SIAM J. MATH. DATA SCI.
Vol. 1, No. 4, pp. 730–758

Singularity Structures and Impacts on Parameter Estimation in Finite Mixtures of Distributions*

Nhat Ho† and XuanLong Nguyen‡

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

Key words. Fisher singularities, system of polynomial equations, semialgebraic set, mixture model, nonlinear partial differential equation, minimax lower bound, maximum likelihood estimation, convergence rates, Wasserstein distances

AMS subject classifications. Primary, 62F15, 62G05; Secondary, 62G20

DOI. 10.1137/18M122947X

1. Introduction.

In the standard asymptotic theory of parametric estimation, a customary regularity assumption is the nonsingularity of the Fisher information matrix defined by the statistical model (see, for example, [43, p. 124] or [60, sect. 5.5]). This condition leads to the cherished root-n consistency and in many cases the asymptotic normality of parameter estimates. When the singularity condition holds, that is, when the true parameters represent a singular point in the statistical model, very little is known about the asymptotic behavior of their estimates.

The singularity situation might have been brushed aside as idiosyncratic by some parametric statistical modelers in the past. As complex and high-dimensional models are increasingly embraced by statisticians and practitioners alike, singularities are no longer a rarity—they start to take a highly visible place in modern statistics and data science. For example, the many zeros present in a high-dimensional linear regression problem represent a type of singularities of the underlying model, points corresponding to rank-deficient Fisher information

*Received by the editors November 28, 2018; accepted for publication (in revised form) July 18, 2019; published electronically October 8, 2019.
https://doi.org/10.1137/18M122947X
Funding: The second author's research was supported in part by grants NSF CAREER DMS-1351362 and NSF CNS-1409303.
†Department of EECS, University of California, Berkeley, CA 94720-1770 (minhnhat@berkeley.edu).
‡Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107 (xuanlong@umich.edu).

matrices [32]. In another example, the zero skewness in the family of skewed distributions represents a singular point [14]. In both examples, singularity points are quite easy to spot—it is the impacts of their presence on improved parameter estimation procedures and the asymptotic properties such procedures entail that are nontrivial matters that have occupied the best efforts of many researchers in the past decade. The textbooks [8, 32], for instance, address such issues for high-dimensional regression problems, while the recent papers [44, 30, 31] investigate statistical inference in the skewed families for distribution. By contrast, with finite mixture models—a popular and rich class of modeling tools for density estimation and heterogeneity inference [49, 47, 48, 37] and a subject of this paper—the singularity phenomenon is not quite well understood, to the best our knowledge, except for specific instances.

One of the simplest instances is the singularity of Fisher information matrix in an (overfitted) finite mixture that includes a homogeneous distribution, a setting studied in [39]. The authors of [41] analyzed a test of heterogeneity based on finite mixtures, addressing the challenge arising from the aforementioned singularity. Recent works on the related topic include [16, 17, 11, 10, 19, 26, 38]. Moreover, model selection under singular models is given a book-length treatment in a seminal contribution by Watanabe [62]. Building upon this framework, Drton and Plummer [20] studied finite mixture–based model selection under singularity. Focusing on the estimation of *parameters* of interest, the authors of [53] investigated likelihood-based estimation methods in a somewhat general parametric modeling framework, subject to the constraint that the Fisher information matrix is one rank deficient. For overfitted finite mixtures, the author of [13] showed that under a condition of strong identifiability, there are estimators which achieve the generic convergence rate $n^{-1/4}$ for parameter estimation. Recent works also established generic behaviors of estimation under somewhat broader settings of overfitted finite mixture models with both maximum likelihood estimation and Bayesian estimation [54, 51, 35]. Under sufficiently strong identifiability conditions for kernel densities, a sharp local minimax lower bound of parameter estimation in overfitted finite mixture models was recently obtained [33].

The family of mixture models is far too rich to submit a uniform kind of behavior of parameter estimation, due to a weak identifiability phenomenon induced by underlying singularities that are much more pervasive than previously thought. In fact, it was shown recently that even classical models such as the location-scale Gaussian mixtures, and the shape-rate Gamma mixtures, do not admit such a generic rate of convergence for an estimation method such as the MLE or Bayesian estimation with a noninformative prior [34]. For instance, singularities arise in the finite mixtures of Gamma distributions, even when the number of mixing components is known—this phenomenon results in an extremely slow convergence behavior for the model parameters lying in the vicinity of singular points, even though such parameters are (perfectly) identifiable. Finite mixtures of Gaussian distributions, though identifiable, exhibit both minimax lower bounds and maximum likelihood estimation rates that are directly linked to the solvability of a system of real polynomial equations, rates which deteriorate quickly with the increasing number of extra mixing components. The results obtained for such specific instances contain considerable insights about parameter estimation in finite mixture models, but they only touch upon the surface of a general and complex phenomenon. Indeed, as we shall see in this paper there is a rich spectrum of asymptotic behavior in which regular (nonsingular) mixtures, strongly identifiable mixtures, and weakly identifiable mixture

models such as the ones studied by the aforementioned works occupy only a small spot.

1.1. Main results. The goal of this paper is to present a general and theoretical framework for analyzing parameter estimation behavior in finite mixture models. We address directly the situations where the nonsingularity condition of the Fisher information matrix may not hold. Our approach is to take on a systematic investigation of the singularity structure of a compact and multidimensional parameter space of mixture models, to identify such singularities and then study the impacts of their presence on parameter estimation. There is a remarkable heterogeneity of the mixture model parameter space on which we can shed some light: it will be shown that different parts of the parameter space may admit different convergence rates by several standard estimation methods. Parameters of different types may possess different estimation rates, e.g., location vs. scale of the same mixture component. Even parameters of the same type may carry distinct rates of estimation, such as shape parameters associated with different mixture components.

To obtain such a fine-grained picture of the parameter space, several fundamental concepts will be introduced. In particular, the natural-valued *singularity level* will be useful in describing the convergence behavior of the (discrete) mixing measure that arises in the mixture model. Specifically, a mixture density of the form $p_G(x) = \int f(x|\eta)\mathrm{d}G(\eta)$, where f denotes a kernel density, corresponds to mixing measure G on a suitable parameter space. If $G = \sum_{i=1}^{k} p_i\delta_{\eta_i}$, then it is often denoted that $p_G(x) = \sum_{i=1}^{k} p_i f(x|\eta_i)$. The singularity level for a mixing measure G describes in a precise manner the variation of the mixture likelihood $p_G(x)$ with respect to changes in mixing measure G. Now, Fisher information singularities simply correspond to points in the parameter space which identify a mixing measure whose singularity level is nonzero. Within the set of Fisher information singularities, the parameter space can be partitioned into disjoint subsets determined by different singularity levels.

Given an i.i.d. n-sample from a (true) mixture density p_{G_0}, where G_0 admits a singularity level r, this will imply, under some mild conditions on f, that a standard estimation method such as the MLE and Bayesian estimation with a noninformative prior carries the rate of convergence $n^{-1/2(r+1)}$, which is also a minimax lower bound (up to a logarithmic factor). Here, the convergence rate is expressed in terms of a suitable Wasserstein metric on the space of mixing measures. Thus, singularity level 0 results in the root-n convergence rate for mixing measure estimation. Fisher singularity corresponds to singularity level 1 or greater than 1, resulting in convergence rates $n^{-1/4}, n^{-1/6}, n^{-1/8}$, and so on.

Convergence in a Wasserstein metric on mixing measures is easily translated into convergence of the supporting atoms [51]. But each atom of the mixing measure may be composed of different types (e.g., location, scale, shape). To anticipate the heterogeneity of parameters of different types, we introduce vector-valued *singularity index*, which extends the notion of the natural-valued singularity level described earlier. The singularity index describes the variation of the mixture likelihood with respect to changes of individual parameter of each type. A singularity index κ corresponds to singularity level $(\|\kappa\|_\infty - 1)$ for the mixing measure, but it tells us much more: the convergence rate for estimating the jth component of the atoms η via the MLE or the Bayesian method will be $n^{-1/2\kappa_j}$. One can in fact go further to capture "complete heterogeneity": via *singularity matrix*, it can be shown that each parameter may allow a possibly different convergence rate depending on the parameter's values. The com-

plete picture of the distribution of singularity structure, however, can be extremely complex to derive. Remarkably, there are examples of finite mixtures for which the compact parameter space can be partitioned into disjoint subsets whose singularity level or elements of singularity index and singularity matrix range from 0 to 1 to 2,..., up to infinity. As a result, if we were to vary the true parameter values, we would encounter a phenomenon akin to that of "phase transition" on the statistical efficiency of parameter estimation occurring within the same model class.

1.2. Techniques. A major component of our general framework is a procedure for identifying subsets of points having the same singularity structure, via common singularity level and so on. It will be shown that these points are in fact a subset of a real affine variety. A real affine variety is a set of solutions to a system of real polynomial equations. The polynomial equations can be derived explicitly by the kernel density functions that define a given mixture distribution. The study of the solutions of polynomial equations is a central subject of algebraic geometry [56, 15]. The connections between specific statistical models and algebraic geometry have received a steady stream of contributions in the last two decades, including the analysis of latent class analysis models [27], factor analysis [21], algebraic statistical models [23], discrete Markov random fields [22], finite mixtures of categorical data [1], Gaussian graphical models [58], and the expectation-minimization (EM) algorithms [40]. As mentioned earlier, singular mixture models were studied by the authors of [62, 20], who focused primarily on aspects of density estimation and model selection (e.g., estimation of the number of parameters), not the estimation of parameters per se. By focusing on the statistical efficiency of parameter estimation for finite mixtures of continuous distributions, we have found that the link to algebraic geometry is distilled from a new source of algebraic structure, in addition to the presence of mixing measures: it is traced to the partial differential equations satisfied by the mixture model's kernel density function. For Gaussian mixtures, it is the relation captured by (3.3) for the Gaussian kernel. The partial differential equations can be nonlinear, with coefficients given by rational functions defined in terms of model parameters. It is this relation that is primarily responsible for the complexity of the singularity structure. A quintessential example of such a relation is given by (3.2) for the skew-normal kernel densities.

Starting from the aforementioned partial differential structure, we seek to represent likelihood function $p_G(x)$ in terms of linearly independent functions, which lead to a representation we call *minimal forms*. These forms provide the basis for studying the behavior of the likelihood as G varies in a suitable neighborhood of mixing measures. It turns out that, as we move through increasingly sophisticated concepts of singularity structure (e.g., level, index), there are correspondingly structured notions of transportation distance for the space of the underlying mixing measures. These distances, which generalize the Wasserstein metric and behave asymptotically as *semipolynomials* of the parameter perturbations, prove to be the right object for linking up the information culled from a data sample (via its likelihood function) and the algebraic structure of the parameter space of inferential interest.

Although our method for the analysis of singularity structure and the asymptotic theory for parameter estimation can be used to rederive old and refine existing results such as those of [13, 34]; a substantial outcome is to establish fresh new results on mixture models for which no asymptotic theory has hitherto been achieved. This leads us to a story of finite

mixtures of skew-normal distributions. The skew-normal distribution was originally proposed in [5, 7, 6]. The skew-normal generalizes normal (Gaussian) distribution, which is enhanced by the capability of handling asymmetric (skewed) data distributions. Due to its more realistic modeling capability for multimodality and asymmetric components, skew-normal mixtures have been increasingly adopted in recent years for model-based inference of heterogeneity by many researchers [46, 3, 4, 45, 55, 28, 42, 52, 9, 64]. Due to its usefulness, a thorough understanding of the asymptotic behavior of parameter estimation for skew-normal mixtures is also of interest in its own right.

The source of complexity of skew-normal mixtures is the structure of the skew-normal kernel density. The evidence for the latter was already made clear in [14, 44, 30, 31], in which a thorough picture of the singularities for the class of skew-normal densities was provided, as well as their impacts on the nonstandard rates of convergence of the MLE. Not only can we recover the results of [30, 31] in terms of rates of convergence, because they correspond to a trivial "mixture" that has exactly one skew-normal component, but an entirely new set of results is established for mixtures of two or more components. It is in this setting that new types of singularities arise out of the interactions between distinct skew-normal components. These interactions define the subset of singular points of a given structure that can be characterized by a system of real polynomial equations. This algebraic geometric characterization allows us to establish either the precise singularity structure or an upper bound for the mixture model's entire parameter space. Due to space constraint, we shall describe only a small set of results pertaining to the skew-normal mixtures in this manuscript.

