Bernoulli **26**(2), 2020, 828–857
https://doi.org/10.3150/18-BEJ1087

Robust estimation of mixing measures in finite mixture models

NHAT HO[1], XUANLONG NGUYEN[2,*] and YA'ACOV RITOV[2,**]

[1]*Department of EECS, University of California, Berkeley, USA. E-mail: minhnhat@berkeley.edu*
[2]*Department of Statistics, University of Michigan, Ann Arbor, USA.*
*E-mail: *xuanlong@umich.edu; **yritov@umich.edu*

In finite mixture models, apart from underlying mixing measure, true kernel density function of each subpopulation in the data is, in many scenarios, unknown. Perhaps the most popular approach is to choose some kernel functions that we empirically believe our data are generated from and use these kernels to fit our models. Nevertheless, as long as the chosen kernel and the true kernel are different, statistical inference of mixing measure under this setting will be highly unstable. To overcome this challenge, we propose flexible and efficient robust estimators of the mixing measure in these models, which are inspired by the idea of minimum Hellinger distance estimator, model selection criteria, and superefficiency phenomenon. We demonstrate that our estimators consistently recover the true number of components and achieve the optimal convergence rates of parameter estimation under both the well- and misspecified kernel settings for any fixed bandwidth. These desirable asymptotic properties are illustrated via careful simulation studies with both synthetic and real data.

Keywords: convergence rates; Fisher singularities; minimum distance estimator; mixture models; model misspecification; model selection; strong identifiability; superefficiency; Wasserstein distances

1. Introduction

Finite mixture models have long been a popular modeling tool for making inference about the heterogeneity in data, starting, at least, with the classical work of Pearson [32] on biometrical ratios on crabs. They have been used in numerous domains arising from biological, physical, and social sciences. For a comprehensive introduction of statistical inference in mixture models, we refer the readers to the books of McLachlan and Basford [29], Lindsay [26], McLachlan and Peel [28].

In finite mixture models, we have our data $X_1, X_2, \dots, X_n \in \mathcal{X} \subset \mathbb{R}^d$ $(d \geq 1)$ to be i.i.d. observations from a finite mixture with density function

$$f_0(x|G_0) := \int f_0(x|\theta)\,\mathrm{d}G_0(\theta) = \sum_{i=1}^{k_0} p_i^0 f_0\bigl(x|\theta_i^0\bigr),$$

where $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ is a true but *unknown* mixing measure with exactly $k_0 < \infty$ non-zero components and $\{f_0(x|\theta), \theta \in \Theta \subset \mathbb{R}^{d_1}\}$ is a true family of density functions, possibly *partially unknown* where $d_1 \geq 1$. There are essentially three principal challenges to the models that have

1350-7265 © 2020 ISI/BS

attracted a great deal of attention from various researchers. They include estimating the true number of components k_0, understanding the behaviors of parameter estimation, that is, the atoms and weights of true mixing measure G_0, and determining the underlying kernel density function f_0 of each subpopulation in the data. The first topic has been an intense area of research recently, see, for example, Roeder [34], Escobar and West [12], Dacunha-Castelle and Gassiat [8,9], Richardson and Green [33], Keribin [24], James, Priebe and Marchette [20], Chen, Li and Fu [5], Chen and Khalili [4], Kasahara and Shimotsu [23]. However, the second and third topic have received much less attention due to their mathematical difficulty.

When the kernel density function f_0 is assumed to be known and k_0 is bounded by some fixed positive integer number, there have been considerable recent advances in the understanding of parameter estimation. In particular, when k_0 is known, that is, the exact-fitted setting of finite mixtures, Ho and Nguyen [16] introduced a stronger version of classical parameter identifiability condition, which is first order identifiability notion, see Definition 2.2 below, to guarantee the standard convergence rate $n^{-1/2}$ of parameter estimation. When k_0 is unknown and bounded by a given number, that is, the over-fitted setting of finite mixtures, Chen [6], Nguyen [31], Ho and Nguyen [16] utilized a notion of second order identifiability to establish convergence rate $n^{-1/4}$ of parameter estimation, which is achieved under some minimum distance based estimator and the maximum likelihood estimator. Sharp minimax rates of parameter estimation for finite mixtures under strong identifiability conditions in sufficiently high orders have been obtained by Heinrich and Kahn [13]. On the other hand, Ho and Nguyen [14,15] studied the singularity structure of finite mixture's parameter space and its impact on rates of parameter estimation when either the first or the second order identifiability condition fails to hold. When the kernel density function f_0 is unknown, there have been some work utilizing the semiparametric approaches (Bordes, Mottelet and Vandekerkhove [3], Hunter, Wang and Hettmansperger [18]). The salient feature of these work is to estimate f_0 from certain classes of functions with infinite dimension and achieve parameter estimation accordingly. However, it is usually very difficult to establish a strong guarantee for the identifiability of the parameters, even when the parameter space is simple (Hunter, Wang and Hettmansperger [18]). Therefore, semiparametric approaches for estimating true mixing measure G_0 are usually not reliable.

Perhaps, the most common approach to avoid the identifiability issue of f_0 is to choose some kernel function f that we tactically believe the data are generated from, and utilize that kernel function to fit the model to obtain an estimate of the true mixing measure G_0. In view of its simplicity and prevalence, this is also the approach that we consider in this paper. However, there is a fundamental challenge with that approach. It is likely that we are subject to a misspecified kernel setting, that is, the chosen kernel f and the true kernel f_0 are different. Hence, parameter estimation under this approach will be potentially unstable. The robustness question is unavoidable. Our principal goal in the paper therefore, is the construction of robust estimators of G_0 where the estimation of both its number of components and its parameters is of interest. Moreover, these estimators should achieve the best possible convergence rates under various assumptions of both the chosen kernel f and the true kernel f_0. When the true number of components k_0 is known, various robust methods had been proposed in the literature, see, for example, Woodward *et al.* [39], Donoho and Liu [10], Cutler and Cordero-Braña [7]. However, there are scarce work for robust estimators when the true number of components k_0 is unknown. Recently, Woo and Sriram [38] proposed a robust estimator of the number of components based on the idea of minimum Hellinger distance estimator (Beran [2], Lindsay [27], Lin and He [25], Karunamuni and Wu [22]). However, their work faced certain limitations. First, their estimator greatly relied upon the choice of kernel bandwidth. In particular, in order to achieve the consistency of the number of components under the well-specified kernel setting, that is, when $\{f\} = \{f_0\}$, the bandwidth should vanish to 0 sufficiently slowly (cf. Theorem 3.1 in Woo and Sriram [38]). Secondly, the behaviors of parameter estimation from their estimator are difficult to interpret due to the subtle choice of bandwidth. Last but not least, they also argued that their method achieved the robust estimation of the number of components under the misspecified kernel setting, i.e., when $\{f\} \neq \{f_0\}$. Not only did their statement lack theoretical guarantee, their argument turned out to be also erroneous (see Section 5.3 in Woo and Sriram [38]). More specifically, they considered the chosen kernel f to be Gaussian kernel while the true kernel f_0 to be Student's t-kernel with a given fixed degree of freedom. The parameter space Θ only consists of mean and scale parameter while the true number of components k_0 is 2. They demonstrated that their estimator still maintained the correct number of components of G_0, that is, $k_0 = 2$, under that setting of f and f_0. Unfortunately, their argument is not clear as their estimator should maintain the number of components of a mixing measure G_* which minimizes the appropriate Hellinger distance to the true model. Of course, establishing the consistency of their parameter estimation under the misspecified kernel setting is also a non-trivial problem.

Inspired by the idea of minimum Hellinger distance estimator, we propose flexible and efficient robust estimators of mixing measure G_0 that address all the limitations from the estimator in Woo and Sriram [38]. Not only our estimators are computationally feasible and robust but they also possess various desirable properties, such as the flexible choice of bandwidth, the consistency of the number of components, and the best possible convergence rates of parameter estimation. In particular, the main contributions in this paper can be summarized as follows

(i) We treat the well-specified kernel setting, that is, $\{f\} = \{f_0\}$, and the misspecified kernel setting, that is, $\{f\} \neq \{f_0\}$, separately. Under both settings, we achieve the consistency of our estimators regarding the true number of components for any fixed bandwidth. Furthermore, when the bandwidth vanishes to 0 at an appropriate rate, the consistency of estimating the true number of components is also guaranteed.

(ii) For any fixed bandwidth, when f_0 is identifiable in the first order, the optimal convergence rate $n^{-1/2}$ of parameter estimation is established under the well-specified kernel setting. Additionally, when f_0 is not identifiable in the first order, we also demonstrate that our estimators still achieve the best possible convergence rates of parameter estimation.

(iii) Under the misspecified kernel setting, we prove that our estimators converge to a mixing measure G_* that is closest to the true model under the Hellinger metric for any fixed bandwidth. When f is first order identifiable and G_* has finite number of components, the optimal convergence rate $n^{-1/2}$ is also established under mild conditions of both kernels f and f_0. Furthermore, when G_* has infinite number of components, some analyses about the consistency of our estimators are also discussed.

Finally, our argument, so far, has mostly focused on the setting when the true mixing measure G_0 is fixed with the sample size n. However, we note in passing that in a proper asymptotic model, G_0 may also vary with n and converge to some probability distribution in the limit. Under the well-specified kernel setting, we verify that our estimators also achieve the minimax convergence rate of estimating G_0 under sufficiently strong condition on the identifiability of kernel density function f_0.

Paper organization. The rest of the paper is organized as follows. Section 2 provides preliminary backgrounds and facts. Section 3 presents an algorithm to construct a robust estimator of mixing measure based on minimum Hellinger distance estimator idea and model selection perspective. Theoretical results regarding that estimator are treated separately under both the well- and mis-specified kernel setting. Section 4 introduces another algorithm to construct a robust estimator of mixing measure based on superefficiency idea. Section 5 addresses the performance of our estimators developed in previous sections under non-standard setting of our models. The theoretical results are illustrated via careful simulation studies with both synthetic and real data in Section 6. Discussions regarding possible future work are presented in Section 7 while self-contained proofs of key results in the paper are given in Ho, Nguyen and Ritov [17].

Notation. Given two densities p, q (with respect to the Lebesgue measure μ), the total variation distance is given by $V(p,q) = \frac{1}{2}\int |p(x) - q(x)|\, d\mu(x)$. Additionally, the square Hellinger distance is given by $h^2(p,q) = \frac{1}{2}\int (\sqrt{p(x)} - \sqrt{q(x)})^2\, d\mu(x)$.

For any $\kappa = (\kappa_1, \ldots, \kappa_{d_1}) \in \mathbb{N}^{d_1}$, we denote $\frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta) = \frac{\partial^{|\kappa|} f}{\partial \theta_1^{\kappa_1} \cdots \partial \theta_{d_1}^{\kappa_{d_1}}}(x|\theta)$ where $\theta = (\theta_1, \ldots, \theta_{d_1})$. Additionally, the expression $a_n \gtrsim b_n$ will be used to denote the inequality up to a constant multiple where the value of the constant is independent of n. We also denote $a_n \asymp b_n$ if both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. Finally, for any $a, b \in \mathbb{R}$, we denote $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

2. Background

Throughout the paper, we assume that the parameter space Θ is a compact subset of $\mathbb{R}^{d_1}$. For any kernel density function f and mixing measure G, we define

$$f(x|G) := \int f(x|\theta)\, dG(\theta).$$

Additionally, we denote $\mathcal{E}_{k_0} := \mathcal{E}_{k_0}(\Theta)$ the space of discrete mixing measures with exactly k_0 distinct support points on Θ and $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ the space of discrete mixing measures with at most k distinct support points on Θ. Additionally, denote $\mathcal{G} := \mathcal{G}(\Theta) = \bigcup_{k \in \mathbb{N}_+} \mathcal{E}_k$ the set of all discrete measures with finite supports on Θ. Finally, $\overline{\mathcal{G}}$ denotes the space of all discrete measures (including those with countably infinite supports) on Θ.

