On strong identifiability and optimal 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 optimal rates of convergence for the maximum likelihood estimates are established for such classes, which are also confirmed by simulation studies.


Keywords and phrases: mixture models, strong identifiability, Wasserstein distances, minimax bounds, maximum likelihood estimation.

1. Introduction

Mixture models are popular modeling tools for making inference about heterogeneous data [15, 18]. Under the mixture modeling, data are viewed as samples from a collection of unobserved or latent subpopulations, each posits 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 [21, 22, 25] and others, and continues to attract much interest [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 model. This approach is well-developed in the context of nonparametric deconvolution [3, 27, 8], but these results are confined to only a specific type of model – the 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 [4].

* This research is supported in part by grants NSF CCF-1115769, NSF CAREER DMS-1351362, and NSF CNS-1409303.
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. Chen’s work, however, was restricted to models that have only a single scalar parameter. This restriction was effectively removed by Nguyen, who showed that Wasserstein distances (cf. [24]) provide a natural source of metrics for deriving rates of convergence of mixing distributions [19]. He established rates of convergence of mixing distributions for a number of finite and infinite mixture models with multi-dimensional parameters. Rousseau and Mengersen studied over-fitted mixtures in a Bayesian estimation setting [20]. Although they did not focus on mixing distributions per se, they showed that the mixing probabilities associated with extra mixing components vanish at a standard \( n^{-1/2} \) rate, subject to the parameter prior distribution and a strong identifiability condition on the density class. 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. Moreover, the convergence rates that we derive will be shown to be optimal in the minimax sense. Assume that each subpopulation is distributed by a density function (with respect to Lebesgue measure on an Euclidean space \( \mathcal{X} \)) that belongs to a known density class \( \{ f(x|\theta, \Sigma), \theta \in \Theta \subset \mathbb{R}^{d_1}, \Sigma \in \Omega \subset \mathbb{R}^{d_2}_{++}, x \in \mathcal{X} \} \).

Here, \( d_1 \geq 1, d_2 \geq 0, \mathbb{R}^{d_2}_{++} \) 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 Wasserstein distance on the space of mixing measures \( G \). Let \( G = \sum_{i=1}^{k} p_i \delta(\theta_i, \Sigma_i) \) and \( G_0 = \sum_{i=1}^{k_0} p_i^0 \delta(\theta_i^0, \Sigma_i^0) \) be two discrete probability measures on \( \Theta \times \Omega \), which is equipped with metric \( \rho \). Recall the Wasserstein distance of order \( r \), for a given \( r \geq 1 \):

\[
W_r(G, G_0) = \left( \inf_q \sum_{i,j} q_{ij} \rho^r( (\theta_i, \Sigma_i), (\theta_j^0, \Sigma_j^0) ) \right)^{1/r},
\]

where the infimum is taken over all joint probability distributions \( q \) on \( [1, \ldots, k] \times [1, \ldots, k_0] \) such that, when expressing \( q \) as a \( k \times k_0 \) matrix, the marginal constraints
hold: $\sum_j q_{ij} = p_i$ and $\sum_i q_{ij} = p'_j$. Suppose that a sequence of mixing measures $G_n \to G_0$ under $W_\tau$ 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 $\omega_n$ under $\rho$ 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 $\omega_n$. 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 $\omega'_n$.

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 Wasserstein distances $W_\tau(G, G')$ (shorthanded by $W_\tau$). Establishing (sharp) lower bounds for $V$ in terms of $W_\tau$ is the main challenge. Such a bound 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 $W_\tau \neq 0$.

**General theory of strong identifiability** The classical identifiability condition requires that $p_G = p_{G'}$ entails $G = G'$. This amounts to the linear independence of elements $f$ in the density class [22]. In order to establish quantitative lower bounds on a distance of mixture densities, we introduce 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 the function $f$ with respect to all parameters in the model, and insisting that these quantities be linearly independent in 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, 20, 19]. We show that for exact-fitted mixtures, the first-order identifiability condition (along with additional and mild regularity conditions) suffices for obtaining that

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

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,

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

(2)
when $W_2(G, G_0)$ is sufficiently small. The lower bound $W_2^2(G, G_0)$ is sharp, i.e. we can not improve the lower bound to $W_1^2$ for any $r < 2$ (notably, $W_2 \geq W_1$). A consequence of this result is, take any standard estimation method (such that the MLE) which yields $n^{-1/2}$ convergence rate for $p_G$, the induced rate of convergence for the mixing measure $G$ is the minimax optimal $n^{-1/4}$ under $W_2$. This means the mixing probability mass converge at $n^{-1/2}$ rate (which recovers the result of [20]), in addition to having that 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 kind 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 $p_G$, and let $\hat{G}_n$ be the MLE estimate of the mixing distribution $G$ ranging among all discrete probability distributions with at most $k$ support points in $\Theta \times \Omega$ under the over-fitted setting or among all discrete probability distributions with exactly $k_0$ support points in $\Theta \times \Omega$ under the exact-fitted setting. The inequalities (1) and (2), and the fact that these bounds are sharp enable us to easily establish the convergence rates of the mixing measures, and to show that these rates are minimax optimal. Such results are stated in Theorem 4.1, Theorem 4.2, and Theorem 4.3. 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.

Summarizing, the novel 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.

(iii) Characterization results showing the applicability of our theory and the convergence rates to a broad range of mixture models with parameters of multiple types, including matrix-variate ones.

(iv) Another novelty of this work is that the settings of exact-fitted and over-fitted mixtures are treated separately, by highlighting the role of the first-order identifiability criterion and $W_1$ metric in the former, as opposed to the second-order identifiability and $W_2$ metric in the latter.

Finally, we note in passing that both the first and second-order identifiability are in some sense necessary in deriving the optimal 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 bounds and the convergence rates of the maximum likelihood estimation, which are optimal in many cases. 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 remaining results are presented in the Appendix.

**Notation** Divergence distances studied in this paper include the total variational distance $V(p_G, p_{G'}) = \frac{1}{2} \int |p_G(x) - p_{G'}(x)|d\mu(x)$ and the Hellinger distance $h^2(p_G, p_{G'}) = \frac{1}{2} \int \left( \sqrt{p_G(x)} - \sqrt{p_{G'}(x)} \right)^2 d\mu(x)$. As $K, L \in \mathbb{N}$, the first derivative of real function $g : \mathbb{R}^{K \times L} \to \mathbb{R}$ of matrix $\Sigma$ 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 $(\theta, \Sigma)$, 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 $\frac{\partial}{\partial \theta} \left( \frac{\partial g_2}{\partial \Sigma_{ij}} \right)$. Finally, 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.

2. Preliminaries

First of all, we need to define our notion of distances on the space of mixing measures $G$. 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 mixing measure $G = \sum_{i=1}^{k} p_i \delta_{(\theta_i, \Sigma_i)}$, where $p = (p_1, p_2, \ldots, p_k)$ denotes the proportion vector. Likewise, let $G' = \sum_{i=1}^{k'} p'_i \delta_{(\theta'_i, \Sigma'_i)}$. A coupling between $p$ and $p'$ is a joint distribution $q$ on $[1, \ldots, k] \times [1, \ldots, k']$, which is expressed as a matrix $q = (q_{ij})_{1 \leq i \leq k, 1 \leq j \leq k'} \in [0, 1]^{k \times k'}$ and admits marginal constraints $\sum_{i=1}^{k} q_{ij} = \sum_{j=1}^{k'} q_{ij} = 1$.
of \( p' \) and \( \sum_{i=1}^{k'} q_{ij} = p_i \) for any \( i = 1, 2, \ldots, k \) and \( j = 1, 2, \ldots, k' \). We call \( q \) a coupling of \( p \) and \( p' \), and use \( \mathcal{Q}(p, p') \) 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 mass from one probability measure to another [24]. For any \( r \geq 1 \), the \( r \)-th order Wasserstein distance between \( G \) and \( G' \) is given by

\[
W_r(G,G') = \left( \inf_{q \in \mathcal{Q}(p,p')} \sum_{i,j} q_{ij} (\|\theta_i - \theta_j\| + \|\Sigma_i - \Sigma_j\|)^r \right)^{1/r}.
\]

In both equations 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 distances of 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 an application of 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 matrix-variate ones, along with the aforementioned upper bounds.

**Example 2.1.** (Multivariate generalized Gaussian distribution [28])

Denote \( f(x|\theta, m, \Sigma) = \frac{m \Gamma(d/2)}{\pi^{d/2} \Gamma((d/(2m))\|\Sigma\|^{1/2})} \exp(-[(x-\theta)^\top \Sigma^{-1}(x-\theta)]^m) \), where \( \theta \in \mathbb{R}^d, m > 0, \) and \( \Sigma \in S_d^{++} \). If \( \Theta_1 \) is bounded subset of \( \mathbb{R}^d \), \( \Theta_2 = \{m \in \mathbb{R}^{+} : 1 \leq m \leq m \leq m \} \), and \( \Omega = \{ \Sigma \in S_d^{++} : \lambda \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq \bar{\lambda} \} \), then for any mixing measures \( G_1, G_2 \), we obtain \( h^2(p_{G_1}, p_{G_2}) \lesssim W_2^2(G_1, G_2) \) and \( V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2) \).

**Example 2.2.** (Multivariate Student’s t-distribution)

Let \( f(x|\Sigma) = C_\nu \left( \nu + (x-\theta)^\top \Sigma^{-1}(x-\theta) \right)^{-\nu + d/2} \), where \( \nu \) is a fixed positive degree of freedom and \( C_\nu = \frac{\Gamma(\nu+d/2)}{\Gamma(\nu/2)\pi^{d/2}} \). If \( \Theta \) is bounded subset of \( \mathbb{R}^d \) and \( \Omega = \{ \Sigma \in S_d^{++} : \lambda \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq \bar{\lambda} \} \), then for any mixing measures \( G_1, G_2 \), we obtain \( h^2(p_{G_1}, p_{G_2}) \lesssim W_2^2(G_1, G_2) \) and \( V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2) \).

