

Robust estimation of mixing measures in finite mixture models

Nhat Ho*, XuanLong Nguyen**, and Ya'acov Ritov**

*Department of EECS, University of California, Berkeley

**Department of Statistics, University of Michigan, Ann Arbor

September 23, 2017

Abstract

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 super-efficiency 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 mis-specified kernel settings for any fixed bandwidth. These desirable asymptotic properties are illustrated via careful simulation studies with both synthetic and real data.¹

AMS 2000 subject classification: Primary 62F15, 62G05; secondary 62G20.

Keywords and phrases: model misspecification, convergence rates, mixture models, Fisher singularities, strong identifiability, minimum distance estimator, model selection, super-efficiency, 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 [1894] 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, 1988, Lindsay, 1995, McLachlan and Peel, 2000].

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

$$p_{G_0, f_0}(x) := \int f_0(x|\theta) dG_0(\theta) = \sum_{i=1}^{k_0} p_i^0 f_0(x|\theta_i^0),$$

¹This 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.

where $G_0 = \sum_{i=1}^{k_0} p_i^0 \delta_{\theta_i^0}$ is a true but *unknown* mixing measure with exactly $k_0 < \infty$ non-zero components and $\left\{ f_0(x|\theta), \theta \in \Theta \subset \mathbb{R}^{d_1} \right\}$ is a true family of density functions, possibly *partially unknown* where $d_1 \geq 1$. There are essentially three principal challenges to the models that have attracted a great deal of attention from various researchers. They include estimating the true number of components k_0 , understanding the behaviors of parameter estimation, i.e., 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, 1994, Escobar and West, 1995, Dacunha-Castelle and Gassiat, 1997, Richardson and Green, 1997, Dacunha-Castelle and Gassiat, 1999, Keribin, 2000, James et al., 2001, Chen et al., 2012, Chen and Khalili, 2012, Kasahara and Shimotsu, 2014]. 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, i.e., the exact-fitted setting of finite mixtures, Ho and Nguyen [2016c] 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, i.e., the over-fitted setting of finite mixtures, Chen [1995], Nguyen [2013], Ho and Nguyen [2016c] 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 [2015]. On the other hand, Ho and Nguyen [2016a,b] 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 et al., 2006, Hunter et al., 2007]. 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 et al., 2007]. 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, i.e., 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., 1984, Donoho and Liu, 1988, Cutler and Cordero-Brana, 1996]. However, there are scarce work for robust estimators when the true number of components k_0 is unknown. Recently, Woo and Sriram [2006] proposed a robust estimator of the number of components based on the idea of minimum Hellinger distance estimator [Beran, 1977, Lindsay, 1994, Lin and He, 2006, Karunamuni and Wu, 2009]. However, their work faced certain limitations. Firstly, 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, i.e., when $\{f\} = \{f_0\}$, the bandwidth should vanish to 0 sufficiently slowly (cf. Theorem 3.1 in [Woo and Sriram, 2006]). Secondly, the behaviors of parameter 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, 2006]). 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 , i.e., $k_0 = 2$, under that setting of f and f_0 . Unfortunately, their argument is not clear as their estimator should maintain the number of components of 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, 2006]. Not only our estimators are computationally feasible and robust but they also possess various desirable properties, such as the flexible choice of bandwidth, the consistency of the number of components, and the best possible convergence rates of parameter estimation. In particular, the main contributions in this paper can be summarized as follows

- (i) We treat the well-specified kernel setting, i.e., $\{f\} = \{f_0\}$, and the misspecified kernel setting, i.e., $\{f\} \neq \{f_0\}$, separately. Under both settings, we achieve the consistency of our estimators regarding the true number of components for any fixed bandwidth. Furthermore, when the bandwidth vanishes to 0 at 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 are given in Section 8 and proofs of the remaining results are given in the Appendices.

Notation Given two densities p, q (with respect to the Lebesgue measure μ), the total variation distance is given by $V(p, q) = \frac{1}{2} \int |p(x) - q(x)| d\mu(x)$. 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 $\kappa = (\kappa_1, \dots, \kappa_{d_1}) \in \mathbb{N}^{d_1}$, we denote $\frac{\partial^{|\kappa|} f}{\partial \theta^\kappa}(x|\theta) = \frac{\partial^{|\kappa|} f}{\partial \theta_1^{\kappa_1} \dots \partial \theta_{d_1}^{\kappa_{d_1}}}(x|\theta)$ where $\theta = (\theta_1, \dots, \theta_{d_1})$. Additionally, the expression $a_n \gtrsim b_n$ will be used to denote the inequality up to a constant multiple where the value of the constant is independent of n . 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

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

Additionally, we denote $\mathcal{E}_{k_0} := \mathcal{E}_{k_0}(\Theta)$ the space of discrete mixing measures with exactly k_0 distinct support points on Θ and $\mathcal{O}_k := \mathcal{O}_k(\Theta)$ the space of discrete mixing measures with at most k distinct support points on Θ . Additionally, denote $\mathcal{G} := \mathcal{G}(\Theta) = \bigcup_{k \in \mathbb{N}_+} \mathcal{E}_k$ the set of all discrete measures with finite supports on Θ . Finally, $\overline{\mathcal{G}}$ 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 [2013], 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, 2008]. In particular, consider a mixing measure $G = \sum_{i=1}^k p_i \delta_{\theta_i}$, where $\mathbf{p} = (p_1, p_2, \dots, p_k)$ denotes the proportion vector. Likewise, let $G' = \sum_{i=1}^{k'} p'_i \delta_{\theta'_i}$. A coupling between \mathbf{p} and \mathbf{p}' is a joint distribution \mathbf{q} on $[1, \dots, k] \times [1, \dots, k']$, which is expressed as a matrix $\mathbf{q} = (q_{ij})_{1 \leq i \leq k, 1 \leq j \leq k'} \in [0, 1]^{k \times k'}$ with margins $\sum_{m=1}^k q_{mj} = p'_j$ and $\sum_{m=1}^{k'} q_{im} = p_i$ for any $i = 1, 2, \dots, k$ and $j = 1, 2, \dots, k'$. We use $\mathcal{Q}(\mathbf{p}, \mathbf{p}')$ to denote the space of all such couplings. For any $r \geq 1$, the r -th order Wasserstein distance between G

and G' is given by

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

where $\|\cdot\|$ denotes the l_2 norm for elements in \mathbb{R}^{d_1} . It is simple to argue that if a sequence of probability measures $G_n \in \mathcal{O}_{k_0}$ 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, 2016b, Heinrich and Kahn, 2015]). We start with

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

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

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 $\kappa = (\kappa_1, \dots, \kappa_{d_1})$.

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

The classical identifiability condition entails that the family of density function $\{f(x|\theta), \theta \in \Theta\}$ is identifiable if for any $G_1, G_2 \in \mathcal{G}$, $p_{G_1, f}(x) = p_{G_2, f}(x)$ almost surely implies that $G_1 \equiv G_2$ [Teicher, 1961]. To be able to establish convergence rates of parameters, we have to utilize the following stronger notion of identifiability:

Definition 2.2. *For any $r \geq 1$, 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 $k \geq 1$, given k different elements $\theta_1, \dots, \theta_k \in \Theta$. 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 $1 \leq i \leq k$ and $|\eta| \leq r$.

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, $l_G := \frac{\partial}{\partial G} \log p_{G, f}(x)$ is the score function, where $\frac{\partial}{\partial G}$ denotes the derivatives with respect to all the components and masses of G . The first order identifiability of f is an equivalent way to say that the Fisher information matrix $I(G, f)$ is non-singular for any G . Now, under the first order identifiability and the first order uniform Lipschitz condition on f , a careful investigation of Theorem 3.1 and Corollary 3.1 in Ho and Nguyen [2016c] yields the following result:

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

$$h(p_{G,f}, p_{G_0,f}) \geq C_0 W_1(G, G_0).$$

Note that, the result of Proposition 2.1 is slightly stronger than that of Theorem 3.1 and Corollary 3.1 in Ho and Nguyen [2016c] as it holds for any $G \in \mathcal{O}_{k_0}$ instead of only for any $G \in \mathcal{E}_{k_0}$ as in these later results. The first order identifiability property of kernel density function f implies that any estimation method that yields the convergence rate $n^{-1/2}$ for $p_{G_0,f}$ under the Hellinger distance, the induced rate of convergence for the mixing measure G_0 is $n^{-1/2}$ under W_1 distance.

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 Lipschitz condition up to the first order. Now, let K be any fixed multivariate density function and $K_\sigma(x) = \frac{1}{\sigma^d} K\left(\frac{x}{\sigma}\right)$ for any $\sigma > 0$. We define

$$f * K_\sigma(x|\theta) := \int f(x - y|\theta) K_\sigma(y) dy$$

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

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

for any discrete mixing measure $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ in $\bar{\mathcal{G}}$. For the convenience of our argument later, we also denote that $p_{G,f} * K_\sigma := p_{G,f}$ as long as $\sigma = 0$. Now, our approach to define a robust estimator of G_0 is inspired by the minimum Hellinger distance estimator [Beran, 1977] and the model selection criteria. Indeed, we have the following algorithm

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

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

- Step 2: Choose

$$\hat{m}_n = \inf \left\{ m \geq 1 : h(p_{\hat{G}_{n,m},f} * K_{\sigma_1}, P_n * K_{\sigma_0}) \leq h(p_{\hat{G}_{n,m+1},f} * K_{\sigma_1}, P_n * K_{\sigma_0}) + C_n n^{-1/2} \right\},$$

- Step 3: Let $\hat{G}_n = \hat{G}_{n,\hat{m}_n}$ 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. The choice of C_n in Algorithm 1 is to guarantee that \hat{m}_n is finite. Additionally, it 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 [2006] 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 [2006] can be stated as follows.

Woo-Sriram (WS) Algorithm:

- Step 1: Determine $\bar{G}_{n,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G,f}, P_n * K_{\sigma_0})$ for any $n, m \geq 1$.

- Step 2: Choose

$$\bar{m}_n = \inf \left\{ m \geq 1 : h(p_{\bar{G}_{n,m},f}, P_n * K_{\sigma_0}) \leq h(p_{\bar{G}_{n,m+1},f}, P_n * K_{\sigma_0}) + C'_n n^{-1/2} \right\},$$

where $C'_n n^{-1/2} \rightarrow 0$.

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

The main distinction between our estimator and Woo-Sriram's (WS) estimator is that we also allow the convolution of mixture density $p_{G,f}$ with K_{σ_1} . This double convolution trick in Algorithm 1 was also considered in James et al. [2001] to construct the consistent estimation of mixture complexity. However, their work was based on the Kullback-Leibler (KL) divergence rather than the Hellinger distance and was restricted to only the choice that $\sigma_1 = \sigma_0$. Under the misspecified kernel setting, i.e., $\{f\} \neq \{f_0\}$, the estimation of mixing measure G_0 from KL divergence can be highly unstable. Additionally, James et al. [2001] 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 \hat{G}_n over WS estimator \bar{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 \hat{G}_n and \bar{G}_n .

Under the well-specified setting, i.e., $\{f\} = \{f_0\}$, 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 \hat{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 $P_n * K_{\sigma_0}(x)$ is an unbiased estimator of $p_{G_0,f_0} * K_{\sigma_0}(x)$ for all $x \in \mathcal{X}$. It guarantees that $h(P_n * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0}) = O_p(n^{-1/2})$ 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 $\sigma_0 > 0$, \bar{G}_n converges to \bar{G}_0 where $\bar{G}_0 = \arg \min_{G \in \bar{\mathcal{G}}} h(p_{G,f_0}, p_{G_0,f_0} * K_{\sigma_0})$

under certain conditions of f_0 , K , and \bar{G}_0 . Unfortunately, \bar{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 $n \rightarrow \infty$ under the well-specified kernel setting with $\sigma_1 = \sigma_0$, we are able to guarantee that $\hat{m}_n \rightarrow k_0$ almost surely when $n\sigma_0^d \rightarrow 0$ (cf. Proposition 3.1). This result is also consistent with the result $\bar{m}_n \rightarrow k_0$ almost surely from Theorem 1 in Woo and

Sriram [2006] 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 $n \rightarrow \infty$. 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 $h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0})$, which may not converge to 0 at the exact rate $n^{-1/2}$ when $\sigma_0 \rightarrow 0$. 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 $h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0})$ 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.

Under the misspecified kernel setting, i.e., $\{f\} \neq \{f_0\}$, the double convolution technique in Algorithm 1 continues to be useful for studying the convergence rate of \widehat{G}_n to G_* where $G_* = \arg \min_{G \in \overline{\mathcal{G}}} h(p_{G, f} * K_{\sigma_1}, p_{G_0, f_0} * K_{\sigma_0})$. 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 $\{f * K_{\sigma_1}\}$ and $\{f_0 * K_{\sigma_0}\}$ 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 $f * K_{\sigma_1} \approx f_0 * K_{\sigma_0}$. 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.

3.1 Well-specified kernel setting

In this section, we consider the setting that f_0 is known, i.e., $\{f\} = \{f_0\}$. 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 $f_0 * K_{\sigma_0}$ in the limit, we would like to guarantee that $f_0 * K_{\sigma_0}$ 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 $f_0 * K_{\sigma_0}$ 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 $f_0 * K_{\sigma_0}$ 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 $f_0 * K_{\sigma_0}$ is identifiable, then $\widehat{m}_n \rightarrow k_0$ almost surely.*
- (ii) *Assume further the following conditions*

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

$$(P.2) \quad \Psi(G_0, \sigma_0) := \int \frac{g(x|G_0, \sigma_0)}{p_{G_0, f_0} * K_{\sigma_0}(x)} dx < \infty \text{ where we have that}$$

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

Then, we obtain

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

$$\text{where } C_1(\sigma_0) := \inf_{G \in \mathcal{O}_{k_0}} \frac{h(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0})}{W_1(G, G_0)}.$$

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

$$h(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \gtrsim W_1(G, G_0)$$

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

- (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

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

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 $\sigma_0 \rightarrow 0$: 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 $\sigma_{0,n} \rightarrow 0$.