As described in the introduction, a principal goal of our paper is to construct robust estimators that maintain the consistency of the number of components and the best possible convergence rates of parameter estimation. As in Nguyen [31], our tool-kit for analyzing the identifiability and convergence of parameter estimation in mixture models is based on Wasserstein distance, which can be defined as the optimal cost of moving masses transforming one probability measure to another (Villani [36]). In particular, consider a mixing measure $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$, where $\mathbf{p} = (p_1, p_2, \ldots, p_k)$ denotes the proportion vector. Likewise, let $G' = \sum_{i=1}^{k'} p'_i \delta_{\theta'_i}$. A coupling between $\boldsymbol{p}$ and $\boldsymbol{p'}$ is a joint distribution $\boldsymbol{q}$ on $[1, \ldots, k] \times [1, \ldots, k']$, which is expressed as a matrix $\boldsymbol{q} = (q_{ij})_{1 \le i \le k, 1 \le j \le k'} \in [0, 1]^{k \times k'}$ with margins $\sum_{m=1}^{k} q_{mj} = p'_j$ and $\sum_{m=1}^{k'} q_{im} = p_i$ for any $i = 1, 2, \ldots, k$ and $j = 1, 2, \ldots, k'$. We use $\mathcal{Q}(\boldsymbol{p}, \boldsymbol{p'})$ to denote the space of all such couplings. For any $r \ge 1$, the r-th order Wasserstein distance between G and G' is given by

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

where $\| \cdot \|$ denotes the l_2 norm for elements in $\mathbb{R}^{d_1}$. It is simple to argue that if a sequence of probability measures $G_n \in \mathcal{O}_{k_0}$ converges to $G_0 \in \mathcal{E}_{k_0}$ under the W_r metric at a rate $\omega_n = o(1)$ then there exists a subsequence of G_n such that the set of atoms of G_n converges to the k_0 atoms of G_0, up to a permutation of the atoms, at the same rate ω_n.

We recall now the following key definitions that are used to analyze the behavior of mixing measures in finite mixture models (cf. Ho and Nguyen [15]). We start with the following.

Definition 2.1. We say the family of densities $\{f(x|\theta), \theta \in \Theta\}$ is *uniformly Hölder* up to the order r, for some $r \ge 1$, if f as a function of θ is differentiable up to the order r and its partial derivatives with respect to θ satisfy the following inequality

$$\sum_{|\kappa|=r} \left| \left(\frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta_1) - \frac{\partial^{|\kappa|} f}{\partial \theta^{\kappa}}(x|\theta_2) \right) \gamma^{\kappa} \right| \le C \|\theta_1 - \theta_2\|_r^{\delta} \|\gamma\|_r^r,$$

for any $\gamma \in \mathbb{R}^{d_1}$ and for some positive constants δ and C independent of x and $\theta_1, \theta_2 \in \Theta$. Here, $\gamma^{\kappa} = \prod_{i=1}^{d_1} \gamma_i^{\kappa_i}$ where $\kappa = (\kappa_1, \ldots, \kappa_{d_1})$.

We can verify that many popular classes of density functions, including Gaussian, Student's t, and skewnormal family, satisfy the uniform Hölder condition up to any order $r \ge 1$.

The classical identifiability condition entails that the family of density function $\{f(x|\theta), \theta \in \Theta\}$ is identifiable if for any $G_1, G_2 \in \mathcal{G}$, $f(x|G_1) = f(x|G_2)$ almost surely implies that $G_1 \equiv G_2$ (Teicher [35]). To be able to establish convergence rates of parameters, we have to utilize the following stronger notion of identifiability.

Definition 2.2. For any $r \ge 1$, we say that the family $\{f(x|\theta), \theta \in \Theta\}$ (or in short, $\{f\}$) is *identifiable in the r-th order* if $f(x|\theta)$ is differentiable up to the r-th order in θ and the following holds

A1. For any $k \ge 1$, given k different elements $\theta_1, \ldots, \theta_k \in \Theta$. If we have $\alpha_{\eta}^{(i)}$ such that for almost all x

$$\sum_{l=0}^{r} \sum_{|\eta|=l} \sum_{i=1}^{k} \alpha_{\eta}^{(i)} \frac{\partial^{|\eta|} f}{\partial \theta^{\eta}}(x|\theta_i) = 0$$

then $\alpha_{\eta}^{(i)} = 0$ for all $1 \le i \le k$ and $|\eta| \le r$.

Rationale of the first order identifiability: Throughout the paper, we denote $I(G, f) := E(l_G l_G^T)$ the Fisher information matrix of the kernel density f at a given mixing measure G. Here, $l_G := \frac{\partial}{\partial G} \log f(x|G)$ is the score function, where $\frac{\partial}{\partial G}$ denotes the vector of derivatives with respect to all the components and masses of G. The first order identifiability of f is an equivalent way to say that the Fisher information matrix $I(G, f)$ is non-singular for any G. Now, under the first order identifiability and the first order uniform Hölder condition on f, we have the following result.

Proposition 2.1. *Suppose that the density family $\{f(x|\theta), \theta \in \Theta\}$ is identifiable in the first order and uniformly Hölder up to the first order. Then, there is a positive constant C_0 depending on G_0, Θ, and f such that as long as $G \in \mathcal{O}_{k_0}$ we have*

$$h\big(f(\cdot|G), f(\cdot|G_0)\big) \geq C_0 W_1(G, G_0).$$

The proof of the above result can be found in Appendix C. Note that, the result of Proposition 2.1 is slightly stronger than that of Theorem 3.1 and Corollary 3.1 in Ho and Nguyen [16] as it holds for any $G \in \mathcal{O}_{k_0}$ instead of only for any $G \in \mathcal{E}_{k_0}$ as in these later results. The first order identifiability property of kernel density function f implies that for any estimation method that yields the convergence rate $n^{-1/2}$ for $f(\cdot|G_0)$ under the Hellinger distance, the induced rate of convergence for the mixing measure G_0 is $n^{-1/2}$ under W_1 distance.

3. Minimum Hellinger distance estimator with non-singular Fisher information matrix

Throughout this section, we assume that two families of density functions $\{f_0(x|\theta), \theta \in \Theta\}$ and $\{f(x|\theta), \theta \in \Theta\}$ are identifiable in the first order and admit the uniform Hölder condition up to the first order. Now, let K be any fixed multivariate density function and $K_\sigma(x) = \frac{1}{\sigma^d} K(\frac{x}{\sigma})$ for any $\sigma > 0$. We define

$$f^\sigma(x|\theta) := f * K_\sigma(x|\theta) := \int f(x - y|\theta) K_\sigma(y)\,\mathrm{d}y$$

for any $\theta \in \Theta$. The notation $f * K_\sigma$ can be thought as the convolution of the density family $\{f(x|\theta), \theta \in \Theta\}$ with the kernel function K_σ. From that definition, we further define

$$f^\sigma(x|G) := f * K_\sigma(x|G) := \sum_{i=1}^{k} p_i f * K_\sigma(x|\theta_i) = \sum_{i=1}^{k} p_i \int f(x - y|\theta_i) K_\sigma(y)\,\mathrm{d}y,$$

for any discrete mixing measure $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ in $\overline{\mathcal{G}}$. For the convenience of our argument later, we denote that $f^\sigma(\cdot|G) := f(\cdot|G)$ as long as $\sigma = 0$. Furthermore, we also define that

$$P_n^\sigma(x) := P_n * K_\sigma(x) := \frac{1}{n} \sum_{i=1}^{n} K_\sigma(x - X_i),$$

where $P_n := \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}$ is the empirical measure associated with the sample $X_1,\dots,X_n$. Now, our approach to define a robust estimator of G_0 is inspired by the minimum Hellinger distance estimator (Beran [2]) and the model selection criteria. Indeed, we have the following algorithm.

Algorithm 1. Let $C_n n^{-1/2} \to 0$ as $n \to \infty$.

- Step 1: Determine $\widehat{G}_{n,m} = \arg\min_{G\in\mathcal{O}_m} h(f^{\sigma_1}(\cdot|G), P_n^{\sigma_0}(\cdot))$ for any $m \geq 1$.
- Step 2: Choose

$$\widehat{m}_n = \inf\bigl\{m \geq 1 : h\bigl(f^{\sigma_1}(\cdot|\widehat{G}_{n,m}), P_n^{\sigma_0}(\cdot)\bigr) \leq h\bigl(f^{\sigma_1}(\cdot|\widehat{G}_{n,m+1}), P_n^{\sigma_0}(\cdot)\bigr) + C_n n^{-1/2}\bigr\},$$

- Step 3: Let $\widehat{G}_n = \widehat{G}_{n,\widehat{m}_n}$ for each n.

Note that, $\sigma_1 \geq 0$ and $\sigma_0 > 0$ are two chosen bandwidths that control the amount of smoothness that we would like to add to f and $f_{0,}$ respectively. As the parameter space Θ is a compact subset of $\mathbb{R}^{d_1}$ and the Hellinger metric is continuous with respect to the mixing measure, the existence of optimal solution $\widehat{G}_{n,m}$ is guaranteed for all n, m. Additionally, the choice of C_n in Algorithm 1 is to guarantee that $\widehat{m}_n$ is finite as n is sufficiently large under the well-specified kernel setting and $\sigma_1 = \sigma_0$. It is due to the result that $A = h(f^{\sigma_1}(\cdot|\widehat{G}_{n,m}), P_n^{\sigma_0}(\cdot)) - h(f^{\sigma_1}(\cdot|\widehat{G}_{n,m+1}), P_n^{\sigma_0}(\cdot)) \to 0$ almost surely as $n \to \infty$ when $m \geq k_0$ under these settings (cf. the proof of Theorem 3.1 in Appendix A). Under the misspecified kernel setting, with an appropriate choice of σ_1 and σ_0 (cf. conditions in Theorem 3.2), the current choice of C_n is also sufficient to ensure that $\widehat{m}_n$ is finite as n is sufficiently large, which is also because $A \to 0$ as m is sufficiently large. Furthermore, C_n can be chosen based on certain model selection criteria. For instance, if we use BIC, then $C_n = \sqrt{(d_1+1)\log n/2}$ where d_1 is the dimension of parameter space. Algorithm 1 is in fact the generalization of the algorithm considered in Woo and Sriram [38] when $\sigma_1 = 0$ and $\sigma_0 > 0$. In particular, with the adaptation of notations as those in our paper, the algorithm in Woo and Sriram [38] can be stated as follows.

Woo–Sriram (WS) Algorithm.

- Step 1: Determine $\overline{G}_{n,m} = \arg\min_{G\in\mathcal{O}_m} h(f(\cdot|G), P_n^{\sigma_0}(\cdot))$ for any $n, m \geq 1$.
- Step 2: Choose

$$\overline{m}_n = \inf\bigl\{m \geq 1 : h\bigl(f(\cdot|\overline{G}_{n,m}), P_n^{\sigma_0}(\cdot)\bigr) \leq h\bigl(f(\cdot|\overline{G}_{n,m+1}), P_n^{\sigma_0}(\cdot)\bigr) + C_n' n^{-1/2}\bigr\},$$

  where $C_n' n^{-1/2} \to 0$.
- Step 3: Let $\overline{G}_n = \overline{G}_{n,\overline{m}_n}$ for each n.

The main distinction between our estimator and Woo–Sriram's (WS) estimator is that we also allow the convolution of mixture density $f(\cdot|G)$ with K_{σ_1}. This double convolution trick in Algorithm 1 was also considered in James, Priebe and Marchette [20] to construct the consistent estimation of mixture complexity. However, their work was based on the Kullback–Leibler (KL) divergence rather than the Hellinger distance and was restricted to only the choice that $\sigma_1 = \sigma_0$.

Under the misspecified kernel setting, that is, $\{f\} \neq \{f_0\}$, the estimation of mixing measure G_0 from KL divergence can be highly unstable. Additionally, James, Priebe and Marchette [20] only worked with the Gaussian case of true kernel function f_0, while in many applications, it is not realistic to expect that f_0 is Gaussian. To demonstrate the advantages of our proposed estimator $\widehat{G}_n$ over WS estimator $\overline{G}_n$, we will provide careful theoretical studies of these estimators in the paper. For readers' convenience, we provide now a brief summary of our analyses of the convergence behaviors of $\widehat{G}_n$ and $\overline{G}_n$ as well as our measure of robustness.

Summary of results for well-specified setting. Under the well-specified setting, that is, $\{f\} = \{f_0\}$, the optimal choice of σ_1 and σ_0 in Algorithm 1 is $\sigma_1 = \sigma_0 > 0$, which guarantees that G_0 is the exact mixing measure that we seek for. Now, the double convolution trick in Algorithm 1 is sufficient to yield the optimal convergence rate $n^{-1/2}$ of $\widehat{G}_n$ to G_0 for any fixed bandwidth $\sigma_0 > 0$ (cf. Theorem 3.1). The core idea of this result comes from the fact that $P_n^{\sigma_0}(x)$ is an unbiased estimator of $f_0^{\sigma_0}(x|G_0)$ for all $x \in \mathcal{X}$. It guarantees that $h(P_n^{\sigma_0}(\cdot), f_0^{\sigma_0}(\cdot|G_0)) = O_p(n^{-1/2})$ under suitable conditions of f_0 when the bandwidth σ_0 is fixed. However, it is not the case for WS Algorithm. Indeed, we demonstrate later in Section 3.3 that for any fixed bandwidth $\sigma_0 > 0$, $\overline{G}_n$ converges to $\overline{G}_0$ where $\overline{G}_0 = \arg\min_{G \in \overline{\mathcal{G}}} h(f_0(\cdot|G), f_0^{\sigma_0}(\cdot|G_0))$ under certain conditions of f_0, K, and $\overline{G}_0$. Unfortunately, $\overline{G}_0$ can be very different from G_0 even if they may have the same number of components. Therefore, even though we may still be able to recover the true number of components with WS Algorithm, we hardly can obtain exact estimation of true parameters. It shows that Algorithm 1 is more appealing than WS Algorithm under the well-specified kernel setting with fixed bandwidth $\sigma_0 > 0$.

