Weak separation in mixture models and implications for principal stratification*

Avi Feller, Evan Greif, Nhat Ho, Luke Miratrix, and Natesh Pillai

UC Berkeley and Harvard University

August 4, 2019

Abstract

Principal stratification is a widely used framework for addressing post-randomization complications. After using principal stratification to define causal effects of interest, researchers are increasingly turning to finite mixture models to estimate these quantities. Unfortunately, standard estimators of mixture parameters, like the MLE, are known to exhibit pathological behavior. We study this behavior in a simple but fundamental example, a two-component Gaussian mixture model in which only the component means and variances are unknown, and focus on the setting in which the components are weakly separated. In this case, we show that the asymptotic convergence rate of the MLE is quite poor, such as $O(n^{-1/6})$ or even $O(n^{-1/8})$. We then demonstrate via theoretical arguments as well as extensive simulations that, in finite samples, the MLE behaves like a threshold estimator, in the sense that the MLE can give strong evidence that the means are equal when the truth is otherwise. We also explore the behavior of the MLE when the MLE is non-zero, showing that it is difficult to estimate both the sign and magnitude of the means in this case. We provide diagnostics for all of these pathologies and apply these ideas to re-analyzing two randomized evaluations of job training programs, JOBS II and Job Corps. Our results suggest that the corresponding maximum likelihood estimates should be interpreted with caution in these cases.

*email: afeller@berkeley.edu. AF and LM gratefully acknowledge financial support from the Spencer Foundation through a grant entitled "Using Emerging Methods with Existing Data from Multi-site Trials to Learn About and From Variation in Educational Program Effects," and from the Institute for Education Science (IES Grant #R305D150040). NSP is partially supported by an ONR grant. We would like to thank Isaiah Andrews, Peter Aronow, Peter Bickel, Alex D'Amour, Peng Ding, Fabrizia Mealli, Christian Robert, Don Rubin, Dylan Small, Aaron Smith, Weixin Yao, and members of the Spencer group for helpful comments and discussion, as well as seminar participants at the Atlantic Causal Inference Conference and Joint Statistical Meetings. All opinions expressed in the paper and any errors that it might contain are solely the responsibility of the authors.

1 Introduction

Finite mixture models are notorious for giving pathological results (Redner and Walker, 1984); indeed, Larry Wasserman has called finite mixtures the "Twilight Zone of Statistics" (Wasserman, 2012). Our motivation for this paper is to understand how the pathological features of *weakly separated* finite mixture models affect inference for component means, especially with respect to estimating causal effects in the principal stratification framework, an important example of such inference.

Principal stratification is a widely used approach for addressing post-randomization complications, including noncompliance with treatment assignment (Frangakis and Rubin, 2002). Typically, the goal is to estimate causal effects within partially latent subgroups known as principal strata. While there are many possible ways to estimate these principal causal effects, the most common approach is via finite mixture models, treating the unknown principal strata as mixture components (Imbens and Rubin, 1997). To date, scores of applied and methodological papers have relied on finite mixtures to estimate causal effects, both explicitly and implicitly.

To present our main results, we construct a simple two-parameter model that captures the essential features of the problem: maximum likelihood estimation for the component means and variances in a two-component location-scale mixture of Gaussian distributions,

$$Y_i \overset{\text{iid}}{\sim} \pi N(\mu_0, \sigma_0) + (1-\pi)N(\mu_1, \sigma_1), \tag{1.1}$$

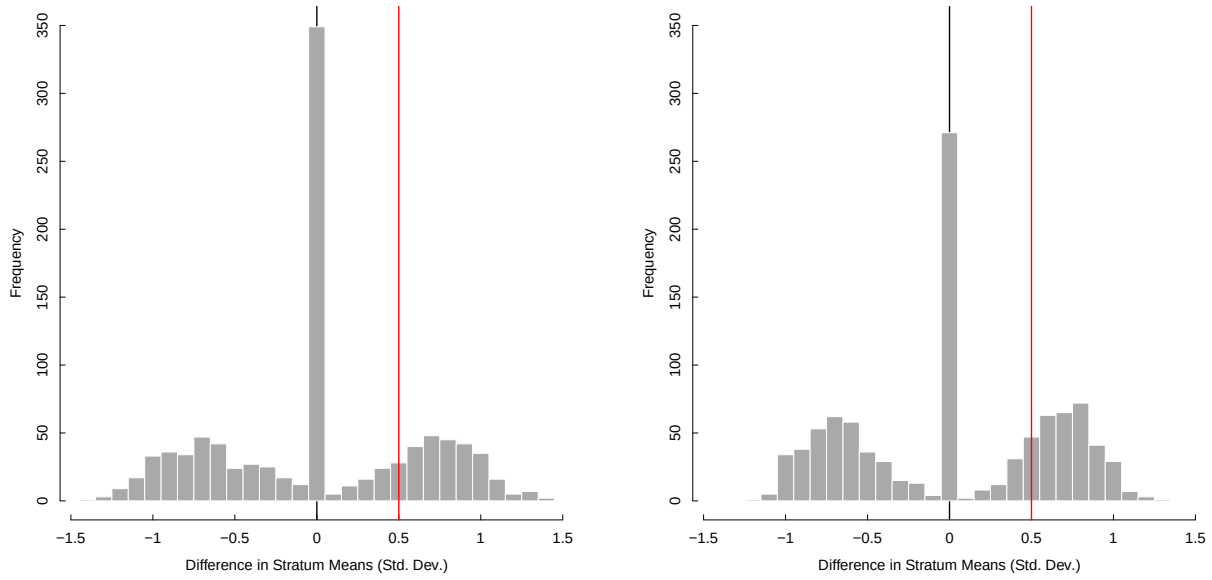
where the mixing proportion, $\pi \in (0, 1)$, is assumed to be known.

While the two-component finite mixture model in (1.1) is a toy example in some settings, it is a fundamental structure in many causal inference problems. For instance, in the canonical example of noncompliance in a randomized trial (Angrist et al., 1996), individuals randomly assigned to the treatment group who actually receive the treatment are

a mixture of Compliers and Always Takers. Assuming that individual outcomes follow a Normal distribution yields the mixture model in (1.1). Thus understanding the difficulties of component-specific inference are vital to estimating parametric principal stratification models.

The asymptotic properties of the MLE for the component means in Equation (1.1) are well established in two settings. First, when the difference in means, $\Delta \equiv \mu_1 - \mu_0$, is fixed, the MLE has strong asymptotic guarantees, including consistency and parametric convergence (Everitt and Hand, 1981; Chen, 2017). Second, when the mixture is degenerate, *i.e.*, $\Delta = 0$, the MLE has at most $O(n^{-1/4})$ convergence (Chen, 1995; Heinrich and Kahn, 2018). This is closely related to the problem of testing the number of components in a finite mixture (McLachlan and Peel, 2004).

In this paper, we focus on the behavior of the MLE when Δ is *small but not zero*. This "intermediate sample size regime" is an important case in practice and is especially relevant for principal stratification models. To set the stage, Figure 1 shows the distribution of the MLE of Δ for 1000 synthetic data sets generated from Equation (1.1) for two settings. The sample sizes and mixing proportions match those in our two key principal stratification examples, JOBS II and Job Corps. The assumed difference in component means is $\Delta = 0.5$ standard deviations, which is quite large for many social science applications but smaller than in textbook examples of well-separated components. In both cases, the distribution of the simulated MLEs is markedly non-Normal. Both distributions have three notable features. First, there is a large point mass at zero. Second, a considerable portion of simulated MLEs have the opposite sign from the truth. Finally, simulated MLEs that are non-zero and have correct sign are not centered at the true value. To emphasize, these features are not due to model mis-specification: we estimate the MLE using the true model.



(a) *JOBS II:* $N = 132$, $\pi = 0.45$ (b) *JobCorps:* $N = 3,371$, $\pi = 0.06$

Figure 1. Distribution of $\widehat{\Delta}^{\text{mle}}$ for 1000 fake data sets designed to reflect the JOBS II and JobCorps studies. Data sets were generated from the two-component homoskedastic Normal mixture model in Equation (1.1) with $\Delta = 0.5$ and, respectively, (a) $N = 132$ and $\pi = 0.45$ and (b) $N = 3,371$ and $\pi = 0.06$.

1.1 Main contributions of our paper

In this paper, we give theoretical explanations for some of the practical difficulties encountered in estimation in two component finite mixture models, as shown in Figure (1), and, based on our findings, suggest guidance for practice.

We first, in Section 2, study the asymptotic properties of the MLE of the two component model (1.1) in the "intermediate sample size regime" when $\Delta \to 0$ as $n \to \infty$. This framework adequately captures weakly separated mixture components in relation to the sample size. Even for the basic model (1.1), not much seems to be known about the convergence rate of the MLE in this regime, especially when σ_0 and σ_1 are unknown. We first establish the convergence rate of the MLE, resulting in several interesting findings for the model in (1.1) when $\sigma_0 = \sigma_1$. When $\sigma \equiv \sigma_0 = \sigma_1$ is known and $\pi \neq \frac{1}{2}$, the convergence rate can and does reach $O(n^{-1/6})$ up to logarithmic factors. This is *worse* than the rate set for the degenerate

case where $\Delta = 0$, suggesting that small but non-zero separations are particularly difficult to estimate well. In such scenarios, our theoretical results explain the empirically observed difficulties in estimating Δ shown in Figure 1. For $\pi = \frac{1}{2}$, we can only estimate the difference up to a sign due to identifiability issues. In this case, the convergence rate for estimating the magnitude of the parameter is a more rapid — yet still slow — $O(n^{-1/4})$.

When σ is not known the worse-case convergence rate of the MLE remains $O(n^{-1/6})$ for the $\pi \neq \frac{1}{2}$ setting but falls to $O(n^{-1/8})$ for the $\pi = \frac{1}{2}$ setting — an order of magnitude worse than when σ was assumed known. These results are quite novel and delicate to derive, as we have to carefully account for the interaction between the location and scale parameters. Interestingly, the results together show that while the convergence is faster for the symmetric case than the asymmetric case in the known variance regime, it is slower in the unknown variance regime.

After presenting our convergence results, we turn to the practical difficulties in estimating Δ and formalize the phenomenon of the large point mass at zero shown in Figure 1. We call this phenomenon *pile up*. Specifically, we show via a mix of simulations and theoretical arguments that, in certain intermediate sample size regimes, $\widehat{\Delta}^{\text{mle}} = 0$ with very high probability even though $\Delta \neq 0$. Thus, the MLE behaves like a threshold estimator analogous to the classic Hodges estimator (see Van der Vaart, 2000). We then show that pile up occurs when the overall mixture variance is less than the within-component variance. To the best of our knowledge, we are the first to document this pile-up phenomenon in finite mixtures.

Next, we turn to using higher-order mixture moments for diagnosing pathologies with the MLE. First, we use these moments to bound the probability of pile up given either the realized data set or population parameters. We then discuss the classic problem of choosing the correct mode in a bimodal likelihood and argue that it is particularly difficult here. We show that this problem corresponds to estimating the sign of Δ (*i.e.*, the relative ordering of μ_0 and μ_1) and demonstrate how to use the third moment of the mixture distribution to

assess the probability that this occurs. We combine these results with extensive simulations to show that, across a range of reasonable settings, the sign of the MLE for Δ is no better at predicting the true sign than a coin flip.

We finally apply these mixture results to estimating principal stratification models in two randomized evaluations of job training programs, JOBS II (Vinokur et al., 1995) and JobCorps (Schochet et al., 2008). These two examples have been the focus of several prominent papers using finite mixtures for principal stratification (*e.g.*, Zhang et al., 2009; Mealli and Pacini, 2013; Frumento et al., 2012) and highlight two main use cases for this framework. For both data sets, we slightly simplify the problem to isolate the pathologies of the finite mixtures. We then assess the observed mixture distributions using the diagnostics we propose and find that pathologies are quite likely. Consequently, we do not have high confidence in the quality of the maximum likelihood estimates of $\widehat{\Delta}^{\text{mle}} = 0$ for JOBS II and an implausibly large $\widehat{\Delta}^{\text{mle}}$ for JobCorps. Our overall findings suggest that finite mixture models should be used with caution in settings such as these.

Overall, the implications for parameter estimation in finite mixtures are both novel and important. In particular, there is a longstanding consensus in finite mixture modeling that the MLE can behave poorly when components are not well separated (Redner and Walker, 1984). Indeed, several experienced researchers have told us that estimating component-specific parameters is "hopeless" in the settings we consider. While we agree with this assessment, we argue that there are no clear guidelines for researchers in practice. In particular, how do researchers know when components are separated "enough" and what happens if they are not? This is especially important because, in settings with insufficient information in the data, the MLE gives a very plausible value of zero rather than 'NA.' We believe that the framework we lay out here is an important next step towards deeper understanding of these issues.

Paper plan. Section 2 describes the asymptotic behavior of the MLE under weak separation. Section 3 explores the non-asymptotic behavior of the MLE and characterizes pile up. Section 4 uses the mixture moments for constructing diagnostics for the MLE. Section 5 gives a brief overview of the principal stratification framework and the connection to finite mixture models as well as an analysis of JOBS II. Section 6 provides additional discussion on implications for practice and possible research directions. Finally, the supplementary materials address several points that go beyond the main text, including proofs.

Notation. For any two densities p and q (with respect to Lebesgue measure μ), the variational distance between p and q is given by $V(p,q) = (1/2)\int |p-q|\, d\mu$. Additionally, the squared Hellinger distance between p and q is given by $h^2(p,q) = (1/2)\int \left(p^{1/2} - q^{1/2}\right)^2 d\mu$. Furthermore, the expression $a_n \gtrsim b_n$ is used to denote $a_n \geq C b_n$ for some C that is independent of n.

1.2 Related literature and previous work

There is a vast literature on inference in finite mixture models, dating back to the seminal work of Pearson (1894). For thorough reviews, see Everitt and Hand (1981), Redner and Walker (1984), Titterington et al. (1985), McLachlan and Peel (2004), and McLachlan et al. (2019). Frühwirth-Schnatter (2006) focuses on the Bayesian paradigm; Lindsay (1995) gives an overview of moment estimators; and Moitra (2014) discusses relevant results from machine learning. We briefly highlight several relevant aspects of this literature.

First, there has been extensive research on the asymptotic behavior of finite mixtures models. Chen (2017) gives a recent, comprehensive review. Much of this literature, however, is about the problem of testing the order of the finite mixture (see McLachlan and Peel, 2004). There are several recent papers that instead address estimation. Chen et al. (2014) focuses on estimating the mixing proportion when components are only weakly separated. Ho and

Nguyen (2016) gives results for the over-specified location-scale Gaussian mixtures. Gadat et al. (2016) study the convergence rate of $\mathbb{L}^2$-norm estimators for a few settings of two component models. Finally, Anandkumar et al. (2012); Hardt and Price (2015); Wu and Yang (2018) explore the asymptotic properties of method of moments estimators in rather general settings of Gaussian mixtures.

Second, the problem of weak separation is a special case of the *weak identification* problem especially common in econometrics. There are many examples of weak identification in other settings, including the *weak instruments problem* (Staiger and Stock, 1997) and the *moving average unit root problem*, which is the source of the term *pile up* (Shephard and Harvey, 1990; Andrews and Cheng, 2012). See also Chen et al. (2014).

Finally, although the technical discussion focuses narrowly on finite mixtures, our motivation remains the broader question of inference for causal effects within principal strata. To date, only a handful of papers have directly addressed the finite sample properties of mixtures for causal inference. Griffin et al. (2008) conduct extensive simulations and conclude that principal stratification models are generally impractical in social science settings. Mattei et al. (2013) caution that univariate mixture models often yield poor results and suggest jointly estimating effects for multiple outcomes, such as by assuming multivariate Normality. Mercatanti (2013) proposes an approach for inference with a multimodal likelihood in the principal stratification setting. Frumento et al. (2016) explore methods for quantifying uncertainty in principal stratification problems when the likelihood is non-ellipsoidal. See also Chung et al. (2004), Zhang et al. (2008), Richardson et al. (2011), and Frumento et al. (2012).

2 Asymptotic properties of the MLE: Phase transition

In this section, we study the asymptotic behavior of the MLE under two distinct but representative settings of model (1.1): first, when the variances σ_0 and σ_1 are assumed known and equal; second, when the variances σ_0 and σ_1 are unknown but assumed to be equal. Overall, we demonstrate that worst-case convergence when the components are close together is generally slow.