Proposition 3.1. *Given a sequence of bandwidths $\{\sigma_{0,n}\}$ such that $\sigma_{0,n} \rightarrow 0$ and $n\sigma_{0,n}^d \rightarrow \infty$ as $n \rightarrow \infty$. If f_0 is identifiable, then $\widehat{m}_n \rightarrow k_0$ 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 $\sigma_{0,n} \rightarrow 0$. 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 $\lim_{\sigma_{0,n} \rightarrow 0} C_1(\sigma_{0,n}) > 0$ (cf. Lemma .2 in Appendix B), 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 $\sigma_{0,n} \rightarrow 0$. As a consequence, whatever the sequence of bandwidths $\sigma_{0,n} \rightarrow 0$ 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 $\sigma_{0,n} \rightarrow 0$ for the future work.

3.2 Misspecified kernel setting

In the previous section, we assume the well-specified kernel setting, i.e., $\{f\} = \{f_0\}$, 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, i.e., the chosen kernel f may be different from the true kernel f_0 . Motivated by this challenge, in this section we consider the setting when $\{f\} \neq \{f_0\}$. 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 $\{f * K_{\sigma_1}(x|\theta), \theta \in \Theta\}$ and $\{f_0 * K_{\sigma_0}(x|\theta), \theta \in \Theta\}$ 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 $p_{G,f} * K_{\sigma_1}$ and $p_{G_0,f_0} * K_{\sigma_0}$, i.e.,

$$G_* := \arg \min_{G \in \overline{\mathcal{G}}} h(p_{G,f} * K_{\sigma_1}, p_{G_0,f_0} * K_{\sigma_0}).$$

As G_* may not be unique, we denote

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

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 $f, f_0, \sigma_1, \sigma_0$, 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*

$$\int p_{G,f} * K_{\sigma_1}(x) \sqrt{\frac{p_{G_0,f_0} * K_{\sigma_0}(x)}{p_{G_*,f} * K_{\sigma_1}(x)}} dx \leq \int \sqrt{p_{G_*,f} * K_{\sigma_1}(x)} \sqrt{p_{G_0,f_0} * K_{\sigma_0}(x)} dx. \quad (1)$$

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 $p_{G_{1,*},f} * K_{\sigma_1}(x) = p_{G_{2,*},f} * K_{\sigma_1}(x)$ 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 $f * K_\sigma$ is identifiable. Furthermore, \mathcal{M}_k is empty for all $k_* < k < \infty$. However, it is possible that \mathcal{M}_∞ still contains various elements. 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}_∞ .

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 $\{f * K_{\sigma_1}\} = \{f_0 * K_{\sigma_0}\}$. 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 $1 \leq i \leq m$. Under that setting, if we choose $\sigma_1 = \sigma_0$, then we can check that $k_* \leq mk_0$ and $p_{G_*, f}(x) = p_{G_0, f_0}(x)$ almost surely. The explicit formulation of G_* , therefore, can be found from the combinations of G_0 and $(p_i^*, \eta_i^*, \tau_i^*)$ where $1 \leq i \leq m$.

From inequality (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 $h^*(p_{G_1, f} * K_{\sigma_1}, p_{G_2, f} * K_{\sigma_1})$ by

$$\begin{aligned} \left(h^*(p_{G_1, f} * K_{\sigma_1}, p_{G_2, f} * K_{\sigma_1}) \right)^2 &= \\ &= \frac{1}{2} \int \left(\sqrt{p_{G_1, f} * K_{\sigma_1}(x)} - \sqrt{p_{G_2, f} * K_{\sigma_1}(x)} \right)^2 \times \sqrt{\frac{p_{G_0, f_0} * K_{\sigma_0}(x)}{p_{G_*, f} * K_{\sigma_1}(x)}} dx. \end{aligned}$$

The notable feature of h^* is the presence of term $\sqrt{p_{G_0, f_0} * K_{\sigma_0}(x) / p_{G_*, f} * K_{\sigma_1}(x)}$ 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 $h^*(p_{G_1, f} * K_{\sigma_1}, p_{G_2, f} * K_{\sigma_1}) \equiv h(p_{G_1, f} * K_{\sigma_1}, p_{G_2, f} * K_{\sigma_1})$ for any $G_1, G_2 \in \overline{\mathcal{G}}$, i.e., 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 $\sigma_1 \geq 0$ 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 (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 $f * K_{\sigma_1}$ is identifiable in the first order and admits uniform Lipschitz property up to the first order. If $k_* < \infty$, then for any $G \in \mathcal{O}_{k_*}$ there holds

$$h^*(p_{G, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) \gtrsim W_1(G, G_*).$$

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 $f * K_{\sigma_1}$ is identifiable, then $\widehat{m}_n \rightarrow k_*$ almost surely.

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

(M.1) The kernel K is chosen such that $f * K_{\sigma_1}$ is identifiable in the first order and admits the uniform Lipschitz property up to the first order.

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

(M.3) $\sup_{\theta \in \Theta} \left\| \frac{\partial f * K_{\sigma_1}(x|\theta)}{\partial \theta} / (f * K_{\sigma_1}(x|\theta))^{3/4} \right\|_{\infty} \leq M_2(\sigma_1)$ 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 $C_{*,1}(\sigma_1) := \inf_{G \in \mathcal{O}_{k_*}} \frac{h^*(p_{G,f} * K_{\sigma_1}, p_{G_*,f} * K_{\sigma_1})}{W_1(G, G_*)}$ 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 $\{f * K_{\sigma_1}\} = \{f_0 * K_{\sigma_0}\}$, i.e., $G_* \equiv G_0$, 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 $\widehat{m}_n \rightarrow k_* < \infty$ almost surely, which is eventually a consequence of the identifiability of kernel density function $f * K_{\sigma_1}$. 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 [2013] 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 $G_2 \in \overline{\mathcal{G}}$, $|p_{G_1,f}(x) - p_{G_2,f}(x)| = 0$ 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 $f * K_{\sigma_1}$ is finitely identifiable. If $k_* = \infty$, we achieve $\widehat{m}_n \rightarrow \infty$ 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 [2013] 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, i.e., 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, i.e., $\{f\} = \{f_0\}$, 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 Lipschitz property up to the first order. Assume now we can find

$$\overline{G}_0 := \arg \min_{G \in \overline{\mathcal{G}}} h(p_{G, f_0}, p_{G_0, f_0} * K_{\sigma_0}),$$

i.e., \overline{G}_0 is the discrete mixing measure that minimizes the Hellinger distance between p_{G, f_0} and $p_{G_0, f_0} * K_{\sigma_0}$. 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 $1 \leq i \leq k_0$ 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 $\overline{C} := \inf_{G \in \mathcal{O}_{\overline{k}_0}} \frac{h^*(p_{G, f_0}, p_{\overline{G}_0, f_0})}{W_1(G, \overline{G}_0)}$ 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 bandwidth 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 $\overline{G}_n \rightarrow G_0$ 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 $\{f * K_{\sigma_1}\} = \{f_0 * K_{\sigma_0}\}$, 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.

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

$$\widetilde{m}_n = \inf \left\{ m \geq 1 : h(p_{\widehat{G}_{n,m},f} * K_{\sigma_1}, P_n * K_{\sigma_0}) < \epsilon \right\},$$

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 $h(p_{\widehat{G}_{n,m},f} * K_{\sigma_1}, P_n * K_{\sigma_0})$ and $h(p_{\widehat{G}_{n,m+1},f} * K_{\sigma_1}, P_n * K_{\sigma_0})$, here we consider solely the evaluation of $h(p_{\widehat{G}_{n,m},f} * K_{\sigma_1}, P_n * K_{\sigma_0})$ 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 [2015]. 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 $n \rightarrow \infty$. Although of theoretical interest, their estimator appears difficult to compute efficiently and may be unstable due to the use of the supremum norm.

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 $\sigma_1 = \sigma_0$, $\tilde{m}_n \rightarrow k_0$ almost surely if and only if

$$\epsilon < h(p_{G_0, k_0-1, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \quad (2)$$

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

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

$$h(p_{G_*, f} * K_{\sigma_1}, p_{G_0, f_0} * K_{\sigma_0}) \leq \epsilon < h(p_{G_*, k_*-1, f} * K_{\sigma_1}, p_{G_0, f_0} * K_{\sigma_0}) \quad (3)$$

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

If we allow $\epsilon \rightarrow 0$ in Algorithm 2, we achieve the inconsistency of \tilde{m}_n under the misspecified kernel setting when $k_* < \infty$. Hence, the choice of threshold ϵ from Heinrich and Kahn [2015] 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 mis-specified 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 $f_0 * K_{\sigma_0}$ admits uniform Lipschitz property up to the first order and is identifiable. If we have

$$\inf_{G \in \mathcal{E}_{k_0-1}} \frac{h(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0})}{W_1(G, G_0)} \min_{1 \leq i \leq k_0} p_i^0 \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \geq \epsilon,$$

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 \tilde{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 p_{G_0, f_0} and $p_{G_*, f}$ when $G_* \in \mathcal{M}$, i.e., both $I(G_0, f_0)$ and $I(G_*, f)$ are non-singular. Therefore, we achieve the cherished convergence rate $n^{-1/2}$ of \hat{G}_n . Unfortunately, these assumptions do not always hold. For instance, both the Gamma and skewnormal kernel are not identifiable in the first order [Ho and Nguyen, 2016a,b]. According to Azzalini and Valle [1996], Wiper et al. [2001], these kernels are particularly useful for modelling various kinds of data: the Gamma kernel is used for modeling non-negative valued data and the 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 p_{G_0, f_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 [2016b] 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). 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:

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

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 [2016b], 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, 2016b]. 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 p_{G_0, f_0} under the Hellinger distance, the induced best possible rate of convergence for the mixing measure G_0 is $n^{-1/2(r+1)}$ under W_{r+1} distance. If $r = \infty$ is the singularity level of $I(G_0, f_0)$, all the estimation methods will yield 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, i.e., $\{f\} = \{f_0\}$, 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, i.e., $\Psi(G_0, \sigma_0) < \infty$. Furthermore, the kernel K is chosen such that the Fisher information matrix $I(G_0, f_0 * K_{\sigma_0})$ has r -th singularity level and $f_0 * K_{\sigma_0}$ admits a uniform Lipschitz property up to the r -th order. Then, we have*

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

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

Remarks:

- (i) A mild condition such that $I(G_0, f_0)$ and $I(G_0, f_0 * K_\sigma)$ have the same singularity level is $\widehat{K}(t) \neq 0$ for all $t \in \mathbb{R}^d$ where $\widehat{K}(t)$ denotes the Fourier transformation of K (cf. Lemma .3 in Appendix B).
- (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 $n \rightarrow \infty$. 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 $n \rightarrow \infty$. 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 $f_0 * K_{\sigma_0}$, the lower bound

$$h(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \gtrsim W_1(G, G_0) \quad (5)$$

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 $\limsup_{n \rightarrow \infty} k_0^n = k < \infty$. 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 $f_0 * K_{\sigma_0}$ is identifiable, then $|\widehat{m}_n - k_0^n| \rightarrow 0$ almost surely as $n \rightarrow \infty$.*

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 [2015], we obtain the following:

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

$$h(p_{G, f_0} * K_{\sigma_0}, p_{G_0^n, f_0} * K_{\sigma_0}) \geq C_v(\sigma) W_1^{2k - 2\widetilde{k}_0 + 1}(G, G_0^n) \quad (6)$$

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 $(2k - 2k_0)$ -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*

$$W_1(\widehat{G}_n, G_0^n) = O_p\left(\sqrt{\frac{\Psi(G_0^n, \sigma_0)}{C_v^2(\sigma_0)}} n^{-1/(4k-4\bar{k}_0+2)}\right)$$

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 $\Psi(G_0^n, \sigma_0) \rightarrow \Psi(\widetilde{G}_0, \sigma_0)$ as $n \rightarrow \infty$.
- (ii) If $W_1(G_0^n, \widetilde{G}_0) = O(n^{-1/(4k-4\bar{k}_0+2)+\kappa})$ for some $\kappa > 0$, then the convergence rate $n^{-1/(4k-4\bar{k}_0+2)}$ of \widehat{G}_n to G_0^n is sharp in the sense of minimax (cf. Theorem 3.2 in [Heinrich and Kahn, 2015]). 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 [2015] 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 $\widehat{G}_{n,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G,f} * K_{\sigma_1}, P_n * K_{\sigma_0})$, we utilize the HMIX algorithm developed in Section 4.1 of [Cutler and Cordero-Brana, 1996]. This algorithm is essentially similar to the EM algorithm and ultimately gives us local solutions to the previous minimization problem.

6.1 Synthetic Data

We start with testing Algorithm 1 using synthetic data. The discussion is divided into separate enquiries of the well- and mis-specified kernel setups.

Well-specified kernel setting Under this setting, we assess the performance of 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:

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

$$G_0 = \frac{1}{2}\delta_{(0, \sqrt{10}, 0)} + \frac{1}{4}\delta_{(-0.3, \sqrt{0.05}, 0)} + \frac{1}{4}\delta_{(0.3, \sqrt{0.05}, 0)}.$$

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, 2016b]); therefore, from the result of Proposition 5.1, the convergence rate of \hat{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 \hat{m}_n to be the estimated number of components with the highest probability of appearing. Afterwards, we take the average among all the replications with the estimated number of components \hat{m}_n to obtain $W_1(\hat{G}_n, G_0)$. See Figure 1 where the Wasserstein distances $W_1(\hat{G}_n, G_0)$ and the percentage of time $\hat{m}_n = 3$ are plotted against increasing sample size n along with the error bars. The simulation results regarding Gaussian and Cauchy family match well with the standard $n^{-1/2}$ convergence rate from Theorem 3.1 while the simulation results regarding skewnormal family also fit with the best possible convergence rate $n^{-1/6}$ as we argued earlier.

