$ for any $n, m \geq 1$.
- Step 2: Choose

$$\tilde{m}_n = \inf \left\{ m \geq 1 : h(p_{\hat{G}_{n,m}^f} * K_\sigma, P_n * K_\sigma) < \epsilon \right\},$$

where $\epsilon > 0$ is any given positive constant.

- Step 3: Let $\tilde{G}_n = \hat{G}_{n,\tilde{m}_n}$ for each n .

Unlike Step 2 in Algorithm 1 where we consider the difference $h(p_{\hat{G}_{n,m}^f} * K_\sigma, P_n * K_\sigma) - h(p_{\hat{G}_{n,m+1}^f} * K_\sigma, P_n * K_\sigma)$, here we consider solely $h(p_{\hat{G}_{n,m}^f} * K_\sigma, P_n * K_\sigma)$. The above robust estimator of mixing measure is based on the idea of minimum Hellinger distance estimator and sufficiency phenomenon. The superefficiency idea was also considered in [Heinrich and Kahn, 2016+]; however, their construction was based on minimum distance estimator without the convolution kernel K_σ and the threshold ϵ

was allowed to go to 0 as $n \rightarrow \infty$. Needless to mention, minimum distance estimator is neither computationally simple nor robust.

Our focus with Algorithm 2 in this section will be mainly about its attractive theoretical performance. As we observe from Algorithm 2, the choices of f , f_0 , and G_0 play crucial roles in determining the convergence rate of \tilde{G}_n to G_0 for any $\epsilon > 0$. Similar to the argument of Theorem 5.3.1 and Theorem 5.3.2, one of the key ingredients to fulfill that goal is to find the conditions of f , f_0 , and G_0 such that we obtain the consistency of \tilde{m}_n , i.e., $\tilde{m}_n \rightarrow k_0$ under the well-specified kernel setting or $\tilde{m}_n \rightarrow k_*$ under the misspecified kernel setting where k_* is defined as in Section 5.3.2. The following proposition yields the sufficient and necessary conditions to answer that consistency question

Theorem 5.4.1. *For any $\sigma > 0$, we have*

- Under the well-specified kernel setting, $\tilde{m}_n \rightarrow k_0$ almost surely if and only if

$$\epsilon < h(p_{G_{0,k_0-1}^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \quad (5.4)$$

where $G_{0,k_0-1} = \arg \min_{G \in \mathcal{E}_{k_0-1}} h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$.

- Under the misspecified kernel setting, if $k_* < \infty$, then $\tilde{m}_n \rightarrow k_*$ almost surely if and only if

$$h(p_{G_*^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \leq \epsilon < h(p_{G_{*,k_*-1}^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \quad (5.5)$$

where $G_{*,k_*-1} = \arg \min_{G \in \mathcal{E}_{k_*-1}} h(p_{G^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$ and $G_* \in \mathcal{M}$ with exactly k_* components.

If we allow $\epsilon \rightarrow 0$ in Algorithm 2, we achieve the inconsistency of \tilde{m}_n under the misspecified kernel setting when $k_* < \infty$. Hence, the choice of threshold ϵ from Heinrich and Kahn [2016+] is not optimal regarding the misspecified kernel setting.

Unfortunately, the conditions (5.4) and (5.5) are rather cryptic as in general, it is hard to determine the exact formulation of G_{0,k_0-1} , G_{*,k_*-1} , and G_* . Thus, we would like to have simple conditions on f , f_0 , and G_0 . Under the well-specified setting, it appears that condition (5.4) can be recasted as condition regarding the lower bound on the smallest mass of G_0 and the minimal distance between its point masses as follows

Proposition 5.4.1. (Well-specified kernel setting) *For any given $\sigma > 0$, assume that $f_0 * K_\sigma$ is uniformly Lipschitz up to the first order and identifiable. Then, there exists a positive constant C depending on f_0, G_0, K, Θ and σ such that if we have*

$$\min_{1 \leq i \leq k_0} p_i^0 \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \geq C\epsilon, \quad (5.6)$$

then we obtain the inequality in (5.4).

Unlike (5.4), it is tricky to derive simple conditions for (5.5) to hold under the misspecified kernel setting due to the wide range of possibility of f . Before arriving at these conditions, we define the following norm $\|\cdot\|$ between any two classes of density functions $\{f_1(x|\theta) : \theta \in \Theta\}$ and $\{f_2(x|\theta) : \theta \in \Theta\}$ as follows

$$\|f_1 - f_2\| = \sup_{\theta \in \Theta} \int |f_1(x|\theta) - f_2(x|\theta)| dx.$$

When $\|f_1 - f_2\| = 0$, for each $\theta \in \Theta$ we have $f_1(x|\theta) = f_2(x|\theta)$ for almost surely $x \in \mathcal{X}$. Now, we have the following definition regarding the *distinguishability* of any two classes of density functions $\{f_1(x|\theta)\}$ and $\{f_2(x|\theta)\}$

Definition 5.4.1. *Given any two classes of density functions $\{f_i(x|\theta), \theta \in \Theta\}$ where $1 \leq i \leq 2$. We say that f_1 and f_2 are **distinguishable** if we have $h(p_{G_1^{f_1}}, p_{G_2^{f_2}}) > 0$ for any finite discrete mixing measures G_1, G_2 in Θ .*

Example 5.4.1. If f_1 is location-scale univariate Gaussian family and f_2 is location-scale univariate Student's t family with fixed degree of freedom $\nu > 1$, then f_1 and f_2 are distinguishable.

The proof for this example is deferred to Appendix B. Equipped with all the above definitions, we have the following sufficient condition regarding the inequality in (5.5)

Proposition 5.4.2. (Misspecified kernel setting) Given $\sigma > 0$. Assume that K is chosen such that $f_0 * K_\sigma$ and $f * K_\sigma$ are distinguishable and are uniformly Lipschitz up to the first order. If $k_* \leq k_0$, there exists a positive constant C_1 depending only on f, f_0, G_0, K, Θ , and σ such that as long as $\|f - f_0\| \leq 2\epsilon^2$ and

$$\min_{1 \leq i \leq k_0} p_i^0 \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \geq C_1 \epsilon,$$

we obtain the inequalities in (5.5).

Remarks:

- (i) The distinguishability of $f_0 * K_\sigma$ and $f * K_\sigma$ is needed for the following lower bound (cf. Lemma 5.9.2 in Section 5.8)

$$h(p_{G^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \geq C'_1 W_1(G, G_0)$$

for any $G \in \mathcal{E}_{k_*-1}$ where C'_1 is some positive constant depending only on f, f_0, G_0, Θ , and σ .

- (ii) The assumption $k_* \leq k_0$ is purely for the argument of the proposition to go through. Indeed, from the lower bound in part (i), our proof relies on the evaluation of the quantity $\inf_{G \in \mathcal{E}_{k_*-1}} W_1(G, G_0)$ as in the proof of Proposition 5.4.1. As $k_* > k_0$, this quantity becomes zero, which is not informative to further lower bound the right hand side of the inequality in part (i). Therefore, the current

approach employed in our proof of Proposition 5.4.2 is not effective to deal with the setting $k_* > k_0$.

5.5 Extension to non-standard settings

In this section, we briefly demonstrate that the robust estimator from Algorithm 1 (similarly Algorithm 2) also achieves the most desirable convergence rates under non-standard settings. In particular, we consider first the situation where f_0 or f may be not identifiable in the first order. In the second setting, the true mixing measure G_0 changes with the sample size n and converges to some discrete distribution \tilde{G}_0 under W_1 distance.

5.5.1 A singular Fisher information matrix

The results in the previous sections are under the assumption that both the true kernel f_0 and chosen kernel f are identifiable in the first order. This is equivalent to the non-singularity of the Fisher information matrix of $p_{G_0^{f_0}}$ and $p_{G_*^f}$ when $G_* \in \mathcal{M}$, i.e., both $I(G_0, f_0)$ and $I(G_*, f)$ are non-singular. Therefore, we achieve the cherished convergence rate $n^{-1/2}$ of \hat{G}_n . Unfortunately, these assumptions do not always hold. For instance, both the Gamma and skew normal kernel are not identifiable in the first order [Ho and Nguyen, 2016a,b]. According to Azzalini and Valle [1996], Wiper et al. [2001], these kernels are particularly useful for modelling various kinds of data: the Gamma kernel is used for modeling non-negative valued data and the skew normal kernel is prevalently used to model asymmetric data. Therefore, it is worth considering the performance of Algorithm 1 under the nonidentifiability in the first order of both kernels f_0 and f . Throughout this section, for the simplicity of the argument we consider only the well-specified kernel setting and the setting that f_0 may be not identifiable in the first order. The argument for the misspecified kernel setting and not first order identifiability of both f or f_0 can be argued in the similar fashion.

The non-identifiability in the first order of f_0 implies that the Fisher information matrix $I(G_0, f_0)$ of $p_{G_0^{f_0}}$ is singular at some particular values of G_0 . Therefore, the convergence rate of \widehat{G}_n to G_0 will be much slower than the standard convergence rate $n^{-1/2}$. In order to precisely determine the convergence rates of parameters under the singular Fisher information matrix setting, [Ho and Nguyen \[2016b\]](#) introduced the notion of *singularity levels* of the Fisher information matrix $I(G_0, f_0)$ (cf. Definition 3.1 and Definition 3.3). Here, we adapt the definitions of singularity levels according to the notations in our paper for the convenience of readers. In particular, we say that G_0 is r -singular relative to the ambient space \mathcal{O}_{k_0} and the kernel f_0 as long as $I(G_0, f_0)$ admits r -th level of singularity level for $0 \leq r < \infty$, i.e., we have

$$\inf_{G \in \mathcal{O}_{k_0}} TV(p_{G^{f_0}}, p_{G_0^{f_0}})/W_s^s(G, G_0) = 0, \quad s = 1, \dots, r.$$

$$TV(p_{G^{f_0}}, p_{G_0^{f_0}}) \gtrsim W_{r+1}^{r+1}(G, G_0), \quad \text{for all } G \in \mathcal{O}_{k_0}. \quad (5.7)$$

The infinite singularity level of the Fisher information matrix $I(G_0, f_0)$ implies that inequality (5.7) will not hold for any $r \geq 0$.

When f_0 is identifiable in the first order, $I(G_0, f_0)$ will only have zero order singularity level for all $G_0 \in \mathcal{E}_{k_0}$, i.e., $r = 0$ in (5.7). However, the singularity levels of the Fisher information matrix $I(G_0, f_0)$ are generally not uniform over G_0 when $I(G_0, f_0)$ is singular. For example, when f_0 is skew normal kernel, $I(G_0, f_0)$ will admit any order of singularity levels, ranging from 0 to ∞ depending on the interaction of atoms and masses of G_0 [[Ho and Nguyen, 2016b](#)]. The notion of singularity level allows us to establish the convergence rates of any estimator of G_0 immediately. In fact, if $r < \infty$ is the singularity level of $I(G_0, f_0)$, for any estimation method that yields the convergence rate $n^{-1/2}$ for $p_{G_0^{f_0}}$ under the Hellinger distance, the induced best possible rate of convergence for the mixing measure G_0 is $n^{-1/2(r+1)}$ under W_{r+1} distance. If $r = \infty$ is the singularity level of $I(G_0, f_0)$, all the estimation methods

will yield the non-polynomial convergence rate of G_0 , i.e., not of the form $n^{-1/2s}$ for any $s \geq 1$.

Now, by using the same line of argument as that of Theorem 5.3.1 we have the following result regarding the convergence rate of \widehat{G}_n to G_0 when the Fisher information matrix $I(G_0, f_0)$ has r -th singularity level

Proposition 5.5.1. *Given $\sigma > 0$. Assume that the Fisher information $I(G_0, f_0)$ has r -th singularity level where $r < \infty$ and that condition (P.2) in Theorem 5.3.1 holds, i.e., $\Psi(G_0, \sigma) < \infty$. Furthermore, the kernel K is chosen such that the Fisher information matrix $I(G_0, f_0 * K_\sigma)$ has r -th singularity level and $f_0 * K_\sigma$ admits a uniform Lipschitz property up to the r -th order. Then, we have*

$$W_{r+1}(\widehat{G}_n, G_0) = O_p\left(\sqrt{\frac{\Psi(G_0, \sigma)}{C_r^2(\sigma)}} n^{-1/2(r+1)}\right)$$

$$\text{where } C_r(\sigma) = \inf_{G \in \mathcal{O}_{k_0}} \frac{h(p_{Gf_0} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)}{W_{r+1}^{r+1}(G, G_0)}.$$

Remarks:

- (i) A mild condition such that $I(G_0, f_0)$ and $I(G_0, f_0 * K_\sigma)$ have the same singularity level is $\widehat{K}(t) \neq 0$ for all $t \in \mathbb{R}^d$ where $\widehat{K}(t)$ denotes the Fourier transformation of K (cf. Lemma 5.10.2 in the Appendix B).
- (ii) Some examples of f_0 that are not identifiable in the first order and satisfy $\Psi(G_0, \sigma) < \infty$ are skew normal and exponential kernel while K is chosen to be Gaussian or exponential kernel respectively.
- (iii) The result of Proposition 5.5.1 implies that under suitable choice of kernel K , Algorithm 1 still achieves the best possible convergence rate for estimating G_0 even when the Fisher information matrix $I(G_0, f_0)$ is singular.

5.5.2 Extension to varying true parameters

So far, our analysis has relied upon the assumption that G_0 is fixed as $n \rightarrow \infty$. However, there are situations where in proper asymptotic models the true mixing measure G_0 also varies with n and converges to some distribution \tilde{G}_0 under W_1 distance as $n \rightarrow \infty$. In this section, we will demonstrate that the estimator in Algorithm 1 still achieves the optimal convergence rate.

Denote the number of components of \tilde{G}_0 by \tilde{k}_0 . For the clarity of our argument we only work with the well-specified kernel setting and with the setting that f_0 is identifiable in the first order. As we have seen from the analysis of Section 5.3.1, when G_0 does not change with n , the key steps used to establish the standard convergence rate $n^{-1/2}$ of \hat{G}_n to G_0 are through the combination of the convergence of \hat{m}_n to k_0 almost surely and, under the first order identifiability of $f_0 * K_\sigma$, the lower bound

$$h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \gtrsim W_1(G, G_0) \quad (5.8)$$

for any $G \in \mathcal{O}_{k_0}$. Unfortunately, these two results no longer hold as G_0 varies with n . In this section, to avoid the ambiguity of the later argument, we denote by G_0^n the true mixing distribution when the sample size is n . Similarly, let k_0^n be the number of components of G_0^n . Assume that $\limsup_{n \rightarrow \infty} k_0^n = k < \infty$. We start with the following result regarding the convergence rate of \hat{m}_n under that setting of G_0^n

Proposition 5.5.2. *Given $\sigma > 0$. If $f_0 * K_\sigma$ is identifiable, then $|\hat{m}_n - k_0^n| \rightarrow 0$ almost surely as $n \rightarrow \infty$.*

According to the above proposition, \hat{m}_n will not converge to \tilde{k}_0 almost surely when $k > \tilde{k}_0$. Additionally, from that proposition, inequality (5.8) no longer holds since both the number of components of \hat{G}_n and G_0^n vary. In fact, we need to impose a much stronger condition on the identifiability of $f_0 * K_\sigma$.

Throughout the rest of this section, we assume that $d = d_1 = 1$, i.e., we work with the univariate setting of f_0 . Equipped with the result of Theorem 4.6 in [Heinrich and Kahn \[2016+\]](#), we have the following bound

Proposition 5.5.3. *Given $\sigma > 0$. Let K be chosen such that $f_0 * K_\sigma$ is identifiable up to the $(2k - 2\tilde{k}_0)$ -order and admits a uniform Lipschitz condition up to $(2k - 2\tilde{k}_0)$ -order. Then, there exist $\epsilon_0 > 0$ and $N(\epsilon_0) \in \mathbb{N}$ such that*

$$h(p_{G^{f_0}} * K_\sigma, p_{G_0^n, f_0} * K_\sigma) \geq C_v(\sigma) W_1^{2k-2\tilde{k}_0+1}(G, G_0^n) \quad (5.9)$$

for any $n \geq N(\epsilon_0)$ and for any $G \in \mathcal{O}_{k_0^n}$ such that $W_1(G, \tilde{G}_0) \leq \epsilon_0$. Here, $C_v(\sigma)$ is some positive constant depending only on \tilde{G}_0 and σ .

Similar to the argument of Lemma 5.3.1, an easy example of K and f_0 for the assumptions of Proposition 5.5.3 to hold is $\hat{K}(t) \neq 0$ for all $t \in \mathbb{R}^d$ and f_0 is strongly identifiable up to the $(2k - 2\tilde{k}_0)$ -order, which is satisfied by location family of density functions (cf. Theorem 2.4 in [\[Heinrich and Kahn, 2016+\]](#)). Now, a combination of Proposition 5.5.2 and Proposition 5.5.3 yields the following result regarding the convergence rate of \hat{G}_n to G_0^n

Corollary 5.5.1. *Given the assumptions in Proposition 5.5.3. Assume that $\Psi(G_0^n, \sigma) < \infty$ for all $n \geq 1$. Then, we have*

$$W_1(\hat{G}_n, G_0^n) = O_p\left(\sqrt{\frac{\Psi(G_0^n, \sigma)}{C_v^2(\sigma)}} n^{-1/(4k-4\tilde{k}_0+2)}\right)$$

where $C_v(\sigma)$ is the constant in inequality (5.9).

Remark:

- (i) As $k = \tilde{k}_0$, we recover the result in Theorem 5.3.1.

- (ii) If f_0 is univariate Gaussian or Cauchy family and K is standard Gaussian or Cauchy kernel respectively, then $\Psi(G_0^n, \sigma) \rightarrow \Psi(\tilde{G}_0, \sigma)$ as $n \rightarrow \infty$.
- (iii) If $W_1(G_0^n, \tilde{G}_0) = O(n^{-1/(4k-4\bar{k}_0+2)+\kappa})$ for some $\kappa > 0$, then the convergence rate $n^{-1/(4k-4\bar{k}_0+2)}$ of \hat{G}_n to G_0^n is minimax (cf. Theorem 3.2 in [Heinrich and Kahn, 2016+]). Therefore, Algorithm 1 also achieves the minimax rate of convergence for estimating G_0^n , which is consistent with the minimax rate of convergence of minimum distance estimator in Heinrich and Kahn [2016+]. However, our estimator from Algorithm 1 is more appealing than that from Heinrich and Kahn [2016+] as it is computationally feasible and robust while the minimum distance estimator is neither computationally feasible nor robust. We will illustrate the result of Corollary 5.5.1 in the simulation studies in Section 5.6.

5.6 Empirical studies

We present in this section numerous numerical studies to validate our theoretical results in the previous sections. To find the mixing measure $\hat{G}_{n,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G^f} * K_\sigma, P_n * K_\sigma)$, we utilize the HMIX algorithm developed in Section 4.1 of [Cutler and Cordero-Brana, 1996]. This algorithm is essentially similar to the EM algorithm and ultimately gives us local solutions to the minimization problem.

5.6.1 Synthetic data

We start with testing Algorithm 1 using synthetic data. The discussion is divided into separate enquiries of the well- and mis-specified kernel setups.

Well-specified kernel setting Under this setting, we assess the performance of estimator in Algorithm 1 under two cases of G_0 :

Case 1: G_0 is fixed with the sample size. Under this case, we consider three choices of f_0 : Gaussian and Cauchy distribution for the first order identifiability, and skew

normal kernel for the fail of first order identifiability.

- Case 1.1 - Gaussian family:

$$\begin{aligned} f_0(x|\eta, \tau) &= \frac{1}{\sqrt{2\pi}\tau} \exp\left(-\frac{(x-\eta)^2}{2\tau^2}\right) \\ G_0 &= \frac{1}{2}\delta_{(0,\sqrt{10})} + \frac{1}{4}\delta_{(-0.3,\sqrt{0.05})} + \frac{1}{4}\delta_{(0.3,\sqrt{0.05})}. \end{aligned}$$

- Case 1.2 - Cauchy family:

$$\begin{aligned} f_0(x|\eta, \tau) &= \frac{1}{\pi\tau(1+(x-\eta)^2/\tau^2)} \\ G_0 &= \frac{1}{2}\delta_{(0,\sqrt{10})} + \frac{1}{4}\delta_{(-0.3,\sqrt{0.05})} + \frac{1}{4}\delta_{(0.3,\sqrt{0.05})}. \end{aligned}$$

- Case 1.3 - Skew normal family:

$$\begin{aligned} f_0(x|\eta, \tau, m) &= \frac{2}{\sqrt{2\pi}\tau} \exp\left(-\frac{(x-\eta)^2}{2\tau^2}\right) \Phi(m(x-\eta)/\tau) \\ G_0 &= \frac{1}{2}\delta_{(0,\sqrt{10},0)} + \frac{1}{4}\delta_{(-0.3,\sqrt{0.05},0)} + \frac{1}{4}\delta_{(0.3,\sqrt{0.05},0)}. \end{aligned}$$

where Φ is the cumulative function of standard normal distribution.

For the Gaussian case and skew normal case of f_0 , we choose K to be the standard Gaussian kernel while K is chosen to be the standard Cauchy kernel for the Cauchy case of f_0 . Note that, regarding skew normal case it was shown that the Fisher information matrix $I(G_0, f_0)$ has second level singularity (cf. Theorem 5.3 in [Ho and Nguyen, 2016b]); therefore, from the result of Proposition 5.5.1, the convergence rate of \widehat{G}_n to G_0 will be at most $n^{-1/6}$. Now for the bandwidth, we choose $\sigma = 1$. The sample sizes will be $n = 200 * i$ where $1 \leq i \leq 20$. The tuning parameter C_n is chosen according to BIC criterion. More specifically, $C_n = \sqrt{3 \log n} / \sqrt{2}$ for Gaussian and Cauchy family while $C_n = \sqrt{2 \log n}$ for skew normal family. For each sample size n ,

we perform Algorithm 1 exactly 100 times and then choose \hat{m}_n to be the estimated number of components with the highest probability of appearing. Afterwards, we take the average among all the replications with the estimated number of components \hat{m}_n to obtain $W_1(\hat{G}_n, G_0)$. See Figure 5.1 where the Wasserstein distances $W_1(\hat{G}_n, G_0)$ and the percentage of time $\hat{m}_n = 3$ are plotted against increasing sample size n along with the error bars. The simulation results regarding Gaussian and Cauchy family match well with the standard $n^{-1/2}$ convergence rate from Theorem 5.3.1 while the simulation results regarding skew normal family also fit with the best possible convergence rate $n^{-1/6}$ as we argued earlier.

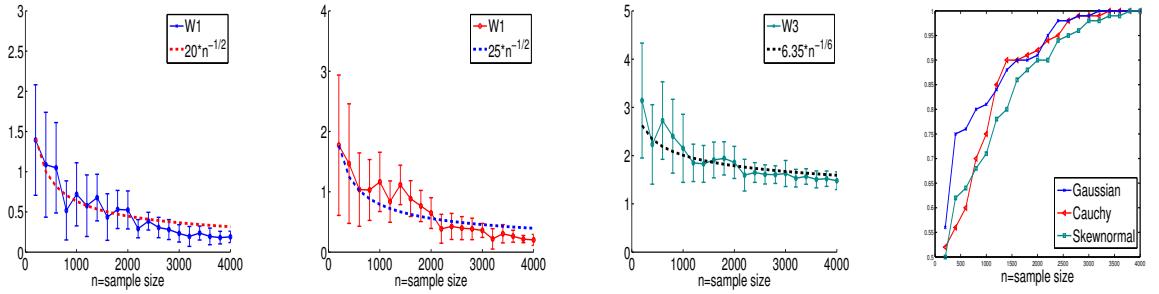


Figure 5.1: Performance of \hat{G}_n in Algorithm 1 under the well-specified kernel setting and fixed G_0 . Top figures - left to right: (1) $W_1(\hat{G}_n, G_0)$ under Gaussian case. (2) $W_1(\hat{G}_n, G_0)$ under Cauchy case. Bottom figures - left to right: (1) $W_1(\hat{G}_n, G_0)$ under Skew normal case. (2) Percentage of time $\hat{m}_n = 3$ obtained from 100 runs.

Case 2: G_0 is varied with the sample size. Under this case, we consider two choices of f_0 : Gaussian and Cauchy distribution with only location parameter.

- Case 2.1 - Gaussian family:

$$\begin{aligned} f_0(x|\eta) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\eta)^2}{2}\right) \\ G_0 &= \frac{1}{4}\delta_{1-1/n} + \frac{1}{4}\delta_{1+1/n} + \frac{1}{2}\delta_2, \end{aligned}$$

where n is the sample size.

- Case 2.2 - Cauchy family:

$$f_0(x|\eta) = \frac{1}{\pi(1 + (x - \eta)^2)}$$

$$G_0 = \frac{1}{4}\delta_{1-1/\sqrt{n}} + \frac{1}{4}\delta_{1+1/\sqrt{n}} + \frac{1}{2}\delta_{1+2/\sqrt{n}}.$$

With these settings, we can verify that $\tilde{G}_0 = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ for the Gaussian case and $\tilde{G}_0 = \delta_1$ for the Cauchy case. Additionally, $W_1(G_0, \tilde{G}_0) \asymp 1/n$ for the Gaussian case and $W_1(G_0, \tilde{G}_0) \asymp 1/\sqrt{n}$ for the Cauchy case. According to the result of Corollary 5.5.1, the convergence rate of $W_1(\hat{G}_n, G_0)$ is $n^{-1/6}$ for the Gaussian case and is $n^{-1/10}$ for the Cauchy case, which are also minimax according to the values of $W_1(G_0, \tilde{G}_0)$. The procedure for choosing K, σ, n , and \hat{m}_n is similar to that of Case 1. See Figure 5.2 where the Wasserstein distances $W_1(\hat{G}_n, G_0)$ and the percentage of time $\hat{m}_n = 3$ are plotted against increasing sample size n along with the error bars. The simulation results for both Gaussian and Cauchy family agree with the convergence rates $n^{-1/6}$ and $n^{-1/10}$ respectively.