1.3. Implications. Mixture models are one of the most popular and versatile tools in modern statistics and data science. Despite numerous efforts, our understanding of the parameter estimation behavior in mixture models remains woefully incomplete from both theoretical and computational standpoints. As noted by in a recent textbook, "mixture models are riddled with difficulties such as non-identifiability" [18]. Perhaps, as we shall show, the complexity lies not in nonidentifiability per se, but the varying levels of identifiability and the roles they play, which are captured precisely by the concepts of singularity level and index introduced in this paper. We note that our theory of singularity structures also carries important consequences on the computational complexity of parameter estimation procedures, including both optimization- and sampling-based methods. Indeed, the inhomogeneous nature of the singularity structures reveals a complex picture of the likelihood function: regions in parameter space that carry low singularity levels/indices may observe a relatively high curvature of the likelihood surface, while high singularity levels imply a "flatter" likelihood surface along a certain subspace of the parameters. Such a subspace is manifested by our construction of sequences of mixing measures that attest to the condition of singularities (e.g., r-singularity or κ-singularity) in general. Reposing upon this insight, the authors of [24, 25, 50] recently established the slow convergence rates of EM updates for approximating the MLE or developed polynomial mixing time MCMC algorithm for approximating the power posterior distribution for several specific settings of mixture models. It is of interest to further exploit the explicit knowledge of singularity structures obtained for a given mixture model class, so as to improve upon the computational efficiency of the optimization and sampling procedures that operate on the model's parameter space.

The concepts of singularity level and index in the paper can be seen as complementary to the theory of singular models in general and applications to mixture models in particular [62, 2, 20]. Indeed, the theory of [62] aims for a systematic approximation of functionals of a parametric density of interest, including the Bayes generalization error or Kullback–Leibler distance. Although the rate of estimation for a parametric density function generally remains root-n under Hellinger distance or a related functional, accounting for singularities yields a more accurate approximation for the marginal likelihood, which results in an improved information criterion for model choice [20]. On the other hand, we study how the model's likelihood and associated distance functionals vary with respect to model parameters. This study requires an elaborate excursion into the singularity structures of the parameter space, because convergence behavior of individual parameters is considerably more complex than that of a density function. Our singularity concepts provide an efficient way to characterize such structures, which directly yield nonregular parameter estimation rates in mixture models.

The plan for the remainder of our paper is as follows. Section 2 lays out the notations and relevant concepts, such as parameter spaces and the underlying geometries. Section 3 presents the general framework of analysis of singularity structure and the impact on convergence rates of parameter estimation for singular points of a given singularity structure. Section 4 illustrates the theory on the finite mixture of skew-normal distributions, which results in minimax bounds and MLE convergence rates for this class of models for the first time. We conclude with a discussion in section 5. Additional details and proofs are given in the supplementary materials, linked from the main article webpage.

Notation. We utilize several notions of distance for mixture densities with respect to Lebesgue measure μ: total variation distance $V(p_G, p_{G_0}) = (1/2)\int |p_G(x) - p_{G_0}x)|\mathrm{d}\mu(x)$ and squared Hellinger distance $h^2(p_G, p_{G_0}) = (1/2)\int(\sqrt{p_G(x)} - \sqrt{p_{G_0}(x)})^2\mathrm{d}\mu(x)$. Additionally, for any $\kappa_1, \kappa_2 \in \mathbb{R}^d$, we denote $\kappa_1 \preceq \kappa_2$ if and only if all the components of κ_1 are less than or equal to the corresponding components of κ_2. Furthermore, $\kappa_1 \prec \kappa_2$ if and only if $\kappa_1 \preceq \kappa_2$ and $\kappa_1 \neq \kappa_2$. Additionally, the expression "$\gtrsim$" will be used to denote the inequality up to a constant multiple where the value of the constant is fixed within our setting. We write $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Finally, for any $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the greatest integer that is less than or equal to x.

2. Preliminaries.

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

2.1. Parameter spaces and geometries.

There are different kinds of parameter space and geometries that are carried which are relevant to the analysis of mixture models. We proceed to describe them in the following.

Natural parameter space. The customarily defined parameter space of the k-mixture of distributions is that of the *mixing component parameters* η_i and *mixing probabilities* p_i. Throughout this paper, it is assumed that $\eta_i \in \Theta$, which is a compact subset of $\mathbb{R}^d$ for some $d \geq 1$ and for $i = 1, \ldots, k$. The mixing probability vector $\boldsymbol{p} = (p_1, \ldots, p_k) \in \Delta^{k-1}$, the $(k-1)$-probability

simplex. For the remainder of the paper, we also use Ω to denote the compact subset of the Euclidean space for parameters $(\boldsymbol{p}, \boldsymbol{\eta})$.

Example 2.1. The skew-normal density kernel on the real line has three parameters $\eta = (\theta, \sigma, m) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{R}$, namely the location, scale, and skewness (shape) parameters. It is given by, for $x \in \mathbb{R}$,

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

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

Semialgebraic sets. The singularity structure of the parameter space submits to a different geometry. It will be described in terms of the zero sets (sets of solutions) of systems of real polynomial equations. The zero set of a system of real polynomial equations is called a (real) affine variety [15]. In fact, the sets which describe the singularity structure of finite mixtures are not affine varieties per se. We will see that they are the intersection between real affine varieties—the real-valued solutions of a finite collection of equations of the form $P(\boldsymbol{p}, \boldsymbol{\eta}) = 0$, and the set of parameter values satisfying $Q(\boldsymbol{p}, \boldsymbol{\eta}) > 0$, for some real polynomials P and Q. The feasible set of polynomial equalities and inequalities, such as the intersection of the above sets, is often referred to as a *semialgebraic set*.

Example 2.2. Continuing on the example of skew-normal mixtures, we will see that the first two types of singularities that arise from the mixture of skew-normals are solutions of the following polynomial equations:

(i) Type A: $P_1(\boldsymbol{\eta}) = \prod_{j=1}^{k} m_j$.

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

These are just two among many more polynomials and types of singularities that we will be able to enumerate in what follows. We quickly note that Type A refers to the one (and only) kind of singularity intrinsic to the skew-normal kernel: $P_1 = 0$ if either one of the $m_j = 0$—one of the skew-normal components is actually normal (symmetric). This type of singularity has received in-depth treatments by a number of authors [14, 44, 30, 31]. On the other hand, Type B is intrinsic to a mixture model, as it describes the relation of parameters of distinct mixing components i and j.

Space of mixing measures and transportation distance. As described in the introduction, a goal of this work is to turn the knowledge about the nature of singularities of parameter space Ω into that of statistical efficiency of parameter estimation procedures. For this purpose, the convergence of parameters in a mixture model is most naturally analyzed in terms of the convergence in the space of mixing measures endowed by transportation distance (Wasserstein distance) metrics [51]. This is because the role played by parameters $\boldsymbol{p}, \boldsymbol{\eta}$ in the mixture model is via mixing measure G. It is mixing measure G that determines the mixture density p_G according to which the data are drawn from. Since the map $(\boldsymbol{p}, \boldsymbol{\eta}) \mapsto G(\boldsymbol{p}, \boldsymbol{\eta}) = G = \sum p_i \delta_{\boldsymbol{\eta}_i}$

is many-to-one, we shall treat a pair of parameter vectors $(\boldsymbol{p}, \boldsymbol{\eta}) = (p_1, \ldots, p_k; \eta_1, \ldots, \eta_k)$ and $(\boldsymbol{p}', \boldsymbol{\eta}') = (p'_1, \ldots, p'_{k'}; \eta'_1, \ldots, \eta'_{k'})$ as being equivalent if the corresponding mixing measures are equal, $G(\boldsymbol{p}, \boldsymbol{\eta}) = G(\boldsymbol{p}', \boldsymbol{\eta}')$. For ease of notation, we often omit the arguments when the context is clear for $G = G(\boldsymbol{p}, \boldsymbol{\eta})$ and $G' = G(\boldsymbol{p}', \boldsymbol{\eta}')$.

For $r \geq 1$, the Wasserstein distance of order r between G and G' takes the form (cf. [61])

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

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

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

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

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

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

satisfy the following: for any $\gamma \in \mathbb{R}^d$,

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

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

3. Singularity structure in finite mixture models.

3.1. Beyond Fisher information.

Given a mixture model

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

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

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

Example 3.1. To illustrate in the context of skew-normal mixtures, where the parameters $\eta = (\theta, v, m)$, observe that the mixture density structure allows the following characterization: $I(G)$ is degenerate if and only if the collection of partial derivatives

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

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

$$\sum_{j=1}^k \alpha_{1j} f(x|\eta_j) + \alpha_{2j} \frac{\partial f}{\partial \theta}(x|\eta_j) + \alpha_{3j} \frac{\partial f}{\partial v}(x|\eta_j) + \alpha_{4j} \frac{\partial f}{\partial m}(x|\eta_j) = 0 \tag{3.1}$$

for almost all $x \in \mathbb{R}$. Lemma 4.1 later shows that the (Fisher information matrix's) singular points are the zeros of some polynomial equations.

We have seen that for the e-mixtures, G is nonsingular if the collection of density kernel functions $f(x|\eta)$ and their first partial derivatives with respect to each model parameter is linearly independent. This condition is also known as the first-order identifiability. For o-mixtures, the relevant notion is the second-order identifiability [13, 51, 35]: the condition that the collection of kernel density functions $f(x|\eta)$, along with their first and second partial derivatives, is linearly independent. This condition fails to hold for skew-normal kernel densities, whose first and second partial derivatives are linked by the following nonlinear partial differential equations:

$$\begin{cases} \dfrac{\partial^2 f}{\partial \theta^2} - 2\dfrac{\partial f}{\partial v} + \dfrac{m^3+m}{v}\dfrac{\partial f}{\partial m} = 0, \\ 2m\dfrac{\partial f}{\partial m} + (m^2+1)\dfrac{\partial^2 f}{\partial m^2} + 2vm\dfrac{\partial^2 f}{\partial v \partial m} = 0. \end{cases} \tag{3.2}$$

It is straightforward to verify these identities (see Appendix A of the supplementary materials). Note that if $m = 0$, the skew-normal kernel becomes normal kernel, which admits a (simpler) linear relationship:

$$\frac{\partial^2 f}{\partial \theta^2} = 2\frac{\partial f}{\partial v}. \tag{3.3}$$

This relation, also noted previously in [10, 38], plays a fundamental role in the analysis of finite mixtures of location-scale normal distributions [34]. Unlike Gaussian kernels, the nonlinear relations expressed by the above PDEs for skew-normal density kernels underscore the exceptional complexity of general mixture models—the inhomogeneity of the parameter space. Analyzing this requires the development of a general theory that we now embark on.