When we allow the bandwidth σ_0 to vanish to 0 as $n \to \infty$ under the well-specified kernel setting with $\sigma_1 = \sigma_0$, we are able to guarantee that $\widehat{m}_n \to k_0$ almost surely when $n\sigma_0^d \to 0$ (cf. Proposition 3.1). This result is also consistent with the result $\overline{m}_n \to k_0$ almost surely from Theorem 1 in Woo and Sriram [38] under the same assumptions of σ_0. Moreover, under these conditions of bandwidth σ_0, both the estimators $\widehat{G}_n$ and $\overline{G}_n$ converge to G_0 as $n \to \infty$. However, instead of obtaining the exact convergence rate $n^{-1/2}$ of $\widehat{G}_n$ to G_0, we are only able to achieve its convergence rate to be $n^{-1/2}$ up to some logarithmic factor when the bandwidth σ_0 vanishes to 0 sufficiently slowly. It is mainly due to the fact that our current technique is based on the evaluation of the term $h(P_n^{\sigma_0}(\cdot), f_0^{\sigma_0}(\cdot|G_0))$, which may not converge to 0 at the exact rate $n^{-1/2}$ when $\sigma_0 \to 0$. The situation is even worse for the convergence rate of $\overline{G}_n$ to G_0 as it relies not only on the evaluation of $h(P_n^{\sigma_0}(\cdot), f_0^{\sigma_0}(\cdot|G_0))$ but also on the convergence rate of $\overline{G}_0$ to G_0, which depends strictly on the vanishing rate of σ_0 to 0. Therefore, the convergence rate of $\overline{G}_n$ in WS Algorithm may be much slower than $n^{-1/2}$. As a consequence, our estimator in Algorithm 1 may be also more efficient than that in WS Algorithm when the bandwidth σ_0 is allowed to vanish to 0.

Summary of results for misspecified setting (robustness to kernel misspecification). Under the misspecified kernel setting, that is, $\{f\} \neq \{f_0\}$, the double convolution technique in Algorithm 1 continues to be useful for studying the convergence rate of $\widehat{G}_n$ to G_* where we define $G_* = \arg\min_{G \in \overline{\mathcal{G}}} h(f^{\sigma_1}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0))$. Unlike the well-specified kernel setting, we allow σ_1 and σ_0 to be different under the misspecified kernel setting. It is particularly useful if we can choose σ_1 and σ_0 such that two families $\{f^{\sigma_1}(\cdot|\theta)\}$ and $\{f_0^{\sigma_0}(\cdot|\theta)\}$ are identical under Hellinger distance. The consequence is that G_* and G_0 will be identical under Wasserstein distance, which means that our estimator is still able to recover true mixing measure even though we choose the wrong kernel to fit our data. Granted, the misspecified setting means that we are usually *not* in such a fortunate situation, but our theory entails a good performance for our estimate when $f^{\sigma_1} \approx f_0^{\sigma_0}$. Now, for the general choice of σ_1 and σ_0, as long as G_* has finite number of components, we are able to establish the convergence rate $n^{-1/2}$ of $\widehat{G}_n$ to G_* under sufficient conditions on f, f_0, and K (cf. Theorem 3.2). However, when the number of components of G_* is infinite, we are only able to achieve the consistency of the number of components of $\widehat{G}_n$ (cf. Proposition 3.2). Even though we do not have specific result regarding the convergence rate of $\widehat{G}_n$ to G_* under that setting of G_*, we also provide important insights regarding that convergence in Section 3.2.2. Summarizing, Algorithm 1 is robust to kernel misspecification since it yields consistent estimates of the number of components as well as the best possible convergence rates for parameter estimation under various settings of f, f_0, σ_0, and σ_1.

3.1. Well-specified kernel setting

In this section, we consider the setting that f_0 is known, that is, $\{f\} = \{f_0\}$. Under that setting, the optimal choice of σ_1 and σ_0 is $\sigma_1 = \sigma_0 > 0$ to guarantee that G_0 is the exact mixing measure that we estimate. As we have seen from the discussion in Section 2, the first order identifiability condition plays an important role to obtain the convergence rate $n^{-1/2}$ of parameter estimation. Since Algorithm 1 relies on investigating the variation around kernel function $f_0^{\sigma_0}$ in the limit, we would like to guarantee that $f_0^{\sigma_0}$ is identifiable in the first order for any $\sigma_0 > 0$. It appears that we have a mild condition of K such that the first order identifiability of $f_0^{\sigma_0}$ is maintained.

Lemma 3.1. *Assume that $\widehat{K}(t) \neq 0$ for almost all $t \in \mathbb{R}^d$ where $\widehat{K}(t)$ is the Fourier transform of kernel function K. Then, as long as f_0 is identifiable in the first order, we obtain that $f_0^{\sigma_0}$ is identifiable in the first order for any $\sigma_0 > 0$.*

The assumption $\widehat{K}(t) \neq 0$ is very mild. Indeed, popular choices of K to satisfy that assumption include the Gaussian and Student's t kernel. Inspired by the result of Lemma 3.1, we have the following result establishing the convergence rate of $\widehat{G}_n$ to G_0 under W_1 distance for any fixed bandwidth $\sigma_0 > 0$.

Theorem 3.1. *Let $\sigma_0 > 0$ be given.*

(i) *If $f_0^{\sigma_0}$ is identifiable, then $\widehat{m}_n \to k_0$ almost surely.*

(ii) *Assume further the following conditions*

(P.1) *The kernel function K is chosen such that $f_0^{\sigma_0}$ is also identifiable in the first order and admits a uniform Hölder property up to the first order.*

(P.2) $\Psi(G_0, \sigma_0) := \int \frac{g(x|G_0,\sigma_0)}{f_0^{\sigma_0}(x|G_0)}\,\mathrm{d}x < \infty$ *where we have that*

$$g(x|G_0, \sigma_0) := \int K_{\sigma_0}^2(x - y) f_0(y|G_0)\,\mathrm{d}y.$$

Then, we obtain

$$W_1(\widehat{G}_n, G_0) = O_p\left(\sqrt{\frac{\Psi(G_0,\sigma_0)}{C_1^2(\sigma_0)}}n^{-1/2}\right),$$

where $C_1(\sigma_0) := \inf_{G\in\mathcal{O}_{k_0}} \frac{h(f_0^{\sigma_0}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0))}{W_1(G,G_0)}$.

Remarks.

(i) Condition (P.1) is satisfied by many kernel functions K according to Lemma 3.1. By assumption (P.1) and Proposition 2.1, we obtain the following bound

$$h\big(f_0^{\sigma_0}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0)\big) \gtrsim W_1(G, G_0)$$

for any $G \in \mathcal{O}_{k_0}$, that is, $C_1(\sigma_0) > 0$.

(ii) Condition (P.2) is mild. One easy example for such setting is when f_0 and K are both Gaussian kernels. In fact, when $\{f_0(x|\eta,\tau), (\eta,\tau) \in \Theta\}$ is a family of univariate Gaussian distributions where η and τ are location and scale parameter, respectively and K is a standard univariate Gaussian kernel, we achieve

$$\begin{aligned}\Psi(G_0,\sigma_0) &= \sum_{i=1}^{k_0}\int \frac{p_i^0 \int K_{\sigma_0}^2(x-y) f_0(y|\eta_i^0,\tau_i^0)\,\mathrm{d}y}{f_0^{\sigma_0}(x|G_0)}\,\mathrm{d}x \\ &< \sum_{i=1}^{k_0}\int \frac{\int K_{\sigma_0}^2(x-y) f_0(y|\eta_i^0,\tau_i^0)\,\mathrm{d}y}{f_0^{\sigma_0}(x|\eta_i^0,\tau_i^0)}\,\mathrm{d}x \\ &\propto \sum_{i=1}^{k_0}\big((\tau_i^0)^2 + \sigma_0^2\big)/\sigma_0^2 < \infty.\end{aligned}$$

Another specific example is when f_0 and K are both Cauchy kernels or generally Student's t kernels with odd degree of freedom. However, assumption (P.2) may fail when K has much shorter tails than f_0. For example, if f_0 is Laplacian kernel and K is Gaussian kernel, then $\Psi(G_0,\sigma_0) = \infty$.

Comments on $\widehat{G}_n$ *as* $\sigma_0 \to 0$: To avoid the ambiguity, we now denote $\{\sigma_{0,n}\}$ as the sequence of varied bandwidths σ_0. The following result shows the consistency of $\widehat{m}_n$ under specific conditions on $\sigma_{0,n} \to 0$.

Proposition 3.1. *Given a sequence of bandwidths* $\{\sigma_{0,n}\}$ *such that* $\sigma_{0,n} \to 0$ *and* $n\sigma_{0,n}^d \to \infty$ *as* $n \to \infty$. *If* f_0 *is identifiable, then* $\widehat{m}_n \to k_0$ *almost surely.*

Our previous result with Theorem 3.1 shows that the parametric $n^{-1/2}$ rate of convergence of $\widehat{G}_n$ to G_0 is achieved for any fixed $\sigma_0 > 0$. It would be more elegant to argue that this rate is achieved for some sequence $\sigma_{0,n} \to 0$. However, this cannot be done with the current technique employed in the proof of Theorem 3.1. In particular, even though we still can guarantee that $\lim_{\sigma_{0,n}\to 0} C_1(\sigma_{0,n}) > 0$ (cf. Lemma A.2 in Appendix C in Ho, Nguyen and Ritov [17]), the technical difficulty is that $\Psi(G_0, \sigma_{0,n}) = O(\sigma_{0,n}^{-\beta(d)})$ for some $\beta(d) > 0$ depending on d as $\sigma_{0,n} \to 0$. As a consequence, whatever the sequence of bandwidths $\sigma_{0,n} \to 0$ we choose, we will be only able to obtain the convergence rate $n^{-1/2}$ up to the logarithmic term of $\widehat{G}_n$ to G_0. It can be thought as the limitation of the elegant technique employed in Theorem 3.1. We leave the exact convergence rate $n^{-1/2}$ of $\widehat{G}_n$ to G_0 under the setting $\sigma_{0,n} \to 0$ for the future work.

3.2. Misspecified kernel setting

In the previous section, we assume the well-specified kernel setting, that is, $\{f\} = \{f_0\}$, and achieve the convergence rate $n^{-1/2}$ of $\widehat{G}_n$ to G_0 under mild conditions on f_0 and K and the choice that $\sigma_1 = \sigma_0$ for any fixed bandwidth $\sigma_0 > 0$. However, the well-specified kernel assumption is often violated in practice, that is, the chosen kernel f may be different from the true kernel f_0. Motivated by this challenge, in this section we consider the setting when $\{f\} \neq \{f_0\}$. Additionally, we also take into account the case when the chosen bandwidths σ_1 and σ_0 may be different. We will demonstrate that the convergence rate of $\widehat{G}_n$ is still desirable under certain assumptions on f, f_0, and K. Furthermore, we also argue that the choice that σ_1 and σ_0 are different can be very useful under the case when two families of density functions $\{f^{\sigma_1}(x|\theta), \theta \in \Theta\}$ and $\{f_0^{\sigma_0}(x|\theta), \theta \in \Theta\}$ are identical. Due to the complex nature of misspecified kernel setting, we will only study the behavior of $\widehat{G}_n$ when the bandwidth $\sigma_1 \geq 0$ and $\sigma_0 > 0$ are fixed in this section. Now, for fixed bandwidths σ_1, σ_0 assume that there exists a discrete mixing measure G_* that minimizes the Hellinger distance between $f^{\sigma_1}(\cdot|G)$ and $f_0^{\sigma_0}(\cdot|G_0)$, i.e.,

$$G_* := \mathop{\arg\min}_{G \in \overline{\mathcal{G}}} h\bigl(f^{\sigma_1}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0)\bigr).$$

As G_* may not be unique, we denote

$$\mathcal{M} := \bigl\{G_* \in \overline{\mathcal{G}} : G_* \text{ is a minimizer of } h\bigl(f^{\sigma_1}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0)\bigr)\bigr\}.$$

When $f * K_{\sigma_1} = f_0 * K_{\sigma_0}$, it is clear that G_0 is an element of $\mathcal{M}$ such that it has the minimum number of components among all the elements in $\mathcal{M}$. To further investigate $\mathcal{M}$ under general setting of f, f_0, σ_1, σ_0, and K, we start with the following key property of elements G_* in $\mathcal{M}$:

Lemma 3.2. *For any $G \in \overline{\mathcal{G}}$ and $G_* \in \mathcal{M}$, there holds*

$$\int f^{\sigma_1}(x|G)\sqrt{\frac{f_0^{\sigma_0}(x|G_0)}{f^{\sigma_1}(x|G_*)}}\,\mathrm{d}x \leq \int \sqrt{f^{\sigma_1}(x|G_*)}\sqrt{f_0^{\sigma_0}(x|G_0)}\,\mathrm{d}x. \tag{1}$$

Equipped with this bound, we have the following important property of $\mathcal{M}$.