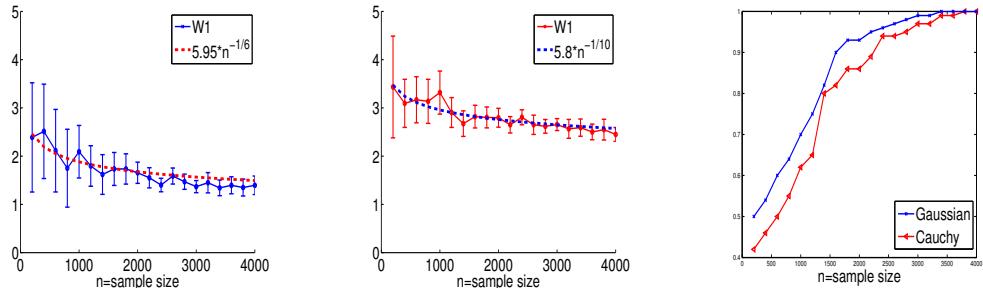


Figure 5.2: Performance of \hat{G}_n in Algorithm 1 under the well-specified kernel setting and varied G_0 . Left to right: (1) $W_1(\hat{G}_n, G_0)$ under Gaussian case. (2) $W_1(\hat{G}_n, G_0)$ under Cauchy case. (3) Percentage of time $\hat{m}_n = 3$ obtained from 100 runs.

Misspecified kernel setting Under that setting, we assess the performance of Algorithm 1 under two cases of f, f_0 , and G_0 .

- Case 2.1 - Gaussian distribution: f is normal kernel,

$$\begin{aligned} f_0(x|\eta, \tau) &= \frac{1}{2}f(x-2|\eta, \tau) + \frac{1}{2}f(x+2|\eta, \tau) \\ G_0 &= \frac{1}{3}\delta_{(0,2)} + \frac{2}{3}\delta_{(1,3)}. \end{aligned}$$

- Case 2.2 - Cauchy distribution: f is Cauchy kernel,

$$\begin{aligned} f_0(x|\eta, \tau) &= \frac{1}{2}f(x-2|\eta, \tau) + \frac{1}{2}f(x+2|\eta, \tau) \\ G_0 &= \frac{1}{3}\delta_{(0,2)} + \frac{2}{3}\delta_{(1,3)}. \end{aligned}$$

With these settings of f, f_0, G_0 , we can verify that $G_* = \frac{1}{6}\delta_{(-2,2)} + \frac{1}{3}\delta_{(-1,3)} + \frac{1}{6}\delta_{(2,2)} + \frac{1}{3}\delta_{(3,3)}$ for any $\sigma > 0$. The procedure for choosing K, σ, n , and \hat{m}_n is similar to that of Case 1 in the well-specified kernel setting. Figure 5.3 illustrates the Wasserstein distances $W_1(\hat{G}_n, G_*)$ and the percentage of time $\hat{m}_n = 4$ along with the increasing sample size n and the error bars. The simulation results under that simple misspecified setting of both families suit with the standard $n^{-1/2}$ rate from Theorem 5.3.2.

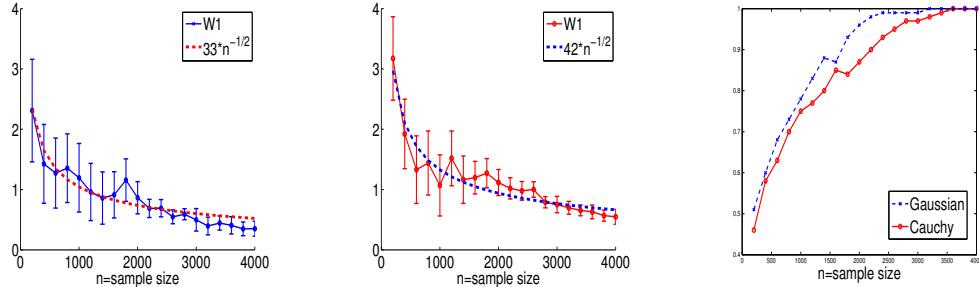


Figure 5.3: Performance of \hat{G}_n in Algorithm 1 under misspecified kernel setting. L to R: (1) $W_1(\hat{G}_n, G_*)$ under Gaussian case. (2) $W_1(\hat{G}_n, G_*)$ under Cauchy case. (3) Percentage of time $\hat{m}_n = 4$ obtained from 100 runs.

5.6.2 Real data

We begin investigating the performance of Algorithm 1 on the well-known data set of the Sodium-lithium countertransport (SLC) data [Dudley et al., 1991, Roeder,

1994, Ishwaran et al., 2001]. This simple dataset includes red blood cell sodium-lithium countertransport (SLC) activity data collected from 190 individuals. As being argued by Roeder [1994], the SLC activity data were believed to be derived from either mixture of two normal distributions or mixture of three normal distributions. Therefore, we will fit this data by using mixture of normal distributions with unknown mean and variance. We choose the bandwidth $\sigma = 0.05$ and the tuning parameter $C_n = \sqrt{3 \log n} / \sqrt{2}$ where n is the sample size. This follows BIC, which is the criterion appropriate for modelling parameters estimation. The simulation result yields $\hat{m}_n = 2$ while the values of \hat{G}_n are reported in Table 5.1.

The SLC activity data was also considered in Woo and Sriram [2006] when the authors achieved $\bar{m}_n = 2$. In particular, they allowed the bandwidth σ to go to 0 and chose the tuning parameter $C_n = 3/n$, which is inspired by AIC criterion. They also obtained similar result of estimating the true number of components when utilizing the minimum Kulback-Leibler divergence estimator (MKE) from [James et al., 2001]. The values of parameter estimates from these two algorithms were presented in Table 7 in Woo and Sriram [2006] where we will use them for the comparison purpose with the results from Algorithm 1. Moreover, we also run the EM algorithm to determine the parameter estimates when we assume the data are from mixture of two normal distributions. All the values of parameter estimates from these three algorithms are included in Table 5.1. Finally, Figure 5.4 represents the fits from parameter estimates of all the mentioned algorithms to SLC data. Even though the weights from Algorithm 1 are not very close to those from WS algorithm and EM algorithm, the fit from Algorithm 1 is comparable to those from these algorithms, i.e., their fits look fairly similar. As a consequence, the results from Algorithm 1 with SLC data are in agreement with those from several state-of-the-art algorithms in the literature.

	p_1	p_2	η_1	η_2	τ_1	τ_2
Algorithm 1	0.264	0.736	0.368	0.231	0.118	0.065
WS algorithm	0.305	0.695	0.352	0.222	0.106	0.060
MKE algorithm	0.246	0.754	0.378	0.225	0.102	0.060
EM algorithm	0.328	0.672	0.363	0.227	0.115	0.058

Table 5.1: Summary of parameter estimates in SLC activity data from mixture of two normal distributions with Algorithm 1, WS algorithm, MKE algorithm, and EM algorithm. Here, p_i, η_i, τ_i represents the weights, means, and variance respectively.

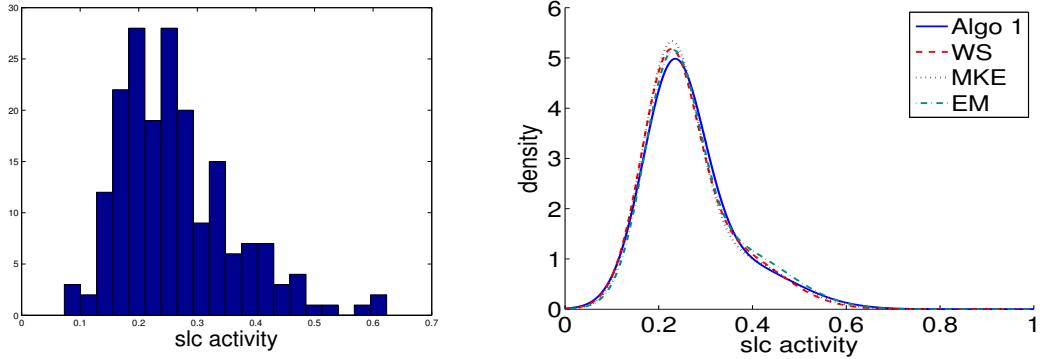


Figure 5.4: From left to right: (1) Histogram of SLC activity data. (2) Density plot from mixture of two normals based on Algorithm 1, WS algorithm, MKE algorithm, and MLE.

5.7 Summaries and discussions

We propose flexible robust estimators of mixing measures in finite mixture models based on minimum Hellinger distance idea. Our estimators are shown to exhibit the consistency of the number of components under both the well-and mis-specified kernel setting. Additionally, the best possible convergence rates of parameter estimates are achieved under various settings of both kernel f and f_0 . Another salient feature of our work is the flexible choice of bandwidth, which circumvents the subtle choice of bandwidth from many proposed estimators in the literature. However, there are still open questions relating to the performance or the extension of our robust estimators in the chapter. We give several examples:

- As being mentioned in the previous sections, the estimators in Algorithm 1 and WS algorithm achieve the consistency of the number of components when the bandwidth goes to 0 sufficiently slow. Can we determine the setting of bandwidth such that the convergence rates of parameter estimates from these algorithms are optimal, at least under the well-specified kernel setting?
- Our analysis is based on the assumption that the components of G_0 belong to compact set Θ . When G_0 is finitely supported, this is always the case, but the set is unknown in advance and, in practice, we often do not know the range of the true parameters. Therefore, it would be interesting to see whether estimators in Algorithm 1 and 2 still achieve both the consistency of the number of components and optimal convergence rates of parameter estimates when $\Theta = \mathbb{R}^{d_1}$.
- Bayesian robust inference of mixing measures in finite mixture models has been of interest recently, see for example [[Miller and Dunson, 2015](#)]. Whether the idea of minimum Hellinger distance can be adapted to that setting is an interesting direction to consider in the future.

5.8 Proofs of key results

We provide now the proofs of Theorem 5.3.1 and Theorem 5.3.2 in Section 5.3. The remaining proofs are given in the Appendices.

PROOF OF THEOREM 5.3.1 We divide the main argument into three key steps:

Step 1: $\hat{m}_n \rightarrow k_0$ almost surely. The proof of this step follows the argument from [Leroux, 1992]. In fact, for any positive integer m we denote

$$G_{0,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma).$$

Now, as $n \rightarrow \infty$ we have almost surely that

$$h(p_{\widehat{G}_{n,m}^{f_0}} * K_\sigma, P_n * K_\sigma) - h(p_{\widehat{G}_{n,m+1}^{f_0}} * K_\sigma, P_n * K_\sigma) \rightarrow d_m,$$

where $d_m = h(p_{G_{0,m}^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) - h(p_{G_{0,m+1}^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$ and the limit is due to the fact that $h(P_n * K_\sigma, p_{G_0} * K_\sigma) \rightarrow 0$ almost surely for all $\sigma > 0$. Now, to demonstrate that $\hat{m}_n \rightarrow k_0$ almost surely, it is sufficient to prove that $d_m = 0$ as $m \geq k_0$ and $d_m > 0$ as $m < k_0$. In fact, as $m \geq k_0$, we have $\inf_{G \in \mathcal{O}_m} h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) = 0$. Therefore, $d_m = 0$ as $m \geq k_0$.

When $m < k_0$, we assume that $d_m = 0$, i.e., $h(p_{G_{0,m}^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) = h(p_{G_{0,m+1}^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$. It implies that

$$h(p_{G_{0,m}^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \leq h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \quad \forall G \in \mathcal{O}_{m+1}.$$

For any $\epsilon > 0$, we choose $G = (1 - \epsilon)G_{0,m} + \epsilon\delta_\theta$ where $\theta \in \Theta$ is some component. The inequality in the above display implies that

$$\begin{aligned} \int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} \left(\left[(1 - \epsilon)p_{G_{0,m}^{f_0}} * K_\sigma(x) + \epsilon f * K_\sigma(x|\theta) \right]^{1/2} - \right. \\ \left. (p_{G_{0,m}^{f_0}} * K_\sigma(x))^{1/2} \right) dx \leq 0. \end{aligned}$$

As $\epsilon \rightarrow 0$, the above inequality divided by ϵ becomes

$$\begin{aligned} & \int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} (p_{G_{0,m}^{f_0}} * K_\sigma(x))^{1/2} dx \geq \\ & \int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} f_0 * K_\sigma(x|\theta) (p_{G_{0,m}^{f_0}} * K_\sigma(x))^{-1/2} dx. \end{aligned}$$

Now, by choosing $\theta = \theta_i^0$ for all $1 \leq i \leq k_0$, as we sum up the right hand side of the above inequality, we obtain

$$\begin{aligned} \int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} (p_{G_{0,m}^{f_0}} * K_\sigma(x))^{1/2} dx & \geq \int (p_{G_0^{f_0}} * K_\sigma(x))^{3/2} (p_{G_{0,m}^{f_0}} * K_\sigma(x))^{-1/2} dx \\ & \geq 1. \end{aligned}$$

Therefore, we have $h(p_{G_{0,m}^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) = 0$. Due to the identifiability assumption of $f_0 * K_\sigma$, the previous equation implies that $G_{0,m} \equiv G_0$, which is a contradiction as $m < k_0$. Thus, we have $d_m > 0$ for any $m < k_0$. We achieve the conclusion that $\hat{m}_n \rightarrow k_0$ almost surely.

Step 2: $h(P_n * K_\sigma, p_{G_0} * K_\sigma) = O_p\left(\sqrt{\frac{\Psi(G_0, \sigma)}{n}}\right)$. Indeed, by means of Taylor expansion up to the first order, we have

$$\begin{aligned} h^2(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma) &= \int \left(1 - \sqrt{1 + \frac{P_n * K_\sigma(x) - p_{G_0^{f_0}} * K_\sigma(x)}{p_{G_0^{f_0}} * K_\sigma(x)}}\right)^2 p_{G_0^{f_0}} * K_\sigma(x) dx. \\ &\simeq \frac{1}{4} \int \frac{(P_n * K_\sigma(x) - p_{G_0^{f_0}} * K_\sigma(x))^2}{p_{G_0^{f_0}} * K_\sigma(x)} dx. \end{aligned}$$

Notice that,

$$E\left(\int \frac{(P_n * K_\sigma(x) - p_{G_0^{f_0}} * K_\sigma(x))^2}{p_{G_0^{f_0}} * K_\sigma(x)} dx\right) = \int \frac{\text{Var}(P_n * K_\sigma(x))}{p_{G_0^{f_0}} * K_\sigma(x)} dx,$$

From assumption (P.2), we obtain $\int \frac{\text{Var}(P_n * K_\sigma(x))}{p_{G_0^{f_0}} * K_\sigma(x)} dx = O\left(\frac{\Psi(G_0, \sigma)}{n}\right)$. It follows that

$$E\left(\int \frac{(P_n * K_\sigma(x) - p_{G_0^{f_0}} * K_\sigma(x))^2}{p_{G_0^{f_0}} * K_\sigma(x)} dx\right) = O\left(\frac{\Psi(G_0, \sigma)}{n}\right).$$

Therefore, we achieve $h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma) = O_p\left(\sqrt{\frac{\Psi(G_0, \sigma)}{n}}\right)$. It implies that for any $\epsilon > 0$, we can find $M_\epsilon > 0$ and the index $N_1(\epsilon) \geq 1$ such that

$$P\left(h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma) > M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma)}{n}}\right) < \epsilon/2 \quad (5.10)$$

for all $n \geq N_1(\epsilon)$.

Step 3: Now, denote the event $A = \{\hat{m}_n \rightarrow k_0 \text{ as } n \rightarrow \infty\}$. Under this event, for each $\omega \in A$, we can find $N(\omega)$ such that as $n \geq N(\omega)$, we have $\hat{m}_n = k_0$. It suggests that $\hat{G}_n \in \mathcal{O}_{k_0}$ as $n \geq N(\omega)$. Define $A_m = \{\omega \in A : \forall n \geq m \text{ we have } \hat{m}_n = k_0\}$. From this definition, we obtain $A_1 \subset A_2 \dots \subset A_m \subset \dots$ and $\bigcup_{m=1}^{\infty} A_m = A$. Therefore, $\lim_{m \rightarrow \infty} P(A_m) = P(A) = 1$. Therefore, for any $\epsilon > 0$ we can find the corresponding index $N_2(\epsilon)$ such that $P(A_{N_2(\epsilon)}) > 1 - \epsilon/2$.

Now, for any $\omega \in A_{N_2(\epsilon)}$, we have $\hat{m}_n = k_0$ as $n \geq N_2(\epsilon)$. From assumptions (P.1) and the definition of $C_1(\sigma)$, we obtain

$$\begin{aligned} C_1(\sigma)W_1(\hat{G}_n, G_0) &\leq h(p_{\hat{G}_n} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \\ &\leq h(p_{\hat{G}_n^{f_0}} * K_\sigma, P_n * K_\sigma) + h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \\ &\leq 2h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma). \end{aligned} \quad (5.11)$$

Using the inequalities from (5.10) and (5.11), we have

$$\begin{aligned}
P\left(W_1(\widehat{G}_n, G_0) > 2M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma)}{C_1^2(\sigma)n}}\right) &= P\left(\left(W_1(\widehat{G}_n, G_0) > 2M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma)}{C_1^2(\sigma)n}}\right) \mathbb{1}_{A_{N_2(\epsilon)}^c}\right) \\
&\quad + P\left(\left(W_1(\widehat{G}_n, G_0) > 2M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma)}{C_1^2(\sigma)n}}\right) \mathbb{1}_{A_{N_2(\epsilon)}}\right) \\
&\leq \epsilon/2 + P\left(\left(W_1(\widehat{G}_n, G_0) > 2M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma)}{C_1^2(\sigma)n}}\right) \mathbb{1}_{A_{N_2(\epsilon)}}\right) < \epsilon
\end{aligned}$$

for all $n \geq \max\{N_1(\epsilon), N_2(\epsilon)\}$. We achieve the conclusion of the theorem.

PROOF OF THEOREM 5.3.2 We also divide our argument into two key steps

Step 1 $\widehat{m}_n \rightarrow k_*$ almost surely. Indeed, by carrying out the same argument as that of Step 1 in the proof of Theorem 5.3.1 (here, we replace f_0 by f and $G_{0,m}$ by $G_{*,m}$, as $m < k_*$), we eventually obtain the following inequality

$$\begin{aligned}
&\int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} (p_{G_{*,m}^f} * K_\sigma(x))^{1/2} dx \geq \\
&\int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} f * K_\sigma(x|\theta) (p_{G_{*,m}^f} * K_\sigma(x))^{-1/2} dx.
\end{aligned}$$

for any $\theta \in \Theta$. By choosing $\theta \in \text{supp}(G_*)$, the set of all support points of G_* , and sum over all of these components, we achieve

$$\begin{aligned}
&\int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} (p_{G_{*,m}^f} * K_\sigma(x))^{1/2} dx \geq \\
&\int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} p_{G_*^f} * K_\sigma(x) (p_{G_{*,m}^f} * K_\sigma(x))^{-1/2} dx.
\end{aligned}$$

From the above inequality, we have

$$\begin{aligned} & \int \left(\sqrt{p_{G_{*,m}^f} * K_\sigma(x)} - \sqrt{p_{G_*^f} * K_\sigma(x)} \right)^2 \sqrt{\frac{p_{G_0^{f_0}} * K_\sigma(x)}{p_{G_{*,m}^f} * K_\sigma(x)}} dx \leq \\ & 2 \left(\int \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} \sqrt{p_{G_{*,m}^f} * K_\sigma(x)} dx - \int \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} \sqrt{p_{G_*^f} * K_\sigma(x)} dx \right) \leq 0, \end{aligned}$$

where the second inequality is due to the fact that G_* minimizes $h(p_{G^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$ among all $G \in \bar{\mathcal{G}}$. The above inequality implies that $p_{G_{*,m}^f} * K_\sigma(x) = p_{G_*^f} * K_\sigma(x)$ for almost surely $x \in \mathcal{X}$. Due to the identifiability of $f * K_\sigma$, we obtain $G_*^m \equiv G_*$, which is a contradiction to the fact that $m < k_*$. Therefore, we achieve $\bar{m}_n \rightarrow k_*$ almost surely.

Step 2 Now, since $\hat{m}_n \rightarrow k_*$ almost surely, using the same argument as Step 3 in the proof of Theorem 5.3.1, we can find $N(\epsilon)$ such that $\hat{m}_n = k_*$ for any $n \geq N(\epsilon)$ and such that $P(A_{N(\epsilon)}) > 1 - \epsilon/2$ for any $\epsilon > 0$. Additionally, since $\hat{G}_n = \hat{G}_{n,\hat{m}_n}$ minimizes $h(p_{G^f} * K_\sigma, P_n * K_\sigma)$ among all $G \in \mathcal{O}_{\hat{m}_n}$, it implies that

$$\int \sqrt{p_{\hat{G}_n^f} * K_\sigma(x)} \sqrt{P_n * K_\sigma(x)} dx \geq \int \sqrt{p_{G_*^f} * K_\sigma(x)} \sqrt{P_n * K_\sigma(x)} dx.$$

when $n \geq N(\epsilon)$. From this inequality, we obtain

$$\begin{aligned} & \int \left(\sqrt{p_{\hat{G}_n^f} * K_\sigma(x)} - \sqrt{p_{G_*^f} * K_\sigma(x)} \right) \left(\sqrt{P_n * K_\sigma(x)} - \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} \right) dx \geq \\ & \int \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} \sqrt{p_{G_*^f} * K_\sigma(x)} dx - \int \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} \sqrt{p_{\hat{G}_n^f} * K_\sigma(x)} dx := B. \end{aligned} \quad (5.12)$$

By means of the inequality in Lemma 5.3.2, we have

$$\begin{aligned} B &\geq \int \sqrt{p_{\widehat{G}_n^f} * K_\sigma(x)} \sqrt{\frac{p_{G_0^{f_0}} * K_\sigma(x)}{p_{G_*^f} * K_\sigma(x)}} dx - \int \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} \sqrt{p_{\widehat{G}_n^f} * K_\sigma(x)} dx \\ &= 2 \left(h^*(p_{\widehat{G}_n^f} * K_\sigma, p_{G_*^f} * K_\sigma) \right)^2 - B. \end{aligned}$$

It implies that $B \geq \left(h^*(p_{\widehat{G}_n^f} * K_\sigma, p_{G_*^f} * K_\sigma) \right)^2$. Plugging this inequality to (5.12) leads to

$$C := \int \left(\sqrt{p_{\widehat{G}_n^f} * K_\sigma(x)} - \sqrt{p_{G_*^f} * K_\sigma(x)} \right) \left(\sqrt{P_n * K_\sigma(x)} - \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} \right) dx \geq \left(h^*(p_{\widehat{G}_n^f} * K_\sigma, p_{G_*^f} * K_\sigma) \right)^2 \quad (5.13)$$

For the left hand side (LHS) of (5.13), we have the following inequality

$$\begin{aligned} C &\leq \left\| (p_{\widehat{G}_n^f} * K_\sigma)^{1/4} - (p_{G_*^f} * K_\sigma)^{1/4} \right\|_\infty \int \left((p_{\widehat{G}_n^f} * K_\sigma(x))^{1/4} + (p_{G_*^f} * K_\sigma(x))^{1/4} \right) \times \\ &\quad \left| \sqrt{P_n * K_\sigma(x)} - \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} \right| dx \\ &\leq \left\| (p_{\widehat{G}_n^f} * K_\sigma)^{1/4} - (p_{G_*^f} * K_\sigma)^{1/4} \right\|_\infty \left\| (p_{\widehat{G}_n^f} * K_\sigma)^{1/4} + (p_{G_*^f} * K_\sigma)^{1/4} \right\|_2 \times \\ &\quad \sqrt{2} h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \quad (5.14) \end{aligned}$$

where the last inequality is due to Holder's inequality. Now, our next argument will be divided into two small key steps

Step 2.1 With assumption (M.3), we will show that

$$D := \left\| (p_{G^f} * K_\sigma)^{1/4} - (p_{G_*^f} * K_\sigma)^{1/4} \right\|_\infty \leq M_3(\sigma) W_1(G, G_0) \quad (5.15)$$

for any $G \in \mathcal{O}_{k_*}$ where $M_3(\sigma)$ is some positive constant.