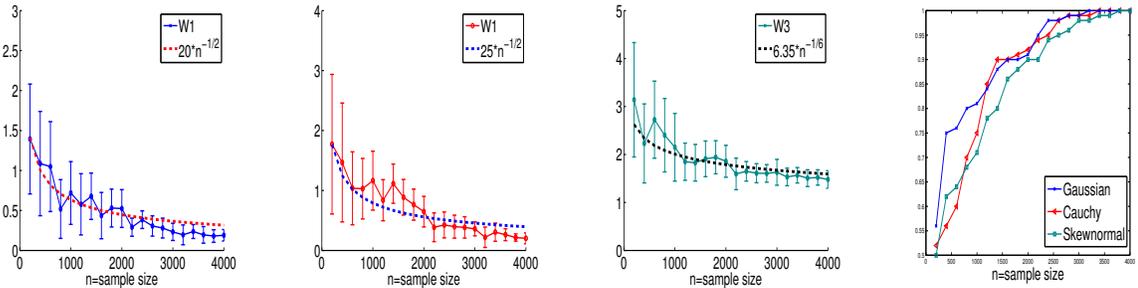


Figure 1: Performance of \hat{G}_n in Algorithm 1 under the well-specified kernel setting and fixed G_0 . Left to right: (1) $W_1(\hat{G}_n, G_0)$ under Gaussian case. (2) $W_1(\hat{G}_n, G_0)$ under Cauchy case. (3) $W_1(\hat{G}_n, G_0)$ under Skewnormal case. (4) Percentage of time $\hat{m}_n = 3$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

Case 2: G_0 is varied with the sample size. Under this case, we consider two choices of f_0 : Gaussian and Cauchy 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 $\tilde{G}_0 = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ for the Gaussian case and $\tilde{G}_0 = \delta_1$ for the Cauchy case. Additionally, $W_1(G_0, \tilde{G}_0) \asymp 1/n$ for the Gaussian case and $W_1(G_0, \tilde{G}_0) \asymp 1/\sqrt{n}$ for the Cauchy case. According to the result of Corollary 5.1, the convergence rate of $W_1(\hat{G}_n, G_0)$ is $n^{-1/6}$ for the Gaussian case and is $n^{-1/10}$ for the Cauchy case, which are also minimax according to the values of $W_1(G_0, \tilde{G}_0)$. The procedure for choosing K, σ_1, σ_0, n , and \hat{m}_n is similar to that of Case 1. See Figure 2 where the Wasserstein distances $W_1(\hat{G}_n, G_0)$ and the percentage of time $\hat{m}_n = 3$ are plotted against increasing sample size n along with the error bars. The simulation results for both Gaussian and Cauchy family agree with the convergence rates $n^{-1/6}$ and $n^{-1/10}$ respectively.

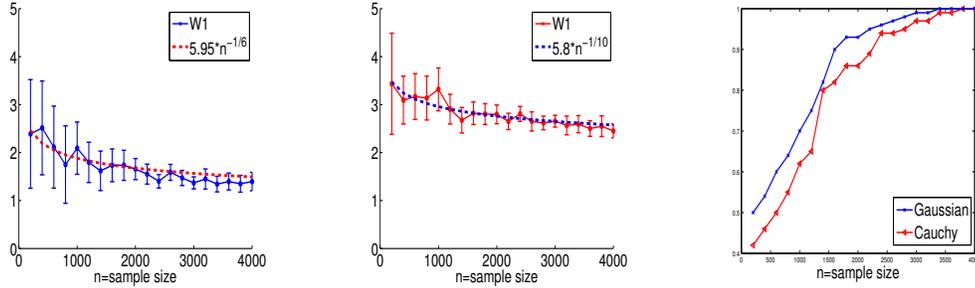


Figure 2: Performance of \hat{G}_n in Algorithm 1 under the well-specified kernel setting and varied G_0 . Left to right: (1) $W_1(\hat{G}_n, G_0)$ under Gaussian case. (2) $W_1(\hat{G}_n, G_0)$ under Cauchy case. (3) Percentage of time $\hat{m}_n = 3$ obtained from 100 runs.

Misspecified kernel setting Under that setting, we assess the performance of Algorithm 1 under two cases of $f, f_0, K, \sigma_1, \sigma_0$, 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 \hat{m}_n is similar to that of Case 1 in the well-specified kernel setting. Figure 3 illustrates the Wasserstein distances $W_1(\hat{G}_n, G_*)$ and the percentage of time $\hat{m}_n = 4$ along with the increasing sample size n and the error bars. The simulation results under that simple misspecified setting of both families suit with the standard $n^{-1/2}$ rate from Theorem 3.2.

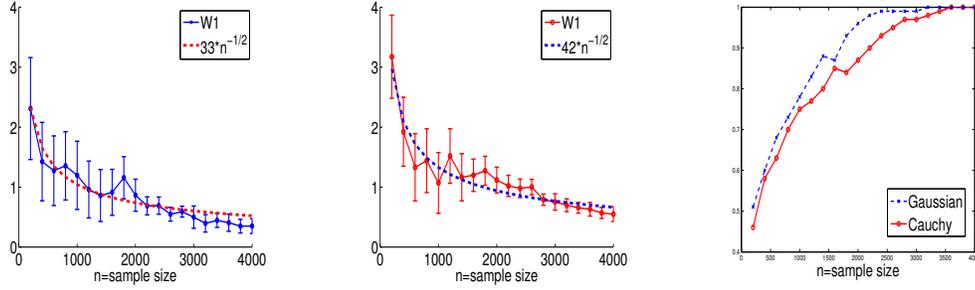


Figure 3: Performance of \hat{G}_n in Algorithm 1 under misspecified kernel setting and f_0 is a finite mixture of f . Left to right: (1) $W_1(\hat{G}_n, G_*)$ under Gaussian case. (2) $W_1(\hat{G}_n, G_*)$ under Cauchy case. (3) Percentage of time $\hat{m}_n = 4$ obtained from 100 runs.

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

- Case 4.1 - Gaussian distribution:

$$f(x|\eta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-\eta)^2}{2}\right), f_0(x|\eta) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(x-\eta)^2}{8}\right)$$

$$G_0 = \frac{1}{3}\delta_{-1} + \frac{2}{3}\delta_2$$

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

$$f(x|\eta) = \frac{1}{\pi(1+(x-\eta)^2)}, f_0(x|\eta) = \frac{4}{2\pi(4+(x-\eta)^2)}$$

$$G_0 = \frac{1}{3}\delta_{-1} + \frac{2}{3}\delta_2$$

To ensure that $\{f * K_{\sigma_1}\} = \{f_0 * K_{\sigma_0}\}$, 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 \hat{m}_n is similar to that of Case 1 in the well-specified kernel setting. Figure 4 illustrates the Wasserstein distances $W_1(\hat{G}_n, G_0)$ and the percentage of time $\hat{m}_n = 2$ 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.

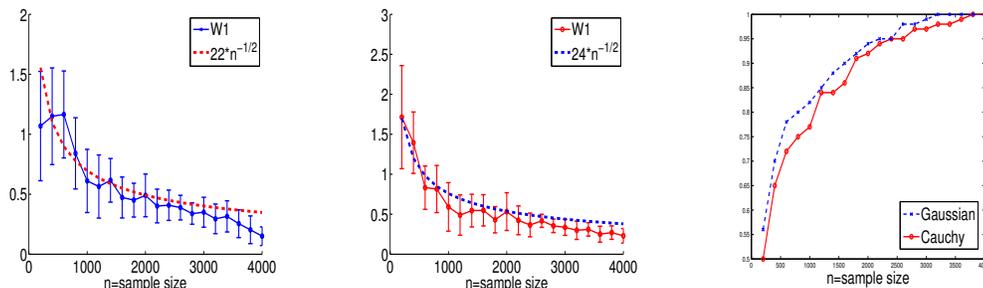


Figure 4: Performance of \hat{G}_n in Algorithm 1 under misspecified kernel setting and $\{f * K_{\sigma_1}\} = \{f_0 * K_{\sigma_0}\}$. Left to right: (1) $W_1(\hat{G}_n, G_*)$ under Gaussian case. (2) $W_1(\hat{G}_n, G_*)$ under Cauchy case. (3) Percentage of time $\hat{m}_n = 2$ obtained from 100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

6.2 Real Data

We begin investigating the performance of Algorithm 1 on the well-known data set of the Sodium-lithium countertransport (SLC) data [Dudley et al., 1991, Roeder, 1994, Ishwaran et al., 2001]. This simple dataset includes red blood cell sodium-lithium countertransport (SLC) activity data collected from 190 individuals. As being argued by Roeder [1994], the SLC activity data were believed to be derived from either mixture of two normal distributions or mixture of three normal distributions. Therefore, we will fit this data by using mixture of normal distributions with unknown mean and variance. We choose the 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 $\hat{m}_n = 2$ while the values of \hat{G}_n are reported in Table 1.

The SLC activity data was also considered in Woo and Sriram [2006] when the authors achieved $\bar{m}_n = 2$. In particular, they allowed the bandwidth σ_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 et al., 2001]. The values of parameter estimation from these two algorithms were presented in Table 7 in Woo and Sriram [2006] where we will use them for the comparison purpose with the results from Algorithm 1. Moreover, we also run the EM Algorithm to determine the parameter 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, i.e., their fits look fairly similar. As a consequence, the results from Algorithm 1 with SLC data are in agreement with those from several state-of-the-art algorithms in the literature.

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

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

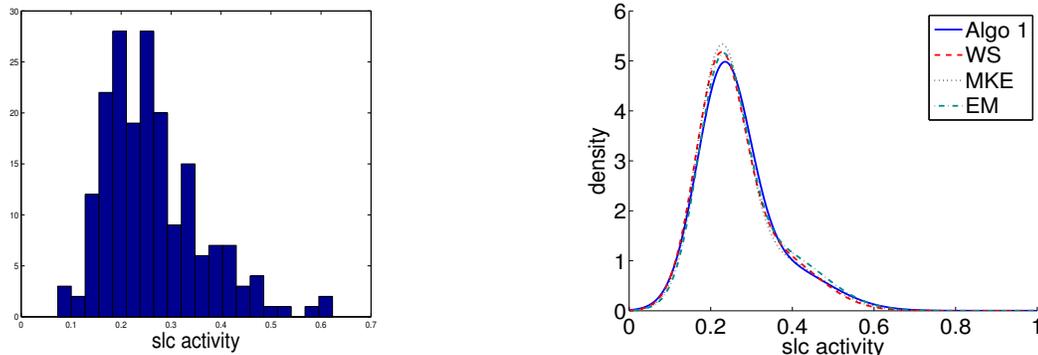


Figure 5: From left to right: (1) Histogram of SLC activity data. (2) Density plot from mixture of two normals based on Algorithm 1, WS Algorithm, MKE Algorithm, and MLE.

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 mis-specified 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 bandwidth 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 set Θ . When G_0 is finitely supported, this is always the case, but the set is unknown in advance and, in practice, we often do not know the range of the true parameters. Therefore, it would be interesting to see whether 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, 2015]. Whether the idea of minimum Hellinger

distance estimator can be adapted to that setting is also an interesting direction to consider in the future.

8 Proofs of key results

In this section, we provide the proofs of Theorem 3.1 and Theorem 3.2 in Section 3. The remaining proofs are given in the Appendices.

PROOF OF THEOREM 3.1 We divide the main argument into three key steps:

Step 1: $\hat{m}_n \rightarrow k_0$ almost surely. The proof of this step follows the argument from [Leroux, 1992]. In fact, for any positive integer m we denote

$$G_{0,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G,f_0} * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0}).$$

Now, as $n \rightarrow \infty$ we have almost surely that

$$h(p_{\hat{G}_{n,m},f_0} * K_{\sigma_0}, P_n * K_{\sigma_0}) - h(p_{\hat{G}_{n,m+1},f_0} * K_{\sigma_0}, P_n * K_{\sigma_0}) \rightarrow d_m,$$

where $d_m = h(p_{G_{0,m},f_0} * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0}) - h(p_{G_{0,m+1},f_0} * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0})$ and the limit is due to the fact that $h(P_n * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0}) \rightarrow 0$ almost surely for all $\sigma_0 > 0$. From the formulation of Step 2 in Algorithm 1 and the fact that $C_n n^{-1/2} \rightarrow 0$ as $n \rightarrow \infty$, we obtain

$$h(p_{\hat{G}_{n,\hat{m}_n},f_0} * K_{\sigma_0}, P_n * K_{\sigma_0}) - h(p_{\hat{G}_{n,\hat{m}_n+1},f_0} * K_{\sigma_0}, P_n * K_{\sigma_0}) \leq C_n n^{-1/2} \rightarrow 0.$$

Therefore, to demonstrate that $\hat{m}_n \rightarrow k_0$ almost surely, it is sufficient to prove that $d_m = 0$ as $m \geq k_0$ and $d_m > 0$ as $m < k_0$. In fact, as $m \geq k_0$, we have $\inf_{G \in \mathcal{O}_m} h(p_{G,f_0} * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0}) = 0$.

Therefore, $d_m = 0$ as $m \geq k_0$.