**Example 2.3.** (Exponentially modified multivariate Student’s t-distribution)

Let \( f(x|\theta, \lambda, \Sigma) \) to be density of \( X = Y + Z \), where \( Y \) follows multivariate t-distribution with location \( \theta \), covariance matrix \( \Sigma \) fixed positive degree of freedom \( \nu \), and \( Z \) is distributed by the product of \( d \) independent exponential distributions with combined shape \( \lambda = (\lambda_1, \ldots, \lambda_d) \). If \( \Theta \) is bounded subset of \( \mathbb{R}^d \times \mathbb{R}_d^+ \), where \( \mathbb{R}_d^+ = \{ x \in \mathbb{R}^d : x_i > 0 \forall i \} \), \( \Omega = \{ \Sigma \in S_d^{++} : \lambda \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq \bar{\lambda} \} \), then for any \( G_1, G_2 \), \( h^2(p_{G_1}, p_{G_2}) \lesssim W_2^2(G_1, G_2) \) and \( V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2) \).
\[ W^2(G_1, G_2) \text{ and } V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2). \]

**Example 2.4. (Modified Gaussian-Gamma distribution)**

Let \( f(x|\theta, \alpha, \beta, \Sigma) \) to be density function of \( X = Y + Z \), where \( Y \) is distributed by multivariate Gaussian distribution with mean \( \theta \), covariance matrix \( \Sigma \), 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 \( \beta = (\beta_1, \ldots, \beta_d) \). If \( \Theta \) is bounded subset of \( \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d \), \( \Omega = \{ \Sigma \in S^{d^+}_d : \lambda \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq \bar{\lambda} \} \), then for any \( G_1, G_2 \in \mathcal{G}(\Theta \times \Omega) \), \( h^2(p_{G_1}, p_{G_2}) \lesssim V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2) \).

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 requires that \( p_G = p_{G'} \) entails \( G = G' \), 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 need stronger notions of identifiability, ones which involve higher order derivatives of density function \( f \), taken with respect to the mixture model parameters. The strength of this theory is that it holds generally for a reasonably broad range of mixture models, which allow 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 \( (\theta, \Sigma) \) and the following holds

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

then, \( \alpha_i = 0, \beta_i = 0 \in \mathbb{R}^{d}, \gamma_i = 0 \in \mathbb{R}^{d \times d} \) for \( i = 1, \ldots, k \).

**Remark** The condition that \( \gamma_i \) is symmetric in Definition 3.1 is crucial, without which the identifiability condition would fail for many classes of density. For instance, assume that \( \frac{\partial f}{\partial \Sigma}(x|\theta_i, \Sigma_i) \) are symmetric matrices for all \( i \) – this clearly holds for any elliptical distributions, such as multivariate Gaussian, Student’s t-distribution, and logistic distribution – if we choose \( \gamma_i \) to be anti-symmetric matrices, then by choosing \( \alpha_i = 0, \beta_i = 0, (\gamma_i)_{uu} = 0 \) for all \( 1 \leq u \leq d \) (i.e. all diagonal elements are 0), A1 holds while \( \gamma_i \) can be different from 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 $\delta_1, \delta_2$ such that for any $R_1, R_2, R_3 > 0$, $\gamma_1 \in \mathbb{R}^{d_1}$, $\gamma_2 \in \mathbb{R}^{d_2}$, $R_1 \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_{d_1}(\Sigma)} \leq R_2$, $||\theta|| \leq R_3$, $\theta_1, \theta_2 \in \Theta$, $\Sigma_1, \Sigma_2 \in \Omega$, there are positive constants $C(R_1, R_2)$ and $C(R_3)$ such that for all $x \in \mathcal{X}$

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

(3)

$$\left| \mathrm{tr} \left( \left( \frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma_1) - \frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma_2) \right)^T \gamma_2 \right) \right| \leq C(R_3)||\Sigma_1 - \Sigma_2||^2 ||\gamma_2||. \quad (4)$$  

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 uniform Lipschitz property up to the first order. Then there are positive constants $\epsilon_0$ and $C_0$, both depending on $G_0$, such that as long as $G \in \mathcal{E}_{k_0}(\Theta \times \Omega)$ 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 $\Theta$ or $\Omega$ is required, the lower bound is of local nature, in the sense that it holds only for those $G$ sufficiently close to $G_0$ by a Wassertein distance at most $\epsilon_0$, 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.** Given all assumptions of Theorem 3.1. 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 $V(p_{G_1}, p_{G_2}) \lesssim W_1(G_1, G_2)$. Then, for a fixed compact subset $\mathcal{G}$ of $\mathcal{E}_{k_0}(\Theta \times \Omega)$, there is a positive constant $C_0 = C_0(G_0)$ such that

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

We shall verify in the sequel that the classes of densities $f$ described in Examples 2.1, 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 (see also (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 $(\theta, \Sigma)$ and the following assumption holds
A2. For any finite $k$ different pairs $(\theta_1, \Sigma_1), \ldots, (\theta_k, \Sigma_k) \in \Theta \times \Omega$, if we have $\alpha_i \in \mathbb{R}, \beta_i, \nu_i \in \mathbb{R}^{d_1}, \gamma_i, \eta_i, \eta \in \Theta$ and symmetric matrices in $\mathbb{R}^{d_2 \times d_2}$ as $i = 1, \ldots, k$ such that for almost all $x$

$$\sum_{i=1}^{k} \left\{ \alpha_i f(x|\theta_i, \Sigma_i) + \beta_i \frac{\partial f}{\partial \theta}(x|\theta_i, \Sigma_i) + \nu_i \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i, \Sigma_i) \nu_i \right\} + \n\n\n\n$$

$$\text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta_i, \Sigma_i)^T \gamma_i \right) + 2\nu_i \left[ \frac{\partial}{\partial \theta} \left( \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta_i, \Sigma_i)^T \eta_i \right) \right) \right] + \n\n\n\n$$

$$\text{tr} \left( \frac{\partial}{\partial \Sigma} \left( \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta_i, \Sigma_i)^T \eta_i \right) \right)^T \eta_i \right) \right\} = 0,$$

then, $\alpha_i = 0, \beta_i = \nu_i = 0 \in \mathbb{R}^{d_1}, \gamma_i = \eta_i = 0 \in \mathbb{R}^{d_2 \times d_2}$ 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 $\delta_3, \delta_4$ such that for any $R_1, R_5, R_6 > 0, \gamma_1 \in \mathbb{R}^{d_1}, \gamma_2 \in \mathbb{R}^{d_2 \times d_2}, R_4 \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_{d_2}(\Sigma)} \leq R_5,$ $||\theta|| \leq R_6, \theta_1, \theta_2 \in \Theta, \Sigma_1, \Sigma_2 \in \Omega,$ there are positive constants $C_1$ depending on $(R_1, R_5)$ and $C_2$ depending on $R_6$ such that for all $x \in \mathcal{X}$

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

$$\left| \text{tr} \left( \left[ \frac{\partial}{\partial \Sigma} \left( \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma_1)^T \gamma_2 \right) \right) - \frac{\partial}{\partial \Sigma} \left( \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta, \Sigma_2)^T \gamma_2 \right) \right) \right] \gamma_2 \right) \right| \leq C_2 \|\Sigma_1 - \Sigma_2\|_2 \||\gamma_2\|_2^2.$$
Here and elsewhere, ratio $V/W$ (iii) Part (b) demonstrates the sharpness of the bound in part (a). In particular, we (iv) It is also worth mentioning that the boundedness of mild condition $\lim_{\epsilon \to 0} V(p_G, p_{G_0})/W^r(G, G_0) : W_1(G, G_0) \leq \epsilon = 0$.

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

$$\sup_{\|\theta - \theta_0\| + \|\Sigma - \Sigma_0\| \leq c_0} \int_{\mathcal{X}} \left( \frac{\partial^2 f}{\partial \theta \partial \Sigma} (x|\theta, \Sigma) \right)^2 / f(x|\theta', \Sigma') \, dx < \infty$$

where $\alpha_1, \alpha_2$ are defined as that of part (b). Then, for any $1 \leq r < 2$:

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

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

(i) We note that a counterpart of part (a) for finite mixtures with multivariate parameters was given in [19] (Proposition 1). The proof in that paper has a problem: it relies on Nguyen’s Theorem 1, which holds only for the exact-fitted setting, but not for the over-fitted setting. This was pointed out to the second author by Elisabeth Gassiat who attributed it to Jonas Kahn. Fortunately, this error can be simply corrected by replacing Nguyen’s Theorem 1 with a weaker version, which holds for the over-fitted setting and suffices for our purpose, for which his method of proof continues to apply: it suffices to prove only the following weaker version:

$$\lim_{\epsilon \to 0} \inf_{\theta \in \Theta \times \Omega} \left\{ V(p_G, p_{G_0})/W_2^r(G, G_0) : W_2(G, G_0) \leq \epsilon \right\} > 0.$$

(ii) The mild condition $\lim_{\lambda_1(\Sigma) \to 0} f(x|\theta, \Sigma) = 0$ is important for the matrix-variate parameter $\Sigma$. In particular, it is useful for addressing the scenario when the smallest eigenvalue of matrix parameter $\Sigma$ is not bounded away from 0. This condition, however, can be removed if we impose that $\Sigma$ is a positive definite matrix 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 $G$. Part (c) implies that $n^{-1/4}$ is minimax optimal. See Section 4.1 for formal statements of such a result.