In fact, denote $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ where $k \leq k_*$ and $G_* = \sum_{i=1}^{k_*} p_i^* \delta_{\theta_i^*}$. Using the same proof argument as that of (5.31) in the proof of Proposition 5.4.1, there exists a positive number ϵ_0 depending only on G_* such that as long as $W_1(G, G_*) \leq \epsilon_0$, G will have exactly k_* components, i.e., $k = k_*$. Additionally, up to the relabelling of the components of G , we also obtain $|p_i - p_i^*| \leq c_0 W_1(G, G_*)$ where c_0 is some positive constant depending only on G_* . Therefore, by choosing G such that $W_1(G, G_*) \leq C_0 = \min \left\{ \epsilon_0, \min_{1 \leq i \leq k_*} \frac{p_i^*}{2c_0} \right\}$, we achieve $|p_i - p_i^*| \leq \min_{1 \leq i \leq k_*} p_i^*/2$. Hence, $p_i \geq \min_{1 \leq i \leq k_*} p_i^*/2$ for all $1 \leq i \leq k_*$. Under this setting of G , for any coupling \mathbf{q} of $\mathbf{p} = (p_1, \dots, p_k)$ and $\mathbf{p}^* = (p_1^*, \dots, p_{k_*}^*)$, by means of triangle inequality we obtain

$$\begin{aligned} D &= \left\| \frac{p_{G^f} * K_\sigma - p_{G_*^f} * K_\sigma}{\{(p_{G^f} * K_\sigma)^{1/4} + (p_{G_*^f} * K_\sigma)^{1/4}\} \{(p_{G^f} * K_\sigma)^{1/2} + (p_{G_*^f} * K_\sigma)^{1/2}\}} \right\|_\infty \\ &\leq \sum_{i,j} q_{ij} \left\| \frac{f * K_\sigma(x|\theta_i) - f * K_\sigma(x|\theta_j^*)}{\{(p_{G^f} * K_\sigma)^{1/4} + (p_{G_*^f} * K_\sigma)^{1/4}\} \{(p_{G^f} * K_\sigma)^{1/2} + (p_{G_*^f} * K_\sigma)^{1/2}\}} \right\|_\infty. \end{aligned}$$

where the ranges of i, j in the above sum satisfy $1 \leq i, j \leq k_*$. It is clear that for any $\alpha \in \{1/2, 1/4\}$

$$\begin{aligned} (p_{G^f} * K_\sigma(x))^\alpha + (p_{G_*^f} * K_\sigma(x))^\alpha &> \min \{p_i^\alpha, (p_j^*)^\alpha\} \{(f * K_\sigma(x|\theta_i))^\alpha + (f * K_\sigma(x|\theta_j^*))^\alpha\} \\ &> \min_{1 \leq i \leq k_*} \left(\frac{p_i^*}{2} \right)^\alpha \{f * K_\sigma(x|\theta_i))^\alpha + (f * K_\sigma(x|\theta_j^*))^\alpha\}. \end{aligned}$$

Therefore, we eventually achieve that

$$D \lesssim \sum_{i,j} q_{ij} \left\| (f * K_\sigma(x|\theta_i))^{1/4} - (f * K_\sigma(x|\theta_j^*))^{1/4} \right\|_\infty.$$

Now, due to assumption (M.3) and mean value theorem, we achieve for any $x \in \mathcal{X}$ that

$$|(f * K_\sigma(x|\theta_i))^{1/4} - (f * K_\sigma(x|\theta_j^*))^{1/4}| \leq M_2(\sigma) |\theta_i - \theta_j^*|.$$

Thus, for any coupling \mathbf{q} of \mathbf{p} and \mathbf{p}^*

$$D \lesssim \sum_{i,j} q_{ij} \|\theta_i - \theta_j^*\|.$$

As a consequence, we eventually have

$$D \lesssim \inf_{\mathbf{q} \in \mathcal{Q}(\mathbf{p}, \mathbf{p}^*)} \sum_{i,j} q_{ij} \|\theta_i - \theta_j^*\| = W_1(G, G_*)$$

for any $G \in \mathcal{O}_{k_*}$ such that $W_1(G, G_*) \leq C_0$. Now, for any $G \in \mathcal{O}_{k_*}$ such that $W_1(G, G_*) > C_0$, as D is bounded, it is clear that $D \lesssim W_1(G, G_*)$. In sum, we achieve inequality (5.15).

Step 2.2 Due to assumption (M.2), we also can quickly verify that

$$\left\| (p_{\widehat{G}_n^f} * K_\sigma)^{1/4} + (p_{G_*^f} * K_\sigma)^{1/4} \right\|_2 \leq 2\sqrt{k_* M_1(\sigma)}. \quad (5.16)$$

Combining (5.14), (5.15), (5.16), we ultimately achieve that

$$\left(h^*(p_{\widehat{G}_n^f} * K_\sigma, p_{G_*^f} * K_\sigma) \right)^2 \leq M(\sigma) W_1(\widehat{G}_n, G_*) h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$$

where $M(\sigma)$ is some positive constant. Due to assumption (M.1), from the result of Lemma 5.3.4 and definition of $C_{*,1}(\sigma)$, we have

$$h^*(p_{\widehat{G}_n^f} * K_\sigma, p_{G_*^f} * K_\sigma) \gtrsim C_{*,1}(\sigma) W_1(\widehat{G}_n, G_*).$$

Combining the above results with the bound $h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma) = O_p\left(\sqrt{\frac{\Psi(G_0, \sigma)}{n}}\right)$ from Step 2 in the proof of Theorem 5.3.1, we quickly obtain the conclusion of the theorem.

5.9 Appendix A

In this Appendix, we provide the proofs of several key results in Section 5.3 and Section 5.4.

PROOF OF LEMMA 5.3.1 The proof of this lemma is a straightforward application of the Fourier transform. In fact, for any finite k different elements $\theta_1, \dots, \theta_k \in \Theta$, assume that we have $\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^{d_1}$ (for all $i = 1, \dots, k$) such that for almost all x

$$\sum_{i=1}^k \alpha_i f_0 * K_\sigma(x|\theta_i) + \beta_i^T \frac{\partial f_0 * K_\sigma}{\partial \theta}(x|\theta_i) = 0,$$

By means of Fourier transformation on both sides of the above equation, we obtain for all $t \in \mathbb{R}^d$ that

$$\widehat{K}_\sigma(t) \left(\sum_{i=1}^k \alpha_i \widehat{f}_0(t|\theta_i) + \beta_i^T \frac{\widehat{f}_0}{\partial \theta}(t|\theta_i) \right) = 0.$$

Since $\widehat{K}_\sigma(t) = \widehat{K}(\sigma t) \neq 0$ for almost all $t \in \mathbb{R}^d$ and f is identifiable in the first order, we obtain that $\alpha_i = 0, \beta_i = \mathbf{0} \in \mathbb{R}^{d_1}$ for all $1 \leq i \leq k$. We achieve the conclusion of this lemma.

PROOF OF LEMMA 5.3.4 We denote the following modification of the total variation distance

$$TV^*(p_{G_1^f} * K_\sigma, p_{G_2^f} * K_\sigma) = \frac{1}{2} \int \left| p_{G_1^f} * K_\sigma(x) - p_{G_2^f} * K_\sigma(x) \right| \left(\frac{p_{G_0^{f_0}} * K_\sigma(x)}{p_{G_*^f} * K_\sigma(x)} \right)^{1/4} dx.$$

for any two mixing measures $G_1, G_2 \in \bar{\mathcal{G}}$. By Holder's inequality, we have

$$TV^*(p_{G^f} * K_\sigma, p_{G_*^f} * K_\sigma) \leq \frac{1}{\sqrt{2}} h^*(p_{G^f} * K_\sigma, p_{G_*^f} * K_\sigma) \times \\ \left(\int \left(\sqrt{p_{G^f} * K_\sigma(x)} + \sqrt{p_{G_*^f} * K_\sigma(x)} \right)^2 dx \right)^{1/2} \leq \sqrt{2} h^*(p_{G^f} * K_\sigma, p_{G_*^f} * K_\sigma). \quad (5.17)$$

Therefore, in order to obtain the conclusion of the lemma it suffices to demonstrate that

$$\inf_{G \in \mathcal{O}_{k_*}} TV^*(p_{G^f} * K_\sigma, p_{G_*^f} * K_\sigma) / W_1(G, G_*) > 0. \quad (5.18)$$

Firstly, we will show that

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_{k_*}} \left\{ \frac{TV^*(p_{G^f} * K_\sigma, p_{G_*^f} * K_\sigma)}{W_1(G, G_*)} : W_1(G, G_*) \leq \epsilon \right\} > 0.$$

Assume that the above inequality does not hold. There exists a sequence $G_n \in \mathcal{O}_{k_*}$ such that $W_1(G_n, G_*) \rightarrow 0$ and $TV^*(p_{G_n^f} * K_\sigma, p_{G_*^f} * K_\sigma) / W_1(G_n, G_*) \rightarrow 0$. By means of Fatou's lemma, we obtain

$$0 = \liminf_{n \rightarrow \infty} \frac{TV^*(p_{G_n^f} * K_\sigma, p_{G_*^f} * K_\sigma)}{W_1(G_n, G_*)} \geq \frac{1}{2} \int \liminf_{n \rightarrow \infty} \frac{\left| p_{G_n^f} * K_\sigma - p_{G_*^f} * K_\sigma \right| \left(\frac{p_{G_0^f} * K_\sigma}{p_{G_*^f} * K_\sigma} \right)^{1/4}}{W_1(G_n, G_*)} dx.$$

Therefore, for almost surely $x \in \mathcal{X}$, we have

$$\liminf_{n \rightarrow \infty} \frac{\left| p_{G_n^f} * K_\sigma - p_{G_*^f} * K_\sigma \right| \left(\frac{p_{G_0^f} * K_\sigma}{p_{G_*^f} * K_\sigma} \right)^{1/4}}{W_1(G_n, G_*)} = 0. \quad (5.19)$$

Since $W_1(G_n, G_*) \rightarrow 0$ and $G_n \in \mathcal{O}_{k_*}$, we can find a subsequence of k_n such that $k_n = k_*$. Without loss of generality, we replace that subsequence of k_n by its whole sequence. Then, G_n will have exactly k_* components for all $n \geq 1$. From here, by

using the same argument as that in the proof of Theorem 3.1 in [Ho and Nguyen \[2016c\]](#), equality (5.19) cannot happen - a contradiction.

Therefore, we can find a positive constant number ϵ_0 such that $TV^*(p_{G^f} * K_\sigma, p_{G_*^f} * K_\sigma) \gtrsim W_1(G, G_*)$ for any $W_1(G, G_*) \leq \epsilon_0$. Now, to obtain the conclusion of (5.18), we only need to verify that

$$\inf_{G \in \mathcal{O}_{k_*}: W_1(G, G_*) > \epsilon_0} TV^*(p_{G^f} * K_\sigma, p_{G_*^f} * K_\sigma) / W_1(G, G_*) > 0.$$

In fact, if the above statement does not hold, we can find a sequence $G'_n \in \mathcal{O}_{k_*}$ such that $W_1(G_n, G_*) > \epsilon_0$ and $TV^*(p_{G'_n} * K_\sigma, p_{G_*^f} * K_\sigma) / W_1(G'_n, G_*) \rightarrow 0$. Since Θ is a closed bounded set, we can find $G' \in \mathcal{O}_{k_*}$ such that a subsequence of G'_n satisfies $W_1(G'_n, G') \rightarrow 0$ and $W_1(G', G_*) > \epsilon_0$. Without loss of generality, we replace that subsequence of G'_n by its whole sequence. Therefore, $TV^*(p_{G'_n} * K_\sigma, p_{G_*^f} * K_\sigma) \rightarrow 0$ as $n \rightarrow \infty$. Since $W_1(G'_n, G') \rightarrow 0$, due to the first order Lipschitz continuity of $f * K_\sigma$ we obtain $p_{G'_n} * K_\sigma(x) \rightarrow p_{G_*^f} * K_\sigma(x)$ for any $x \in \mathcal{X}$ when $n \rightarrow \infty$. Now, by means of Fatou's lemma

$$\begin{aligned} 0 = \lim_{n \rightarrow \infty} TV^*(p_{G'_n} * K_\sigma, p_{G_*^f} * K_\sigma) &\geq \int \liminf_{n \rightarrow \infty} \left| p_{G'_n} * K_\sigma - p_{G_*^f} * K_\sigma \right| \left(\frac{p_{G_0^{f_0}} * K_\sigma}{p_{G_*^f} * K_\sigma} \right)^{1/4} dx \\ &= TV^*(p_{G_*^f} * K_\sigma, p_{G_*^f} * K_\sigma), \end{aligned}$$

which only happens when $p_{G'_n} * K_\sigma(x) = p_{G_*^f} * K_\sigma(x)$ for almost surely x . Due to the identifiability of $f * K_\sigma$, the former equality implies that $G' \equiv G_*$, which is a contradiction to $W_1(G', G_*) > \epsilon_0$. We achieve the conclusion of the lemma.

PROOF OF PROPOSITION 5.3.1 In this proof, to avoid the ambiguity we denote $\widehat{G}_{n,m,\sigma_n} = \widehat{G}_{n,m}$ and $G_{0,m,\sigma_n} = \arg \min_{G \in \mathcal{O}_m} h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$ for any $\sigma > 0$.

Now, as $n \rightarrow \infty$, we will prove for almost surely that

$$h(p_{\widehat{G}_{n,m,\sigma_n}^{f_0}} * K_{\sigma_n}, P_n * K_{\sigma_n}) - h(p_{\widehat{G}_{n,m+1,\sigma_n}^{f_0}} * K_{\sigma_n}, P_n * K_{\sigma_n}) \rightarrow d'_m, \quad (5.20)$$

where $d'_m = h(p_{G_{0,m}^{f_0}}, p_{G_0^{f_0}}) - h(p_{G_{0,m+1}^{f_0}}, p_{G_0^{f_0}})$ where $G_{0,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G^{f_0}}, p_{G_0^{f_0}})$. To achieve this result, we start with the following lemma

Lemma 5.9.1. *For any sequence G_n and $\sigma_n \rightarrow 0$, we have as $n \rightarrow \infty$ that*

$$h(p_{G_n^{f_0}} * K_{\sigma_n}, p_{G_n^{f_0}}) \rightarrow 0.$$

The proof of this lemma is deferred to the Appendix. Now, applying the result of Lemma 5.9.1 to the sequences G_{0,m,σ_n} and σ_n , we have

$$\lim_{n \rightarrow \infty} h(p_{G_{0,m,\sigma_n}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n}) = \lim_{n \rightarrow \infty} h(p_{G_{0,m,\sigma_n}^{f_0}}, p_{G_0^{f_0}}) \geq h(p_{G_{0,m}^{f_0}}, p_{G_0^{f_0}}). \quad (5.21)$$

On the other hand, from the definition of G_{0,m,σ_n} , we have $h(p_{G_{0,m,\sigma_n}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n}) \leq h(p_{G_{0,m}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n})$. Therefore,

$$\begin{aligned} \lim_{n \rightarrow \infty} h(p_{G_{0,m,\sigma_n}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n}) &\leq \lim_{n \rightarrow \infty} h(p_{G_{0,m}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n}) \\ &= h(p_{G_{0,m}^{f_0}}, p_{G_0^{f_0}}). \end{aligned} \quad (5.22)$$

Combining the results from (5.21) and (5.22), we have

$$\lim_{n \rightarrow \infty} h(p_{G_{0,m,\sigma_n}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n}) = h(p_{G_{0,m}^{f_0}}, p_{G_0^{f_0}}). \quad (5.23)$$

Now, we will demonstrate that

$$\lim_{n \rightarrow \infty} h(p_{\widehat{G}_{n,m,\sigma_n}^{f_0}} * K_{\sigma_n}, P_n * K_{\sigma_n}) = \lim_{n \rightarrow \infty} h(p_{G_{0,m,\sigma_n}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n}). \quad (5.24)$$

In fact, from the definition of $\widehat{G}_{n,m,\sigma_n}$ we quickly obtain that

$$\begin{aligned}\lim_{n \rightarrow \infty} h(p_{\widehat{G}_{n,m,\sigma_n}^{f_0}} * K_{\sigma_n}, P_n * K_{\sigma_n}) &\leq \lim_{n \rightarrow \infty} h(p_{G_{0,m,\sigma_n}^{f_0}} * K_{\sigma_n}, P_n * K_{\sigma_n}) \\ &= \lim_{n \rightarrow \infty} h(p_{G_{0,m,\sigma_n}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n})\end{aligned}\quad (5.25)$$

where the last equality is due to the fact that $h(P_n * K_{\sigma_n}, p_{G_0^{f_0}}) \rightarrow 0$ almost surely as $n \rightarrow \infty$, $\sigma_n \rightarrow 0$ and $n\sigma_n^d \rightarrow \infty$. On the other hand, from the formulation of G_{0,m,σ_n} we have

$$\begin{aligned}\lim_{n \rightarrow \infty} h(p_{G_{0,m,\sigma_n}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n}) &\leq \lim_{n \rightarrow \infty} h(p_{\widehat{G}_{n,m,\sigma_n}^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n}) \\ &= \lim_{n \rightarrow \infty} h(p_{\widehat{G}_{n,m,\sigma_n}^{f_0}} * K_{\sigma_n}, P_n * K_{\sigma_n})\end{aligned}\quad (5.26)$$

Combining (5.25) and (5.26), we obtain equality (5.24). Now, the combination of (5.23) and (5.24) leads to

$$\lim_{n \rightarrow \infty} h(p_{\widehat{G}_{n,m,\sigma_n}^{f_0}} * K_{\sigma_n}, P_n * K_{\sigma_n}) = h(p_{G_{0,m}^{f_0}}, p_{G_0^{f_0}}).$$

Therefore, we obtain the conclusion of (5.20). From here, by using the same argument as Step 1 of Theorem 5.3.1, we ultimately get $d'_m = 0$ as $m \geq k_0$ and $d'_m > 0$ as $m < k_0$. As a consequence, $\widehat{m}_n \rightarrow k_0$ almost surely as $n \rightarrow \infty$. The conclusion of the proposition follows.

PROOF OF LEMMA 5.3.2 The proof proceeds by using the idea from Leroux's argument [Leroux, 1992]. In fact, from the definition of G_* , we have $h(p_{G_*^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \leq h(p_{G^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$ for any $G \in \overline{\mathcal{G}}$. Now, for any $\theta \in \Theta$, by choosing $G = (1 - \epsilon)G_* + \epsilon\delta_\theta$ and letting $\epsilon \rightarrow 0$ as Step 1 in the proof of Theorem 5.3.1, we

eventually obtain

$$\begin{aligned} & \int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} (p_{G_*^f} * K_\sigma(x))^{1/2} dx \\ & \geq \int (p_{G_0^{f_0}} * K_\sigma(x))^{1/2} f * K_\sigma(x|\theta) (p_{G_*^f} * K_\sigma(x))^{-1/2} dx. \end{aligned}$$

By choosing $\theta \in \text{supp}(G_*)$ and summing over all of these components, we readily obtain inequality (5.1), which concludes the result of the lemma.

PROOF OF LEMMA 5.3.3 By means of Holder inequality, we obtain

$$\begin{aligned} \left(\int \sqrt{p_{G_{1,*}^f} * K_\sigma(x)} \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} dx \right)^2 & \leq \int p_{G_{1,*}^f} * K_\sigma(x) \sqrt{\frac{p_{G_0^{f_0}} * K_\sigma(x)}{p_{G_{2,*}^f} * K_\sigma(x)}} dx \times \\ & \quad \times \int \sqrt{p_{G_{2,*}^f} * K_\sigma(x)} \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} dx. \end{aligned}$$

From the definition of $G_{1,*}$ and $G_{2,*}$, we achieve

$$\int \sqrt{p_{G_{1,*}^f} * K_\sigma(x)} \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} dx = \int \sqrt{p_{G_{2,*}^f} * K_\sigma(x)} \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} dx.$$

Therefore, the above inequality along with inequality (5.1) in Lemma 5.3.2 lead to

$$\int p_{G_{1,*}^f} * K_\sigma(x) \sqrt{\frac{p_{G_0^{f_0}} * K_\sigma(x)}{p_{G_{2,*}^f} * K_\sigma(x)}} dx = \int \sqrt{p_{G_{2,*}^f} * K_\sigma(x)} \sqrt{p_{G_0^{f_0}} * K_\sigma(x)} dx.$$

It eventually implies that

$$\int \left(\sqrt{p_{G_{1,*}^f} * K_\sigma(x)} - \sqrt{p_{G_{2,*}^f} * K_\sigma(x)} \right)^2 \sqrt{\frac{p_{G_0^{f_0}} * K_\sigma(x)}{p_{G_{2,*}^f} * K_\sigma(x)}} dx = 0.$$

Therefore, $p_{G_{1,*}^f} * K_\sigma(x) = p_{G_{2,*}^f} * K_\sigma(x)$ almost surely $x \in \mathcal{X}$. We obtain the conclusion of the lemma.

PROOF OF THEOREM 5.3.3 The proof of the theorem is rather similar to that in Theorem 5.3.2. Therefore, we only give a sketch of this proof. In fact, to demonstrate that $\bar{m}_n \rightarrow \bar{k}_0$, we only need to show that $\bar{d}_m > 0$ when $m < \bar{k}_0$ where

$$\bar{d}_m = h(p_{\bar{G}_{0,m}^{f_0}}, p_{G_0} * K_\sigma) - h(p_{\bar{G}_{0,m+1}^{f_0}}, p_{G_0} * K_\sigma) \quad (5.27)$$

and $\bar{G}_{0,m} = \arg \min_{G \in \mathcal{O}_m} h(p_G, p_{G_0} * K_\sigma)$. If $d_m = 0$ for some $m < k_0$, following the technique in Step 1 in the proof of Theorem 5.3.2 we eventually achieve

$$\begin{aligned} & \int \left(\sqrt{p_{\bar{G}_{0,m}^{f_0}}} - \sqrt{p_{\bar{G}_0^{f_0}}} \right)^2 \sqrt{\frac{p_{G_0^{f_0}} * K_\sigma}{p_{\bar{G}_0^{f_0}}}} dx \leq \\ & 2 \int \sqrt{p_{G_0^{f_0}} * K_\sigma} \left(\sqrt{p_{\bar{G}_{0,m}^{f_0}} * K_\sigma} - \sqrt{p_{\bar{G}_0^{f_0}} * K_\sigma} \right) dx \leq 0, \end{aligned}$$

which is a contradiction. Therefore, $\bar{m}_n \rightarrow \bar{k}_0$ almost surely.

To establish the convergence rate of \bar{G}_n to \bar{G}_0 , by using inequality (5.2) we ultimately get the following inequality

$$\int \left(\sqrt{p_{\bar{G}_n^{f_0}}} - \sqrt{p_{\bar{G}_0^{f_0}}} \right) \left(\sqrt{P_n * K_\sigma} - \sqrt{p_{G_0^{f_0}} * K_\sigma} \right) dx \geq \left(\bar{h}(p_{\bar{G}_n^{f_0}}, p_{\bar{G}_0^{f_0}}) \right)^2$$

Couple with condition (S.1), (S.2), and inequality (5.3), by using the same technique as that from Step 2 in the proof of Theorem 5.3.2, we have

$$\{\bar{C}(\sigma)\}^2 W_1^2(\bar{G}_n, \bar{G}_0) \leq \bar{M}(\sigma) W_1(\bar{G}_n, \bar{G}_0) h(P_n * K_\sigma, p_{G_0^{f_0}} * K_\sigma),$$

which immediately yields the conclusion of the theorem.

PROOF OF THEOREM 5.4.1 Here, we provide the proof for part (b) only as it is the generalization of part (a). The proof is similar to that in Step 1 of Theorem

5.3.2. In fact, as $n \rightarrow \infty$ we have for almost surely that

$$h(p_{\widehat{G}_{n,m}} * K_\sigma, P_n * K_\sigma) \rightarrow h(p_{G_{*,m}^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \quad (5.28)$$

where $G_{*,m} = \arg \min_{G \in \mathcal{O}_m} h(p_G * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$. From the argument of Step 1 in the proof of Theorem 5.3.2, we have

$$h(p_{G_{*,m+1}^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) < h(p_{G_{*,m}^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \quad (5.29)$$

for any $1 \leq m \leq k_* - 1$. It implies that $G_{*,m} \in \mathcal{E}_m$ for all $1 \leq m \leq k_*$. Now, if we would like to have $\tilde{m}_n \rightarrow k_*$ as $n \rightarrow \infty$, the sufficient and necessary condition is

$$h(p_{G_*^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \leq \epsilon < h(p_{G_{*,k_*-1}^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma),$$

which is precisely the conclusion of the theorem.

PROOF OF PROPOSITION 5.4.1 Using the argument from Step 1 in the proof of Theorem 5.3.1, we obtain that G_{0,k_0-1} has exactly $k_0 - 1$ elements. Now, since $f_0 * K_\sigma$ is uniformly Lipschitz up to the first order and identifiable, we obtain

$$\inf_{G \in \mathcal{E}_{k_0-1}} h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) / W_1(G, G_0) = C'$$

where C' is some positive constant depending only on f_0, G_0, Θ , and σ . Therefore, we get

$$h(p_{G_{0,k_0-1}^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \geq C' W_1(G_0, G_{0,k_0-1}) \geq C' \inf_{G \in \mathcal{E}_{k_0-1}} W_1(G, G_0). \quad (5.30)$$

Now, for any $G = \sum_{i=1}^{k_0-1} p_i \delta_{\theta_i} \in \mathcal{E}_{k_0-1}$, we can find the index $j^* \in [1, k_0]$ such that

$$\|\theta_i - \theta_{j^*}^0\| \geq \min_{1 \leq j \neq j^* \leq k_0} \|\theta_i - \theta_j^0\|$$

for any $1 \leq i \leq k_0 - 1$. Therefore, we obtain

$$2\|\theta_i - \theta_{j^*}^0\| \geq \|\theta_i - \theta_{j^*}^0\| + \min_{1 \leq j \neq j^* \leq k_0} \|\theta_i - \theta_j^0\| \geq \min_{1 \leq u \neq v \leq k_0} \|\theta_u^0 - \theta_v^0\|$$

for any $1 \leq i \leq k_0 - 1$. From the definition of $W_1(G, G_0)$, we can find the optimal coupling $\mathbf{q} \in \mathcal{Q}(\mathbf{p}, \mathbf{p}^0)$ such that $W_1(G, G_0) = \sum q_{ij} \|\theta_i - \theta_j^0\|$. Hence, we get

$$\begin{aligned} W_1(G, G_0) &\geq \sum_{i=1}^{k_0} q_{ij^*} \|\theta_i - \theta_{j^*}^0\| \geq p_{j^*}^0 \min_{1 \leq i \leq k_0-1} \|\theta_i - \theta_{j^*}^0\| \\ &\geq \left(\min_{1 \leq i \leq k_0} p_i^0 \times \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \right) / 2 \end{aligned}$$

for all $G \in \mathcal{E}_{k_0-1}$. It implies that

$$\inf_{G \in \mathcal{E}_{k_0-1}} W_1(G, G_0) \geq \left(\min_{1 \leq i \leq k_0} p_i^0 \times \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \right) / 2. \quad (5.31)$$

By combining (5.30) and (5.31), if we choose $\min_{1 \leq i \leq k_0} p_i^0 \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \geq 2\epsilon/C'$, then $\epsilon < h(p_{G_{0,k_0-1}^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$. As a consequence, by defining $C = 1/C'$ we obtain the conclusion of the lemma.

PROOF OF PROPOSITION 5.4.2 The proof proceeds by treating two sides of (5.5) separately.

Inequality $h(p_{G_*^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \leq \epsilon$: From the definition of G_* , we obtain

$$\begin{aligned} h^2(p_{G_*^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) &\leq h^2(p_{G_0^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \leq V(p_{G_0^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \\ &\leq \|f - f_0\|/2 \end{aligned}$$

It implies that as long as we choose f, f_0 such that $\|f - f_0\| \leq 2\epsilon^2$, we achieve $h(p_{G_*^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \leq \epsilon$.

