Electronic Journal of Statistics
Vol. 10 (2016) 271–307
ISSN: 1935-7524
DOI: 10.1214/16-EJS1105

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

Nhat Ho and XuanLong Nguyen[*]

Department of Statistics
University of Michigan
e-mail: minhnhat@umich.edu; xuanlong@umich.edu

Abstract: This paper 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.

MSC 2010 subject classifications: Primary 62F15, 62G05; secondary 62G20.
Keywords and phrases: Mixture models, strong identifiability, Wasserstein distances, minimax bounds, maximum likelihood estimation.

Received February 2015.

1. Introduction

Mixture models are a popular modeling tool for making inference about heterogeneous data [15, 18]. 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. Theoretical issues related to parameter estimation in mixture models, however, remain poorly understood — as noted in a recent textbook [5] (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 [22, 23, 26] and others, and continues to attract much inter-

[*] The authors would like to thank Elisabeth Gassiat, Xuming He, Judith Rousseau, Naisying Wang and Shuheng Zhou for valuable discussions related to this work. This research is supported in part by grants NSF CCF-1115769, NSF CAREER DMS-1351362, and NSF CNS-1409303.

est [11, 10, 7, 1]. 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 [3, 28, 8], 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 [4]. 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 [19]. 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 [21]. 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., [6, 2, 13]).

Setting The goal of this paper 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 $\big\{f(x|\theta,\Sigma), \theta \in \Theta \subset \mathbb{R}^{d_1}, \Sigma \in \Omega \subset S_{d_2}^{++}, x \in \mathcal{X}\big\}$. 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 paper 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 [19], the convergence of mixture model parameters can be measured in terms of a Wassertein 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. [25])

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

where the infimum is taken over all joint probability distributions $\vec{q}$ on $[1, \ldots, k] \times [1, \ldots, k_0]$ such that, when expressing $\vec{q}$ 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 $G_n \to G_0$ under W_r metric at a rate $\omega_n = o(1)$. If all G_n have the same number of atoms $k = k_0$ 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 [23]. 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 [4] and [19] 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 [4, 16, 21, 19]. We show that for exact-fitted mixtures, the first-order identifiability condition (along with additional and mild regularity conditions) suffices for obtaining that

$$V(p_G, p_{G_0}) \gtrsim W_1(G, G_0), \tag{1}$$

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,

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

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 [21]), 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 3.2.

Convergence of MLE estimators and minimax lower bounds Assuming that n-iid sample $X_1, \ldots, X_n$ are generated according to p_{G_0} and let $\widehat{G}_n$ 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 (1) and (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 4.2, Theorem 4.3, and Theorem 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 4.2 and Theorem 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 paper 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 described 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 paper. Such a theory is fundamentally different from the strong identifiability theory, and will be reported elsewhere.

Paper organization The rest of the paper is organized as follows. Section 2 provides some preliminary backgrounds and facts. Section 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 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 4.2. Self-contained proofs of the key theorems are given in Section 5 while proofs of the remaining results are presented in the Appendices.

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 $g : \mathbb{R}^{K \times L} \to \mathbb{R}$ 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 $g_2 : \mathbb{R}^N \times \mathbb{R}^{K \times L} \to \mathbb{R}$ 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 the paper, 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 "$\gtrsim$" 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. Preliminaries

First of all, we need to define our notion of distances on the space of mixing measures. In this paper, 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 $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ denotes the proportion vector. Likewise, let $G' = \sum_{i=1}^{k'} p'_i \delta_{(\theta'_i, \Sigma'_i)}$. A coupling between $\vec{p}$ and $\vec{p'}$ is a joint distribution $\vec{q}$ on $[1 \ldots, k] \times [1, \ldots, k']$, which is expressed as a matrix $\vec{q} = (q_{ij})_{1 \leq i \leq k, 1\ \leq j \leq k'} \in [0,1]^{k \times k'}$ and admits marginal constraints $\sum_{i=1}^{k} q_{ij} = p'_j$ and $\sum_{j=1}^{k'} q_{ij} = p_i$ for any $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k'$. We call $\vec{q}$ a coupling of $\vec{p}$ and $\vec{p'}$, and use $\mathcal{Q}(\vec{p}, \vec{p'})$ to denote the space of all such couplings.

As in [19], 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 [25]. For any $r \geq 1$, the r-th order Wasserstein distance between G and G' is given by

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

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 paper 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 [19] (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 2.1 (Multivariate generalized Gaussian distribution [29])**.**
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 $\mathbb{R}^d$, $\Theta_2=\left\{m\in\mathbb{R}^+ : 1\leq \underline{m}\leq m\leq\overline{m}\right\}$, and $\Omega=\left\{\Sigma\in S_d^{++}:\underline{\lambda}\leq\sqrt{\lambda_1(\Sigma)}\leq\sqrt{\lambda_d(\Sigma)}\leq\overline{\lambda}\right\}$, where $\underline{\lambda},\overline{\lambda}>0$, 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 (Multivariate Student's t-distribution [20])**.**
Let $f(x|\theta,\Sigma)=\frac{C_\nu}{|\Sigma|^{1/2}}\left(\nu+(x-\theta)^T\Sigma^{-1}(x-\theta)\right)^{-(\nu+d)/2}$, 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 $\Omega=\left\{\Sigma\in S_d^{++}:\underline{\lambda}\leq\sqrt{\lambda_1(\Sigma)}\leq\sqrt{\lambda_d(\Sigma)}\leq\overline{\lambda}\right\}$, 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 2.3 (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 $\lambda=(\lambda_1,\ldots,\lambda_d)$. If Θ is a bounded subset of $\mathbb{R}^d\times\mathbb{R}_+^d$ and $\Omega=\left\{\Sigma\in S_d^{++}:\underline{\lambda}\leq\sqrt{\lambda_1(\Sigma)}\leq\sqrt{\lambda_d(\Sigma)}\leq\overline{\lambda}\right\}$, 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 2.4 (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 $\alpha=(\alpha_1,\ldots,\alpha_d)$ and combined rate vector $\beta=(\beta_1,...,\beta_d)$. If Θ is a bounded subset of $\mathbb{R}^d\times\mathbb{R}_+^d\times\mathbb{R}_+^d$, $\Omega=\left\{\Sigma\in S_d^{++}:\underline{\lambda}\leq\sqrt{\lambda_1(\Sigma)}\leq\sqrt{\lambda_d(\Sigma)}\leq\overline{\lambda}\right\}$, then for any mixing measures G_1,G_2, we have $h^2(p_{G_1},p_{G_2})\lesssim V(p_{G_1},p_{G_2})\lesssim W_1(G_1,G_2)$.