(iv) It is also worth mentioning that the boundedness of $\Theta$, as well as the boundedness from above of the eigenvalues of elements of $\Omega$ are both necessary conditions. 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 distance $h \geq V$ can vanish much faster than $W_r(G, G_0)$, as can be seen by:
Proposition 3.1. Let $\Theta$ be a subset of $\mathbb{R}^d$ and $\Omega = S_{d_2}^{++}$. Then for any $r \geq 1$ and $\beta > 0$ we have

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

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

Corollary 3.2. Given all assumptions of Theorem 3.2 (part (a)), in addition to $\Theta$ and $\Omega$ being compact. 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 $V(p_{G_1},p_{G_2}) \lesssim W_2^2(G_1,G_2)$. 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

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

3.2. Characterization of strong identifiability

In this subsection we identify a 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 shows how strong identifiability conditions continue to be preserved under certain transformations with respect to the parameter space. First, we consider univariate density functions with parameters of multiple types:

Theorem 3.3. (Densities with multiple scalar parameters)

(a) Generalized univariate logistic densities: Let $f(x|\theta,\sigma) := \frac{1}{\sigma} f((x - \theta)/\sigma)$, where $f(x) = \frac{\Gamma(p + q)}{\Gamma(p)\Gamma(q)} \exp(px) (1 + \exp(x))^{p+q}$, 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 $f(x|\theta,\lambda,\sigma) := \frac{1}{\sigma} f((x - \theta)/\sigma,\lambda)$, where $f(x,\lambda) = \frac{\lambda^x}{\Gamma(\lambda)} \exp(-\lambda(x+\exp(-x)))$ as $\lambda > 0$. Then the family $\{f(x|\theta,\lambda,\sigma), \theta \in \mathbb{R}, \sigma \in \mathbb{R}_+\}$ 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 density function classes with matrix-variate parameter spaces, as introduced in Section 2:
Theorem 3.4. (Densities with matrix-variate parameters)

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

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

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

(d) The family \( \{ f(x|\theta, a, b), \theta \in \mathbb{R}^d, \Sigma \in S^+_d, a \in \mathbb{R}^d_+, b \in \mathbb{R}^d_+ \} \) of modified Gaussian-Gamma distribution is not identifiable in the first order.

These theorems are the matrix-variate or multiple type parameters versions of results proven for density classes with a single parameter [4]. As the proofs of these results are technically involved, we present only 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 is established by exploiting how the corresponding characteristics 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 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 family of Gamma distribution is not identifiable in the first order.

Before ending this section, we state results in response to a question posed by Xuming He on strong identifiability in transformed parameter spaces. The following theorem states 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 \( J(\eta, \Lambda) \in \mathbb{R}^{(d_1 + d_2) \times (d_1 + d_2)} \) be the modified Jacobian matrix of \( T(\eta, \Lambda) \), i.e. the usual Jacobian matrix when \( (\eta, \Lambda) \) is taken as a \( d_1 + d_2 \) 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 identifiability (either in the first or second order) yields sharp lower bounds of $V(p_G, p_{G_0})$ in terms of Wasserstein distances $W_1(G, G_0)$. It is useful to know that in the transformed parameter space, one may still enjoy the same inequality. Specifically, for any discrete probability measure $Q = \sum_{i=1}^{k_0} p_i \delta_{(\eta_i, \Lambda_i)} \in E_{k_0}(\Theta^* \times \Omega^*)$, 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 $E_{k_0}(\Theta^* \times \Omega^*)$, while probability measure $Q$ varies in $E_{k_0}(\Theta^* \times \Omega^*)$.

**Corollary 3.3.** Assume that the conditions of Theorem 3.5 hold. Further, suppose that the first derivative of $f$ in terms of $\theta, \Sigma$ and the first derivative of $T$ in terms of $\eta, \Lambda$ are $\alpha$-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 $Q \in E_{k_0}(\Theta^* \times \Omega^*)$ 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 $\Theta$ and $\Omega$ are bounded sets, the condition on the boundedness of the first derivative of $f$ in terms of $\theta, \Sigma$ and the first derivative of $g$ in terms of $\eta, \Lambda$ can be left out. Additionally, the restriction that these derivatives should be $\alpha$-Hölder continuous can be relaxed to only that the first derivative of $f$ and the first derivative of $g$ are $\alpha_1$-Hölder continuous and $\alpha_2$-Hölder continuous where $\alpha_1, \alpha_2 > 0$ can be different. Finally, the conclusion of Corollary 3.3 still holds for lower bound $W_2^2(Q, Q_0)$ if we impose second-order identifiability on kernel density $f$ as well as 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 mixture density $p_{G_0}$, where $G_0$ is unknown true mixing distribution with exactly $k_0$ support points, and 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 of $G_0$ is $\Theta \times \Omega$. In this section we shall assume that $\Theta$ is a compact subset of $\mathbb{R}^{d_1}$ and $\Omega = \left\{ \Sigma \in S_{d_2}^+ : \underline{\Lambda} \leq \sqrt{\lambda_1(\Sigma)} \leq \sqrt{\lambda_d(\Sigma)} \leq \bar{\Lambda} \right\}$, where $0 < \underline{\Lambda}, \bar{\Lambda}$ are known and $d_1, d_2 \geq 1$. We denote $\Theta^* = \Theta \times \Omega$. The maximum likelihood estimator for $G_0$ in the over-fitted mixture setting is given by

$$\hat{G}_n = \arg \max_{G \in \mathcal{O}_k(\Theta \times \Omega)} \frac{1}{n} \sum_{i=1}^{n} \log(p_G(X_i)).$$
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 transport 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 \( W_1 \) 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 \( W_2 \).

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., [23]). For any positive integer number \( k_1 \), define several classes of mixture densities
\[
\mathcal{P}_{k_1}(\Theta^*) = \left\{ p_{G} : G \in \mathcal{O}_{k_1}(\Theta^*) \right\},
\]
\[
\mathcal{P}_{k_1}^{1/2}(\Theta^*) = \left\{ \left( p_{G + G_0} \right)^{1/2} : G \in \mathcal{O}_{k_1}(\Theta^*) \right\}.
\]

For any \( \delta > 0 \), define the intersection of a Hellinger ball centering at \( p_{G_0} \) and put
\[
\mathcal{P}_{k_1}^{1/2}(\Theta^*, \delta) = \left\{ \left( p_{G + G_0} \right)^{1/2} \in \mathcal{P}_{k_1}^{1/2} : h(p_{G + G_0}, p_{G_0}) \leq \delta \right\}.
\]

The size of this set is captured by the entropy integral:
\[
J_B(\delta, \mathcal{P}_{k_1}^{1/2}(\Theta^*, \delta), \mu) = \int_{\delta^2/2^{13}}^{\delta} H_B^{1/2}(u, \mathcal{P}_{k_1}^{1/2}(\Theta^*, u), \mu) du \vee \delta,
\]

where \( H_B \) denotes the bracketing entropy of \( \mathcal{P}_{k_1}^{1/2}(\Theta^*) \) under \( L_2 \) distance (cf. [23] for a definition of the bracket entropy).

We are ready to state a general result of MLE under the exact-fitted mixture setting:

\textbf{Theorem 4.1. (Exact-fitted mixtures)} Assume that \( f \) satisfies the conditions of Theorem 3.1. Take \( \Psi(\delta) \geq J_B(\delta, \mathcal{P}_{k_0}^{1/2}(\Theta^*, \delta), \mu_0) \) in such a way that \( \frac{\Psi(\delta)}{\delta^2} \) is a non-increasing function of \( \delta \). Then for a universal constant \( c \), constant \( C_1 = C_1(\Theta^*) \), \( \{\delta_n\} \) is a non-negative sequence such that
\[
\sqrt{m \delta_n^2} \geq c \Psi(\delta_n),
\]
and for all \( \delta \geq \frac{\delta_n}{\sqrt{C_1}} \), we have
\[
P(W_1(\hat{G}_n, G_0) > \delta) \leq c \exp \left( -\frac{nC_1^2\delta^2}{c^2} \right).
\]

\textbf{Proof.} By Theorem 3.1,
\[
C_1(\Theta^*) W_1^2(G, G_0) \leq V^2(p_G, p_{G_0}) \leq h^2(p_G, p_{G_0}) \text{ for all } G \in \mathcal{E}_{k_0}(\Theta^*),
\]
(5)
where $C_1(\Theta^*)$ is a positive constant depending only on $\Theta^*$ and $G_0$. Now, invoking Theorem 7.4 of [23], as $\delta \geq \delta_n$, there is a universal constant $c > 0$ such that

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

(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 4.2. (Over-fitted mixtures)** Assume that $f$ satisfies the conditions in part (a) of Theorem 3.2. Take $\Psi(\delta) \geq \mathcal{J}_B(\delta, \bar{\mathcal{P}}_k^{1/2}(\Theta^*, \delta), p_{G_0})$ in such a way that $\frac{\Psi(\delta)}{\delta^2}$ is a non-increasing function of $\delta$. Then for a universal constant $c$, constant $C_1 = C_1(\Theta^*)$, \{\delta_n\} is a non-negative sequence such that

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

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

$$P(W_2(\hat{G}_n, G_0) > \delta^{1/2}) \leq c \exp \left( -\frac{nC_2^2 \delta^2}{c^2} \right).$$

To derive concrete rates $\delta_n$, one simply need to verify the condition 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 $\Theta = [-a_n, a_n]^d \times [m, \bar{m}]$ where $a_n \leq L(\log(n))^{\gamma}$ as $L$ is some positive constant, $\gamma > 0$, and $1 < m \leq \bar{m}$ are two known positive numbers. Let $\{f(x|\theta, m, \Sigma) | (\theta, m) \in \Theta, \Sigma \in \Omega\}$ to be the class of multivariate generalized Gaussian distributions.

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

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

**Remark:** (i) The condition $m > 1$ can be relaxed to $m \geq 1$ under exact-fitted setting; however, it is crucial under over-fitted setting that $m > 1$. 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 from 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 second-order identifiability leads to very slow convergence rate of the MLE under the
over-fitted Gaussian mixture setting. This theory is fundamentally different from the theory of strong identifiability, and will be reported elsewhere. (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 MLE rates \((\log n/n)^{1/2}\) and \((\log n/n)^{1/4}\) for the exact-fitted and over-fitted finite mixtures, respectively, are in fact optimal up to a logarithmic term of \(n\). Under exact-fitted finite mixture setting, it is simple to establish the standard \(n^{-1/2}\) minimax lower bound:

\[
\inf_{\hat{G}_n \in \hat{G}_0} \sup_{G_0 \in G_0} E_{P_{G_0}} (W_1(\hat{G}_n, G_0)) \gtrsim n^{-1/2},
\]

where the infimum is taken over all possible estimators \(\hat{G}_n\) (not just the MLE). More surprising is the following minimax lower bound result for the over-fitted mixture setting.