Inequality $\epsilon < h(p_{G_{*,k_*-1}^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$: We start with the following lemma

Lemma 5.9.2. *Under the hypothesis of Proposition 5.4.2, we obtain*

$$h(p_{G^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \geq C_1 W_1(G, G_0)$$

for any $G \in \mathcal{E}_{k_*-1}$ where C_1 is some positive constant depending only on $f, f_0, G_0, \Theta, \Omega$, and σ .

The proof of Lemma 5.9.2 is deferred to the Appendix. Now, from the above lemma, we have

$$\begin{aligned} h(p_{G_{*,k_*-1}^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) &\geq C_1 W_1(G_0, G_{*,k_*-1}) \geq C_1 \inf_{G \in \mathcal{E}_{k_*-1}} W_1(G, G_0) \\ &\geq C_1 \inf_{G \in \mathcal{E}_{k_0-1}} W_1(G, G_0) \end{aligned}$$

where the last inequality is due to $k_* \leq k_0$. By utilizing the same argument as that of the well-specified setting in the proof of Proposition 5.4.1, if we choose

$$\min_{1 \leq i \leq k_0} p_i^0 \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \geq 2\epsilon/C_1,$$

then $\epsilon < h(p_{G_{*,k_*-1}^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)$. Combining all of the above argument, we achieve the conclusion of the proposition.

PROOF OF PROPOSITION 5.5.3 The proof of this proposition is a straightforward combination of Fatou's lemma and the argument from Theorem 4.6 in Heinrich and Kahn [2016+]. In fact, for any $\epsilon > 0$, as $W_1(G_0^n, \tilde{G}_0) \rightarrow 0$ we can find $M(\epsilon) \in \mathbb{N}$ such that $W_1(G_0^n, \tilde{G}_0) < \epsilon$ for any $n \geq M(\epsilon)$. Additionally, as $\limsup_{n \rightarrow \infty} k_0^n = k$, we can find $T(\epsilon) \in \mathbb{N}$ such that $k_0^n \leq k$ for all $n \geq T(\epsilon)$. Denote $N(\epsilon) = \max \{M(\epsilon), T(\epsilon)\}$ for any $\epsilon > 0$. Now, assume that for any $\epsilon > 0$, we have

$$\inf_{G \in \mathcal{O}_{k_0^n}: W_1(G, \tilde{G}_0) < \epsilon} h(p_{G^{f_0}} * K_\sigma, p_{G_0^{n,f_0}} * K_\sigma) / W_1^{2k-2\tilde{k}_0+1}(G, G_0^n) = 0.$$

as long as $n \geq N(\epsilon)$. For each $n \geq N(\epsilon)$, it implies that we have a sequence $G_m^n \in \mathcal{O}_{k_0^n} \subset \mathcal{O}_k$ such that $W_1(G_m^n, \tilde{G}_0) < \epsilon$ for all $m \geq 1$ and

$$h(p_{G_m^{n,f_0}} * K_\sigma, p_{G_0^{n,f_0}} * K_\sigma) / W_1^{2k-2\tilde{k}_0+1}(G_m^n, G_0^n) \rightarrow 0$$

as $m \rightarrow \infty$. By means of Fatou's lemma, we eventually have

$$\liminf_{m \rightarrow \infty} \left(p_{G_m^{n,f_0}} * K_\sigma(x) - p_{G_0^{n,f_0}} * K_\sigma(x) \right) / W_1^{2k-2\tilde{k}_0+1}(G_m^n, G_0^n) \rightarrow 0$$

almost surely $x \in \mathcal{X}$. However, from the argument of Theorem 4.6 in Heinrich and Kahn [2016+], we can find $\epsilon_0 > 0$ such that for all $G_m^n, G_0^n \in \mathcal{O}_k$ where $W_1(G_m^n, \tilde{G}_0) \vee W_1(G_0^n, \tilde{G}_0) < \epsilon_0$, not for almost surely $x \in \mathcal{X}$ that

$$\left(p_{G_m^{n,f_0}} * K_\sigma(x) - p_{G_0^{n,f_0}} * K_\sigma(x) \right) / W_1^{2k-2\tilde{k}_0+1}(G_m^n, G_0^n) \rightarrow 0$$

for each $n \geq N(\epsilon_0)$, which is a contradiction. Therefore, we achieve the conclusion of the proposition.

5.10 Appendix B

This appendix contains remaining proofs of the main results in the chapter.

PROOF OF PROPOSITION 5.2.1 A careful investigation of the proof of Theorem 3.1 in [Ho and Nguyen, 2016c] implies that

$$h(p_{G^f}, p_{G_0^f}) \gtrsim W_1(G, G_0), \quad (5.32)$$

for any $G \in \mathcal{O}_{k_0}$ such that $W_1(G, G_0)$ is sufficiently small. The latter restriction means that the result in (5.32) is of a local nature. We also would like to extend this lower bound of $h(p_{G^f}, p_{G_0^f})$ for any $G \in \mathcal{O}_{k_0}$. It appears that the first order Lipschitz continuity of f is sufficient to extend (5.32) for any $G \in \mathcal{O}_{k_0}$. In fact, by the result in (5.32), we can find a positive constant ϵ_0 such that

$$\inf_{G \in \mathcal{O}_{k_0}: W_1(G, G_0) \leq \epsilon_0} h(p_{G^f}, p_{G_0^f}) / W_1(G, G_0) > 0$$

Therefore, to extend (5.32) for any $G \in \mathcal{O}_{k_0}$, it is sufficient to demonstrate that

$$\inf_{G \in \mathcal{O}_{k_0}: W_1(G, G_0) > \epsilon_0} h(p_{G^f}, p_{G_0^f}) / W_1(G, G_0) > 0$$

Assume by the contrary that the above result does not hold. It implies that we can find a sequence $G_n \in \mathcal{O}_{k_0}$ such that $W_1(G_n, G_0) > \epsilon_0$ and $h(p_{G_n^f}, p_{G_0^f}) / W_1(G_n, G_0) \rightarrow 0$ as $n \rightarrow \infty$. Since Θ is a compact set, we can find $G' \in \mathcal{O}_{k_0}$ such that a subsequence of G_n satisfies $W_1(G_n, G') \rightarrow 0$ and $W_1(G', G_0) > \epsilon_0$. Without loss of generality, we replace that subsequence by its whole sequence. Therefore, $h(p_{G_n^f}, p_{G_0^f}) \rightarrow 0$ as $n \rightarrow \infty$. Due to the first order Lipschitz continuity of f , we obtain $p_{G_n^f}(x) \rightarrow p_{G'^f}(x)$

for any $x \in \mathcal{X}$ when $n \rightarrow \infty$. Now, by means of Fatou's lemma, we have

$$0 = \lim_{n \rightarrow \infty} h(p_{G_n^f}, p_{G_0^f}) \geq \int \liminf_{n \rightarrow \infty} \left(\sqrt{p_{G_n^f}(x)} - \sqrt{p_{G_0^f}(x)} \right)^2 dx = h(p_{G'^f}, p_{G_0^f}).$$

Since f is identifiable, the above result implies that $G' \equiv G_0$, which contradicts the assumption that $W_1(G', G_0) > \epsilon_0$. As a consequence, we can extend inequality (5.32) for any $G \in \mathcal{O}_{k_0}$ when f is uniformly Lipschitz up to the first order.

PROOF OF EXAMPLE 5.4.1 Assume by the contrary that f_1 and f_2 are not distinguishable. We denote $\theta = (\eta, \tau)$ where η represents the location parameter and τ represents the variance parameter. Now, the assumption implies that we can find $G_1 = \sum_{i=1}^{t_1} \alpha_i \delta_{(\eta_i, \tau_i)}$ and $G_2 = \sum_{i=1}^{t_2} \beta_i \delta_{(\eta'_i, \tau'_i)}$ such that $h(p_{G_1^{f_1}}, p_{G_2^{f_2}}) = 0$. Therefore, we have

$$\sum_{i=1}^{t_1} \alpha_i f_1(x|\eta_i, \tau_i) = \sum_{i=1}^{t_2} \beta_i f_2(x|\eta'_i, \tau'_i) \text{ for almost surely } x \in \mathbb{R}$$

The above equation can be rewritten as

$$\sum_{i=1}^{t_1} \alpha'_i \exp\left(- (a_i x_1^2 + b_i x_1 + c_i)\right) = \sum_{i=1}^{t_2} \beta'_i \left(\nu + a'_i x_1^2 + b'_i x_1 + c'_i\right)^{-(\nu+1)/2}, \quad (5.33)$$

where $\alpha'_i = \frac{\alpha_i}{\sqrt{2\pi}\tau_i}$, $a_i = \frac{1}{2\tau_i^2}$, $b_i = \frac{\eta_i}{\tau_i^2}$, $c_i = \frac{\eta_i^2}{2\tau_i^2}$ as $1 \leq i \leq t_1$ and $\beta'_j = \frac{C_\nu \beta_j}{\tau'_j}$, $a'_i = \frac{1}{\nu(\tau'_i)^2}$, $b'_i = -\frac{2\eta'_i}{\nu(\tau'_i)^2}$, $c'_i = \frac{(\eta'_i)^2}{\nu(\tau'_i)^2}$ for all $1 \leq j \leq t_2$ with $C_\nu = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})}$.

Choose $a_{i_1} = \min_{1 \leq i \leq t_1} \{a_i\}$. Denote $J = \{1 \leq i \leq t_1 : a_i = a_{i_1}\}$. Choose $1 \leq i_2 \leq t_1$ such that $b_{i_2} = \max_{i \in J} \{b_i\}$. Now, multiply both sides of (5.33) with $\exp(a_{i_2} x_1^2 + b_{i_2} x_1 +$

c_{i_2}), we obtain

$$\begin{aligned} \alpha'_{i_2} + \sum_{i \neq i_2} \alpha'_i \exp\left(a_{i_2}x_1^2 + b_{i_2}x_1 + c_{i_2} - (a_i x_1^2 + b_i x_1 + c_i)\right) = \\ \sum_{i=1}^{t_2} \beta'_i \exp(a_{i_2}x_1^2 + b_{i_2}x_1 + c_{i_2}) \left(\nu + a'_i x_1^2 + b'_i x_1 + c'_i\right)^{-(\nu+1)/2}. \end{aligned}$$

As $x \rightarrow \infty$, the left hand side of the above equation goes to α'_{i_2} while the right hand side of the above equation either goes to 0 if $\beta'_i = 0$ for all $1 \leq i \leq t_2$ or goes to $\pm\infty$ if at least one of β'_i differs from 0. As a consequence, we obtain $\alpha'_{i_2} = 0$ and $\beta'_i = 0$ for all $1 \leq i \leq t_2$. It leads to $\beta_i = 0$ for all $1 \leq i \leq t_2$, which is a contradiction. As a consequence, we achieve the conclusion of the example.

PROOF OF LEMMA 5.9.1 The proof idea of this lemma is similar to that of Theorem 1 in Chapter 2 of [Devroye and Gyorfi, 1985]. However, it is slightly more complex than that of this theorem as we allow G_n to vary when σ_n vary. Here, we provide the proof of this lemma for the completeness. Since the Hellinger distance is upper bound by the total variation distance, it is sufficient to demonstrate that $V(p_{G_n^{f_0}} * K_{\sigma_n}, p_{G_n^{f_0}}) \rightarrow 0$ as $n \rightarrow \infty$. Firstly, assume that $f_0(x|\theta)$ is continuous and vanishes outside a compact set which is independent of θ and Σ . For any large number M , we split $K = K' + K''$ where $K' = K1_{\|x\| \leq M}$ and $K'' = K1_{\|x\| > M}$. Now, by using Young's inequality we obtain

$$\begin{aligned} \int |p_{G_n^{f_0}} * K_{\sigma_n}(x) - p_{G_n^{f_0}}(x)| dx &\leq \int \left| p_{G_n^{f_0}} * K'_{\sigma_n}(x) - p_{G_n^{f_0}}(x) \int K'_{\sigma_n}(y) dy \right| dx + \\ &\quad \int |p_{G_n^{f_0}} * K''_{\sigma_n}(x)| dx + \int p_{G_n^{f_0}}(x) dx \int K''_{\sigma_n}(x) dx \\ &\leq \int_A \left| p_{G_n^{f_0}} * K_{\sigma_n}(x) - p_{G_n^{f_0}}(x) \int K'_{\sigma_n}(y) dy \right| dx + 2 \int K''_{\sigma_n}(x) dx. \end{aligned}$$

for some compact set A . It is clear that for any $\epsilon > 0$, we can choose $M(\epsilon)$ such that as $M > M(\epsilon)$, $\int K''_{\sigma_n}(x)dx = \int K''(x)dx < \epsilon$. Regarding the first term in the right hand side of the above display, by denoting $G_n = \sum_{i=1}^{m_n} p_i^n \delta_{\theta_i^n}$ we obtain

$$\begin{aligned} \int_A \left| p_{G_n^{f_0}} * K_{\sigma_n}(x) - p_{G_n^{f_0}}(x) \int K'_{\sigma_n}(y)dy \right| dx &\leq \int_A \int |p_{G_n^{f_0}}(x-y) - p_{G_n^{f_0}}(x)| K'_{\sigma_n}(y) dy dx \\ &\leq \int_A \int \sum_{i=1}^{m_n} p_i^n |f_0(x-y|\theta_i^n) - f_0(x|\theta_i^n)| K'_{\sigma_n}(y) dy dx \\ &\leq \omega(M\sigma_n) \int_A \int |K'_{\sigma_n}(y)| dy dx \leq \omega(M\sigma_n)\mu(A) \rightarrow 0 \end{aligned}$$

where $\omega(t) = \sup_{||x-y|| \leq t} |f_0(x|\theta) - f_0(y|\theta)|$ denotes the modulus of continuity of f_0 and μ denotes the Lebesgue measure. Therefore, the conclusion of this lemma holds for that setting of $f_0(x|\theta)$.

Regarding the general setting of $f_0(x|\theta)$, for any $\epsilon > 0$ since Θ is a bounded set, we can find a continuous function $g(x|\theta)$ being supported on a compact set $B(\epsilon)$ that is independent of $\theta \in \Theta$ such that $\int |f_0(x|\theta) - g(x|\theta)| dx < \epsilon$. Hence, we obtain

$$\begin{aligned} \int |p_{G_n^{f_0}} * K_{\sigma_n}(x) - p_{G_n^{f_0}}(x)| dx &\leq \int \left| \left(p_{G_n^{f_0}} - p_{G_n^g} \right) * K_{\sigma_n}(x) \right| dx + \\ &\quad \int |p_{G_n^{f_0}}(x) - p_{G_n^g}(x)| dx + \int |p_{G_n^g} * K_{\sigma_n}(x) - p_{G_n^g}(x)| dx \\ &\leq 2\epsilon + \int |p_{G_n^g} * K_{\sigma_n}(x) - p_{G_n^g}(x)| dx \end{aligned}$$

where $\int |p_{G_n^g} * K_{\sigma_n}(x) - p_{G_n^g}(x)| dx \rightarrow 0$ as $n \rightarrow \infty$. We achieve the conclusion of the lemma.

Lemma 5.10.1. *Assume that f_0 and K satisfy condition (P.1) in Theorem 5.3.1.*

Furthermore, K has an integrable radial majorant $\Psi \in L_1(\mu)$ where $\Psi(x) = \sup_{||y|| \geq ||x||} |K(y)|$. Then, we can find a positive constant ϵ_1^0 such that as $\sigma \leq \epsilon_1^0$, for any $G \in \mathcal{O}_{k_0}$ we

have

$$h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \geq TV(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \gtrsim W_1(G, G_0).$$

, i.e., $C_1(\sigma) \geq C_1$ as $\sigma \rightarrow 0$ where C_1 only depends on G_0 .

Proof. We divide the proof of this lemma into two key steps

Step 1: We firstly demonstrate the following result

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_{k_0}, \sigma > 0} \left\{ \frac{TV(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)}{W_1(G, G_0)} : W_1(G, G_0) \vee \sigma \leq \epsilon \right\} > 0. \quad (5.34)$$

The proof idea of the above inequality is essentially similar to that from the proof of Theorem 3.1 in [Ho and Nguyen \[2016c\]](#). Here, we provide such proof for the completeness. Assume that the conclusion of inequality (5.34) does not hold. Therefore, we can find two sequences $\{G_n\}$ and $\{\sigma_n\}$ such that $TV(p_{G_n^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n})/W_1(G_n, G_0) \rightarrow 0$ where $W_1(G_n, G_0) \rightarrow 0$ and $\sigma_n \rightarrow 0$ as $n \rightarrow \infty$. As $G_n \in \mathcal{O}_{k_0}$, it implies that there exists a subsequence $\{G_{n_m}\}$ of $\{G_n\}$ such that G_{n_m} has exactly k_0 elements for all m . Without loss of generality, we replace this subsequence by the whole sequence $\{G_n\}$. Now, we can represent G_n as $G_n = \sum_{i=1}^{k_0} p_i^n \delta_{\theta_i^n}$ such that $(p_i^n, \theta_i^n) \rightarrow (p_i^0, \theta_i^0)$. Similar to the argument in Step 1 from the proof of Theorem 3.1 in [Ho and Nguyen \[2016c\]](#), we have $W_1(G_n, G_0) \lesssim d(G_n, G_0)$ where $d(G_n, G_0) = \sum_{i=1}^{k_0} p_i^n \|\Delta \theta_i^n\| + |\Delta p_i^n|$ and $\Delta p_i^n = p_i^n - p_i^0$, $\Delta \theta_i^n = \theta_i^n - \theta_i^0$ for all $1 \leq i \leq k_0$. It implies that $V(p_{G_n^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n})/d(G_n, G_0) \rightarrow 0$.

Now, we denote $g_n(x|\theta) = \int f_0(x-y|\theta) K_{\sigma_n}(y) dy$ for all $\theta \in \Theta$. Similar to Step 2 from the proof of Theorem 3.1 in [Ho and Nguyen \[2016c\]](#), by means of Taylor expansion up to the first order we can represent

$$\frac{p_{G_n^{f_0}} * K_{\sigma_n}(x) - p_{G_0^{f_0}} * K_{\sigma_n}(x)}{d(G_n, G_0)} \asymp \frac{1}{d(G_n, G_0)} \left(\sum_{i=1}^{k_0} \Delta p_i^n g_n(x|\theta_i^0) + p_i^n \frac{\partial g_n}{\partial \theta}(x|\theta_i^0) \right)$$

which are the linear combinations of the elements of $g_n(x|\theta_i^0)$, $\frac{\partial g_n}{\partial \theta}(x|\theta_i^0)$ for $1 \leq i \leq k_0$.

Denote m_n to be the maximum of the absolute values of these coefficients. We can argue that $m_n \not\rightarrow 0$ as $n \rightarrow \infty$. Additionally, since K satisfies condition (P.3), from Theorem 3 in Chapter 2 of [Devroye and Gyorfi \[1985\]](#), for any $\theta \in \Theta$, we have $g_n(x|\theta) \rightarrow f_0(x|\theta)$ and $\frac{\partial g_n}{\partial \theta}(x|\theta) \rightarrow \frac{\partial f_0}{\partial \theta}(x|\theta)$ for almost surely x . Therefore, we obtain

$$\frac{1}{m_n} \frac{d_n \left(p_{G_n^{f_0}} * K_{\sigma_n}(x) - p_{G_0^{f_0}} * K_{\sigma_n}(x) \right)}{d(G_n, G_0)} \rightarrow \sum_{i=1}^{k_0} \alpha_i f_0(x|\theta_i^0) + \beta_i^T \frac{\partial f_0}{\partial \theta}(x|\theta_i^0)$$

where not all the elements of α_i, β_i equal to 0. Due to the first order identifiability of f_0 and the Fatou's lemma, $TV(p_{G_n^{f_0}} * K_{\sigma_n}, p_{G_0^{f_0}} * K_{\sigma_n})/d(G_n, G_0) \rightarrow 0$ will lead to $\alpha_i = 0, \beta_i = \mathbf{0} \in \mathbb{R}^{d_1}$ for all $1 \leq i \leq k_0$, which is a contradiction. We achieve the conclusion of (5.34).

Step 2: The result of (5.34) implies that we can find a positive number ϵ_1^0 such that as $W_1(G, G_0) \vee \sigma \leq \epsilon_1^0$, we have

$$h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \geq TV(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \gtrsim W_1(G, G_0). \quad (5.35)$$

In order to extend the above inequality to any $G \in \mathcal{O}_{k_0}$, it is sufficient to demonstrate that

$$\inf_{\sigma < \epsilon_1^0, W_1(G, G_0) > \epsilon_1^0} \frac{h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)}{W_1(G, G_0)} > 0.$$

In fact, if the above result does not hold, we can find two sequences $G'_n \in \mathcal{O}_{k_0}$ and σ'_n such that $W_1(G'_n, G_0) > \epsilon_1^0$, $\sigma'_n \leq \epsilon_1^0$ and $h(p_{G'_n} * K_{\sigma'_n}, p_{G_0^{f_0}} * K_{\sigma'_n})/W_1(G'_n, G_0) \rightarrow 0$ as $n \rightarrow \infty$. Since Θ is closed bounded set, we can find two subsequences $\{G'_{n_m}\}$ and $\{\sigma'_{n_m}\}$ of $\{G'_n\}$ and $\{\sigma'_n\}$ respectively such that $W_1(G'_{n_m}, G') \rightarrow 0$ and $|\sigma'_{n_m} - \sigma'| \rightarrow 0$ as $m \rightarrow \infty$ where $G' \in \mathcal{O}_{k_0}$ and $\sigma' \in [0, \epsilon_1^0]$.

Due to the first order Lipschitz continuity of $f * K_{\sigma_{nm}}$ for any $m \geq 1$, we achieve $p_{G'^{f_0}} * K_{\sigma'_{nm}}(x) \rightarrow p_{G'^{f_0}} * K_{\sigma'}(x)$ for any $x \in \mathcal{X}$. Here, $p_{G'^{f_0}} * K_{\sigma'} = p_{G'^{f_0}}$ when $\sigma' = 0$. Therefore, by utilizing the Fatou's argument, we obtain $h(p_{G'^{f_0}} * K_{\sigma'}, p_{G_0^{f_0}} * K_{\sigma'}) = 0$, which implies $G' \equiv G_0$, a contradiction. As a consequence, when $\sigma \leq \epsilon_1^0$, for any $G \in \mathcal{O}_{k_0}$ we have

$$h(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \geq V(p_{G^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \gtrsim W_1(G, G_0).$$

We achieve the conclusion of the lemma. \square

PROOF OF LEMMA 5.9.2 To obtain the conclusion of this lemma, it is equivalent to demonstrate that

$$\inf_{G \in \mathcal{E}_{k_*-1}} h(p_{G^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) / W_1(G, G_0) > 0.$$

Assume by the contrary that the above conclusion does not hold. It implies that we can find sequence of measures $G_n \in \mathcal{E}_{k_*-1}$ such that $h(p_{G_n^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) / W_1(G_n, G_0) \rightarrow 0$ as $n \rightarrow \infty$. Since Θ is closed bounded set, it implies that we can find a subsequence of G_n that converge to $G' \in \mathcal{E}_{k_*-1}$ in W_1 distance. Without loss of generality, we replace this subsequence by the whole sequence of G_n , i.e $W_1(G_n, G') \rightarrow 0$. Since $k_* \leq k_0$, it implies that $W_1(G', G_0) \neq 0$. Therefore, $W_1(G_n, G_0) \not\rightarrow 0$ as $n \rightarrow \infty$. Hence, $h(p_{G_n^{f_0}} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \rightarrow 0$ as $n \rightarrow \infty$. Now, from the triangle inequality, we obtain

$$|h(p_{G_n^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) - h(p_{G'^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma)| \leq h(p_{G_n^f} * K_\sigma, p_{G'^f} * K_\sigma).$$

As $W_1(G_n, G') \rightarrow 0$, from the uniform Lipschitz continuity of f and Holder's inequality, we obtain $h(p_{G_n^f} * K_\sigma, p_{G'^f} * K_\sigma) \leq h(p_{G_n^f}, p_{G'^f}) \rightarrow 0$ as $n \rightarrow \infty$. Thus,

$h(p_{G_n^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) \rightarrow h(p_{G'^f} * K_\sigma, p_{G_0^{f_0}} * K_\sigma) = 0$. Since f and f_0 are distinguishable, we achieve $G' \equiv G_0$, which is a contradiction. As a consequence, we achieve the conclusion of the lemma.