When $m < k_0$, we assume that $d_m = 0$, i.e., $h(p_{G_{0,m},f_0} * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0}) = h(p_{G_{0,m+1},f_0} * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0})$. It implies that

$$h(p_{G_{0,m},f_0} * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0}) \leq h(p_{G,f_0} * K_{\sigma_0}, p_{G_0,f_0} * K_{\sigma_0}) \quad \forall G \in \mathcal{O}_{m+1}.$$

For any $\epsilon > 0$, we choose $G = (1 - \epsilon)G_{0,m} + \epsilon\delta_\theta$ where $\theta \in \Theta$ is some component. The inequality in the above display implies that

$$\int (p_{G_0,f_0} * K_{\sigma_0}(x))^{1/2} \left(\left[(1 - \epsilon)p_{G_{0,m},f_0} * K_{\sigma_0}(x) + \epsilon f_0 * K_{\sigma_0}(x|\theta) \right]^{1/2} - (p_{G_{0,m},f_0} * K_{\sigma_0}(x))^{1/2} \right) dx \leq 0.$$

As $\epsilon \rightarrow 0$, the above inequality divided by ϵ becomes

$$\int (p_{G_0,f_0} * K_{\sigma_0}(x))^{1/2} (p_{G_{0,m},f_0} * K_{\sigma_0}(x))^{1/2} dx \geq \int (p_{G_0,f_0} * K_{\sigma_0}(x))^{1/2} f_0 * K_{\sigma_0}(x|\theta) (p_{G_{0,m},f_0} * K_{\sigma_0}(x))^{-1/2} dx.$$

Now, by choosing $\theta = \theta_i^0$ for all $1 \leq i \leq k_0$, as we sum up the right hand side of the above inequality, we obtain

$$\begin{aligned} & \int (p_{G_0, f_0} * K_{\sigma_0}(x))^{1/2} (p_{G_0, m, f_0} * K_{\sigma_0}(x))^{1/2} dx \\ & \geq \int (p_{G_0, f_0} * K_{\sigma_0}(x))^{1/2} \left(\sum_{i=1}^{k_0} p_i^0 f_0 * K_{\sigma_0}(x | \theta_i^0) \right) (p_{G_0, m, f_0} * K_{\sigma_0}(x))^{-1/2} \\ & \geq \int (p_{G_0, f_0} * K_{\sigma_0}(x))^{3/2} (p_{G_0, m, f_0} * K_{\sigma_0}(x))^{-1/2} dx \geq 1 \end{aligned}$$

where the final inequality is due to the inequality $\int q_1^{3/2}(x) q_2^{-1/2}(x) dx \geq 1$ for any two density functions $q_1(x)$ and $q_2(x)$. Therefore, we have $h(p_{G_0, m, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) = 0$. Due to the identifiability assumption of $f_0 * K_{\sigma_0}$, the previous equation implies that $G_{0, m} \equiv G_0$, which is a contradiction as $m < k_0$. Thus, we have $d_m > 0$ for any $m < k_0$. We achieve the conclusion that $\hat{n}_n \rightarrow k_0$ almost surely.

Step 2: $h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) = O_p\left(\sqrt{\frac{\Psi(G_0, \sigma_0)}{n}}\right)$. Indeed, by means of Taylor expansion up to the first order, we have

$$\begin{aligned} & h^2(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \\ & = \int \left(1 - \sqrt{1 + \frac{P_n * K_{\sigma_0}(x) - p_{G_0, f_0} * K_{\sigma_0}(x)}{p_{G_0, f_0} * K_{\sigma_0}(x)}} \right)^2 p_{G_0, f_0} * K_{\sigma_0}(x) dx \\ & \simeq \frac{1}{4} \int \frac{(P_n * K_{\sigma_0}(x) - p_{G_0, f_0} * K_{\sigma_0}(x))^2}{p_{G_0, f_0} * K_{\sigma_0}(x)} dx. \end{aligned}$$

Notice that,

$$E\left(\int \frac{(P_n * K_{\sigma_0}(x) - p_{G_0, f_0} * K_{\sigma_0}(x))^2}{p_{G_0, f_0} * K_{\sigma_0}(x)} dx\right) = \int \frac{\text{Var}(P_n * K_{\sigma_0}(x))}{p_{G_0, f_0} * K_{\sigma_0}(x)} dx,$$

From assumption (P.2), we obtain $\int \frac{\text{Var}(P_n * K_{\sigma_0}(x))}{p_{G_0, f_0} * K_{\sigma_0}(x)} dx = O\left(\frac{\Psi(G_0, \sigma_0)}{n}\right)$. It follows that

$$E\left(\int \frac{(P_n * K_{\sigma_0}(x) - p_{G_0, f_0} * K_{\sigma_0}(x))^2}{p_{G_0, f_0} * K_{\sigma_0}(x)} dx\right) = O\left(\frac{\Psi(G_0, \sigma_0)}{n}\right).$$

Therefore, we achieve $h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) = O_p\left(\sqrt{\frac{\Psi(G_0, \sigma_0)}{n}}\right)$. It implies that for any $\epsilon > 0$, we can find $M_\epsilon > 0$ and the index $N_1(\epsilon) \geq 1$ such that

$$P\left(h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) > M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma_0)}{n}}\right) < \epsilon/2 \quad (7)$$

for all $n \geq N_1(\epsilon)$.

Step 3: Now, denote the event $A = \{\widehat{m}_n \rightarrow k_0 \text{ as } n \rightarrow \infty\}$. Under this event, for each $\omega \in A$, we can find $N(\omega)$ such that as $n \geq N(\omega)$, we have $\widehat{m}_n = k_0$. It suggests that $\widehat{G}_n \in \mathcal{O}_{k_0}$ as $n \geq N(\omega)$. Define $A_m = \{\omega \in A : \forall n \geq m \text{ we have } \widehat{m}_n = k_0\}$. From this definition, we obtain $A_1 \subset A_2 \dots \subset A_m \subset \dots$ and $\bigcup_{m=1}^{\infty} A_m = A$. Therefore, $\lim_{m \rightarrow \infty} P(A_m) = P(A) = 1$. Therefore, for any $\epsilon > 0$ we can find the corresponding index $N_2(\epsilon)$ such that $P(A_{N_2(\epsilon)}) > 1 - \epsilon/2$.

Now, for any $\omega \in A_{N_2(\epsilon)}$, we have $\widehat{m}_n = k_0$ as $n \geq N_2(\epsilon)$. From assumptions (P.1) and the definition of $C_1(\sigma_0)$ in Theorem 3.1, we obtain

$$\begin{aligned} C_1(\sigma)W_1(\widehat{G}_n, G_0) &\leq h(p_{\widehat{G}_n, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \\ &\leq h(p_{\widehat{G}_n, f_0} * K_{\sigma_0}, P_n * K_{\sigma_0}) + h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \\ &\leq 2h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}). \end{aligned} \quad (8)$$

Using the inequalities from (7) and (8), we have

$$\begin{aligned} P\left(W_1(\widehat{G}_n, G_0) > 2M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma_0)}{C_1^2(\sigma_0)n}}\right) &= P\left(\left(W_1(\widehat{G}_n, G_0) > 2M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma_0)}{C_1^2(\sigma_0)n}}\right) \mathbb{1}_{A_{N_2(\epsilon_2)}^c}\right) \\ &\quad + P\left(\left(W_1(\widehat{G}_n, G_0) > 2M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma_0)}{C_1^2(\sigma_0)n}}\right) \mathbb{1}_{A_{N_2(\epsilon)}}\right) \\ &\leq \epsilon/2 + P\left(\left(W_1(\widehat{G}_n, G_0) > 2M_\epsilon \sqrt{\frac{\Psi(G_0, \sigma_0)}{C_1^2(\sigma_0)n}}\right) \mathbb{1}_{A_{N_2(\epsilon)}}\right) < \epsilon \end{aligned}$$

for all $n \geq \max\{N_1(\epsilon), N_2(\epsilon)\}$. We achieve the conclusion of the theorem.

PROOF OF THEOREM 3.2 We divide our argument in the proof of this theorem into two key steps.

Step 1 $\widehat{m}_n \rightarrow k_*$ almost surely. Indeed, by carrying out the same argument as that of Step 1 in the proof of Theorem 3.1 (here, we replace f_0 by f and $G_{0,m}$ by $G_{*,m}$, as $m < k_*$), we eventually obtain the following inequality

$$\begin{aligned} &\int (p_{G_0, f_0} * K_{\sigma_0}(x))^{1/2} (p_{G_{*,m}, f} * K_{\sigma_1}(x))^{1/2} dx \geq \\ &\int (p_{G_0, f_0} * K_{\sigma_0}(x))^{1/2} f * K_{\sigma_1}(x|\theta) (p_{G_{*,m}, f} * K_{\sigma_1}(x))^{-1/2} dx. \end{aligned}$$

for any $\theta \in \Theta$. By choosing $\theta \in \text{supp}(G_*)$, which is the set of all support points of G_* , and sum over all of these components, we achieve

$$\begin{aligned} &\int (p_{G_0, f_0} * K_{\sigma_0}(x))^{1/2} (p_{G_{*,m}, f} * K_{\sigma_1}(x))^{1/2} dx \geq \\ &\int (p_{G_0, f_0} * K_{\sigma_0}(x))^{1/2} p_{G_{*,f}} * K_{\sigma_1}(x) (p_{G_{*,m}, f} * K_{\sigma_1}(x))^{-1/2} dx. \end{aligned}$$

From the above inequality, we have

$$\begin{aligned} & \int \left(\sqrt{p_{G_{*,m},f} * K_{\sigma_1}(x)} - \sqrt{p_{G_{*,f}} * K_{\sigma_1}(x)} \right)^2 \sqrt{\frac{p_{G_0,f_0} * K_{\sigma_0}(x)}{p_{G_{*,m},f} * K_{\sigma_1}(x)}} dx \\ & \leq 2 \left(\int \sqrt{p_{G_0,f_0} * K_{\sigma_0}(x)} \sqrt{p_{G_{*,m},f} * K_{\sigma_1}(x)} dx \right. \\ & \quad \left. - \int \sqrt{p_{G_0,f_0} * K_{\sigma_0}(x)} \sqrt{p_{G_{*,f}} * K_{\sigma_1}(x)} dx \right) \leq 0, \end{aligned}$$

where the second inequality is due to the fact that G_* minimizes $h(p_{G,f} * K_{\sigma_1}, p_{G_0,f_0} * K_{\sigma_0})$ among all $G \in \overline{\mathcal{G}}$. The above inequality implies that $p_{G_{*,m},f} * K_{\sigma_1}(x) = p_{G_{*,f}} * K_{\sigma_1}(x)$ for almost surely $x \in \mathcal{X}$. Due to the identifiability of $f * K_{\sigma_1}$, we obtain $G_{*,m} \equiv G_*$, which is a contradiction to the fact that $m < k_*$. Therefore, we achieve $\widehat{m}_n \rightarrow k_*$ almost surely.

Step 2 Now, since $\widehat{m}_n \rightarrow k_*$ almost surely, using the same argument as Step 3 in the proof of Theorem 3.1, we can find $N(\epsilon)$ such that $\widehat{m}_n = k_*$ for any $n \geq N(\epsilon)$ and such that $P(A_{N(\epsilon)}) > 1 - \epsilon/2$ for any $\epsilon > 0$. Additionally, since $\widehat{G}_n = \widehat{G}_{n,\widehat{m}_n}$ minimizes $h(p_{G,f} * K_{\sigma_1}, P_n * K_{\sigma_0})$ among all $G \in \mathcal{O}_{\widehat{m}_n}$, it follows that

$$\int \sqrt{p_{\widehat{G}_n,f} * K_{\sigma_1}(x)} \sqrt{P_n * K_{\sigma_0}(x)} dx \geq \int \sqrt{p_{G_{*,f}} * K_{\sigma_1}(x)} \sqrt{P_n * K_{\sigma_0}(x)} dx.$$

when $n \geq N(\epsilon)$. From this inequality, we obtain

$$\begin{aligned} & \int \left(\sqrt{p_{\widehat{G}_n,f} * K_{\sigma_1}(x)} - \sqrt{p_{G_{*,f}} * K_{\sigma_1}(x)} \right) \left(\sqrt{P_n * K_{\sigma_0}(x)} - \sqrt{p_{G_0,f_0} * K_{\sigma_0}(x)} \right) dx \geq \quad (9) \\ & \int \sqrt{p_{G_0,f_0} * K_{\sigma_0}(x)} \sqrt{p_{G_{*,f}} * K_{\sigma_1}(x)} dx - \int \sqrt{p_{G_0,f_0} * K_{\sigma_0}(x)} \sqrt{p_{\widehat{G}_n,f} * K_{\sigma_1}(x)} dx := B. \end{aligned}$$

By means of the key inequality in Lemma 3.2, we have

$$\begin{aligned} B & \geq \int \sqrt{p_{\widehat{G}_n,f} * K_{\sigma_1}(x)} \sqrt{\frac{p_{G_0,f_0} * K_{\sigma_0}(x)}{p_{G_{*,f}} * K_{\sigma_1}(x)}} dx - \int \sqrt{p_{G_0,f_0} * K_{\sigma_0}(x)} \sqrt{p_{\widehat{G}_n,f} * K_{\sigma_1}(x)} dx \\ & = 2 \left(h^*(p_{\widehat{G}_n,f} * K_{\sigma_1}, p_{G_{*,f}} * K_{\sigma_1}) \right)^2 - B. \end{aligned}$$

It implies that $B \geq \left(h^*(p_{\widehat{G}_n,f} * K_{\sigma_1}, p_{G_{*,f}} * K_{\sigma_1}) \right)^2$. Plugging this inequality to (9) leads to

$$\begin{aligned} C & := \int \left(\sqrt{p_{\widehat{G}_n,f} * K_{\sigma_1}(x)} - \sqrt{p_{G_{*,f}} * K_{\sigma_1}(x)} \right) \left(\sqrt{P_n * K_{\sigma_0}(x)} - \sqrt{p_{G_0,f_0} * K_{\sigma_0}(x)} \right) dx \\ & \geq \left(h^*(p_{\widehat{G}_n,f} * K_{\sigma_1}, p_{G_{*,f}} * K_{\sigma_1}) \right)^2. \quad (10) \end{aligned}$$

For the left hand side (LHS) of (10), we have the following inequality

$$\begin{aligned}
C &\leq \left\| (p_{\widehat{G}_n, f} * K_{\sigma_1})^{1/4} - (p_{G_*, f} * K_{\sigma_1})^{1/4} \right\|_{\infty} \int \left((p_{\widehat{G}_n, f} * K_{\sigma_1}(x))^{1/4} \right. \\
&\quad \left. + (p_{G_*, f} * K_{\sigma_1}(x))^{1/4} \right) \left| \sqrt{P_n * K_{\sigma_0}(x)} - \sqrt{p_{G_0, f_0} * K_{\sigma_0}(x)} \right| dx \\
&\leq \left\| (p_{\widehat{G}_n, f} * K_{\sigma_1})^{1/4} - (p_{G_*, f} * K_{\sigma_1})^{1/4} \right\|_{\infty} \left\| (p_{\widehat{G}_n, f} * K_{\sigma_1})^{1/4} + (p_{G_*, f} * K_{\sigma_1})^{1/4} \right\|_2 \\
&\quad \times \sqrt{2} h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \quad (11)
\end{aligned}$$

where the last inequality is due to Holder's inequality. Now, our next argument will be divided into two small key steps.