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.

3.1. Definitions and general identifiability bounds

Definition 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 $(\theta_1,\Sigma_1),...,(\theta_k,\Sigma_k) \in \Theta \times \Omega$, 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 $i = 1,\ldots,k$) such that for almost all x*

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

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

Remark The condition that γ_i is symmetric in Definition 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, $\alpha_i = 0$, $\beta_i = \vec{0}$, then the equation in (A1) holds even when $\gamma_i \neq \vec{0}$ 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 $x \in \mathcal{X}$

$$\left|\gamma_1^T \left(\frac{\partial f}{\partial \theta}(x|\theta_1,\Sigma) - \frac{\partial f}{\partial \theta}(x|\theta_2,\Sigma)\right)\right| \leq C(R_1,R_2)\|\theta_1 - \theta_2\|^{\delta_1}\|\gamma_1\|, \tag{3}$$

$$\left|\operatorname{tr}\left(\left(\frac{\partial f}{\partial \Sigma}(x|\theta,\Sigma_1) - \frac{\partial f}{\partial \Sigma}(x|\theta,\Sigma_2)\right)^T \gamma_2\right)\right| \leq C(R_3)\|\Sigma_1 - \Sigma_2\|^{\delta_2}\|\gamma_2\|. \tag{4}$$

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 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 Wassertein 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 3.1. *Suppose that all assumptions of Theorem 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*

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

We shall verify in the sequel that the classes of densities f described in Examples 2.1, 2.2, and 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 [4]:

Definition 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 $(\theta_1, \Sigma_1), ..., (\theta_k, \Sigma_k) \in \Theta \times \Omega$, if we have $\alpha_i \in \mathbb{R}, \beta_i, \nu_i \in \mathbb{R}^{d_1}$, γ_i, η_i ***symmetric matrices*** *in $\mathbb{R}^{d_2 \times d_2}$ as $i = 1, \ldots, k$ such that for almost all x*

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

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

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 $x \in \mathcal{X}$

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

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

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 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 $\lim_{\lambda_1(\Sigma) \to 0} f(x|\theta, \Sigma) = 0$. 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 $1 \leq r < 2$:

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

(c) (Optimality of bound for Hellinger distance) Assume f is second-order differentiable with respect to θ, Σ and for some sufficiently small $c_0 > 0$,

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

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 $1 \leq r < 2$:

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

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 [19] (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. [1] 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 $\lim_{\lambda_1(\Sigma)\to 0} f(x|\theta, \Sigma) = 0$ 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 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 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 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*

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

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:

[1] A 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 [4], which admits the same correction described above.

Corollary 3.2. *Suppose that all assumptions of Theorem 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 [21]), in addition to having that the component parameters converge at the $n^{-1/4}$ rate.

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 identifiablity 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 3.3. *(Densities with multiple scalar parameters)*

(a) Generalized univariate logistic densities: Let $f(x|\theta,\sigma) := \frac{1}{\sigma} f\left((x-\theta)/\sigma\right)$, 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 [12, 14, 17]: 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:

Theorem 3.4. *(Densities with matrix-variate parameters)*

(a) The family $\left\{f(x|\theta,\Sigma,m), \theta \in \mathbb{R}^d, \Sigma \in S_d^{++}, m \geq 1\right\}$ of multivariate generalized Gaussian distribution is identifiable in the first order.

(b) The family $\left\{f(x|\theta,\Sigma), \theta \in \mathbb{R}^d, \Sigma \in S_d^{++}\right\}$ of multivariate t-distribution with fixed odd degree of freedom is identifiable in the second order.

(c) *The family* $\left\{f(x|\theta,\Sigma,\lambda),\theta\in\mathbb{R}^d,\Sigma\in S_d^{++},\lambda\in\mathbb{R}_+^d\right\}$ *of exponentially modified multivariate t-distribution with fixed odd degree of freedom is identifiable in the second order.*

(d) *The family* $\left\{f(x|\theta,\Sigma,a,b),\theta\in\mathbb{R}^d,\Sigma\in S_d^{++},a\in\mathbb{R}_+^d,b\in\mathbb{R}_+^d\right\}$ *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 [4]. As the proofs of these results are technically involved, we present only the proof of Theorem 3.4 in the Appendix. 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 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 the family of Gamma distributions is not identifiable in the first order.

The results of Theorem 3.4 shed light on which classes of distribution satisfy the inequality being established in Theorem 3.1 and Theorem 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 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 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 identifiablity (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 3.3. *Assume that the conditions of Theorem 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 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.

4. Minimax lower bounds, MLE rates and illustrations

4.1. Minimax lower bounds and MLE rates of convergence

Given n-iid sample $X_1, X_2, ..., X_n$ 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 $\Omega = \Big\{\Sigma\in S_{d_2}^{++} : \underline{\lambda}\leq\sqrt{\lambda_1(\Sigma)}\leq\sqrt{\lambda_{d_2}(\Sigma)}\leq\overline{\lambda}\Big\}$, where $0<\underline{\lambda},\overline{\lambda}$ 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

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

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

The inequalities established by Theorem 3.1 and Theorem 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 convergence 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., [24]). For any positive integer number k_1, define several classes of mixture densities $\mathcal{P}_{k_1}(\Theta^*)=\{p_G : G\in\mathcal{O}_{k_1}(\Theta^*)\}$, $\overline{\mathcal{P}}_{k_1}(\Theta^*)=\Big\{p_{\frac{G+G_0}{2}} : G\in\mathcal{O}_{k_1}(\Theta^*)\Big\}$, and $\overline{\mathcal{P}}_{k_1}^{1/2}(\Theta^*)=\Big\{\Big(p_{\frac{G+G_0}{2}}\Big)^{1/2} : G\in\mathcal{O}_{k_1}(\Theta^*)\Big\}$. For any $\delta>0$, define the intersection with a Hellinger ball centered at p_{G_0} via

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

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

$$\mathcal{J}_B(\delta,\overline{\mathcal{P}}_{k_1}^{1/2}(\Theta^*,\delta),\mu)=\int\limits_{\delta^2/2^{13}}^{\delta}H_B^{1/2}(u,\overline{\mathcal{P}}_{k_1}^{1/2}(\Theta^*,u),\mu)\mathrm{d}u\vee\delta,$$

where H_B denotes the bracketing entropy of $\overline{\mathcal{P}}_{k_1}^{1/2}(\Theta^*)$ under L_2 distance (cf. [24] 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 [24] with the adaption of notations as those in our paper