Lemma 3.3. *For any two elements $G_{1,*}, G_{2,*} \in \mathcal{M}$, we obtain $f^{\sigma_1}(x|G_{1,*}) = f^{\sigma_1}(x|G_{2,*})$ for almost surely $x \in \mathcal{X}$.*

Now, we consider the partition of $\mathcal{M}$ into the union of $\mathcal{M}_k = \{G_* \in \mathcal{M} : G_* \text{ has } k \text{ elements}\}$ where $k \in [1, \infty]$. Let $k_* := k_*(\mathcal{M})$ be the minimum number $k \in [1, \infty]$ such that $\mathcal{M}_k$ is non-empty. We divide our argument into two distinct settings of k_*: k_* is finite and k_* is infinite.

3.2.1. *Finite k_**

By Lemma 3.3, $\mathcal{M}_{k_*}$ will have exactly one element G_* provided that f^{σ_1} is identifiable. Furthermore, $\mathcal{M}_k$ is empty for all $k_* < k < \infty$. However, it is possible that $\mathcal{M}_\infty$ still contains various elements. Due to the parsimonious nature of Algorithm 1 and the result of Theorem 3.2, we will be able to demonstrate that $\widehat{G}_n$ still converges to the unique element $G_* \in \mathcal{M}_{k_*}$ at the optimal rate $n^{-1/2}$ regardless of the behavior of $\mathcal{M}_\infty$.

For the simplicity of our later argument under that setting of k_*, we denote by G_* the unique element in $\mathcal{M}_{k_*}$. As we mentioned earlier, one simple example for $k_* < \infty$ is when $\{f^{\sigma_1}\} = \{f_0^{\sigma_0}\}$. Another example is when f is a location-scale family and f_0 is a finite mixture of f while $\sigma_1 = \sigma_0 > 0$. In particular, $f(x|\eta, \tau) = \frac{1}{\tau} f((x - \eta)/\tau)$ where η and τ are location and scale parameters respectively. Additionally, $f_0(x) = \sum_{i=1}^m p_i^* f(x|\eta_i^*, \tau_i^*)$ for some fixed positive integer m and fixed pairwise distinct components $(p_i^*, \eta_i^*, \tau_i^*)$ where $1 \le i \le m$. Under that setting, if we choose $\sigma_1 = \sigma_0$, then we can check that $k_* \le mk_0$ and $f(x|G_*) = f_0(x|G_0)$ almost surely. The explicit formulation of G_*, therefore, can be found from the combinations of G_0 and $(p_i^*, \eta_i^*, \tau_i^*)$ where $1 \le i \le m$.

From inequality (1) in Lemma 3.2, we have the following well-defined weighted version of Hellinger distance.

Definition 3.1. Given $\sigma_1 > 0$. For any two mixing measures $G_1, G_2 \in \overline{\mathcal{G}}$, we define the weighted Hellinger distance $h^*(f^{\sigma_1}(\cdot|G_1), f^{\sigma_1}(\cdot|G_2))$ by

$$\begin{aligned}&\big(h^*\big(f^{\sigma_1}(\cdot|G_1), f^{\sigma_1}(\cdot|G_2)\big)\big)^2\\&\quad = \frac{1}{2}\int \big(\sqrt{f^{\sigma_1}(x|G_1)} - \sqrt{f^{\sigma_1}(x|G_2)}\big)^2 \times \sqrt{\frac{f_0^{\sigma_0}(x|G_0)}{f^{\sigma_1}(x|G_*)}}\,\mathrm{d}x.\end{aligned}$$

The notable feature of h^* is the presence of term $\sqrt{f_0^{\sigma_0}(x|G_0)/f^{\sigma_1}(x|G_*)}$ in its formulation, which makes it different from the traditional Hellinger distance. As long as $\{f\} = \{f_0\}$ and $\sigma_1 = \sigma_0$, we obtain $h^*(f^{\sigma_1}(\cdot|G_1), f^{\sigma_1}(\cdot|G_2)) \equiv h(f^{\sigma_1}(\cdot|G_1), f^{\sigma_1}(\cdot|G_2))$ for any $G_1, G_2 \in \overline{\mathcal{G}}$, that is, the traditional Hellinger distance is a special case of h^* under the well-specified kernel setting and the choice that $\sigma_1 = \sigma_0$. The weighted Hellinger distance h^* is particularly useful for studying the convergence rate of $\widehat{G}_n$ to G_* for any fixed $\sigma_1 \ge 0$ and $\sigma_0 > 0$.

Note that, in the context of the well-specified kernel setting in Section 3.1, the key step that we utilized to obtain the convergence rate $n^{-1/2}$ of $\widehat{G}_n$ to G_0 is based on the lower bound of the Hellinger distance and the first order Wasserstein distance in inequality (3.1). With the modified Hellinger distance h^*, it turns out that we still have the similar kind of lower bound as long as $k_* < \infty$.

Lemma 3.4. *Assume that f^{σ_1} is identifiable in the first order and admits uniform Hölder property up to the first order. If $k_* < \infty$, then for any $G \in \mathcal{O}_{k_*}$ there holds*

$$h^*\big(f^{\sigma_1}(\cdot|G), f^{\sigma_1}(\cdot|G_*)\big) \gtrsim W_1(G, G_*).$$

Equipped with the above inequality, we have the following result regarding the convergence rate of $\widehat{G}_n$ to G_*.

Theorem 3.2. *Assume $k_* < \infty$ for some $\sigma_1 \geq 0$ and $\sigma_0 > 0$.*

(i) *If f^{σ_1} is identifiable, then $\widehat{m}_n \to k_*$ almost surely.*

(ii) *Assume further that condition* (P.2) *in Theorem* 3.1 *holds, that is,* $\Psi(G_0, \sigma_0) < \infty$ *and the following conditions hold*:

(M.1) *The kernel K is chosen such that f^{σ_1} is identifiable in the first order and admits the uniform Hölder property up to the first order.*

(M.2) $\sup_{\theta \in \Theta} \int \sqrt{f^{\sigma_1}(x|\theta)}\, \mathrm{d}x \leq M_1(\sigma_1)$ *for some positive constant* $M_1(\sigma_1)$.

(M.3) $\sup_{\theta \in \Theta} \|\frac{\partial f^{\sigma_1}}{\partial \theta}(x|\theta)/(f^{\sigma_1}(x|\theta))^{3/4}\|_\infty \leq M_2(\sigma_1)$ *for some positive constant* $M_2(\sigma_1)$.

Then, we have

$$W_1(\widehat{G}_n, G_*) = O_p\left(\sqrt{\frac{M^2(\sigma_1)\Psi(G_0, \sigma_0)}{C_{*,1}^4(\sigma_1)}} n^{-1/2}\right),$$

where $C_{*,1}(\sigma_1) := \inf_{G \in \mathcal{O}_{k_*}} \frac{h^*(f^{\sigma_1}(\cdot|G), f^{\sigma_1}(\cdot|G_*))}{W_1(G, G_*)}$ *and* $M(\sigma_1)$ *is some positive constant.*

Remarks.

(i) As being mentioned in Lemma 3.4, condition (M.1) is sufficient to guarantee that $C_{*,1}(\sigma_1) > 0$.

(ii) Conditions (M.2) and (M.3) are mild. An easy example is when f is Gaussian kernel and K is standard Gaussian kernel.

(iii) When f_0 is indeed a finite mixture of f, a close investigation of the proof of Theorem 3.2 reveals that we can relax condition (M.2) and (M.3) for the conclusion of this theorem to hold.

(iv) Under the setting that $\{f^{\sigma_1}\} = \{f_0^{\sigma_0}\}$, that is, $G_* \equiv G_0$, the result of Theorem 3.2 implies that $\widehat{G}_n$ converges to the true mixing measure G_0 at optimal rate $n^{-1/2}$ even though we are under the misspecified kernel setting.

3.2.2. *Infinite k_**

So far, we have assumed that k_* has finite number of support points and achieve the cherished convergence rate $n^{-1/2}$ of $\widehat{G}_n$ to unique element $G_* \in \mathcal{M}_{k_*}$ under certain conditions on f, f_0, and K. It is due to the fact that $\widehat{m}_n \to k_* < \infty$ almost surely, which is eventually a consequence of the identifibility of kernel density function f^{σ_1}. However, for the setting $k_* = \infty$, to establish the consistency of $\widehat{m}_n$, we need to resort to a slightly stronger version of identifiability, which is finitely identifiable condition. We adapt Definition 3 in Nguyen [31] as follows.

Definition 3.2. The family $\{f(x|\theta), \theta \in \Theta\}$ is finitely identifiable if for any $G_1 \in \mathcal{G}$ and $G_2 \in \overline{\mathcal{G}}$, $|f(x|G_1) - f(x|G_2)| = 0$ for almost all $x \in \mathcal{X}$ implies that $G_1 \equiv G_2$.

An example of finite identifiability is when f is Gaussian kernel with both location and variance parameter. Now, a close investigation of the proof of Theorem 3.2 quickly yields the following result.

Proposition 3.2. *Given $\sigma_1 > 0$ such that f^{σ_1} is finitely identifiable. If $k_* = \infty$, we achieve $\widehat{m}_n \to \infty$ almost surely.*

Even though we achieve the consistency result of $\widehat{m}_n$ when $k_* = \infty$, the convergence rate of $\widehat{G}_n$ to G_* still remains an elusive problem. However, an important insight from Proposition 3.2 indicates that the convergence rate of $\widehat{G}_n$ to some element $G_* \in \mathcal{M}_\infty$ may be much slower than $n^{-1/2}$ when $k_* = \infty$. It is due to the fact that both $\widehat{G}_n$ and $G_* \in \mathcal{M}_\infty$ have unbounded numbers of components in which the kind of bound in Lemma 3.4 is no longer sufficient. Instead, something akin to the bounds given in Theorem 2 of Nguyen [31] in the misspecified setting is required. We leave the detailed analyses of $\widehat{G}_n$ under that setting of k_* for the future work.

3.3. Comparison to WS Algorithm

In the previous sections, we have established a careful study regarding the behaviors of $\widehat{G}_n$ in Algorithm 1, that is, we achieved the consistency of the number of components as well as the convergence rates of parameter estimation under various settings of f and f_0 when the bandwidths σ_1 and σ_0 are fixed. As we mentioned at the beginning of Section 3, Algorithm 1 is the generalization of WS Algorithm when $\sigma_1 = 0$ and $\sigma_0 > 0$. Therefore, the general results with estimator $\widehat{G}_n$ in Theorem 3.2 are still applicable to $\overline{G}_n$ under that special case of σ_1 and σ_0. To rigorously demonstrate the flexibilities and advantages of our estimator $\widehat{G}_n$ over WS estimator $\overline{G}_n$, we firstly discuss the behaviors of estimator $\overline{G}_n$ from WS Algorithm under the well-specified kernel setting, that is, $\{f\} = \{f_0\}$, and the fixed bandwidth setting of σ_0. Remember that f_0 is assumed to be identifiable in the first order and to have uniform Hölder property up to the first order. Assume now we can find

$$\overline{G}_0 := \arg\min_{G \in \overline{\mathcal{G}}} h\big(f_0(\cdot|G), f_0^{\sigma_0}(\cdot|G_0)\big),$$

that is, $\overline{G}_0$ is the discrete mixing measure that minimizes the Hellinger distance between $f_0(\cdot|G)$ and $f_0^{\sigma_0}(\cdot|G_0)$. Note that, $\overline{G}_0$ is a special case of G_* when $\{f\} = \{f_0\}$ and $\sigma_1 = 0$. The form of $\overline{G}_0$ can be determined explicitly under various settings of f_0 and K. For instance, assume that f_0 are either univariate Gaussian kernel or Cauchy kernel with parameters $\theta = (\eta, \tau)$ where η and τ are location and variance parameter and K are either standard univariate Gaussian kernel or Cauchy kernel, respectively. Then, a simple calculation shows that $\overline{G}_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{(\theta_i^0, \overline{\tau}_i^0)}$ where $\overline{\tau}_i^0 = \sqrt{(\tau_i^0)^2 + \sigma_0^2}$ for any $1 \le i \le k_0$ and $\sigma_0 > 0$.