Step 2.1 With assumption (M.3), we will show that

$$D = \left\| (p_{G, f} * K_{\sigma_1})^{1/4} - (p_{G_*, f} * K_{\sigma_1})^{1/4} \right\|_{\infty} \leq M_3(\sigma_1) W_1(G, G_0) \quad (12)$$

for any $G \in \mathcal{O}_{k_*}$ where $M_3(\sigma_1)$ is some positive constant.

In fact, denote $G = \sum_{i=1}^k p_i \delta_{\theta_i}$ where $k \leq k_*$ and $G_* = \sum_{i=1}^{k_*} p_i^* \delta_{\theta_i^*}$. Using the same proof argument as that of (27) in the proof of Proposition 4.1, there exists a positive number ϵ_0 depending only G_* such that as long as $W_1(G, G_*) \leq \epsilon_0$, G will have exactly k_* components, i.e., $k = k_*$. Additionally, up to the relabelling of the components of G , we also obtain that $|p_i - p_i^*| \leq c_0 W_1(G, G_*)$ where c_0 is some positive constant depending only on G_* . Therefore, by choosing G such that $W_1(G, G_*) \leq C_0 = \min \left\{ \epsilon_0, \min_{1 \leq i \leq k_*} \frac{p_i^*}{2c_0} \right\}$, we achieve $|p_i - p_i^*| \leq \min_{1 \leq i \leq k_*} p_i^*/2$. Hence, $p_i \geq \min_{1 \leq i \leq k_*} p_i^*/2$ for all $1 \leq i \leq k_*$. Under this setting of G , for any coupling \mathbf{q} of $\mathbf{p} = (p_1, \dots, p_k)$ and $\mathbf{p}^* = (p_1^*, \dots, p_{k_*}^*)$, by means of triangle inequality we obtain

$$\begin{aligned}
D &= \left\| \frac{p_{G, f} * K_{\sigma_1} - p_{G_*, f} * K_{\sigma_1}}{\{(p_{G, f} * K_{\sigma_1})^{1/4} + (p_{G_*, f} * K_{\sigma_1})^{1/4}\} \{(p_{G, f} * K_{\sigma_1})^{1/2} + (p_{G_*, f} * K_{\sigma_1})^{1/2}\}} \right\|_{\infty} \\
&\leq \sum_{i, j} q_{ij} \left\| \frac{f * K_{\sigma_1}(x|\theta_i) - f * K_{\sigma_1}(x|\theta_j^*)}{\{(p_{G, f} * K_{\sigma_1})^{1/4} + (p_{G_*, f} * K_{\sigma_1})^{1/4}\} \{(p_{G, f} * K_{\sigma_1})^{1/2} + (p_{G_*, f} * K_{\sigma_1})^{1/2}\}} \right\|_{\infty}.
\end{aligned}$$

where the ranges of i, j in the above sum satisfy $1 \leq i, j \leq k_*$. It is clear that for any $\alpha \in \{1/2, 1/4\}$

$$\begin{aligned}
&(p_{G, f} * K_{\sigma_1}(x))^{\alpha} + (p_{G_*, f} * K_{\sigma_1}(x))^{\alpha} \\
&> \min \{p_i^{\alpha}, (p_j^*)^{\alpha}\} \{(f * K_{\sigma_1}(x|\theta_i))^{\alpha} + (f * K_{\sigma_1}(x|\theta_j^*))^{\alpha}\} \\
&> \min_{1 \leq i \leq k_*} \left(\frac{p_i^*}{2} \right)^{\alpha} \{(f * K_{\sigma_1}(x|\theta_i))^{\alpha} + (f * K_{\sigma_1}(x|\theta_j^*))^{\alpha}\}.
\end{aligned}$$

Therefore, we eventually achieve that

$$D \lesssim \sum_{i, j} q_{ij} \left\| (f * K_{\sigma_1}(x|\theta_i))^{1/4} - (f * K_{\sigma_1}(x|\theta_j^*))^{1/4} \right\|_{\infty}.$$

Now, due to assumption (M.3) and mean value theorem, we achieve for any $x \in \mathcal{X}$ that

$$\left| (f * K_{\sigma_1}(x|\theta_i))^{1/4} - (f * K_{\sigma_1}(x|\theta_j^*))^{1/4} \right| \leq M_2(\sigma_1) \|\theta_i - \theta_j^*\|.$$

Thus, for any coupling q of p and p^*

$$D \lesssim \sum_{i,j} q_{ij} \|\theta_i - \theta_j^*\|.$$

As a consequence, we eventually have

$$D \lesssim \inf_{q \in \mathcal{Q}(p, p^*)} \sum_{i,j} q_{ij} \|\theta_i - \theta_j^*\| = W_1(G, G_*)$$

for any $G \in \mathcal{O}_{k_*}$ such that $W_1(G, G_*) \leq C_0$. Now, for any $G \in \mathcal{O}_{k_*}$ such that $W_1(G, G_*) > C_0$, as D is bounded, it is clear that $D \lesssim W_1(G, G_*)$. In sum, we achieve inequality (12).

Step 2.2 Due to assumption (M.2), we also can quickly verify that

$$\left\| (p_{\widehat{G}_n, f} * K_{\sigma_1})^{1/4} + (p_{G_*, f} * K_{\sigma_1})^{1/4} \right\|_2 \leq 2\sqrt{k_* M_1(\sigma_1)}. \quad (13)$$

Combining (11), (12), (13), we ultimately achieve that

$$\left(h^*(p_{\widehat{G}_n, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) \right)^2 \leq M(\sigma_1) W_1(\widehat{G}_n, G_*) h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0})$$

where $M(\sigma_1)$ is some positive constant. Due to assumption (M.1), from the result of Lemma 3.4 and the definition of $C_{*,1}(\sigma_1)$, we have

$$h^*(p_{\widehat{G}_n, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) \gtrsim C_{*,1}(\sigma_1) W_1(\widehat{G}_n, G_*).$$

Combining the above results with the bound $h(P_n * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) = O_p\left(\sqrt{\frac{\Psi(G_0, \sigma_0)}{n}}\right)$ from Step 2 in the proof of Theorem 3.1, we quickly obtain the conclusion of the theorem.

References

- A. Azzalini and A. D. Valle. The multivariate skew-normal distribution. *Biometrika*, 83:715–726, 1996.
- R. Beran. Minimum hellinger distance estimates for parametric models. *Annals of Statistics*, 5:445–453, 1977.
- L. Bordes, S. Mottelet, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model. *Annals of Statistics*, 34:1204–1232, 2006.
- J. Chen. Optimal rate of convergence for finite mixture models. *Annals of Statistics*, 23(1):221–233, 1995.
- J. Chen and A. Khalili. Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association*, 103:1674–1683, 2012.
- J. Chen, P. Li, and Y. Fu. Inference on the order of a normal mixture. *Journal of the American Statistical Association*, 107:1096–1105, 2012.

- A. Cutler and O. I. Cordero-Brana. Minimum Hellinger distance estimation for finite mixture models. *Journal of the American Statistical Association*, 91:1716–1723, 1996.
- D. Dacunha-Castelle and E. Gassiat. The estimation of the order of a mixture model. *Bernoulli*, 3: 279–299, 1997.
- D. Dacunha-Castelle and E. Gassiat. Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Annals of Statistics*, 27:1178–1209, 1999.
- L. Devroye and Laszlo Györfi. *Nonparametric density estimation: the L1 view*. John Wiley and Sons, 1985.
- D. Donoho and R. C. Liu. The automatic robustness of minimum distance functionals. *Annals of Statistics*, 16:552–586, 1988.
- C. R. K. Dudley, L. A. Giuffra, A. E. G. Raine, and S. T. Reeders. Assessing the role of apnh, a gene encoding for a human amiloride-sensitive Na^+/H^+ antiporter, on the interindividual variation in red cell Na^+/Li^+ countertransport. *Journal of the American Society of Nephrology*, 2:937–943, 1991.
- M. D. Escobar and M. West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588, 1995.
- P. Heinrich and J. Kahn. Optimal rates for finite mixture estimation. *arXiv:1507.04313*, 2015.
- N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44:2726–2755, 2016a.
- N. Ho and X. Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *arXiv:1609.02655*, 2016b.
- N. Ho and X. Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10:271–307, 2016c.
- D. R. Hunter, S. Wang, and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. *Annals of Statistics*, 35:224–251, 2007.
- H. Ishwaran, L. F. James, and J. Sun. Bayesian model selection in finite mixtures by marginal density decompositions. *Journal of the American Statistical Association*, 96:1316–1332, 2001.
- L. F. James, C. E. Priebe, and D. J. Marchette. Consistent estimation of mixture complexity. *Annals of Statistics*, 29:1281–1296, 2001.
- R.J. Karunamuni and J. Wu. Minimum hellinger distance estimation in a nonparametric mixture model. *Journal of Statistical Planning and Inference*, 139:1118–1133, 2009.
- H. Kasahara and K. Shimotsu. Non-parametric identification and estimation of the number of components in multivariate mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 76:97–111, 2014.
- C. Keribin. Consistent estimation of the order of mixture models. *Sankhya Series A*, 62:49–66, 2000.
- B. G. Leroux. Consistent estimation of a mixing distribution. *Annals of Statistics*, 20:1350–1360, 1992.
- N. Lin and X. He. Robust and efficient estimation under data grouping. *Biometrika*, 93:99–112, 2006.

- B. Lindsay. *Mixture models: Theory, Geometry and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, CA, 1995.
- B. G. Lindsay. Efficiency versus robustness: The case for minimum hellinger distance and related methods. *Annals of Statistics*, 22:1081–1114, 1994.
- G. J. McLachlan and K. E. Basford. *Mixture models: Inference and Applications to Clustering. Statistics: Textbooks and Monographs*. New York, 1988.
- G. J. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics, 2000.
- J. Miller and D. Dunson. Robust bayesian inference via coarsening. *arXiv:1506.06101*, 2015.
- X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400, 2013.
- K. Pearson. Contributions to the theory of mathematical evolution. *Philosophical Transactions of the Royal Society of London A*, 185:71–110, 1894.
- S. Richardson and P. J. Green. On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B (Methodological)*, 59:731–792, 1997.
- K. Roeder. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statistical Association*, 89:487–495, 1994.
- H. Teicher. Identifiability of mixtures. *Annals of Statistics*, 32:244–248, 1961.
- Cédric Villani. *Optimal transport: Old and New*. Springer, 2008.
- M. Wiper, D. R. Insua, and F. Ruggeri. Mixtures of gamma distributions with applications. *Journal of Computational and Graphical Statistics*, 10:440–454, 2001.
- M. Woo and T. N. Sriram. Robust estimation of mixture complexity. *Journal of the American Statistical Association*, 101:1475–1486, 2006.
- W. A. Woodward, W. C. Parr, W. R. Schucany, and H. Lindsey. A comparison of minimum distance and maximum likelihood estimation of a mixture proportion. *Journal of the American Statistical Association*, 79:590–598, 1984.

Appendix A

In this Appendix, we provide the proofs of several key results in Section 3 and Section 4.

PROOF OF LEMMA 3.1 The proof of this lemma is a straightforward application of the Fourier transform. In fact, for any finite k different elements $\theta_1, \dots, \theta_k \in \Theta$, assume that we have $\alpha_i \in \mathbb{R}, \beta_i \in \mathbb{R}^{d_1}$ (for all $i = 1, \dots, k$) such that for almost all x

$$\sum_{i=1}^k \alpha_i f_0 * K_{\sigma_0}(x|\theta_i) + \beta_i^T \frac{\partial f_0 * K_{\sigma_0}}{\partial \theta}(x|\theta_i) = 0,$$

By means of the Fourier transformation on both sides of the above equation, we obtain for all $t \in \mathbb{R}^d$ that

$$\widehat{K_{\sigma_0}}(t) \left(\sum_{i=1}^k \alpha_i \widehat{f_0}(t|\theta_i) + \beta_i^T \frac{\partial \widehat{f_0}}{\partial \theta}(t|\theta_i) \right) = 0.$$

Since $\widehat{K_{\sigma_0}}(t) = \widehat{K}(\sigma_0 t) \neq 0$ for almost all $t \in \mathbb{R}^d$ and f is identifiable in the first order, we obtain that $\alpha_i = 0, \beta_i = \mathbf{0} \in \mathbb{R}^{d_1}$ for all $1 \leq i \leq k$. We achieve the conclusion of this lemma.