Lemma 5.10.2. *Assume that $\widehat{K}(t) \neq 0$ for almost all $t \in \mathbb{R}^d$ where $\widehat{K}(t)$ is the Fourier transform of kernel function K . Then if $I(G_0, f_0)$ has r -th singularity level for some $r \geq 0$, then $I(G_0, f_0 * K_\sigma)$ also has r -th singularity level for any $\sigma > 0$.*

Proof. Remind that, $I(G_0, f_0)$ has r -th singularity level is equivalent to G_0 is r -singular relative to the ambient space \mathcal{O}_{k_0} and kernel function f_0 in [Ho and Nguyen, 2016b]. Now, for any $\rho \in \mathbb{N}$, given any sequence $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{\theta_i^n} \in \mathcal{O}_{k_0}$ and $G_n \rightarrow G_0$ in W_ρ metric. We can find a subsequence of G_n such that $k_n = k_0$ and each atoms of G_0 will have exactly one component of G_n converges to. Without loss of generality, we replace the subsequence of G_n by its whole sequence and relabel the atoms of G_n such that $(p_i^n, \theta_i^n) \rightarrow (p_i^0, \theta_i^0)$ for all $1 \leq i \leq k_0$. Denote $\Delta\theta_i^n = \theta_i^n - \theta_i^0$ and $\Delta p_i^n = p_i^n - p_i^0$ for all $1 \leq i \leq k_0$. From Definition 3.1 in Ho and Nguyen [2016b], a ρ -minimal form of G_n from Taylor expansion up to the order ρ satisfies

$$\frac{p_{G_n^{f_0}}(x) - p_{G_0^{f_0}}(x)}{W_\rho^\rho(G_n, G_0)} = \sum_{l=1}^{T_\rho} \left(\frac{\xi_l^{(\rho)}(G_n)}{W_\rho^\rho(G_0, G_n)} \right) H_l^{(\rho)}(x) + o(1),$$

for all x . Here, $H_l^{(\rho)}$ are linearly independent functions of x for all l , and coefficients $\xi_l^{(\rho)}(G)$ are polynomials of the components of $\Delta\theta_i$, and p_i for l ranges from 1 to a finite T_ρ . From the above representation, we achieve

$$\frac{p_{G_n^{f_0 * K_\sigma}}(x) - p_{G_0^{f_0 * K_\sigma}}(x)}{W_\rho^\rho(G_n, G_0)} = \sum_{l=1}^{T_\rho} \left(\frac{\xi_l^{(\rho)}(G_n)}{W_\rho^\rho(G_0, G_n)} \right) H_l^{(\rho)} * K_\sigma(x) + o(1), \quad (5.36)$$

where $H_l^{(\rho)} * K_\sigma(x) = \int H_l^{(\rho)}(x - y) K_\sigma(y) dy$ for all $1 \leq l \leq T_\rho$. We will show that $H_l^{(\rho)} * K_\sigma(x)$ are linearly independent functions of x for all $1 \leq l \leq T_\rho$. In fact, assume

that we can find the coefficients $\alpha_l \in \mathbb{R}$ such that

$$\sum_{l=1}^{T_\rho} \alpha_l H_l^{(\rho)} * K_\sigma(x) = 0$$

for all x . By means of Fourier transformation in both sides of the above equation, we obtain

$$\widehat{K}(t) \left(\sum_{l=1}^{T_\rho} \alpha_l \widehat{H}_l^{(\rho)}(t) \right) = 0$$

for all $t \in \mathbb{R}^d$. As $\widehat{K}(t) \neq 0$ for all $t \in \mathbb{R}^d$ and $H_l^{(\rho)}(x)$ for all l are linearly independent functions of x for all $1 \leq l \leq T_\rho$, the above equation implies that $\alpha_l = 0$ for all $1 \leq l \leq T_\rho$. Therefore, $H_l^{(\rho)} * K_\sigma(x)$ are linearly independent functions of x for all $1 \leq l \leq T_\rho$ and $\rho \in \mathbb{N}$.

From the hypothesis, since $I(G_0, f_0)$ has r -th singularity level, it implies that for any sequence $G_n \in \mathcal{O}_{k_0}$ such that $W_{r+1}^{r+1}(G_n, G_0) \rightarrow 0$, we do not have all the ratios $\xi_l^{(r+1)}(G_n)/W_{r+1}^{r+1}(G_0, G_n)$ in (5.36) go to 0 for $1 \leq l \leq T_{r+1}$. It in turns also means that not all the ratios $\xi_l^{(r)}(G'_n)/W_r^r(G_0, G'_n)$ in (5.36) go to 0. Additionally, as $I(G_0, f_0)$ has r -th singularity level, we can find a sequence $G'_n \in \mathcal{O}_{k_0}$ such that $W_r^r(G'_n, G_0) \rightarrow 0$ and $\xi_l^{(r)}(G'_n)/W_r^r(G_0, G'_n)$ in (5.36) go to 0 for $1 \leq l \leq T_r$. It in turns also means that all the ratios $\xi_l^{(r)}(G'_n)/W_r^r(G_0, G'_n)$ in (5.36) go to 0. As a consequence, from Definition 3.3 in [Ho and Nguyen \[2016b\]](#), we achieve the conclusion of the lemma. \square

CHAPTER VI

Multilevel clustering via Wasserstein means

We propose a novel approach to the problem of multilevel clustering, which aims to simultaneously partition data in each group and discover grouping patterns among groups in a potentially large hierarchically structured corpus of data. Our method involves a joint optimization formulation over several spaces of discrete probability measures, which are endowed with Wasserstein distance metrics. We propose a number of variants of this problem, which admit fast optimization algorithms, by exploiting the connection to the problem of finding Wasserstein barycenters. Consistency properties are established for the estimates of both local and global clusters. Finally, experiment results with both synthetic and real data are presented to demonstrate the flexibility and scalability of the proposed approach.¹

6.1 Introduction

In numerous applications in engineering and sciences, data are often organized in a multilevel structure. For instance, a typical structural view of text data in machine learning is to have words grouped into documents, documents are grouped into corpora. A prominent strand of modeling and algorithmic works in the past couple decades has been to discover latent multilevel structures from these hierarchically

¹This chapter has been published in [Ho et al., 2017].

structured data. For specific clustering tasks, one may be interested in simultaneously partitioning the data in each group (to obtain local clusters) and partitioning a collection of data groups (to obtain global clusters). Another concrete example is the problem of clustering images (i.e., global clusters) where each image contains partitions of multiple annotated regions (i.e., local clusters) [Oliva and Torralba, 2001]. While hierarchical clustering techniques may be employed to find a tree-structed clustering given a collection of data points, they are not applicable to discovering the nested structure of multilevel data. Bayesian hierarchical models provide a powerful approach, exemplified by influential works such as Blei et al. [2003], Pritchard et al. [2000], Teh et al. [2006]. More specific to the simultaneous and multilevel clustering problem, we mention the paper of Rodriguez et al. [2008]. In this interesting work, a Bayesian nonparametric model, namely the nested Dirichlet process (NDP) model, was introduced that enables the inference of clustering of a collection of probability distributions from which different groups of data are drawn. With suitable extensions, this modeling framework has been further developed for simultaneous multilevel clustering, see for instance, [Wulsin et al., 2016, Nguyen et al., 2014, Huynh et al., 2016].

The focus of this chapter is on the multilevel clustering problem motivated in the aforementioned modeling works, but we shall take a purely optimization approach. We aim to formulate optimization problems that enable the discovery of multilevel clustering structures hidden in grouped data. Our technical approach is inspired by the role of optimal transport distances in hierarchical modeling and clustering problems. The optimal transport distances, also known as Wasserstein distances [Villani, 2003], have been shown to be the natural distance metric for the convergence theory of latent mixing measures arising in both mixture models [Nguyen, 2013] and hierarchical models [Nguyen, 2016]. They are also intimately connected to the problem of clustering — this relationship goes back at least to the work of [Pollard, 1982], where it is pointed out that the well-known K-means clustering algorithm can be

directly linked to the quantization problem — the problem of determining an optimal finite discrete probability measure that minimizes its second-order Wasserstein distance from the empirical distribution of given data [Graf and Luschgy, 2000].

If one is to perform simultaneous K-means clustering for hierarchically grouped data, both at the global level (among groups), and local level (within each group), then this can be achieved by a joint optimization problem defined with suitable notions of Wasserstein distances inserted into the objective function. In particular, multilevel clustering requires the optimization in the space of probability measures defined in *different* levels of abstraction, including the space of measures of measures on the space of grouped data. Our goal, therefore, is to formulate this optimization precisely, to develop algorithms for solving the optimization problem efficiently, and to make sense of the obtained solutions in terms of statistical consistency.

The algorithms that we propose address directly a multilevel clustering problem formulated from a purely optimization viewpoint, but they may also be taken as a fast approximation to the inference of latent mixing measures that arise in the nested Dirichlet process of [Rodriguez et al., 2008]. From a statistical viewpoint, we shall establish a consistency theory for our multilevel clustering problem in the manner achieved for K-means clustering [Pollard, 1982]. From a computational viewpoint, quite interestingly, we will be able to explicate and exploit the connection between our optimization and that of finding the Wasserstein barycenter [Aguech and Carlier, 2011], an interesting computational problem that have also attracted much recent interests, e.g., [Cuturi and Doucet, 2014].

In summary, the main contributions offered in this work include (i) a new optimization formulation to the multilevel clustering problem using Wasserstein distances defined on different levels of the hierarchical data structure; (ii) fast algorithms by exploiting the connection of our formulation to the Wasserstein barycenter problem; (iii) consistency theorems established for proposed estimates under very mild condition of

data's distributions; (iv) several flexible alternatives by introducing constraints that encourage the borrowing of strength among local and global clusters, and (v) finally, demonstration of efficiency and flexibility of our approach in a number of simulated and real data sets.

The chapter is organized as follows. Section 6.2 provides preliminary background on Wasserstein distance, Wasserstein barycenter, and the connection between K-means clustering and the quantization problem. Section 6.3 presents several optimization formulations of the multilevel clustering problem, and the algorithms for solving them. Section 6.4 establishes consistency results of the estimators introduced in Section 6.4. Section 6.5 presents careful simulation studies with both synthetic and real data. Finally, we conclude the chapter with a discussion in Section 6.6. Additional technical details, including all proofs, are given in the Supplement.

6.2 Background

For any given subset $\Theta \subset \mathbb{R}^d$, let $\mathcal{P}(\Theta)$ denote the space of Borel probability measures on Θ . The Wasserstein space of order $r \in [1, \infty)$ of probability measures on Θ is defined as $\mathcal{P}_r(\Theta) = \left\{ G \in \mathcal{P}(\Theta) : \int \|x\|^r dG(x) < \infty \right\}$, where $\|\cdot\|$ denotes Euclidean metric in \mathbb{R}^d . Additionally, for any $k \geq 1$ the probability simplex is denoted by $\Delta_k = \left\{ u \in \mathbb{R}^k : u_i \geq 0, \sum_{i=1}^k u_i = 1 \right\}$. Finally, let $\mathcal{O}_k(\Theta)$ (resp., $\mathcal{E}_k(\Theta)$) be the set of probability measures with at most (resp., exactly) k support points in Θ .

Wasserstein distances For any elements G and G' in $\mathcal{P}_r(\Theta)$ where $r \geq 1$, the Wasserstein distance of order r between G and G' is defined as (cf. [Villani, 2003]):

$$W_r(G, G') = \left(\inf_{\pi \in \Pi(G, G')} \int_{\Theta^2} \|x - y\|^r d\pi(x, y) \right)^{1/r}$$

where $\Pi(G, G')$ is the set of all probability measures on $\Theta \times \Theta$ that have marginals G and G' . In words, $W_r^r(G, G')$ is the optimal cost of moving mass from G to G' , where the cost of moving unit mass is proportional to r -power of Euclidean distance in Θ . When G and G' are two discrete measures with finite number of atoms, fast computation of $W_r(G, G')$ can be achieved (see, e.g., [Cuturi \[2013\]](#)). The details of this are deferred to the Supplement.

By a recursion of concepts, we can speak of measures of measures, and define a suitable distance metric on this abstract space: the space of Borel measures on $\mathcal{P}_r(\Theta)$, to be denoted by $\mathcal{P}_r(\mathcal{P}_r(\Theta))$. This is also a Polish space (that is, complete and separable metric space) as $\mathcal{P}_r(\Theta)$ is a Polish space. It will be endowed with a Wasserstein metric of order r that is induced by a metric W_r on $\mathcal{P}_r(\Theta)$ as follows (cf. Section 3 of [Nguyen \[2016\]](#)): for any $\mathcal{D}, \mathcal{D}' \in \mathcal{P}_r(\mathcal{P}_r(\Theta))$

$$W_r(\mathcal{D}, \mathcal{D}') := \left(\inf_{\mathcal{P}_r(\Theta)^2} \int W_r^r(G, G') d\pi(G, G') \right)^{1/r}$$

where the infimum in the above ranges over all $\pi \in \Pi(\mathcal{D}, \mathcal{D}')$ such that $\Pi(\mathcal{D}, \mathcal{D}')$ is the set of all probability measures on $\mathcal{P}_r(\Theta) \times \mathcal{P}_r(\Theta)$ that has marginals \mathcal{D} and \mathcal{D}' . In words, $W_r(\mathcal{D}, \mathcal{D}')$ corresponds to the optimal cost of moving mass from \mathcal{D} to \mathcal{D}' , where the cost of moving unit mass in its space of support $\mathcal{P}_r(\Theta)$ is proportional to the r -power of the W_r distance in $\mathcal{P}_r(\Theta)$. Note a slight notational abuse — W_r is used for both $\mathcal{P}_r(\Theta)$ and $\mathcal{P}_r(\mathcal{P}_r(\Theta))$, but it should be clear which one is being used from context.

Wasserstein barycenter Next, we present a brief overview of Wasserstein barycenter problem, first studied by [Aguech and Carlier, 2011](#) and subsequently many others (e.g., [\[Benamou et al., 2015, Solomon et al., 2015, Álvarez Estebana et al., 2016\]](#)). Given probability measures $P_1, P_2, \dots, P_N \in \mathcal{P}_2(\Theta)$ for $N \geq 1$, their Wasserstein barycenter $\overline{P}_{N,\lambda}$ is such that

$$\bar{P}_{N,\lambda} = \arg \min_{P \in \mathcal{P}_2(\Theta)} \sum_{i=1}^N \lambda_i W_2^2(P, P_i) \quad (6.1)$$

where $\lambda \in \Delta_N$ denote weights associated with P_1, \dots, P_N . When P_1, \dots, P_N are discrete measures with finite number of atoms and the weights λ are uniform, it was shown by [Anderes et al., 2015] that the problem of finding Wasserstein barycenter $\bar{P}_{N,\lambda}$ over the space $\mathcal{P}_2(\Theta)$ in (6.1) is reduced to search only over a much simpler space $\mathcal{O}_l(\Theta)$ where $l = \sum_{i=1}^N s_i - N + 1$ and s_i is the number of components of P_i for all $1 \leq i \leq N$. Efficient algorithms for finding local solutions of the Wasserstein barycenter problem over $\mathcal{O}_k(\Theta)$ for some $k \geq 1$ have been studied recently in [Cuturi and Doucet, 2014]. These algorithms will prove to be a useful building block for our method as we shall describe in the sequel. The notion of Wasserstein barycenter has been utilized for approximate Bayesian inference [Srivastava et al., 2015].

K-means as quantization problem The well-known K -means clustering algorithm can be viewed as solving an optimization problem that arises in the problem of quantization, a simple but very useful connection [Pollard, 1982, Graf and Luschgy, 2000]. The connection is the following. Given n unlabelled samples $Y_1, \dots, Y_n \in \Theta$. Assume that these data are associated with at most k clusters where $k \geq 1$ is some given number. The K -means problem finds the set S containing at most k elements $\theta_1, \dots, \theta_k \in \Theta$ that minimizes the following objective

$$\inf_{S: |S| \leq k} \frac{1}{n} \sum_{i=1}^n d^2(Y_i, S). \quad (6.2)$$

Let $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{Y_i}$ be the empirical measure of data Y_1, \dots, Y_n . Then, problem (6.2) is equivalent to finding a discrete probability measure G which has finite number of support points and solves:

$$\inf_{G \in \mathcal{O}_k(\Theta)} W_2^2(G, P_n). \quad (6.3)$$

Due to the inclusion of Wasserstein metric in its formulation, we call this a *Wasserstein means problem*. This problem can be further thought of as a Wasserstein barycenter problem where $N = 1$. In light of this observation, as noted by [Cutturi and Doucet, 2014], the algorithm for finding the Wasserstein barycenter offers an alternative for the popular Loyd's algorithm for determining local minimum of the K-means objective.

6.3 Clustering with multilevel structure data

Given m groups of n_j exchangeable data points $X_{j,i}$ where $1 \leq j \leq m, 1 \leq i \leq n_j$, i.e., data are presented in a two-level grouping structure, our goal is to learn about the two-level clustering structure of the data. We want to obtain simultaneously local clusters for each data group, and global clusters among all groups.

6.3.1 Multilevel Wasserstein Means (MWM) Algorithm

For any $j = 1, \dots, m$, we denote the empirical measure for group j by $P_{n_j}^j := \frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{X_{j,i}}$. Throughout this section, for simplicity of exposition we assume that the number of both local and global clusters are either known or bounded above by a given number. In particular, for local clustering we allow group j to have at most k_j clusters for $j = 1, \dots, m$. For global clustering, we assume to have M group (Wasserstein) means among the m given groups.

High level idea For local clustering, for each $j = 1, \dots, m$, performing a K-means clustering for group j , as expressed by (6.3), can be viewed as finding a finite discrete measure $G_j \in \mathcal{O}_{k_j}(\Theta)$ that minimizes squared Wasserstein distance $W_2^2(G_j, P_{n_j}^j)$. For global clustering, we are interested in obtaining clusters out of m groups, each of which is now represented by the discrete measure G_j , for $j = 1, \dots, m$. Adopting again the viewpoint of Eq. (6.3), provided that all of G_j s are given, we can apply K -

means quantization method to find their distributional clusters. The global clustering in the space of measures of measures on Θ can be succinctly expressed by

$$\inf_{\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))} W_2^2 \left(\mathcal{H}, \frac{1}{m} \sum_{j=1}^m \delta_{G_j} \right).$$

However, G_j are not known — they have to be optimized through local clustering in each data group.

MWM problem formulation We have arrived at an objective function for jointly optimizing over both local and global clusters

$$\inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} \sum_{j=1}^m W_2^2(G_j, P_{n_j}^j) + W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^m \delta_{G_j}). \quad (6.4)$$

We call the above optimization the problem of *Multilevel Wasserstein Means* (*MWM*). The notable feature of MWM is that its loss function consists of two types of distances associated with the hierarchical data structure: one is distance in the space of measures, e.g., $W_2^2(G_j, P_{n_j}^j)$, and the other in space of measures of measures, e.g., $W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^m \delta_{G_j})$. By adopting K-means optimization to both local and global clustering, the multilevel Wasserstein means problem might look formidable at the first sight. Fortunately, it is possible to simplify this original formulation substantially, by exploiting the structure of \mathcal{H} .

Indeed, we can show that formulation (6.4) is equivalent to the following optimization problem, which looks much simpler as it involves only measures on Θ :

$$\inf_{G_j \in \mathcal{O}_{k_j}(\Theta), \mathbf{H}} \sum_{j=1}^m W_2^2(G_j, P_{n_j}^j) + \frac{d_{W_2}^2(G_j, \mathbf{H})}{m} \quad (6.5)$$

where $d_{W_2}^2(G, \mathbf{H}) := \min_{1 \leq i \leq M} W_2^2(G, H_i)$ and $\mathbf{H} = (H_1, \dots, H_M)$, with each $H_i \in \mathcal{P}_2(\Theta)$. The proof of this equivalence is deferred to Proposition B.4 in the Sup-

plement. Before going into to the details of the algorithm for solving (6.5) in Section 6.3.1.2, we shall present some simpler cases, which help to illustrate some properties of the optimal solutions of (6.5), while providing insights of subsequent developments of the MWM formulation. Readers may proceed directly to Section 6.3.1.2 for the description of the algorithm in the first reading.

6.3.1.1 Properties of MWM in special cases

Example 1. Suppose $k_j = 1$ and $n_j = n$ for all $1 \leq j \leq m$, and $M = 1$. Write $\mathbf{H} = H \in \mathcal{P}_2(\Theta)$. Under this setting, the objective function (6.5) can be rewritten as

$$\inf_{\substack{\theta_j \in \Theta, \\ H \in \mathcal{P}_2(\Theta)}} \sum_{j=1}^m \sum_{i=1}^n \|\theta_j - X_{j,i}\|^2 + W_2^2(\delta_{\theta_j}, H)/m, \quad (6.6)$$

where $G_j = \delta_{\theta_j}$ for any $1 \leq j \leq m$. From the result of Theorem A.1 in the Supplement,

$$\begin{aligned} \inf_{\theta_j \in \Theta} \sum_{j=1}^m W_2^2(\delta_{\theta_j}, H) &\geq \inf_{H \in \mathcal{E}_1(\Theta)} \sum_{j=1}^m W_2^2(G_j, H) \\ &= \sum_{j=1}^m \|\theta_j - (\sum_{i=1}^m \theta_i)/m\|^2, \end{aligned}$$

where second infimum is achieved when $H = \delta_{(\sum_{j=1}^m \theta_j)/m}$. Thus, objective function (6.6) may be rewritten as

$$\inf_{\theta_j \in \Theta} \sum_{j=1}^m \sum_{i=1}^n \|\theta_j - X_{j,i}\|^2 + \|m\theta_j - (\sum_{l=1}^m \theta_l)\|^2/m^3.$$

Write $\bar{X}_j = (\sum_{i=1}^n X_{j,i})/n$ for all $1 \leq j \leq m$. As $m \geq 2$, we can check that the unique optimal solutions for the above optimization problem are $\theta_j = ((m^2n+1)\bar{X}_j + \sum_{i \neq j} \bar{X}_i)/(m^2n+m)$ for any $1 \leq j \leq m$. If we further assume that our data $X_{j,i}$ are i.i.d samples from probability measure P^j having mean $\mu_j = E_{X \sim P^j}(X)$ for any

$1 \leq j \leq m$, the previous result implies that $\theta_i \not\rightarrow \theta_j$ for almost surely as long as $\mu_i \neq \mu_j$. As a consequence, if μ_j are pairwise different, the multi-level Wasserstein means under that simple scenario of (6.5) will not have identical centers among local groups.

On the other hand, we have $W_2^2(G_i, G_j) = \|\theta_i - \theta_j\|^2 = \left(\frac{mn}{mn+1}\right)^2 \|\bar{X}_i - \bar{X}_j\|^2$.

Now, from the definition of Wasserstein distance

$$\begin{aligned} W_2^2(P_n^i, P_n^j) &= \min_{\sigma} \frac{1}{n} \sum_{l=1}^n \|X_{i,l} - X_{j,\sigma(l)}\|^2 \\ &\geq \|\bar{X}_i - \bar{X}_j\|^2, \end{aligned}$$

where σ in the above sum varies over all the permutation of $\{1, 2, \dots, n\}$ and the second inequality is due to Cauchy-Schwarz's inequality. It implies that as long as $W_2^2(P_n^i, P_n^j)$ is small, the optimal solution G_i and G_j of (6.6) will be sufficiently close to each other. By letting $n \rightarrow \infty$, we also achieve the same conclusion regarding the asymptotic behavior of G_i and G_j with respect to $W_2(P^i, P^j)$.

Example 2. $k_j = 1$ and $n_j = n$ for all $1 \leq j \leq m$ and $M = 2$. Write $\mathbf{H} = (H_1, H_2)$.

Moreover, assume that there is a strict subset A of $\{1, 2, \dots, m\}$ such that

$$\max \left\{ \max_{i,j \in A} W_2(P_n^i, P_n^j), \right. \\ \left. \max_{i,j \in A^c} W_2(P_n^i, P_n^j) \right\} \ll \min_{i \in A, j \in A^c} W_2(P_n^i, P_n^j),$$

i.e., the distances of empirical measures P_n^i and P_n^j when i and j belong to the same set A or A^c are much less than those when i and j do not belong to the same set. Under this condition, by using the argument from part (i) we can write the objective function (6.5) as

$$\inf_{\substack{\theta_j \in \Theta, \\ H_1 \in \mathcal{P}_2(\Theta)}} \sum_{j \in A} \sum_{i=1}^n \|\theta_j - X_{j,i}\|^2 + \frac{W_2^2(\delta_{\theta_j}, H_1)}{|A|} +$$

$$\inf_{\substack{\theta_j \in \Theta, \\ H_2 \in \mathcal{P}_2(\Theta)}} \sum_{j \in A^c} \sum_{i=1}^n \|\theta_j - X_{j,i}\|^2 + \frac{W_2^2(\delta_{\theta_j}, H_2)}{|A^c|}.$$

The above objective function suggests that the optimal solutions θ_i, θ_j (equivalently, G_i and G_j) will not be close to each other as long as i and j do not belong to the same set A or A^c , i.e., P_n^i and P_n^j are very far. Therefore, the two groups of “local” measures G_j do not share atoms under that setting of empirical measures.

The examples examined above indicate that the MWM problem in general do not “encourage” the local measures G_j to share atoms among each other in its solution. Additionally, when the empirical measures of local groups are very close, it may also suggest that they belong to the same cluster and the distances among optimal local measures G_j can be very small.

6.3.1.2 Algorithm Description

Now we are ready to describe our algorithm in the general case. This is a procedure for finding a local minimum of Problem (6.5) and is summarized in Algorithm 1. We prepare the following details regarding the initialization and updating steps required by the algorithm:

- The initialization of local measures $G_j^{(0)}$ (i.e., the initialization of their atoms and weights) can be obtained by performing K -means clustering on local data $X_{j,i}$ for $1 \leq j \leq m$. The initialization of elements $H_i^{(0)}$ of $H^{(0)}$ is based on a simple extension of the K-means algorithm. Details are given in Algorithm 3 in the Supplement;
- The updates $G_j^{(t+1)}$ can be computed efficiently by simply using algorithms from [Cuturi and Doucet \[2014\]](#) to search for local solutions of these barycenter

Algorithm 1 Multilevel Wasserstein Means (MWM)

Input: Data $X_{j,i}$, Parameters k_j, M .
Output: prob. measures G_j and elements H_i of \mathbf{H} .
Initialize measures $G_j^{(0)}$, elements $H_i^{(0)}$ of $\mathbf{H}^{(0)}$, $t = 0$.

while $Y_j^{(t)}, b_j^{(t)}, H_i^{(t)}$ have not converged **do**

1. Update $Y_j^{(t)}$ and $b_j^{(t)}$ for $1 \leq j \leq m$:
- for** $j = 1$ **to** m **do**
- $i_j \leftarrow \arg \min_{1 \leq u \leq M} W_2^2(G_j^{(t)}, H_u^{(t)})$.
- $G_j^{(t+1)} \leftarrow \arg \min_{G_j \in \mathcal{O}_{k_j}(\Theta)} W_2^2(G_j, P_{n_j}^j) +$
- $+ W_2^2(G_j, H_{i_j}^{(t)})/m$.
- end for**
2. Update $H_i^{(t)}$ for $1 \leq i \leq M$:
- for** $j = 1$ **to** m **do**
- $i_j \leftarrow \arg \min_{1 \leq u \leq M} W_2^2(G_j^{(t+1)}, H_u^{(t)})$.
- end for**
- for** $i = 1$ **to** M **do**
- $C_i \leftarrow \{l : i_l = i\}$ for $1 \leq l \leq M$.
- $H_i^{(t+1)} \leftarrow \arg \min_{H_i \in \mathcal{P}_2(\Theta)} \sum_{l \in C_i} W_2^2(H_i, G_l^{(t+1)})$.
- end for**
3. $t \leftarrow t + 1$.

end while

problems within the space $\mathcal{O}_{k_j}(\Theta)$ from the atoms and weights of $G_j^{(t)}$;

- Since all $G_j^{(t+1)}$ are finite discrete measures, finding the updates for $H_i^{(t+1)}$ over the whole space $\mathcal{P}_2(\Theta)$ can be reduced to searching for a local solution within space $\mathcal{O}_{l^{(t)}}$ where $l^{(t)} = \sum_{j \in C_i} |\text{supp}(G_j^{(t+1)})| - |C_i|$ from the global atoms $H_i^{(t)}$ of $\mathbf{H}^{(t)}$ (Justification of this reduction is derived from Theorem A.1 in the Supplement). This again can be done by utilizing algorithms from Cuturi and Doucet [2014]. Note that, as $l^{(t)}$ becomes very large when m is large, to speed up the computation of Algorithm 1 we impose a threshold L , e.g., $L = 10$, for $l^{(t)}$ in its implementation.