2.1 Known variances setting

Motivated by the illustrative simulations in Figure 1, we now explore the properties of the MLE, $\widehat{\Delta}^{\text{mle}}$, when Δ is small but non-zero. In the classical asymptotic regime, where Δ is fixed as in Equation (1.1), it is immediate that $\widehat{\Delta}^{\text{mle}}$ has a parametric rate of convergence in this simple example (see Redner and Walker, 1984; Chen, 1995). However, as shown in Figure 1a, this asymptotic regime can be a poor approximation to reality when components only have moderate separation. We therefore consider an asymptotic regime in which Δ_n shrinks as n increases. Our core finding is that, under this regime in which the two components are only slightly separated and the variance is known, the convergence rate of the MLE for the difference in means is quite poor.

Under the assumption that variances are known, we re-parametrize Equation (1.1) and assume that $Y_i, i \in \{1, \ldots, n\}$, are i.i.d. samples from the model:

$$Y_i \overset{\text{iid}}{\sim} \pi N\left(\mu - \delta_n, \sigma\right) + (1-\pi) N\left(\mu + c\delta_n, \sigma\right), \tag{2.1}$$

where $c := \frac{\pi}{1-\pi}$ and $\delta_n \in \Theta$ is a free parameter that varies with n. We assume the equal variance case of $\sigma_0 = \sigma_1 = \sigma$ for a known σ. Relative to Equation (1.1), $\mu_0 = \mu - \delta_n$, $\mu_1 = \mu + c\delta_n$, $\Delta = (1+c)\delta_n$, and μ is the overall mean, $\mathbb{E}Y_i = \mu$. For simplicity, we set

$\mu = 0$; all of the results in this section are applicable for any $\mu \in \mathbb{R}$. When $\mu = 0$ then the δ_n parameter is both the (negative) location of the first component as well as scaling of the separation of components Δ; it thus corresponds to both a location and a separation parameter. We focus on this separation parameter δ_n for ease of mathematical derivations; because Δ from Equation (1.1) is a constant re-scaling of δ, all the asymptotic results equally apply. We further assume that $\delta_n \in \Theta$ where Θ is a compact subset of $\mathbb{R}$ and $0 \in \Theta$. Finally, define $\widehat{\delta}_n^{\text{mle}}$ as the MLE for δ_n for the model in (2.1).

The following result shows the convergence rates of MLE for (2.1) where the variances are assumed to be known:

Theorem 2.1. *For the model (2.1), the following holds for any $\epsilon > 0$*

(a) (Asymmetric regime) When $\pi \in (0, 1/2)$, then

$$C_1(\epsilon)\left(\frac{1}{n}\right)^{1/6} \leq \sup_{\delta_n \in \Theta_{1,n}(\epsilon)} \mathbb{E}_{\delta_n}\left(|\widehat{\delta}_n^{mle} - \delta_n|\right) \leq C_2(\epsilon)\left(\frac{\log n}{n}\right)^{1/6},$$

where $\Theta_{1,n}(\epsilon) = \left\{\delta : |\delta| \leq n^{-1/6+\epsilon}\right\}$.

(b) (Symmetric regime) When $\pi = 1/2$, then

$$C_1(\epsilon)\left(\frac{1}{n}\right)^{1/4} \leq \sup_{\delta_n \in \Theta_{2,n}(\epsilon)} \mathbb{E}_{\delta_n}\left(\left||\widehat{\delta}_n^{mle}| - |\delta_n|\right|\right) \leq C_2(\epsilon)\left(\frac{\log n}{n}\right)^{1/4},$$

where $\Theta_{2,n}(\epsilon) = \left\{\delta : |\delta| \leq n^{-1/4+\epsilon}\right\}$.

Here, $\mathbb{E}_{\delta_n}$ denotes the expectation taken with respect to the product measure with mixture density of $Y_1, \ldots, Y_n$ under the model (2.1). Furthermore, $C_1(\epsilon)$ and $C_2(\epsilon)$ are two positive constants depending only on ϵ. Symmetry gives an analogous result for $\pi \in (1/2, 1)$.

The proof of Theorem 2.1 is provided in Appendix G.1. The variance parameter, σ is subsumed in the constants and does not impact the rates.

Prior work (Chen, 1995) has shown that when $\delta_n = 0$ the rate is of order $n^{-1/4}$ for the asymmetric case; the above therefore shows that there exists some $\delta_n \neq 0$ in a neighborhood of 0 where convergence is even worse than this degenerate case. In particular, an immediate consequence of this theorem is that, for $\pi \neq 1/2$, there exists a sequence of δ_n going to 0 at no more than a $n^{-1/6}$ rate such that the error of the MLE is also of order $n^{-1/6}$.

For the symmetric regime we are simply looking at difference in magnitude, not sign. This is because when $\pi = 1/2$ the sign of δ_n is not identifiable, and we find that

$$\sup_{\delta_n \in \Theta} \mathbb{E}_{\delta_n} |\widehat{\delta}_n^{\text{mle}} - \delta_n| \gtrsim n^{-1/r},$$

for any $r \geq 2$ and for any fixed parameter space Θ. Here, $\mathbb{E}_{\delta_n}$ denotes the expectation taken with respect to product measure with mixture density of $Y_1, \ldots, Y_n$ under the model (2.1); see the Appendix G.3 for the proof.

Connections to the Wasserstein metric. The above connects to the Wasserstein metric, which has recently been used to study parameter estimation in mixture models (Nguyen, 2013; Ho and Nguyen, 2016; Heinrich and Kahn, 2018), for additional interpretation of the results in Theorem 2.1. In particular, let $\widehat{G}_n^{\text{mle}}$ denote a probability measure (or equivalently mixing measure) with two atoms $(-\widehat{\delta}_n^{\text{mle}}, c\widehat{\delta}_n^{\text{mle}})$ whose weights are $(\pi, 1-\pi)$ and G_n a probability measure with two atoms $(-\delta_n, c\delta_n)$ whose weights are $(\pi, 1-\pi)$, then we can verify that the results of Theorem 2.1 are equivalent to

$$C_1(\epsilon) n^{-1/6} \leq \sup_{\delta_n \in \Theta_{1,n}(\epsilon)} \mathbb{E}_{\delta_n} \left(W_3(\widehat{G}_n^{\text{mle}}, G_n) \right) \asymp \sup_{\delta_n \in \Theta_{1,n}(\epsilon)} \mathbb{E}_{\delta_n} \left(|\widehat{\delta}_n^{\text{mle}} - \delta_n| \right) \leq C_2(\epsilon) \left(\frac{\log n}{n} \right)^{1/6}$$

under the asymmetric regime and

$$C_1(\epsilon) n^{-1/4} \leq \sup_{\delta_n \in \Theta_{2,n}(\epsilon)} \mathbb{E}_{\delta_n} \left(W_2(\widehat{G}_n^{\text{mle}}, G_n) \right) \asymp \sup_{\delta_n \in \Theta_{2,n}(\epsilon)} \mathbb{E}_{\delta_n} \left(\left| |\widehat{\delta}_n^{\text{mle}}| - |\delta_n| \right| \right) \leq C_2(\epsilon) \left(\frac{\log n}{n} \right)^{1/4}$$

under the symmetric regime.

2.2 Unknown equal variances setting

We now show that our previous results still generally hold when we relax the restriction that the variances are known. For the unknown equal variances setting, we assume that $Y_1, \ldots, Y_n$ are i.i.d. samples from a two component location-scale Gaussian mixture with density

$$Y_i \overset{\text{iid}}{\sim} \pi N(\mu - \delta_n, \sigma_n) + (1-\pi) N(\mu + c\delta_n, \sigma_n). \tag{2.2}$$

Here, δ_n and σ_n change with the sample size n and converge to some limit points. We assume $\sigma_n \in \Omega$, a compact subset of $\mathbb{R}_+$. We set the overall mean of $\mu = 0$ for convenience as before; δ_n is again a scaling of the gap between the two mixture means. We define $(\widehat{\delta}_n^{\text{mle}}, \widehat{\sigma}_n^{\text{mle}})$ as the MLE for the separation and scale parameters for the model in (2.2). Unlike the previous convergence results with $\widehat{\delta}_n$ in the case with known variance, the convergence rates of $\widehat{\delta}_n$ and $\widehat{\sigma}_n$ are much harder to establish due to the strong dependence between the seperation parameter δ and scale parameter σ, which is determined by the following partial differential equation (PDE):

$$\frac{\partial^2 f}{\partial \delta^2}(x, \delta, \sigma) = 2\frac{\partial f}{\partial \sigma^2}(x, \delta, \sigma), \tag{2.3}$$

for all x, δ, σ and Normal density f. This dependence leads to worse convergence rates for parameter estimation for over-fit location-scale Gaussian mixtures (Ho and Nguyen, 2016) and for hypothesis testing for the number of components of location-scale Gaussian mixtures (Chen and Chen, 2003). Under the specific setting that we consider, this dependence leads to a new characterization of the asymptotic behavior of $\widehat{\delta}_n^{\text{mle}}$, $|\widehat{\delta}_n^{\text{mle}}|$, and $\widehat{\sigma}_n^{\text{mle}}$ under the two regimes $\pi \in (0, 1/2)$ and $\pi = 1/2$. To the best of our knowledge, these have not been

previously addressed in the literature.

Theorem 2.2. *Take $\pi \in (0, 1/2]$. Under the unknown equal variances setting (2.2), the following holds*

(a) (Asymmetric regime) When $\pi \in (0, 1/2)$, then

$$C_1(\epsilon)\left(\frac{1}{n}\right)^{1/3} \leq \sup_{(\delta_n,\sigma_n)\in\mathcal{S}_{1,n}(\epsilon)} \mathbb{E}_{(\delta_n,\sigma_n)}\left(|\widehat{\delta}_n^{mle} - \delta_n|^2 + |(\widehat{\sigma}_n^{mle})^2 - \sigma_n^2|\right) \leq C_2(\epsilon)\left(\frac{\log n}{n}\right)^{1/3},$$

where $\mathcal{S}_{1,n}(\epsilon) = \left\{(\delta_n, \sigma_n) : |\delta_n|^2 + |(\sigma_n)^2 - (\overline{\sigma})^2| \leq n^{-1/3+\epsilon}\right\}$ for any $\epsilon > 0$ and some positive constant $\overline{\sigma}$.

(b) (Symmetric regime) When $\pi = 1/2$, then

$$C_1(\epsilon)\left(\frac{1}{n}\right)^{1/4} \leq \sup_{(\delta_n,\sigma_n)\in\mathcal{S}_{2,n}(\epsilon)} \mathbb{E}_{(\delta_n,\sigma_n)}\left(\left||\widehat{\delta}_n^{mle}| - |\delta_n|\right|^2 + |(\widehat{\sigma}_n^{mle})^2 - \sigma_n^2|\right) \leq C_2(\epsilon)\left(\frac{\log n}{n}\right)^{1/4},$$

where $\mathcal{S}_{2,n}(\epsilon) = \left\{(\delta_n, \sigma_n) : |\delta_n|^2 + |(\sigma_n)^2 - (\overline{\sigma})^2| \leq n^{-1/4+\epsilon}\right\}$ for any $\epsilon > 0$ and some positive constant $\overline{\sigma}$.

Here, $\mathbb{E}_{(\delta_n,\sigma_n)}$ denotes the expectation taken with respect to a product measure with a mixture density of $Y_1, \ldots, Y_n$ under the unknown equal variances setting (2.2). Furthermore, $C_1(\epsilon)$ and $C_2(\epsilon)$ are two positive constants depending only on ϵ.

The proof of Theorem 2.1 is provided in Appendix G.2.

A few comments are in order. First, under the asymmetric regime, the convergence rate of the separation parameter $\widehat{\delta}_n^{\text{mle}}$ to δ_n is of an order no more than $n^{-1/6}$ (due to the squared term within the expectation) while that of scale parameter $(\widehat{\sigma}_n^{\text{mle}})^2$ to $(\sigma_n)^2$ is no more than order $n^{-1/3}$, as long as the true parameters δ_n and σ_n belong to $S_{1,n}(\epsilon)$. The PDE of the distribution in (2.3) suggests the faster apparent convergence rate of the scale parameter relative to the separation parameter.

Second, under the symmetric regime, the worse-case convergence rate of $|\widehat{\delta}_n^{\text{mle}}|$ to $|\delta_n|$ is $n^{-1/8}$, which is slower than the worst-case rate $n^{-1/4}$ of $(\widehat{\sigma}_n^{\text{mle}})^2$ to $(\sigma_n)^2$, when the true parameters δ_n and σ_n belong to $S_{2,n}(\epsilon)$. Here, we consider the absolute value of the separation parameter for the convergence as the sign of separation parameter is not identifiable under the symmetric setting. Furthermore, in contrast to the know variance setting (2.1), the worse-case convergence rate of separation parameter under the symmetric regime is slower than that of separation parameter under the asymmetric regime. That fundamental difference can be again explained by the PDE of the location-scale Gaussian distribution.

3 Non-asymptotic properties of the MLE: Pile Up

Thus far, we have established rigorous asymptotic (minimax) behaviors of MLE under the asymmetric and symmetric cases of model (2.1) and model (2.2). The goal of this section is to shed some light on the non-asymptotic sample properties of the MLE. To facilitate the discussion, we focus solely on the known variances setting (2.1), *i.e.*, we want to analyze the non-asymptotic behavior of MLE when δ_n is near zero. We work with the likelihood function of our re-parameterized model (again, setting $\mu = 0$). This allows us to directly obtain statements regarding the points of the maximum likelihood which in turn allows for the characterization of the MLE's behavior. In particular, we first show that under our parameterization, zero (corresponding to no separation) will always be an inflection point if not a local mode. Finally, we show that, in general, the local mode is in fact the MLE when the estimated overall variance is less than σ, the assumed component variance.

3.1 Zero as a local mode of the likelihood

Given an observation $Y = y$ from the mixture model (2.1), the log-likelihood for δ_n is

$$\ell(\delta_n|Y = y) = \log\left(\pi e^{-0.5(y-\delta_n)^2} + (1-\pi)e^{-0.5(y-c\delta_n)^2}\right), \tag{3.1}$$

where we set $\sigma = 1$, though these results immediately extend to arbitrary σ. The score function is then

$$\ell'(\delta_n|Y = y) = -\frac{\pi e^{-0.5(y+\delta_n)^2}(y-\mu+\delta_n) - c(1-\pi)e^{-0.5(y-c\delta_n)^2}(y-c\delta_n)}{\pi e^{-0.5(y+\delta_n)^2} + (1-\pi)e^{-0.5(y-c\delta_n)^2}}. \tag{3.2}$$

Since $c = \frac{\pi}{1-\pi}$ with $\pi \in (0, 1/2]$, it follows from (3.2) that

$$\ell'(0|Y = y) = 0, \text{ for all } y \in \mathbb{R}. \tag{3.3}$$

Given the samples $\boldsymbol{Y}_n = (Y_1, Y_2, \ldots Y_n)$ from model (2.1), Equation (3.3) yields the following approximation of the log-likelihood given samples $\boldsymbol{Y}_n$:

$$\ell(\delta_n|\boldsymbol{Y}_n) = \ell(0|\boldsymbol{Y}_n) + \frac{1}{2}\ell''(0|\boldsymbol{Y}_n)\delta_n^2 + O(\delta_n^2). \tag{3.4}$$

In the event that $\ell''(0|\boldsymbol{Y}_n) < 0$, zero is a local mode for the log-likelihood function $\ell(\delta_n|\boldsymbol{Y}_n)$; we call this event

$$\mathcal{E} \equiv \{\ell''(0|\boldsymbol{Y}_n) < 0\}. \tag{3.5}$$

Direct calculation yields that

$$\ell''(0|\boldsymbol{Y}_n) = c\left(\sum_{i=1}^{n} Y_i^2 - n\right), \tag{3.6}$$

and thus $\ell''(0|\boldsymbol{Y}_n) < 0$ when $\sum_{i=1}^{n} Y_i^2 < n$. Equivalently, $\ell''(0|\boldsymbol{Y}_n) < 0$ when $\widehat{m}_2 < 1$,

where $\widehat{m}_2 \equiv \frac{1}{n}\sum_{i=1}^n Y_i^2$ is the observed second moment of the mixture distribution, and the assumed within-component variance is 1. We return to this connection to higher-order moments below.

3.2 Zero as the global mode of the likelihood

After establishing that zero is a local mode of the likelihood when $\ell''(0|Y_n) < 0$, an important question is whether zero is also a global mode in this case. Let $\mathcal{F} \equiv \{\widehat{\delta}_n^{\text{mle}} = 0\}$ be the event that zero is also the global mode for the likelihood function $\ell(\delta|Y)$, where $\widehat{\delta}_n^{\text{mle}}$ is the MLE under the setting of model (2.1). We refer to the event $\mathcal{F}$ as *pile up* throughout the paper. While it is clear that $\mathcal{F} \subset \mathcal{E}$, the reverse implication is not trivial. We divide our analysis into two cases: $\pi = 1/2$ and $\pi \in (0, 1/2)$. We again denote $\widehat{m}_2 := \frac{1}{n}\sum_{i=1}^n Y_i^2$.