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

Fix $r \geq 1$, and consider an arbitrary sequence of $G_n \in \mathcal{O}_k$, such that $W_r(G_n, G_0) \to 0$. Let $k_n \leq k$ be the number of distinct support points of G_n. There exists a subsequence of G_n for which k_n is constant in n and each supporting atom η_i^0 as $i \in \{1, \ldots, k_0\}$ of G_0 is the limit point of exactly s_i atoms of G_n. Additionally, there may be also a subset of atoms of G_n whose limits are not among the atoms of G_0. Without loss of generality, we assume that there are $\bar{l} \geq 0$ such limit points, which are denoted by η_i^0 for $k_0 + 1 \leq i \leq k_0 + \bar{l}$. By relabelling its atoms, we can express G_n as

$$G_n = \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}^n \delta_{\eta_{ij}^n}, \tag{3.4}$$

where $\eta_{ij}^n \to \eta_i^0$ for all $i = 1, \ldots, k_0 + \bar{l}$, $j = 1, \ldots, s_i$. Additionally, $\sum_{i=1}^{k_0+\bar{l}} s_i = k_n$. Thus, from here on we replace the sequence of G_n by this subsequence. To simplify the notation, n will be dropped from the superscript when the context is clear. In addition, we use the notation $\Delta\eta_{ij} := \eta_{ij} - \eta_i^0$ for $i = 1, \ldots, k_0 + \bar{l}$, $j = 1, \ldots, s_i$. Also, $p_{i\cdot} := \sum_{j=1}^{s_i} p_{ij}$, and $\Delta p_{i\cdot} := p_{i\cdot} - p_i^0$, for $i = 1, \ldots, k_0 + \bar{l}$, where $p_i^0 = 0$ as $k_0 + 1 \leq i \leq k_0 + \bar{l}$. For the setting

of e-mixtures, the sequence of elements G_n is restricted to $\mathcal{E}_{k_0} \subset \mathcal{O}_k$, so $k_n = k_0$ for all n. It follows that $s_i^n = 1$ for all $i = 1, \ldots, k_0$ and $\bar{l} = 0$, so the notation is simplified further: let $\Delta\eta_i := \Delta\eta_{i1} = \eta_i - \eta_i^0, \Delta p_i := \Delta p_{i\cdot} = p_i - p_i^0$ for all $i = 1, \ldots, k_0$.

We define the following distance:

$$D_r(G, G_0) := \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij} \|\Delta\eta_{ij}\|_r^r + \sum_{i=1}^{k_0+\bar{l}} |\Delta p_{i\cdot}|. \tag{3.5}$$

Then, as $W_r(G, G_0) \downarrow 0$, it is simple to see that

$$W_r^r(G, G_0) \asymp D_r(G, G_0). \tag{3.6}$$

The above relation relates a Wasserstein metric to a *semipolynomial* of degree r (recall that a semipolynomial is a polynomial of a collection of variables and/or the absolute value of some of the variables). To investigate the behavior of likelihood function $p_G(x)$ as G tends to G_0 in Wasserstein distance W_r, by representation (3.4),

$$p_G(x) - p_{G_0}(x) = \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij}(f(x|\eta_{ij}) - f(x|\eta_i^0)) + \sum_{i=1}^{k_0+\bar{l}} \Delta p_{i\cdot} f(x|\eta_i^0). \tag{3.7}$$

By Taylor expansion up to order r, we find that

$$\begin{aligned} p_G(x) - p_{G_0}(x) = &\sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij} \sum_{|\kappa|=1}^{r} \frac{(\Delta\eta_{ij})^\kappa}{\kappa!} \frac{\partial^{|\kappa|} f}{\partial \eta^\kappa}(x|\eta_i^0) \\ &+ \sum_{i=1}^{k_0+\bar{l}} \Delta p_{i\cdot} f(x|\eta_i^0) + R_r(x), \end{aligned} \tag{3.8}$$

where $R_r(x)$ is the Taylor remainder. Moreover, we have $\sup_x |R_r(x)/W_r^r(G, G_0)| \to 0$, since f is uniform Lipschitz up to the order r. We arrive at the following formula, which measures the speed of change of the likelihood function as G varies in the Wasserstein neighborhood of G_0:

$$\begin{aligned} \frac{p_G(x) - p_{G_0}(x)}{W_r^r(G, G_0)} = &\sum_{|\kappa|=1}^{r} \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} \left(\frac{p_{ij}(\Delta\eta_{ij})^\kappa/\kappa!}{W_r^r(G, G_0)} \right) \frac{\partial^{|\kappa|} f}{\partial \eta^\kappa}(x|\eta_i^0) \\ &+ \sum_{i=1}^{k_0+\bar{l}} \frac{\Delta p_{i\cdot}}{W_r^r(G, G_0)} f(x|\eta_i^0) + o(1). \end{aligned} \tag{3.9}$$

The right-hand side of (3.9) is a linear combination of the partial derivatives of f evaluated at G_0. It is crucial to note, by the relation (3.6), that each coefficient of this linear representation is asymptotically equivalent to the ratio of two semipolynomials.

Equation (3.9) highlights the distinct roles of model parameters and the kernel density function in the formation of a mixture model's likelihood. The former appears only in the

coefficients, while the latter provides the partial derivatives which appear as if they are basis functions for the linear combination. We wrote "as if," because the partial derivatives of kernel f may not be linearly independent functions (recall the examples in section 3.1). When a partial derivative of f can be represented as a linear combination of other partial derivatives, it can be eliminated from the expression in the right-hand side. This reduction process may be repeatedly applied until all partial derivatives that remain are linearly independent functions. This motivates the following concept.

Definition 3.2. *The following representation is called an* r-minimal *form of the mixture likelihood for a sequence of mixing measures* G *tending to* G_0 *in the* W_r *metric:*

$$\frac{p_G(x) - p_{G_0}(x)}{W_r^r(G, G_0)} = \sum_{l=1}^{T_r} \left(\frac{\xi_l^{(r)}(G)}{W_r^r(G, G_0)} \right) H_l^{(r)}(x) + o(1), \tag{3.10}$$

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

(1) $H_l^{(r)}(x)$ *for all* l *are linearly independent functions of* x*, and*

(2) *coefficients* $\xi_l^{(r)}(G)$ *are polynomials of the components of* $\Delta\eta_{ij}$*, and* $\Delta p_{i\cdot}, p_{ij}$*.*

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

Definition 3.3. *Fix* $r \geq 1$*, and let* $\mathcal{G}$ *be a class of discrete probability measures which has a bounded number of support points in* Θ*. We say that* G_0 *is* r*-singular relative to* $\mathcal{G}$ *if* G_0 *admits an* r*-minimal form given by* (3.10)*, according to which there exists a sequence of* $G \in \mathcal{G}$ *tending to* G_0 *under* W_r *such that*

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

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

Lemma 3.4. (a) *(Invariance) The existence of the sequence of* G *in the statement of Definition 3.3 holds for all* r*-minimal forms once it holds for at least one* r*-minimal form.*

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

The monotonicity of r-singularity naturally leads to the notion of singularity level of a mixing measure G_0 (and the corresponding parameters) relative to an ambient space $\mathcal{G}$.

Definition 3.5. *The singularity level of* G_0 *relative to a given class* $\mathcal{G}$*, denoted by* $\ell(G_0|\mathcal{G})$*, is*

0 *if* G_0 *is not* r*-singular for any* $r \geq 1$*;*
∞ *if* G_0 *is* r*-singular for all* $r \geq 1$*;*
otherwise, the largest natural number $r \geq 1$ *for which* G_0 *is* r*-singular.*

The role of the ambient space $\mathcal{G}$ is critical in determining the singularity level of $G_0 \in \mathcal{G}$. Clearly, by definition, if $G_0 \in \mathcal{G} \subset \mathcal{G}'$, r-singularity relative to $\mathcal{G}$ entails r-singularity relative to $\mathcal{G}'$. This means $\ell(G_0|\mathcal{G}) \leq \ell(G_0|\mathcal{G}')$. Let us look at the following examples:

- Take $\mathcal{G} = \mathcal{E}_{k_0}$, i.e., the setting of e-mixtures. If the collection of $\{\partial^\kappa f/\partial\eta^\kappa(x|\eta_j)|j = 1,\ldots,k_0; |\kappa| \leq 1\}$ evaluated at G_0 is linearly independent, then G_0 is not 1-singular relative to $\mathcal{E}_{k_0}$. It follows that $\ell(G_0|\mathcal{G}) = 0$.
- Take $\mathcal{G} = \mathcal{O}_k$ for any $k > k_0$, i.e., the setting of o-mixtures. It is not difficult to check that G_0 is always 1-singular relative to $\mathcal{O}_k$ for any $k > k_0$. Thus, $\ell(G_0|\mathcal{G}) \geq 1$. Now, if the collection of $\{\partial^\kappa f/\partial\eta^\kappa(x|\eta_j)|j = 1,\ldots,k_0; |\kappa| \leq 2\}$ evaluated at G_0 is linearly independent, then it can be observed that G_0 is not 2-singular relative to $\mathcal{O}_k$. Thus, $\ell(G_0|\mathcal{G}) = 1$.

The conditions described in the two examples above are in fact referred to as strong identifiability conditions studied in [13, 51, 35]. The notion of singularity level generalizes such identifiability conditions by allowing us to consider situations where such conditions fail to hold. This is the situation where $\ell(G_0|\mathcal{G}) = 2, 3, \ldots, \infty$. The significance of this concept can be appreciated by the following theorem.

Theorem 3.6. *Let $\mathcal{G}$ be a class of probability measures on Θ that have a bounded number of support points, and fix $G_0 \in \mathcal{G}$. Suppose that $\ell(G_0|\mathcal{G}) = r$ for some $0 \leq r < \infty$:*

(i) *There holds that $\inf_{G\in\mathcal{G}} \frac{\|p_G - p_{G_0}\|_\infty}{W_s^s(G,G_0)} > 0$ for any $s \geq r+1$.*

(ii) *In addition, $\inf_{G\in\mathcal{G}} \frac{V(p_G,p_{G_0})}{W_s^s(G,G_0)} > 0$ for any $s \geq r+1$.*

The following establishes that the bounds obtained above are tight under some conditions.

Theorem 3.7. *Consider the same setting as that of Theorem 3.6.*

(i) *If $1 \leq r < \infty$ and, in addition, the following hold:*

(a) *f is $(r+1)$-order differentiable with respect to η and for some constant $c_0 > 0$,*

$$\sup_{\|\eta-\eta'\|\leq c_0} \int_{x\in\mathcal{X}} \left(\frac{\partial^s f}{\partial\eta^\alpha}(x|\eta)\right)^2 / f(x|\eta')dx < \infty \tag{3.11}$$

for any index α such that $|\alpha| = s$ and $s \in \{1,\ldots,r\}$.

(b) *There is a sequence of $G \in \mathcal{G}$ tending to G_0 in Wasserstein metric W_r, and the coefficients of the r-minimal form $\xi_l^{(r)}(G)$ satisfy $\xi_l^{(r)}(G)/W_1^s(G,G_0) \to 0$ for all real number $s \in [1, r+1)$ and $l = 1,\ldots,T_r$. Additionally, all the masses p_{ij} in the representation (3.4) of G are bounded away from 0.*

Then, for any real number $s \in [1, r+1)$,

$$\liminf_{G\in\mathcal{G}:W_1(G,G_0)\to 0} \frac{h(p_G,p_{G_0})}{W_1^s(G,G_0)} = \liminf_{G\in\mathcal{G}:W_s(G,G_0)\to 0} \frac{V(p_G,p_{G_0})}{W_s^s(G,G_0)} = 0.$$

(ii) *If $\ell(G_0|\mathcal{G}) = \infty$ and the conditions (a), (b) in part (i) hold for any $l \in \mathbb{N}$ (here, the parameter r in these conditions is replaced by l), then the conclusion of part (i) holds for any $s \geq 1$.*

We make a few remarks.

- Parts (i) and (ii) of Theorem 3.6 show how the singularity level of G_0 relative to ambient space $\mathcal{G}$ may be used to translate the convergence of mixture densities (under the sup-norm and the total variation distance) into the convergence of mixing measures under a suitable Wasserstein metric W_s. Part (i) of Theorem 3.7 shows a sufficient condition under which the power $r+1$ in the bounds from Theorem 3.6 cannot be improved.
- In part (i) of Theorem 3.7, the condition regarding the integrand of the partial derivative of f (cf. (3.11)) can be easily checked to be satisfied by many kernels, such as the Gaussian kernel, Gamma kernel, Student t's kernel, and skew-normal kernel.
- Condition (b) regarding the sequence of G appears somewhat opaque in general, but it will be illustrated in specific examples for skew-normal mixtures in what follows. It is sufficient, but not necessary, for verifying the r-singularity of G_0 to construct the sequence of G so that $\xi_l^{(r)}(G) = 0$ for all $1 \le l \le T_r$, provided such a sequence exists. This requires finding an appropriate parameterization of a sequence of G tending toward G_0 that satisfies a number of polynomial equations defined in terms of the parameter perturbations.