The following guarantee for Algorithm 1 can be established:

Theorem 6.3.1. *Algorithm 1 monotonically decreases the objective function (6.4) of the MWM formulation.*

6.3.2 Multilevel Wasserstein Means with Sharing

As we have observed from the analysis of several specific cases, the **multilevel Wasserstein means** formulation may not encourage the sharing components locally among m groups in its solution. However, enforced sharing has been demonstrated to be a very useful technique, which leads to the “borrowing of strength” among different parts of the model, consequentially improving the inferential efficiency [Teh et al., 2006, Nguyen, 2016]. In this section, we seek to encourage the borrowing of strength among groups by imposing additional constraints on the atoms of G_1, \dots, G_m in the original MWM formulation (6.4). Denote $\mathcal{A}_{M,\mathcal{S}_K} = \left\{ G_j \in \mathcal{O}_K(\Theta), \mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta)) : \text{supp}(G_j) \subseteq \mathcal{S}_K \forall 1 \leq j \leq m \right\}$ for any given $K, M \geq 1$ where the constraint set \mathcal{S}_K has exactly K elements. To simplify the exposition, let us assume that $k_j = K$ for all $1 \leq j \leq m$. Consider the following locally constrained version of the multilevel Wasserstein means problem

$$\inf \sum_{j=1}^m W_2^2(G_j, P_{n_j}^j) + W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^m \delta_{G_j}). \quad (6.7)$$

where $\mathcal{S}_K, G_j, \mathcal{H} \in \mathcal{A}_{M,\mathcal{S}_K}$ in the above infimum. We call the above optimization the problem of *Multilevel Wasserstein Means with Sharing (MWMS)*. The local constraint assumption $\text{supp}(G_j) \subseteq \mathcal{S}_K$ had been utilized previously in the literature — see for example the work of [Kulis and Jordan, 2012], who developed an optimization-based approach to the inference of the HDP [Teh et al., 2006], which also encourages explicitly the sharing of local group means among local clusters. Now, we can rewrite objective function (6.7) as follows

$$\inf_{\mathcal{S}_K, G_j, \mathbf{H} \in \mathcal{B}_{M, \mathcal{S}_K}} \sum_{j=1}^m W_2^2(G_j, P_{n_j}^j) + \frac{d_{W_2}^2(G_j, \mathbf{H})}{m} \quad (6.8)$$

where $\mathcal{B}_{M, \mathcal{S}_K} = \left\{ G_j \in \mathcal{O}_K(\Theta), \mathbf{H} = (H_1, \dots, H_M) : \text{supp}(G_j) \subseteq \mathcal{S}_K \forall 1 \leq j \leq m \right\}$.

The high level idea of finding local minimums of objective function (6.8) is to first, update the elements of constraint set \mathcal{S}_K to provide the supports for local measures G_j and then, obtain the weights of these measures as well as the elements of global set H by computing appropriate Wasserstein barycenters. Due to space constraint, the details of these steps of the MWMS Algorithm (Algorithm 2) are deferred to the Supplement.

6.4 Consistency results

We proceed to establish consistency for the estimators introduced in the previous section. For the brevity of the presentation, we only focus on the MWM method; consistency for MWMS can be obtained in a similar fashion. Fix m , and assume that P^j is the true distribution of data $X_{j,i}$ for $j = 1, \dots, m$. Write $\mathbf{G} = (G_1, \dots, G_m)$ and $\mathbf{n} = (n_1, \dots, n_m)$. We say $\mathbf{n} \rightarrow \infty$ if $n_j \rightarrow \infty$ for $j = 1, \dots, m$. Define the following functions

$$f_{\mathbf{n}}(\mathbf{G}, \mathcal{H}) = \sum_{j=1}^m W_2^2(G_j, P_{n_j}^j) + W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^m \delta_{G_j}),$$

$$f(\mathbf{G}, \mathcal{H}) = \sum_{j=1}^m W_2^2(G_j, P^j) + W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^m \delta_{G_j}),$$

where $G_j \in \mathcal{O}_{k_j}(\Theta)$, $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta))$ as $1 \leq j \leq m$. The first consistency property of the WMW formulation:

Theorem 6.4.1. *Given that $P^j \in \mathcal{P}_2(\Theta)$ for $1 \leq j \leq m$. Then, there holds almost surely, as $\mathbf{n} \rightarrow \infty$*

$$\inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} f_{\mathbf{n}}(\mathbf{G}, \mathcal{H}) - \inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} f(\mathbf{G}, \mathcal{H}) \rightarrow 0.$$

The next theorem establishes that the “true” global and local clusters can be recovered. To this end, assume that for each \mathbf{n} there is an optimal solution $(\widehat{G}_1^{n_1}, \dots, \widehat{G}_m^{n_m}, \widehat{\mathcal{H}}^{\mathbf{n}})$ or in short $(\widehat{\mathbf{G}}^{\mathbf{n}}, \widehat{\mathcal{H}}^{\mathbf{n}})$ of the objective function (6.4). Moreover, there exist a (not necessarily unique) optimal solution minimizing $f(\mathbf{G}, \mathcal{H})$ over $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$. Let \mathcal{F} be the collection of such optimal solutions. For any $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$, define

$$d(\mathbf{G}, \mathcal{H}, \mathcal{F}) = \inf_{(\mathbf{G}^0, \mathcal{H}^0) \in \mathcal{F}} \sum_{j=1}^m W_2^2(G_j, G_j^0) + W_2^2(\mathcal{H}, \mathcal{H}^0).$$

Given the above assumptions, we have the following result regarding the convergence of $(\widehat{\mathbf{G}}^{\mathbf{n}}, \widehat{\mathcal{H}}^{\mathbf{n}})$:

Theorem 6.4.2. *Assume that Θ is bounded and $P^j \in \mathcal{P}_2(\Theta)$ for all $1 \leq j \leq m$. Then, we have $d(\widehat{\mathbf{G}}^{\mathbf{n}}, \widehat{\mathcal{H}}^{\mathbf{n}}, \mathcal{F}) \rightarrow 0$ as $\mathbf{n} \rightarrow \infty$ almost surely.*

Remark: (i) The assumption Θ is bounded is just for the convenience of proof argument. We believe that the conclusion of this theorem may still hold when $\Theta = \mathbb{R}^d$. (ii) If $|\mathcal{F}| = 1$, i.e., there exists an unique optimal solution $\mathbf{G}^0, \mathcal{H}^0$ minimizing $f(\mathbf{G}, \mathcal{H})$ over $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$, the result of Theorem 6.4.2 implies that $W_2(\widehat{G}_j^{n_j}, G_j^0) \rightarrow 0$ for $1 \leq j \leq m$ and $W_2(\widehat{\mathcal{H}}^{\mathbf{n}}, \mathcal{H}^0) \rightarrow 0$ as $\mathbf{n} \rightarrow \infty$.

6.5 Empirical studies

6.5.1 Synthetic data

In this section, we are interested in evaluating the effectiveness of both MWM and MWMS clustering algorithms by considering different synthetic data generating processes. Unless otherwise specified, we set the number of groups $m = 50$, number

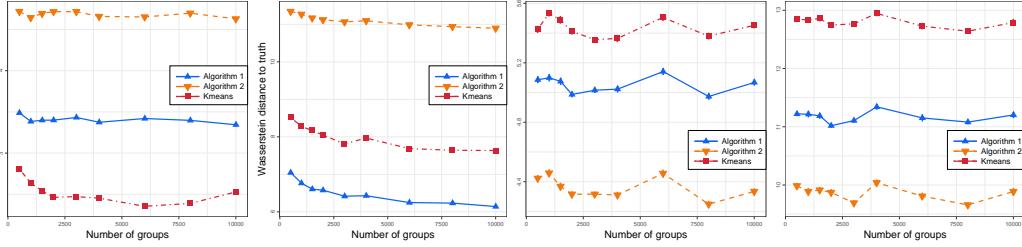


Figure 6.1: Data with a lot of small groups: (a) NC data with constant variance; (b) NC data with non-constant variance; (c) LC data with constant variance; (d) LC data with non-constant variance

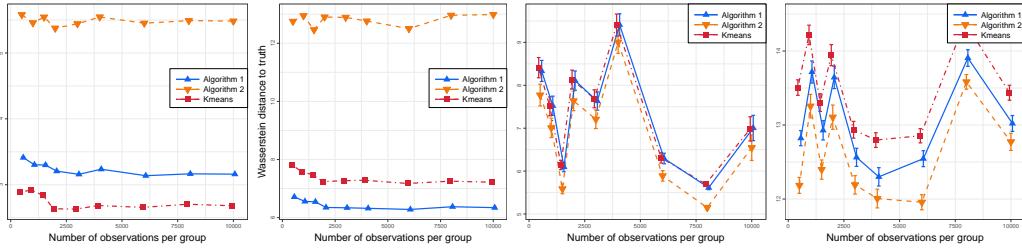


Figure 6.2: Data with few big groups: (a) NC data with constant variance; (b) NC data with non-constant variance; (c) LC data with constant variance; (d) LC data with non-constant variance

of observations per group $n_j = 50$ in $d = 10$ dimensions, number of global clusters $M = 5$ with 6 atoms. For Algorithm 1 (MWM) local measures G_j have 5 atoms each; for Algorithm 2 (MWMS) number of atoms in constraint set S_K is 50. As a benchmark for the comparison we will use a basic 3-stage K-means approach (the details of which can be found in the Supplement). The Wasserstein distance between the estimated distributions (i.e. $\hat{G}_1, \dots, \hat{G}_m; \hat{H}_1, \dots, \hat{H}_M$) and the data generating ones will be used as the comparison metric.

Recall that the MWM formulation does not impose constraints on the atoms of G_i , while the MWMS formulation explicitly enforces the sharing of atoms across these measures. We used multiple layers of mixtures while adding Gaussian noise at each layer to generate global and local clusters and the no-constraint (NC) data. We varied number of groups m from 500 to 10000. We notice that the 3-stage K-means algorithm performs the best when there is no constraint structure *and* variance is constant across clusters (Fig. 6.1(a) and 6.2(a)) — this is, not surprisingly, a favorable setting for the

basic K-means method. As soon as we depart from the (unrealistic) constant-variance, no-sharing assumption, both of our algorithms start to outperform the basic three-stage K-means. The superior performance is most pronounced with local-constraint (LC) data (with or without constant variance conditions). See Fig. 6.1(c,d). It is worth noting that even when group variances are constant, the 3-stage K-means is no longer longer effective because now fails to account for the shared structure. When $m = 50$ and group sizes are larger, we set $S_K = 15$. Results are reported in Fig. 6.2 (c), (d). These results demonstrate the effectiveness and flexibility of our both algorithms.

6.5.2 Real data analysis

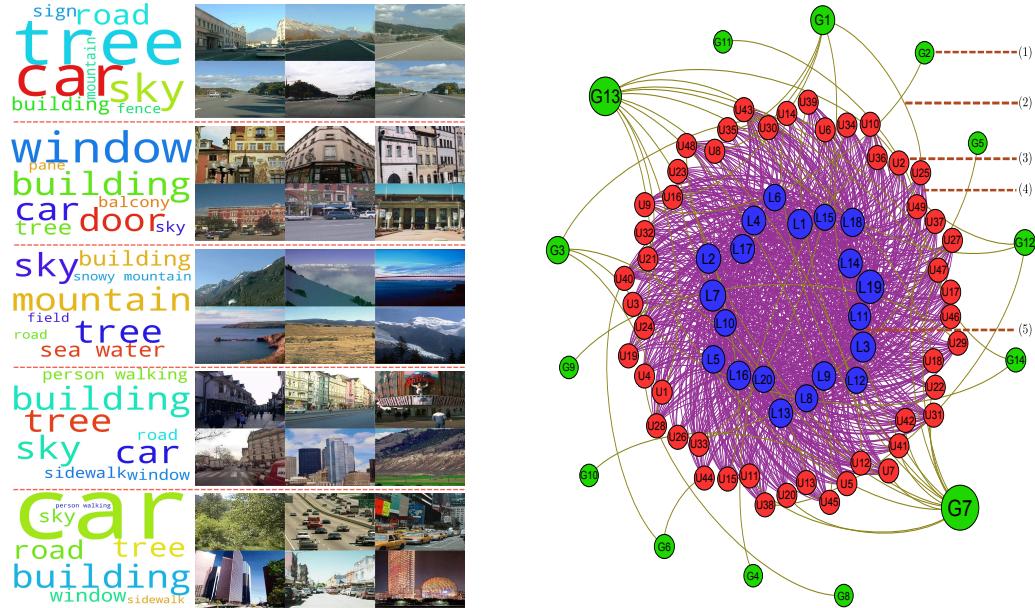


Figure 6.3: Clustering representation for two datasets: (a) Five image clusters from *Labelme* data discovered by MWMS algorithm: tag-clouds on the left are accumulated from all images in the clusters while six images on the right are randomly chosen images in that cluster; (b) StudentLife discovered network with three node groups: (1) discovered student clusters, (3) student nodes, (5) discovered activity location (from Wifi data); and two edge groups: (2) Student to cluster assignment, (4) Student involved to activity location. Node sizes (of discovered nodes) depict the number of element in clusters while edge sizes between *Student* and *activity location* represent the popularity of student's activities.

We applied our multilevel clustering algorithms to two real-world datasets: LabelMe and StudentLife.

LabelMe dataset consists of 2,688 annotated images which are classified into 8 scene categories including *tall buildings*, *inside city*, *street*, *highway*, *coast*, *open country*, *mountain*, and *forest* Oliva and Torralba [2001] . Each image contains multiple annotated regions. Each region, which is annotated by users, represents an object in the image. As shown in Figure 6.4, the left image is an image from *open country* category and contains 4 regions while the right panel denotes an image of *tall buildings* category including 16 regions. Note that the regions in each image can be overlapped. We remove the images containing less than 4 regions and obtain 1,800 images.

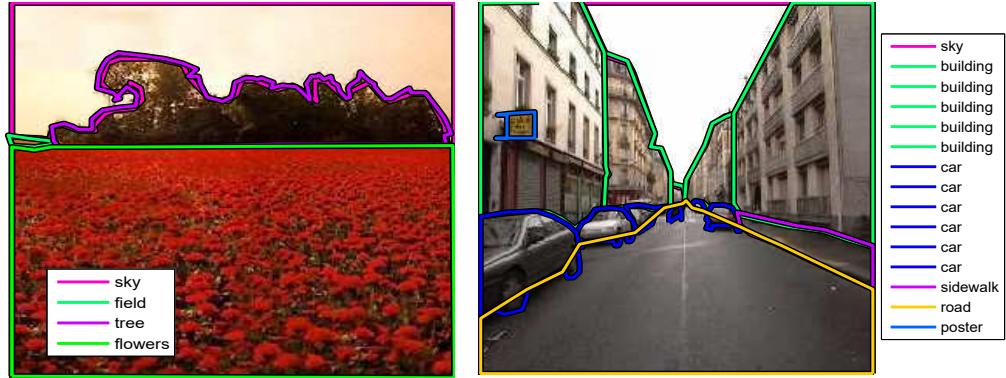


Figure 6.4: Examples of images used in LabelMe dataset. Each image consists of different annotated regions.

We then extract GIST feature Oliva and Torralba [2001] for each region in a image. GIST is a visual descriptor to represent perceptual dimensions and oriented spatial structures of a scene. Each GIST descriptor is a 512-dimensional vector. We further use PCA to project GIST features into 30 dimensions. Finally, we obtain 1,800 “documents”, each of which contains regions as observations. Each region now is represented by a 30-dimensional vector. We now can perform clustering regions in every image since they are visually correlated. In the next level of clustering, we can cluster images into scene categories.

Table 6.1: Clustering performance for LabelMe dataset.

Methods	NMI	ARI	AMI	Time (s)
K-means	0.349	0.237	0.324	0.3
TSK-means	0.236	0.112	0.22	218
MC2	0.315	0.206	0.273	4.2
MWM	0.373	0.263	0.352	332
MWMS	0.391	0.284	0.368	544

StudentLife dataset is a large dataset frequently used in pervasive and ubiquitous computing research. Data signals consist of multiple channels (e.g., WiFi signals, Bluetooth scan, etc.), which are collected from smartphones of 49 students at Dartmouth College over a 10-week spring term in 2013. However, in our experiments, we use only WiFi signal strengths. We applied a similar procedure described in [Nguyen et al. \[2016\]](#) to pre-process the data. We aggregate the number of scans by each Wifi access point and select 500 Wifi Ids with the highest frequencies. Eventually, we obtain 49 “documents” with totally approximately 4.6 million 500-dimensional data points.

Experimental results. To quantitatively evaluate our proposed methods, we compare our algorithms with several base-line methods: K-means, three-stage K-means (TSK-means) as described in the Supplement, MC2-SVI without context [Huynh et al. \[2016\]](#). Clustering performance in Table 6.1 is evaluated with the image clustering problem for *LabelMe dataset*. With *K-means*, we average all data points to obtain a single vector for each images. K-means needs much less time to run since the number of data points is now reduced to 1,800. For MC2-SVI, we used stochastic variational and a parallelized Spark-based implementation in [Huynh et al. \[2016\]](#) to carry out experiments. This implementation has the advantage of making use of all of 16 cores on the test machine. The running time for MC2-SVI is reported after scanning one epoch. In terms of clustering accuracy, MWM and MWMS algorithms perform the best.

Fig. 6.3 demonstrates five representative image clusters with six randomly chosen images in each (on the right) which are discovered by our MWMS algorithm. We also accumulate labeled tags from all images in each cluster to produce the tag-cloud on the left. These tag-clouds can be considered as visual ground truth of clusters. Our algorithm can group images into clusters which are consistent with their tag-clouds.

We use StudentLife dataset to demonstrate the capability of multilevel clustering with large-scale datasets. This dataset not only contains a large number of data points but presents in high dimension. Our algorithms need approximately 1 hour to perform multilevel clustering on this dataset. Fig. 6.3 presents two levels of clusters discovered by our algorithms. The innermost (blue) and outermost (green) rings depict local and global clusters respectively. Global clusters represent groups of students while local clusters shared between students (“documents”) may be used to infer locations of students’ activities. From these clusteing we can dissect students’ shared location (activities), e.g. Student 49 (*U49*) mainly takes part in activity location 4 (*L4*).

6.6 Discussion

We have proposed an optimization based approach to multilevel clustering using Wasserstein metrics. There are several possible directions for extensions. Firstly, we have only considered continuous data; it is of interest to extend our formulation to discrete data. Secondly, our method requires knowledge of the numbers of clusters both in local and global clustering. When these numbers are unknown, it seems reasonable to incorporate penalty on the model complexity. Thirdly, our formulation does not directly account for the “noise” distribution away from the (Wasserstein) means. To improve the robustness, it may be desirable to make use of the first-order Wasserstein metric instead of the second-order one. Finally, we are interested in extending our approach to richer settings of hierarchical data, such as one when group level-context is available.

6.7 Appendix A

In this appendix, we collect relevant information on the Wasserstein metric and Wasserstein barycenter problem, which were introduced in Section 6.2 in this chapter. For any Borel map $g : \Theta \rightarrow \Theta$ and probability measure G on Θ , the push-forward measure of G through g , denoted by $g\#G$, is defined by the condition that $\int_{\Theta} f(y)d(g\#G)(y) = \int_{\Theta} f(g(x))dG(x)$ for every continuous bounded function f on Θ .

Wasserstein metric When $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ and $G' = \sum_{i=1}^{k'} p'_i \delta_{\theta'_i}$ are discrete measures with finite support, i.e., k and k' are finite, the Wasserstein distance of order r between G and G' can be represented as

$$W_r^r(G, G') = \min_{T \in \Pi(G, G')} \langle T, M_{G, G'} \rangle \quad (6.9)$$

where we have

$$\Pi(G, G') = \left\{ T \in \mathbb{R}_+^{k \times k'} : T \mathbb{1}_{k'} = \mathbf{p}, \quad T \mathbb{1}_k = \mathbf{p}' \right\}$$

such that $\mathbf{p} = (p_1, \dots, p_k)^T$ and $\mathbf{p}' = (p'_1, \dots, p'_{k'})^T$, $M_{G, G'} = \{\|\theta_i - \theta'_j\|\}_{i,j} \in \mathbb{R}_+^{k \times k'}$ is the cost matrix, i.e. matrix of pairwise distances of elements between G and G' , and $\langle A, B \rangle = \text{tr}(A^T B)$ is the Frobenius dot-product of matrices. The optimal $T \in \Pi(G, G')$ in optimization problem (6.9) is called the optimal coupling of G and G' , representing the **optimal transport** between these two measures. When $k = k'$, the complexity of best algorithms for finding the optimal transport is $O(k^3 \log k)$. Currently, [Cuturi \[2013\]](#) proposed a regularized version of (6.9) based on Sinkhorn distance where the complexity of finding an approximation of the optimal transport is $O(k^2)$. Due to its favorably fast computation, throughout the chapter we shall utilize Cuturi's algorithm to compute the Wasserstein distance between G and G' as well as

their optimal transport in (6.9).

Wasserstein barycenter As introduced in Section 6.2 in this chapter, for any probability measures $P_1, P_2, \dots, P_N \in \mathcal{P}_2(\Theta)$, their Wasserstein barycenter $\bar{P}_{N,\lambda}$ is such that

$$\bar{P}_{N,\lambda} = \arg \min_{P \in \mathcal{P}_2(\Theta)} \sum_{i=1}^N \lambda_i W_2^2(P, P_i)$$

where $\lambda \in \Delta_N$ denote weights associated with P_1, \dots, P_N . According to [Agueh and Carlier, 2011], $P_{N,\lambda}$ can be obtained as a solution to so-called multi-marginal optimal transporation problem. In fact, if we denote T_k^1 as the measure preseving map from P_1 to P_k , i.e., $P_k = T_k^1 \# P_1$, for any $1 \leq k \leq N$, then

$$\bar{P}_{N,\lambda} = \left(\sum_{k=1}^N \lambda_k T_k^1 \right) \# P_1.$$

Unfortunately, the forms of the maps T_k^1 are analytically intractable, especially if no special constraints on P_1, \dots, P_N are imposed.

Recently, [Anderes et al., 2015] studied the Wasserstein barycenters $\bar{P}_{N,\lambda}$ when P_1, P_2, \dots, P_N are finite discrete measures and $\lambda = (1/N, \dots, 1/N)$. They demonstrate the following sharp result (cf. Theorem 2 in [Anderes et al., 2015]) regarding the number of atoms of $\bar{P}_{N,\lambda}$

Theorem A.1. *There exists a Wasserstein barycenter $\bar{P}_{N,\lambda}$ such that $\text{supp}(\bar{P}_{N,\lambda}) \leq \sum_{i=1}^N s_i - N + 1$.*

Therefore, when P_1, \dots, P_N are indeed finite discrete measures and the weights are uniform, the problem of finding Wasserstein barycenter $\bar{P}_{N,\lambda}$ over the (computationally large) space $\mathcal{P}_2(\Theta)$ is reduced to a search over a smaller space $\mathcal{O}_l(\Theta)$ where $l = \sum_{i=1}^N s_i - N + 1$.

6.8 Appendix B

In this appendix, we provide proofs for the remaining results in this chapter. We start by giving a proof for the transition from multilevel Wasserstein means objective function (6.4) to objective function (6.5) in Section 6.3.1 in this chapter. All the notations in this appendix are similar to those in the main text. For each closed subset $\mathcal{S} \subset \mathcal{P}_2(\Theta)$, denote the Voronoi region generated by \mathcal{S} on the space $\mathcal{P}_2(\Theta)$ by the collection of subsets $\{V_P\}_{P \in \mathcal{S}}$, where $V_P := \{Q \in \mathcal{P}_2(\Theta) : W_2^2(Q, P) = \min_{G \in \mathcal{S}} W_2^2(Q, G)\}$. We define the projection mapping $\pi_{\mathcal{S}}$ as: $\pi_{\mathcal{S}} : \mathcal{P}_2(\Theta) \rightarrow \mathcal{S}$ where $\pi_{\mathcal{S}}(Q) = P$ as $Q \in V_P$. Note that, for any $P_1, P_2 \in \mathcal{S}$ such that V_{P_1} and V_{P_2} share the boundary, the values of $\pi_{\mathcal{S}}$ at the elements in that boundary can be chosen to be either P_1 or P_2 . Now, we start with the following useful lemmas.

Lemma B.1. *For any closed subset \mathcal{S} on $\mathcal{P}_2(\Theta)$, if $\mathcal{Q} \in \mathcal{P}_2(\mathcal{P}_2(\Theta))$, then*

$$E_{X \sim \mathcal{Q}}(d_{W_2}^2(X, \mathcal{S})) = W_2^2(\mathcal{Q}, \pi_{\mathcal{S}} \# \mathcal{Q})$$

where $d_{W_2}^2(X, \mathcal{S}) = \inf_{P \in \mathcal{S}} W_2^2(X, P)$.