As being argued in Section 3.2, $\overline{G}_0$ may have infinite number of components in general; however, for the sake of simplicity, we assume that there exists $\overline{G}_0$ having finite number of components, which is also unique according to the argument in Section 3.2. Under the assumptions of Theorem 3.2 when $\sigma_1 = 0$, we eventually achieve that

$$W_1(\overline{G}_n, \overline{G}_0) = O_p\left(\sqrt{\frac{\overline{M}^2 \Psi(G_0, \sigma_0)}{[\overline{C}]^4}} n^{-1/2}\right),$$

where $\overline{C} := \inf_{G \in \mathcal{O}_{\overline{k}_0}} \frac{h^*(f_0(\cdot|G), f_0(\cdot|\overline{G}_0))}{W_1(G, \overline{G}_0)}$ and $\overline{M}$ is some positive constant. The above result implies that the estimator $\overline{G}_n$ from WS Algorithm will not converge to the true mixing measure G_0 for any fixed bandwith σ_0. It demonstrates that Algorithm 1 is more appealing than WS Algorithm under the well-specified kernel setting with fixed bandwidth $\sigma_0 > 0$. For the setting when the bandwidth σ_0 is allowed to vanish to 0, our result indicates that the convergence rate of $\overline{G}_n$ to G_0 will depend not only on the vanishing rate of the term $\Psi(G_0, \sigma_0)$ to 0 but also on the convergence rate of $\overline{G}_0$ to G_0. Intuitively, to ensure that the convergence of $\overline{G}_n$ to G_0 is $n^{-1/2}$, we also need to achieve that of $\overline{G}_0$ to G_0 to be $n^{-1/2}$. Under the specific case that f_0 and K are univariate Gaussian kernels, the convergence rate of $\overline{G}_0$ to G_0 is $n^{-1/2}$ only when σ_0 goes to 0 at the same rate $n^{-1/2}$. However, it will lead to a strong convergence of $\Psi(G_0, \sigma_0)$ to ∞, which makes the convergence rate of $\overline{G}_n \to G_0$ become much slower than $n^{-1/2}$. Therefore, it is possible that the convergence rate of WS estimator $\overline{G}_n$ to G_0 may be much slower than $n^{-1/2}$ regardless of the choice of bandwidth σ_0. As a consequence, our estimator in Algorithm 1 may also be more efficient than WS estimator under that regime of vanishing bandwidth σ_0.

Under the misspecified kernel setting, we would like to emphasize that our estimator $\widehat{G}_n$ is also more flexible than WS estimator $\overline{G}_n$ as we provide more freedom with the choice of bandwidth σ_1 in Algorithm 1, instead of specifically fixing $\sigma_1 = 0$ as that in WS Algorithm. If there exists $\sigma_1 > 0$ such that $\{f^{\sigma_1}\} = \{f_0^{\sigma_0}\}$, then our estimator $\widehat{G}_n$ will converge to G_0 while WS estimator $\overline{G}_n$ will converge to $\overline{G}_0$ that can be very different from G_0. Therefore, the performance of our estimator is also better than that of WS estimator under that specific misspecified kernel setting.

3.4. Remarks with deconvolution problems

In this section, we would like to take an opportunity for explaining the differences between our setting in the paper and some deconvolution problems in the literature. In particular, we consider the following two setups of deconvolution problems:

(1) (Partial information with noise) In Johannes [21], the author considered the setting that we have i.i.d. samples $Y_1, \ldots, Y_n$ from density function of the form $f_Y = f_X * f_\epsilon$ where f_ϵ is unknown and i.i.d. samples $\epsilon_1, \ldots, \epsilon_m$ from f_ϵ are observed. The goal of such setting is to estimate f_X given partial information about the distribution of ϵ.

(2) (Repeated structure) The next setting is an extension of the above deconvolution problem to the setting with repeated structure of Y. More specifically, the outcomes Y of each individual i satisfy $Y_{ij} = a_i + \epsilon_{ij}$ for $j = 1, \ldots, M$ where M stands for the number of

repeated measures, a_i are i.i.d. draws from some unknown distribution G, and ϵ_{ij} are also i.i.d. draws from some distribution F with zero mean and unit variance. Additionally, a_i and ϵ_{ij} are independent.

Even though the above deconvolution settings may share certain similarity with the setup we consider in the paper, there are certain challenges for extending the results of our paper to these settings. For the clarity of our argument, we would like to summarize the setup considered in our paper. In particular, the current results with well- and misspecified settings in the paper rely on the assumption that data are i.i.d. samples from mixture density of the following form

$$f_0(x|G_0) = \int f_0(x|\theta)\,\mathrm{d}G_0(\theta) = \sum_{i=1}^{k_0} p_i^0 f_0\big(y|\theta_i^0\big),$$

where $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ is a true but unknown discrete mixing measure with exactly $k_0 < \infty$ non-zero components and $\{f_0(y|\theta), \theta \in \Theta \subset \mathbb{R}^{d_1}\}$ is a true family of density functions (not need to be location family), possibly unknown where $d_1 \geq 1$. Since the density family f_0 is generally unknown in practice, we choose some density family f based on prior knowledge with the data and use this family f to fit the model. Governed by the applications such as clustering, our principal goal in the paper is the robustness of the minimum Hellinger distance estimator of G_0, that is, the number of components of G_0 and its parameters.

As being indicated in our setup, the discrete mixing measure G_0 and density family f_0 are both unknown and we do not observe any i.i.d. draws from them. On the other hand, the repeated measure deconvolution setup has data generation via F as well as i.i.d. draws from mixture components via mixing measure G. The additional repeated measures per individual should help in terms of statistical efficiency, but how to quantify this is not clear to us. In particular, there seem to be some technical hurdles as to how to adapt the results in our paper to the repeated measures deconvolution problem. The same technical challenges also hold for partial information with noise setup of deconvolution problem in Johannes [21]. We leave an extension of our results to the aforementioned deconvolution problems for future exploration.

4. Different approach with minimum Hellinger distance estimator

Thus far, we have developed a robust estimator of mixing measure G_0 based on the idea of minimum Hellinger distance estimator and model selection criteria. That estimator is shown to attain various desirable properties, including the consistency of number of components $\widehat{m}_n$ and the optimal convergence rates of $\widehat{G}_n$. In this section, we take a rather different approach of constructing such robust estimator. In fact, we have the following algorithm:

Algorithm 2.

- Step 1: Determine $\widehat{G}_{n,m} = \arg\min_{G \in \mathcal{O}_m} h(f^{\sigma_1}(\cdot|G), P_n^{\sigma_0}(\cdot))$ for any $n, m \geq 1$.

- Step 2: Choose

$$\widetilde{m}_n = \inf\big\{m \geq 1 : h\big(f^{\sigma_1}(\cdot|\widehat{G}_{n,m}), P_n^{\sigma_0}(\cdot)\big) < \epsilon\big\},$$

  where $\epsilon > 0$ is any given positive constant and σ_1, σ_0 are two chosen bandwidths.
- Step 3: Let $\widetilde{G}_n = \widehat{G}_{n,\widetilde{m}_n}$ for each n.

Unlike Step 2 in Algorithm 1 where we consider the difference between $h(f^{\sigma_1}(\cdot|\widehat{G}_{n,m}), P_n^{\sigma_0}(\cdot))$ and $h(f^{\sigma_1}(\cdot|\widehat{G}_{n,m+1}), P_n^{\sigma_0}(\cdot))$, here we consider the evaluation of $h(f^{\sigma_1}(\cdot|\widehat{G}_{n,m}), P_n^{\sigma_0}(\cdot))$ in Algorithm 2. The above robust estimator of mixing measure is based on the idea of minimum Hellinger distance estimator and superefficiency phenomenon. A related approach considered in the well-specified setting was taken by Heinrich and Kahn [13]. Their construction was based on minimizing supremum norm based distance, without using the convolution kernels K_{σ_1} and K_{σ_0}; moreover, the threshold ϵ was set to vanish as $n \to \infty$. However, the supremum norm based estimator may be numerically unstable under the misspecified setting.

Our focus with Algorithm 2 in this section will be mainly about its attractive theoretical performance. As we observe from Algorithm 2, the values of f, f_0, K, and G_0 along with the bandwidths σ_1, σ_0 play crucial roles in determining the convergence rate of $\widetilde{G}_n$ to G_0 for any given $\epsilon > 0$. Similar to the argument of Theorem 3.1 and Theorem 3.2, one of the key ingredients to fulfill that goal is to find the conditions of these factors such that we obtain the consistency of $\widetilde{m}_n$. The following theorem yields the sufficient and necessary conditions to address the consistency question.

Theorem 4.1. *Given $\sigma_1 \geq 0$ and $\sigma_0 > 0$. Then, we have*

(i) *Under the well-specified kernel setting and the case that $\sigma_1 = \sigma_0$, $\widetilde{m}_n \to k_0$ almost surely if and only if*

$$\epsilon < h\big(f_0^{\sigma_0}(\cdot|G_{0,k_0-1}), f_0^{\sigma_0}(\cdot|G_0)\big), \tag{2}$$

where $G_{0,k_0-1} = \arg\min_{G\in\mathcal{E}_{k_0-1}} h(f_0^{\sigma_0}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0))$.

(ii) *Under the misspecified kernel setting, if $k_* < \infty$, then $\widetilde{m}_n \to k_*$ almost surely if and only if*

$$h\big(f^{\sigma_1}(\cdot|G_*), f_0^{\sigma_0}(\cdot|G_0)\big) \leq \epsilon < h\big(f^{\sigma_1}(\cdot|G_{*,k_*-1}), f_0^{\sigma_0}(\cdot|G_0)\big), \tag{3}$$

where $G_{*,k_*-1} = \arg\min_{G\in\mathcal{E}_{k_*-1}} h(f^{\sigma_1}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0))$ and $G_* \in \mathcal{M}$ with exactly k_* components.

If we allow $\epsilon \to 0$ in Algorithm 2, we achieve the inconsistency of $\widetilde{m}_n$ under the misspecified kernel setting when $k_* < \infty$. Hence, the choice of threshold ϵ from Heinrich and Kahn [13] is not optimal regarding the misspecified kernel setting. Unfortunately, conditions (2) and (3) are rather cryptic as in general, it is hard to determine the exact formulation of G_{0,k_0-1}, G_{*,k_*-1}, and G_*. It would be of interest to find relatively simple sufficient conditions on f, f_0, K, G_0, σ_1, and σ_0 according to which either (2) or (3) holds. Unfortunately, this seems to be a difficult task in the misspecified setting. Under the well-specified kernel setting, a sufficient condition for (2) can be reformulated as a condition regarding the lower bound on the smallest mass of G_0, the minimal distance between its point masses, and the lower bound between the Hellinger distance and Wasserstein distance:

Proposition 4.1 (Well-specified kernel setting). *For any given $\sigma_0 > 0$, assume that $f_0^{\sigma_0}$ admits uniform Hölder property up to the first oder and is identifiable. If we have*

$$\inf_{G \in \mathcal{E}_{k_0-1}} \frac{h(f_0^{\sigma_0}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0))}{W_1(G, G_0)} \min_{1 \le i \le k_0} p_i^0 \min_{1 \le i \ne j \le k_0} \|\theta_i^0 - \theta_j^0\| \ge \epsilon,$$

then we obtain the inequality in (2).

5. Non-standard settings

In this section, we briefly demonstrate that our robust estimator in Algorithm 1 (similarly Algorithm 2) also achieves desirable convergence rates under non-standard settings. In particular, in the first setting, either f_0 or f may not be identifiable in the first order. In the second setting, the true mixing measure G_0 changes with the sample size n and converges to some discrete distribution $\widetilde{G}_0$ under W_1 distance.

5.1. Singular Fisher information matrix

The results in the previous sections are under the assumption that both the true kernel f_0 and the chosen kernel f are identifiable in the first order. This is equivalent to the non-singularity of the Fisher information matrix of $f_0(x|G_0)$ and $f(x|G_*)$ when $G_* \in \mathcal{M}$, that is, both $I(G_0, f_0)$ and $I(G_*, f)$ are non-singular. Therefore, we achieve the cherished convergence rate $n^{-1/2}$ of $\widehat{G}_n$. Unfortunately, these assumptions do not always hold. For instance, both the Gamma and skewnormal kernel are not identifiable in the first order (Ho and Nguyen [14], Ho and Nguyen [15]). According to Azzalini and Dalla Valle [1], Wiper, Rios Insua and Ruggeri [37], these kernels are particularly useful for modelling various kinds of data: the Gamma kernel is used for modeling non-negative valued data and the skewnormal kernel is used for modeling asymmetric data. Therefore, it is worth considering the performance of our estimator in Algorithm 1 under the nonidentifiability in the first order of both kernels f_0 and f. Throughout this section, for the simplicity of the argument we consider only the well-specified kernel setting and the setting that f_0 may not be identifiable in the first order. Additionally, we also choose $\sigma_1 = \sigma_0 > 0$. The argument for the misspecified kernel setting, the non-identifiability in the first order setting of either f or f_0, and the general choices of σ_1, σ_0 can be argued in the similar fashion.