**Theorem 4.3. (Minimax lower bound for over-fitted mixtures)** If the class of densities \(f\) satisfies condition (c) of Theorem 3.2, then for any positive \(r < 4\),

\[
\inf_{\hat{G}_n \in \hat{G}_0} \sup_{G_0 \in G_0 \setminus \hat{G}_0} E_{P_{G_0}} (W_2(\hat{G}_n, G_0)) \gtrsim n^{-1/r}.
\]

**Proof.** For any \(G_0 = \sum_{i=1}^{k_0} \sum_{i=1}^{k_0+1} p_i^{0} \delta_{(\theta_i^{0}, \Sigma_i^{0})}\) having \(k_0 < k\) support points in \(\Theta \times \Omega\), we construct \(G'_0 = \sum_{i=1}^{k_0+1} p_i^{1} \delta_{(\theta_i^{1}, \Sigma_i^{1})}\) such that \(p_i^{1} = p_i^{0}/2, p_i^{1} = p_i^{0}\) as \(3 \leq i \leq k_0+1, \theta_i^{1} = \theta_i^{0} - \epsilon I_{d_1}, \theta_i^{0} + \epsilon I_{d_1}, \theta_i^{1} = \theta_i^{0} - \epsilon I_{d_2}, \Sigma_i^{0} = \Sigma_i^{0} - \epsilon I_{d_2}, \Sigma_i^{1} = \Sigma_i^{0} + \epsilon I_{d_2}\). It is easy to verify that \(W_2(\hat{G}_n, G'_0) \asymp \epsilon\). Using the same argument as that of the proof of Theorem 3.2, part (c), we obtain \(h(p_{G_0}^{n}, p_{G'_0}^{n}) \lesssim \epsilon^r\) for any \(1 \leq r < 2\). Applying Le Cam’s method (cf. [26]), for any given estimator \(\hat{G}_n\), we obtain

\[
\sup_{G \in \{G_0, G'_0\}} E_{P_{G_0}} (W_2(\hat{G}_n, G)) \gtrsim \epsilon \left(1 - \frac{V(p_{G_0}^{n}, p_{G'_0}^{n})}{2}\right),
\]

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

\[
V(p_{G_0}^{n}, p_{G'_0}^{n}) \leq h(p_{G_0}^{n}, p_{G'_0}^{n}) = \sqrt{1 - (1 - h^2(p_{G_0}^{n}, p_{G'_0}^{n}))^n} \leq \sqrt{1 - (1 - C\epsilon^{2r})^n},
\]

where \(C\) is some positive constant independent of \(\epsilon\) and \(n\). As a consequence, we obtain

\[
\sup_{G \in \{G_0, G'_0\}} E_{P_{G_0}} (W_2(\hat{G}_n, G)) \gtrsim \epsilon \left(1 - \frac{1}{2} \sqrt{1 - (1 - C\epsilon^{2r})^n}\right).
\]
By choosing $\epsilon^2 = \frac{1}{C_1 n^{1/2}}$, the right hand side of the above inequality is bounded below by $C_1 \epsilon \simeq n^{-1/2r}$ 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 a certain 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 first-order identifiability but not second-order identifiability condition.

Strong identifiability bounds  Bound $V \gtrsim W_1$ for exact-fitted mixtures and $V \gtrsim W_2^2$ for over-fitted mixtures are illustrated for the class of Student’s t-distributions, 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, \( \Omega = \left\{ \Sigma \in S^+_2 : \sqrt{2} \leq \lambda_1(\Sigma) \leq \sqrt{\lambda_3(\Sigma)} \leq 2 \right\} \). 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 \( \theta_1^0 = (-2,2), \theta_0^0 = (-4,4) \), covariances \( \Sigma_1^0 = \left( \frac{9}{4} , \frac{1}{5} \right) \), \( \Sigma_2^0 = \left( \frac{5}{2} , \frac{2}{5} \right) \), \( m_1^0 = 2, m_2^0 = 3 \) (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 base on sufficiently small Wasserstein distance, we generate \( G \) by adding small random noise \( \epsilon \) to the mixing coefficients and support points of \( G_0 \).

It can be observed that both lower bounds and upper bounds match exactly our theory for strongly identifiable classes of distributions such as t-distribution and 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 exact-fitted setting, but the under the over-fitted setting it can be observed that the lower bound \( h \gtrsim W_2 \) no longer holds. In fact, if the Gaussian mixture is over-fitted by one extra component, it can be shown that \( h \gtrsim W_4^4 \gtrsim W_2^4 \) (see illustrations in the middle and right panel), and that this bound is sharp. This has a drastic consequence on the convergence rate of the mixing measure, which we discuss next.

**Convergence rates of MLE**

First, we generate \( n \)-iid samples from a bivariate location-covariance Gaussian mixture with three components with an arbitrarily fixed choice of \( G_0 \). The true parameters for the mixing measure \( G_0 \) are: \( \theta_1^0 = (0,3), \theta_0^0 = (1, -4), \theta_0^0 = (5, 2) \), \( \Sigma_1^0 = \left( \begin{array}{cc} 4.2824 & 1.7324 \\ 1.7324 & 0.81759 \end{array} \right) \), \( \Sigma_2^0 = \left( \begin{array}{cc} 1.75 & -1.25 \\ -1.25 & 1.75 \end{array} \right) \), \( \Sigma_3^0 = \left( \begin{array}{cc} 1 & 0 \\ 0 & 4 \end{array} \right) \), and \( p_1^0 = 0.3, p_2^0 = 0.4, p_3^0 = 0.3 \). MLE \( \hat{G}_n \) are obtained by the EM algorithm as we assume that the data come from a mixture of \( k \) Gaussians where \( k \geq k_0 = 3 \). See Figure 4 where the Wasserstein distances between \( \hat{G}_n \) 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 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 \( W_4(\hat{G}_n, G_0) \) (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.
Fig 3: Mixture of location-scale Gaussian distributions, which satisfy first-order identifiability but not second-order identifiability condition. Left panel: Exact-fitted setting. Middle and right panels are for over-fitted setting by one extra component. Right panel shows that $h \approx W^2$ no longer holds as $h \to 0$.

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

5. Proofs of key theorems

In this section, we present self-contained proofs for two key technical 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 important insights underlying the theory — they are organized in a sequence of steps to help the reader. The proofs of remaining results are deferred to the Appendix.

5.1. Strong identifiability in exact-fitted mixtures

**Proof of Theorem 3.1** It suffices to show that

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

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 sequence of $G_n = \sum_{i=1}^{k_0} p^n_i \delta(\theta^n_i, \Sigma^n_i) \in \mathcal{E}_{k_0} (\Theta \times \Omega)$ converging to $G_0$ in $W_1$ 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 $i = 1, \ldots, k_0$, $(\theta^n_i, \Sigma^n_i) \to (\theta^0_i, \Sigma^0_i)$. For each pair $(G_n, G_0)$, let $\{q^n_{ij}\}$ denote the corresponding probabilities of the optimal coupling for $(G_n, G_0)$ pair, so we can write:

$$W_1(G_n, G_0) = \sum_{1 \leq i, j \leq k_0} q^n_{ij} (\|\theta^n_i - \theta^0_j\| + \|\Sigma^n_i - \Sigma^0_j\|).$$

Since $(\theta^n_i, \Sigma^n_i) \to (\theta^0_i, \Sigma^0_i)$ and $G_n$ and $G_0$ have the same number of support points, it is an easy observation that for sufficiently large $n$, $q^n_{ii} = \min(p^n_i, p^0_i)$. And so, \[
\sum_{i \neq j} q^n_{ij} = \sum_{i=1}^{k_0} |p^n_i - p^0_i|.
\] Adopting the notations that $\Delta \theta^n_i := \theta^n_i - \theta^0_i$, $\Delta \Sigma^n_i := \Sigma^n_i - \Sigma^0_i$, and $\Delta p^n_i := p^n_i - p^0_i$ for all $1 \leq i \leq k_0$, we have

$$W_1(G_n, G_0) \lesssim \sum_{i=1}^{k_0} p^n_i (\|\Delta \theta^n_i\| + \|\Delta \Sigma^n_i\|) + |\Delta p^n_i| =: d(G_n, G_0).$$

The inequality in the above display is due to $q^n_{ii} \leq p^n_i$, and the observation that $\|\theta^n_i - \theta^0_j\|, \|\Sigma^n_i - \Sigma^0_j\|$ 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:

$$p_{G_n}(x) - p_{G_0}(x) = \sum_{i=1}^{k_0} \Delta p^n_i f(x|\theta^n_i, \Sigma^n_i) + \sum_{i=1}^{k_0} p^n_i (f(x|\theta^n_i, \Sigma^n_i) - f(x|\theta^0_i, \Sigma^0_i)).$$

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

$$\sum_{i=1}^{k_0} p^n_i (f(x|\theta^n_i, \Sigma^n_i) - f(x|\theta^0_i, \Sigma^0_i)) = \sum_{i=1}^{k_0} p^n_i \left\{ (\Delta \theta^n_i)^T \frac{\partial f}{\partial \theta}(x|\theta^0_i, \Sigma^0_i) + \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta^0_i, \Sigma^0_i)^T \Delta \Sigma^n_i \right) \right\} + R_n(x),$$

where $R_n(x) = O \left( \sum_{i=1}^{k_0} p^n_i (\|\Delta \theta^n_i\|^{1+\delta_1} + \|\Delta \Sigma^n_i\|^{1+\delta_2}) \right)$, where the appearance of $\delta_1$ and $\delta_2$ 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 $n \to \infty$. 