Symmetric case. When $\pi = \frac{1}{2}$, conditioning on the event $\mathcal{E}$ (equivalently $\widehat{m}_2 < 1$), we can check that

$$\ell''(\delta|\boldsymbol{Y}_n) = \frac{4}{n}\sum_{i=1}^n \frac{Y_i^2}{\left(\exp(-\delta Y_i) + \exp(\delta Y_i)\right)^2} - 1 \leq \widehat{m}_2 - 1 < 0$$

where the inequality is due to applying Cauchy-Schwarz $\exp(-\delta Y_i) + \exp(\delta Y_i) \geq 2$ for all $i \in \{1, \ldots, n\}$. The above inequality implies that the log-likehood function $\ell(\delta|\boldsymbol{Y}_n)$ is strictly concave under the event $\mathcal{E}$. Therefore, zero is the global maximum of the log-likelihood function under the event $\mathcal{E}$. This leads to the following result regarding pile up.

Proposition 1. *Under the symmetric setting of location-scale Gaussian mixtures with known variances, $\mathcal{E} \equiv \mathcal{F}$, i.e., pile up occurs as long as 0 is a local maxima of the log-likelihood function.*

The result of Proposition 1 suggests that we can rewrite the representation of MLE under

symmetric setting with known variances as

$$\widehat{\delta}_n^{\text{mle}} = \begin{cases} 0, & \text{if } \widehat{m}_2 < 1 \\ O_p(n^{-1/4}), & \text{if } \widehat{m}_2 \geq 1 \end{cases}.$$

Thus, at least in the symmetric case, the MLE behaves like a threshold estimator analogous to the classic Hodges estimator (see Van der Vaart, 2000).

Asymmetric case. Unlike the symmetric case, we can see via simulations that there are instances for which $\mathcal{E} \neq \mathcal{F}$ in relatively small samples. Nonetheless, these counter-examples are fairly rare; for $\Delta = (1+c)\delta_n = 0.25$, $\{\mathcal{E} \cap \mathcal{F}^c\}$ occurs in fewer than 3 percent of simulation draws with sample sizes less than $N = 500$, decreasing to below 1 percent with samples sizes of $N = 1000$ or more. Extensive simulation studies seem to imply that $\mathbb{P}_n(\mathcal{F}) \nearrow \mathbb{P}_n(\mathcal{E})$.[1] We do not have a rigorous proof of this and therefore state it as a conjecture:

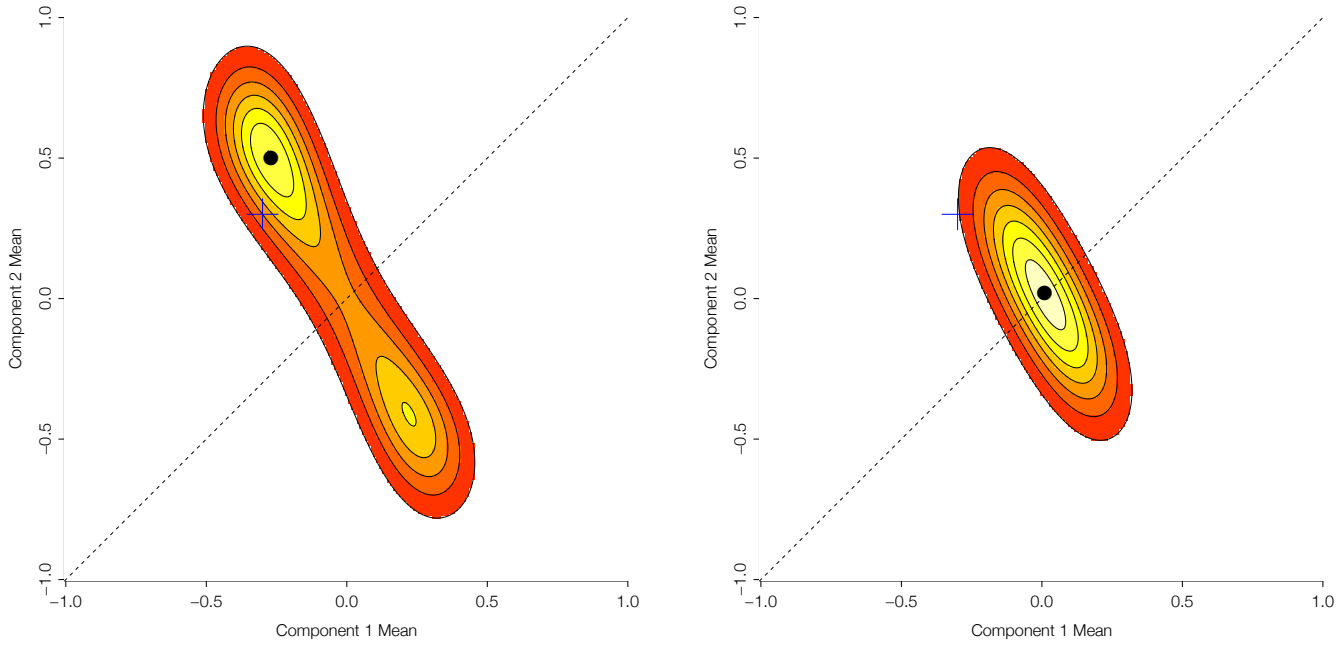
Conjecture 3.1. *Under the asymmetric setting of location-scale Gaussian mixtures with known variances, if* $\delta_n = O_p(n^{-1/6})$, *then* $\lim_{n\to\infty} \mathbb{P}_n(\mathcal{E} \cap \mathcal{F}) = 1$.

Thus Conjecture 3.1, if true, implies that, for the asymmetric setting of location-scale Gaussian mixtures with known variances, the probability that pile up occurs, i.e., $\widehat{\delta}_n = 0$, can be well approximated by the event $\{\ell''(0|\boldsymbol{Y}_n) < 0\}$. In other words, we can safely ignore the case in which zero is a local but not a global mode of the likelihood.

Figure 2 shows this pile up phenomenon in practice. Specifically, Figures 2a and 2b show the likelihood surfaces for two data sets generated via Equation (1.1), with $N = 200$, $\pi = 0.35$, and $\Delta = (1 + c)\delta = 0.6$. In Figure 2a, the likelihood is bimodal and the global mode is close to the truth, albeit more extreme.[2] In Figure 2b, the likelihood is unimodal

[1] The index n denotes the fact that the sampling distribution in (2.1) changes with n.

[2] The characterization of $\widehat{\delta}_n^{\text{mle}}$ as a Hodges-like estimator suggests that the MLE will be biased away from zero when $\widehat{\delta}_n^{\text{mle}} \neq 0$. This is closely related to the bias induced by introducing identifiability constraints,



(a) Example bimodal likelihood (b) Example unimodal likelihood

Figure 2. Two example likelihoods for component means, with data generated via Equation (1.1) with parameters $N = 200$, $\pi = 0.35$, and $\Delta = 0.6$. The '+' denotes the true component means.

and centered at zero, which is far from the truth.

4 Diagnostics for MLE pathologies

The results above suggest that the higher-order moments of the mixture distribution play an important role in the finite sample properties of the MLE. We now construct diagnostics for the MLE using these moments. First, we use these higher-order moments to construct diagnostics for pile up for the MLE, specifically the probability that pile up will occur given a set of moments, either observed moments or assumed moments. We then construct similar diagnostics for the relative order of the components, as captured by the sign of Δ. Throughout, we consider the setting with known variances, since the corresponding moment

such as $\delta > 0$ (Jasra et al., 2005; Frühwirth-Schnatter, 2006). In both cases, the MLE is the maximum of a truncated likelihood surface, truncated at the line $\delta = 0$.

equations are tractable in this case.

4.1 Probability of pile up

The probability of pile up can be characterized by using the sampling distribution of the second moment, Y^2. In particular, we can determine $\mathbb{P}\{\widehat{m}_2 < 1\}$ using the first three moments of Y^2:

$$m_2 = \mathbb{E}[Y^2] = 1 + c\delta_n^2 \tag{4.1}$$

$$v_2 = \mathbb{V}[Y^2] = 3 + 3(\pi + c^4(1-\pi))\delta_n^4 - m_2^2 \tag{4.2}$$

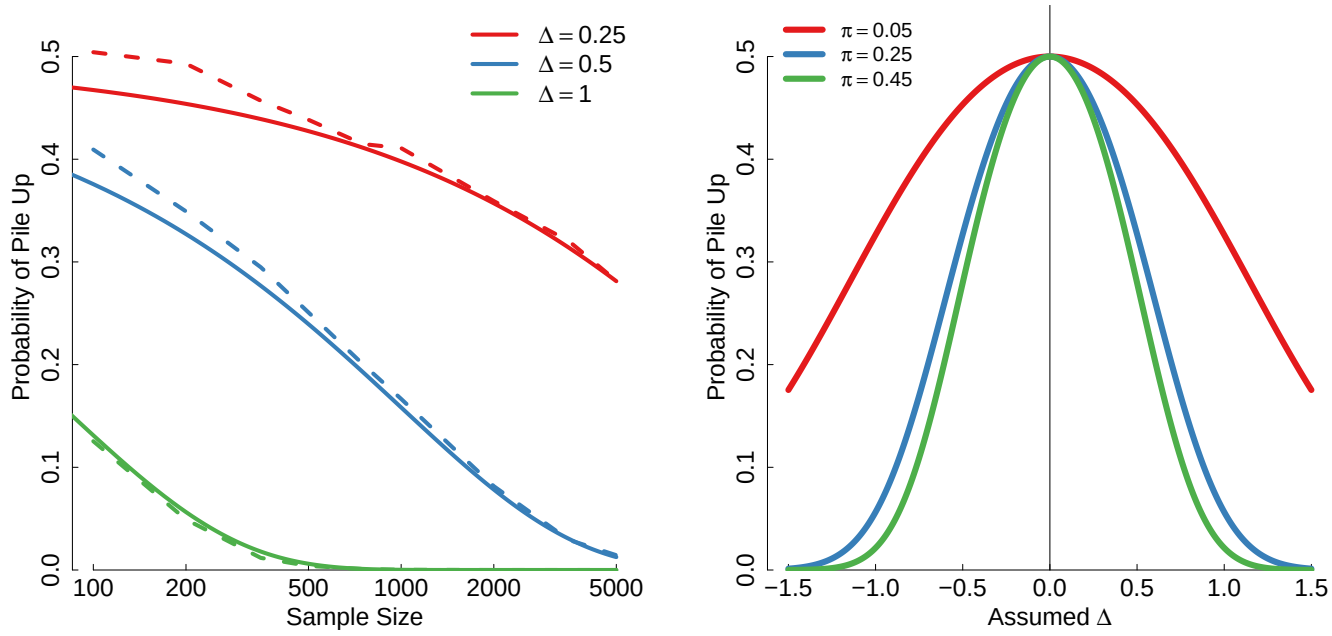
$$\Gamma_2 = \frac{1}{v_2^{3/2}}\mathbb{E}|Y^2 - m_2|^3, \tag{4.3}$$

where we can obtain Γ_2 via Monte Carlo methods. Using the Berry-Essen theorem for the convergence rates of a CLT, and assuming Conjecture 3.1, we can obtain the following bound for the probability of pile up:

$$|\mathbb{P}_n(\mathcal{E}) - \Phi(b_n)| \leq 0.7915\frac{\Gamma_2}{\sqrt{n}}. \tag{4.4}$$

As we show in simulations, $\Phi(b_n)$ appears to be an excellent approximation to the empirical pile up probability, even though the bound, which depends on the sixth mixture moment, can be wide in practice. See supplementary materials.

We can use this result for practical diagnostics, both for planning a future analysis and for assessing a particular data set. Figure 3a shows the pile up probability computed via simulation and via Equation (4.4), with $\pi = 0.325$, $\Delta = (1+c)\delta_n = 0.25$, and varying n. First, there is excellent agreement between the simulations and the Normal approximation, though $\Phi(b_n)$ slightly under-states the probabilities obtained via simulation. Second, while the probability of pile up is decreasing in both n and Δ, it is hardly a "small sample" issue.



(a) $\pi = 0.25$ (b) $N = 200$

Figure 3. Probability of pileup given sample size and separation of means. Dotted lines are simulated values across 5,000 simulations; solid lines use the Normal approximation, $\Phi(b_n)$.

For $\Delta = 0.25$, which would be quite large in many social science applications, pile up remains a meaningful possibility even with sample sizes in the thousands. For $\Delta = 1.0$, which would be an implausibly large separation in many settings, the probability of pile up is still greater than 1 in 4 for $n = 5,000$. Finally, Figure 3b shows similar results for a moderate sample size of $N = 200$ but varying mixing proportions. In this case, the probability of pile up decreases as π approaches 0.5. We believe that figures such as these are useful diagnostics before observing the mixing distribution itself.

We can also incorporate information from the observed mixture distribution. First, we can plug in the observed empirical moments, $\widehat{m}_2$ and $\widehat{v}_2$, to calculate $\widehat{b} = \frac{1-\widehat{m}_2}{\sqrt{\widehat{v}_2/n}}$ and $\Phi(\widehat{b})$. This relies on the Normal approximation for the sampling distribution as well as precisely estimating $\widehat{v}_2$, which is the fourth moment of the observed mixture distribution and might be noisy in practice. Alternatively, we could use a case-resampling bootstrap to estimate

$\mathbb{P}\{\widehat{m}_2 < 1\}$. Note that this is not the same as using the case-resampling bootstrap to estimate standard errors, which we advise against (see supplementary materials). Rather, this is analogous to the use of the bootstrap as a diagnostic tool in finite mixtures; see, for example, Grün and Leisch (2004). Finally, we note that an estimated MLE of zero still provides some information about the unknown parameter. For instance, if $\widehat{\Delta}^{\text{mle}} = (c+1)\widehat{\delta}_n = 0$, $\Delta = 0.2$ is a much more plausible value than $\Delta = 2.0$. We discuss this in the supplementary materials.

4.2 Probability of a sign error

We now turn to the sign of $\widehat{\Delta}^{\text{mle}}$ when $\pi \neq 1/2$ (the sign is not estimable when $\pi = 1/2$). Specifically, we define a *sign error* as $\text{sgn}\left(\widehat{\Delta}^{\text{mle}}\right) \neq \text{sgn}\left(\Delta\right)$. This is a well-studied issue in mixture modeling; for example, choosing the true mode in a multimodal likelihood is a classic problem (see Gan and Jiang, 1999; Biernacki, 2005). Redner and Walker (1984) give a foundational review of *asymptotic* versus *local* identifiability in mixtures. For a more recent perspective, see Kim and Lindsay (2015), who introduce the concept of *empirical* identifiability.

As with pile up, we use higher order moments for diagnosis. This is slightly more complicated than for pile up because $\text{sgn}\left(\widehat{\Delta}\right)$ is undefined when $\widehat{\Delta} = 0$. Thus, we need to consider the joint sampling distribution of both the second and third moments. In the setting with known, equal variances in Equation (2.1), we have the following moment equations:

$$m_2 = \mathbb{E}[Y^2] = 1 + \pi(1-\pi)\Delta^2$$
$$m_3 = \mathbb{E}[Y^3] = \pi(1-\pi)(1-2\pi)\Delta^3.$$

Following Tan and Chang (1972), the corresponding sample moments have the following

distribution:

$$\begin{pmatrix} \widehat{m}_2 \\ \widehat{m}_3 \end{pmatrix} \dot{\sim} \mathcal{N}\left(\begin{pmatrix} m_2 \\ m_3 \end{pmatrix}, \frac{1}{n} \begin{pmatrix} \kappa_{11}\Delta^4 + 2m_2^2 & \kappa_{12}\Delta^5 + 6m_2 m_3 \\ & \kappa_{22a}\Delta^6 + \kappa_{22b} m_2 \Delta^4 + 6m_2^3 \end{pmatrix} \right), \tag{4.5}$$

with constants $\kappa_{11} = \pi(1-\pi)(1-6\pi(1-\pi))$; $\kappa_{12} = \pi(1-\pi)(1-2\pi)(1-12\pi(1-\pi))$; $\kappa_{22a} = \pi(1-\pi)(1-30\pi(1-\pi)+120\pi^2(1-\pi)^2) + 9\pi^2(1-\pi)^2(1-2\pi)^2$; and $\kappa_{22b} = 9\pi(1-\pi)(1-6\pi(1-\pi))$. Thus, we can approximate the joint probability of pile up, sign error, or neither for a given Δ, n, and π, where we set $\Delta > 0$ for illustration:

$$\begin{aligned} &\mathbb{P}\left(\{\text{pile up; sign error; neither}\}\right) \approx \\ &\quad \mathbb{P}\left(\{\widehat{m}_2 < 1;\ \widehat{m}_2 > 1 \cap \widehat{m}_3 < 0;\ \widehat{m}_2 > 1 \cap \widehat{m}_3 > 0\}\right) \end{aligned} \tag{4.6}$$

If desired, we could apply a similar Berry-Essen bound for these probabilities, as in Equation (4.4). Instead, we simply invoke the Central Limit Theorem and use the Normal approximation in Equation (4.5).