Theorem 4.1. *Take $\Psi(\delta)\geq J_B(\delta,\overline{\mathcal{P}}_k^{1/2}(\Theta^*,\delta),\mu)$ 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 $\delta \geq \delta_n$

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

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

Theorem 4.2. (Exact-fitted mixtures) *Assume that* f *satisfies the conditions of Theorem 3.1. Take* $\Psi(\delta) \geq \mathcal{J}_B(\delta, \overline{\mathcal{P}}_{k_0}^{1/2}(\Theta^*, \delta), \mu)$ *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*

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

Proof. By Theorem 3.1,

$$C_1(\Theta^*)W_1^2(G, G_0) \leq V^2(p_G, p_{G_0}) \leq 2h^2(p_G, p_{G_0}) \text{ for all } G \in \mathcal{E}_{k_0}(\Theta^*), \tag{5}$$

where $C_1(\Theta^*)$ is a positive constant depending only on Θ^* and G_0. Now, invoking Theorem 4.1, as $\delta \geq \delta_n$, there is a universal constant $c > 0$ such that

$$P(h(p_{\widehat{G}_n}, p_{G_0}) > \delta) \leq c \exp\left(-\frac{n\delta^2}{c^2}\right). \tag{6}$$

Combining (5) and (6), we readily achieve the conclusion of our theorem. □

Using the same argument we arrive at a general convergence rate result of the MLE under the over-fitted setting:

Theorem 4.3. (Over-fitted mixtures) *Assume that* f *satisfies the conditions in part (a) of Theorem 3.2. Take* $\Psi(\delta) \geq \mathcal{J}_B(\delta, \overline{\mathcal{P}}_{k}^{1/2}(\Theta^*, \delta), \mu)$ *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*

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

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 4.1. (Mixtures of multivariate generalized Gaussian distributions) *Given $\Theta = [-a_n, a_n]^d \times [\underline{m}, \overline{m}]$ where $a_n \leq L(\log(n))^{\gamma}$ as L is some positive constant, $\gamma > 0$, and $1 < \underline{m} \leq \overline{m}$ 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 $P(W_1(\widehat{G}_n, G_0) > \delta_n) \lesssim \exp(-c\log(n))$, where δ_n is a sufficiently large multiple of $(\log(n)/n)^{1/2}$ and c is positive constant depending only on $L, \gamma, \underline{m}, \overline{m}, \underline{\lambda}, \overline{\lambda}$.

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

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 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 4.4. (Minimax lower bound for over-fitted mixtures) *If the class of densities f satisfies condition (c) of Theorem 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 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 [27], for any sequence of estimates $\widehat{G}_n$ ranging in $\mathcal{O}_k$, we obtain that

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

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

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

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. □

4.2. Illustrations

For the remainder of this section, we shall illustrate via simulations the strong identifiability bounds established in Section 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 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 1 and Figure 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.

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 locations $\theta_1^0 = (-2, 2)$, $\theta_2^0 = (-4, 4)$, covariances

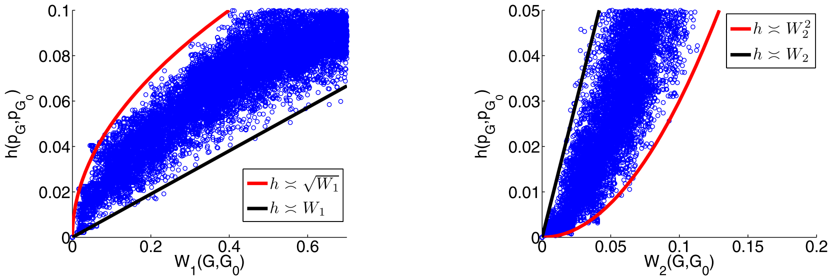

FIG 1. *Mixture of Student's t-distributions. Left: Exact-fitted setting. Right: Over-fitted setting.*

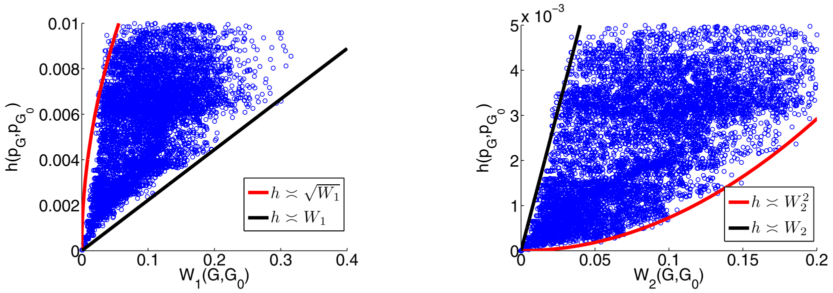

FIG 2. *Mixture of multivariate generalized Gaussian distributions. Left: Exact-fitted setting. Right: Over-fitted setting.*

$\Sigma_1^0 = \begin{pmatrix} 9/4 & 1/5 \\ 1/5 & 13/6 \end{pmatrix}$, $\Sigma_2^0 = \begin{pmatrix} 5/2 & 2/5 \\ 2/5 & 7/3 \end{pmatrix}$, $m_1^0 = 2$, $m_2^0 = 3$ (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), 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.

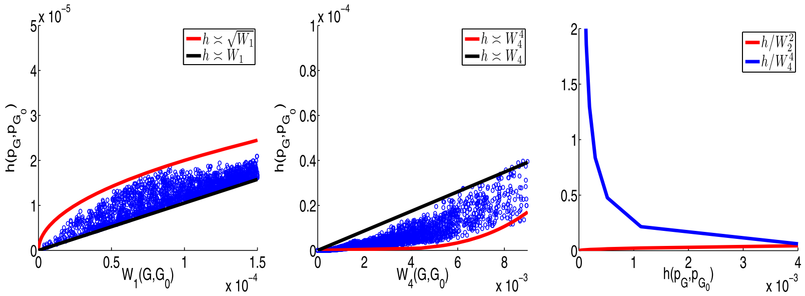

Fig 3. *Mixture of location-scale Gaussian distributions, which satisfy first-order identifiablity but not second-order identifiability condition. Left panel: Exact-fitted setting. Middle and right panels are for over-fitted setting by one extra component. Right panel shows that $h \gtrsim W_2^2$ no longer holds as $h \to 0$.*