N. Ho and X. Nguyen/Optimal rates of parameter estimation 20
Denote $A_n(x) = \sum_{i=1}^{k_0} p_i^n \left[ (\Delta \theta^n_i)^T \frac{\partial f}{\partial \theta}(x|\theta^n_i, \Sigma^n_i) + \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta^n_i, \Sigma^n_i)^T \Delta \Sigma^n_i \right) \right]$ and $B_n(x) = \sum_{i=1}^k \Delta p_i^n f(x|\theta^n_i, \Sigma^n_i)$. 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 the linear combination 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:
\[
\left\{ \sum_{i=1}^{k_0} |\Delta p_i^n| + p_i^n (\|\Delta \theta^n_i\|_1 + \|\Delta \Sigma^n_i\|_1) \right\}/d(G_n, G) \to 0.
\]

Since the entrywise $\ell_1$ and $\ell_2$ norms are equivalent, the above entails $\left\{ \sum_{i=1}^{k_0} |\Delta p_i^n| + p_i^n (\|\Delta \theta^n_i\|_1 + \|\Delta \Sigma^n_i\|_1) \right\}/d(G_n, G_0) \to 0$, 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 $n \to \infty$,
\[
d_n A_n(x)/d(G_n, G_0) \to \sum_{i=1}^{k_0} \beta_i \frac{\partial f}{\partial \theta}(x|\theta^n_i, \Sigma^n_i) + \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta^n_i, \Sigma^n_i)^T \gamma_i \right),
\]
\[
d_n B_n(x)/d(G_n, G_0) \to \sum_{i=1}^{k_0} \alpha_i f(x|\theta^n_i, \Sigma^n_i),
\]

such that not all scalar elements of $\alpha_i$, $\beta_i$ and $\gamma_i$ vanish. Moreover, $\gamma_i$ are symmetric matrices because $\Sigma^n_i$ are symmetric matrices for all $n, i$. Note that
\[
d_n V(p_{G_n}, p_{G_0})/d(G_n, G_0) = \int d_n |p_{G_n}(x) - p_{G_0}(x)|/d(G_n, G_0)
\]
\[= \int d_n |A_n(x) + B_n(x) + R_n(x)|/d(G_n, G_0) \, dx \to 0.
\]

By Fatou’s lemma, the integrand in the above display vanishes for almost all $x$. Thus, for almost all $x$
\[
\sum_{i=1}^{k_0} \alpha_i f(x|\theta^n_i, \Sigma^n_i) + \beta_i \frac{\partial f}{\partial \theta}(x|\theta^n_i, \Sigma^n_i) + \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta^n_i, \Sigma^n_i)^T \gamma_i \right) = 0.
\]

By the first-order identifiability criteria of $f$, we have $\alpha_i = 0$, $\beta_i = 0 \in \mathbb{R}^{d_1}$, and $\gamma_i = 0 \in \mathbb{R}^{d_2 \times d_2}$ for all $i = 1, 2, \ldots, k$, 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

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

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 $\sup_{x \in X} |p_{G_n}(x) - p_{G_0}(x)|/W_2^2(G_n, G_0) \to 0$ 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, $k_0 + 1 \leq k^* \leq k$.

Write $G_n = \sum_{i = 1}^{k^*} p_i^n \delta(\theta_i^n, \Sigma_i^n)$ and $G_0 = \sum_{i = 1}^{k_0} p_i^0 \delta(\theta_i^0, \Sigma_i^0)$. Since $W_2(G_n, G_0) \to 0$, there exists a subsequence of $G_n$, such that each support point $(\theta_i^n, \Sigma_i^n)$ of $G_0$ is the limit of a subset of $s_k \geq 1$ support points of $G_n$. There may also be a subset of support points of $G_n$ whose limits are not among the support points of $G_0$ — we assume there are $m \geq 0$ such limit points. To avoid notational cluttering, we replace the subsequence of $G_n$ by the whole sequence $\{G_n\}$. By re-labeling the support points, $G_n$ can be expressed by

$$
G_n = \sum_{i = 1}^{k_0 + m} \sum_{j = 1}^{s_i} p^n_{ij} \delta(\theta^n_{ij}, \Sigma^n_{ij}) \xrightarrow{W_2} G_0 = \sum_{i = 1}^{k_0 + m} p_i^0 \delta(\theta_i^0, \Sigma_i^0)
$$

where $(\theta^n_{ij}, \Sigma^n_{ij}) \to (\theta_i^0, \Sigma_i^0)$ 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 $\Sigma$ is (strictly) positive definite whose maximum eigenvalue is bounded (from above) by constant $M$, then $\Sigma$ is also bounded under the entrywise $\ell_2$ norm. However if $\Sigma$ is only positive semidefinite, it can be singular and its $\ell_2$ norm potentially unbounded. In our context, for $i \geq k_0 + 1$ it is possible that the limiting matrices $\Sigma_i^n$ can be singular. It comes from the fact that the same eigenvalues of $\Sigma_i^n$ can go to 0 as $n \to \infty$, which implies $\det(\Sigma_i^n) \to 0$ and hence $\det(\Sigma_i^n) = 0$. By re-labeling the support points, we may assume without loss of generality that $\Sigma_i^n \to 0$ are (strictly) positive definite matrices and $\Sigma_{k_0+1}^0, \ldots, \Sigma_{k_0+m}^0$ 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 $\lim_{\lambda_1(\Sigma) \to 0} f(x|\theta, \Sigma) = 0$: accordingly, for each $x$,

$$
f(x|\theta^n_{ij}, \Sigma^n_{ij}) \to 0 as n \to \infty for all k_0 + m_1 + 1 \leq i \leq k_0 + m, 1 \leq j \leq s_i.
$$

**Step 2** Using shorthand notations $\Delta\theta^n_{ij} := \theta^n_{ij} - \theta_i^0$, $\Delta\Sigma^n_{ij} := \Sigma^n_{ij} - \Sigma_i^0$ for $i = 1, \ldots, k_0 + m_1$ and $j = 1, \ldots, s_i$, it is simple to see that

$$
W_2^2(G_n, G_0) \lesssim d(G_n, G_0) := \sum_{i = 1}^{k_0 + m_1} \sum_{j = 1}^{s_i} p_{ij}^n \left| \Delta\theta^n_{ij} \right|^2 + \left| \Delta\Sigma^n_{ij} \right|^2 + \sum_{i = 1}^{k_0 + m} \left| p_i^n - p_i^0 \right|.
$$
because $W^2_n(G_n, G_0)$ 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 $(\theta^n_{ij}, \Sigma^n_{ij})$ for $j = 1, \ldots, s_i$ with $(\theta^0, \Sigma^0)$ by mass $\min(p^n_i, p^0)$, while the remaining masses are coupled arbitrarily. From the assumption, $\sup_{x \in \mathcal{X}} |p_{G_n}(x) - p_{G_0}(x)| / W^2_n(G_n, G_0)$ 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:

$$p_{G_n}(x) - p_{G_0}(x) = \sum_{i=1}^{k_0 + m_1} \sum_{j=1}^{s_i} p^n_{ij} (f(x|\theta^n_{ij}, \Sigma^n_{ij}) - f(x|\theta^0_i, \Sigma^0_i))$$

$$+ \sum_{i=1}^{k_0 + m_1} (p^n_{ij} - p^0_{ij}) f(x|\theta^0_i, \Sigma^0_i) + \sum_{i=k_0 + m_1 + 1}^{k_0 + m} s_i p^n_{ij} f(x|\theta^n_{ij}, \Sigma^n_{ij}) := A_n(x) + B_n(x) + C_n(x).$$

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

$$A_n(x) = \sum_{i=1}^{k_0 + m_1} \sum_{j=1}^{s_i} p^n_{ij} (f(x|\theta^n_{ij}, \Sigma^n_{ij}) - f(x|\theta^0_i, \Sigma^0_i)) = \sum_{i=1}^{k_0 + m_1} \sum_{\alpha} A^n_{\alpha, \alpha}(\theta^0_i, \Sigma^0_i) + R_n(x),$$

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

$$A^n_{1,0}(\theta^0_i, \Sigma^0_i) = \sum_{j=1}^{s_i} p^n_{ij} (\Delta \theta^n_{ij})^T \frac{\partial f}{\partial \theta}(x|\theta^0_i, \Sigma^0_i),$$

$$A^n_{0,1}(\theta^0_i, \Sigma^0_i) = \sum_{j=1}^{s_i} p^n_{ij} \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta^0_i, \Sigma^0_i)^T \Delta \Sigma_{ij} \right),$$

$$A^n_{2,0}(\theta^0_i, \Sigma^0_i) = \sum_{j=1}^{s_i} p^n_{ij} \text{tr} \left( \frac{\partial^2 f}{\partial \theta^2}(x|\theta^0_i, \Sigma^0_i) \Delta \theta^n_{ij} \right),$$

$$A^n_{0,2}(\theta^0_i, \Sigma^0_i) = \sum_{j=1}^{s_i} p^n_{ij} \text{tr} \left( \frac{\partial}{\partial \Sigma}(x|\theta^0_i, \Sigma^0_i)^T \Delta \Sigma^n_{ij} \right),$$

$$A^n_{1,1}(\theta^0_i, \Sigma^0_i) = \sum_{j=1}^{s_i} \Delta \theta^n_{ij} \left( \frac{\partial}{\partial \theta} \left( \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta^0_i, \Sigma^0_i)^T \Delta \Sigma^n_{ij} \right) \right) \right).$$

In addition, $R_n(x) = O \left( \sum_{i=1}^{k_0 + m_1} \sum_{j=1}^{s_i} p^n_{ij} (\|\Delta \theta^n_{ij}\|^2 + \|\Delta \Sigma^n_{ij}\|^2) \right)$ 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 $\theta$ and $\Sigma$ up to the second order, and evaluated at the distinct pairs $(\theta_i, \Sigma_i)$ for $i = 1, \ldots, k_0 + m_1$. (To be specific, the elements of $f(x|\theta, \Sigma)$, $\partial f/\partial \theta(x|\theta, \Sigma)$, $\partial^2 f/\partial \Sigma(x|\theta, \Sigma)$, $\partial^2 f/\partial \theta^2(x|\theta, \Sigma)$, $\partial^2 f/\partial \theta \partial \Sigma(x|\theta, \Sigma)$). 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 $\partial f/\partial \theta_i$ as $1 \leq i \leq d_1$ and $\partial^2 f/\partial \Sigma_{uv}$ for $1 \leq u, v \leq d_2$, we obtain