PROOF OF LEMMA 3.4 We denote the following weighted version of the total variation distance

$$V^*(p_{G_1, f} * K_{\sigma_1}, p_{G_2, f} * K_{\sigma_1}) = \frac{1}{2} \int |p_{G_1, f} * K_{\sigma_1}(x) - p_{G_2, f} * K_{\sigma_1}(x)| \times \left(\frac{p_{G_0, f_0} * K_{\sigma_0}(x)}{p_{G_*, f} * K_{\sigma_1}(x)} \right)^{1/4} dx.$$

for any two mixing measures $G_1, G_2 \in \overline{\mathcal{G}}$. By Holder's inequality, we have

$$\begin{aligned} V^*(p_{G, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) &\leq \frac{1}{\sqrt{2}} h^*(p_{G, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) \\ &\quad \times \left(\int \left(\sqrt{p_{G, f} * K_{\sigma_1}(x)} + \sqrt{p_{G_*, f} * K_{\sigma_1}(x)} \right)^2 dx \right)^{1/2} \\ &\leq \sqrt{2} h^*(p_{G, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}). \end{aligned} \quad (14)$$

Therefore, in order to obtain the conclusion of the lemma it suffices to demonstrate that

$$\inf_{G \in \mathcal{O}_{k_*}} V^*(p_{G, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) / W_1(G, G_*) > 0. \quad (15)$$

Firstly, we will show that

$$\liminf_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_{k_*}} \left\{ \frac{V^*(p_{G, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1})}{W_1(G, G_*)} : W_1(G, G_*) \leq \epsilon \right\} > 0.$$

Assume that the above inequality does not hold. There exists a sequence $G_n \in \mathcal{O}_{k_*}$ such that $W_1(G_n, G_*) \rightarrow 0$ and $V^*(p_{G_n, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) / W_1(G_n, G_*) \rightarrow 0$. By means of Fatou's lemma, we obtain

$$\begin{aligned} 0 &= \liminf_{n \rightarrow \infty} \frac{V^*(p_{G_n, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1})}{W_1(G_n, G_*)} \\ &\geq \frac{1}{2} \int \liminf_{n \rightarrow \infty} \frac{|p_{G_n, f} * K_{\sigma_1}(x) - p_{G_*, f} * K_{\sigma_1}(x)| \left(\frac{p_{G_0, f_0} * K_{\sigma_0}(x)}{p_{G_*, f} * K_{\sigma_1}(x)} \right)^{1/4}}{W_1(G_n, G_*)} dx. \end{aligned}$$

Therefore, for almost surely $x \in \mathcal{X}$, we have

$$\liminf_{n \rightarrow \infty} \frac{|p_{G_n, f} * K_{\sigma_1}(x) - p_{G_*, f} * K_{\sigma_1}(x)| \left(\frac{p_{G_0, f_0} * K_{\sigma_0}(x)}{p_{G_*, f} * K_{\sigma_1}(x)} \right)^{1/4}}{W_1(G_n, G_*)} = 0. \quad (16)$$

Since $W_1(G_n, G_*) \rightarrow 0$ and $G_n \in \mathcal{O}_{k_*}$, we can find a subsequence of k_n such that $k_n = k_*$. Without loss of generality, we replace that subsequence of k_n by its whole sequence. Then, G_n will have exactly k_* components for all $n \geq 1$. From here, by using the same argument as that in the proof of Theorem 3.1 in Ho and Nguyen [2016c], equality (16) cannot happen – a contradiction.

Therefore, we can find a positive constant number ϵ_0 such that $V^*(p_{G, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) \gtrsim W_1(G, G_*)$ for any $W_1(G, G_*) \leq \epsilon_0$. Now, to obtain the conclusion of (15), we only need to verify that

$$\inf_{G \in \mathcal{O}_{k_*}: W_1(G, G_*) > \epsilon_0} V^*(p_{G, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) / W_1(G, G_*) > 0.$$

In fact, if the above statement does not hold, we can find a sequence $G'_n \in \mathcal{O}_{k_*}$ such that $W_1(G'_n, G_*) > \epsilon_0$ and $V^*(p_{G'_n, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) / W_1(G'_n, G_*) \rightarrow 0$. Since Θ is a closed bounded set, we can find $G' \in \mathcal{O}_{k_*}$ such that a subsequence of G'_n satisfies $W_1(G'_n, G') \rightarrow 0$ and $W_1(G', G_*) > \epsilon_0$. Without loss of generality, we replace that subsequence of G'_n by its whole sequence. Therefore, $V^*(p_{G'_n, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) \rightarrow 0$ as $n \rightarrow \infty$. Since $W_1(G'_n, G') \rightarrow 0$, due to the first order uniform Lipschitz property of $f * K_{\sigma_1}$ we obtain $p_{G'_n, f} * K_{\sigma_1}(x) \rightarrow p_{G', f} * K_{\sigma_1}(x)$ for almost all $x \in \mathcal{X}$ when $n \rightarrow \infty$. Now, by means of Fatou's lemma

$$\begin{aligned} 0 &= \lim_{n \rightarrow \infty} V^*(p_{G'_n, f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}) \\ &\geq \int \liminf_{n \rightarrow \infty} |p_{G'_n, f} * K_{\sigma_1}(x) - p_{G_*, f} * K_{\sigma_1}(x)| \left(\frac{p_{G_0, f_0} * K_{\sigma_0}(x)}{p_{G_*, f} * K_{\sigma_1}(x)} \right)^{1/4} dx \\ &= V^*(p_{G', f} * K_{\sigma_1}, p_{G_*, f} * K_{\sigma_1}), \end{aligned}$$

which only happens when $p_{G', f} * K_{\sigma_1}(x) = p_{G_*, f} * K_{\sigma_1}(x)$ for almost surely x . Due to the identifiability of $f * K_{\sigma_1}$, the former equality implies that $G' \equiv G_*$, which is a contradiction to the condition that $W_1(G', G_*) > \epsilon_0$. As a consequence, we achieve the conclusion of the lemma.

PROOF OF PROPOSITION 3.1 For the simplicity of presentation we implicitly denote $\widehat{G}_{n, m} = \widehat{G}_{n, m}(\sigma_{0, n})$ and $G_{0, m} = G_{0, m}(\sigma_{0, n}) = \arg \min_{G \in \mathcal{O}_m} h(p_{G, f_0} * K_{\sigma_{0, n}}, p_{G_0, f_0} * K_{\sigma_{0, n}})$ for any element

$\sigma_{0, n} > 0$, i.e., both $\widehat{G}_{n, m}$ and $G_{0, m}$ strictly depend on $\sigma_{0, n}$ and will vary as long as $\sigma_{0, n} \rightarrow 0$. Now, as $n \rightarrow \infty$, we will prove for almost surely that

$$h(p_{\widehat{G}_{n, m}, f_0} * K_{\sigma_{0, n}}, P_n * K_{\sigma_{0, n}}) - h(p_{\widehat{G}_{n, m+1}, f_0} * K_{\sigma_{0, n}}, P_n * K_{\sigma_{0, n}}) \rightarrow d'_m, \quad (17)$$

where $d'_m = h(p_{\widetilde{G}_{0, m}, f_0}, p_{G_0, f_0}) - h(p_{\widetilde{G}_{0, m+1}, f_0}, p_{G_0, f_0})$ where $\widetilde{G}_{0, m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G, f_0}, p_{G_0, f_0})$. To achieve this result, we start with the following lemma

Lemma .1. *For any sequence G_n and $\sigma_n \rightarrow 0$, we have as $n \rightarrow \infty$ that*

$$h(p_{G_n, f_0} * K_{\sigma_n}, p_{G_n, f_0}) \rightarrow 0.$$

The proof of this lemma is deferred to Appendix B. Now, applying the result of Lemma .1 to the sequences $G_{0,m}$ and $\sigma_{0,n}$, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} h(p_{G_{0,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}) &= \lim_{n \rightarrow \infty} h(p_{G_{0,m},f_0}, p_{G_{0,f_0}}) \\ &\geq h(p_{\tilde{G}_{0,m},f_0}, p_{G_{0,f_0}}). \end{aligned} \quad (18)$$

On the other hand, from the definition of $G_{0,m}$, we have

$$h(p_{G_{0,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}) \leq h(p_{\tilde{G}_{0,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}).$$

Therefore,

$$\begin{aligned} \lim_{n \rightarrow \infty} h(p_{G_{0,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}) &\leq \lim_{n \rightarrow \infty} h(p_{\tilde{G}_{0,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}) \\ &= h(p_{\tilde{G}_{0,m},f_0}, p_{G_{0,f_0}}). \end{aligned} \quad (19)$$

Combining the results from (18) and (19), we have

$$\lim_{n \rightarrow \infty} h(p_{G_{0,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}) = h(p_{\tilde{G}_{0,m},f_0}, p_{G_{0,f_0}}). \quad (20)$$

Now, we will demonstrate that

$$\lim_{n \rightarrow \infty} h(p_{\hat{G}_{n,m},f_0} * K_{\sigma_{0,n}}, P_n * K_{\sigma_{0,n}}) = \lim_{n \rightarrow \infty} h(p_{G_{0,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}). \quad (21)$$

In fact, from the definition of $\hat{G}_{n,m}$ we quickly obtain that

$$\begin{aligned} \lim_{n \rightarrow \infty} h(p_{\hat{G}_{n,m},f_0} * K_{\sigma_{0,n}}, P_n * K_{\sigma_{0,n}}) &\leq \lim_{n \rightarrow \infty} h(p_{G_{0,m},f_0} * K_{\sigma_{0,n}}, P_n * K_{\sigma_{0,n}}) \\ &= \lim_{n \rightarrow \infty} h(p_{G_{0,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}) \end{aligned} \quad (22)$$

where the last equality is due to the fact that $h(P_n * K_{\sigma_{0,n}}, p_{G_{0,f_0}}) \rightarrow 0$ almost surely as $n \rightarrow \infty$, $\sigma_{0,n} \rightarrow 0$ and $n\sigma_{0,n}^d \rightarrow \infty$. On the other hand, from the formulation of $G_{0,m}$ we have

$$\begin{aligned} \lim_{n \rightarrow \infty} h(p_{G_{0,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}) &\leq \lim_{n \rightarrow \infty} h(p_{\hat{G}_{n,m},f_0} * K_{\sigma_{0,n}}, p_{G_{0,f_0}} * K_{\sigma_{0,n}}) \\ &= \lim_{n \rightarrow \infty} h(p_{\hat{G}_{n,m},f_0} * K_{\sigma_{0,n}}, P_n * K_{\sigma_{0,n}}) \end{aligned} \quad (23)$$

Combining (22) and (23), we obtain equality (21). Now, the combination of (20) and (21) leads to

$$\lim_{n \rightarrow \infty} h(p_{\hat{G}_{n,m},f_0} * K_{\sigma_{0,n}}, P_n * K_{\sigma_{0,n}}) = h(p_{\tilde{G}_{0,m},f_0}, p_{G_{0,f_0}}).$$

Therefore, we obtain the conclusion of (17). From here, by using the same argument as Step 1 of Theorem 3.1, we ultimately get $d'_m = 0$ as $m \geq k_0$ and $d'_m > 0$ as $m < k_0$. As a consequence, $\hat{m}_n \rightarrow k_0$ almost surely as $n \rightarrow \infty$. The conclusion of the proposition follows.

PROOF OF LEMMA 3.2 The proof proceeds by using the idea from Leroux's argument [Leroux, 1992]. In fact, from the definition of G_* , we have $h(p_{G_*,f} * K_{\sigma_1}, p_{G_{0,f_0}} * K_{\sigma_0}) \leq h(p_{G,f} * K_{\sigma_1}, p_{G_{0,f_0}} * K_{\sigma_0})$ for any $G \in \bar{\mathcal{G}}$. Now, for any $\theta \in \Theta$, by choosing $G = (1 - \epsilon)G_* + \epsilon\delta_\theta$ and letting $\epsilon \rightarrow 0$ as in Step 1 of the proof of Theorem 3.1, we eventually obtain

$$\begin{aligned} &\int (p_{G_{0,f_0}} * K_{\sigma_0}(x))^{1/2} (p_{G_*,f} * K_{\sigma_1}(x))^{1/2} dx \\ &\geq \int (p_{G_{0,f_0}} * K_{\sigma_0}(x))^{1/2} f * K_{\sigma_1}(x|\theta) (p_{G_*,f} * K_{\sigma_1}(x))^{-1/2} dx. \end{aligned}$$

By choosing $\theta \in \text{supp}(G_*)$ and summing over all of these components, we readily obtain inequality (1), which concludes the result of the lemma.

PROOF OF LEMMA 3.3 By means of Holder inequality, we obtain

$$\begin{aligned} \left(\int \sqrt{p_{G_{1,*},f} * K_{\sigma_1}(x)} \sqrt{p_{G_{0,f_0}} * K_{\sigma_0}(x)} dx \right)^2 &\leq \int p_{G_{1,*},f} * K_{\sigma_1}(x) \sqrt{\frac{p_{G_{0,f_0}} * K_{\sigma_0}(x)}{p_{G_{2,*},f} * K_{\sigma_1}(x)}} dx \\ &\quad \times \int \sqrt{p_{G_{2,*},f} * K_{\sigma_1}(x)} \sqrt{p_{G_{0,f_0}} * K_{\sigma_0}(x)} dx. \end{aligned}$$

From the definition of $G_{1,*}$ and $G_{2,*}$, we have

$$\int \sqrt{p_{G_{1,*},f} * K_{\sigma_1}(x)} \sqrt{p_{G_{0,f_0}} * K_{\sigma_0}(x)} dx = \int \sqrt{p_{G_{2,*},f} * K_{\sigma_1}(x)} \sqrt{p_{G_{0,f_0}} * K_{\sigma_0}(x)} dx.$$

Therefore, the above inequality along with innequality (1) in Lemma 3.2 lead to

$$\int p_{G_{1,*},f} * K_{\sigma_1}(x) \sqrt{\frac{p_{G_{0,f_0}} * K_{\sigma_0}(x)}{p_{G_{2,*},f} * K_{\sigma_1}(x)}} dx = \int \sqrt{p_{G_{2,*},f} * K_{\sigma_1}(x)} \sqrt{p_{G_{0,f_0}} * K_{\sigma_0}(x)} dx.$$

It eventually implies that

$$\int \left(\sqrt{p_{G_{1,*},f} * K_{\sigma_1}(x)} - \sqrt{p_{G_{2,*},f} * K_{\sigma_1}(x)} \right)^2 \sqrt{\frac{p_{G_{0,f_0}} * K_{\sigma_0}(x)}{p_{G_{2,*},f} * K_{\sigma_1}(x)}} dx = 0.$$

Therefore, $p_{G_{1,*},f} * K_{\sigma_1}(x) = p_{G_{2,*},f} * K_{\sigma_1}(x)$ almost surely $x \in \mathcal{X}$. As a consequence, we obtain the conclusion of the lemma.