Figure 4 shows the conditional probability of sign error given no pile up across values of N and Δ found by two methods: (1) direct simulation (simulations are restricted to draws in which $\widehat{\Delta}^{\text{mle}} \neq 0$); and (2) the tail probability of Equation (4.6) based on the Normal approximation in Equation (4.5). While the probability of a sign error decreases in both n and Δ, it remains remarkably high over plausible parameter values. Indeed, for $\Delta = 0.25$ the sign of Δ is essentially a coin flip, even with a sample size of 5,000. Importantly, conventional approaches for standard errors in the MLE (McLachlan and Peel, 2004) typically ignore this uncertainty. For additional discussion, see Kim and Lindsay (2015).

As in Section 4.1, we can assess the probability of sign error in practice. Based only on the sample size and mixing proportion, we can re-create Figure 4 across plausible parameter values. We can also plug observed values into Equation (4.5). Alternatively, we can count

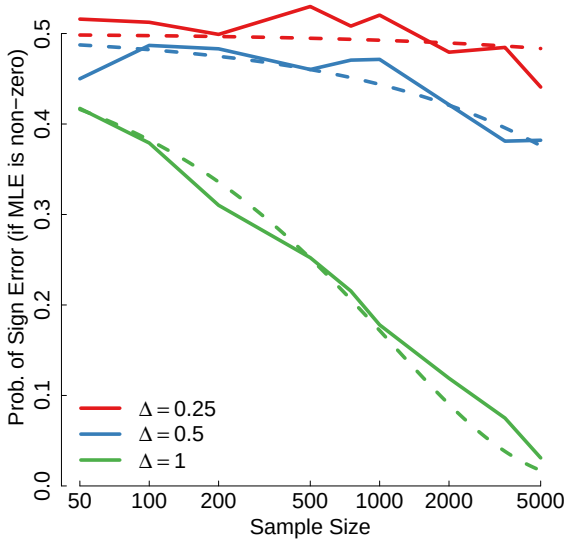


Figure 4. Probability that $\text{sgn}\left(\widehat{\Delta}\right) \neq \text{sgn}(\Delta)$ based on simulations (solid line) and the method of moments approximation in Equation (4.5) (dotted line); based on $\pi = 0.25$ and 1000 simulations at each set of parameter values

the proportion of bootstrap replicates in which the sign of the bootstrapped third moment differs from the observed sign and $\widehat{m}_2 > 1$.

5 Application to principal stratification

We now motivate the use of finite mixtures in principal stratification. For our primary running example, we re-analyze the Job Search Intervention Study (JOBS II), a randomized field experiment of a mental health and job training intervention among unemployed workers (Vinokur et al., 1995) that has been extensively studied in the causal inference literature (Jo and Stuart, 2009; Mattei et al., 2013). This is an example of one-sided noncompliance and is a simple but non-trivial example of the principal stratification setup. In the supplementary materials, we also re-analyze a randomized evaluation of JobCorps, the largest job training program in the US (Schochet et al., 2008). We briefly discuss these results at the end of this section.

5.1 Setup

We begin with the canonical example of a randomized experiment with noncompliance, such as JOBS II, and set up the problem using the potential outcomes framework (Neyman, 1923; Rubin, 1974). We observe N individuals who are randomly assigned to a treatment group, $T_i = 1$, or control group, $T_i = 0$, with observed outcome, Y. For JOBS II, the primary outcome is a measure of depression six months after randomization. As usual, we assume that randomization is valid and that the Stable Unit Treatment Value Assumption holds (SUTVA; Rubin, 1980; Imbens and Rubin, 2015). This allows us to define potential outcomes for individual i, $Y_i(0)$ and $Y_i(1)$, under control and treatment respectively, with observed outcome, $Y_i^{\text{obs}} = T_i Y_i(1) + (1 - T_i) Y_i(0)$. The fundamental problem of causal inference is that we observe only one potential outcome for each unit. Finally, we define the Intent-to-Treat (ITT) effect as the impact of randomization on the outcome, $\text{ITT} = \mathbb{E}[Y_i(1) - Y_i(0)]$. Throughout, we take expectations and probabilities to be over a hypothetical super-population.

The main complication is that only 55% of those individuals assigned to treatment actually enrolled in the program. Let D_i be an indicator for whether individual i receives the treatment, with corresponding compliance $D_i(0)$ and $D_i(1)$ for control and treatment respectively. For simplicity, we assume that only individuals assigned to treatment can receive the active intervention (*i.e.*, there is one-sided noncompliance), which is the case in the JOBS II evaluation. Formally, $D_i(0) = 0$ for all i. This gives two subgroups of interest: Never Takers, $D_i(1) = 0$, and Compliers, $D_i(1) = 1$. Following Angrist et al. (1996) and Frangakis and Rubin (2002), we refer to these subgroups interchangeably as *compliance types* or *principal strata*, $U_i \in \{\text{c}, \text{n}\}$, with "c" denoting Compliers and "n" denoting Never Takers. Table 1 shows the relationship between observed groups and principal strata.

Table 1: Summary statistics for observed groups in JOBS II

Z	D^{obs}	**Observed Mean**	**Observed SD**	**Possible Principal Strata**
1	1	-0.16	1.03	Compliers
1	0	0.05	0.96	Never Takers
0	0	0.14	0.99	Compliers and Never Takers

The two main estimands are the ITT effects for Compliers and Never Takers:

$$\text{ITT}_\text{c} = \mathbb{E}[Y_i(1) - Y_i(0) \mid U_i = \text{c}] = \mu_{\text{c}1} - \mu_{\text{c}0},$$
$$\text{ITT}_\text{n} = \mathbb{E}[Y_i(1) - Y_i(0) \mid U_i = \text{n}] = \mu_{\text{n}1} - \mu_{\text{n}0},$$

in which μ_{ut} represents the outcome mean for $U_i = u$ and $T_i = t$. We are primarily interested in ITT_c, the impact of randomization on Compliers, which measures the impact of actually enrolling in JOBS II. Since we observe stratum membership for individuals assigned to treatment, we can immediately estimate $\mu_{\text{c}1}$ and $\mu_{\text{n}1}$. Moreover, due to randomization, the observed proportion of Compliers in the treatment group is, in expectation, equal to the overall proportion of Compliers in the population, $\pi \equiv \mathbb{P}\{U_i = \text{c}\}$. Thus, we treat π as essentially known or, at least, directly estimable. The main inferential challenge is that we do not observe stratum membership in the control group. Rather we observe a mixture of Compliers and Never Takers assigned to control:

$$Y_i^{\text{obs}} \mid T_i = 0 \sim \pi f_{\text{c}0}(y_i) + (1 - \pi) f_{\text{n}0}(y_i), \tag{5.1}$$

where $f_{u0}(y)$ is the distribution of potential outcomes for individuals in stratum u assigned to control.

The standard solution for this problem is to invoke the exclusion restriction for Never Takers, which states that $\text{ITT}_\text{n} = 0$, or equivalently, $\mu_{\text{n}1} = \mu_{\text{n}0}$. Substantively, this states that the only impact of randomization on the outcome is by changing the intermediate variable, D. This is often a reasonable assumption, since actual program participation—rather than

the randomization itself—is typically the important factor in practice. With this assumption, we can then estimate ITT_c with the usual instrumental variables approach (Angrist et al., 1996). In JOBS II, however, there is a concern that randomization has a negative impact on depression levels for Never Takers (see Mattei et al., 2013). Thus, assuming that $\text{ITT}_\text{n} = 0$ could lead to biased estimates for ITT_c.

5.2 Model-based estimation

In a seminal paper, Imbens and Rubin (1997) outlined a model-based instrumental variables framework, proposing a parametric model for the outcome distribution conditional on stratum membership and treatment assignment, such as $f_{ut}(y_i) = \mathcal{N}(\mu_{ut}, \sigma^2_{ut})$. While the exclusion restriction can strengthen inference in this setting, it is not strictly necessary. Instead, identification is based entirely on standard results for mixture models.

Since Imbens and Rubin (1997), dozens of papers have used finite mixtures for estimating causal effects.[3] For one-sided noncompliance, we can write the observed data likelihood with mean-shifted standard Normal component distributions as:

$$\mathcal{L}_{\text{obs}}(\theta) = \prod_{i:\ T_i=1,\ D_i^{\text{obs}}=1} \pi\phi(y_i;\mu_{\text{c1}}) \ \times \prod_{i:\ T_i=1,\ D_i^{\text{obs}}=0} (1-\pi)\phi(y_i;\mu_{\text{n1}}) \ \times \prod_{i:\ T_i=0} \left[\pi\phi(y_i;\mu_{\text{c0}}) + (1-\pi)\phi(y_i;\mu_{\text{n0}})\right],$$

where θ represents the vector of parameters and $\phi(y_i;\mu)$ is the Normal density with mean μ and variance 1. In practice, we often relax the assumption of known, common variance. Since the observed data likelihood for individuals with $T_i = 1$ immediately factors into the likelihood for the Compliers and the likelihood for the Never Takers, we can directly

[3] Some examples of other relevant papers are Little and Yau (1998); Hirano et al. (2000); Barnard et al. (2003); Ten Have et al. (2004); Gallop et al. (2009); Zhang et al. (2009); Elliott et al. (2010); Zigler and Belin (2011); Frumento et al. (2012); Page (2012); Schochet (2013).

estimate μ_{c1} and μ_{n1}. With one-sided noncompliance, we can also directly estimate π among individuals assigned to treatment.

The challenge is therefore to estimate μ_{c0} and μ_{n0} via a two-component homoskedastic Gaussian mixture with known mixing proportion, π.[4] See Mattei et al. (2013) for further discussion of parametric mixture modeling in this setting.

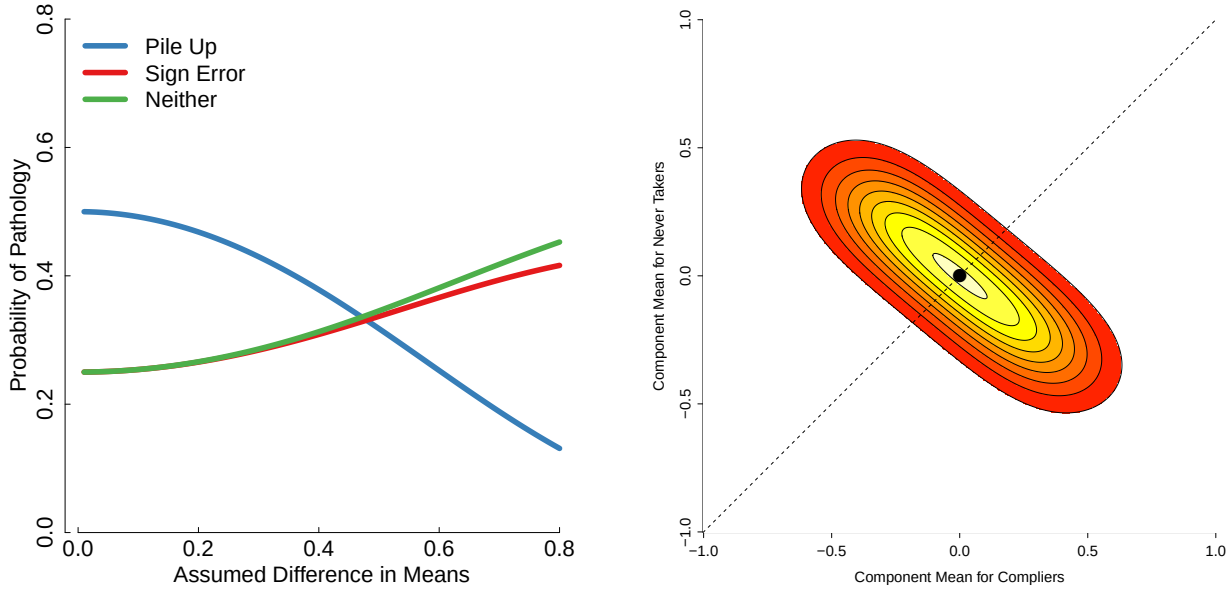
5.3 Application to JOBS II

We now turn to using the non-asymptotic results in Section 4 for estimation and diagnostics for JOBS II. We focus on a subset of $N = 410$ high risk individuals, with $N_1 = 278$ randomly assigned to treatment and $N_0 = 132$ to control. The finite mixture consists of the $N_0 = 132$ individuals assigned to control with mixing proportion $\widehat{\pi} = 0.45$.

Table 1 shows summary statistics for the three observed groups. We standardize the outcome by subtracting off the grand mean and dividing by $\widehat{\sigma}_1 = \sqrt{\pi\widehat{\sigma}^2_{n1} + (1-\pi)\widehat{\sigma}^2_{c1}}$, the estimated within-component standard deviation under treatment. Based on the group means, it is clear that workers who are observed to enroll in the program have lower depression, on average, than those who do not. Note that the point estimates for $\widehat{\sigma}_{c1}$ and $\widehat{\sigma}_{n1}$ are quite close, which is consistent with the equal variance assumption.

First, we consider the expected performance of the mixture MLE based solely on the observed sample size and mixing proportion. Figure 5a gives the probability of pile up and sign error over a range of plausible values of Δ using the Normal approximation in Equation (4.5) and the observed JOBS II values of $N = 132$ and $\widehat{\pi} = 0.45$. The pattern is striking. For values of $\Delta < 0.5$, the most likely estimate of the MLE is zero, regardless of the true value of Δ. If the MLE is non-zero, the probability of correctly estimating the sign of Δ is only slightly better than a coin flip.

[4]Note that there is a very small amount of information about π from the mixture model among those assigned to the control group. Given the other complications that arise in mixture modeling, we ignore this and regard π as if it were estimated directly from the treatment group.



(a) Prob. of MLE Pathology

(b) Observed likelihood

Figure 5. Quality of Maximum Likelihood Estimation for the finite mixture model in JOBS II, with parameters $N = 132$ and $\pi = 0.45$. Panels (a) and (b) show the probability of MLE pathology and expected bias of the MLE if non-zero; Panel (c) shows the observed likelihood for the JOBS II mixture, with a maximum at $\mu_{c0} = \mu_{n0}$.

Second, we incorporate information from the mixture distribution itself. First, the observed second and third moments are $\widehat{m}_2 = 0.96$ and $\widehat{m}_3 = 0.17$ (after centering the mixture distribution). When we plug the observed values into the Normal approximations in Equation (4.5), the probability of pile up is 0.63 and the probability of a sign error is 0.31. The corresponding probabilities based on the case-resampling bootstrap are nearly identical, 0.64 and 0.29 respectively. Thus, prior to any estimation, we believe that the probability of a pathological MLE is high.

Figure 5b shows the observed likelihood surface for Equation (1.1) fit to the JOBS II data. The likelihood is unimodal and centered at zero, which is consistent with the univariate results in Mattei et al. (2013).[5] Given the high probability of pile up *ex ante*, our analysis

[5]We can see this using the summary statistics in Mattei et al. (2013). For the univariate model without the exclusion restriction, their Table 1 gives point estimates $\widehat{\mu}_{c1} = 1.96$ and $\widehat{\mu}_{n1} = 2.08$ on the depression

suggests that we should interpret the MLE of $\widehat{\Delta}^{\text{mle}} = 0$ with caution.

5.4 Application to Job Corps

In the supplementary materials, we provide a detailed re-analysis of a randomized evaluation of JobCorps, the largest job training program in the US (Schochet et al., 2008). Following Lee (2009) and Zhang et al. (2009), we are interested in the impact of Job Corps on (log) hourly wages, which is a measure of job quality. This quantity, however, is only well defined for a certain sub-population, known as *always employed* individuals. This is a principal causal effect and is sometimes referred to as the *Survivor Average Causal Effect* (SACE). While more complicated than non-compliance in JOBS II, we can again formulate the question as estimating the component means in a Normal finite mixture model. We focus on a mixture of $N = 3,371$ individuals with $\pi = 0.06$. Thus, while the mixing proportion is relatively extreme, the sample size is considerable.