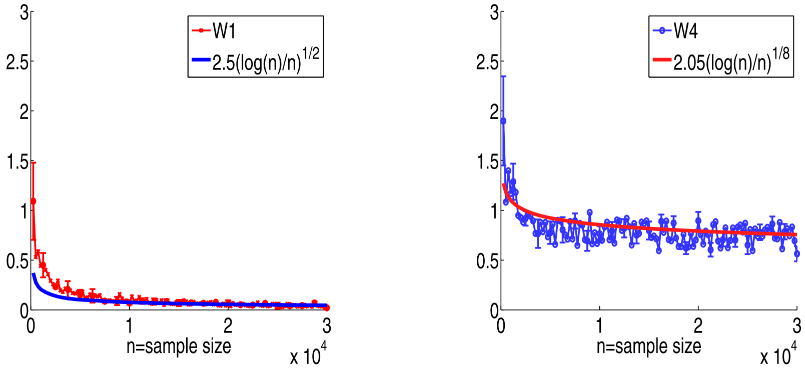

Fig 4. *MLE rates for location-covariance mixtures of Gaussians. Left: Exact-fitted — $W_1 \asymp n^{-1/2}$. Right: Over-fitted by one — $W_4 \asymp n^{-1/8}$.*

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: $\theta_1^0 = (0,3), \theta_2^0 = (1,-4), \theta_3^0 = (5,2)$, $\Sigma_1^0 = \left(\begin{smallmatrix} 4.2824 & 1.7324 \\ 1.7324 & 0.81759 \end{smallmatrix}\right)$, $\Sigma_2^0 = \left(\begin{smallmatrix} 1.75 & -1.25 \\ -1.25 & 1.75 \end{smallmatrix}\right)$, $\Sigma_3^0 = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 4 \end{smallmatrix}\right)$, and $p_1^0 = 0.3$, $p_2^0 = 0.4$, $p_3^0 = 0.3$. The parameter spaces Θ, Ω are identical 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 4 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.

5. Proofs of key theorems

In this section, we present self-contained proofs for two key theorems: Theorem 3.1 for strongly identifiable mixtures in the exact-fitted setting and Theorem 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 sequence of steps to help the reader. The proofs of the remaining results are deferred to the Appendices.

5.1. Strong identifiability in exact-fitted mixtures

Proof of Theorem 3.1 It suffices to show that

$$\lim_{\epsilon \to 0} \inf \left\{ V(p_G, p_{G_0})/W_1(G,G_0) | W_1(G,G_0) \leq \epsilon \right\} > 0, \tag{7}$$

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

Step 1 Suppose that (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 $V(p_{G_n}, p_{G_0})/W_1(G_n, G_0) \to 0$ as $n \to \infty$. As $W_1(G_n, G_0) \to 0$, 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 $i = 1, \ldots, k_0$, $(\theta_i^n, \Sigma_i^n) \to (\theta_i^0, \Sigma_i^0)$. 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 $(\theta_i^n, \Sigma_i^n) \to (\theta_i^0, \Sigma_i^0)$ 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 $V(p_{G_n}, p_{G_0})/d(G_n, G_0) \to 0$.

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

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

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 $\sup_x |R_n(x)/d(G_n, G_0)| \to 0$ as $n \to \infty$.

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 $n \to \infty$. Indeed, if the opposite is true, then the summation of the absolute values of these coefficients must also tend to 0:

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

Since we have the entrywise ℓ_1 and ℓ_2 norms are equivalent, the above entails $\left\{\sum_{i=1}^{k_0} |\Delta p_i^n| + p_i^n (\|\Delta\theta_i^n\| + \|\Delta\Sigma_i^n\|)\right\}/d(G_n, G_0) \to 0$, 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 $n \to \infty$.

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 $n \to \infty$,

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

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

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

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

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

By the first-order identifiability criteria of f, we have $\alpha_i = 0, \beta_i = \vec{0} \in \mathbb{R}^{d_1}$, and $\gamma_i = \vec{0} \in \mathbb{R}^{d_2 \times d_2}$ for all $i = 1, 2, ..., k$, which is a contradiction. Hence, (7) is proved.

5.2. Strong identifiability in over-fitted mixtures

Proof of Theorem 3.2 (a) We only need to establish that

$$\lim_{\epsilon \to 0} \inf_{G \in \mathcal{O}_k(\Theta)} \left\{ \sup_{x \in \mathcal{X}} |p_G(x) - p_{G_0}(x)|/W_2^2(G, G_0) : W_2(G, G_0) \le \epsilon \right\} > 0. \qquad (8)$$

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 3.1.

Step 1 Suppose that (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 $\sup_{x \in \mathcal{X}} |p_{G_n}(x) - p_{G_0}(x)|/W_2^2(G_n, G_0) \to 0$ as $n \to \infty$. 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 3.1 and the fact that $W_2^2(G_n, G_0) \lesssim W_1(G_n, G_0)$ for all n. Thus, $k_0 + 1 \le k^* \le k$.

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 $W_2(G_n, G_0) \to 0$, 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 $s_i \ge 1$ 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 $m \ge 0$ 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 $(\theta_{ij}^n, \Sigma_{ij}^n) \to (\theta_i^0, \Sigma_i^0)$ for each $i = 1, \ldots, k_0 + m$, $j = 1, \ldots, s_i$, $p_i^0 = 0$ for $i < k_0$, and we have that $p_{i.}^n := \sum_{j=1}^{s_i} p_{ij}^n \to p_i^0$ 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 $n \to \infty$, which implies $\det(\Sigma_{ij}^n) \to 0$ and hence $\det(\Sigma_i^0) = 0$. By re-labeling the support points, we may assume without loss of generality that $\Sigma_{k_0+1}^0, \ldots, \Sigma_{k_0+m_1}^0$ are (strictly) positive definite matrices and $\Sigma_{k_0+m_1+1}^0, \ldots, \Sigma_{k_0+m}^0$ 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 $\lim_{\lambda_1(\Sigma) \to 0} f(x|\theta, \Sigma) = 0$ 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 $\mathcal{X}$, $f(x|\theta_{ij}^n, \Sigma_{ij}^n) \to 0$ as $n \to \infty$ 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 $i = 1, \ldots, k_0 + m_1$ and $j = 1, \ldots, s_i$, it is simple to see that

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

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 $j = 1, \ldots, s_i$ with (θ_i^0, Σ_i^0) by mass $\min(p_{i\cdot}^n, p_i^0)$, 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 $\sup_{x \in \mathcal{X}} |p_{G_n}(x) - p_{G_0}(x)|/d(G_n, G_0) \to 0$.