PROOF OF THEOREM 4.1 Here, we only provide the proof for part (b) as it is the generalization of part (a). The proof is similar to that in Step 1 of Theorem 3.2. In fact, as $n \rightarrow \infty$ we have for almost surely that

$$h(p_{\hat{G}_{n,m},f} * K_{\sigma_1}, P_n * K_{\sigma_0}) \rightarrow h(p_{G_{*,m},f} * K_{\sigma_1}, p_{G_{0,f_0}} * K_{\sigma_0}) \quad (24)$$

where $G_{*,m} = \arg \min_{G \in \mathcal{O}_m} h(p_{G,f} * K_{\sigma_1}, p_{G_{0,f_0}} * K_{\sigma_0})$. From the argument of Step 1 in the proof of Theorem 3.2, we have

$$h(p_{G_{*,m+1},f} * K_{\sigma_1}, p_{G_{0,f_0}} * K_{\sigma_0}) < h(p_{G_{*,m},f} * K_{\sigma_1}, p_{G_{0,f_0}} * K_{\sigma_0}) \quad (25)$$

for any $1 \leq m \leq k_* - 1$. It implies that $G_{*,m} \in \mathcal{E}_m$ for all $1 \leq m \leq k_*$. Now, if we would like to have $\tilde{m}_n \rightarrow k_*$ as $n \rightarrow \infty$, the sufficient and necessary condition is

$$h(p_{G_{*,f}} * K_{\sigma_1}, p_{G_{0,f_0}} * K_{\sigma_0}) \leq \epsilon < h(p_{G_{*,k_*-1},f} * K_{\sigma_1}, p_{G_{0,f_0}} * K_{\sigma_0}),$$

which is precisely the conclusion of the theorem.

PROOF OF PROPOSITION 4.1 Using the argument from Step 1 in the proof of Theorem 3.1, we obtain that G_{0,k_0-1} has exactly $k_0 - 1$ elements. Now, since $f_0 * K_{\sigma_0}$ is uniformly Lipschitz up to the first order and identifiable, we obtain

$$\inf_{G \in \mathcal{E}_{k_0-1}} h(p_{G,f_0} * K_{\sigma_0}, p_{G_{0,f_0}} * K_{\sigma_0}) / W_1(G, G_0) = C'$$

where C' is some positive constant depending only on f_0, G_0, Θ , and σ_0 . Therefore, we get

$$h(p_{G_0, k_0-1, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \geq C' W_1(G_0, G_{0, k_0-1}) \geq C' \inf_{G \in \mathcal{E}_{k_0-1}} W_1(G, G_0). \quad (26)$$

Now, for any $G = \sum_{i=1}^{k_0-1} p_i \delta_{\theta_i} \in \mathcal{E}_{k_0-1}$, we can find the index $j^* \in [1, k_0]$ such that

$$\|\theta_i - \theta_{j^*}^0\| \geq \min_{1 \leq j \neq j^* \leq k_0} \|\theta_i - \theta_j^0\|$$

for any $1 \leq i \leq k_0 - 1$. Therefore, we obtain

$$2\|\theta_i - \theta_{j^*}^0\| \geq \|\theta_i - \theta_{j^*}^0\| + \min_{1 \leq j \neq j^* \leq k_0} \|\theta_i - \theta_j^0\| \geq \min_{1 \leq u \neq v \leq k_0} \|\theta_u^0 - \theta_v^0\|$$

for any $1 \leq i \leq k_0 - 1$. From the definition of $W_1(G, G_0)$, we can find the optimal coupling $\mathbf{q} \in \mathcal{Q}(\mathbf{p}, \mathbf{p}^0)$ such that $W_1(G, G_0) = \sum_{i,j} q_{ij} \|\theta_i - \theta_j^0\|$. Hence, we get

$$\begin{aligned} W_1(G, G_0) &\geq \sum_{i=1}^{k_0} q_{ij^*} \|\theta_i - \theta_{j^*}^0\| \geq p_{j^*}^0 \min_{1 \leq i \leq k_0-1} \|\theta_i - \theta_{j^*}^0\| \\ &\geq \left(\min_{1 \leq i \leq k_0} p_i^0 \times \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \right) / 2 \end{aligned}$$

for all $G \in \mathcal{E}_{k_0-1}$. It implies that

$$\inf_{G \in \mathcal{E}_{k_0-1}} W_1(G, G_0) \geq \left(\min_{1 \leq i \leq k_0} p_i^0 \times \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \right) / 2. \quad (27)$$

By combining (26) and (27), if we choose $\min_{1 \leq i \leq k_0} p_i^0 \min_{1 \leq i \neq j \leq k_0} \|\theta_i^0 - \theta_j^0\| \geq 2\epsilon/C'$, then $\epsilon < h(p_{G_0, k_0-1, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0})$. As a consequence, by defining $C = 1/C'$ we obtain the conclusion of the lemma.

PROOF OF PROPOSITION 5.3 The proof of this proposition is a straightforward combination of Fatou's lemma and the argument from Theorem 4.6 in [Heinrich and Kahn, 2015]. In fact, for any $\epsilon > 0$, as $W_1(G_0^n, \tilde{G}_0) \rightarrow 0$ we can find $M(\epsilon) \in \mathbb{N}$ such that $W_1(G_0^n, \tilde{G}_0) < \epsilon$ for any $n \geq M(\epsilon)$. Additionally, as $\limsup_{n \rightarrow \infty} k_0^n = k$, we can find $T(\epsilon) \in \mathbb{N}$ such that $k_0^n \leq k$ for all $n \geq T(\epsilon)$. Denote $N(\epsilon) = \max \{M(\epsilon), T(\epsilon)\}$ for any $\epsilon > 0$. Now, assume that for any $\epsilon > 0$, we have

$$\inf_{G \in \mathcal{O}_{k_0^n}: W_1(G, \tilde{G}_0) < \epsilon} h(p_{G, f_0} * K_{\sigma_0}, p_{G_0^n, f_0} * K_{\sigma_0}) / W_1^{2k-2\tilde{k}_0+1}(G, G_0^n) = 0.$$

as long as $n \geq N(\epsilon)$. For each $n \geq N(\epsilon)$, it implies that we have a sequence $G_m^n \in \mathcal{O}_{k_0^n} \subset \mathcal{O}_k$ such that $W_1(G_m^n, \tilde{G}_0) < \epsilon$ for all $m \geq 1$ and

$$h(p_{G_m^n, f_0} * K_{\sigma_0}, p_{G_0^n, f_0} * K_{\sigma_0}) / W_1^{2k-2\tilde{k}_0+1}(G_m^n, G_0^n) \rightarrow 0$$

as $m \rightarrow \infty$. By means of Fatou's lemma, we eventually have

$$\liminf_{m \rightarrow \infty} (p_{G_m^n, f_0} * K_{\sigma_0}(x) - p_{G_0^n, f_0} * K_{\sigma_0}(x)) / W_1^{2k-2\tilde{k}_0+1}(G_m^n, G_0^n) \rightarrow 0$$

almost surely $x \in \mathcal{X}$. However, from the argument of Theorem 4.6 in Heinrich and Kahn [2015], we can find $\epsilon_0 > 0$ such that for all $G_m^n, G_0^n \in \mathcal{O}_k$ where $W_1(G_m^n, \tilde{G}_0) \vee W_1(G_0^n, \tilde{G}_0) < \epsilon_0$, not almost surely $x \in \mathcal{X}$ that $(p_{G_m^n, f_0} * K_{\sigma_0}(x) - p_{G_0^n, f_0} * K_{\sigma_0}(x)) / W_1^{2k-2\tilde{k}_0+1}(G_m^n, G_0^n) \rightarrow 0$ for each $n \geq N(\epsilon_0)$, which is a contradiction. Therefore, we achieve the conclusion of the proposition.

Appendix B

This appendix contains remaining proofs of the main results in the paper.

PROOF OF PROPOSITION 2.1 A careful investigation of the proof of Theorem 3.1 in Ho and Nguyen [2016c] implies that

$$h(p_{G,f}, p_{G_0,f}) \gtrsim W_1(G, G_0), \quad (28)$$

for any $G \in \mathcal{O}_{k_0}$ such that $W_1(G, G_0)$ is sufficiently small. The latter restriction means that the result in (28) is of a local nature. We also would like to extend this lower bound of $h(p_{G,f}, p_{G_0,f})$ for any $G \in \mathcal{O}_{k_0}$. It appears that the first order Lipschitz continuity of f is sufficient to extend (28) for any $G \in \mathcal{O}_{k_0}$. In fact, by the result in (28), we can find a positive constant ϵ_0 such that

$$\inf_{G \in \mathcal{O}_{k_0} : W_1(G, G_0) \leq \epsilon_0} h(p_{G,f}, p_{G_0,f}) / W_1(G, G_0) > 0$$

Therefore, to extend (28) for any $G \in \mathcal{O}_{k_0}$, it is sufficient to demonstrate that

$$\inf_{G \in \mathcal{O}_{k_0} : W_1(G, G_0) > \epsilon_0} h(p_{G,f}, p_{G_0,f}) / W_1(G, G_0) > 0$$

Assume by the contrary that the above result does not hold. It implies that we can find a sequence $G_n \in \mathcal{O}_{k_0}$ such that $W_1(G_n, G_0) > \epsilon_0$ and $h(p_{G_n,f}, p_{G_0,f}) / W_1(G_n, G_0) \rightarrow 0$ as $n \rightarrow \infty$. Since Θ is a compact set, we can find $G' \in \mathcal{O}_{k_0}$ such that a subsequence of G_n satisfies $W_1(G_n, G') \rightarrow 0$ and $W_1(G', G_0) > \epsilon_0$. Without loss of generality, we replace that subsequence by its whole sequence. Therefore, $h(p_{G_n,f}, p_{G_0,f}) \rightarrow 0$ as $n \rightarrow \infty$. Due to the first order Lipschitz continuity of f , we obtain $p_{G_n,f}(x) \rightarrow p_{G',f}(x)$ for any $x \in \mathcal{X}$ when $n \rightarrow \infty$. Now, by means of Fatou's lemma, we have

$$0 = \lim_{n \rightarrow \infty} h(p_{G_n,f}, p_{G_0,f}) \geq \int \liminf_{n \rightarrow \infty} \left(\sqrt{p_{G_n,f}(x)} - \sqrt{p_{G_0,f}(x)} \right)^2 dx = h(p_{G',f}, p_{G_0,f}).$$

Since f is identifiable, the above result implies that $G' \equiv G_0$, which contradicts the assumption that $W_1(G', G_0) > \epsilon_0$. As a consequence, we can extend inequality (28) for any $G \in \mathcal{O}_{k_0}$ when f is uniformly Lipschitz up to the first order.

PROOF OF LEMMA .1 The proof idea of this lemma is similar to that of Theorem 1 in Chapter 2 of [Devroye and Györfi, 1985]. However, it is slightly more complex than the proof of this theorem as we allow G_n to vary when σ_n vary. Here, we provide the proof of this lemma for the completeness. Since the Hellinger distance is upper bound by the total variation distance, it is sufficient to demonstrate that $V(p_{G_n, f_0} * K_{\sigma_n}, p_{G_n, f_0}) \rightarrow 0$ as $n \rightarrow \infty$. Firstly, assume that $f_0(x|\theta)$ is continuous and vanishes outside a compact set which is independent of θ and Σ . For any large number M , we split $K = K' + K''$ where $K' = K1_{\|x\| \leq M}$ and $K'' = K1_{\|x\| > M}$. Now, by using Young's inequality we obtain

$$\begin{aligned} & \int |p_{G_n, f_0} * K_{\sigma_n}(x) - p_{G_n, f_0}(x)| dx \\ & \leq \int \left| p_{G_n, f_0} * K'_{\sigma_n}(x) - p_{G_n, f_0}(x) \int K'_{\sigma_n}(y) dy \right| dx \\ & \quad + \int |p_{G_n, f_0} * K''_{\sigma_n}(x)| dx + \int p_{G_n, f_0}(x) dx \int K''_{\sigma_n}(x) dx \\ & \leq \int \left| p_{G_n, f_0} * K_{\sigma_n}(x) - p_{G_n, f_0}(x) \int K'_{\sigma_n}(y) dy \right| dx + 2 \int K''_{\sigma_n}(x) dx. \end{aligned}$$

for some compact set A . It is clear that for any $\epsilon > 0$, we can choose $M(\epsilon)$ such that as $M > M(\epsilon)$, $\int K''_{\sigma_n}(x)dx = \int K''(x)dx < \epsilon$. Regarding the first term in the right hand side of the above display, by denoting $G_n = \sum_{i=1}^{m_n} p_i^n \delta_{\theta_i^n}$ we obtain

$$\begin{aligned} & \int_A \left| p_{G_n, f_0} * K_{\sigma_n}(x) - p_{G_n, f_0}(x) \int K'_{\sigma_n}(y)dy \right| dx \\ & \leq \int_A \int |p_{G_n, f_0}(x-y) - p_{G_n, f_0}(x)| K'_{\sigma_n}(y) |dy| dx \\ & \leq \int_A \int \sum_{i=1}^{m_n} p_i^n |f_0(x-y|\theta_i^n) - f_0(x|\theta_i^n)| K'_{\sigma_n}(y) |dy| dx \\ & \leq \omega(M\sigma_n) \int_A \int |K'_{\sigma_n}(y)| dy dx \leq \omega(M\sigma_n) \mu(A) \rightarrow 0 \end{aligned}$$

where $\omega(t) = \sup_{\|x-y\| \leq t} |f_0(x|\theta) - f_0(y|\theta)|$ denotes the modulus of continuity of f_0 and μ denotes the Lebesgue measure. Therefore, the conclusion of this lemma holds for that setting of $f_0(x|\theta)$.