The non-identifiability in the first order of f_0 implies that the Fisher information matrix $I(G_0, f_0)$ of $f_0(x|G_0)$ is singular at particular values of G_0. Therefore, the convergence rate of $\widehat{G}_n$ to G_0 will be much slower than the standard convergence rate $n^{-1/2}$. In order to precisely determine the convergence rates of parameter estimation under the singular Fisher information matrix setting, Ho and Nguyen [15] introduced a notion of *singularity level* of the mixing measure G_0 relative to the mixture model class; alternatively we say the singularity level of Fisher information matrix $I(G_0, f_0)$ (cf. Definition 3.1 and Definition 3.3 in [15]). Here, we briefly summarize the high level idea of singularity level according to the notations in our paper for the convenience of readers. In particular, we say that $I(G_0, f_0)$ admits r-th level of singularity relative to the ambient space $\mathcal{O}_{k_0}$ for $0 \leq r < \infty$ if we have:

$$\begin{aligned} &\inf_{G \in \mathcal{O}_{k_0}} V\big(f_0(\cdot|G), f_0(\cdot|G_0)\big)/W_s^s(G, G_0) = 0, \quad s = 1, \ldots, r, \\ &V\big(f_0(\cdot|G), f_0(\cdot|G_0)\big) \gtrsim W_{r+1}^{r+1}(G, G_0), \quad \text{for all } G \in \mathcal{O}_{k_0}. \end{aligned} \tag{4}$$

The infinite singularity level of the Fisher information matrix $I(G_0, f_0)$ implies that inequality (4) will not hold for any $r \geq 0$. (Actually, these are consequences, not the original definition of singularity level in Ho and Nguyen [15], but this is sufficient for our purpose.)

When f_0 is identifiable in the first order, $I(G_0, f_0)$ will only have singularity level zero for all $G_0 \in \mathcal{E}_{k_0}$, i.e., $r = 0$ in (4). However, the singularity levels of the Fisher information matrix $I(G_0, f_0)$ are generally not uniform over G_0 when $I(G_0, f_0)$ is singular. For example, when f_0 is skewnormal kernel, $I(G_0, f_0)$ will admit any level of singularity, ranging from 0 to ∞ depending on the interaction of atoms and masses of G_0 (Ho and Nguyen [15]). The notion of singularity level allows us to establish precisely the convergence rate of any estimator of G_0. In fact, if $r < \infty$ is the singularity level of $I(G_0, f_0)$, for any estimation method that yields the convergence rate $n^{-1/2}$ for $f_0(x|G_0)$ under the Hellinger distance, the induced best possible rate of convergence for the mixing measure G_0 is $n^{-1/2(r+1)}$ under W_{r+1} distance. If $r = \infty$ is the singularity level of $I(G_0, f_0)$, all the estimation methods will yield a non-polynomial convergence rate of G_0, one that is slower than $n^{-1/2s}$ for any $s \geq 1$.

Now, by using the same line of argument as that of Theorem 3.1 we have the following result regarding the convergence rate of $\widehat{G}_n$ to G_0 when the Fisher information matrix $I(G_0, f_0)$ has r-th singularity level for some $r < \infty$.

Proposition 5.1. *Given the well-specified kernel setting, that is,* $\{f\} = \{f_0\}$, *and the choice that* $\sigma_1 = \sigma_0 > 0$. *Assume that the Fisher information* $I(G_0, f_0)$ *has* r*-th singularity level where* $r < \infty$ *and condition* (P.2) *in Theorem* 3.1 *holds, that is,* $\Psi(G_0, \sigma_0) < \infty$. *Furthermore, the kernel* K *is chosen such that the Fisher information matrix* $I(G_0, f_0^{\sigma_0})$ *has* r*-th singularity level and* $f_0^{\sigma_0}$ *admits a uniform Hölder property up to the* r*-th order. Then, we have*

$$W_{r+1}(\widehat{G}_n, G_0) = O_p\left(\sqrt{\frac{\Psi(G_0, \sigma_0)}{C_r^2(\sigma_0)}} n^{-1/2(r+1)}\right),$$

where $C_r(\sigma_0) = \inf_{G \in \mathcal{O}_{k_0}} \frac{h(f_0^{\sigma_0}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0))}{W_{r+1}^{r+1}(G, G_0)}$.

Remarks.

(i) A mild condition such that $I(G_0, f_0)$ and $I(G_0, f_0^{\sigma})$ have the same singularity level is $\widehat{K}(t) \neq 0$ for all $t \in \mathbb{R}^d$ where $\widehat{K}(t)$ denotes the Fourier transformation of K (cf. Lemma A.3 in Appendix C in Ho, Nguyen and Ritov [17]).

(ii) Examples of f_0 that are not identifiable in the first order and satisfy $\Psi(G_0, \sigma) < \infty$ are skewnormal and exponential kernel while K is chosen to be Gaussian or exponential kernel, respectively.

(iii) The result of Proposition 5.1 implies that under suitable choices of kernel K, our estimator in Algorithm 1 still achieves the best possible convergence rate for estimating G_0 even when the Fisher information matrix $I(G_0, f_0)$ is singular.

5.2. Varying true parameters

So far, our analysis has relied upon the assumption that G_0 is fixed as $n \to \infty$. However, there are situations such as in an asymptotic minimax analysis the true mixing measure G_0 is allowed to vary with n and converge to some distribution $\widetilde{G}_0$ under W_1 distance as $n \to \infty$. In this section, we will demonstrate that our estimator in Algorithm 1 still achieves the optimal convergence rate under that setting of G_0.

Denote the number of components of $\widetilde{G}_0$ by $\widetilde{k}_0$. For the clarity of our argument, we only work with the well-specified kernel setting and with the setting that f_0 is identifiable in the first order. As we have seen from the analysis of Section 3.1, when G_0 does not change with n, the key steps used to establish the standard convergence rate $n^{-1/2}$ of $\widehat{G}_n$ to G_0 are through the combination of the convergence of $\widehat{m}_n$ to k_0 almost surely and, under the first order identifiability of $f_0^{\sigma_0}$, the lower bound

$$h\big(f_0^{\sigma_0}(\cdot|G), f_0^{\sigma_0}(\cdot|G_0)\big) \gtrsim W_1(G, G_0) \tag{5}$$

holds for any $G \in \mathcal{O}_{k_0}$. Unfortunately, these two results no longer hold as G_0 varies with n. The varying G_0 is now denoted by G_0^n, the true mixing distribution when the sample size is n. Let k_0^n be the number of components of G_0^n. Assume moreover that $\limsup_{n\to\infty} k_0^n = k < \infty$. We start with the following result regarding the convergence rate of $\widehat{m}_n$ under that setting of G_0^n.

Proposition 5.2. *Given $\sigma_0 > 0$, $\widehat{m}_n$ obtained by Algorithm 1. If $f_0^{\sigma_0}$ is identifiable, then $|\widehat{m}_n - k_0^n| \to 0$ almost surely as $n \to \infty$.*

According to the above proposition, $\widehat{m}_n$ will not converge to $\widetilde{k}_0$ almost surely when $k > \widetilde{k}_0$. Additionally, from that proposition, inequality (5) no longer holds since both the number of components of $\widehat{G}_n$ and G_0^n vary. To account for that problem, we need to impose a much stronger condition on the identifiability of $f_0 * K_{\sigma_0}$.

Throughout the rest of this section, we assume that $d = d_1 = 1$, i.e., we specifically work with the univariate setting of f_0, and $k > \widetilde{k}_0$. Using a bound of Heinrich and Kahn [13], we obtain the following.

Proposition 5.3. *Given $\sigma_0 > 0$. Let K be chosen such that $f_0^{\sigma_0}$ is identifiable up to the $(2k - 2\widetilde{k}_0)$-order and admits a uniform Hölder condition up to $(2k - 2\widetilde{k}_0)$-order. Then, there exist $\epsilon_0 > 0$ and $N(\epsilon_0) \in \mathbb{N}$ such that*

$$h\big(f_0^{\sigma_0}(\cdot|G), f_0^{\sigma_0}\big(\cdot|G_0^n\big)\big) \geq C_v(\sigma) W_1^{2k-2\widetilde{k}_0+1}\big(G, G_0^n\big), \tag{6}$$

for any $n \geq N(\epsilon_0)$ and for any $G \in \mathcal{O}_{k_0^n}$ such that $W_1(G, \widetilde{G}_0) \leq \epsilon_0$. Here, $C_v(\sigma)$ is some positive constant depending only on $\widetilde{G}_0$ and σ.

Similar to the argument of Lemma 3.1, a simple example of K and f_0 for the assumptions of Proposition 5.3 to hold is $\widehat{K}(t) \neq 0$ for all $t \in \mathbb{R}^d$ and f_0 is identifiable up to the $(2k - 2\widetilde{k}_0)$-order. Now, a combination of Proposition 5.2 and Proposition 5.3 yields the following result regarding the convergence rate of $\widehat{G}_n$ to G_0^n.

Corollary 5.1. *Given the assumptions in Proposition 5.3. Assume that $\Psi(G_0^n, \sigma_0) < \infty$ for all $n \geq 1$. Then, we have*

$$W_1\bigl(\widehat{G}_n, G_0^n\bigr) = O_p\biggl(\sqrt{\frac{\Psi(G_0^n, \sigma_0)}{C_v^2(\sigma_0)}} n^{-1/(4k-4\widetilde{k}_0+2)}\biggr),$$

where $C_v(\sigma_0)$ is the constant in inequality (6).

Remark.

(i) If f_0 and K are univariate Gaussian kernels or Cauchy kernel respectively, then $\Psi(G_0^n, \sigma_0) \to \Psi(\widetilde{G}_0, \sigma_0)$ as $n \to \infty$.

(ii) If $W_1(G_0^n, \widetilde{G}_0) = O(n^{-1/(4k-4\overline{k}_0+2)+\kappa})$ for some $\kappa > 0$, then the convergence rate $n^{-1/(4k-4\overline{k}_0+2)}$ of $\widehat{G}_n$ to G_0^n is sharp in the sense of minimax (cf. Theorem 3.2 in Heinrich and Kahn [13]). Therefore, our estimator in Algorithm 1 also achieves the minimax rate of convergence for estimating G_0^n. However, our estimator from Algorithm 1 may be more appealing than that from Heinrich and Kahn [13] for computational reasons. We will illustrate the result of Corollary 5.1 via careful simulation studies in Section 6.

6. Empirical studies

We present in this section numerous numerical studies to validate our theoretical results in the previous sections. To find the mixing measure $\widehat{G}_{n,m} = \arg\min_{G \in \mathcal{O}_m} h(f^{\sigma_1}(\cdot|G), P_n^{\sigma_0}(\cdot))$, we utilize the HMIX algorithm developed in Section 4.1 of Cutler and Cordero-Braña [7]. This algorithm is essentially similar to the EM algorithm and ultimately gives us local solutions to the previous minimization problem.

6.1. Synthetic data

We start with evaluating Algorithm 1 using synthetic data. To avoid local minima from finding $\widehat{G}_{n,m}$ with synthetic data from HMIX algorithm, we initialize the parameters in a relatively small neighborhood around the true parameters to guarantee that the updates from HMIX algorithm converge to local optimal solutions that are close to true parameters after a certain number of iterations. Now, our discussion is divided into separate enquiries of the well- and misspecified kernel setups.

Well-specified kernel setting

Under this setting, we assess the performance of our estimator in Algorithm 1 under two cases of G_0:

Case 1. G_0 is fixed with the sample size. Under this case, we consider three choices of f_0: Gaussian and Cauchy kernel for satisfying first order identifiability condition, and skewnormal kernel for failing the first order identifiability condition.

- Case 1.1 – Gaussian family:

$$f_0(x|\eta,\tau) = \frac{1}{\sqrt{2\pi}\tau}\exp\left(-\frac{(x-\eta)^2}{2\tau^2}\right),$$
$$G_0 = \frac{1}{2}\delta_{(0,\sqrt{10})} + \frac{1}{4}\delta_{(-0.3,\sqrt{0.05})} + \frac{1}{4}\delta_{(0.3,\sqrt{0.05})}.$$

- Case 1.2 – Cauchy family:

$$f_0(x|\eta,\tau) = \frac{1}{\pi\tau(1+(x-\eta)^2/\tau^2)},$$
$$G_0 = \frac{1}{2}\delta_{(0,\sqrt{10})} + \frac{1}{4}\delta_{(-0.3,\sqrt{0.05})} + \frac{1}{4}\delta_{(0.3,\sqrt{0.05})}.$$

- Case 1.3 – Skewnormal family:

$$f_0(x|\eta,\tau,m) = \frac{2}{\sqrt{2\pi}\tau}\exp\left(-\frac{(x-\eta)^2}{2\tau^2}\right)\Phi\big(m(x-\eta)/\tau\big),$$
$$G_0 = \frac{1}{2}\delta_{(0,\sqrt{10},0)} + \frac{1}{4}\delta_{(-0.3,\sqrt{0.05},0)} + \frac{1}{4}\delta_{(0.3,\sqrt{0.05},0)},$$

where Φ is the cumulative function of standard normal distribution.