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

$$\begin{aligned} p_{G_n}(x) - p_{G_0}(x) &= \sum_{i=1}^{k_0+m_1} \sum_{j=1}^{s_i} p_{ij}^n (f(x|\theta_{ij}^n, \Sigma_{ij}^n) - f(x|\theta_i^0, \Sigma_i^0)) \\ &+ \sum_{i=1}^{k_0+m_1} (p_{i.}^n - p_i^0) f(x|\theta_i^0, \Sigma_i^0) \\ &+ \sum_{i=k_0+m_1+1}^{k_0+m} \sum_{j=1}^{s_i} p_{ij}^n f(x|\theta_{ij}^n, \Sigma_{ij}^n) \\ &:= A_n(x) + B_n(x) + C_n(x). \qquad (9) \end{aligned}$$

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

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

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

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

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

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 $i = 1, \ldots, k_0 + m$. (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 3.1, we shall argue that *not* all such coefficients vanish as $n \to \infty$. 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

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

Therefore, $\sum_{i=1}^{k_0+m} |p_{i.}^n - p_i^0|/D_n \to 1$ as $n \to \infty$. 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 $n \to \infty$, 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 $n \to \infty$, we obtain

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

where $\alpha_i \in \mathbb{R}, \beta_i, \nu_{i1}, \ldots, \nu_{is_i} \in \mathbb{R}^{d_1}$, $\gamma_i, \eta_{i1}, \ldots, \eta_{is_i}$ are symmetric matrices in $\mathbb{R}^{d_2 \times d_2}$ for all $1 \le i \le k_0 + m_1, 1 \le j \le s_i$. Additionally, $d_n C_n(x)/D_n = D_n^{-1} \sum_{i=k_0+m_1+1}^{k_0+m} \sum_{j=1}^{s_i} d_n p_{ij}^n f(x|\theta_{ij}^n, \Sigma_{ij}^n) \to 0$ due to the fact that for almost all x, $f(x|\theta_{ij}^n, \Sigma_{ij}^n) \to 0$ for all $k_0 + m_1 + 1 \le i \le k_0 + m, 1 \le j \le s_i$ and the fact that $d_n p_{ij}^n / D_n \le 1$ for all $k_0 + m_1 + 1 \le i \le k_0 + m$, $1 \le j \le s_i$. As a consequence, we obtain for almost all x that

$$\begin{aligned}
\sum_{i=1}^{k_0+m_1} \Bigg\{ \alpha_i f(x|\theta_i^0, \Sigma_i^0) + \beta_i^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0) + \sum_{j=1}^{s_i} \nu_{ij}^T \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^0, \Sigma_i^0)\nu_{ij} &\quad + \\
\text{tr}\left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \gamma_i\right) + 2\sum_{j=1}^{s_i} \nu_{ij}^T \left[\frac{\partial}{\partial \theta}\left(\text{tr}\left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \eta_{ij}\right)\right)\right] &\quad + \\
\sum_{j=1}^{s_i} \text{tr}\left(\frac{\partial}{\partial \Sigma}\left(\text{tr}\left(\frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0)^T \eta_{ij}\right)\right)^T \eta_{ij}\right) \Bigg\} &\quad = \quad 0. \quad (10)
\end{aligned}$$

Now, in this paragraph we will argue that not all coefficients in (10) go to 0 as $n \to \infty$. 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 $k_0+m_1+1 \le i \le k_0+m, 1 \le j \le s_i$ for infinitely many n. Then, it indicates that at least one coefficient in (10) should be 1. Our observation is proved. Case 2: Otherwise, m_n lies in the set $\{p_{ij}^n/D_n\}$ as $k_0+m_1+1 \le i \le k_0+m, 1 \le j \le s_i$ 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 coefficents in (10) vanish to 0. Therefore, $d_n|p_{i.}^n - p_i^0|/D_n = |p_{i.}^n - p_i^0|/p_{i^*j^*}^n \to 0$ 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 $|p_{i.}^n - p_i^0|/\sum_{i=1}^{k_0+m_1} |p_{i.}^n - p_i^0| \to 0$ for all $1 \leq i \leq k_0 + m_1$ as $n \to \infty$, which is a contradiction. Our observation is proved.

Therefore, at least one coefficient in (10) is different from 0. However, from the second-order identifiability of $\{f(x|\theta,\Sigma), \theta \in \Theta, \Sigma \in \Omega\}$, we obtain $\alpha_i = 0, \beta_i = \nu_{i1} = \ldots = \nu_{is_i} = \vec{0} \in \mathbb{R}^{d_1}, \gamma_i = \eta_{i1} = \ldots = \eta_{is_i} = \vec{0} \in \mathbb{R}^{d_2 \times d_2}$ for all $1 \leq i \leq k_0 + m_1$, which is a contradiction. This concludes the proof of Eq. (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 $\theta_1^n = \theta_1^0 - \frac{1}{n}\vec{1}_{d_1}, \theta_2^n = \theta_1^0 + \frac{1}{n}\vec{1}_{d_1}, \Sigma_1^n = \Sigma_1^0 - \frac{1}{n} I_{d_2}$ 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 $\vec{1}_n$ 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 $i = 2, \ldots, k_0$. Also, $p_1^n = p_2^n = \frac{p_1^0}{2}$ and $p_{i+1}^n = p_i^0$ for all $i = 2, \ldots, k_0$. 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 $n \to \infty$

$$V(p_{G_n}, p_{G_0}) \asymp \int\limits_{x \in \mathcal{X}} \left| \sum_{i=1}^{2} \sum_{\alpha_1, \alpha_2} (\Delta\theta_{1i}^n)^{\alpha_1} (\Delta\Sigma_{1i}^n)^{\alpha_2} \frac{\partial f}{\partial \theta^{\alpha_1} \partial \Sigma^{\alpha_2}}(x|\theta_1^0, \Sigma_1^0) + R_1(x) \right| \mathrm{d}x$$
$$= \int\limits_{x \in \mathcal{X}} |R_1(x)| \, \mathrm{d}x,$$

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

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

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 $\int |R_1(x)| \, \mathrm{d}x = O(n^{-2})$. 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

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

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).

Appendix

In this appendix, we give proofs of the following results: Theorem 3.4 regarding the characterization of strong identifiability in mixture models with matrix-variate parameters and most of the remained propositions and corollaries. For the transparency of our argument, the proofs for Theorem 3.4 are restricted to only first-order identifiability. The proof techniques are similar for the second-order identifiability. The proof of Theorem 3.3, which concerns the characterization of strong identifiability in multiple scalar parameters, issomewhat similar to that of Theorem 3.4 and therefore is omitted. Several easy proofs are also omitted. They include that of Theorem 3.5, which follows from an application of chain rule. Proof of Corollary 3.3 follows immediately from the triangle inequality. Proof of Corollary 4.1 relies on calculations of the bracket entropy integral, which is a straightforward extension of the argument of [9] to the multivariate setting.