Despite the large sample size, we continue to find pathological estimates from the Normal mixture model. First, based on the diagnostics we propose above, the probability of pile up is around one-third, which is surprising given the large sample size. Rather than find that $\widehat{\Delta}^{\text{mle}} = 0$, however, we estimate an implausibly large $\widehat{\Delta}^{\text{mle}} = -4.5$ standard deviations. This estimate is well outside outside the minimax bounds, $\Delta \in [-2.4, 2.2]$, suggesting that bias might be substantial.[6] See the supplementary materials for additional analysis. In practice, the simplest explanation for these results is that the simple Normal mixture model in Equation (1.1) is a poor fit to the data. At the same time, it is difficult to imagine a different parametric mixture model that would be a better fit. This suggests that parametric

scale. The treatment effect point estimates are $\widehat{\text{ITT}}_{\text{c}} = -0.206$ and $\widehat{\text{ITT}}_{\text{n}} = -0.084$, which imply $\widehat{\mu}_{\text{c0}} = 1.96 + 0.206 = 2.166$ and $\widehat{\mu}_{\text{n0}} = 2.08 + 0.084 = 2.164$. Therefore, $\widehat{\Delta} \approx 0$. By contrast, the implied estimate for Δ from their bivariate model is $\widehat{\Delta} = 0.261$, which is roughly three-quarters of a standard deviation on the depression scale. Finally, note that the model in Mattei et al. (2013) assumes unknown, unequal variances.

[6] Following Lee (2009), we calculate minimax bounds via trimmed means of the mixture distribution. Specifically, we bound μ_{NE1} via the mean of the $\pi = 0.06$ individuals with, respectively, the lowest and highest values of hourly wages, with similar bounds for μ_{EE1}.

finite mixtures might not be an effective strategy here.

6 Discussion

We find that maximum likelihood estimates for component-specific means in finite mixtures can yield pathological results in a range of practical settings. These pathologies are particularly relevant for estimating causal effects in principal stratification models, which are often based on estimates of component means. Echoing previous work (*e.g.*, Griffin et al., 2008), we therefore caution researchers on the use and interpretation of model-based estimates of component-specific parameters, especially for causal inference.

First, we suggest that, whenever possible, researchers consider alternative approaches to inference that do not rely on model-based estimation. In the context of principal stratification, these alternatives often rely on constant treatment effect assumptions or on conditional independence across multiple outcomes (e.g., Jo, 2002; Jo and Stuart, 2009; Ding et al., 2011). When such restrictions are not possible, we recommend that researchers first compute nonparametric bounds (see Zhang and Rubin, 2003; Grilli and Mealli, 2008; Lee, 2009; Miratrix et al., 2018).

Second, researchers might nonetheless be interested in leveraging parametric assumptions for estimation. In this case, we suggest that researchers use our results to assess the probability of pathological results for different parameter values. Similar to design analysis, these calculations can provide practical guidance on whether mixture modeling will yield useful inference. One possibility is to incorporate multiple outcomes, such as in Mattei et al. (2013). This can greatly improve inference; intuitively, the distance between components will be greater in multivariate space, in effect, giving larger Δ and easier separation (see also Mercatanti et al., 2015).

Third, we have focused on maximum likelihood rather than Bayesian methods (Frühwirth-

Schnatter, 2006). The Bayesian approach offers some distinct advantages over likelihood-based inference.[7] For example, the Bayesian can incorporate informative prior information, which can be especially important in finite mixture modeling; see, for example, Aitkin and Rubin (1985); Hirano et al. (2000); Chung et al. (2004); Lee et al. (2009); Gelman (2010). Moreover, our concern about sign error is trivial in the Bayesian setting: the global mode is simply a poor summary of a multi-modal posterior. More broadly, the weak identification issues we highlight in this paper are not necessarily relevant to a strict Bayesian. Imbens and Rubin (1997) and Mattei et al. (2013), for example, characterize weak identification as substantial regions of flatness in the posterior, which increases uncertainty but does not lead to any fundamental challenges.[8] Nonetheless, we argue that our results are highly relevant for Bayesians who are also interested in good frequency properties (Rubin, 1984). In the supplementary materials, we offer evidence that the pathological behaviors we document for the MLE also hold for the posterior mean and median with some "default" prior values. In this sense, we conduct a Frequentist evaluation of a Bayesian procedure (*e.g.*, Rubin, 2004) and find poor frequency properties overall. More generally, we agree that informative prior information can be a powerful tool for improving inference in this setting. Finding suitable priors for finite mixture models is a topic for future research.

Going forward, we hope that the approach outlined here can serve as a useful template for studying the behavior of mixture model estimates in finite samples. Moreover, we considered only a very simple case in this paper; in the future, we plan to assess inference for much richer models, especially those common in principal stratification. Finally, we are actively exploring alternative estimation strategies, particularly those that more directly leverage

[7] The Bayesian approach also introduces some unique challenges that we do not address here, namely the label-switching problem (Celeux et al., 2000; Jasra et al., 2005) and the difficulty of specifying vague prior distributions for finite mixtures (Grazian and Robert, 2015).

[8] Imbens and Rubin (1997) note that "issues of identification [in the Bayesian perspective] are quite different from those in the frequentist perspective because with proper prior distributions, posterior distributions are always proper. The effect of adding or dropping assumptions is directly addressed in the phenomenological Bayesian approach by examining how the posterior predictive distributions for causal estimands change."

Bayesian methods and that can give sensible point estimates. In the end, inference in the Twilight Zone is possible. But we must proceed with caution.

References

Aitkin, M. and Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B*, 47(1):67–75.

Anandkumar, A., Hsu, D., , and Kakade, S. M. (2012). A method of moments for mixture models and hidden markov models. In *COLT*.

Andrews, D. W. (1993). Exactly median-unbiased estimation of first order autoregressive/unit root models. *Econometrica*, 61(1):139–165.

Andrews, D. W. K. (2000). Inconsistency of the Bootstrap when a Parameter is on the Boundary of the Parameter Space. *Econometrica*, 68(2):399–405.

Andrews, D. W. K. and Cheng, X. (2012). Estimation and Inference With Weak, Semi-Strong, and Strong Identification. *Econometrica*, 80(5):2153–2211.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Barnard, J., Frangakis, C. E., Hill, J. L., and Rubin, D. B. (2003). Principal stratification approach to broken randomized experiments. *Journal of the American Statistical Association*, 98(462):299–323.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89:1012–1016.

Bickel, P. J. and Freedman, D. A. (1981). Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics*, 9(6):1196–1217.

Biernacki, C. (2005). Testing for a global maximum of the likelihood. *Journal of Computational and Graphical Statistics*, 14(3):657–674.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95:957–970.

Chen, H. and Chen, J. (2003). Tests for homogeneity in normal mixtures in the presence of a structural parameter. *Statistica Sinica*, 13:351–365.

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

Chen, J. (2017). Consistency of the MLE under mixture models. *Statistical Science*, 32(1):47–63.

Chen, X., Ponomareva, M., and Tamer, E. (2014). Likelihood inference in some finite mixture models. *Journal of Econometrics*, 182(1):87–99.

Chung, H., Loken, E., and Schafer, J. L. (2004). Difficulties in drawing inferences with finite-mixture models. *The American Statistician*, 58(2):152–158.

Ding, P., Feller, A., and Miratrix, L. (2016). Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society, Series B*.

Ding, P., Geng, Z., Yan, W., and Zhou, X.-H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *Journal of the American Statistical Association*, 106:1578–1591.

Elliott, M. R., Raghunathan, T. E., and Li, Y. (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics*, 11(2):353–372.

Everitt, B. S. and Hand, D. J. (1981). *Finite mixture distributions*. Chapman and Hall, London, New York.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58(1):21–29.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models: Modeling and applications to random processes*. Springer Science & Business Media.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *Journal of the American Statistical Association*, 107(498):450–466.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2016). The fragility of standard inferential approaches in principal stratification models relative to direct likelihood approaches. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(1):58–70.

Gadat, S., Marteau, C., and Maugis-Rabusseau, C. (2016). Parameter recovery in two-component contamination mixtures: the l2 strategy. *arXiv preprint arXiv:1604.00306*.

Gallop, R., Small, D. S., Lin, J. Y., Elliott, M. R., Joffe, M., and Ten Have, T. R. (2009). Mediation analysis with principal stratification. *Statistics in Medicine*, 28(7):1108–1130.

Gan, L. and Jiang, J. (1999). A test for global maximum. *Journal of the American Statistical Association*, 94(447):847–854.

Gelman, A. (2010). Bayesian inference in political science, finance, and marketing research. *Frontiers of Statistical Decision Making and Bayesian Analysis: In Honor of James O. Berger*, pages 377–417.

Gelman, A. (2011). Why it doesn't make sense in general to form confidence intervals by inverting hypothesis tests. http://andrewgelman.com/2011/08/25/why_it_doesnt_m/.

Ghosal, S. and van der Vaart, A. (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Annals of Statistics*, 29:1233–1263.

Grazian, C. and Robert, C. P. (2015). Jeffreys priors for mixture estimation. pages 37–48.

Griffin, B. A., McCaffrey, D. F., and Morral, A. R. (2008). An application of principal stratification to control for institutionalization at follow-up in studies of substance abuse treatment programs. *The Annals of Applied Statistics*, 2:1034–1055.

Grilli, L. and Mealli, F. (2008). Nonparametric bounds on the causal effect of university studies on job opportunities using principal stratification. *Journal of Educational and Behavioral Statistics*, 33(1):111–130.

Grün, B. and Leisch, F. (2004). *Bootstrapping Finite Mixture Models*. 2004 Proceedings in Computational Statistics.

Hansen, B. E. (1999). The grid bootstrap and the autoregressive model. *Review of Economics and Statistics*, 81(4):594–607.

Hardt, M. and Price, E. (2015). Tight bounds for learning a mixture of two gaussians. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 753–760. ACM.

Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *Annals of Statistics*, 46.

Hirano, K., Imbens, G. W., Rubin, D. B., and Zhou, X. H. (2000). Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88.

Ho, N. and Nguyen, X. (2016). Convergence rates of parameter estimation for some weakly identifiable finite mixtures. *Annals of Statistics*, 44.

Hodges, J. L. and Lehmann, E. L. (1963). Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34:598–611.

Imbens, G. and Rubin, D. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, 25(1):305–327.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1):50–67.

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics*, 27:385–409.

Jo, B. and Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine*, 28(23):2857–2875.

Kang, H., Cai, T. T., and Small, D. S. (2015). Robust confidence intervals for causal effects with possibly invalid instruments. *arXiv*, page 1504.03718.

Kim, D. and Lindsay, B. G. (2015). Empirical identifiability in finite mixture models. *Annals of the Institute of Statistical Mathematics*, 67(4):745–772.

Laber, E. B. and Murphy, S. A. (2011). Adaptive Confidence Intervals for the Test Error in Classification. *Journal of the American Statistical Association*, 106(495):904–913.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Lee, K., Mengersen, K., Marin, J.-M., and Robert, C. P. (2009). Bayesian Inference on Mixtures of Distributions. *Perspectives in Mathematical Sciences. Stat. Sci. Interdiscip. Res.*, 7:165–202.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 5.

Little, R. J. and Yau, L. H. Y. (1998). Statistical techniques for analyzing data from prevention trials: Treatment of no-shows using Rubin's causal model. *Psychological Methods*, 3(2):147–159.

Mattei, A., Li, F., and Mealli, F. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *The Annals of Applied Statistics*, 7(4):2336–2360.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6:355–378.

Mealli, F. and Pacini, B. (2013). Using Secondary Outcomes to Sharpen Inference in Randomized Experiments With Noncompliance. *Journal of the American Statistical Association*, 108:1120–1131.

Mercatanti, A. (2013). A Likelihood-based analysis for relaxing the exclusion restriction in randomized experiments with noncompliance. *Australian & New Zealand Journal of Statistics*, 55(2):129–153.

Mercatanti, A., Li, F., and Mealli, F. (2015). Improving inference of gaussian mixtures using auxiliary variables. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 8(1):34–48.

Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica*, 75(5):1411–1452.

Miratrix, L., Furey, J., Feller, A., Grindal, T., and Page, L. C. (2018). Bounding, an accessible method for estimating principal causal effects, examined and explained. *Journal of Research on Educational Effectiveness*, 11(1):133–162.

Moitra, A. (2014). Algorithmic aspects of machine learning. http://people.csail.mit.edu/moitra/docs/bookex.pdf.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Annals of Statistics*, 4(1):370–400.

Nolen, T. L. and Hudgens, M. G. (2011). Randomization-Based Inference Within Principal Strata. *Journal of the American Statistical Association*, 106(494):581–593.

Page, L. C. (2012). Principal stratification as a framework for investigating mediational processes in experimental settings. *Journal of Research on Educational Effectiveness*, 5(3):215–244.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM review*.

Richardson, T. S., Evans, R. J., and Robins, J. M. (2011). Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610.

Rosenbaum, P. R. (2002). *Observational studies*. Springer.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:668–701.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher Randomization Test. *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12(4):1151–1172.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Schochet, P. Z. (2013). Student mobility, dosage, and principal stratification in school-based RCTs. *Journal of Educational and Behavioral Statistics*, 38(4):323–354.

Schochet, P. Z., Burghardt, J., and McConnell, S. (2008). Does Job Corps Work? Impact Findings from the National Job Corps Study. *The American Economic Review*, 98(5):1864–1886.

Shephard, N. G. and Harvey, A. C. (1990). On the probability of estimating a deterministic component in the local level model. *Journal of Time Series Analysis*, 11(4):339–347.

Staiger, D. and Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, 65(3):557–586.

Tan, W. Y. and Chang, W. (1972). Some comparisons of the method of moments and the method of maximum likelihood in estimating parameters of a mixture of two normal densities. *Journal of the American Statistical Association*, 67(339):702–708.

Ten Have, T. R., Elliott, M. R., Joffe, M., Zanutto, E., and Datto, C. (2004). Causal models for randomized physician encouragement trials in treating primary care depression. *Journal of the American Statistical Association*, 99(465):16–25.

Titterington, D., Smith, A., and Makov, U. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons.

van de Geer, S. (2000). *Empirical Processes in M-estimation*. Cambridge University Press.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vinokur, A. D., Price, R. H., and Schul, Y. (1995). Impact of the jobs intervention on unemployed workers varying in risk for depression. *American Journal of Community Psychology*, 23(1):39–74.

Wasserman, L. (2012). Mixture Models: The Twilight Zone of Statistics. http://normaldeviate.wordpress.com/2012/08/04/mixture-models-the-twilight-zone-of-statistics/.

Wu, Y. and Yang, P. (2018). Optimal estimation of Gaussian mixtures via denoised method of moments. *arXiv preprint arXiv:1807.07237*.

Yu, B. (1997). Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam*, pages 423–435.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by "death". *Journal of Educational and Behavioral Statistics*, 28(4):353–368.

Zhang, J. L., Rubin, D. B., and Mealli, F. (2008). Evaluating the effects of job training programs on wages through principal stratification. *Advances in Econometrics*, 21:117–145.

Zhang, J. L., Rubin, D. B., and Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. *Journal of the American Statistical Association*, 104(485):166–176.

Zigler, C. M. and Belin, T. R. (2011). The potential for bias in principal causal effect estimation when treatment received depends on a key covariate. *The Annals of Applied Statistics*, 5(3):1876–1892.

Supplementary Materials for "Weak separation in mixture models and implications for principal stratification"

A Robust estimation via method of moments

Rather than use higher order moments as diagnostics, we can instead use the method of moments directly for estimation. Several recent papers have highlighted the attractive properties of method of moment estimators for general mixture models (Anandkumar et al., 2012; Wu and Yang, 2018). Applying these results, we show that the method of moments approach has similar asymptotic properties to the MLE but better finite sample properties; in particular, the method of moments is not susceptible to pile up.

First, in the setting with known, equal variances in Equation (2.1), we have the following moment equations:

$$\begin{aligned} m_1 = \mathbb{E}[Y] &= \mu \\ m_2 = \mathbb{E}[Y^2] &= 1 + c\delta^2 \\ m_3 = \mathbb{E}[Y^3] &= \frac{1-2\pi}{1-\pi} c\delta^3, \end{aligned} \tag{A.1}$$

where $\Delta = (1+c)\delta$. Since there is no information in the first moment about δ, we consider two estimators based on the second and third moments:[9]

$$|\widehat{\delta}_{m_2}| := \left| \frac{\widehat{m}_2 - 1}{c} \right|^{1/2} \qquad \widehat{\delta}_{m_3} := \left[\frac{(1-\pi)}{c\,(1-2\pi)}\, \widehat{m}_3 \right]^{1/3},$$

where $\hat{m}_2$ and $\hat{m}_3$ are the sample second and third (non-central) moments, respectively. First, the absolute value for $|\widehat{\delta}_{m_2}|$ is necessary because there is no information about sign of δ in the second moment. Thus, $\widehat{\delta}_{m_2}$ is a natural estimator when $\pi = 1/2$. By contrast, when $\pi \in (0, 1/2)$, $\hat{\delta}_{m_3}$ will estimate both the magnitude and sign of δ.