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

Therefore, $\sum_{i=1}^{k_0 + m_1} |p_{ij}^n - p_{ij}^0|/D_n \to 1$ as $n \to \infty$. It implies that we should have at least one coefficient associated with a $f(x|\theta)$ (appearing in $B_n(x)/D_n$) does not converge 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

$$
d_n B_n(x)/D_n \to \sum_{i=1}^{k_0 + m_1} \alpha_i f(x|\theta_i^0, \Sigma_i^0),
$$

$$
d_n \sum_{i=1}^{k_0 + m_1} A_{i,0}^n(\theta_i^0, \Sigma_i^0)/D_n \to \sum_{i=1}^{k_0 + m_1} \beta_i^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0),
$$

$$
d_n \sum_{i=1}^{k_0 + m_1} A_{i,1}^n(\theta_i^0, \Sigma_i^0)/D_n \to \sum_{i=1}^{k_0 + m_1} \sum_{j=1}^{s_i} \nu_{ij}^T \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^0, \Sigma_i^0),
$$

$$
d_n \sum_{i=1}^{k_0 + m_1} A_{i,2}^n(\theta_i^0, \Sigma_i^0)/D_n \to \sum_{i=1}^{k_0 + m_1} \sum_{j=1}^{s_i} \nu_{ij}^T \frac{\partial^2 f}{\partial \theta \partial \Sigma}(x|\theta_i^0, \Sigma_i^0),
$$

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 $1 \leq i \leq k_0 + m_1, 1 \leq j \leq s_i$. Additionally, $d_n C_n(x)/D_n = \ldots$
\[ D_n^{-1} \sum_{i=k_0, m+1}^{k_0+m} \sum_{j=1}^{s_i} d_n p_{ij} f(x|\theta_{ij}^0, \Sigma_{ij}^0) \to 0 \] due to the fact that \( f(x|\theta_{ij}^0, \Sigma_{ij}^0) \to 0 \) for all \( k_0 + m + 1 \leq i \leq k_0 + m, 1 \leq j \leq s_i \). As a consequence, we obtain for all \( x \) that

\[
\sum_{i=1}^{k_0+m_1} \left\{ \alpha_i f(x|\theta_i^0, \Sigma_i^0) + \beta_i^T \frac{\partial f}{\partial \theta}(x|\theta_i^0, \Sigma_i^0) + \sum_{j=1}^{s_i} \nu_{ij}^T \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i^0, \Sigma_i^0) \nu_{ij} + \right.
\]

\[
\text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0) T \gamma_i \right) + 2 \sum_{j=1}^{s_i} \nu_{ij}^T \left( \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0) T \eta_j \right) \right) \left] + \right.
\]

\[
\sum_{j=1}^{s_i} \text{tr} \left( \frac{\partial f}{\partial \Sigma}(x|\theta_i^0, \Sigma_i^0) T \eta_j \right) \right\} = 0.
\]

From the second-order identifiability of \( \{ f(x|\theta, \Sigma), \theta \in \Theta, \Sigma \in \Omega \} \), we obtain \( \alpha_i = 0, \beta_i = \nu_{i1} = \ldots = \nu_{is_i} = 0 \in \mathbb{R}^{d_1}, \gamma_i = \eta_{i1} = \ldots = \eta_{is_i} = 0 \in \mathbb{R}^{d_2 \times d_2} \) for all \( 1 \leq i \leq k_0 + m_1 \), which is a contradiction to the fact that not all coefficients go to 0 as \( n \to \infty \). 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 \( \theta_i^n = \theta_i^0 - \frac{1}{n} \mathbf{1}_{d_1}, \theta_i^n = \theta_i^0 - \frac{1}{n} \mathbf{1}_{d_1}, \Sigma_i^n = \Sigma_i^0 - \frac{1}{n} I_{d_2} \) and \( \Sigma_i^n = \Sigma_i^0 + \frac{1}{n} I_{d_2} \). Here, \( I_{d_2} \) denotes identity matrix in \( \mathbb{R}^{d_1 \times d_2} \) and \( \mathbf{1}_n \) a vector with all elements being equal to 1.

In addition, \( (\theta_{i+1}^n, \Sigma_{i+1}^n) = (\theta_i^0, \Sigma_i^0) \) for all \( i = 2, \ldots, k_0 \). Also, \( p_i^n = p_i^0 = \frac{p_i^0}{2} \) and \( p_{i+1}^n = p_{i}^0 \) for all \( i = 2, \ldots, k_0 \). It is simple to verify that \( E_n := W_n^r(G_n, G_0) = \frac{(p_0^r)^r}{2r} (\| \theta_1^n - \theta_1^0 \|^r + \| \theta_2^n - \theta_2^0 \|^r + \| \Sigma_1^n - \Sigma_1^0 \|^r + \| \Sigma_2^n - \Sigma_2^0 \|^r)^r \frac{1}{n^r} \to 1 \).

By means of Taylor’s expansion up to the first order, we get that as \( n \to \infty \)

\[
V(p_{G_n}, p_{G_0}) \leq \int_{x \in X} \left| R_1(x) \right| dx,
\]

where \( \alpha_1 \in \mathbb{N}^{d_1}, \alpha_2 \in \mathbb{N}^{d_1} \times \mathbb{N}^{d_2} \times d_2 \) 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 \( \sum_{i=1}^{2} (\Delta \theta_i^n)^{\alpha_1} (\Delta \Sigma_i^n)^{\alpha_2} = 0 \) for each \( \alpha_1 \), \( \alpha_2 \) such that \( |\alpha_1| + |\alpha_2| = 1 \). Since \( f \) is
second-order differentiable with respect to \( \theta, \Sigma, R_1(x) \) takes the form

\[
R_1(x) = \sum_{i=1}^{2} \sum_{|\alpha| = 2} \frac{2}{\alpha!} (\Delta \theta_i^n)^{\alpha_1} (\Delta \Sigma_i^n)^{\alpha_2} \times \\
\times \int_0^1 (1-t) \frac{\partial^2 f}{\partial \theta^{\alpha_1} \partial \Sigma^{\alpha_2}}(x|\theta_0^n + t \Delta \theta_i^n, \Sigma_0^n + t \Delta \Sigma_i^n) dt,
\]

where \( \alpha = (\alpha_1, \alpha_2) \). Note that, \( \sum_{i=1}^{2} |\Delta \Sigma_i^n|^{\alpha_2} = O(n^{-2}) \). Additionally, from the hypothesis,

\[
\sup_{t \in [0,1]} \int_{x \in X} \frac{\partial^2 f}{\partial \theta^{\alpha_1} \partial \Sigma^{\alpha_2}}(x|\theta_0^n + t \Delta \theta_i^n, \Sigma_0^n + t \Delta \Sigma_i^n) \ dx < \infty.
\]

It follows that \( \int |R_1(x)| \ dx = O(n^{-2}) \). So for any \( r < 2 \), \( V(p_{G_n}, p_{G_0}) = o(W_r^n(G_n, G_0)) \).

This concludes the proof.

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

\[
h^2(p_{G_n}, p_{G_0}) \leq \frac{1}{2p_0} \int_{x \in X} \frac{(p_{G_n}(x) - p_{G_0}(x))^2}{f(x|\theta_0^n, \Sigma_0^n)} \ dx \leq \int_{x \in X} \frac{R_1^n(x)}{f(x|\theta_0^n, \Sigma_0^n)} \ dx.
\]

where first inequality is due to \( \sqrt{p_{G_n}(x)} + \sqrt{p_{G_0}(x)} > \sqrt{p_{G_0}(x)} > \sqrt{p_{G_0}(x)} > \sqrt{p_{G_0}(x)} \) 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).

References


**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 param-
eters 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 proofs of Theorem 3.3, which concerns the characterization of strong identifiability in multiple scalar parameters, are somewhat similar to those of Theorem 3.4 and therefore are 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 triangle inequalities. Proof of Corollary 4.1 relies on calculations of the bracket 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 \( \inf_{G \in G'} V(p_{G}, p_{G_0})/W_1(G, G_0) > 0 \). Assume the contrary, then we can find a sequence of \( G_n \in G' \) and \( W_1(G_n, G_0) > \epsilon \) such that \( V(p_{G_n}, p_{G_0})/W_1(G_n, G_0) \to 0 \) as \( n \to \infty \). Since \( G \) 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 \( G' \neq G_0 \), 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, \( V(p_{G_n}, p_{G'}) \leq C(\Theta, \Omega)W_1(G_n, G') \), 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 \( G' = G_0 \), which is a contradiction. This completes the proof.

6.2. The importance of boundedness conditions under over-fitted setting

PROOF OF PROPOSITION 3.1 We choose \( G_n = \sum_{i=1}^{k_0+1} \exp(n/\alpha) I_{\theta_i}(\theta_i, \Sigma_n^i) \in \mathcal{O}_k(\Theta \times \Omega) \) such that \( (\theta_i^n, \Sigma_i^n) = (\theta_i^0, \Sigma_i^0) \) for \( i = 1, \ldots, k_0 \), \( \theta_{k_0+1}^n = \theta_1^0 \), \( \Sigma_{k_0+1}^n = \Sigma_1^0 + \exp(n/\alpha) I_{\theta_1}(\theta_1, \Sigma_1) \), where \( \alpha = \frac{1}{2\beta} \). Additionally, \( p_i^n = p_i^0 - \exp(-n) \), \( p_i^n = p_i^0 \) for all \( 2 \leq i \leq k_0 \), and \( p_i^0 = \exp(-n) \). With this construction, we can check that \( W_1^\beta(G, G_0) = d_2^{\beta/2}/\sqrt{n} \). Now, as \( h^2(p_{G_n}, p_{G_0}) \lesssim V(p_{G_n}, p_{G_0}) \), we have

\[
\exp \left( \frac{2}{W_1^\beta(G_n, G_0)} \right) h^2(p_{G}, p_{G_0}) \lesssim \exp \left( -n + \frac{2\sqrt{n}}{d_2^{\beta/2}} \right) \times 
\int_{x \in \mathcal{X}} |f(x|\theta_1^0, \Sigma_i^0) - f(x|\theta_1^n, \Sigma_i^n)|dx,
\]
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**  We only present the proof for part (a) and part (b). The proofs for part (c) and (d) are somewhat similar and is omitted.