6. Proofs of other results

6.1. Extension to the whole domain in exact-fitted mixtures

Proof of Corollary 3.1 By Theorem 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 $\frac{V(p_{G_n}, p_{G_0})}{W_1(G_n, G_0)} \to 0$ as $n \to \infty$. Since $\mathcal{G}$ is a compact set, we can find $G' \in \mathcal{G}$ and $W_1(G', G_0) > \epsilon$ such that $G_n \to G'$ under W_1 metric. It implies that $W_1(G_n, G_0) \to W_1(G', G_0)$ as $n \to \infty$. As $G' \not\equiv G_0$, we have $\lim_{n \to \infty} W_1(G_n, G_0) > 0$. As a consequence, $V(p_{G_n}, p_{G_0}) \to 0$ as $n \to \infty$.

From the hypothesis, $V(p_{G_n}, p_{G'}) \leq C(\Theta, \Omega) W_1^{\alpha}(G_n, G')$, so $V(p_{G_n}, p_{G'}) \to 0$ as $W_1(G_n, G') \to 0$. 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.

6.2. The importance of boundedness conditions in the over-fitted setting

Proof of Proposition 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 $i = 1, \ldots, k_0$, $\theta_{k_0+1}^n = \theta_1^0$, $\Sigma_{k_0+1}^n = \Sigma_1^0 + \frac{\exp(n/r)}{n^{\alpha}} I_{d_2}$ 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

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

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

6.3. Characterization of strong identifiability

Proof of Theorem 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 $(\theta_1, \Sigma_1, m_1), \ldots, (\theta_k, \Sigma_k, m_k)$, 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 $j = 1, \ldots, k$ such that:

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

Substituting the first derivatives of f to get

$$\sum_{j=1}^{k} \Bigg\{ \alpha_j' + \left((\beta_j')^T (x - \theta_j) + (x - \theta_j)^T \gamma_j' (x - \theta_j)\right) \times \left[(x - \theta_j)^T \Sigma_j^{-1} (x - \theta_j)\right]^{m_j - 1} + \eta_j' \log[(x - \theta_j)^T \Sigma_j^{-1} (x - \theta_j)] \Bigg\} \times \exp\left(-\left[(x - \theta_j)^T \Sigma_j^{-1} (x - \theta_j)\right]^{m_j}\right) = 0, \quad (11)$$

where

$$\alpha_j' = \frac{2\alpha_j m_j \Gamma(d/2) - m_j \Gamma(d/2) \operatorname{tr}(\Sigma_j^{-1} \gamma_j) + 2\eta_j \Gamma(d/2) \left(1 - \frac{d}{2m_j} \psi\left(\frac{d}{2m_j}\right)\right)}{2\pi^{d/2} \Gamma(d/(2m_j)) |\Sigma_j|^{1/2}},$$

$$\beta'_j = \frac{2m_j^2\Gamma(d/2)}{\pi^{d/2}\Gamma(d/(2m_j))|\Sigma_j|^{1/2}}\Sigma_j^{-1}\beta_j, \quad \gamma'_j = \frac{m_j^2\Gamma(d/2)}{\pi^{d/2}\Gamma(d/(2m_j))|\Sigma_j|^{1/2}}\Sigma_j^{-1}\gamma_j\Sigma_j^{-1},$$

$$\text{and} \quad \eta'_j = \frac{-m_j\eta_j\Gamma(d/2)}{\pi^{d/2}\Gamma(d/(2m_j))|\Sigma_j|^{1/2}}.$$

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

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

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

Borrowing a technique from [26], since $(\theta_1, \Sigma_1), \ldots, (\theta_{\overline{i}}, \Sigma_{\overline{i}})$ are distinct, we have two possibilities:

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

Possibility 2 If Σ_j are not the same for all $1 \leq j \leq \overline{i}$, then we assume without loss of generality that $\Sigma_1, \ldots, \Sigma_m$ are the only distinct matrices from $\Sigma_1, \ldots, \Sigma_{\overline{i}}$, where $m \leq \overline{i}$. Denote $\delta_{ij} = \Sigma_i - \Sigma_j$ as $1 \leq i < j \leq m$, then as x' does not belong to $\bigcup_{1 \leq i < j \leq m} \left\{u \in \mathbb{R}^d : u^T\delta_{ij}u = 0\right\}$, we will have $(x')^T\Sigma_1 x', \ldots, (x')^T\Sigma_m x'$ are distinct. Therefore, if x' does not belong to $\bigcup_{1 \leq i < j \leq m} \left\{u \in \mathbb{R}^d : u^T\delta_{ij}u = 0\right\}$, which is a finite union of conics, then we have $((x')^T\theta_1, (x')^T\Sigma_1 x'), \ldots, ((x')^T\theta_m, (x')^T\Sigma_m x')$ are distinct. Additionally, for any θ_j where $m+1 \leq j \leq \overline{i}$ 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, $((x')^T\theta_1, (x')^T\Sigma_1 x'), \ldots, ((x')^T\theta_{\overline{i}}, (x')^T\Sigma_{\overline{i}}x')$ 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 $x' \notin D$, $((x')^T\theta_1, (x')^T\Sigma_1 x'), \ldots ((x')^T\theta_{\overline{i}}, (x')^T\Sigma_{\overline{i}}x')$ are distinct. Thus, (d_i, e_i) are different as $1 \leq i \leq \overline{i}$.

Choose $d_{i_1} = \min_{1 \leq i \leq \overline{i}} \{d_i\}$. Denote $J = \left\{1 \leq i \leq \overline{i} : d_i = d_{i_1}\right\}$. Choose $1 \leq i_2 \leq \overline{i}$ 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 (11) with $\exp -(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}}$, we get

$$A_{i_2}(x_1) + \sum_{j \neq i_2} A_j(x_1) \exp\Big[(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} \quad - \\ (d_j x_1^2 + e_j x_1 + f_j)^{m_j}\Big] \quad = \quad 0. \qquad (12)$$

Note that if $j \in J \backslash \{i_2\}$, $d_j = d_{i_2}$, $m_j = m_{i_2}$, and $e_j > e_{i_2}$. So, $(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} \lesssim -x_1$ as x_1 is large enough. This implies that when $x_1 \to \infty$,

$$B_1(x_1) := \sum_{j \neq J\backslash\{i_2\}} A_j(x_1) \exp\Big[(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} \quad - \\ (d_j x_1^2 + e_j x_1 + f_j)^{m_j}\Big] \quad \to \quad 0.$$