The following result establishes that these estimators have asymptotic behavior similar to the MLE, as described in Theorem 2.1.

Proposition 2. *Given the formulations of estimators $\widehat{\delta}_{m_2}$ and $\widehat{\delta}_{m_3}$, for the setting of known equal variances (2.1), the following holds*

[9] In principle, we could also consider a generalized method of moments estimator based on both the second and third moments, though this is less transparent than the estimators we discuss below. See Anandkumar et al. (2012); Hardt and Price (2015); Wu and Yang (2018).

(a) (Asymmetric regime) When $\pi \in (0, 1/2)$, then

$$\sup_{\delta_n \in \Theta} \left| \left| \widehat{\delta}_{m_2} \right| - |\delta_n| \right| = O_p\left(n^{-1/4}\right), \tag{A.2}$$

$$\sup_{\delta_n \in \Theta} \left| \widehat{\delta}_{m_3} - \delta_n \right| = O_p\left(n^{-1/6}\right). \tag{A.3}$$

(b) (Symmetric regime) When $\pi = 1/2$, then

$$\sup_{\delta_n \in \Theta} \left| \left| \widehat{\delta}_{m_2} \right| - |\delta_n| \right| = O_p\left(n^{-1/4}\right), \tag{A.4}$$

where $\widehat{\delta}_{m_3}$ is undefined when $\pi = 1/2$.

While these simple estimators have the same asymptotic behavior as the MLE, neither $\widehat{\delta}_{m_2}$ nor $\widehat{\delta}_{m_3}$ are susceptible to pile up. It suggests that the moment estimators under the simple setting of known equal variances are more robust than the MLE.

B Analysis of Job Corps

B.1 Setup.

Following (Zhang et al., 2009), we use the principal stratification framework to define the impact of Job Corps on hourly wages. Let S be an indicator for employment, with corresponding potential outcomes $S_i(0)$ and $S_i(1)$ and observed employment status S_i^{obs} for individual i. We then define principal strata, U, based on the joint distribution, $\{S_i(0), S_i(1)\}$:

$$U_i = \begin{cases} EE & \text{if } S_i(1) = 1, S_i(0) = 1 \\ EN & \text{if } S_i(1) = 1, S_i(0) = 0 \\ NE & \text{if } S_i(1) = 0, S_i(0) = 1 \\ NN & \text{if } S_i(1) = 0, S_i(0) = 0 \end{cases}.$$

We are interested in the impact of randomization on the *always employed* strata, EE. This is sometimes known as a *Survival Average Causal Effect* and is closely related to the idea of "truncation due to death" (see Zhang et al., 2009). Finally, following (Lee, 2009), we invoke the *monotonicity* assumption, which states that random encouragement to enroll in a job training program can only increase employment, $S_i(1) \geq S_i(0)$; thus the NE group does not exist.[10]

[10] While this simplifies the analysis and allows us to highlight the role of finite mixture modeling, Zhang et al. (2009) argue against this assumption. In particular, they argue that enrolling in a job training program might raise an individual's *reservation wage* and, as a result, make that individual less likely to accept a lower paying job. We merely note that relaxing this assumption further complicates the analysis, since the mixing proportions are no longer identified non-parametrically.

Table 2: Summary statistics for observed groups in Job Corps

Z	S^{obs}	**Observed Mean**	**Observed SD**	**Possible Principal Strata**
1	1	0.03	1.013	EE and EN
1	0	—	—	NN
0	1	-0.05	1	EE
0	0	—	—	NN and EN

Table 2 shows the relationship between principal strata and the observed groups, based on Z and S^{obs}. Under monotonicity, we directly observe *always employed* individuals (EE) assigned to the control group. We can therefore directly estimate the average outcome for this group, μ_{EE0}. We can also directly estimate the proportion of EE individuals via $\widehat{\pi}_{EE} = \mathbb{P}[S \mid Z_i = 0]$, the proportion of *never employed* individuals (NN) via $\widehat{\pi}_{\text{NN}} = 1 - \mathbb{P}[S \mid Z_i = 1]$, and the proportion of the induced to employment individuals (EN) via $\widehat{\pi}_{\text{EN}} = 1 - \widehat{\pi}_{\text{NN}} - \widehat{\pi}_{\text{EE}}$. Without additional assumptions, however, we cannot estimate μ_{EE1}, instead observing a mixture of EE and EN individuals. Consistent with (Zhang et al., 2009) and (Frumento et al., 2012), we therefore assume that log-hourly wages follow a mixture of Gaussians with known mixing proportion, as in Equation (1.1) in the main text. Note that this mixture is much simpler than the full model considered in (Zhang et al., 2009), which accounts for some important additional complications.

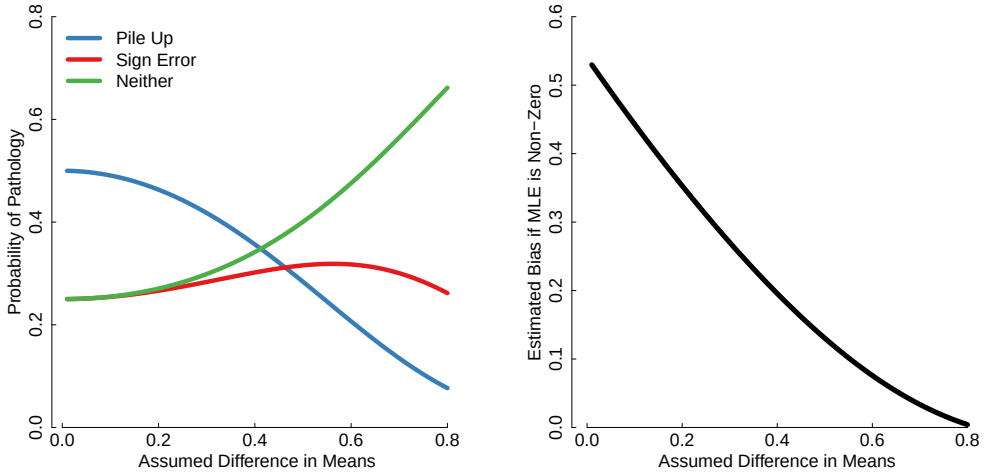
B.2 Diagnostics.

We focus on a complete case subset used by (Lee, 2009) of $N = 9,145$ individuals, with $N_1 = 5,546$ randomly assigned to treatment and $N_0 = 3,599$ to control. The mixture model consists of the $N_{11} = 3,371$ individuals assigned to treatment who are employed, with mixing proportion $\widehat{\pi} = 0.06$.

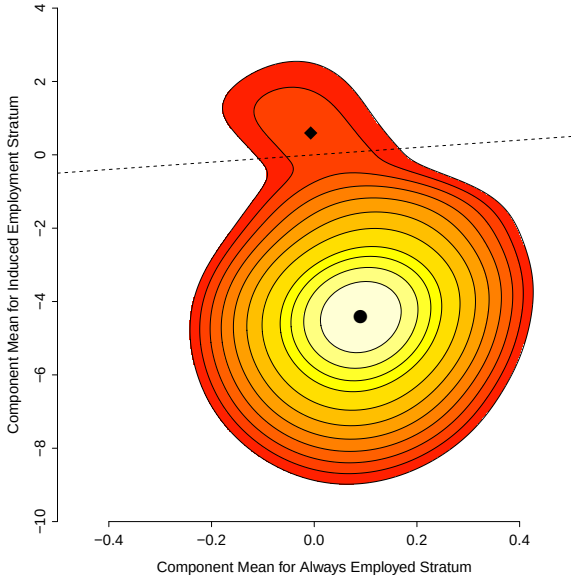
Table 2 shows summary statistics for observable groups. We standardize the outcome by subtracting off the grand mean and dividing by $\widehat{\sigma}_0$, the estimated standard deviation for individuals assigned to control who are employed. This is also the standard deviation for EE individuals assigned to control. Since hourly wage is only defined for employed workers, the rows with $S^{\text{obs}} = 0$ have undefined outcomes.

Figure B.6a gives the probability of pile up and sign error over a range of plausible values of Δ using the Normal approximation in Equation (4.5) and the observed Job Corps mixtures parameters of $N = 3,371$ and $\widehat{\pi} = 0.06$. As in Figure 5a, pile up is a major concern, though the probability of a sign error is somewhat less *ex ante*, in part because the mixing proportion is much closer to 0. Figure B.6b shows the bias of the MLE if the MLE is non-zero and the sign is correct. As with JOBS II, the bias can be severe.

We can also incorporate the higher order moments of the mixture distribution. In this case, the observed second and third moments are $\widehat{m}_2 = 1.03$ and $\widehat{m}_3 = -0.87$, respectively (after centering the mixture distribution). Plugging the observed values into the Normal approximations in Equation (4.5), the pile up probability of 0.34 and the sign error probability is 0.03. The corresponding probabilities based on the case-resampling bootstrap are nearly



(a) Prob. of MLE Pathology (b) Expected Bias if MLE is non-zero



(c) Observed likelihood

Figure B.6. Quality of Maximum Likelihood Estimation for the finite mixture model in Job Corps, with parameters $N = 3,371$ and $\pi = 0.06$. Panels (a) and (b) show the probability of MLE pathology and expected bias of the MLE if non-zero; Panel (c) shows the observed likelihood for the Job Corps mixture, with a global mode and a local mode. The dotted line denotes equal component means.

identical, 0.34 and 0.04 respectively.

Figure B.6c shows the observed likelihood for the mixture model. The MLE is at $\widehat{\mu}_{\text{EE1}}^{\text{mle}} = 0.09$ and $\widehat{\mu}_{\text{NE1}}^{\text{mle}} = -4.40$, which implies $\widehat{\Delta}^{\text{mle}} = -4.49$ standard deviations. This is clearly an extreme estimate. Transforming these estimates to \$ per hour shows that $\widehat{\mu}_{\text{EE1}}^{\text{mle}} = \$8.24$ per hour and $\widehat{\mu}_{\text{NE1}}^{\text{mle}} = \$0.09$ per hour, which is far below feasible hourly wages in this sample. This estimate is also outside the minimax bounds, $\Delta \in [-2.4, 2.2]$.[11] There is also a local mode centered at $\widehat{\mu}_{\text{EE1}}^{\text{mle}} = -0.01$ and $\widehat{\mu}_{\text{NE1}}^{\text{mle}} = 0.59$, which implies $\widehat{\Delta}^{\text{mle}} = 0.60$ standard deviations. In units of \$ per hour, this is $\widehat{\mu}_{\text{EE1}}^{\text{mle}} = \$7.47$ per hour and $\widehat{\mu}_{\text{NE1}}^{\text{mle}} = \$13.64$ per hour. While far more feasible than the global mode, these estimates are still worrisome, since it is unlikely that the group induced to employment by Job Corps would have hourly wages nearly twice those of the always employed group; see Figure B.6b. Regardless, the likelihood at the MLE is considerably higher than at the local mode, with $-2 * (\ell(+0.60|Y) - \ell(-4.49|Y)) = 296$. Taken together, these results suggest that maximum likelihood does not give practically useful results in this example.

In practice, the simplest explanation for these results is that the simple Normal mixture model in Equation (1.1) in the main text is a poor fit to the data. At the same time, however, it is difficult to imagine a more plausible parametric mixture model in this setting. Thus parametric finite mixtures might not be an effective strategy in this example.

C Validating the Normal approximations

We present figures testing the agreement of the moment-based Normal approximations with their corresponding pathologies assessed via simulation. Figure C.7 compares the incidence of pile up and $\widehat{m}_2 < 1$ for a range of values of π, Δ, and N. The blue line indicates the probability the method of moments estimator indicator of pile up ($1\{\widehat{m}_2 < 1\}$) agrees with whether or not pile up was observed in simulation. The results are averaged over 1000 simulated data sets. Unsurprisingly, the correspondence improves as N increases and is worst when $\pi = 0.1$, the case in which the mixture is its most asymmetric. Overall, however, the Normal approximation provides an excellent estimator for whether pile up has occurred in the sample.

Figure C.8 shows the corresponding plots for assessing the sign of Δ. Here, due to the extra noise in m_3, the correspondence is much less sharp. The discrepancies are most noticeable when π is close to 0 and Δ is small.

D Confidence sets via inverting tests

Given the poor performance of the MLE, we are interested in methods that perform well even when Δ is small. Based on the large literature on weak identification in other settings, we

[11] Following (Lee, 2009), we calculate minimax bounds via trimmed means of the mixture distribution. Specifically, we bound μ_{NE1} via the mean of the $\pi = 0.06$ individuals with, respectively, the lowest and highest values of hourly wages, with similar bounds for μ_{EE1}.

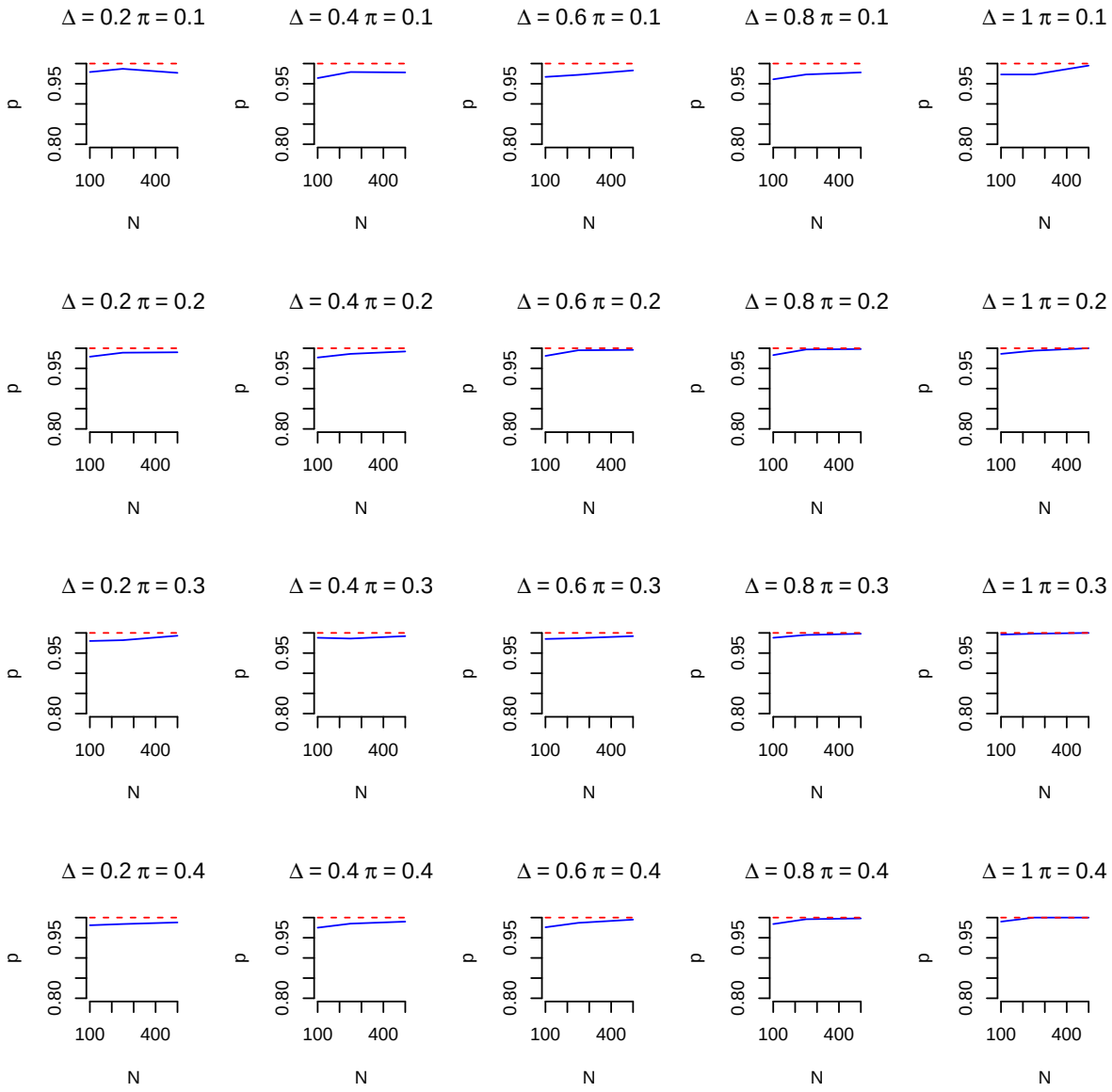


Figure C.7. Probability that the diagnostic based on the second moment ($1\{\widehat{m}_2 < 1\}$) agrees with whether or not pile up was observed in simulation. The dotted red line perfect correspondence at each tested N. The blue line is the average agreement probability over 1000 simulated data sets.