We are ready to state the impact of the singularity level of mixing measure G_0 relative to an ambient space $\mathcal{G}$ on the rate of convergence for an estimate of G_0, where $\mathcal{G} = \mathcal{E}_{k_0}$ in e-mixtures, and $\mathcal{G} = \mathcal{O}_k$ in o-mixtures. Let $\mathcal{G}$ be structured into a sieve of subsets defined by the maximum singularity level relative to $\mathcal{G}$:

$$\mathcal{G} = \bigcup_{r=1}^{\infty} \mathcal{G}_r, \quad \text{where} \quad \mathcal{G}_r := \left\{ G \in \mathcal{G} \middle| \ell(G|\mathcal{G}) \le r \right\}, \quad r = 0, 1, \ldots, \infty.$$

The first part of the following theorem gives a minimax lower bound for the estimation of the mixing measure G_0, given that the singularity level of G_0 is known up to a constant $r \ge 1$. The second part gives a result on the convergence rate of a point estimate such as the MLE.

Theorem 3.8. (a) *Fix $r \ge 1$. Assume that for any $G_0 \in \mathcal{G}_r$, the conclusion of part* (i) *of Theorem 3.7 holds for $\mathcal{G}_r$ (i.e., $\mathcal{G}$ is replaced by $\mathcal{G}_r$ in that theorem). Then, for any real number $s \in [1, r+1)$, there holds that*

$$\inf_{\widehat{G}_n \in \mathcal{G}_r} \sup_{G_0 \in \mathcal{G}_r} E_{p_{G_0}} W_s(\widehat{G}_n, G_0) \gtrsim n^{-1/2s}.$$

Here, the infimum is taken over all sequences of estimates $\widehat{G}_n \in \mathcal{G}_r$ and $E_{p_{G_0}}$ denotes expectation taken with respect to product measure with mixture density $p_{G_0}^n$.

(b) *Let $G_0 \in \mathcal{G}_r$ for some fixed $r \ge 1$. Let $\widehat{G}_n \in \mathcal{G}_r$ be a point estimate for G_0, which is obtained from an n-sample of i.i.d. observations drawn from p_{G_0}. As long as $h(p_{\widehat{G}_n}, p_{G_0}) = O_P(n^{-1/2})$, we have*

$$W_{r+1}(\widehat{G}_n, G_0) = O_P(n^{-1/2(r+1)}).$$

Proof. Part (a) of this theorem is a consequence of the conclusion of Theorem 3.7. The proof of this deduction is rather standard and similar to that of Theorem 1.1 of [34], and hence it is omitted. Part (b) follows immediately from part (ii) of Theorem 3.6, as we have $h(p_{\widehat{G}_n}, p_{G_0}) \geq V(p_{\widehat{G}_n}, p_{G_0}) \gtrsim W_{r+1}^{r+1}(\widehat{G}_n, G_0)$. ∎

We conclude this section with some comments. It is well known that many density estimation methods, such as the MLE and Bayesian estimation applied to a compact parameter space for parametric mixture models, guarantee a root-n rate (up to a logarithmic term) of convergence under the Hellinger distance metric on the density functions (cf. [59, 29, 18]). It follows that, as far as we are concerned, the remaining work in establishing the convergence behavior of mixing measure estimation (as opposed to density estimation) lies in the calculation of the singularity levels, i.e., the identification of sets $\mathcal{G}_r$. For skew-normal mixtures, such calculations will be carried out in section 4. For the settings of G_0 where we are able to obtain the exact singularity levels, we can also construct the sequence of G required by part (i) of Theorem 3.7. Whenever the exact singularity level is obtained, we automatically obtain a (local) minimax lower bound and a matching upper bound for MLE convergence rate under a Wasserstein distance metric, thanks to the above theorem. In some cases, however, the singularity level of G_0 may be not determined exactly, but only an upper bound given. In such cases, only an upper bound to the convergence rate of the MLE can be obtained, while minimax lower bounds may be unknown.

3.3. Inhomogeneity of parameter space. Singularity level is a useful concept for deriving rates of convergence for the mixing measure under a suitable Wasserstein metric, which in turn entails an upper bound on the rate of convergence for individual model parameters (which make up the atoms of the mixing measure). However, different parameters may actually admit different convergence behaviors. To study this phenomenon of inhomogeneity, a more elaborate notion is required to describe the singularity structure of the parameter space. In this section, we shall introduce such a notion, which we call *singularity index*, along with *generalized transportation distance* for the mixing measures.

3.3.1. Generalized transportation distance. As we have seen in the discussion after the definition of a Wasserstein metric in section 2.1, the convergence rate of mixing measures G_n under Wasserstein metric W_r induces the same rate of convergence for the atoms of G_n, denoted by η. By way of example, suppose that η is in fact made up of three parameters $\eta = (\theta, v, m)$, as illustrated in the case of skew-normal mixtures; this implies that the same upper bound ω_n holds for all individual components θ, v, and m of the model parameter. Thus, this fails to demonstrate the situations in which different parameter components may in fact exhibit distinct convergence behavior. For instance, later we will see that in normal and skew-normal mixtures, location parameters may converge more slowly than the scale parameters.

To derive inhomogeneous convergence rates of different model parameters, we introduce a general version of the optimal transport distance, which can be formulated as follows.

Definition 3.9. *For any $\kappa = (\kappa_1, \ldots, \kappa_d) \in \mathbb{N}^d$, let*

$$d_\kappa(\theta_1, \theta_2) := \left(\sum_{i=1}^{d} |\theta_{1,i} - \theta_{2,i}|^{\kappa_i} \right)^{1/\|\kappa\|_\infty},$$

where $\theta_i = (\theta_{i,1}, \ldots, \theta_{i,d}) \in \mathbb{R}^d$, *as* $i = 1, 2$, *and* $\|\kappa\|_\infty = \max_{1 \le i \le d}\{\kappa_i\}$. *The generalized transportation distance with respect to* κ *is given by*

$$\widetilde{W}_\kappa(G, G') := \left(\inf \sum_{i,j} q_{ij} d_\kappa^{\|\kappa\|_\infty}(\eta_i, \eta'_j) \right)^{1/\|\kappa\|_\infty},$$

where the infimum is taken over all couplings $\boldsymbol{q}$ *between* $\boldsymbol{p}$ *and* $\boldsymbol{p'}$.

For instance, in skew-normal mixtures, if $\eta = (\theta, v, m)$, $\eta' = (\theta', v', m')$, and $\kappa = (2, 1, 1)$, then $d_\kappa(\eta, \eta') = (|\theta - \theta'|^2 + |v - v'| + |m - m'|)^{1/2}$. For general κ, d_κ is a semimetric—it satisfies a "weak" triangle inequality; i.e., it only satisfies the triangle inequality up to some positive constant less than 1, except when all κ_i are identical. This implies that $\widetilde{W}_\kappa(G, G_0)$ is a semimetric that only satisfies a weak triangle inequality. There is an easy relation between generalized transportation distance and the standard Wasserstein distance, which is given by

$$\widetilde{W}_\kappa(G, G') \gtrsim W_r(G, G') \tag{3.12}$$

for any $\kappa \in \mathbb{N}^d$ such that $\|\kappa\|_\infty = r \ge 1$. The equality holds when $\kappa_i = r$ for all $1 \le i \le d$. Note that the role of κ in the definition of generalized transportation distance is to capture the inhomogeneous behaviors of different parameters. In particular, assume that a sequence of probability measures $G_n \in \mathcal{O}_k$ tending to G_0 under the $\widetilde{W}_\kappa$ metric at a rate $\omega_n = o(1)$ for some $\kappa \in \mathbb{N}^d$. If all G_n have the same number of atoms $k_n = k_0$ as that of G_0, then the set of atoms of G_n converges to the k_0 atoms of G_0, up to a permutation of the atoms, at the same rate ω_n under the d_κ metric. Therefore, the ith component of each atom of G_n will converge to the ith component of the corresponding atom of G_0 at rate $(\omega_n)^{\|\kappa\|_\infty/\kappa_i}$ for any $i = 1, \ldots, d$. A similar implication holds when G_n has its number of components varying from k_0 to k.

3.3.2. Singularity index. As in section 3.2, we adopt the strategy of investigating the behavior of the likelihood function $p_G(x)$ as G varies in a generalized transportation neighborhood of G_0. In particular, for any fixed $\kappa \in \mathbb{N}^d$, consider a sequence of $G_n \in \mathcal{O}_k$ such that $\widetilde{W}_\kappa(G_n, G_0) \to 0$. Let G_n be represented as in (3.4). To avoid notational cluttering, we again drop n from the superscript when the context is clear. Similar to the relation (3.6), we can also relate the generalized transportation distance to a semipolynomial of order $\|\kappa\|_\infty$. In particular, for any fixed $\kappa \in \mathbb{N}^d$ and any element G represented by (3.4), we define the following distance:

$$D_\kappa(G, G_0) := \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} p_{ij} d_\kappa^{\|\kappa\|_\infty}(\eta_{ij}, \eta_i^0) + \sum_{i=1}^{k_0+\bar{l}} |\Delta p_{i\cdot}|. \tag{3.13}$$

Then, as $\widetilde{W}_\kappa(G, G_0) \downarrow 0$, we find that

$$\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0) \asymp D_\kappa(G, G_0). \tag{3.14}$$

Now, we denote $r := \|\kappa\|_\infty$ (this notational choice is deliberate, as we will see shortly). By virtue of the argument from section 3.2, using Taylor expansion up to the r-order we have

$$\frac{p_G(x) - p_{G_0}(x)}{\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0)} = \sum_{|\alpha|=1}^{r} \sum_{i=1}^{k_0+\bar{l}} \sum_{j=1}^{s_i} \left(\frac{p_{ij}(\Delta\eta_{ij})^\alpha/\alpha!}{\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0)} \right) \frac{\partial^{|\alpha|} f}{\partial \eta^\alpha}(x|\eta_i^0) + \sum_{i=1}^{k_0+\bar{l}} \frac{\Delta p_{i\cdot}}{\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0)} f(x|\eta_i^0) + \frac{R_r(x)}{\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0)},$$

where $R_r(x)$ is the Taylor remainder. By the inequality (3.12), we can verify that

$$\sup_x \left| R_r(x)/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0) \right| \lesssim \sup_x |R_r(x)/W_r^r(G, G_0)| \to 0$$

as long as f is uniform Lipschitz up to order r. Thus,

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

We are ready to define the following concept.

Definition 3.10. *For any $\kappa \in \mathbb{N}^d$, the following representation is called the κ-*minimal *form of the mixture likelihood for a sequence of mixing measures G tending to G_0 in $\widetilde{W}_\kappa$ distance:*

$$\frac{p_G(x) - p_{G_0}(x)}{\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0)} = \sum_{l=1}^{T_\kappa} \left(\frac{\xi_l^{(\kappa)}(G)}{\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0)} \right) H_l^{(\kappa)}(x) + o(1), \tag{3.15}$$

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

(1) *$H_l^{(\kappa)}(x)$ for all l are linearly independent functions of x, and*

(2) *coefficients $\xi_l^{(\kappa)}(G)$ are polynomials of the components of $\Delta\eta_{ij}$ and $\Delta p_{i\cdot}, p_{ij}$.*

It is clear that the κ-minimal form is a general version of the r-minimal form when $\kappa = (r, \ldots, r)$. The procedure for constructing κ-minimal forms is similar to that of r-minimal forms, and will be given in section 3.5, where we specifically search for a subset of linearly independent partial derivatives up to the order $\|\kappa\|_\infty$. The multi-index κ-form provides the basis for the notion of multi-index singularity that we now define.

Definition 3.11. *Let $\mathcal{G}$ be a class of probability measures which has a bounded number of support points in Θ. For any $\kappa \in \mathbb{N}^d$, we say that G_0 is κ-singular relative to $\mathcal{G}$ if G_0 admits a κ-minimal form given by (3.15), according to which there exists a sequence of $G \in \mathcal{G}$ tending to G_0 under $\widetilde{W}_\kappa$ distance such that*

$$\xi_l^{(\kappa)}(G)/\widetilde{W}_\kappa^{\|\kappa\|_\infty}(G, G_0) \to 0 \quad \text{for all } l = 1, \ldots, T_\kappa.$$

Like r-singularity, the notion of κ-singularity possesses a crucial monotonic property in terms of partial order with vector.