(a) Assume that for given \( k \geq 1 \) and \( k \) different tuples \((\theta_1, \Sigma_1, m_1), \ldots, (\theta_k, \Sigma_k, m_k)\), 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:

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

Substituting the first derivatives of \( f \) to get

\[
\sum_{j=1}^{k} \left( \alpha_j \left( (\beta_j^T(x - \theta_j) + (x - \theta_j)^T \gamma_j (x - \theta_j) \right) \right) \times \left[ (x - \theta_j)^T \Sigma_j^{-1}(x - \theta_j) \right]^{m_j-1} + \eta_j \log \left( (x - \theta_j)^T \Sigma_j^{-1}(x - \theta_j) \right) \right\} \times \exp \left( - \left[ (x - \theta_j)^T \Sigma_j^{-1}(x - \theta_j) \right]^{m_j} \right) = 0, \tag{10}
\]

where

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

\[
\beta_j' = \frac{2m_j^2 \Gamma(d/2) - \pi^{d/2} \Gamma(d/(2m_j)) |\Sigma_j|^{1/2} \Sigma_j^{-1} \beta_j \gamma_j + \frac{m_j^2 \Gamma(d/2) - \pi^{d/2} \Gamma(d/(2m_j)) |\Sigma_j|^{1/2} \Sigma_j^{-1} \gamma_j \Sigma_j^{-1}}{m_j \eta_j \Gamma(d/2)}}{\pi^{d/2} \Gamma(d/(2m_j)) |\Sigma_j|^{1/2}}, \quad \gamma_j' = \frac{m_j^2 \Gamma(d/2) - \pi^{d/2} \Gamma(d/(2m_j)) |\Sigma_j|^{1/2} \Sigma_j^{-1} \gamma_j \Sigma_j^{-1}}{m_j \eta_j \Gamma(d/2)}.
\]

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

\[
a_i = (x')^T \gamma_i' x', \quad b_i = \left[ \left( (\beta_i')^T - 2\theta_i^T \gamma_i' \right) x' \right], \quad c_i = \theta_i^T \gamma_i' \theta_i - (\beta_i')^T \theta_i,
\]

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

Borrowing a technique from [25], since \((\theta_1, \Sigma_1), \ldots, (\theta_{\tilde{i}}, \Sigma_{\tilde{i}})\) are distinct, we have two possibilities:
Possibility 1  If \( \Sigma_j \) are the same for all \( 1 \leq j \leq \tilde{\ell} \), then \( \theta_1, \ldots, \theta_\tilde{\ell} \) are distinct. For any \( i < j \), denote \( \Delta_{ij} = \theta_i - \theta_j \). Note that if \( x' \notin \bigcup_{1 \leq i < j \leq \tilde{\ell}} \{ u \in \mathbb{R}^d : u^T \Delta_{ij} = 0 \} \), which is a finite union of hyperplanes, then \( (x')^T \theta_1, \ldots, (x')^T \theta_\tilde{\ell} \) are distinct. Hence, if we choose \( x' \in \mathbb{R}^d \) outside this union of hyperplanes, we have \( (x')^T \theta_1, (x')^T \Sigma_1 x' \), \( \ldots, (x')^T \theta_\tilde{\ell}, (x')^T \Sigma_\tilde{\ell} x' \) are distinct.

Possibility 2  If \( \Sigma_j \) are not the same for all \( 1 \leq j \leq \tilde{\ell} \), then we assume without loss of generality that \( \Sigma_1, \ldots, \Sigma_m \) are the only distinct matrices from \( \Sigma_1, \ldots, \Sigma_\tilde{\ell} \), where \( m \leq \tilde{\ell} \). Denote \( \delta_{ij} = \Sigma_i - \Sigma_j \) as \( 1 \leq i < j \leq m \), then as \( x' \) does not belong to \( \bigcup_{1 \leq i < j \leq m} \{ u \in \mathbb{R}^d : u^T \delta_{ij} u = 0 \} \), we have \( (x')^T \Sigma_1 x', \ldots, (x')^T \Sigma_m x' \) are distinct. Therefore, if \( x' \) does not belong to \( \bigcup_{1 \leq i < j \leq m} \{ u \in \mathbb{R}^d : u^T \delta_{ij} u = 0 \} \), which is finite union of conics, then we have \( (x')^T \theta_1, (x')^T \Sigma_1 x' \), \( \ldots, (x')^T \theta_\tilde{\ell}, (x')^T \Sigma_\tilde{\ell} x' \) are distinct. Additionally, for any \( \theta_j \) where \( m + 1 \leq j \leq \tilde{\ell} \) that shares the same \( \Sigma_i \) 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 \( x' \) outside a finite union of conics and hyperplanes, \( (x')^T \theta_1, (x')^T \Sigma_1 x' \), \( \ldots, (x')^T \theta_\tilde{\ell}, (x')^T \Sigma_\tilde{\ell} x' \) are all different.

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

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

\[
A_i(x_1) = \alpha_i' + (a_i x_1^2 + b_i x_1 + c_i)(d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2} - 1} + \eta_i' \log(d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2}).
\]

Multiply both sides of (10) with \( \exp - (d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} \), we get

\[
A_{i_2}(x_1) + \sum_{j \neq i_2} A_j(x_1) \exp \left[ (d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} \right] - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} = 0. \tag{11}
\]

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

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

This implies that when \( x_1 \to \infty \),

\[
B_1(x_1) := \sum_{j \neq i \setminus \{ i_2 \}} A_j(x_1) \exp \left[ (d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} \right] - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} \to 0.
\]
On the other hand, if \( j \notin J \) and \( 1 \leq j \leq 7 \), then \( d_j > d_{i_2} \) and \( m_{i_2} = m_j \). So,
\[
(d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} \leq -x_1^{2m_j} \text{ as } x_1 \text{ is large enough.}
\]
This implies that when \( x_1 \to \infty \),
\[
B_2(x_1) := \sum_{j \notin J, 1 \leq j \leq 7} A_j(x_1) \exp \left[ (d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} \right] \to 0.
\]
Or else, if \( j > 7 \), then \( m_j > m_{i_2} \). So, \( (d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} \leq -x_1^{2m_j} \). As a result,
\[
B_3(x_1) := \sum_{j > 7} A_j(x_1) \exp \left[ (d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} \right] \to 0.
\]
Now, by letting \( x_1 \to \infty \),
\[
\sum_{j \neq i_2} A_j(x_1) \exp \left[ (d_{i_2} x_1^2 + e_{i_2} x_1 + f_{i_2})^{m_{i_2}} - (d_j x_1^2 + e_j x_1 + f_j)^{m_j} \right] = A_1(x) + A_2(x) + A_3(x) \to 0. \tag{12}
\]
Combing (11) and (12), we obtain that as \( x_1 \to \infty \), \( A_{i_2}(x_1) \to 0 \). The only possibility for this result to happen is \( a_{i_2} = b_{i_2} = \gamma_{i_2} = 0 \). Or, equivalently, \((x')^T \gamma_{i_2}' x' = [(\beta_{i_2}')^T - 2 \beta_{i_2}' \gamma_{i_2}'] x' = 0\). If \( \gamma_{i_2}' \neq 0 \), we can choose the element \( x' \notin D \) lying outside the hyperplane \( \{u \in \mathbb{R}^d : u^T \gamma_{i_2}' u = 0\} \). It means that \((x')^T \gamma_{i_2}' x' \neq 0\), which is a contradiction. Therefore, \( \gamma_{i_2}' = 0 \). It implies that \((\beta_{i_2}')^T x' = 0\). If \( \beta_{i_2}' \neq 0 \), we can choose \( x' \notin D \) such that \((\beta_{i_2}')^T x' \neq 0\). Hence, \( \beta_{i_2}' = 0 \). With these results, \( \alpha_{i_2}' = 0 \). Overall, we obtain \( \alpha_{i_2}' = \beta_{i_2}' = \gamma_{i_2}' = \eta_{i_2}' = 0 \). Repeating the same argument to the remained parameters \( \alpha_j, \beta_j, \gamma_j, \eta_j \) and we get \( \alpha_j = \beta_j = \gamma_j = \eta_j = 0 \) for \( 1 \leq j \leq k \). 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:
\[
\sum_{j=1}^k \alpha_j f(x|\theta_j, \Sigma_j) + \beta_j^T \frac{\partial f}{\partial \theta}(x|\theta_j, \Sigma_j) + \text{tr}(\frac{\partial f}{\partial \Sigma}(x|\theta_j, \Sigma_j)^T \gamma_j) = 0. \tag{13}
\]
Multiply both sides with $\exp(it^T x)$ and take the integral in $\mathbb{R}^d$, after direct calculations, the above equation can be rewritten as

$$
\sum_{j=1}^{k} \left[ \int_{\mathbb{R}^d} \left( \alpha_j' \exp(i(\Sigma_j^{-1/2} t)^T x) \right) + \frac{\exp(i(\Sigma_j^{1/2} t)^T x)(\beta_j')^T x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} \cdot \exp(i(\Sigma_j^{1/2} t)^T x)x^T M_j x \right] \exp(it^T \theta_j) = 0, \quad (14)
$$

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

To simplify the left hand side of equation (14), 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 $C = \int_{\mathbb{R}^d} \frac{\exp(it^T x)x^TM x}{(\nu + \|x\|^2)^{(\nu+d+2)/2}} dx$, where $\beta' \in \mathbb{R}^d$ and $M = (M_{ij}) \in \mathbb{R}^{d \times d}$.