For the Gaussian case and skewnormal case of f_0, we choose K to be the standard Gaussian kernel while K is chosen to be the standard Cauchy kernel for the Cauchy case of f_0. Note that, regarding skewnormal case it was shown that the Fisher information matrix $I(G_0, f_0)$ has second level singularity (cf. Theorem 5.3 in Ho and Nguyen [15]); therefore, from the result of Proposition 5.1, the convergence rate of $\widehat{G}_n$ to G_0 will be at most $n^{-1/6}$. Now for the bandwidths, we choose $\sigma_1 = \sigma_0 = 1$. The sample sizes will be $n = 200 * i$ where $1 \leq i \leq 20$. The tuning parameter C_n is chosen according to BIC criterion. More specifically, $C_n = \sqrt{3\log n}/\sqrt{2}$ for Gaussian and Cauchy family while $C_n = \sqrt{2\log n}$ for skewnormal family. For each sample size n, we perform Algorithm 1 exactly 100 times and then choose $\widehat{m}_n$ to be the estimated number of components with the highest probability of appearing. Afterwards, we take the average among all the replications with the estimated number of components $\widehat{m}_n$ to obtain $W_1(\widehat{G}_n, G_0)$. See Figure 1 where the Wasserstein distances $W_1(\widehat{G}_n, G_0)$ and the percentage of time $\widehat{m}_n = 3$ are plotted against increasing sample size n along with the error bars. The simulation results regarding Gaussian and

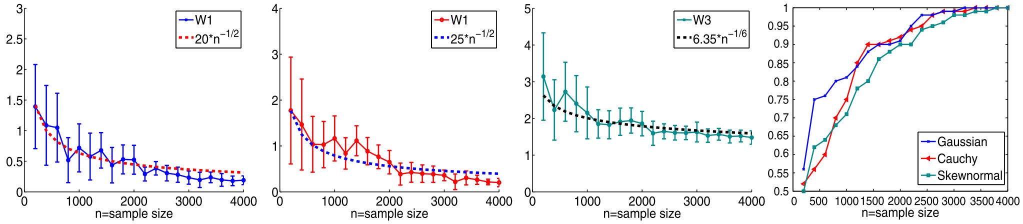

Figure 1. Performance of $\widehat{G}_n$ in Algorithm 1 under the well-specified kernel setting and fixed G_0. Left to right: (1) $W_1(\widehat{G}_n, G_0)$ under Gaussian case. (2) $W_1(\widehat{G}_n, G_0)$ under Cauchy case. (3) $W_3(\widehat{G}_n, G_0)$ under Skewnormal case. (4) Percentage of time $\widehat{m}_n = 3$ obtained from 100 runs.

Cauchy family match well with the standard $n^{-1/2}$ convergence rate from Theorem 3.1 while the simulation results regarding skewnormal family also fit with the best possible convergence rate $n^{-1/6}$ as we argued earlier.

Case 2. G_0 is varied with the sample size. Under this case, we consider two choices of f_0: Gaussian and Cauchy kernel with only location parameter.

- Case 2.1 – Gaussian family:

$$f_0(x|\eta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\eta)^2}{2}\right),$$

$$G_0 = \frac{1}{4}\delta_{1-1/n} + \frac{1}{4}\delta_{1+1/n} + \frac{1}{2}\delta_2,$$

  where n is the sample size.
- Case 2.2 – Cauchy family:

$$f_0(x|\eta) = \frac{1}{\pi(1+(x-\eta)^2)},$$

$$G_0 = \frac{1}{4}\delta_{1-1/\sqrt{n}} + \frac{1}{4}\delta_{1+1/\sqrt{n}} + \frac{1}{2}\delta_{1+2/\sqrt{n}}.$$

With these settings, we can verify that $\widetilde{G}_0 = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ for the Gaussian case and $\widetilde{G}_0 = \delta_1$ for the Cauchy case. Additionally, $W_1(G_0, \widetilde{G}_0) \asymp 1/n$ for the Gaussian case and $W_1(G_0, \widetilde{G}_0) \asymp 1/\sqrt{n}$ for the Cauchy case. According to the result of Corollary 5.1, the convergence rate of $W_1(\widehat{G}_n, G_0)$ is $n^{-1/6}$ for the Gaussian case and is $n^{-1/10}$ for the Cauchy case, which are also minimax according to the values of $W_1(G_0, \widetilde{G}_0)$. The procedure for choosing K, σ_1, σ_0, n, and $\widehat{m}_n$ is similar to that of Case 1. See Figure 2 where the Wasserstein distances $W_1(\widehat{G}_n, G_0)$ and the percentage of time $\widehat{m}_n = 3$ are plotted against increasing sample size n along with the error bars. The simulation results for both Gaussian and Cauchy family agree with the convergence rates $n^{-1/6}$ and $n^{-1/10}$, respectively.

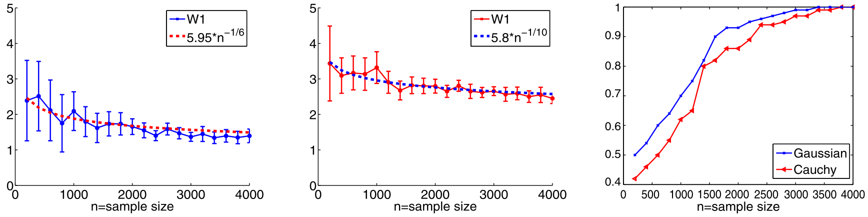

Figure 2. Performance of $\widehat{G}_n$ in Algorithm 1 under the well-specified kernel setting and varied G_0. Left to right: (1) $W_1(\widehat{G}_n, G_0)$ under Gaussian case. (2) $W_1(\widehat{G}_n, G_0)$ under Cauchy case. (3) Percentage of time $\widehat{m}_n = 3$ obtained from 100 runs.

Misspecified kernel setting

Under that setting, we assess the performance of Algorithm 1 under two cases of f, f_0, K, σ_1, σ_0, and G_0.

Case 3. f_0 is a finite mixture of f and $\sigma_1 = \sigma_0 > 0$. Under this case, we consider two choices of f: Gaussian and Cauchy kernel with both location and scale parameter.

- Case 3.1 – Gaussian distribution: f is normal kernel,

$$f_0(x|\eta,\tau) = \frac{1}{2}f(x-2|\eta,\tau) + \frac{1}{2}f(x+2|\eta,\tau),$$
$$G_0 = \frac{1}{3}\delta_{(0,2)} + \frac{2}{3}\delta_{(1,3)}.$$

- Case 3.2 – Cauchy distribution: f is Cauchy kernel,

$$f_0(x|\eta,\tau) = \frac{1}{2}f(x-2|\eta,\tau) + \frac{1}{2}f(x+2|\eta,\tau),$$
$$G_0 = \frac{1}{3}\delta_{(0,2)} + \frac{2}{3}\delta_{(1,3)}.$$

With these settings of f, f_0, G_0, we can verify that $G_* = \frac{1}{6}\delta_{(-2,2)} + \frac{1}{3}\delta_{(-1,3)} + \frac{1}{6}\delta_{(2,2)} + \frac{1}{3}\delta_{(3,3)}$ for any $\sigma_1 = \sigma_0 > 0$. The procedure for choosing K, σ_0, n, and $\widehat{m}_n$ is similar to that of Case 1 in the well-specified kernel setting. Figure 3 illustrates the Wasserstein distances $W_1(\widehat{G}_n, G_*)$ and the percentage of time $\widehat{m}_n = 4$ along with the increasing sample size n and the error bars. The simulation results under that simple misspecified setting of both families suit with the standard $n^{-1/2}$ rate from Theorem 3.2.

Remarks with the results from Case 3

The results from various settings of Case 3 raise an interesting question regarding the practicality of the established consistency results in Section 3.2 for the misspecified kernel cases. In particular, even though the true mixing measure G_0 has only two components, we actually obtain

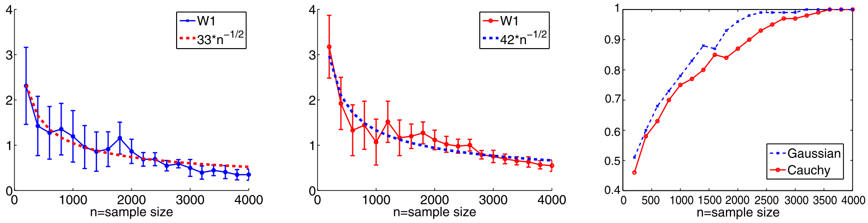

Figure 3. Performance of $\widehat{G}_n$ in Algorithm 1 under misspecified kernel setting and f_0 is a finite mixture of f. Left to right: (1) $W_1(\widehat{G}_n, G_*)$ under Gaussian case. (2) $W_1(\widehat{G}_n, G_*)$ under Cauchy case. (3) Percentage of time $\widehat{m}_n = 4$ obtained from 100 runs.

mixing measure G_* with four components under the choice of K and σ_0, σ_1. Since we do not know the true kernel density function, the natural question is how does finding 4 components of mixing measure reflect the true heterogeneity structure of the population?

The issue that being raised under these specific settings is valid and broadly applied to inference under misspecification: how does the statistician interpret what he has found given that his model is not correct? A satisfactory answer may not be available unless some assumption is accepted about "how misspecified" is our misspecification. We believe that a good approach to address this is by appealing to approximation theory: even though our choice of kernel K is misspecified, we may have some knowledge and reasonable assumption about how well this kernel choice approximates the class of densities that contain the true data-generating density. The question turns to how to relate the estimates obtained under kernel misspecification to the true parameters. We believe that this constitutes a fascinating research direction.

As for the specific settings that we mentioned above, it is not possible to recover G_0 from G_*, given that the statistician does not know anything about the true f_0, because $f(x|G_*) = f_0(x|G_0)$. From a theoretical standpoint, there is not much we can do in this situation due to the indistinguishability between the pair (f, G) and the pair (f_0, G_0), unless some additional knowledge is know about either f_0 or G_0.

From a practical standpoint, we advocate for an extra post-processing step for mixture model based inference: once the algorithm stops, the statistician should perform a testing procedure to verify if some of the obtained clusters should be merged into one or not: clusters that are similar to each other should be merged, small clusters should be discarded and/or merged to a similar one. This also helps with the quality of the parameter estimates for the merged clusters, since we know that fitting with overfitted mixture models can result in highly inefficient estimates. Then, the fact that we find 4 mass points for mixing measure G, along with the parameters associated with these 4 points, should still be useful for understanding about the structure of heterogeneity for the underlying population, even as G_0 is not recovered. We leave a rigorous development of this post-processing methodology for future exploration.

Case 4. σ_1, σ_0 are chosen such that $\{f^{\sigma_1}\} = \{f_0^{\sigma_0}\}$. Under this case, we consider two choices of f and f_0: Gaussian and Cauchy kernel with only location parameter.

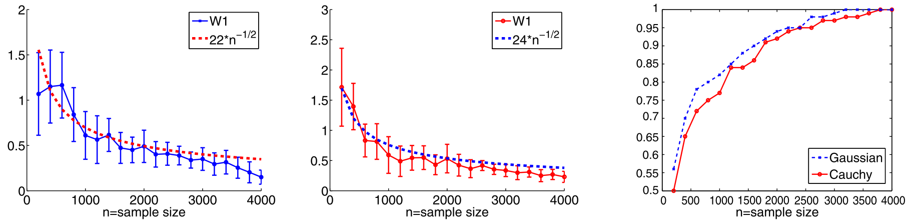

Figure 4. Performance of $\widehat{G}_n$ in Algorithm 1 under misspecified kernel setting and $\{f * K_{\sigma_1}\} = \{f_0 * K_{\sigma_0}\}$. Left to right: (1) $W_1(\widehat{G}_n, G_*)$ under Gaussian case. (2) $W_1(\widehat{G}_n, G_*)$ under Cauchy case. (3) Percentage of time $\widehat{m}_n = 2$ obtained from 100 runs.

- Case 4.1 – Gaussian distribution:

$$f(x|\eta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\eta)^2}{2}\right), \quad f_0(x|\eta) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(x-\eta)^2}{8}\right),$$
$$G_0 = \frac{1}{3}\delta_{-1} + \frac{2}{3}\delta_2.$$

- Case 4.2 – Cauchy distribution: f is Cauchy kernel,

$$f(x|\eta) = \frac{1}{\pi(1+(x-\eta)^2)}, \quad f_0(x|\eta) = \frac{4}{2\pi(4+(x-\eta)^2)},$$
$$G_0 = \frac{1}{3}\delta_{-1} + \frac{2}{3}\delta_2.$$

To ensure that $\{f^{\sigma_1}\} = \{f_0^{\sigma_0}\}$, we need to choose $\sigma_1^2 + 1 = \sigma_0^2 + 4$ for both the cases of Gaussian and Cauchy distribution when K is chosen to be the standard Gaussian and Cauchy kernel respectively. Therefore, with our simulation studies of Algorithm 1 in this case, we choose $\sigma_1 = 2$ while $\sigma_0 = 1$. Under these choices of bandwidths, we quickly have $G_* = G_0$. Note that, since there exists no value of $\sigma_0 > 0$ such that $\sigma_0^2 + 4 = 1$, it implies that the estimator from WS algorithm may not be able to estimate the true mixing measure G_0 regardless the value of σ_0. Now, the procedure for choosing K, n, and $\widehat{m}_n$ is similar to that of Case 1 in the well-specified kernel setting. Figure 4 illustrates the Wasserstein distances $W_1(\widehat{G}_n, G_0)$ and the percentage of time $\widehat{m}_n = 2$ along with the increasing sample size n and the error bars. The simulation results under that misspecified setting of both families fit with the standard $n^{-1/2}$ rate from Theorem 3.2.