Proof. For any element $\pi \in \Pi(\mathcal{Q}, \pi_{\mathcal{S}} \# \mathcal{Q})$:

$$\begin{aligned} \int W_2^2(P, G) d\pi(P, G) &\geq \int d_{W_2}^2(P, \mathcal{S}) d\pi(P, G) \\ &= \int d_{W_2}^2(P, \mathcal{S}) d\mathcal{Q}(P) \\ &= E_{X \sim \mathcal{Q}}(d_{W_2}^2(X, \mathcal{S})) \end{aligned}$$

where the integrations in the first two terms range over $\mathcal{P}_2(\Theta) \times \mathcal{S}$ while that in the

final term ranges over $\mathcal{P}_2(\Theta)$. Therefore, we obtain

$$\begin{aligned} W_2^2(\mathcal{Q}, \pi_{\mathcal{S}} \# \mathcal{Q}) &= \inf_{\mathcal{P}_2(\Theta) \times \mathcal{S}} \int W_2^2(P, G) d\pi(P, G) \\ &\geq E_{X \sim \mathcal{Q}}(d_{W_2}^2(X, \mathcal{S})) \end{aligned} \quad (6.10)$$

where the infimum in the first equality ranges over all $\pi \in \Pi(\mathcal{Q}, \pi_{\mathcal{S}} \# \mathcal{Q})$.

On the other hand, let $g : \mathcal{P}_2(\Theta) \rightarrow \mathcal{P}_2(\Theta) \times \mathcal{S}$ such that $g(P) = (P, \pi_{\mathcal{S}}(P))$ for all $P \in \mathcal{P}_2(\Theta)$. Additionally, let $\mu_{\pi_{\mathcal{S}}} = g \# \mathcal{Q}$, the push-forward measure of \mathcal{Q} under mapping g . It is clear that $\mu_{\pi_{\mathcal{S}}}$ is a coupling between \mathcal{Q} and $\pi_{\mathcal{S}} \# \mathcal{Q}$. Under this construction, we obtain for any $X \sim \mathcal{Q}$ that

$$\begin{aligned} E(W_2^2(X, \pi_{\mathcal{S}}(X))) &= \int W_2^2(P, G) d\mu_{\pi_{\mathcal{S}}}(P, G) \\ &\geq \inf \int W_2^2(P, G) d\pi(P, G) \\ &= W_2^2(\mathcal{Q}, \pi_{\mathcal{S}} \# \mathcal{Q}) \end{aligned} \quad (6.11)$$

where the infimum in the second inequality ranges over all $\pi \in \Pi(\mathcal{Q}, \pi_{\mathcal{S}} \# \mathcal{Q})$ and the integrations range over $\mathcal{P}_2(\Theta) \times \mathcal{S}$. Now, from the definition of $\pi_{\mathcal{S}}$

$$\begin{aligned} E(W_2^2(X, \pi_{\mathcal{S}}(X))) &= \int W_2^2(P, \pi_{\mathcal{S}}(P)) d\mathcal{Q}(P) \\ &= \int d_{W_2}^2(P, \mathcal{S}) d\mathcal{Q}(P) \\ &= E(d_{W_2}^2(X, \mathcal{S})) \end{aligned} \quad (6.12)$$

where the integrations in the above equations range over $\mathcal{P}_2(\Theta)$. By combining (6.11) and (6.12), we would obtain that

$$E_{X \sim \mathcal{Q}}(d_{W_2}^2(X, \mathcal{S})) \geq W_2^2(\mathcal{Q}, \pi_{\mathcal{S}} \# \mathcal{Q}). \quad (6.13)$$

From (6.10) and (6.13), it is straightforward that $E_{X \sim Q}(d(X, S)^2) = W_2^2(Q, \pi_S \# Q)$.

Therefore, we achieve the conclusion of the lemma. \square

Lemma B.2. *For any closed subset $\mathcal{S} \subset \mathcal{P}_2(\Theta)$ and $\mu \in \mathcal{P}_2(\mathcal{P}_2(\Theta))$ with $\text{supp}(\mu) \subseteq \mathcal{S}$, there holds $W_2^2(\mathcal{Q}, \mu) \geq W_2^2(\mathcal{Q}, \pi_{\mathcal{S}} \# \mathcal{Q})$ for any $\mathcal{Q} \in \mathcal{P}_2(\mathcal{P}_2(\Theta))$.*

Proof. Since $\text{supp}(\mu) \subseteq \mathcal{S}$, it is clear that $W_2^2(\mathcal{Q}, \mu) = \inf_{\pi \in \Pi(\mathcal{Q}, \mu)} \int_{\mathcal{P}_2(\Theta) \times \mathcal{S}} W_2^2(P, G) d\pi(P, G)$.

Additionally, we have

$$\begin{aligned} \int W_2^2(P, G) d\pi(P, G) &\geq \int d_{W_2}^2(P, \mathcal{S}) d\pi(P, G) \\ &= \int d_{W_2}^2(P, \mathcal{S}) d\mathcal{Q}(P) \\ &= E_{X \sim Q}(d_{W_2}^2(X, S)) \\ &= W_2^2(\mathcal{Q}, \pi_{\mathcal{S}} \# \mathcal{Q}) \end{aligned}$$

where the last inequality is due to Lemma B.1 and the integrations in the first two terms range over $\mathcal{P}_2(\Theta) \times \mathcal{S}$ while that in the final term ranges over $\mathcal{P}_2(\Theta)$. Therefore, we achieve the conclusion of the lemma. \square

Equipped with Lemma B.1 and Lemma B.2, we are ready to establish the equivalence between multilevel Wasserstein means objective function (5) and objective function (4) in Section 6.3.1 in the main text.

Lemma B.3. *For any given positive integers m and M , we have*

$$\begin{aligned} A &:= \inf_{\mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))} W_2^2(\mathcal{H}, \frac{1}{m} \sum_{j=1}^m \delta_{G_j}) \\ &= \frac{1}{m} \inf_{\mathbf{H}=(H_1, \dots, H_M)} \sum_{j=1}^m d_{W_2}^2(G_j, \mathbf{H}) := B. \end{aligned}$$

Proof. Write $\mathcal{Q} = \frac{1}{m} \sum_{j=1}^m \delta_{G_j}$. From the definition of B , for any $\epsilon > 0$, we can find $\overline{\mathbf{H}}$

such that

$$\begin{aligned}
B &\geq \frac{1}{m} \sum_{j=1}^m d_{W_2}^2(G_j, \bar{\mathbf{H}}) - \epsilon \\
&= E_{X \sim \mathcal{Q}}(d_{W_2}^2(X, \bar{\mathbf{H}})) - \epsilon \\
&= W_2^2(\mathcal{Q}, \pi_{\bar{\mathbf{H}}} \# \mathcal{Q}) - \epsilon \\
&\geq A - \epsilon
\end{aligned}$$

where the second equality in the above display is due to Lemma B.1 while the last inequality is from the fact that $\pi_{\bar{\mathbf{H}}} \# \mathcal{Q}$ is a discrete probability measure in $\mathcal{P}_2(\mathcal{P}_2(\Theta))$ with exactly M support points. Since the inequality in the above display holds for any ϵ , it implies that $B \geq A$. On the other hand, from the formation of A , for any $\epsilon > 0$, we also can find $\mathcal{H}' \in \mathcal{E}_M(\mathcal{P}_2(\Theta))$ such that

$$\begin{aligned}
A &\geq W_2^2(\mathcal{H}', \mathcal{Q}) - \epsilon \\
&\geq W_2^2(\mathcal{Q}, \pi_{\mathbf{H}'} \# \mathcal{Q}) - \epsilon \\
&= \frac{1}{m} \sum_{j=1}^m d_{W_2}^2(G_j, \mathbf{H}') - \epsilon \\
&\geq B - \epsilon
\end{aligned}$$

where $\mathbf{H}' = \text{supp}(\mathcal{H}')$, the second inequality is due to Lemma B.2, and the third equality is due to Lemma B.1. Therefore, it means that $A \geq B$. We achieve the conclusion of the lemma. \square

Proposition B.4. *For any positive integer numbers m, M and k_j as $1 \leq j \leq m$, we*

denote

$$\begin{aligned}
C &:= \inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta) \forall 1 \leq j \leq m, \\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} \sum_{i=1}^m W_2^2(G_j, P_{n_j}^j) \\
&\quad + W_2^2(\mathcal{H}, \frac{1}{m} \sum_{i=1}^m \delta_{G_i}) \\
D &:= \inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta) \forall 1 \leq j \leq m, \\ \mathbf{H} = (H_1, \dots, H_M)}} \sum_{j=1}^m W_2^2(G_j, P_{n_j}^j) \\
&\quad + \frac{d_{W_2}^2(G_j, \mathbf{H})}{m}.
\end{aligned}$$

Then, we have $C = D$.

Proof. The proof of this proposition is a straightforward application of Lemma B.3. Indeed, for each fixed (G_1, \dots, G_m) the infimum w.r.t to \mathcal{H} in C leads to the same infimum w.r.t to \mathbf{H} in D , according to Lemma B.3. Now, by taking the infimum w.r.t to (G_1, \dots, G_m) on both sides, we achieve the conclusion of the proposition. \square

In the remainder of the Supplement, we present the proofs for all remaining theorems stated in the main text.

PROOF OF THEOREM 6.3.1 The proof of this theorem is straightforward from the formulation of Algorithm 1. In fact, for any $G_j \in \mathcal{E}_{k_j}(\Theta)$ and $\mathbf{H} = (H_1, \dots, H_M)$, we denote the function

$$f(\mathbf{G}, \mathbf{H}) = \sum_{j=1}^m W_2^2(G_j, P_n^j) + \frac{d_{W_2}^2(G_j, \mathbf{H})}{m}$$

where $\mathbf{G} = (G_1, \dots, G_m)$. To obtain the conclusion of this theorem, it is sufficient to demonstrate for any $t \geq 0$ that

$$f(\mathbf{G}^{(t+1)}, \mathbf{H}^{(t+1)}) \leq f(\mathbf{G}^{(t)}, \mathbf{H}^{(t)}).$$

This inequality comes directly from $f(\mathbf{G}^{(t+1)}, \mathbf{H}^{(t)}) \leq f(\mathbf{G}^{(t)}, \mathbf{H}^{(t)})$, which is due to the Wasserstein barycenter problems to obtain $G_j^{(t+1)}$ for $1 \leq j \leq m$, and $f(\mathbf{G}^{(t+1)}, \mathbf{H}^{(t+1)}) \leq f(\mathbf{G}^{(t+1)}, \mathbf{H}^{(t)})$, which is due to the optimization steps to achieve elements $H_u^{(t+1)}$ of $\mathbf{H}^{(t+1)}$ as $1 \leq u \leq M$. As a consequence, we achieve the conclusion of the theorem.

PROOF OF THEOREM 6.4.1 To simplify notation, write

$$\begin{aligned} L_{\mathbf{n}} &= \inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} f_{\mathbf{n}}(\mathbf{G}, \mathcal{H}), \\ L_0 &= \inf_{\substack{G_j \in \mathcal{O}_{k_j}(\Theta), \\ \mathcal{H} \in \mathcal{E}_M(\mathcal{P}_2(\Theta))}} f(\mathbf{G}, \mathcal{H}). \end{aligned}$$

For any $\epsilon > 0$, from the definition of L_0 , we can find $G_j \in \mathcal{O}_{k_j}(\Theta)$ and $\mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta))$ such that

$$f(\mathbf{G}, \mathcal{H})^{1/2} \leq L_0^{1/2} + \epsilon.$$

Therefore, we would have

$$\begin{aligned} L_{\mathbf{n}}^{1/2} - L_0^{1/2} &\leq L_n^{1/2} - f(\mathbf{G}, \mathcal{H})^{1/2} + \epsilon \\ &\leq f_{\mathbf{n}}(\mathbf{G}, \mathcal{H})^{1/2} - f(\mathbf{G}, \mathcal{H})^{1/2} + \epsilon \\ &= \frac{f_{\mathbf{n}}(\mathbf{G}, \mathcal{H}) - f(\mathbf{G}, \mathcal{H})}{f_{\mathbf{n}}(\mathbf{G}, \mathcal{H})^{1/2} + f(\mathbf{G}, \mathcal{H})^{1/2}} + \epsilon \\ &\leq \sum_{j=1}^m \frac{|W_2^2(G_j, P_{n_j}^j) - W_2^2(G_j, P^j)|}{W_2(G_j, P_{n_j}^j) + W_2(G_j, P^j)} + \epsilon \\ &\leq \sum_{j=1}^m W_2(P_{n_j}^j, P^j) + \epsilon. \end{aligned}$$

By reversing the direction, we also obtain the inequality $L_n^{1/2} - L_0^{1/2} \geq \sum_{j=1}^m W_2(P_{n_j}^j, P^j) - \epsilon$. Hence, $|L_n^{1/2} - L_0^{1/2} - \sum_{j=1}^m W_2(P_{n_j}^j, P^j)| \leq \epsilon$ for any $\epsilon > 0$. Since $P^j \in \mathcal{P}_2(\Theta)$ for all $1 \leq j \leq m$, we obtain that $W_2(P_{n_j}^j, P^j) \rightarrow 0$ almost surely as $n_j \rightarrow \infty$ (see for

example Theorem 6.9 in [Villani, 2009]). As a consequence, we obtain the conclusion of the theorem.

PROOF OF THEOREM 6.4.2 For any $\epsilon > 0$, we denote

$$\mathcal{A}(\epsilon) = \left\{ G_i \in \mathcal{O}_{k_i}(\Theta), \mathcal{H} \in \mathcal{E}_M(\mathcal{P}(\Theta)) : d(\mathbf{G}, \mathcal{H}, \mathcal{F}) \geq \epsilon \right\}.$$

Since Θ is a compact set, we also have $\mathcal{O}_{k_j}(\Theta)$ and $\mathcal{E}_M(\mathcal{P}_2(\Theta))$ are compact for any $1 \leq i \leq m$. As a consequence, $\mathcal{A}(\epsilon)$ is also a compact set. For any $(\mathbf{G}, \mathcal{H}) \in \mathcal{A}(\epsilon)$, by the definition of \mathcal{F} we would have $f(\mathbf{G}, \mathcal{H}) > f(\mathbf{G}^0, \mathcal{H}^0)$ for any $(\mathbf{G}^0, \mathcal{H}^0) \in \mathcal{F}$. Since $\mathcal{A}(\epsilon)$ is compact, it leads to

$$\inf_{(\mathbf{G}, \mathcal{H}) \in \mathcal{A}(\epsilon)} f(\mathbf{G}, \mathcal{H}) > f(\mathbf{G}^0, \mathcal{H}^0).$$

for any $(\mathbf{G}^0, \mathcal{H}^0) \in \mathcal{F}$. From the formulation of f_n as in the proof of Theorem 6.4.1, we can verify that $\lim_{n \rightarrow \infty} f_n(\widehat{\mathbf{G}}^n, \widehat{\mathcal{H}}^n) = \lim_{n \rightarrow \infty} f(\widehat{\mathbf{G}}^n, \widehat{\mathcal{H}}^n)$ almost surely as $n \rightarrow \infty$. Combining this result with that of Theorem 6.4.1, we obtain $f(\widehat{\mathbf{G}}^n, \widehat{\mathcal{H}}^n) \rightarrow f(\mathbf{G}^0, \mathcal{H}^0)$ as $n \rightarrow \infty$ for any $(\mathbf{G}^0, \mathcal{H}^0) \in \mathcal{F}$. Therefore, for any $\epsilon > 0$, as n is large enough, we have $d(\widehat{\mathbf{G}}^n, \widehat{\mathcal{H}}^n, \mathcal{F}) < \epsilon$. As a consequence, we achieve the conclusion regarding the consistency of the mixing measures.

6.9 Appendix C

In this appendix, we provide details on the algorithm for the Multilevel Wasserstein means with sharing (MWMS) formulation (Algorithm 2). Recall the MWMS

objective function as follows

$$\inf_{\mathcal{S}_K, G_j, \mathbf{H} \in \mathcal{B}_{M, \mathcal{S}_K}} \sum_{j=1}^m W_2^2(G_j, P_{n_j}^j) + \frac{d_{W_2}^2(G_j, \mathbf{H})}{m}$$

where $\mathcal{B}_{M, \mathcal{S}_K} = \left\{ G_j \in \mathcal{O}_K(\Theta), \mathbf{H} = (H_1, \dots, H_M) : \text{supp}(G_j) \subseteq \mathcal{S}_K \forall 1 \leq j \leq m \right\}$.

We make the following remarks regarding the initializations and updates of Algorithm 2:

- (i) An efficient way to initialize global set $S_K^{(0)} = \left\{ a_1^{(0)}, \dots, a_K^{(0)} \right\} \in \mathbb{R}^{d \times K}$ is to perform K -means on the whole data set $X_{j,i}$ for $1 \leq j \leq m, 1 \leq i \leq n_j$;
- (ii) The updates $a_j^{(t+1)}$ are indeed the solutions of the following optimization problems

$$\inf_{a_j^{(t)}} \left\{ \sum_{l=1}^m W_2^2(G_l^{(t)}, P_n^l) + \frac{\sum_{l=1}^m W_2^2(G_l^{(t)}, H_{i_l}^{(t)})}{m} \right\},$$

which is equivalent to find $a_j^{(t)}$ to optimize

$$\begin{aligned} & m \sum_{u=1}^m \sum_{v=1}^{n_j} T_{j,v}^u \|a_j^{(t)} - X_{u,v}\|^2 \\ & + \sum_{u=1}^m \sum_v U_{j,v}^u \|a_j^{(t)} - h_{i_j,v}^{(t)}\|^2. \end{aligned}$$

where T^j is an optimal coupling of $G_j^{(t)}$, P_n^j and U^j is an optimal coupling of $G_j^{(t)}$, $H_{i_j}^{(t)}$. By taking the first order derivative of the above function with respect to $a_j^{(t)}$, we quickly achieve $a_j^{(t+1)}$ as the closed form minimum of that function;

- (iii) Updating the local weights of $G_j^{(t+1)}$ is equivalent to updating $G_j^{(t+1)}$ as the atoms of $G_j^{(t+1)}$ are known to stem from $S_K^{(t+1)}$.

Now, similar to Theorem 3.1 in the main text, we also have the following theoretical

Algorithm 2 Multilevel Wasserstein Means with Sharing (MWMS)

Input: Data $X_{j,i}$, K , M .

Output: global set S_K , local measures G_j , and elements H_i of \mathbf{H} .

Initialize $S_K^{(0)} = \{a_1^{(0)}, \dots, a_K^{(0)}\}$, elements $H_i^{(0)}$ of $\mathbf{H}^{(0)}$, and $t = 0$.

while $S_K^{(t)}, G_j^{(t)}, H_i^{(t)}$ have not converged **do**

 1. Update global set $S_K^{(t)}$:

for $j = 1$ **to** m **do**

$$i_j \leftarrow \arg \min_{1 \leq u \leq M} W_2^2(G_j^{(t)}, H_u^{(t)}).$$

$T^j \leftarrow$ optimal coupling of $G_j^{(t)}$, P_n^j (cf. Appendix A).

$U^j \leftarrow$ optimal coupling of $G_j^{(t)}$, $H_{i_j}^{(t)}$.

end for

for $i = 1$ **to** M **do**

$h_i^{(t)} \leftarrow$ atoms of $H_i^{(t)}$ with $h_{i,v}^{(t)}$ as v-th column.

end for

for $i = 1$ **to** K **do**

$$mD \leftarrow m \sum_{u=1}^m \sum_{v=1}^{n_i} T_{i,v}^u + \sum_{u=1}^m \sum_{v \neq i} U_{i,v}^u.$$

$$a_i^{(t+1)} \leftarrow \left(m \sum_{u=1}^m \sum_{v=1}^{n_i} T_{i,v}^u X_{u,v} + \sum_{u=1}^m \sum_v U_{i,v}^u h_{j_u,v}^{(t)} \right) / mD.$$

end for

 2. Update $G_j^{(t)}$ for $1 \leq j \leq m$:

for $j = 1$ **to** m **do**

$$G_j^{(t+1)} \leftarrow \arg \min_{G_j: \text{supp}(G_j) \equiv S_K^{(t+1)}} W_2^2(G_j, P_{n_j}^j)$$

$$+ W_2^2(G_j, H_{i_j}^{(t)}) / m.$$

end for

 3. Update $H_i^{(t)}$ for $1 \leq i \leq M$ as Algorithm 1.

 4. $t \leftarrow t + 1$.

end while

guarantee regarding the behavior of Algorithm 2 as follows

Theorem C.1. *Algorithm 2 monotonically decreases the objective function of the MWMS formulation.*

Proof. The proof is quite similar to the proof of Theorem 6.3.1. In fact, recall from the proof of Theorem 6.3.1 that for any $G_j \in \mathcal{E}_{k_j}(\Theta)$ and $\mathbf{H} = (H_1, \dots, H_M)$ we denote the function

$$f(\mathbf{G}, \mathbf{H}) = \sum_{j=1}^m W_2^2(G_j, P_n^j) + \frac{d_{W_2}^2(G_j, \mathbf{H})}{m}$$

where $\mathbf{G} = (G_1, \dots, G_m)$. Now it is sufficient to demonstrate for any $t \geq 0$ that

$$f(\mathbf{G}^{(t+1)}, \mathbf{H}^{(t+1)}) \leq f(\mathbf{G}^{(t)}, \mathbf{H}^{(t)}).$$

where the formulation of f is similar as in the proof of Theorem 6.3.1. Indeed, by the definition of Wasserstein distances, we have

$$\begin{aligned} E &= mf(\mathbf{G}^{(t)}, \mathbf{H}^{(t)}) = \\ &\sum_{u=1}^m \sum_{j,v} m T_{j,v}^u \|a_j^{(t)} - X_{u,v}\|^2 + U_{j,v}^u \|a_j^{(t)} - h_{i_{u,v}}^{(t)}\|^2. \end{aligned}$$

Therefore, the update of $a_i^{(t+1)}$ from Algorithm 2 leads to

$$\begin{aligned} E &\geq \sum_{u=1}^m \sum_{j,v} m T_{j,v}^u \|a_j^{(t+1)} - X_{u,v}\|^2 \\ &+ U_{j,v}^u \|a_j^{(t+1)} - h_{i_{u,v}}^{(t)}\|^2 \\ &\geq m \sum_{j=1}^m W_2^2(G_j^{(t)'}, P_n^j) + \sum_{j=1}^m W_2^2(G_j^{(t)'}, H_{i_j}^{(t)}) \\ &\geq m \sum_{j=1}^m W_2^2(G_j^{(t)'}, P_n^j) + \sum_{j=1}^m d_{W_2}^2(G_j^{(t)'}, \mathbf{H}^{(t)}) \\ &= mf(\mathbf{G}'^{(t)}, \mathbf{H}^{(t)}) \end{aligned}$$

where $\mathbf{G}'^{(t)} = (G_1^{(t)'}, \dots, G_m^{(t)'})$, $G_j^{(t)'}$ are formed by replacing the atoms of $G_j^{(t)}$ by the elements of $S_K^{(t+1)}$, noting that $\text{supp}(G_j^{(t)'}) \subseteq \mathcal{S}_K^{(t+1)}$ as $1 \leq j \leq m$, and the second inequality comes directly from the definition of Wasserstein distance. Hence, we obtain

$$f(\mathbf{G}^{(t)}, \mathbf{H}^{(t)}) \geq f(\mathbf{G}'^{(t)}, \mathbf{H}^{(t)}). \quad (6.14)$$

From the formation of $G_j^{(t+1)}$ as $1 \leq j \leq m$, we get

$$\sum_{j=1}^m d_{W_2}^2(G_j^{(t+1)}, \mathbf{H}^{(t)}) \leq \sum_{j=1}^m d_{W_2}^2(G_j^{(t)'}, \mathbf{H}^{(t)}).$$

Thus, it leads to

$$f(\mathbf{G}'^{(t)}, \mathbf{H}^{(t)}) \geq f(\mathbf{G}^{(t+1)}, \mathbf{H}^{(t)}). \quad (6.15)$$

Finally, from the definition of $H_1^{(t+1)}, \dots, H_M^{(t+1)}$, we have

$$f(\mathbf{G}^{(t+1)}, \mathbf{H}^{(t)}) \geq f(\mathbf{G}^{(t+1)}, \mathbf{H}^{(t+1)}). \quad (6.16)$$

By combining (6.14), (6.15), and (6.16), we arrive at the conclusion of the theorem. \square

6.10 Appendix D

In this appendix, we offer details on the data generation processes utilized in the simulation studies presented in Section 6.5 in the main text. The notions of m, n, d, M are given in the main text. Let K_i be the number of supporting atoms of H_i and k_j the number of atoms of G_j . For any $d \geq 1$, we denote $\mathbf{1}_d$ to be d dimensional vector with all components to be 1. Furthermore, \mathcal{I}_d is an identity matrix with d

dimensions.

Comparison metric (Wasserstein distance to truth)

$$W := \frac{1}{m} \sum_{j=1}^m W_2(\hat{G}_j, G_j) + d_{\mathcal{M}}(\hat{\mathbf{H}}, \mathbf{H})$$

where $\hat{\mathbf{H}} := \{\hat{H}_1, \dots, \hat{H}_M\}$, $\mathbf{H} := \{H_1, \dots, H_M\}$ and $d_{\mathcal{M}}(\hat{H}, H)$ is a minimum-matching distance [Tang et al. \[2014\]](#), [Nguyen \[2015\]](#):

$$d_{\mathcal{M}}(\hat{\mathbf{H}}, \mathbf{H}) := \max\{\bar{d}(\hat{\mathbf{H}}, \mathbf{H}), \bar{d}(\mathbf{H}, \hat{\mathbf{H}})\}$$

where

$$\bar{d}(\hat{\mathbf{H}}, \mathbf{H}) := \max_{1 \leq i \leq M} \min_{1 \leq j \leq M} W_2(H_i, \hat{H}_j).$$

Multilevel Wasserstein means setting The global clusters are generated as follows:

means for atoms $\mu_i := 5(i - 1), i = 1, \dots, M$.

atoms of $H_i : \phi_{ij} \sim \mathcal{N}(\mu_i \mathbf{1}_d, \mathcal{I}_d), j = 1, \dots, K_i$.

weights of atoms: $\pi_i \sim \text{Dir}(\mathbf{1}_{K_i})$.

$$\text{Let } H_i := \sum_{j=1}^{K_i} \pi_{ij} \delta_{\phi_{ij}}.$$

For each group $j = 1, \dots, m$, generate local measures and data as follows:

pick cluster label $z_j \sim \text{Unif}(\{1, \dots, M\})$.

mean for atoms : $\tau_{ji} \sim H_{z_j}, i = 1, \dots, k_j$.

atoms of $G_j : \theta_{ji} \sim \mathcal{N}(\tau_{ji}, \mathcal{I}_d), i = 1, \dots, k_j$.

weights of atoms $p_j \sim \text{Dir}(\mathbf{1}_{k_j})$.