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

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

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

$$B_2(x_1) := \sum_{\substack{j \notin J, \\ 1 \leq j \leq \overline{i}}} A_j(x_1) \exp\Big[(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} \quad - \\ (d_j x_1^2 + e_j x_1 + f_j)^{m_j}\Big] \quad \to \quad 0.$$

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

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

As a result,

$$B_3(x_1) := \sum_{j > \overline{i}} A_j(x_1) \exp\Big[(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} \quad - \\ (d_j x_1^2 + e_j x_1 + f_j)^{m_j}\Big] \quad \to \quad 0.$$

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

$$\sum_{j \neq i_2} A_j(x_1) \exp\Big[(d_{i_2}x_1^2 + e_{i_2}x_1 + f_{i_2})^{m_{i_2}} - (d_j x_1^2 + e_j x_1 + f_j)^{m_j}\Big] \quad = \\ A_1(x) + A_2(x) + A_3(x) \quad \to \quad 0. \qquad (13)$$

Combining (12) and (13), we obtain that as $x_1 \to \infty$, $A_{i_2}(x_1) \to 0$. The only possibility for this result to happen is $a_{i_2} = b_{i_2} = \eta'_{i_2} = 0$. Or, equivalently, $(x')^T \gamma'_{i_2} x' = \left[(\beta'_i)^T - 2\theta_{i_2}^T \gamma'_{i_2}\right] x' = 0$. If $\gamma'_{i_2} \neq 0$, we can choose the element $x' \notin D$ lying outside the hyperplane $\left\{u \in \mathbb{R}^d : u^T \gamma'_{i_2} u = 0\right\}$. 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 $(\theta_1, \Sigma_1), ..., (\theta_k, \Sigma_k)$, 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:

$$\sum_{j=1}^{k} \alpha_j f(x|\theta_j, \Sigma_j) + \beta_j^T \frac{\partial f}{\partial \theta}(x|\theta_j, \Sigma_j) + \text{tr}(\frac{\partial f}{\partial \Sigma}(x|\theta_j, \Sigma_j)^T \gamma_j) = 0. \tag{14}$$

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

$$\begin{aligned}\sum_{j=1}^{k} \Big[\int\limits_{\mathbb{R}^d} \Big(\frac{\alpha'_j \exp(i(\Sigma_j^{1/2} t)^T x)}{(\nu + \|x\|^2)^{(\nu+d)/2}} + \frac{\exp(i(\Sigma_j^{1/2} t)^T x)(\beta'_j)^T x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} \quad + \\ \frac{\exp(i(\Sigma_j^{1/2} t)^T x) x^T M_j x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} \Big) dx \Big] \exp(it^T \theta_j) \quad = \quad 0,\end{aligned} \tag{15}$$

where $\alpha'_j = \alpha_j - \frac{\text{tr}(\Sigma_j^{-1}\gamma_j)}{2}, \beta'_j = \frac{(\nu+d)}{2} \Sigma^{-1/2} \beta_j$, 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 (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 $(\frac{t_1}{\|t\|}, ..., \frac{t_d}{\|t\|})^T$, 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:

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

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 $1 \leq j \leq k$

$$\int\limits_{\mathbb{R}^d} \frac{\exp(i(\Sigma_j^{1/2}t)^T x)}{(\nu + \|x\|^2)^{(\nu+d)/2}} dx = C_1 A_1(\|\Sigma_j^{1/2}t\|). \tag{16}$$

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

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

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 $t' \in \mathbb{R}$.

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

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 $1 \leq j \leq d$

$$\int\limits_{\mathbb{R}^d} \frac{\exp(i(\Sigma_j^{1/2}t)^T x)(\beta'_j)^T x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} dx = \frac{C_2(\beta'_j)^T \Sigma_j^{1/2} t A_2(\|\Sigma_j^{1/2}t\|)}{\|t\|}. \tag{17}$$

$$\begin{aligned} \int\limits_{\mathbb{R}^d} \frac{\exp(i(\Sigma_j^{1/2}t)^T x) x^T M_j x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} dx = \frac{1}{\|\Sigma_j^{1/2}t\|^2}(\sum_{u,v} M_{uv}^j [\Sigma_j^{1/2}t]_u [\Sigma_j^{1/2}t]_v) \times \\ \times(C_2 A_3(\|\Sigma_j^{1/2}t\|) - C_3 A_1(\|\Sigma_j^{1/2}t\|)) + C_3(\sum_{l=1}^{d} M_{ll}^j)A_1(\|\Sigma_j^{1/2}t\|), \end{aligned} \tag{18}$$

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 (16), (17), and (18), we can rewrite (15) as:

$$\sum_{j=1}^{k} \left[\alpha'_j A_1(\|\Sigma_j^{1/2}t\|) + C_2 \frac{(\Sigma_j^{1/2}t)^T \beta'_j}{\|\Sigma_j^{1/2}t\|} A_2(\|\Sigma_j^{1/2}t\|) \quad + \right.$$

$$C_3(\sum_{l=1}^{d} M^j_{ll})A_1(\|\Sigma_j^{1/2}t\|) + \left(\sum_{u,v} M^j_{uv}\frac{[\Sigma_j^{1/2}t]_u[\Sigma_j^{1/2}t]_v}{\|\Sigma_j^{1/2}t\|^2}\right)(C_2A_3(\|\Sigma_j^{1/2}t\|) \quad - \quad C_3A_1(\|\Sigma_j^{1/2}t\|))\Big]\exp(it^T\theta_j) \quad = \quad 0.$$

Define $t = t_1t'$, 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 $t' \notin D$, $((t')^T\theta_1, (t')^T\Sigma_1 t'), ...((t')^T\theta_k, (t')^T\Sigma_k t')$ 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:

$$\sum_{j=1}^{k}\Big[\alpha'_jA_1(\sigma_j|t_1|) + C_2\frac{t_1(\Sigma_j^{1/2}t')^T\beta'_j}{|t_1|\sigma_j}A_2(\sigma_j|t_1|) + C_3(\sum_{l=1}^{d}M^j_{ll})A_1(\sigma_j|t_1|) \quad + \left(\sum_{u,v}M^j_{uv}\frac{[\Sigma_j^{1/2}t']_u[\Sigma_j^{1/2}t']_v}{\sigma_j^2}\right)(C_2A_3(\sigma_j|t_1|) - C_3A_1(\sigma_j|t_1|)\Big]\exp(i\theta'_jt_1) \quad = \quad 0.$$