Lemma 3.12. (a) *(Invariance) The existence of the sequence of G in the statement of Definition 3.11 holds for all κ-minimal forms once it holds for at least one κ-minimal form.*

(b) *(Monotonicity) If G_0 is κ-singular for some $\kappa \in \mathbb{N}^d$, then G_0 is κ'-singular for any $\kappa' \preceq \kappa$.*

Let $\overline{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$. The monotonicity of κ-singularity naturally leads to the following notion of singularity index of a mixing measure G_0 (and the corresponding parameters) relative to an ambient space $\mathcal{G}$.

Definition 3.13. *For any $\kappa \in \overline{\mathbb{N}}^d$, we say κ is a singularity index of G_0 relative to a given class $\mathcal{G}$ if and only if G_0 is κ'-singular relative to $\mathcal{G}$ for any $\kappa' \prec \kappa$, and there is no $\kappa' \succeq \kappa$ such that G_0 remains κ'-singular relative to $\mathcal{G}$. Define the* singularity index set

$$\mathcal{L}(G_0|\mathcal{G}) := \left\{\kappa \in \overline{\mathbb{N}}^d : \kappa \text{ is a singularity index of } G_0 \text{ relative to } \mathcal{G}\right\}.$$

This definition suggests that the singularity index set may not be always a singleton in general. The following proposition clarifies the relation between singularity level $\ell(G_0|\mathcal{G})$ and singularity index set $\mathcal{L}(G_0|\mathcal{G})$.

Proposition 3.14. *Assume that $\ell(G_0|\mathcal{G}) = r$ for some $r \geq 0$. Then the following hold:*

(i) *If $r = 0$, then $\mathcal{L}(G_0|\mathcal{G}) = \{(1, \ldots, 1)\}$.*

(ii) *If $r = \infty$, then $\mathcal{L}(G_0|\mathcal{G}) = \{(\infty, \ldots, \infty)\}$.*

(iii) *If $1 \leq r < \infty$, then there exists $\kappa \in \mathcal{L}(G_0|\mathcal{G})$ such that $\kappa \preceq (r+1, \ldots, r+1)$ and at least one component of κ is $r+1$.*

(iv) *If $r \geq 1$, and G_0 is not $\overline{\kappa}$-singular relative to $\mathcal{G}$ for some $\overline{\kappa} \in \mathbb{N}^d$, then there exists $\kappa \in \mathcal{L}(G_0|\mathcal{G})$ such that $\kappa \preceq \overline{\kappa}$.*

(v) *If some finite $\kappa \in \mathcal{L}(G_0|\mathcal{G})$, then $\ell(G_0|\mathcal{G}) \leq \|\kappa\|_\infty - 1$. Moreover, if κ is unique, then $\ell(G_0|\mathcal{G}) = \|\kappa\|_\infty - 1$.*

This proposition establishes that when the singularity index set of a mixing measure G_0 relative to $\mathcal{G}$ is a singleton, one can determine the corresponding singularity level of G_0 immediately. We will give several examples of ambient space $\mathcal{G}$ and kernel f under which this situation holds (cf. the examples in section 3.4). The role of the singularity index is in determining minimax lower bound and convergence rate of estimation for individual parameters that make up the mixing measure G_0. Briefly speaking, provided that κ is a singularity index of G_0, then any density estimation method, such as the MLE or Bayesian estimation, that guarantees, say, a root-n rate of convergence toward density p_{G_0} under the Hellinger metric will lead to the convergence rate $n^{-1/2\|\kappa\|_\infty}$ of estimating G_0 under generalized transportation metric $\widetilde{W}_\kappa$, which is also minimax under additional conditions on kernel density f. The implication of such results is that the ith component of each atom of G_0 can be estimated with rate $n^{-1/2\kappa_i}$ where κ_i is the ith component of κ. Such results will be described in Appendix B of the supplementary materials.

Complete inhomogeneity. Although the singularity index captures the inhomogeneous convergence behavior of different components of an atom of G_0, i.e., parameters of different

types, such as location, scale, and skewness, it is possible that every parameter in a mixture model admits a different convergence rate, including those of the same type but associating with different mixture components. We call this phenomenon "complete inhomogeneity." To characterize this, we shall introduce *blocked generalized transportation distance* by replacing uniform semimetric d_κ in the formulation of generalized transportation distance as possibly different semimetrics d_{K_i} with respect to the ith atom η_i^0 of G_0 where $K_i \in \mathbb{N}^d$ for all $1 \leq i \leq k_0$. The convergence rates of η_i^0, therefore, will be determined by the optimal choices of K_i for any i. To quantify these choices, we define a new notion of *singularity matrix* in terms of a matrix K which treats all K_i as its rows. With this concept in place, we can establish rates of convergence for estimating G_0 and its atoms, as well as components of these atoms based on specific values of the singularity matrix. Due to space constraint, detailed formulation and discussion of singularity matrix are given in the technical report [36].

3.4. Revisiting known results on finite mixtures. In this section, singularity structures of parameter space will be examined to shed some light on and further refine known or recent results on the parameter estimation behavior of several classes of finite mixtures.

O-mixtures with second-order identifiable kernels. As being studied in [13, 51, 35, 54], the second-order identifiability condition of kernel density f simply means that the collection of $\{\partial^\kappa f/\partial\eta^\kappa(x|\eta_j)|j = 1,\ldots,k_0; |\kappa| \leq 2\}$ evaluated at G_0 is linearly independent.

Proposition 3.15. *Assume that f is second-order identifiable and admits uniform Lipschitz condition up to second order. Then $\ell(G_0|\mathcal{O}_k) = 1$ and $\mathcal{L}(G_0|\mathcal{O}_k) = \{(2,\ldots,2)\}$.*

By Proposition 3.15 and Theorem 3.8, the convergence rate of estimating G_0 under o-mixtures of second-order identifiable kernel f is $n^{-1/4}$. Moreover, by Theorem SM2.2 in Appendix B in the supplementary materials, the components of each atom of G_0 also admit uniform convergence rate $n^{-1/4}$.

Univariate Gaussian o-mixtures. Location-scale Gaussian mixtures are among the most popular mixture models in statistics. For simplicity, consider univariate Gaussian o-mixtures and let $G_0 \in \mathcal{E}_{k_0} \subset \mathcal{O}_{k,c_0}$ for some $k > k_0$ and small constant $c_0 > 0$. That is, the ambient space $O_{k,c_0} \subset \mathcal{O}_k$ contains only (discrete) probability measures whose point masses are bounded from below by c_0. Let $\{f(x|\theta, v = \sigma^2)\}$ be the family of univariate location-scale Gaussian distributions. Recall the partial differential equation, (3.3), satisfied by Gaussian kernels [10, 38, 34]. Following [34], denote by $\overline{r}(k-k_0)$ the minimal value of $r > 0$ such that the system of polynomial equations

$$\sum_{j=1}^{k-k_0+1} \sum_{\substack{n_1+2n_2=\alpha \\ n_1,n_2\geq 0}} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1!n_2!} = 0 \quad \text{for each } \alpha = 1,\ldots,s \tag{3.16}$$

does not have any solution for the unknowns $(a_j, b_j, c_j)_{j=1}^{k-k_0+1}$ such that all of the c_j's are nonzeros, and at least one of the a_j's is nonzero. By means of the argument from Proposition 2.2 in [34], we can verify that the singularity level of G_0 is $\ell(G_0|\mathcal{O}_{k,c_0}) = \overline{r}(k-k_0) - 1$. It leads to the convergence rate $n^{-1/2\overline{r}(k-k_0)}$ of estimating mixing measure G_0 when we overfit Gaussian mixture models by k components, as established in [34]. However, we can say more: the location parameters and the scale parameters in the Gaussian mixtures admit different

convergence rates (bounds). This is due to examining the singularity index of G_0.

Proposition 3.16. *For any $G_0 \in \mathcal{E}_{k_0} \cap \mathcal{O}_{k_0,c_0}$, we obtain*

$$\mathcal{L}(G_0|\mathcal{O}_{k,c_0}) = \begin{cases} \left\{\left(\overline{r}(k-k_0), \dfrac{\overline{r}(k-k_0)}{2}\right)\right\} & \text{if } \overline{r}(k-k_0) \text{ is an even number,} \\ \left\{\left(\overline{r}(k-k_0), \dfrac{\overline{r}(k-k_0)+1}{2}\right)\right\}, & \text{if } \overline{r}(k-k_0) \text{ is an odd number.} \end{cases}$$

The result of Proposition 3.16 indicates that under Gaussian o-mixtures the best possible convergence rate of location parameter is $n^{-1/2\overline{r}(k-k_0)}$ while that of scale parameter is $n^{-1/\overline{r}(k-k_0)}$ when $\overline{r}(k-k_0)$ is even or is $n^{-1/(\overline{r}(k-k_0)+1)}$ when $\overline{r}(k-k_0)$ is even. These convergence rates are sharp, thanks to part (a) of Theorem SM2.2 in Appendix B of the supplementary materials. Thus, in an overfitted Gaussian mixture, the more overfitted the model, the slower the estimation rate. Moreover, the scale parameter is generally more efficient to estimate than the location parameter. This phenomenon was also established recently with EM updates for location and scale parameters under a few specific settings of univariate Gaussian o-mixtures [25].

Gamma mixtures. The Gamma family of densities assumes the form $f(x|a,b) := \frac{b^a}{\Gamma(a)} x^{a-1}\exp(-bx)$ for $x > 0$, and 0 otherwise, where a, b are positive shape and rate parameters, respectively. The Gamma kernel admits the following partial differential equation:

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

As demonstrated in [34], this identity leads to two disjoint categories of the parameter values of G_0, which are called "generic cases" and "pathological cases," respectively. In particular, denote $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(a_i^0, b_i^0)}$, where $k_0 \geq 2$. Assume that $a_i^0 \geq 1$ for all $1 \leq i \leq k_0$. Now, we define the following:

(A1) Generic cases: $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} \neq \{1, 0\}$ for all $1 \leq i, j \leq k_0$.

(A2) Pathological cases: $\{|a_i^0 - a_j^0|, |b_i^0 - b_j^0|\} = \{1, 0\}$ for some $1 \leq i, j \leq k_0$.

For the Gamma o-mixtures setting, following [34] we also define the constrained set of $\mathcal{O}_k$:

$$\overline{\mathcal{O}}_{k,c_0} = \Bigg\{ G = \sum_{i=1}^{k'} p_i \delta_{(a_i,b_i)} \Bigg| k' \leq k \text{ and } |a_i - a_j^0| \notin [1-c_0, 1+c_0] \cup [2-c_0, 2+c_0] \quad \text{for all } (i,j) \Bigg\},$$

where $c_0 > 0$. The singularity structure of G_0 is clarified by the following result.

Proposition 3.17. *Fix any $G_0 \in \mathcal{E}_{k_0}$.*

(a) *For generic cases specified by* (A1)*, the following hold:*

(a1) *For e-mixtures: $\ell(G_0|\mathcal{E}_{k_0}) = 0$, $\mathcal{L}(G_0|\mathcal{E}_{k_0}) = \{(1,1)\}$.*

(a2) *For o-mixtures: $\ell(G_0|\overline{\mathcal{O}}_{k,c_0}) = 1$, $\mathcal{L}(G_0|\overline{\mathcal{O}}_{k,c_0}) = \{(2,2)\}$.*

(b) *For pathological cases specified by* (A2)*, the following hold:*

(b1) *For e-mixtures: $\ell(G_0|\mathcal{E}_{k_0}) = \infty$, $\mathcal{L}(G_0|\mathcal{E}_{k_0}) = \{(\infty,\infty)\}$.*

(b2) *For o-mixtures:* $\ell(G_0|\overline{\mathcal{O}}_{k,c_0}) = \infty$, $\mathcal{L}(G_0|\overline{\mathcal{O}}_{k,c_0}) = \{(\infty, \infty)\}$.