Let $G_j := \sum_{i=1}^{k_j} p_{ji} \delta_{\theta_{ji}}$.

data mean $\mu_i \sim G_j, i = 1, \dots, n_j$.

observation $X_{j,i} \sim \mathcal{N}(\mu_i, \mathcal{I}_d)$.

For the case of non-constrained variances, the variance to generate atoms θ_{ji} of G_j is set to be proportional to global cluster label z_j assigned to G_j .

Multilevel Wasserstein means with sharing setting

The global clusters are generated as follows:

means for atoms $\mu_i := 5(i - 1), i = 1, \dots, M$.

atoms of $H_i : \phi_{ij} \sim \mathcal{N}(\mu_i \mathbf{1}_d, \mathcal{I}_d), j = 1, \dots, K_i$.

weights of atoms $\pi_i \sim \text{Dir}(\mathbf{1}_{K_i})$.

Let $H_i := \sum_{j=1}^{K_i} \pi_{ij} \delta_{\phi_{ij}}$.

For each shared atom $k = 1, \dots, K$:

pick cluster label $z_k \sim \text{Unif}(\{1, \dots, M\})$.

mean for atoms : $\tau_k \sim H_{z_k}$.

atoms of $S_K : \theta_k \sim \mathcal{N}(\tau_k, \mathcal{I}_d)$.

For each group $j = 1, \dots, m$ generate local measures and data as follows:

pick cluster label $\tilde{z}_j \sim \text{Unif}(\{1, \dots, M\})$.

select shared atoms $s_j = \{k : z_k = \tilde{z}_j\}$.

weights of atoms $p_{s_j} \sim \text{Dir}(\mathbf{1}_{|s_j|})$; $G_j := \sum_{i \in s_j} p_i \delta_{\theta_i}$.

data mean $\mu_i \sim G_j, i = 1, \dots, n_j$.

observation $X_{j,i} \sim \mathcal{N}(\mu_i, \mathcal{I}_d)$.

For the case of non-constrained variances, the variance to generate atoms θ_i of G_j where $i \in s_j$ is set to be proportional to global cluster label \tilde{z}_j assigned to G_j .

Three-stage K-means First, we estimate G_j for each group $1 \leq j \leq m$ by using K-means algorithm with k_j clusters. Then, we cluster labels using K-means algorithm with M clusters based on the collection of all atoms of G_j s. Finally, we estimate the atoms of each H_i via K-means algorithm with exactly L clusters for each group of local atoms. Here, L is some given threshold being used in Algorithm 1 in Section 6.3.1 in the main text to speed up the computation (see final remark regarding Algorithm 1 in Section 6.3.1). The three-stage K-means algorithm is summarized in Algorithm 3.

Algorithm 3 Three-stage K-means

Input: Data $X_{j,i}$, k_j , M , L .

Output: local measures G_j and global elements H_i of \mathbf{H} .

Stage 1

for $j = 1$ **to** m **do**

$G_j \leftarrow k_j$ clusters of group j with K-means (atoms as centroids and weights as label frequencies).

end for

$\mathcal{C} \leftarrow$ collection of all atoms of G_j .

Stage 2

$\{D_1, \dots, D_M\} \leftarrow M$ clusters from K-means on \mathcal{C} .

Stage 3

for $i = 1$ **to** M **do**

$H_i \leftarrow L$ clusters of D_i with K-means (atoms as centroids and weights as label frequencies).

end for

CHAPTER VII

Conclusions and suggestions

In this thesis, we have investigated several fundamental challenges of mixture and hierarchical models. Our main contributions can be summarized briefly as follows:

- A systematic understanding of statistical efficiency of parameter estimation in finite mixture models.
- Robust estimators of mixing measure in finite mixture models.
- Efficient joint optimization approaches to cluster complex multilevel data.

In the following sections, we will outline several directions that we would like to pursue in the future

7.1 Statistical efficiency, computational complexity, and high dimensionality of mixture and hierarchical models

7.1.1 Statistical efficiency of parameter estimation

In Chapter II, Chapter III, and Chapter IV, the systematic understanding of statistical efficiency regarding parameter estimation is developed thoroughly. It indicates crucial steps toward the development of more efficient model-based inference

procedures. In particular, this raises the following directions at both inference and modeling questions that we intend to pursue in the future

- (1) Methods based on likelihood-based penalization techniques were shown to be quite effective. In many cases, parameter values residing in the vicinity of regions of high singularity levels should be hard to estimate efficiently. Developing a penalization technique generalized to regularize the estimates toward subsets containing singularity points of smaller levels is an interesting problem we hope to address.
- (2) Suitable choices of Bayesian prior have been proposed to induce favorable posterior contraction behavior for overfitted finite mixtures. It is of significant interest to develop an appropriate prior for the mixture model parameters given our knowledge of singular points residing in the parameter space.
- (3) Reparametrization is an effective technique that can be employed to combat singularities present in the class of skewed distributions [Hallin and Ley, 2014]. It would be interesting to study if such reparametrization technique can be systematically developed for the mixture models as well.

7.1.2 Computational complexity of parameter estimation

The improved understanding of statistical efficiency of parameter estimation in finite mixture models carries notable consequences on the computational complexity of parameter estimation procedures, including both optimization and sampling based methods. More specifically, the non-uniform nature of the singularity levels reveals a complex structure of the likelihood function: regions in parameter space that carry low singularity levels may observe a relatively high curvature of the likelihood surface, while high singularity levels imply a “flatter” likelihood surface along a certain subspace of the parameters. Given such interpretation, one of the important direc-

tions is to explore the convergence behaviors of EM algorithm in Gaussian mixture models when both the location and covariance parameter are of consideration. Current studies in the literature demonstrated that when only the location parameter is of interest, a suitable initialization in the neighborhood of global maximizers will guarantee the geometric convergence of EM algorithm to these maximizers [Balakrishnan et al., 2017]. It is our view that the insights from these studies along with our improved understanding of geometric structures of the model’s parameter space will shed light on the performance of EM algorithm under these models.

7.1.3 Efficient models in high dimensional clustering

Apart from the future directions arising from the previous chapters, the general themes of our future research are to move beyond mixture models toward more challenging regimes with several promising applications in practice. In particular, high dimensional data with grouping structures, such as gene microarray data, are omnipresent nowadays. Empirical studies suggested that only a few dimensions in such data are actually influential while the remaining dimensions usually do not contain important information. Motivated by the fact that traditional clustering methods are not effective to capture such phenomena in big data, some models like regularized K-means or mixture models [Pan and Shen, 2007, Sun et al., 2012] have been proposed recently in the literature to address this challenge. Nevertheless, these models used very strong assumptions regarding data structures; fitting them is computationally costly or even infeasible when the dimension and the sample size are rather large. Given the significant impacts of this problem in practice, our principal goals in this research direction are to develop efficient and scalable models such that they perform sufficiently well with various settings of high dimensional data.

7.1.4 Computational complexity of MCMC methods

Sampling techniques based on MCMC have been used extensively in machine learning and statistics applications in the recent years due to the huge advancement in high performance computing. In Bayesian statistics, various MCMC algorithms have been proposed to keep up with increasingly complex structures of hierarchical models. However, current studies demonstrated that certain MCMC algorithms tend to have very slow mixing times, a criterion used to measure the number of iterations needed for the posterior distribution to be within some small distance of the stationary distribution. For instance, the collapsed Gibbs sampling algorithm for the posterior distribution of group labels in Gaussian mixture models was shown to have its mixing times at least of some large power of the sample size [Tosh and Dasgupta, 2014]. Given these computational challenges, our future goals are two-fold: we intend first to explore the computational complexity of contemporary MCMC algorithms in hierarchical models, and secondly to utilize these understandings to develop fast and scalable alternative MCMC algorithms for these models with rapid mixing times.

7.2 Semi-parametric inference of finite mixtures of regression models

Finite mixtures of regression models are utilized when regression data are believed to belong to distinct unobserved categories. One simple instance of such models is when each group shares the same regression relationship but the error distributions among categories are different. Due to their great modeling flexibility, these models have been used extensively in machine learning applications, market segmentation, and social sciences. Nevertheless, most of the previous works with these models in the literature tend to rely on parametric assumptions about dependence of the parameters on covariates, which are usually not realistic. To address these limitations,

an important direction is to explore the semiparametric inference with finite mixture of regression models. This approach had been considered by [Huang and Yao \[2012\]](#) where they made use of kernel regression to obtain parameter estimation; however, their work was only restricted to the univariate setting of covariates. Our current directions are to develop more computationally efficient semiparametric models that can be applied to much broader settings of data. Last but not least, we also intend to extend our current insights of semiparametric inference to more challenging regimes of finite mixture of regression models, such as the high dimensional settings when the number of covariates are much larger than the sample size.

7.3 Statistical applications of optimal transport theory

Given the promising applications of optimal transport to complex multi-level data in Chapter [VI](#), there are two main directions that we would like to pursue in the future

- (1) Firstly, our current work with multi-level data in Chapter [VI](#) has focused mostly on moderate size settings. One worthy yet challenging direction is to scale up our approach to million data points or more. Secondly, we have only considered continuous data; it is of interest to extend our formulation of multilevel Wasserstein means to discrete data. Thirdly, our method requires knowledge of the numbers of clusters both in local and global clustering. When these numbers are unknown, it seems reasonable to incorporate penalty on the model complexity. Fourthly, our formulation does not directly account for the “noise” distribution away from the (Wasserstein) means. To improve the robustness, it may be desirable to make use of the first-order Wasserstein metric instead of the second-order one. Finally, we are interested in extending our approach to richer settings of hierarchical data, such as one when group level-context is available. Another interesting direction is to incorporate the optimal transport

perspective to more complex practical settings of multi-centers data, such as those with center-level contexts.

- (2) Current advances in clustering analysis witness valuable statistical insights about the geometric structures of latent mixing measures arising from hierarchical models based on Wasserstein metric. In particular, the variation of likelihood function in mixture models can be captured effectively by the changes in Wasserstein neighborhood or the borrowing strength phenomenon in hierarchical Dirichlet Process models can be analyzed under optimal transport perspective [Nguyen, 2016]. Motivated by such fruitful connections, there has been a growing interest of extending the understandings of Wasserstein metric to other statistical settings, such as a multi-label classification problem with Wasserstein loss function [Frogner et al., 2015]. Our ultimate goals under this direction concern with exploring the methodological and algorithmic aspects of optimal transport to improve statistical and computational efficiencies of several state of the art models in statistics. We believe that it will be an extraordinarily fertile area with potentially numerous applications in the future.

BIBLIOGRAPHY

BIBLIOGRAPHY

- M. Aguech and G. Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43:904–924, 2011.
- E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37:3099–3132, 2009.
- E. Anderes, S. Borgwardt, and J. Miller. Discrete wasserstein barycenters: optimal transport for discrete data. <http://arxiv.org/abs/1507.07218>, 2015.
- R. B. Arellano-Valle, L. M. Castro, M. C. Genton, and H. W. Gómez. Bayesian inference for shape mixtures of skewed distributions, with application to regression analysis. *Bayesian Analysis*, 3:513–540, 2008.
- R. B. Arellano-Valle, M. C. Genton, and R. H. Loschi. Shape mixtures of multivariate skew-normal distributions. *Journal of Multivariate Analysis*, 100:91–101, 2009.
- A. Azzalini. Further results on a class of distributions which includes the normal ones. *Statistica (Bologna)*, 46:199–208, 1986.
- A. Azzalini and A. Capitanio. Statistical applications of the multivariate skew-normal distribution. *Journal of the Royal Statistical Society, Series B(Methodological)*, 61: 579–602, 1999.
- A. Azzalini and A. D. Valle. The multivariate skew-normal distribution. *Biometrika*, 83:715–726, 1996.
- S. Balakrishnan, M. Wainwright, and B. Yu. Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Annals of Statistics*, 45(1): 77–120, 2017.
- S. Basu, R. Pollack, and M. Roy. *Algorithms in real algebraic geometry*. Springer-Verlag Berlin Heidelberg, 2006.
- M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, 2010.
- J. D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Payré. Iterative bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 2:1111–1138, 2015.

- R. Beran. Minimum hellinger distance estimates for parametric models. *Annals of Statistics*, 5:445–453, 1977.
- P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press, 1993.
- D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
- L. Bordes, S. Mottelet, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model. *Annals of Statistics*, 34:1204–1232, 2006.
- B. Buchberger. *An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal*. PhD thesis, Johannes Kepler University of Linz, 1965.
- P. Bühlmann and S. van de Geer. *Statistics for high dimensional data*. Springer, 2011.
- C. Caillerie, F. Chazal, J. Dedecker, and B. Michel. Deconvolution for the Wasserstein metric and geometric inference. *Electronic Journal of Statistics*, 5:1394–1423, 2011.
- A. Canale and B. Scarpa. Bayesian nonparametric location-scale-shape mixtures. *TEST*, pages 1–18, 2015.
- R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of American Statistical Association*, 83:1184–1186, 1988.
- H. Chen and J. Chen. Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica*, 13:351–365, 2003.
- J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995.
- J. Chen. Consistency of the mle under mixture models. *arXiv preprint arXiv:1607.01251*, 2016.
- J. Chen and P. Li. Hypothesis test for normal mixture models: the em approach. *Annals of Statistics*, 37:2523–2542, 2009.
- J. Chen and X. Tan. Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, 100:1367–1383, 2009.
- J. Chen, X. Tan, and R. Zhang. Inference for normal mixtures in mean and variance. *Statistica Sinica*, 18:443–465, 2008.
- J. Chen, P. Li, and Y. Fu. Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107:1096–1105, 2012.

- M. Chiogna. A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution. *Statistical Methods and Applications*, 14:331–341, 2005.
- D. Cox, J. Little, and D. O’Shea. *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer, 2007.
- A. Cutler and O. I. Cordero-Brana. Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91:1716–1723, 1996.
- M. Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems 26*, 2013.
- M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. *Proceedings of the 31st International Conference on Machine Learning*, 2014.
- D. Dacunha-Castelle and E. Gassiat. The estimation of the order of a mixture model. *Bernoulli*, 3:279–299, 1997.
- D. Dacunha-Castelle and E. Gassiat. Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Annals of Statistics*, 27:1178–1209, 1999.
- A. DasGupta. *Asymptotic Theory of Statistics and Probability*. Springer, 2008.
- S. Dasgupta. Learning mixtures of Gaussians. Technical Report UCB/CSD-99-1047, University of California, Berkeley, 1999.
- N. E. Day. Estimating the components of a mixture of normal distributions. *Biometrika*, 56:463–474, 1969.
- L. Devroye and Laszlo Gyorfi. *Nonparametric density estimation: the L1 view*. John Wiley and Sons, 1985.
- D. Donoho and R. C. Liu. The automatic robustness of minimum distance functionals. *Annals of Statistics*, 16:552–586, 1988.
- M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on Algebraic Statistics*. Birkhauser, 2009.
- C. R. K. Dudley, L. A. Giuffra, A. E. G. Raine, and S. T. Reeders. Assessing the role of apnh, a gene encoding for a human amiloride- sensitive Na^+/H^+ antiporter, on the interindividual variation in red cell Na^+/Li^+ countertransport. *Journal of the American Society of Nephrology*, 2:937–943, 1991.
- R. Elmore, P. Hall, and A. Neeman. An application of classical invariant theory to identifiability in nonparametric mixtures. *Ann. Inst. Fourier (Grenoble)*, 55:1–28, 2005.

- M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.
- J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19(3):1257–1272, 1991.
- C. Frogner, C. Zhang, H. Mabahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. *Advances in Neural Information Processing Systems (NIPS) 28*, 2015.
- C. Genovese and L. Wasserman. Rates of convergence for the gaussian mixture sieve. *Annals of Statistics*, 28:1105–1127, 2000.
- S. Ghosal and A. Roy. Predicting false discovery proportion under dependence. *Journal of the American Statistical Association*, 106:1208–1217, 2011.
- S. Ghosal and A. van der Vaart. Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29:1233–1263, 2001.
- S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Springer-Verlag, New York, 2000.
- P. Hall and X. H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 31:201–224, 2003.
- P. Hall, A. Neeman, R. Pakyari, and R. Elmore. Nonparametric inference in multivariate mixtures. *Biometrika*, 92:667–678, 2005.
- M. Hallin and C. Ley. Skew-symmetric distributions and fisher information - a tale of two densities. *Bernoulli*, 18:747–763, 2012.
- M. Hallin and C. Ley. Skew-symmetric distributions and fisher information: the double sin of skew-normal. *Bernoulli*, 20:1432–1453, 2014.
- T. Hastie, R. Tibshirani, and M. J. Wainwright. *Statistical Learning with Sparsity: The Lasso and generalizations*. CRC Press, 2015.
- R. J. Hathaway. A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Annals of Statistics*, 13:795–800, 1985.
- P. Heinrich and J. Kahn. Optimal rates for finite mixture estimation. *Under review*, 2016+.
- N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016a.
- N. Ho and X. Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *arXiv:1609.02655. Under review, Annals of Statistics*, 2016b.

- N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10:271–307, 2016c.
- N. Ho and X. Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. Technical Report 540, Department of Statistics, University of Michigan, 2016d.
- N. Ho, X. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. Multilevel clustering via Wasserstein means. *To appear, Proceedings of 34th International Conference on Machine Learning*, 2017.
- Y. S. Hsu, M. D. Fraser, and J. J. Walker. Identifiability of finite mixtures of von mises distributions. *Annals of Statistics*, 9:1130–1131, 1981.
- M. Huang and W. Yao. Mixture of regression models with varying mixing proportions: A semiparametric approach. *Journal of the American Statistical Association*, 107: 711–724, 2012.
- D. R. Hunter, S. Wang, and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. *Annals of Statistics*, 35:224–251, 2007.
- V. Huynh, D. Phung, S. Venkatesh, X. Nguyen, M. Hoffman, and H. Bui. Scalable nonparametric bayesian multilevel clustering. *Proceedings of Uncertainty in Artificial Intelligence*, 2016.
- H. Ishwaran, L. F. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96:1316–1332, 2001.
- L. F. James, C. E. Priebe, and D. J. Marchette. Consistent estimation of mixture complexity. *Annals of Statistics*, 29:1281–1296, 2001.
- J. Chen and A. Khalili. Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 103:1674–1683, 2012.
- A. Kalai, A. Moitra, and G. Valiant. Disentangling gaussians. *Communications of the ACM*, 55(2):113–120, 2012.
- R.J. Karunamuni and J. Wu. Minimum hellinger distance estimation in a nonparametric mixture model. *Journal of Statistical Planning and Inference*, 139:1118–1133, 2009.
- H. Kasahara and K. Shimotsu. Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 76:97–111, 2014a.
- H. Kasahara and K. Shimotsu. Testing the number of components in normal mixture regression models. *Journal of the American Statistical Association*, 110:1632–1645, 2014b.

- J. T. Kent. Identifiability of finite mixtures for directional data. *Annals of Statistics*, 11:984–988, 1983.
- C. Keribin. Consistent estimation of the order of mixture models. *Sankhya Series A*, 62:49–66, 2000.
- W. Kruijer, J. Rousseau, and A. van der Vaart. Adaptive Bayesian density estimation with location-scale mixtures. *Electronic Journal of Statistics*, 4:1225–1257, 2010.
- B. Kulis and M. I. Jordan. Revisiting k-means: new algorithm via bayesian nonparametrics. *Proceedings of the 29th International Conference on Machine Learning*, 2012.
- L. F. Lee and A. Chesher. Specification testing when score test statistics are identically zero. *Journal of Econometrics*, 31:33–61, 1986.
- S. X. Lee and G. J. McLachlan. On mixtures of skew normal and skew t -distributions. *Advances in Data Analysis and Classification*, 7:241–266, 2013.
- E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, 1998.
- B. G. Leroux. Consistent estimation of a mixing distribution. *Annals of Statistics*, 20:1350–1360, 1992.
- C. Ley and D. Paindaveine. On the singularity of multivariate skew-symmetric models. *Journal of Multivariate Analysis*, 101:1434–1444, 2010.
- N. Lin and X. He. Robust and efficient estimation under data grouping. *Biometrika*, 93:99–112, 2006.
- T. I. Lin. Maximum likelihood estimation for multivariate skew normal mixture models. *Journal of Multivariate Analysis*, 100:257–265, 2009.
- T. I. Lin, J. C. Lee, and S. Y. Yen. Finite mixture modelling using the skew normal distribution. *Statistica Sinica*, 17:909–927, 2007.
- B. Lindsay. *Mixture models: Theory, geometry and applications*. In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA., 1995.
- B. G. Lindsay. Efficiency versus robustness: The case for minimum hellinger distance and related methods. *Annals of Statistics*, 22:1081–1114, 1994.
- X. Liu and Y. Shao. Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, 31:807–832, 2004.
- K. V. Mardia. Statistics of directional data. *Journal of the Royal Statistical Society. Series B(Methodological)*, 37:349–393, 1975.
- G. J. McLachlan and K. E. Basford. *Mixture models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs*. New York, 1988.

- J. Miller and D. Dunson. Robust bayesian inference via coarsening. *arXiv:1506.06101*, 2015.
- Thanh-Binh Nguyen, Vu Nguyen, Svetha Venkatesh, and Dinh Phung. Mcnc: Multi-channel nonparametric clustering from heterogeneous data. In *Proceedings of ICPR*, 2016.
- V. Nguyen, D. Phung, X. Nguyen, S. Venkatesh, and H. Bui. Bayesian nonparametric multilevel clustering with group-level contexts. *Proceedings of the 31st International Conference on Machine Learning*, 2014.
- X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 41(1):370–400, 2013.
- X. Nguyen. Posterior contraction of the population polytope in finite admixture models. *Bernoulli*, 21:618–646, 2015.
- X. Nguyen. Borrowing strength in hierarchical bayes: Posterior concentration of the dirichlet base measure. *Bernoulli*, 22:1535–1571, 2016.
- A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.
- W. Pan and X. Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8:1145–1164, 2007.
- K. Pearson. Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110, 1894.
- D. Peel and G. J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.
- D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, 28:199–205, 1982.
- M. O. Prates, C. R. B. Cabral, and V. H. Lachos. mixsmsn: fitting finite mixture of scale mixture of skew-normal distributions. *Journal of Statistical Software*, 54, 2013.
- J. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.
- S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Methodological)*, 59:731–792, 1997.
- A. Rodriguez, D. Dunson, and A.E. Gelfand. The nested Dirichlet process. *J. Amer. Statist. Assoc.*, 103(483):1131–1154, 2008.

- K. Roeder. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89:487–495, 1994.
- A. Rotnitzky, D. R. Cox, M. Bottai, and J. Robins. Likelihood-based inference with singular information matrix. *Bernoulli*, 6:243–284, 2000.
- J. Rousseau. Rates of convergence for the posterior distributions of mixtures of Betas and adaptive nonparametric estimation of densities. *Annals of Statistics*, 38(1):146–180, 2010.
- J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B*, 73(5):689–710, 2011.
- S. W. Schnatter and S. Pyne. Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions. *Biostatistics*, 11:317–336, 2009.
- J. Solomon, G. Fernando, G. Payré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. In *The International Conference and Exhibition on Computer Graphics and Interactive Techniques*, 2015.
- S. Srivastava, V. Cevher, Q. Dinh, and D. Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, 2015.
- B. Sturmfels. *Solving system of polynomial equations*. Providence R.I, 2002.
- W. Sun, J. Wang, and Y. Fang. Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics*, 6:148–167, 2012.
- Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of The 31st International Conference on Machine Learning*, pages 190–198. ACM, 2014.
- Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.*, 101:1566–1581, 2006.
- H. Teicher. Identifiability of mixtures. *Annals of Statistics*, 32:244–248, 1961.
- H. Teicher. Identifiability of finite mixtures. *Annals of Statistics*, 34:1265–1269, 1963.
- C. Tosh and S. Dasgupta. Lower bounds for the Gibbs sampler over mixtures of Gaussians. *Proceedings of International Conference on Machine Learning (ICML)*, 2014.

- W. Toussile and E. Gassiat. Variable selection in model-based clustering using multilocus genotype data. *Advances in Data Analysis and Classification*, 3:109–134, 2009.
- S. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.
- A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 1998.
- C. Villani. *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer, Berlin, 2009.
- Cédric Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.
- M. Wiper, D. R. Insua, and F. Ruggeri. Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics*, 10:440–454, 2001.
- M. Woo and T. N. Sriram. Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101:1475–1486, 2006.
- W. A. Woodward, W. C. Parr, W. R. Schucany, and H. Lindsey. A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *Journal of the American Statistical Association*, 79:590–598, 1984.
- D. F. Wulsin, S. T. Jensen, and B. Litt. Nonparametric multi-level clustering of human epilepsy seizures. *Annals of Applied Statistics*, 10:667–689, 2016.
- S. Xiao and G. Zeng. Determination of the limits for multivariate rational functions. *Science China Mathematics*, 57:397–416, 2014.
- S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *Annals of Statistics*, 39(1):209–214, 1968.
- B. Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1997.
- C. B. Zeller, C. R. B. Cabral, and V. H. Lachos. Robust mixture regression modeling based on scale mixtures of skew-normal distributions. *TEST*, pages 1–22, 2015.
- C. Zhang. Fourier methods for estimating mixing densities and distributions. *Annals of Statistics*, 18(2):806–831, 1990.
- T. Zhang, A. Weisel, and M. S. Greco. Multivariate generalized gaussian distribution: Convexity and graphical models. *IEEE Transactions on Signal Processing*, 61:4141–4148, 2013.
- P. C. Álvarez Estebana, E. del Barrio, J.A. Cuesta-Albertos, and C. Matrán. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441:744–762, 2016.