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

$$\sum_{j=1}^{k}\Big[\Big(\alpha'_j + C_3(\sum_{l=1}^{d}M^j_{ll}) - C_3\Big(\sum_{u,v}M^j_{uv}\frac{[\Sigma_j^{1/2}t']_u[\Sigma_j^{1/2}t']_v}{\sigma_j^2}\Big)\Big)| \quad \times \times A_1(\sigma_j|t_1|) + C_2\Big(\sum_{u,v}M^j_{uv}\frac{[\Sigma_j^{1/2}t']_u[\Sigma_j^{1/2}t']_v}{\sigma_j^2}\Big)A_3(\sigma_j|t_1|) \quad + C_2(it_1)\frac{(\Sigma_j^{1/2}t')^T\beta'_j}{\sigma_j}A_1(\sigma_j|t_1|)\Big]\exp(i\theta'_jt_1) \quad = \quad 0. \ (19)$$

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 (19) can be rewritten as:

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

where we have $\alpha_j^{''} = \alpha_j' + C_3(\sum_{l=1}^{d} M_{ll}^j) - C_3(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2})$, $\beta_j^{''} = C_2 \frac{(\Sigma_j^{1/2} t')^T \beta_j'}{\sigma_j}$, and $\gamma_j^{''} = C_2(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2})$. The above equation yields that for all $t_1 \geq 0$

$$\sum_{j=1}^{k} \left[\left(\alpha_j^{''} + \beta_j^{''}(it_1) \right) \sum_{u=0}^{l-1} a_u \sigma_j^u t_1^u + \gamma_j^{''} \sum_{u=0}^{l} b_u \sigma_j^u t_1^u \right] \quad \times$$
$$\exp\left(it_1\theta_j' - \sigma_j \sqrt{2l-1} t_1 \right) \quad = \quad 0. \tag{20}$$

Using the Laplace transformation on both sides of (20) and denoting $c_j = \sigma_j \sqrt{2l-1} - i\theta_j'$ as $1 \leq j \leq k$, we obtain that as $\text{Re}(s) > \max_{1 \leq j \leq k} \left\{ -\sigma_j \sqrt{2l-1} \right\}$

$$\sum_{j=1}^{k} \alpha_j^{''} \sum_{u=0}^{l-1} \frac{u! a_u \sigma_j^u}{(s+c_j)^{u+1}} + i\beta_j^{''} \sum_{u=1}^{l} \frac{u! a_{u-1} \sigma_j^{u-1}}{(s+c_j)^{u+1}} \quad +$$
$$\gamma_j^{''} \sum_{u=0}^{l} \frac{u! b_u \sigma_j^u}{(s+c_j)^{u+1}} \quad = \quad 0. \tag{21}$$

Without loss of generality, we assume that $\sigma_1 \leq \sigma_2 \leq ... \leq \sigma_k$. It demonstrates that $-\sigma_1 \sqrt{2l-1} = \max_{1 \leq j \leq k} \left\{ -\sigma_j \sqrt{2l-1} \right\}$. 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 (21) with $(s+c_1)^{l+1}$, as $\text{Re}(s) > -\sigma_1 \sqrt{2l-1}$ and $s \to -c_1$, 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 (21) with $(s+c_1)^l$ and using the same argument, as $s \to -c_1$, 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

$$\alpha_j' + C_3(\sum_{l=1}^{d} M_{ll}^j) - C_3(\sum_{u,v} M_{uv}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2}) = 0, \; \frac{(\Sigma_j^{1/2} t')^T \beta_j'}{\sigma_j} = 0,$$

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

$$\alpha_j' + C_3(\sum_{l=1}^{d} M_{ll}^j) = 0, (t')^T \Sigma_j^{1/2} \beta_j' = 0, \text{ and } (t')^T \gamma_j t' = 0.$$

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.

References

[1] 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. MR2549554

[2] M. Belkin and K. Sinha. Polynomial learning of distribution families. In *FOCS*, 2010. MR3024780

[3] R. J. Carroll and P. Hall. Optimal rates of convergence for deconvolving a density. *Journal of American Statistical Association*, 83:1184–1186, 1988. MR0997599

[4] J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995. MR1331665

[5] A. DasGupta. *Asymptotic Theory of Statistics and Probability.* Springer, 2008. MR2664452

[6] S. Dasgupta. Learning mixtures of Gaussians. Technical Report UCB/CSD-99-1047, University of California, Berkeley, 1999.

[7] 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. MR2141286

[8] J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19(3):1257–1272, 1991. MR1126324

[9] 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. MR1873329

[10] P. Hall, A. Neeman, R. Pakyari, and R. Elmore. Nonparametric inference in multivariate mixtures. *Biometrika*, 92:667–678, 2005. MR2202653

[11] P. Hall and X. H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 31:201–224, 2003. MR1962504

[12] 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. MR0628771

[13] A. Kalai, A. Moitra, and G. Valiant. Disentangling gaussians. *Communications of the ACM*, 55(2):113–120, 2012.

[14] J. T. Kent. Identifiability of finite mixtures for directional data. *Annals of Statistics*, 11:984–988, 1983. MR0707948

[15] B. Lindsay. *Mixture models: Theory, geometry and applications.* In NSF-CBMS Regional Conference Series in Probability and Statistics. IMS, Hayward, CA., 1995.

[16] X. Liu and Y. Shao. Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics*, 31:807–832, 2004. MR1994731

[17] K. V. Mardia. Statistics of directional data. *Journal of the Royal Statistical Society. Series B(Methodological)*, 37:349–393, 1975. MR0402998

[18] G. J. McLachlan and K. E. Basford. *Mixture models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs.* New York, 1988. MR0926484

[19] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 41(1):370–400, 2013. MR3059422

[20] D. Peel and G. J. McLachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10:339–348, 2000.

[21] 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. MR2867454

[22] H. Teicher. Identifiability of mixtures. *Annals of Statistics*, 32:244–248, 1961. MR0120677

[23] H. Teicher. Identifiability of finite mixtures. *Annals of Statistics*, 34:1265–1269, 1963. MR0155376

[24] S. van de Geer. *Empirical Processes in M-estimation.* Cambridge University Press, 2000. MR1739079

[25] Cédric Villani. *Optimal transport: Old and New.* Springer, 2008. MR2459454

[26] S. J. Yakowitz and J. D. Spragins. On the identifiability of finite mixtures. *Annals of Statistics*, 39(1):209–214, 1968. MR0224204

[27] B. Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435, 1997. MR1462963

[28] C. Zhang. Fourier methods for estimating mixing densities and distributions. *Annals of Statistics*, 18(2):806–831, 1990. MR1056338

[29] 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. MR3085302