Part (a) of Proposition 3.17 entails the $n^{-1/2}$ and $n^{-1/4}$ rates of parameter estimation in the generic cases of e-mixture and o-mixture settings, respectively. If true parameter values belong to pathological cases, however, part (b) of the proposition implies that polynomial rates of parameter estimation are not possible, due to infinite singularity level.

Although nontrivial convergence behavior obtained in previous work can also recovered via our notion of singularity levels and indices, with the exception of location-scale Gaussian mixtures, none of these examples exhibits the complex inhomogeneity of the parameter space. To demonstrate the full spectrum of complexity of finite mixture models, we will apply the theory on finite mixtures of skew-normal distributions starting in section 4.

3.5. Construction of minimal forms. We have seen that minimax rates of parameter estimation for finite mixtures can be read off from the singularity structures, via notions of singularity levels and indices, of the parameter space. For the remainder of section 3, we shall present a general procedure or calculating singularity level/index for a given mixing measure.

To this end, one needs to first construct r-minimal forms and κ-minimal forms. Since the latter can be constructed in the same manner as the former, we will focus our presentation on r-minimal forms. A simple way of constructing an r-minimal form is to select a subset of partial derivatives of f taken up to order r such that all these functions are linearly independent. A simple procedure is to start from the smallest order $r = 1$ and then move up to $r = 2, 3, \ldots$ and so on. For each r, assume that we have obtained a linearly independent subset of partial derivatives up to order $r - 1$. Now we go over the ordered list of rth partial derivatives: $\{\partial^{|\kappa|} f/\partial\eta^{\kappa} | \kappa \in \mathbb{N}^d, |\kappa| = r\}$. For each κ such that $|\kappa| = r$, if the partial derivative of f of order κ can be expressed as a linear combination of other partial derivatives among those already selected, then this one is eliminated. The process goes on until we exhaust the list of the partial derivatives.

Example 3.18. Continuing from Example 3.1, suppose that G_0 satisfies (3.1). According to the proof of Lemma 4.1 (cf. [36]), we can choose $\alpha_{4k} \neq 0$, so the partial derivative may be eliminated via the following reduction:

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

Note that this elimination step is valid only for a subset of $G_0 = G(\boldsymbol{p}^0, \boldsymbol{\eta}^0)$ for which (3.1) holds, that is, only if $P_1(\boldsymbol{\eta}^0) = 0$ or $P_2(\boldsymbol{\eta}^0) = 0$, where these polynomials are defined in Lemma 4.1.

Example 3.19. If $f(x|\eta) = f(x|\theta, v, m)$, where $m = 0$, the skew-normal kernel becomes the Gaussian kernel. Due to (3.3), all partial derivatives with respect to both θ and v can be eliminated via the following reduction: for any $\kappa_1, \kappa_2 \in \mathbb{N}$ and for any $j = 1, \ldots, k_0$,

$$\frac{\partial^{\kappa_1+\kappa_2} f(x|\eta_j^0)}{\partial\theta^{\kappa_1} v^{\kappa_2}} = \frac{1}{2^{\kappa_2}} \frac{\partial^{\kappa_1+2\kappa_2} f(x|\eta_j^0)}{\partial\theta^{\kappa_1+2\kappa_2}}.$$

This elimination is valid for all parameter values $(\boldsymbol{p}^0, \boldsymbol{\eta}^0)$ and r-minimal forms for all orders.

Example 3.20. For the skew-normal kernel density $f(x|\eta) = f(x|\theta, v, m)$, (3.2) yields the following reductions: for any $j = 1, \ldots, k_0$, $\eta = (\theta, v, m) = \eta_j^0 = (\theta_j^0, v_j^0, m_j^0)$ such that $m \neq 0$,

$$\frac{\partial^2 f}{\partial \theta^2} = 2\frac{\partial f}{\partial v} - \frac{m^3 + m}{v}\frac{\partial f}{\partial m}, \tag{3.17}$$

$$\frac{\partial^2 f}{\partial v \partial m} = -\frac{1}{v}\frac{\partial f}{\partial m} - \frac{m^2 + 1}{2vm}\frac{\partial^2 f}{\partial m^2}. \tag{3.18}$$

Differentiating results in a ripple effect on subsequent eliminations at higher orders. For example, partial derivatives up to the third order of f evaluated at $\eta = \eta_j^0 = (\theta_j^0, v_j^0, m_j^0)$ for any $j = 1, \ldots, k_0$, where $m_j^0 \neq 0$, can be expressed as follows:

$$\begin{aligned}
\frac{\partial^3 f}{\partial \theta^3} &= 2\frac{\partial^2 f}{\partial \theta \partial v} - \frac{m^3 + m}{v}\frac{\partial^2 f}{\partial \theta \partial m}, \\
\frac{\partial^3 f}{\partial \theta^2 \partial v} &= 2\frac{\partial^2 f}{\partial v^2} + \frac{m^3 + m}{v^2}\frac{\partial f}{\partial m} - \frac{m^3 + m}{v}\frac{\partial^2 f}{\partial v \partial m}, \\
\frac{\partial^3 f}{\partial \theta^2 \partial m} &= 2\frac{\partial^2 f}{\partial v \partial m} - \frac{3m^2 + 1}{v}\frac{\partial f}{\partial m} - \frac{m^3 + m}{v}\frac{\partial^2 f}{\partial m^2},
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^3 f}{\partial v \partial m^2} &= -\frac{m^2 + 1}{2vm}\frac{\partial^3 f}{\partial m^3} - \frac{3m^2 - 1}{2vm^2}\frac{\partial^2 f}{\partial m^2}, \\
\frac{\partial^3 f}{\partial v^2 \partial m} &= -\frac{2}{v}\frac{\partial^2 f}{\partial v \partial m} - \frac{m^2 + 1}{2vm}\frac{\partial^3 f}{\partial v \partial m^2} \\
&= \frac{(m^2 + 1)^2}{4v^2 m^2}\frac{\partial^3 f}{\partial m^3} + \frac{(m^2 + 1)(7m^2 - 1)}{4m^3 v^2}\frac{\partial^2 f}{\partial m^2} + \frac{2}{v^2}\frac{\partial f}{\partial m}, \\
\frac{\partial^3 f}{\partial \theta \partial v \partial m} &= -\frac{m^2 + 1}{2vm}\frac{\partial^3 f}{\partial \theta \partial m^2} - \frac{1}{v}\frac{\partial^2 f}{\partial \theta \partial m}.
\end{aligned} \tag{3.19}$$

All three examples above demonstrate how the dependence among partial derivatives of kernel density f, among different orders κ, and among those evaluated at different components i, has a deep impact on the representation of r-minimal forms.

In general, r-minimal form (3.10) may be expressed somewhat more explicitly as follows:

$$\frac{p_G(x) - p_{G_0}(x)}{W_r^r(G, G_0)} = \sum_{(i,\kappa) \in \mathcal{I}, \mathcal{K}} \frac{\xi_{i,\kappa}^{(r)}(G)}{W_r^r(G_0, G)} H_{i,\kappa}^{(r)}(x|G_0) + \sum_{i=1}^{k_0} \frac{\zeta_i^{(r)}(G)}{W_r^r(G_0, G)} f(x|\eta_i^0) + o(1),$$

where $\mathcal{I} \subset \{1, \ldots, k_0\}$ and $\mathcal{K} \subset \mathbb{N}^d$ of elements κ such that $|\kappa| \leq r$. It is emphasized that the sets $\mathcal{I}$ and $\mathcal{K}$ are specific to a particular r-minimal form under consideration. $H_{i,\kappa}^{(r)}$ are a collection of linearly independent partial derivatives of f that are also independent of all functions $f(x|\eta_i^0)$. $H_{i,\kappa}^{(r)}$ are taken from the collection of partial derivatives with order at most

r. Moreover, $\xi_{i,\kappa}^{(r)}$ and $\zeta_i^{(r)}$ take the following polynomial forms:

$$\xi_{i,\kappa}^{(r)}(G) = \sum_{j=1}^{s_i} \frac{p_{ij}(\Delta\eta_{ij})^{\kappa}}{\kappa!} + \sum_{i',\kappa'} \beta_{i,\kappa,i',\kappa'}(G_0) \sum_{j=1}^{s_{i'}} \frac{p_{ij}(\Delta\eta_{ij})^{\kappa'}}{\kappa'!}, \tag{3.20}$$

$$\zeta_i^{(r)}(G) = \Delta p_{i\cdot} + \sum_{i',\kappa'} \gamma_{i,i',\kappa'}(G_0) \sum_{j=1}^{s_{i'}} \frac{p_{ij}(\Delta\eta_{ij})^{\kappa'}}{\kappa'!}. \tag{3.21}$$

In the right-hand side of each of the last two equations, i' is taken from a subset of $\{1, \ldots, k_0\}$ and κ' is from a subset of $\mathbb{N}^d$ such that $|\kappa| \leq |\kappa'| \leq r$. The actual details of these subsets depend on the specific elimination scheme leading to the r-minimal form. Likewise, the nonzero coefficients $\beta_{i,\kappa,i',\kappa'}(G_0)$ and $\gamma_{i,\kappa,i',\kappa'}(G_0)$ arise from the specific elimination scheme. We include argument G_0 in these coefficients to highlight the fact that they may be dependent on the atoms of G_0 (cf. Examples 3.18 and 3.20).

By the definition of r-singularity for any $r \geq 1$, G_0 is r-singular relative to $\mathcal{G}$ if there exists a sequence of G tending to G_0 in the ambient space $\mathcal{G}$ such that the sequences of semipolynomial fractions, namely $\xi_{i,\kappa}^{(r)}(G)/W_r^r(G, G_0)$ and $\zeta_i^{(r)}(G)/W_r^r(G, G_0)$ (whose numerators are given by (3.20) and (3.21)), must vanish. As a consequence, the question of r-singularity for a given element G_0 is determined by the limiting behavior of a finite collection of infinite sequences of semipolynomial fractions.

3.6. Polynomial limits of minimal form coefficients. The limiting behavior of the semipolynomial fractions described above is independent of a particular choice of the r-minimal form. Indeed, in part (a) of Lemma 3.12, we established an invariance property of the r-singularity, which does not depend on a specific form of the r-minimal form. Let us restrict the basis functions to be members of the collection of all partial derivatives of f up to order r. In the proof of part (b) of Lemma 3.12, it was shown that the coefficients $\xi_l^{(r)}(G)$ have to be elements of a set of polynomials of $\Delta\eta_{ij}$, $\Delta p_{i\cdot}$, and p_{ij}, which are closed under linear combinations of its elements. Let us denote this set by $\mathcal{P}(G, G_0)$, which is invariant with respect to any specific choice of the basis functions (from the collection of partial derivatives) for the r-minimal form. Moreover, G_0 is r-singular if and only if a sequence of G tending to G_0 in W_r can be constructed such that for any element $\xi_l^{(r)}(G) \in \mathcal{P}(G, G_0)$, we have $\xi_l^{(r)}(G)/W_r^r(G, G_0) \to 0$. Equivalently,

$$\xi_l^{(r)}(G)/D_r(G, G_0) \to 0 \quad \text{for all } \xi_l^r(G) \in \mathcal{P}(G, G_0). \tag{3.22}$$

Extracting the limits of a single multivariate semipolynomial fraction (a.k.a. rational semipolynomial functions) is quite challenging in general, due to the interaction among multiple variables involved [63]. Analyzing the limits of not one but a collection of multivariate rational semipolynomials is considerably more difficult. To obtain meaningful and concrete results, we need to consider specific systems of multivariate rational semipolynomials that arise from the r-minimal form.