Regarding the general setting of $f_0(x|\theta)$, for any $\epsilon > 0$ since Θ is a bounded set, we can find a continuous function $g(x|\theta)$ being supported on a compact set $B(\epsilon)$ that is independent of $\theta \in \Theta$ such that $\int |f_0(x|\theta) - g(x|\theta)| dx < \epsilon$. Hence, we obtain

$$\begin{aligned} \int |p_{G_n, f_0} * K_{\sigma_n}(x) - p_{G_n, f_0}(x)| dx & \leq \int \left| (p_{G_n, f_0} - p_{G_n, g}) * K_{\sigma_n}(x) \right| dx \\ & \quad + \int |p_{G_n, f_0}(x) - p_{G_n, g}(x)| dx + \int |p_{G_n, g} * K_{\sigma_n}(x) - p_{G_n, g}(x)| dx \\ & \leq 2\epsilon + \int |p_{G_n, g} * K_{\sigma_n}(x) - p_{G_n, g}(x)| dx \end{aligned}$$

where $\int |p_{G_n, g} * K_{\sigma_n}(x) - p_{G_n, g}(x)| dx \rightarrow 0$ as $n \rightarrow \infty$. We achieve the conclusion of the lemma.

Lemma .2. Assume that f_0 and K satisfy condition (P.1) in Theorem 3.1. Furthermore, K has an integrable radial majorant $\Psi \in L_1(\mu)$ where $\Psi(x) = \sup_{\|y\| \geq \|x\|} |K(y)|$. Then, we can find a positive constant ϵ_1^0 such that as $\sigma \leq \epsilon_1^0$, for any $G \in \mathcal{O}_{k_0}$ we have

$$h(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \geq V(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \gtrsim W_1(G, G_0).$$

, i.e., $C_1(\sigma_0) \geq C_1$ as $\sigma_0 \rightarrow 0$ where C_1 only depends on G_0 .

Proof. We divide the proof of this lemma into two key steps

Step 1: We firstly demonstrate the following result

$$\lim_{\epsilon \rightarrow 0} \inf_{G \in \mathcal{O}_{k_0}, \sigma_0 > 0} \left\{ \frac{V(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0})}{W_1(G, G_0)} : W_1(G, G_0) \vee \sigma_0 \leq \epsilon \right\} > 0. \quad (29)$$

The proof idea of the above inequality is essentially similar to that from the proof of Theorem 3.1 in [Ho and Nguyen, 2016c]. Here, we provide such proof for the completeness. Assume that the conclusion of inequality (29) does not hold. Therefore, we can find two sequences $\{G_n\}$ and $\{\sigma_{0,n}\}$ such that $V(p_{G_n, f_0} * K_{\sigma_{0,n}}, p_{G_0, f_0} * K_{\sigma_{0,n}})/W_1(G_n, G_0) \rightarrow 0$ where $W_1(G_n, G_0) \rightarrow 0$ and $\sigma_{0,n} \rightarrow 0$ as $n \rightarrow \infty$. As $G_n \in \mathcal{O}_{k_0}$, it implies that there exists a subsequence $\{G_{n_m}\}$ of $\{G_n\}$ such that G_{n_m} has exactly k_0 elements for all m . Without loss of generality, we replace this subsequence by the whole sequence $\{G_n\}$. Now, we can represent G_n as $G_n = \sum_{i=1}^{k_0} p_i^n \delta_{\theta_i^n}$ such that $(p_i^n, \theta_i^n) \rightarrow (p_i^0, \theta_i^0)$. Similar to the argument in Step 1 from the proof of Theorem 3.1 in Ho and Nguyen [2016c], we have $W_1(G_n, G_0) \lesssim d(G_n, G_0)$ where $d(G_n, G_0) = \sum_{i=1}^{k_0} p_i^n \|\Delta \theta_i^n\| + |\Delta p_i^n|$ and $\Delta p_i^n = p_i^n - p_i^0, \Delta \theta_i^n = \theta_i^n - \theta_i^0$ for all $1 \leq i \leq k_0$. It implies that $V(p_{G_n, f_0} * K_{\sigma_{0,n}}, p_{G_0, f_0} * K_{\sigma_{0,n}})/d(G_n, G_0) \rightarrow 0$.

Now, we denote $g_n(x|\theta) = \int f_0(x-y|\theta) K_{\sigma_{0,n}}(y) dy$ for all $\theta \in \Theta$. Similar to Step 2 from the proof of Theorem 3.1 in Ho and Nguyen [2016c], by means of Taylor expansion up to the first order we can represent

$$\frac{p_{G_n, f_0} * K_{\sigma_{0,n}}(x) - p_{G_0, f_0} * K_{\sigma_{0,n}}(x)}{d(G_n, G_0)} \asymp \frac{1}{d(G_n, G_0)} \left(\sum_{i=1}^{k_0} \Delta p_i^n g_n(x|\theta_i^0) + p_i^n \frac{\partial g_n}{\partial \theta}(x|\theta_i^0) \right)$$

which is the linear combinations of the elements of $g_n(x|\theta_i^0), \frac{\partial g_n}{\partial \theta}(x|\theta_i^0)$ for $1 \leq i \leq k_0$. Denote m_n to be the maximum of the absolute values of these coefficients. We can argue that $m_n \not\rightarrow 0$ as $n \rightarrow \infty$. Additionally, since K has an integrable radial majorant $\Psi \in L_1(\mu)$, from Theorem 3 in Chapter 2 of Devroye and Györfi [1985], we have $g_n(x|\theta) \rightarrow f_0(x|\theta)$ and $\frac{\partial g_n}{\partial \theta}(x|\theta) \rightarrow \frac{\partial f_0}{\partial \theta}(x|\theta)$ for almost surely x and for any $\theta \in \Theta$. Therefore, we obtain

$$\frac{1}{m_n} \frac{d_n \left(p_{G_n, f_0} * K_{\sigma_{0,n}}(x) - p_{G_0, f_0} * K_{\sigma_{0,n}}(x) \right)}{d(G_n, G_0)} \rightarrow \sum_{i=1}^{k_0} \alpha_i f_0(x|\theta_i^0) + \beta_i^T \frac{\partial f_0}{\partial \theta}(x|\theta_i^0)$$

where not all the elements of α_i, β_i equal to 0. Due to the first order identifiability of f_0 and the Fatou's lemma, $V(p_{G_n, f_0} * K_{\sigma_{0,n}}, p_{G_0, f_0} * K_{\sigma_{0,n}})/d(G_n, G_0) \rightarrow 0$ will lead to $\alpha_i = 0, \beta_i = \mathbf{0} \in \mathbb{R}^{d_1}$ for all $1 \leq i \leq k_0$, which is a contradiction. We achieve the conclusion of (29).

Step 2: The result of (29) implies that we can find a positive number ϵ_1^0 such that as $W_1(G, G_0) \vee \sigma_0 \leq \epsilon_1^0$, we have

$$h(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \geq V(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \gtrsim W_1(G, G_0). \quad (30)$$

In order to extend the above inequality to any $G \in \mathcal{O}_{k_0}$, it is sufficient to demonstrate that

$$\inf_{\sigma_0 < \epsilon_1^0, W_1(G, G_0) > \epsilon_1^0} \frac{h(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0})}{W_1(G, G_0)} > 0.$$

In fact, if the above result does not hold, we can find two sequences $G'_n \in \mathcal{O}_{k_0}$ and $\sigma'_{0,n}$ such that $W_1(G'_n, G_0) > \epsilon_1^0, \sigma'_{0,n} \leq \epsilon_1^0$ and $h(p_{G'_n, f_0} * K_{\sigma'_{0,n}}, p_{G_0, f_0} * K_{\sigma'_{0,n}})/W_1(G'_n, G_0) \rightarrow 0$ as $n \rightarrow \infty$. Since Θ is closed bounded set, we can find two subsequences $\{G'_{n_m}\}$ and $\{\sigma'_{0, n_m}\}$ of $\{G'_n\}$ and $\{\sigma'_{0,n}\}$

respectively such that $W_1(G'_{n_m}, G') \rightarrow 0$ and $|\sigma'_{0, n_m} - \sigma'| \rightarrow 0$ as $m \rightarrow \infty$ where $G' \in \mathcal{O}_{k_0}$ and $\sigma' \in [0, \epsilon_1^0]$.

Due to the first order Lipschitz continuity of $f * K_{\sigma'_{0, n_m}}$ for any $m \geq 1$, we achieve that

$$p_{G'_{n_m}, f_0} * K_{\sigma'_{0, n_m}}(x) \rightarrow p_{G', f_0} * K_{\sigma'}(x)$$

for any $x \in \mathcal{X}$. Here, $p_{G', f_0} * K_{\sigma'} = p_{G', f_0}$ when $\sigma' = 0$. Therefore, by utilizing the Fatou's argument, we obtain $h(p_{G', f_0} * K_{\sigma'}, p_{G_0, f_0} * K_{\sigma'}) = 0$, which implies that $G' \equiv G_0$, a contradiction. As a consequence, when $\sigma_0 \leq \epsilon_1^0$, for any $G \in \mathcal{O}_{k_0}$ we have

$$h(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \geq V(p_{G, f_0} * K_{\sigma_0}, p_{G_0, f_0} * K_{\sigma_0}) \gtrsim W_1(G, G_0).$$

We achieve the conclusion of the lemma. \square

Lemma .3. 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 . If $I(G_0, f_0)$ has r -th singularity level for some $r \geq 0$, then $I(G_0, f_0 * K_{\sigma_0})$ also has r -th singularity level for any $\sigma_0 > 0$.

Proof. Remind that, $I(G_0, f_0)$ has r -th singularity level is equivalent to the fact that G_0 is r -singular relative to the ambient space \mathcal{O}_{k_0} (cf. Definition 3.1 and Definition 3.2 in [Ho and Nguyen, 2016b]).

Now, for any $\rho \in \mathbb{N}$, given any sequence $G_n = \sum_{i=1}^{k_n} p_i^n \delta_{\theta_i^n} \in \mathcal{O}_{k_0}$ such that $G_n \rightarrow G_0$ in W_ρ metric. We can find a subsequence of G_n such that $k_n = k_0$ and each atoms of G_0 will have exactly one component of G_n converges to. Without loss of generality, we replace the subsequence of G_n by its whole sequence and relabel the atoms of G_n such that $(p_i^n, \theta_i^n) \rightarrow (p_i^0, \theta_i^0)$ for all $1 \leq i \leq k_0$. Denote $\Delta\theta_i^n = \theta_i^n - \theta_i^0$ and $\Delta p_i^n = p_i^n - p_i^0$ for all $1 \leq i \leq k_0$. From Definition 3.1 in Ho and Nguyen [2016b], a ρ -minimal form of G_n from Taylor expansion up to the order ρ satisfies

$$\frac{p_{G_n, f_0}(x) - p_{G_0, f_0}(x)}{W_\rho^\rho(G_n, G_0)} = \sum_{l=1}^{T_\rho} \left(\frac{\xi_l^{(\rho)}(G_n)}{W_\rho^\rho(G_n, G_0)} \right) H_l^{(\rho)}(x) + o(1), \quad (31)$$

for all x . Here, $H_l^{(\rho)}(x)$ are linearly independent functions of x for all l , and coefficients $\xi_l^{(\rho)}(G)$ are polynomials of the components of $\Delta\theta_i$ and Δp_i for l ranges from 1 to a finite T_ρ . From the above representation, we achieve

$$\frac{p_{G_n, f_0 * K_{\sigma_0}}(x) - p_{G_0, f_0 * K_{\sigma_0}}(x)}{W_\rho^\rho(G_n, G_0)} = \sum_{l=1}^{T_\rho} \left(\frac{\xi_l^{(\rho)}(G_n)}{W_\rho^\rho(G_n, G_0)} \right) H_l^{(\rho)} * K_{\sigma_0}(x) + o(1), \quad (32)$$

where $H_l^{(\rho)} * K_{\sigma_0}(x) = \int H_l^{(\rho)}(x - y) K_{\sigma_0}(y) dy$ for all $1 \leq l \leq T_\rho$. We will show that $H_l^{(\rho)} * K_{\sigma_0}(x)$ are linearly independent functions of x for all $1 \leq l \leq T_\rho$. In fact, assume that we can find the coefficients $\alpha_l \in \mathbb{R}$ such that

$$\sum_{l=1}^{T_\rho} \alpha_l H_l^{(\rho)} * K_{\sigma_0}(x) = 0$$

for all x . By means of Fourier transformation in both sides of the above equation, we obtain

$$\widehat{K}_{\sigma_0}(t) \left(\sum_{l=1}^{T_\rho} \alpha_l \widehat{H}_l^{(\rho)}(t) \right) = 0$$

for all $t \in \mathbb{R}^d$. As $\widehat{K}_{\sigma_0}(t) = \widehat{K}(\sigma_0 t) \neq 0$ for all $t \in \mathbb{R}^d$ and $H_l^{(\rho)}(x)$ for all l are linearly independent functions of x for all $1 \leq l \leq T_\rho$, the above equation implies that $\alpha_l = 0$ for all $1 \leq l \leq T_\rho$. Therefore, $H_l^{(\rho)} * K_{\sigma_0}(x)$ are linearly independent functions of x for all $1 \leq l \leq T_\rho$ and $\rho \in \mathbb{N}$.

According to Definition 3.2 in Ho and Nguyen [2016b], since $I(G_0, f_0)$ has r -th singularity level, it implies that for any sequence $G_n \in \mathcal{O}_{k_0}$ such that $W_{r+1}^{r+1}(G_n, G_0) \rightarrow 0$, we do not have all the ratios $\xi_l^{(r+1)}(G_n)/W_{r+1}^{r+1}(G_0, G_n)$ in (31) go to 0 for $1 \leq l \leq T_{r+1}$. It means that not all the ratios $\xi_l^{(r+1)}(G_n)/W_{r+1}^{r+1}(G_0, G_n)$ in (32) go to 0. Additionally, as $I(G_0, f_0)$ has r -th singularity level, we can find a sequence $G'_n \in \mathcal{O}_{k_0}$ such that $W_r^r(G'_n, G_0) \rightarrow 0$ and $\xi_l^{(r)}(G'_n)/W_r^r(G_0, G'_n)$ in (31) go to 0 for $1 \leq l \leq T_r$. It in turns also means that all the ratios $\xi_l^{(r)}(G'_n)/W_r^r(G_0, G'_n)$ in (32) go to 0. As a consequence, from Definition 3.3 in Ho and Nguyen [2016b], we achieve the conclusion of the lemma. \square