6.2. Real data

We begin investigating the performance of Algorithm 1 on the well-known data set of the Sodium-lithium countertransport (SLC) data (Dudley *et al.* [11], Roeder [34], Ishwaran, James and Sun [19]). This simple dataset includes red blood cell sodium-lithium countertransport (SLC) activity data collected from 190 individuals. As being argued by Roeder [34], the SLC activity

Table 1. Summary of parameter estimates in SLC activity data from mixture of two normal distributions with Algorithm 1, WS Algorithm, MKE Algorithm, and EM Algorithm. Here, p_i, η_i, τ_i represents the weights, means, and variance respectively

	p_1	p_2	η_1	η_2	τ_1	τ_2
Algorithm 1	0.264	0.736	0.368	0.231	0.118	0.065
WS Algorithm	0.305	0.695	0.352	0.222	0.106	0.060
MKE Algorithm	0.246	0.754	0.378	0.225	0.102	0.060
EM Algorithm	0.328	0.672	0.363	0.227	0.115	0.058

data were believed to be derived from either mixture of two normal distributions or mixture of three normal distributions. Therefore, we will fit this data by using mixture of normal distributions with unknown mean and variance. We choose the bandwidths $\sigma_1 = \sigma_0 = 0.05$ and the tuning parameter $C_n = \sqrt{3 \log n}/\sqrt{2}$ where n is the sample size. This follows BIC, which is the criterion appropriate for modelling parameter estimation. The simulation result yields $\widehat{m}_n = 2$ while the values of $\widehat{G}_n$ are reported in Table 1.

The SLC activity data was also considered in Woo and Sriram [38] when the authors achieved $\overline{m}_n = 2$. In particular, they allowed the bandwidth σ_0 in WS Algorithm to go to 0 and chose the tuning parameter $C_n = 3/n$, which is inspired by AIC criterion. They also obtained similar result of estimating the true number of components when utilizing the minimum Kulback–Leibler divergence estimator (MKE) from James, Priebe and Marchette [20]. The values of parameter estimation from these two algorithms were presented in Table 7 in Woo and Sriram [38] where we will use them for the comparison purpose with the results from Algorithm 1. Moreover, we also run the EM Algorithm to determine the parameter estimation when we assume the data come from mixture of two normal distributions. All the values of parameter estimation from these three algorithms are included in Table 1. Finally, Figure 5 represents the fits from parameter estimation of all the aforementioned algorithms to SLC data. Even though the weights from Algorithm 1 are not very close to those from WS Algorithm and EM Algorithm, the fit from Algorithm 1 is comparable to those from these algorithms, that is, their fits look fairly similar. As a consequence,

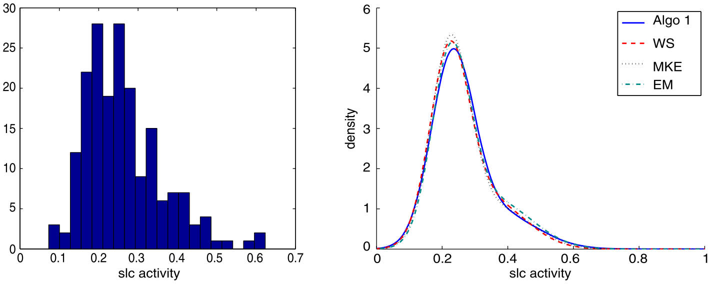

Figure 5. From left to right: (1) Histogram of SLC activity data. (2) Density plot from mixture of two normals based on Algorithm 1, WS Algorithm, MKE Algorithm, and MLE.

the results from Algorithm 1 with SLC data are in agreement with those from several state-of-the-art algorithms in the literature.

7. Summaries and discussions

In this paper, we propose flexible robust estimators of mixing measure in finite mixture models based on the idea of minimum Hellinger distance estimator, model selection criteria, and super-efficiency phenomenon. Our estimators are shown to exhibit the consistency of the number of components under both the well- and misspecified kernel setting. Additionally, the best possible convergence rates of parameter estimation are derived under various settings of both kernel f and f_0. Another salient feature of our estimators is the flexible choice of bandwidths, which circumvents the subtle choices of bandwidth from proposed estimators in the literature. However, there are still many open questions relating to the performance or the extension of our robust estimators in the paper. We give several examples:

- As being mentioned in the paper, our estimator in Algorithm 1 and WS estimator achieve the consistency of the number of components when the bandwidth σ_0 goes to 0 sufficiently slow. Can we determine the setting of bandwidths such that the convergence rates of parameter estimation from these estimators are optimal, at least under the well-specified kernel setting?
- Our analysis is based on the assumption that the parameters of G_0 belong to the compact parameter space Θ. When G_0 is finitely supported, this is always the case, but the set is unknown in advance and, in practice, we often do not know the range of the true parameters. Therefore, it would be interesting to see whether our estimators in Algorithm 1 and Algorithm 2 still achieve both the consistency of the number of components and best possible convergence rates of parameter estimation when $\Theta = \mathbb{R}^{d_1}$.
- Bayesian robust inference of mixing measure in finite mixture models has been of interest recently, see, for example, Miller and Dunson [30]. Whether the idea of minimum Hellinger distance estimator can be adapted to that setting is also an interesting direction to consider in the future.

Acknowledgements

This research is supported in part by grants NSF CCF-1115769, NSF CAREER DMS-1351362, and NSF CNS-1409303 to XN. YR gratefully acknowledges the partial support from NSF DMS-1712962.

Supplementary Material

Supplement to "Robust estimation of mixing measures in finite mixture models" (DOI: 10.3150/18-BEJ1087SUPP; .pdf). In this supplemental material, we provide self-contained proofs of several key results in the paper.

References

[1] Azzalini, A. and Dalla Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83** 715–726. MR1440039 https://doi.org/10.1093/biomet/83.4.715

[2] Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. *Ann. Statist.* **5** 445–463. MR0448700

[3] Bordes, L., Mottelet, S. and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34** 1204–1232. MR2278356 https://doi.org/10.1214/009053606000000353

[4] Chen, J. and Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *J. Amer. Statist. Assoc.* **103** 1674–1683. MR2722574 https://doi.org/10.1198/016214508000001075

[5] Chen, J., Li, P. and Fu, Y. (2012). Inference on the order of a normal mixture. *J. Amer. Statist. Assoc.* **107** 1096–1105. MR3010897 https://doi.org/10.1080/01621459.2012.695668

[6] Chen, J.H. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. MR1331665 https://doi.org/10.1214/aos/1176324464

[7] Cutler, A. and Cordero-Braña, O.I. (1996). Minimum Hellinger distance estimation for finite mixture models. *J. Amer. Statist. Assoc.* **91** 1716–1723. MR1439115 https://doi.org/10.2307/2291601

[8] Dacunha-Castelle, D. and Gassiat, E. (1997). The estimation of the order of a mixture model. *Bernoulli* **3** 279–299. MR1468306 https://doi.org/10.2307/3318593

[9] Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: Population mixtures and stationary ARMA processes. *Ann. Statist.* **27** 1178–1209. MR1740115 https://doi.org/10.1214/aos/1017938921

[10] Donoho, D.L. and Liu, R.C. (1988). The "automatic" robustness of minimum distance functionals. *Ann. Statist.* **16** 552–586. MR0947562 https://doi.org/10.1214/aos/1176350820

[11] Dudley, C.R.K., Giuffra, L.A., Raine, A.E.G. and Reeders, S.T. (1991). Assessing the role of APNH, a gene encoding for a human amiloride-sensitive $\mathrm{Na}^+/\mathrm{H}^+$ antiporter, on the interindividual variation in red cell $\mathrm{Na}^+/\mathrm{Li}^+$ countertransport. *J. Am. Soc. Nephrol.* **2** 937–943.

[12] Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

[13] Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Ann. Statist.* **46** 2844–2870. MR3851757 https://doi.org/10.1214/17-AOS1641

[14] Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Ann. Statist.* **44** 2726–2755. MR3576559 https://doi.org/10.1214/16-AOS1444

[15] Ho, N. and Nguyen, X. (2016). Singularity structures and impacts on parameter estimation in finite mixtures of distributions. Available at arXiv:1609.02655.

[16] Ho, N. and Nguyen, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electron. J. Stat.* **10** 271–307. MR3466183 https://doi.org/10.1214/16-EJS1105

[17] Ho, N., Nguyen, X. and Ritov, Y. (2020). Supplement to "Robust estimation of mixing measures in finite mixture models." https://doi.org/10.3150/18-BEJ1087SUPP.

[18] Hunter, D.R., Wang, S. and Hettmansperger, T.P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35** 224–251. MR2332275 https://doi.org/10.1214/009053606000001118

[19] Ishwaran, H., James, L.F. and Sun, J. (2001). Bayesian model selection in finite mixtures by marginal density decompositions. *J. Amer. Statist. Assoc.* **96** 1316–1332. MR1946579 https://doi.org/10.1198/016214501753382255

[20] James, L.F., Priebe, C.E. and Marchette, D.J. (2001). Consistent estimation of mixture complexity. *Ann. Statist.* **29** 1281–1296. MR1873331 https://doi.org/10.1214/aos/1013203454

[21] Johannes, J. (2009). Deconvolution with unknown error distribution. *Ann. Statist.* **37** 2301–2323. MR2543693 https://doi.org/10.1214/08-AOS652

[22] Karunamuni, R.J. and Wu, J. (2009). Minimum Hellinger distance estimation in a nonparametric mixture model. *J. Statist. Plann. Inference* **139** 1118–1133. MR2479854 https://doi.org/10.1016/j.jspi.2008.07.004

[23] Kasahara, H. and Shimotsu, K. (2014). Non-parametric identification and estimation of the number of components in multivariate mixtures. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 97–111. MR3153935 https://doi.org/10.1111/rssb.12022

[24] Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhya, Ser. A* **62** 49–66. MR1769735

[25] Lin, N. and He, X. (2006). Robust and efficient estimation under data grouping. *Biometrika* **93** 99–112. MR2277743 https://doi.org/10.1093/biomet/93.1.99

[26] Lindsay, B.G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*. Hayward, CA: Institute of Mathematical Statistics.

[27] Lindsay, B.G. (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods. *Ann. Statist.* **22** 1081–1114. MR1292557 https://doi.org/10.1214/aos/1176325512

[28] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics: Applied Probability and Statistics*. New York: Wiley Interscience. MR1789474 https://doi.org/10.1002/0471721182

[29] McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs* **84**. New York: Dekker. MR0926484

[30] Miller, J. and Dunson, D. Robust Bayesian inference via coarsening. *J. Amer. Statist. Assoc.* To appear.

[31] Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.* **41** 370–400. MR3059422 https://doi.org/10.1214/12-AOS1065

[32] Pearson, K. (1894). Contributions to the theory of mathematical evolution. *Philos. Trans. R. Soc. Lond. Ser. A* **185** 71–110.

[33] Richardson, S. and Green, P.J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc. Ser. B* **59** 731–792. MR1483213 https://doi.org/10.1111/1467-9868.00095

[34] Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* **89** 487–495. MR1294074

[35] Teicher, H. (1961). Identifiability of mixtures. *Ann. Math. Stat.* **32** 244–248. MR0120677 https://doi.org/10.1214/aoms/1177705155

[36] Villani, C. (2009). *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften* [*Fundamental Principles of Mathematical Sciences*] **338**. Berlin: Springer. MR2459454 https://doi.org/10.1007/978-3-540-71050-9

[37] Wiper, M., Rios Insua, D. and Ruggeri, F. (2001). Mixtures of gamma distributions with applications. *J. Comput. Graph. Statist.* **10** 440–454. MR1939034 https://doi.org/10.1198/106186001317115054

[38] Woo, M.-J. and Sriram, T.N. (2006). Robust estimation of mixture complexity. *J. Amer. Statist. Assoc.* **101** 1475–1486. MR2279473 https://doi.org/10.1198/016214506000000555

[39] Woodward, W.A., Parr, W.C., Schucany, W.R. and Lindsey, H. (1984). A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *J. Amer. Statist. Assoc.* **79** 590–598. MR0763578

Received September 2017 and revised October 2018