In the remainder of this paper, we will proceed to do just that. By working with specific choices of kernel density f, it will be shown that under the compactness of the parameter spaces, one can extract a subset of limits from the system of rational semipolynomials

$\xi_l^{(r)}(G)/D_r(G, G_0)$. These limits are expressed as a system of polynomials admitting nontrivial solutions. For a given $r \geq 1$, if the extracted system of polynomial limits does not contain admissible solutions, then it means that there does not exist any sequence of mixing measures G for which a valid r-minimal form can be constructed, because (3.22) is not fulfilled. This would entail the upper bound $\ell(G_0|\mathcal{G}) < r$. On the other hand, if the extracted system of polynomial limits does contain at least one admissible solution, this is a hint that the r-singularity level of G_0 relative to the ambient space G *might* hold. Whether this is actually the case or not requires an explicit construction of a sequence of $G \in \mathcal{G}$ (often building upon the admissible solutions of the polynomial limits) and then the verification that condition (3.22) indeed holds. For the verification purpose, it is sufficient (and simpler) to work with a specific choice of r-minimal form, as Definition 3.3 allows.

The foregoing description, along with the presentation in the previous subsection on the construction of canonical forms, provides the outline of a general procedure which links the determination of the singularity structure of parameter space to the solvability of a system of polynomial limits. This procedure will be illustrated carefully in section 4 for the remarkable world of mixtures of skew-normal distributions.

4. O-mixtures of skew-normal distributions. In this section, we provide an illustration of the general theory by presenting some results for skew-normal mixtures. Our motivation is two-fold: First, as discussed in the introduction, skew-normal mixture models are widely embraced in applications despite little or no known theoretical results, so understanding their theoretical properties is of interest in their own right. Second, skew-normal mixtures appear to be an ideal illustration for the theory and tools developed in the previous section, which helps to shed some light on the remarkably complex structure of a mixture distribution's parameter space. On the flip side, our presentation of such examples will be necessarily more technical. To alleviate the technicality and due to space constraints, we focus the presentation only on singularity structures in the overfitted setting subject to certain restrictions on probability mass and other parameters.[1] This case is quite interesting, because it illustrates the full power of the general method of analysis that was described in section 3 in a concrete fashion, yielding a general result while revealing sufficiently complex structures. Readers interested in finer and more complete results of skew-normal mixtures are invited to consult the technical report [36].

Lemma 4.1. *For skew-normal density kernel $f(x|\boldsymbol{\eta})$, the collection of $\{\partial^\kappa f/\partial\eta^\kappa(x|\eta_j)|j = 1, \ldots, k_0; 0 \leq |\kappa| \leq 1\}$ is not linearly independent if and only if $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$ are the zeros of either polynomial P_1 or P_2, which are defined as follows:*

Type A: $P_1(\boldsymbol{\eta}) = \prod_{j=1}^{k_0} m_j$.

Type B: $P_2(\boldsymbol{\eta}) = \prod_{1\leq i\neq j\leq k_0}\{(\theta_i - \theta_j)^2 + [\sigma_i^2(1 + m_j^2) - \sigma_j^2(1 + m_i^2)]^2\}$.

This lemma leads us to consider

$$\mathcal{S}_0 := \left\{ G = G(\boldsymbol{p}, \boldsymbol{\eta}) \,\middle|\, (\boldsymbol{p}, \boldsymbol{\eta}) \in \Omega, P_1(\boldsymbol{\eta}) \neq 0, P_2(\boldsymbol{\eta}) \neq 0 \right\}. \tag{4.1}$$

[1] Specifically, consider $G_0 \in \mathcal{E}_{k_0}$ relative to ambient space $\mathcal{O}_{k,c_0}$ for some $k > k_0$ and small constant $c_0 > 0$ where $O_{k,c_0} \subset \mathcal{O}_k$ contains only (discrete) probability measures whose point masses are bounded from below by c_0. Moreover, we investigate the singularity structure of $G_0 \in \mathcal{S}_0$, a subset to be defined by (4.1).

In o-mixtures, we will see that $\ell(G_0|\mathcal{O}_{k,c_0})$ and $\mathcal{L}(G_0|\mathcal{O}_{k,c_0})$ may grow with $k-k_0$, the number of extra mixing components. The main exercise is to arrive at suitable r-minimal and κ-minimal forms, for which the behavior of their coefficients can be analyzed. Section 3.5 describes a general strategy for the construction of an r-minimal form (or, equivalently, (r,r,r)-minimal form) based on the partial derivatives of the density kernel f with respect to the parameters $\eta=(\theta,v,m)$ up to order r. This is also a strategy that we would like to utilize for κ-minimal forms for any $\kappa\in\mathbb{N}^3$.

For skew-normal kernel density f, the following lemma provides an explicit form for reducing a partial derivative of f to other partial derivatives of lower orders.

Lemma 4.2. *For any $r\geq 1$, denote*

$$\begin{aligned}
A_1^r &= \{(\alpha_1,\alpha_2,\alpha_3):\ \alpha_1\leq 1,\ \alpha_3=0,\ \textit{and}\ |\alpha|\leq r\},\\
A_2^r &= \{(\alpha_1,\alpha_2,\alpha_3):\ \alpha_1\leq 1,\ \alpha_2=0,\alpha_3\geq 1,\ \textit{and}\ |\alpha|\leq r\},\\
\mathcal{F}_r &= A_1^r\cup A_2^r.
\end{aligned}$$

Let $f(x|\eta)=f(x|\theta,v,m)$ denote the skew-normal kernel. Then, for any $\alpha=(\alpha_1,\alpha_2,\alpha_3)\in\mathbb{N}^3$ and $m\neq 0$, there holds that

$$\frac{\partial^{|\alpha|}f}{\partial\theta^{\alpha_1}\partial v^{\alpha_2}\partial m^{\alpha_3}}=\sum\frac{P^{\kappa_1,\kappa_2,\kappa_3}_{\alpha_1,\alpha_2,\alpha_3}(m)}{H^{\kappa_1,\kappa_2,\kappa_3}_{\alpha_1,\alpha_2,\alpha_3}(m)Q^{\kappa_1,\kappa_2,\kappa_3}_{\alpha_1,\alpha_2,\alpha_3}(v)}\frac{\partial^{|\kappa|}f}{\partial\theta^{\kappa_1}\partial v^{\kappa_2}\partial m^{\kappa_3}},$$

where κ in the above sum satisfies $\kappa\in\mathcal{F}_{|\alpha|}$ and $\kappa_1+2\kappa_2+2\kappa_3\leq\alpha_1+2\alpha_2+2\alpha_3$. Additionally, $P^{\kappa_1,\kappa_2,\kappa_3}_{\alpha_1,\alpha_2,\alpha_3}(m)$, $H^{\kappa_1,\kappa_2,\kappa_3}_{\alpha_1,\alpha_2,\alpha_3}(m)$, and $Q^{\kappa_1,\kappa_2,\kappa_3}_{\alpha_1,\alpha_2,\alpha_3}(v)$ are polynomials in terms of m,m, and v, respectively.

Next, we show that the partial derivatives on the right-hand side of the above identity are in fact linearly independent, under additional assumptions on G_0. In particular, following the notation from Lemma 4.2, if $G_0\in\mathcal{S}_0$, then for any $r\geq 1$, the collection of partial derivatives of the skew-normal density kernel $f(x|\eta)$, namely

$$\left\{\frac{\partial^{|\kappa|}f(x|\eta)}{\partial\theta^{\kappa_1}\partial v^{\kappa_2}\partial m^{\kappa_3}}\,\middle|\,\kappa=(\kappa_1,\kappa_2,\kappa_3)\in\mathcal{F}_r,\ \eta=\eta_1^0,\ldots,\eta_{k_0}^0\right\},\tag{4.2}$$

is linearly independent. These relations allow us to obtain suitable minimal forms for the mixture densities of skew-normals.

4.1. A general theorem for skew-normal o-mixtures. In this section, we shall present results on singularity structures of G_0 for the general case $k>k_0$. To do so, we define the system of the limiting polynomials that characterizes both the singularity level and the singularity index of G_0. Recall the notation introduced by the statement of Lemma 4.2, where $P^{\kappa_1,\kappa_2,\kappa_3}_{\alpha_1,\alpha_2,\alpha_3}(m)$, $H^{\kappa_1,\kappa_2,\kappa_3}_{\alpha_1,\alpha_2,\alpha_3}(m)$, and $Q^{\kappa_1,\kappa_2,\kappa_3}_{\alpha_1,\alpha_2,\alpha_3}(v)$ are polynomials in terms of m,m, and v, respectively, that arise in the decomposition of partial derivatives of the skew-normal kernel function.

For given $r\geq 1$ and for each $i=1,\ldots,k_0$, the system of limiting polynomials is given by

the following equations of real unknowns $(a_j, b_j, c_j, d_j)_{j=1}^{k-k_0+1}$:

$$\left\{ \sum_{j=1}^{k-k_0+1} \sum_{\alpha} \frac{P^{\beta_1,\beta_2,\beta_3}_{\alpha_1,\alpha_2,\alpha_3}(m_i^0)}{H^{\beta_1,\beta_2,\beta_3}_{\alpha_1,\alpha_2,\alpha_3}(m_i^0) Q^{\beta_1,\beta_2,\beta_3}_{\alpha_1,\alpha_2,\alpha_3}(v_i^0)} \frac{d_j^2 a_j^{\alpha_1} b_j^{\alpha_2} c_j^{\alpha_3}}{\alpha_1!\alpha_2!\alpha_3!} = 0 : \right.$$

$$\left. \beta \in \mathcal{F}_r \cap \{\beta_1 + 2\beta_2 + 2\beta_3 \leq r\} \right\}, \tag{4.3}$$

where the range of $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in \mathbb{N}^3$ in the above sum satisfies $\alpha_1 + 2\alpha_2 + 2\alpha_3 = \beta_1 + 2\beta_2 + 2\beta_3$.

There are $2r - 1$ equations in the above system of $4(k - k_0 + 1)$ unknowns. A solution of (4.3) is considered *nontrivial* if all of d_j are nonzeros while at least one among $a_1, \ldots, a_i, b_1, \ldots, b_i, c_1, \ldots, c_i$ is nonzero. We say that system (4.3) is unsolvable if it does not have any nontrivial (or admissible) solutions. Note that increasing r makes the system more constrained. The main result of this section is the following.

Theorem 4.3. *For each $i = 1, \ldots, k_0$, let $\rho(v_i^0, m_i^0, k - k_0)$ be the minimum r for which system of polynomial equations (4.3) does not admit nontrivial solutions. Let $G_0 \in \mathcal{S}_0$ and*

$$\mathrm{R}(G_0, k) := \max_{1 \leq i \leq k_0} \rho(v_i^0, m_i^0, k - k_0). \tag{4.4}$$

(i) *Then $\ell(G_0|\mathcal{O}_{k,c_0}) \leq \mathrm{R}(G_0, k) - 1$.*
(ii) *Moreover, there exists $\kappa \in \mathcal{L}(G_0|\mathcal{O}_{k,c_0})$ such that*

$$\kappa \preceq \begin{cases} \left(\mathrm{R}(G_0,k), \dfrac{\mathrm{R}(G_0,k)}{2}, \dfrac{\mathrm{R}(G_0,k)}{2} \right) & \text{if } \mathrm{R}(G_0,k) \text{ is an even number,} \\ \left(\mathrm{R}(G_0,k), \dfrac{\mathrm{R}(G_0,k)+1}{2}, \dfrac{\mathrm{R}(G_0,k)+1}{2} \right) & \text{if } \mathrm{R}(G_0,k) \text{ is an odd number.} \end{cases}$$

Remark. We make the following comments regarding the results of Theorem 4.3:
(i) If $k - k_0 = 1$, we can obtain $\mathrm{R}(G_0, k) = 4$ from the examples given in section SM1.2.1 in Appendix A of the supplementary materials (although in the examples we only worked out the case that $k_0 = 1$; for general $k_0 \geq 1$, the techniques are the same). Since $(4, 2, 2)$ is the unique singularity index of G_0, the bounds with the singularity level index are tight.
(ii) Since the index component for the location parameter dominates that of shape and scale parameters $(4 > 2)$, estimating shape-scale parameters may be more efficient than estimating location parameters in skew-normal o-mixtures.
(iii) In order to determine $\mathrm{R}(G_0, k)$, we need to find the value of $\rho(v_i^0, m_i^0, k - k_0)$ for all $1 \leq i \leq k_0$. One may ask whether the value of $\rho(v_i^0, m_i^0, k - k_0)$ depends on the specific values of v_i^0, m_i^0. The structure of $\rho(v_i^0, m_i^0, k - k_0)$ will be looked at in more detail in subsection SM1.3 in Appendix A of the supplementary materials.