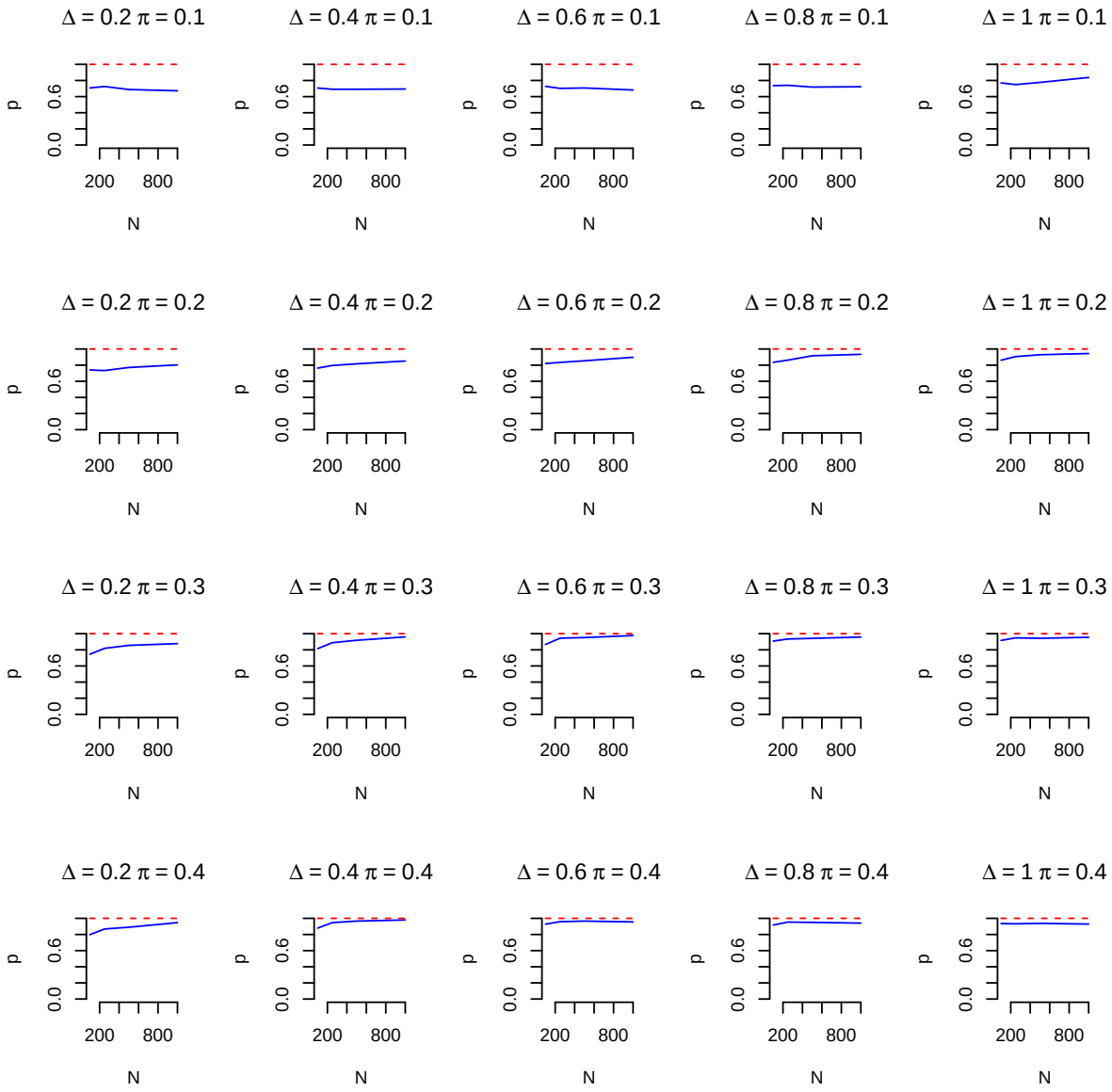


Figure C.8. Probability that the diagnostic based on the third moment agrees with whether or not the wrong sign pathology was observed in simulation. The dotted red line perfect correspondence at each tested N. The blue line is the average agreement probability over 1000 simulated data sets.

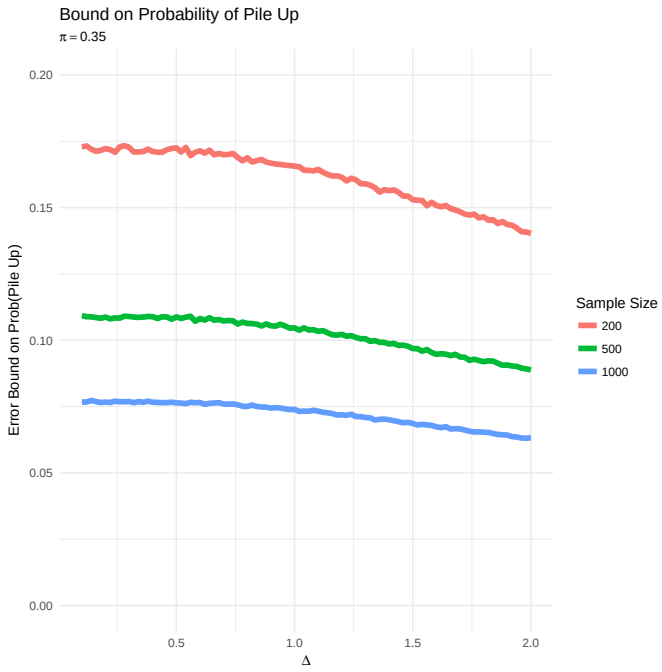


Figure C.9. Berry-Essen bound for probability of pile up for $\pi = 0.35$ and a range of values of N and Δ.

presume that many such methods are possible. As a starting point, we suggest an approach to construct confidence intervals based on inverting a sequence of tests. This approach is widely used in other weak identification settings, namely weak instruments (e.g., Staiger and Stock, 1997; Kang et al., 2015) and the unit root moving average problem (Mikusheva, 2007). It is also closely related to the method of constructing confidence intervals for causal effects by inverting a sequence of Fisher Randomization Tests (Rosenbaum, 2002).

At the same time, this approach has its drawbacks. First, while test inversion yields confidence sets with good coverage properties, it does not necessarily yield good point estimates. In particular, it is possible to construct a Hodges-Lehmann-style estimator via the point on the grid with the highest p-value (Hodges and Lehmann, 1963). But since pile up and sign error remain issues, any point estimator in this case should be interpreted with caution. Second, the coverage guarantees hold only when the model is correctly specified; under even moderate mis-specification, the resulting estimator can cease to exist (Gelman, 2011). Importantly, the MLE performs poorly *even when the model is correctly specified*. Alternatively, researchers uninterested in test inversion for confidence intervals might nonetheless be interested in using this approach to assess model fit. If the proposed procedure rejects everywhere, this is evidence that the Normal mixture model is a poor fit.

We discuss two basic approaches here. Our first approach is a version of the grid bootstrap of (Andrews, 1993) and (Hansen, 1999), which generates Monte Carlo p-values by simulating fake data sets from the null hypothesis. While the grid bootstrap is conceptually straightforward and enjoys theoretical guarantees (Mikusheva, 2007), it is also computationally intensive. Our second approach is therefore a fast approximation that directly uses the

Normal sampling distribution in Equation (4.5) of the main text to derive a χ^2 test at each grid point. To demonstrate these methods, we first outline inference for Δ alone and then extend this to inference for the component-specific means, μ_0 and μ_1.

D.1 Overview of grid bootstrap

To conduct a grid bootstrap, we first need a grid. Define $\Delta = \{\Delta_0, \Delta_1, \ldots, \Delta_n\}$ with $\Delta_i > \Delta_j$ for $i > j$. The immediate goal is then to obtain a p-value for the following null hypotheses for each value $\Delta_j \in \Delta$:

$$H_0 : \Delta = \Delta_j \text{ vs. } H_1 : \Delta \neq \Delta_j. \tag{D.1}$$

For convenience we first center the data (i.e., we set $\mu = 0$ as in the main text). Next, we need a test statistic, $t(\mathbf{y}, \Delta_j)$, that is a function of the observed (or simulated) data and the value of Δ under the null hypothesis, $\Delta = \Delta_j$. For a given N, and initially assuming π and σ^2 are known, we then obtain exact p-values through simulation with the following procedure:

- For each $\Delta_j \in \Delta$
  - Calculate the observed test statistic, $t_j^n = t(\mathbf{y}^n, \Delta_j)$.
  - Generate B data sets of size N from the model

$$\mathbf{y}_j^* \overset{\text{iid}}{\sim} \pi\mathcal{N}\left(\frac{\Delta_j}{2}, \sigma^2\right) + (1-\pi)\mathcal{N}\left(-\frac{\Delta_j}{2}, \sigma^2\right).$$

  - For each simulated $\mathbf{y}_j^*$, compute $t_j^* = t(\mathbf{y}_j^*, \Delta_j)$.
  - Calculate the empirical p-value of t_j^n as a function of the null distribution, t_j^*.
- Calculate the confidence set, $\text{CS}_\alpha(\Delta) = \{\Delta_j : p(\Delta_j) > 1-\alpha\}$ for a specified significance level α, where $p(\Delta_j)$ is the empirical p-value of $\widehat{\Delta}^{\text{mle}}$ assuming that $\Delta = \Delta_j$.

Note that the resulting confidence set might not be continuous, which could occur if the sampling distribution is strongly bimodal.

D.2 Constructing a test statistic

So long as the model is correctly specified, this approach yields an exact p-value for any valid test statistic, up to Monte Carlo error (Mikusheva, 2007). We propose a test statistic based on the joint distribution of $\widehat{m}_2$ and $\widehat{m}_3$.[12] Equation 4.5 suggests a natural combination of the estimated cumulants:

$$t_m(\mathbf{y}, \Delta_j) = (d_2, d_3)\text{Var}(m_2, m_3)^{-1}(d_2, d_3)^T, \tag{D.2}$$

[12]There are many possible alternatives. For example, Frumento et al. (2016) suggest test statistics based on scaled log-likelihood ratios. Another option is to use univariate test statistics based on $\widehat{m}_2$ or $\widehat{m}_3$.

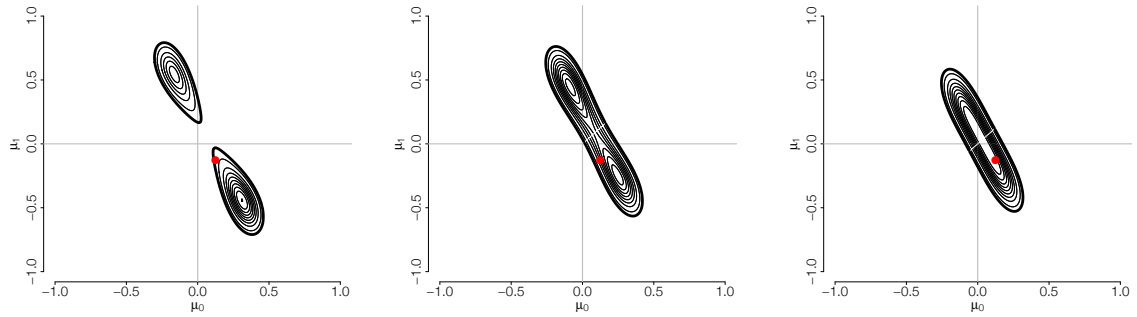


Figure D.10. Three examples of the grid of Wald test p-values from Equation D.3. The three simulated data sets were drawn from Equation (1.1) in the main text with $N = 1000$, $\pi = 0.325$, $\sigma^2 = 1$, $\mu_0 = \frac{1}{8}$, $\mu_1 = -\frac{1}{8}$. The dark line shows the cutoff for $p = 0.05$. The red dot shows the true value. Note that the Wald test is undefined when $\mu_0 = \mu_1$.

where $d_k = \widehat{m}_k - m_k$, and we use the assumed null of $\Delta = \Delta_j$ to obtain (m_2, m_3) and $\text{Var}(m_2, m_3)$. As we saw, the Normal approximation in Equation (4.5) in the main text is excellent, even for modest sample sizes (say $N > 100$). This implies:

$$t_m(\mathbf{y}, \Delta_j) \overset{a}{\sim} \chi^2_2.$$

We can therefore obtain a p-value via a Wald test, rather than via simulation, at each grid point, which is much faster computationally.

Finally, to use these approaches to estimate component means, we need to (1) expand the grid, and (2) expand the test statistic. A natural choice for a grid of points is the two-dimensional grid over μ_0 and μ_1. To expand the test statistic, we directly use the first three cumulants from Equation (4.5) from the main text and from (Tan and Chang, 1972) to obtain a joint test statistic as in Equation (D.2):

$$t_m(\mathbf{y}, \Delta_j) = (d_1, d_2, d_3)\text{Var}(\kappa_1, \kappa_2, \kappa_3)^{-1}(d_1, d_2, d_3)^T \sim \chi^2_3. \tag{D.3}$$

As above, we can obtain p-values via the grid bootstrap rather than via the χ^2 distribution. Figure D.10 shows the distribution of p-values for three different examples from the same data generating process, with $N = 1000$, $\pi = 0.325$, $\sigma^2 = 1$, $\mu_0 = +\frac{1}{8}$, $\mu_1 = -\frac{1}{8}$.[13]

Figure D.11 shows the 95% coverage for the confidence sets obtained through this fast approximation. As expected, the coverage is essentially exact. In particular, 95% coverage for this procedure is far better than the corresponding coverage based on the MLE.

[13]Note that the χ^2 distribution no longer holds when $\mu_0 = \mu_1$. While we can use a univariate Normal distribution to obtain a valid p-value in this case, this additional complication is generally unnecessary in practice.

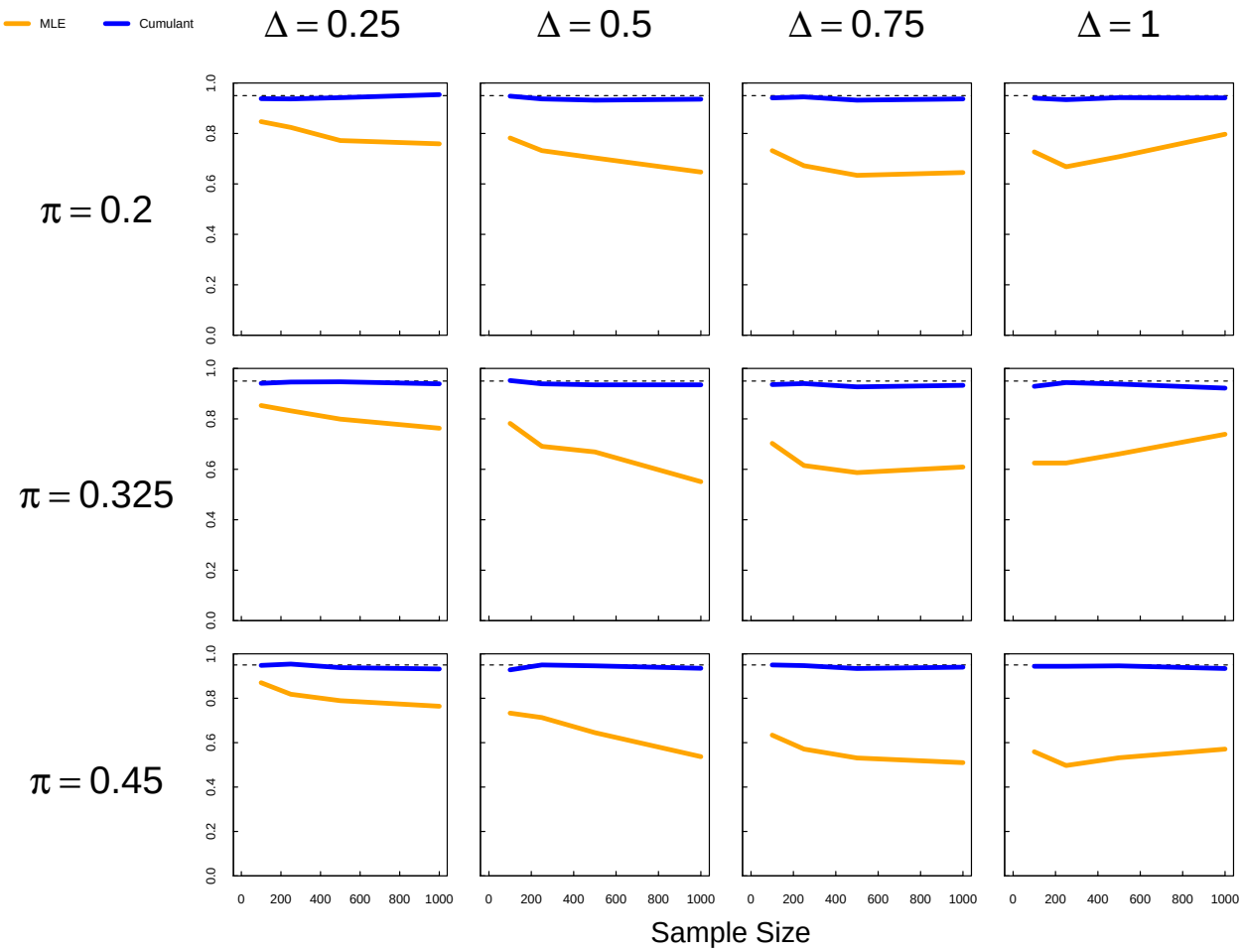


Figure D.11. Coverage for 95% confidence sets based on the test inversion algorithm described in Section D. The results for the MLE are for the standard finite mixtures estimator.