In fact, by using orthogonal transformation $x = O \cdot z$, where $O \in \mathbb{R}^{d \times d}$ and its first column to be $(t_1, ..., t_d)^T$, we can verify that $\exp(it^T x) = \exp(i\|t\|_1 z_1)$, $\|x\|^2 = \|z\|^2$, and $dx = \left| \text{det}(O) \right| dz = dz$, then we obtain the following results:

$$
A = \int_{\mathbb{R}^d} \frac{\exp(i\|t\|_1 z_1)}{(\nu + \|z\|^2)^{(\nu+d)/2}} dz = \int_{\mathbb{R}} \exp(i\|t\|_1 z_1) \prod_{j=2}^{d} \frac{1}{(\nu + \|z\|^2)^{(\nu+d)/2}} dz_4 dz_3 \cdots dz_1 = C_1 A_1(\|t\|_1),
$$

where $C_1 = \prod_{j=2}^{d} \int_{\mathbb{R}} \frac{1}{(1 + z^2)^{(\nu+1)/2}} dz$ and $A_1(t') = \int_{\mathbb{R}} \frac{\exp(i\|t'\| z_1)}{(\nu + z^2)^{(\nu+1)/2}} dz$ for any $t' \in \mathbb{R}$. Hence, for all $1 \leq j \leq k$

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

Turning to $B$ and $C$, with same line of calculations, we obtain

$$
B = \left( \sum_{j=1}^{d} \frac{O_j \beta_j'}{\nu + \|z\|^2} \right) \int_{\mathbb{R}^d} \frac{\exp(it^T z_1)}{(\nu + \|z\|^2)^{(\nu+d)/2}} dz = \left( \sum_{j=1}^{d} \frac{O_j \beta_j'}{\nu + \|z\|^2} \right) C_2 A_2(\|t\|_1) = \frac{C_2(\beta')^T A_2(\|t\|)}{\|t\|}.
$$
A simply means the multivariate generalized Gaussian distribution, we can find $D$ are pairwise distinct. By denoting $j \sum_{l=1}^{d} M_{jl} t_j t_l$, we can rewrite (16) as:

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

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

$$
\int_{\mathbb{R}^d} \frac{\exp(\nu ||\Sigma_j^{1/2} x||^2)}{(1 + ||\Sigma_j^{1/2} x||^2)^{(\nu+2d)/2}} dx = \frac{1}{||\Sigma_j^{1/2} t||^2} (\sum_{u,v} M_{uv}^{j} [\Sigma_j^{1/2}]_u [\Sigma_j^{1/2}]_v) \times
$$

$$
\times (C_2 A_3(||\Sigma_j^{1/2} t||) - C_3 A_1(||\Sigma_j^{1/2} t||)) + C_3(\sum_{l=1}^{d} M_{ll}^{j}) A_1(||\Sigma_j^{1/2} t||),
$$

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

As a consequence, by combining (15),(16), and (17), we can rewrite (14) as:

$$
\sum_{j=1}^{k} \left[ \alpha_j^{j} A_1(||\Sigma_j^{1/2} t||) + C_2 \frac{||\Sigma_j^{1/2} t||^2}{||\Sigma_j^{1/2} t||^2} \beta_j^{j} A_2(||\Sigma_j^{1/2} t||) + C_3(\sum_{l=1}^{d} M_{ll}^{j}) A_1(||\Sigma_j^{1/2} t||) + \right.
\left. \left( \sum_{u,v} M_{uv}^{j} [\Sigma_j^{1/2}]_u [\Sigma_j^{1/2}]_v \right) \left( C_2 A_3(||\Sigma_j^{1/2} t||) - C_3 A_1(||\Sigma_j^{1/2} t||) \right) \right] \exp(it^T \theta_j) = 0.
$$

Define $t = t_1 t'$, where $t_1 \in \mathbb{R}$ and $t' \in \mathbb{R}^d$. By using the same argument as that of multivariate generalized Gaussian distribution, we can find $D$ to be the finite union of conics and hyperplanes such that $t' \notin D, ((t')^T \theta_1, (t')^T \Sigma_1 t'), \ldots, ((t')^T \theta_k, (t')^T \Sigma_k t')$ 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:
\[
\sum_{j=1}^{k} \left[ \alpha_j' A_j(\sigma_j|t_1|) + C_2 \frac{t_1 (\Sigma_j^{1/2} t')^T \beta_j'}{|t_1| \sigma_j} A_2(\sigma_j|t_1|) + C_3 \left( \sum_{l=1}^{d} M_{l l}^j \right) A_1(\sigma_j|t_1|) \right] + \\
\left( \sum_{u,v} M_{u v}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} \right) \left( C_2 A_3(\sigma_j|t_1|) - C_3 A_1(\sigma_j|t_1|) \right) \exp(i \theta_j' t_1) = 0.
\]

Since \( A_2(\sigma_j|t_1|) = (i|t_1|) A_1(\sigma_j|t_1|) \), the above equation can be rewritten as:
\[
\sum_{j=1}^{k} \left[ \left( \alpha_j' + C_3 \left( \sum_{l=1}^{d} M_{l l}^j \right) \right) - C_3 \left( \sum_{u,v} M_{u v}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} \right) \right] \times \\
A_1(\sigma_j|t_1|) + C_2 \left( \sum_{u,v} M_{u v}^j \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} \right) A_3(\sigma_j|t_1|) + \\
C_2 (i|t_1|) \frac{[\Sigma_j^{1/2} t']_u [\Sigma_j^{1/2} t']_v}{\sigma_j^2} A_1(\sigma_j|t_1|) \exp(i \theta_j' t_1) = 0. \tag{18}
\]

As \( \nu \) is odd number, we assume \( \nu = 2l - 1 \). By using classical result in complex analysis, we obtain for any \( m \in \mathbb{N} \) that
\[
\int_{-\infty}^{+\infty} \frac{\exp(i|t_1|z)}{(z^2 + \nu^2)^m} \, dz = \frac{2\pi \exp(-|t_1|\sqrt{2l-1})}{(2\sqrt{2l-1})^{2m-1}} \left[ \sum_{j=1}^{m} \binom{2m-1-j}{m-j} (2|t_1|\sqrt{2l-1})^{j-1} \right].
\]

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 \( C_4 = \frac{2\pi}{(2\sqrt{2l-1})^{2l-1}} \), \( a_u = \binom{2l-u-2}{l-u-1} (2\sqrt{2l-1})^u u! \).

Simultaneously, as \( A_3(t_1) = A_1(t_1) - \nu \int_{\mathbb{R}} \frac{\exp(i|t_1|z)}{(\nu^2 + z^2)^{(\nu+3)/2}} \, dz \), 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 \( b_u = \left[ \binom{2l-u-2}{l-u-1} - \frac{1}{4} \binom{2l-u}{l-u} \right] (2\sqrt{2l-1})^u \) as \( 0 \leq u \leq l-1 \), and \( b_l = -\frac{1}{4} \binom{2l}{l} (2\sqrt{2l-1})^l \). 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 (18) can be rewritten as:
\[
\sum_{j=1}^{k} \left[ \alpha_j'' + \beta_j'' (i|t_1|) \sum_{u=0}^{l-1} a_u \sigma_j^u |t_1|^u + \gamma_j'' \sum_{u=0}^{l} b_u \sigma_j^u |t_1|^u \right] \times \\
\exp \left( i \theta_j'' - \sigma_j \sqrt{2l-1}|t_1| \right) = 0,
\]
where \( \alpha_j'' = \alpha_j' + C_3 \sum_{l=1}^{d} M_{ll}^{j} - C_3 (\sum_{u,v} M_{uv}^{j} \frac{[\Sigma_{j}^{1/2} t']_{u} [\Sigma_{j}^{1/2} t']_{v}}{\sigma_j^2}) \), \( \beta_j'' = C_2 \frac{(\Sigma_{j}^{1/2} t')^T \beta_j'}{\sigma_j} \), 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 \( t_1 \geq 0 \)

\[
\sum_{j=1}^{k} \left[ (\alpha_j'' + \beta_j'' (it_1)) \sum_{u=0}^{l-1} a_u \sigma_j'' t_1^u + \gamma_j'' \sum_{u=0}^{l} b_u \sigma_j'' t_1^u \right] \times \\
\exp \left( it_1 \theta_j'' - \sigma_j \sqrt{2l-1} t_1 \right) = 0. \tag{19}
\]

Using the Laplace transformation on both sides of (19) and denoting \( c_j = \sigma_j \sqrt{2l-1} - i \theta_j'' \) as \( 1 \leq j \leq k \), we obtain that as \( \text{Re}(s) > \max_{1 \leq j \leq k} \left\{ -\sigma_j \sqrt{2l-1} \right\} \)

\[
\sum_{j=1}^{k} \alpha_j'' \sum_{u=0}^{l-1} \frac{u! a_u \sigma_j''}{(s + c_j)^{u+1}} + i \beta_j'' \sum_{u=1}^{l} \frac{u! b_u \sigma_j''}{(s + c_j)^{u+1}} + \gamma_j'' \sum_{u=0}^{l} \frac{u! b_u \sigma_j''}{(s + c_j)^{u+1}} = 0. \tag{20}
\]

Without loss of generality, we assume that \( \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_k \). It demonstrates that \(-\sigma_1 \sqrt{2l-1} = \max_{1 \leq j \leq k} \{ -\sigma_j \sqrt{2l-1} \} \). Denote \( a_u'' = a_u \sigma_j'' \) and \( b_u'' = b_u \sigma_j'' \) for all \( u \). By multiplying both sides of (20) with \((s + c_1)^{l+1}\), as \( \text{Re}(s) > -\sigma_1 \sqrt{2l-1} \) and \( s \to -c_1 \), we obtain \( |i \beta_j'' (l+1) a_{l-1}'' + \gamma_j'' b_l'' b_l''| = 0 \) or equivalently \( \beta_j'' = \gamma_j'' = 0 \) since \( a_{l-1}'' = b''_l \neq 0 \). Likewise, multiply both sides of (20) with \((s + c_1)^{l}\) and using the same argument, as \( s \to -c_1 \), we obtain \( \alpha_j'' = 0 \). Overall, we obtain \( \alpha_j'' = \beta_j'' = \gamma_j'' = 0 \). Continue 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

\[
\alpha_j' + C_3 \sum_{l=1}^{d} M_{ll}^{j} - C_3 (\sum_{u,v} M_{uv}^{j} \frac{[\Sigma_{j}^{1/2} t']_{u} [\Sigma_{j}^{1/2} t']_{v}}{\sigma_j^2}) = 0, \]

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 \( \sum_{u,v} M_{uv}^{j} [\Sigma_{j}^{1/2} t']_{u} [\Sigma_{j}^{1/2} t']_{v} = (t')^T \Sigma_{j}^{1/2} M_{j} \Sigma_{j}^{1/2} t' = (t')^T \gamma_j t' \), it is equivalent that

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

With 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 \( \alpha_j', \beta_j' \), 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.