5. Discussion and concluding remarks. This paper focuses on the mathematics of finite mixture models, introducing in particular a general theory for the identification of singularity structures arising from this popular class of statistical models. It is shown that the singularity

structures of the model's parameter space directly determine minimax lower bounds and MLE convergence rates, under conditions on the compactness of the parameter space.

Understanding the behavior of parameter estimates of mixture models is useful, because the mixing parameters represent explicitly the heterogeneity of the underlying data population for which mixture models are most suitable. The systematic identification of singularity structures and the implications on parameter estimation is only a crucial step toward the development of more efficient model-based inference procedures. It is our view that such procedures must account for the presence of singular points residing in the parameter space of the model.

As a matter of fact, there are quite a few examples of such efforts applied to specific statistical models, even if the picture of the singularity structures associating with those models might not have been discussed explicitly. This raises a question of whether or not it is possible to extend and generalize such techniques in order to address the presence of singularities in a direct fashion. We give several examples:

(1) For overfitted mixture models, methods based on likelihood-based penalization techniques were shown to be quite effective (see, e.g., [57, 12, 20]). Our work shows that parameter values residing in the vicinity of regions of high singularity levels should be hard to estimate efficiently. Can a penalization technique be generalized to regularize the estimates toward subsets containing singularity points of lower levels?

(2) Suitable choices of Bayesian prior have been proposed to induce favorable posterior contraction behavior for overfitted finite mixtures [54]. Can we develop an appropriate prior for the mixture model parameters, given our knowledge of singular points residing in the parameter space?

(3) Reparameterization is an effective technique that can be employed to combat singularities present in the class of skewed distributions [31]. It would be interesting to study whether such a reparameterization technique can be systematically developed for mixture models as well.

Acknowledgments. The authors would like to thank Michael I. Jordan, Antonio Lijoi, Judith Rousseau, Yaácov Ritov, Martin J. Wainwright, Larry Wasserman, and Bin Yu for valuable discussions related to this work.

REFERENCES

[1] E. S. Allman, C. Matias, and J. A. Rhodes, *Identifiability of parameters in latent structure models with many observed variables*, Ann. Statist., 37 (2009), pp. 3099–3132.

[2] M. Aoyagi, *A Bayesian learning coefficient of generalization error and Vandermonde matrix-type singularities*, Comm. Statist. Theory Methods, 39 (2010), pp. 2667–2687.

[3] R. B. Arellano-Valle, L. M. Castro, M. C. Genton, and H. W. Gómez, *Bayesian inference for shape mixtures of skewed distributions, with application to regression analysis*, Bayesian Anal., 3 (2008), pp. 513–540.

[4] R. B. Arellano-Valle, M. C. Genton, and R. H. Loschi, *Shape mixtures of multivariate skew-normal distributions*, J. Multivariate Anal., 100 (2009), pp. 91–101.

[5] A. Azzalini, *Further results on a class of distributions which includes the normal ones*, Statistica (Bologna), 46 (1986), pp. 199–208.

[6] A. Azzalini and A. Capitanio, *Statistical applications of the multivariate skew-normal distribution*, J.

R. Stat. Soc. Ser. B. Stat. Methodol., 61 (1999), pp. 579–602.
[7] A. Azzalini and A. D. Valle, *The multivariate skew-normal distribution*, Biometrika, 83 (1996), pp. 715–726.
[8] P. Bühlmann and S. van de Geer, *Statistics for High Dimensional Data*, Springer, Heidelberg, 2011.
[9] A. Canale and B. Scarpa, *Bayesian nonparametric location-scale-shape mixtures*, TEST, 25 (2016), pp. 113–130.
[10] H. Chen and J. Chen, *Tests for homogeneity in normal mixtures in the presence of a structural parameter*, Statist. Sinica, 13 (2003), pp. 351–365.
[11] H. Chen, J. Chen, and J. D. Kalbfleisch, *A modified likelihood ratio test for homogeneity in finite mixture models*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 63 (2001), pp. 19–29.
[12] J. Chen, *Consistency of the MLE under mixture models*, Statist. Sci., 32 (2017), pp. 47–63.
[13] J. H. Chen, *Optimal rate of convergence for finite mixture models*, Ann. Statist., 23 (1995), pp. 221–233.
[14] M. Chiogna, *A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution*, Stat. Methods Appl., 14 (2005), pp. 331–341.
[15] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*, Springer, New York, 2007.
[16] D. Dacunha-Castelle and E. Gassiat, *Testing in locally conic models and application to mixture models*, ESAIM Probab. Statist., 1 (1997), pp. 285–317.
[17] D. Dacunha-Castelle and E. Gassiat, *Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes*, Ann. Statist., 27 (1999), pp. 1178–1209.
[18] A. DasGupta, *Asymptotic Theory of Statistics and Probability*, Springer, New York, 2008.
[19] M. Drton, *Likelihood ratio tests and singularities*, Ann. Statist., 37 (2009), pp. 979–1012.
[20] M. Drton and M. Plummer, *A Bayesian information criterion for singular models*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 79 (2017), pp. 323–380.
[21] M. Drton, B. Sturmfels, and S. Sullivant, *Algebraic factor analysis: Tetrads, pentads and beyond*, Probab. Theory Related Fields, 138 (2007), pp. 463–493.
[22] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on Algebraic Statistics*, Birkhäuser, Basel, 2009.
[23] M. Drton and S. Sullivant, *Algebraic statistical models*, Statist. Sinica, 17 (2007), pp. 1273–1297.
[24] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu, *Singularity, Misspecification, and the Convergence Rate of EM*, preprint, https://arxiv.org/abs/1810.00828, 2018.
[25] R. Dwivedi, N. Ho, K. Khamaru, M. J. Wainwright, M. I. Jordan, and B. Yu, *Challenges with EM in Application to Weakly Identifiable Mixture Models*, preprint, https://arxiv.org/abs/1902.00194, 2019.
[26] E. Gassiat and R. V. Handel, *The local geometry of finite mixtures*, Trans. Amer. Math. Soc., 366 (2014), pp. 1047–1072.
[27] D. Geiger, D. Heckerman, H. King, and C. Meek, *Stratified exponential families: Graphical models and model selection*, Ann. Statist., 29 (2001), pp. 505–529.
[28] S. Ghosal and A. Roy, *Predicting false discovery proportion under dependence*, J. Amer. Statist. Assoc., 106 (2011), pp. 1208–1217.
[29] S. Ghosal and A. van der Vaart, *Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities*, Ann. Statist., 29 (2001), pp. 1233–1263.
[30] M. Hallin and C. Ley, *Skew-symmetric distributions and Fisher information - a tale of two densities*, Bernoulli, 18 (2012), pp. 747–763.
[31] M. Hallin and C. Ley, *Skew-symmetric distributions and Fisher information: The double sin of skew-normal*, Bernoulli, 20 (2014), pp. 1432–1453.
[32] T. Hastie, R. Tibshrani, and M. J. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, Boca Raton, FL, 2015.
[33] P. Heinrich and J. Kahn, *Optimal Rates for Finite Mixture Estimation*, preprint, https://arxiv.org/abs/1507.04313, 2015.
[34] N. Ho and X. Nguyen, *Convergence rates of parameter estimation for some weakly identifiable finite mixtures*, Ann. Statist., 44 (2016), pp. 2726–2755.
[35] N. Ho and X. Nguyen, *On strong identifiability and convergence rates of parameter estimation in finite mixtures*, Electron. J. Stat., 10 (2016), pp. 271–307.
[36] N. Ho and X. Nguyen, *Singularity Structures and Impacts on Parameter Estimation in Finite Mixtures*

of Distributions, preprint, https://arxiv.org/abs/1609.02655, 2016.
[37] H. Holzmann, A. Munk, and T. Gneiting, *Identifiability of finite mixtures of elliptical distributions*, Scand. J. Stat., 33 (2006), pp. 753–763.
[38] H. Kasahara and K. Shimotsu, *Testing the number of components in normal mixture regression models*, J. Amer. Statist. Assoc., 110 (2015), pp. 1632–1645.
[39] N. M. Kiefer, *A Remark on the Parameterization of a Model for Heterogeneity*, Working paper 278, Department of Economics, Cornell University, Ithaca, NY, 1982.
[40] K. Kubjas, E. Robeva, and B. Sturmfels, *Fixed points of the EM algorithm and nonnegative rank boundaries*, Ann. Statist., 43 (2015), pp. 422–461.
[41] L. F. Lee and A. Chesher, *Specification testing when score test statistics are identically zero*, J. Econometrics, 31 (1986), pp. 33–61.
[42] S. X. Lee and G. J. McLachlan, *On mixtures of skew normal and skew t-distributions*, Adv. Data Anal. Classif., 7 (2013), pp. 241–266.
[43] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, Springer, New York, 1998.
[44] C. Ley and D. Paindaveine, *On the singularity of multivariate skew-symmetric models*, J. Multivariate Anal., 101 (2010), pp. 1434–1444.
[45] T. I. Lin, *Maximum likelihood estimation for multivariate skew normal mixture models*, J. Multivariate Anal., 100 (2009), pp. 257–265.
[46] T. I. Lin, J. C. Lee, and S. Y. Yen, *Finite mixture modelling using the skew normal distribution*, Statist. Sinica, 17 (2007), pp. 909–927.
[47] B. Lindsay, *Mixture models: Theory, geometry and applications*, in NSF-CBMS Regional Conference Series in Probability and Statistics, IMS, Hayward, CA, 1995.
[48] J. Marin, K. Mengersen, and C. P. Robert, *Bayesian modelling and inference on mixtures of distributions*, in Handbook of Statistics, Vol. 25, Elsevier, Amsterdam, 2005, pp. 459–507.
[49] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs*, Marcel Dekker, New York, 1988.
[50] W. Mou, N. Ho, M. J. Wainwright, P. L. Bartlett, and M. I. Jordan, *Polynomial-Time Algorithm for Power Posterior Sampling in Bayesian Mixture Models*, manuscript, 2019.
[51] X. Nguyen, *Convergence of latent mixing measures in finite and infinite mixture models*, Ann. Statist., 4 (2013), pp. 370–400.
[52] M. O. Prates, C. R. B. Cabral, and V. H. Lachos, *mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions*, J. Stat. Softw., 54 (2013).
[53] A. Rotnitzky, D. R. Cox, M. Bottai, and J. Robins, *Likelihood-based inference with singular information matrix*, Bernoulli, 6 (2000), pp. 243–284.
[54] J. Rousseau and K. Mengersen, *Asymptotic behaviour of the posterior distribution in overfitted mixture models*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 73 (2011), pp. 689–710.
[55] S. W. Schnatter and S. Pyne, *Bayesian inference for finite mixtures of univariate and multivariate skew-normal and skew-t distributions*, Biostatistics, 11 (2009), pp. 317–336.
[56] B. Stumfel, *Solving Systems of Polynomial Equations*, AMS, Providence, RI, 2002.
[57] W. Toussile and E. Gassiat, *Variable selection in model-based clustering using multilocus genotype data*, Adv. Data Anal. Classif., 3 (2009), pp. 109–134.
[58] C. Uhler, *Geometry of maximum likelihood estimation in Gaussian graphical models*, Ann. Statist., 40 (2012), pp. 238–261.
[59] S. van de Geer, *Empirical Processes in M-Estimation*, Cambridge University Press, Cambridge, UK, 2000.
[60] A. W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK, 1998.
[61] C. Villani, *Topics in Optimal Transportation*, AMS, Providence, RI, 2003.
[62] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Cambridge University Press, Cambridge, UK, 2009.
[63] S. Xiao and G. Zeng, *Determination of the limits for multivariate rational functions*, Sci. China Math., 57 (2014), pp. 397–416.
[64] C. B. Zeller, C. R. B. Cabral, and V. H. Lachos, *Robust mixture regression modeling based on scale mixtures of skew-normal distributions*, TEST, 25 (2016), pp. 375–396.