D.3 Grid bootstrap for principal stratification model

In the full principal stratification model, we directly estimate the outcome means for Compliers and Never Takers assigned to treatment, $\widehat{\mu}_{\text{c1}}$ and $\widehat{\mu}_{\text{n1}}$, and use the finite mixture model to estimate corresponding outcome means for Compliers and Never Takers assigned to control, $\widehat{\mu}_{\text{c0}}$ and $\widehat{\mu}_{\text{n0}}$. Our goal is inference for $\text{ITT}_{\text{c}} = \widehat{\mu}_{\text{c1}} - \widehat{\mu}_{\text{c0}}$ and $\text{ITT}_{\text{n}} = \widehat{\mu}_{\text{n1}} - \widehat{\mu}_{\text{n0}}$. While this is straightforward given estimates for μ_{c0} and μ_{n0}, we only have confidence sets for these means.

We therefore propose the following approach to obtaining $(1-\alpha)100\%$ confidence sets for ITT_{c} and ITT_{n}:

- Use a grid bootstrap or test inversion to obtain a joint $(1-\alpha/2)100\%$ confidence set for μ_{c0} and μ_{n0}, which we can project into univariate confidence sets, $\text{CS}_{\alpha/2}(\mu_{\text{c0}})$ and $\text{CS}_{\alpha/2}(\mu_{\text{n0}})$
- Directly obtain $(1-\alpha/2)100\%$ confidence intervals via the Normal distribution for μ_{c1} and μ_{n1}, $\text{CS}_{\alpha/2}(\mu_{\text{c1}})$ and $\text{CS}_{\alpha/2}(\mu_{\text{n1}})$
- For ITT_{c} (repeat for ITT_{n}):
  - If $\text{CS}_{\alpha/2}(\mu_{\text{c0}})$ is not disjoint, obtain a $(1-\alpha)100\%$ confidence interval for ITT_{c}:

$$\text{CS}^{UB}_{\alpha}(\text{ITT}_{\text{c}}) = \text{CS}^{UB}_{\alpha/2}(\mu_{\text{c1}}) - \text{CS}^{LB}_{\alpha/2}(\mu_{\text{c0}})$$
$$\text{CS}^{LB}_{\alpha}(\text{ITT}_{\text{c}}) = \text{CS}^{LB}_{\alpha/2}(\mu_{\text{c1}}) - \text{CS}^{UB}_{\alpha/2}(\mu_{\text{c0}})$$

  - If $\text{CS}_{\alpha/2}(\mu_{\text{c0}})$ is disjoint, repeat the above calculations for each separate segment and then take the union

This yields valid confidence sets for both treatment effects of interest. If desired, we could incorporate an additional Bonferroni correction to account for the two separate intervals.

Finally, if desired, we can extend this procedure to account for uncertainty in π and σ, which are nuisance parameters for the desired hypothesis tests. We can therefore use results from (Berger and Boos, 1994) to obtain valid p-values in this context. First, we obtain a $(1-\gamma)$-level joint confidence set for $CS_{\gamma}(\pi, \sigma^2)$, such as via case-resampling bootstrap, with γ very small, such as $\gamma = 0.001$. We obtain a valid p-value for, say, Δ, by taking the maximum p-value over $CS_{\gamma}(\pi, \sigma^2)$ plus a correction for the added uncertainty:

$$p_{\gamma}(\Delta_0) = \sup_{(\pi,\sigma^2) \in CS_{\gamma}(\pi,\sigma^2)} p(\Delta_0) + \gamma.$$

See (Nolen and Hudgens, 2011) and (Ding et al., 2016) for further discussion of the validity of this approach.

E Failure of resampling methods

Resampling methods, such as the case-resampling bootstrap, are common in finite mixture model settings. For example, (McLachlan and Peel, 2004, Sec. 2.16.2) recommend using the bootstrap to improve estimation of standard errors when the Fisher information yields a poor approximation (see also Grün and Leisch, 2004). Others have suggested subsampling in similar settings (Andrews, 2000). Figure E.12 shows the coverage for 95% confidence sets based on the case-resampling and subsampling intervals. Clearly, the coverage is far from nominal.

The form of $\widehat{\Delta}^{\text{mom}}$ shows why the performance of these methods is so poor. As (Bickel and Freedman, 1981) prove, for the bootstrap to be consistent in the iid context, the mapping from the underlying distribution of the data to the distribution of the statistic must be continuous (see also Andrews, 2000). Clearly,

$$\widehat{\Delta}^{\text{mom}} = \text{sgn}(\widehat{m}_3)\sqrt{\frac{\widehat{m}_2 - 1}{\pi(1-\pi)}}$$

is not a continuous mapping from the sample to $\widehat{\Delta}^{\text{mom}}$, with a boundary at $m_2 \geq 1$ and a discontinuity at $m_3 = 0$.[14] In the related case of the unit root problem, (Mikusheva, 2007) shows that other resampling methods also fail, including subsampling and the m of n bootstrap. In the context of principal stratification, (Zhang et al., 2009) note that confidence intervals based on the bootstrap often fail when the likelihood is multimodal. (Frumento et al., 2016) offer additional discussion in this setting.

F Frequency Performance of the Posterior Mean and Median

Bayesian inference for finite mixtures introduces some unique challenges for specifying priors (e.g., Grazian and Robert, 2015). Nonetheless, inference for a posterior with a sufficiently vague prior should be broadly similar to inference based on the likelihood alone. Thus, without an informative prior for $\{\mu_0, \mu_1\}$ in the two-component Gaussian mixture, the posterior mean and median should exhibit similar pathologies to those exhibited by the MLE. We test this intuition using the `bayesm` package in R. Figure F.13 shows histograms of the posterior mean of Δ when the true Δ is 0.5 and 1, $\pi = 0.3$, and $N = 100$. We use the default priors of the `bayesm` package except in the case of the Dirichlet parameter, which is set to reflect that $\pi = 0.3$ is known (i.e., we assume a very informative prior). The histograms exhibit the same behavior as the MLE of Δ. In particular, the estimator concentrates around 0 and seems unable to differentiate between $\Delta > 0$ and $\Delta < 0$.

Figure F.14 shows the corresponding plot for the distribution of the posterior median of

[14] In some promising recent work, (Laber and Murphy, 2011) explore bootstrap-type methods with non-continuous mappings. We hope to explore this more in the future.

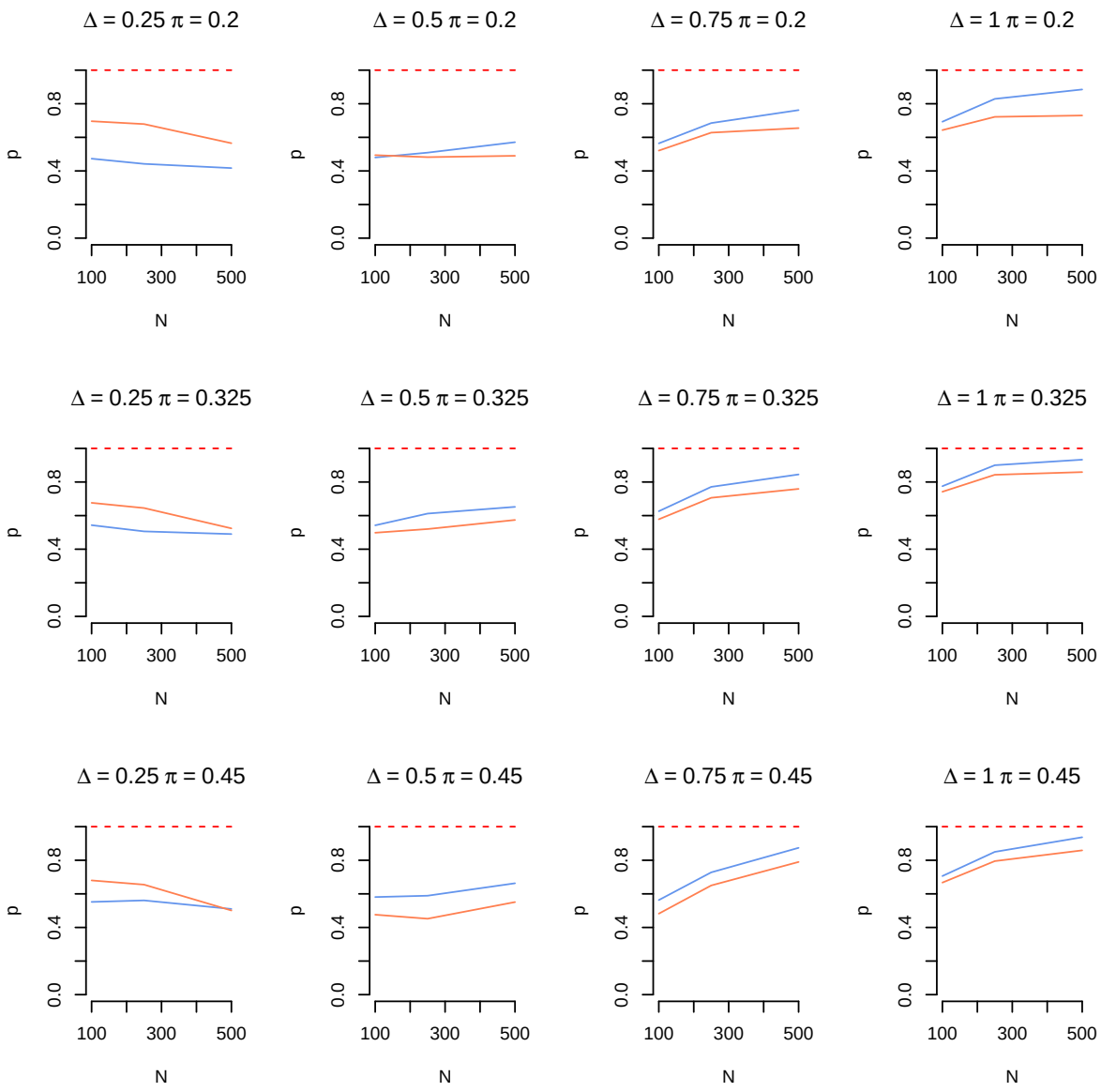
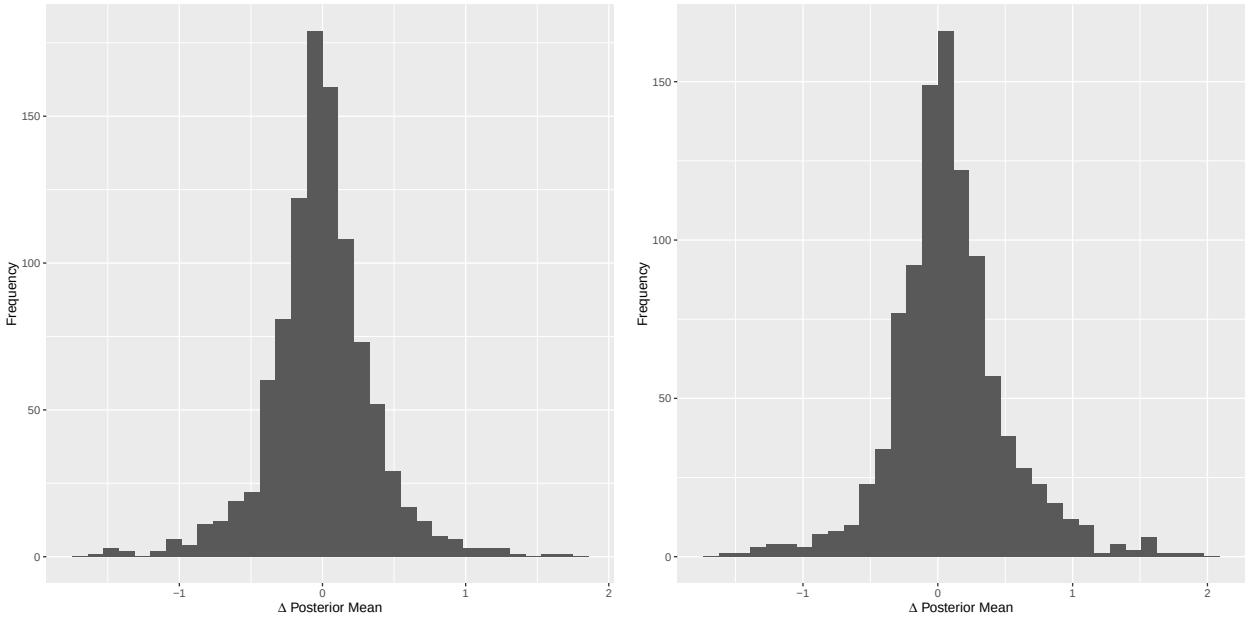


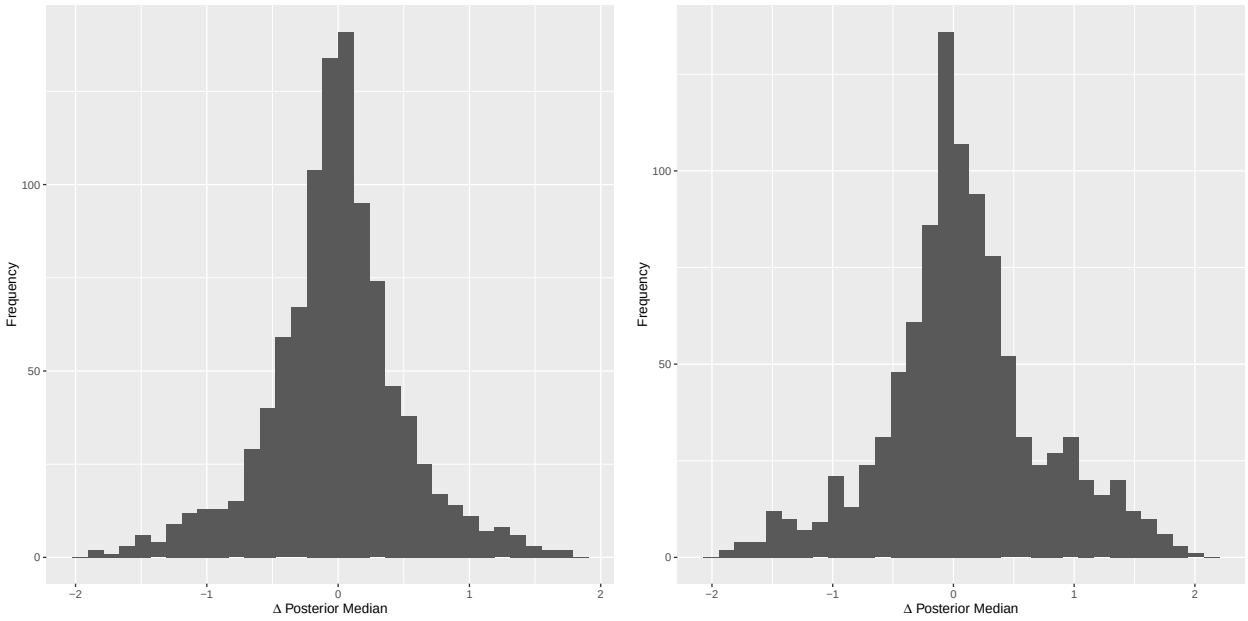
Figure E.12. Coverage probabilities for 95% confidence sets based on the case-resampling and subsampling intervals. The blue line represents the case-resampling coverage probability, while the blue line represents the subsampling coverage probability.



(a) $\Delta = 0.5$ (b) $\Delta = 1$

Figure F.13. Histograms of the posterior mean for Δ calculated via MCMC draws from `bayesm`. The histogram on the left is for $\Delta = 0.5$, while the histogram on the right is for $\Delta = 1$. Both histograms have $N = 100$, $\pi = 0.3$, and $\sigma = 1$.

Δ. As we can see, the median also concentrates about 0 and appears unable to determine the sign of Δ.



(a) $\Delta = 0.5$ (b) $\Delta = 1$

Figure F.14. Histograms of the posterior median for Δ calculated via MCMC draws from `bayesm`. The histogram on the left is for $\Delta = 0.5$, while the histogram on the right is for $\Delta = 1$. Both histograms have $N = 100$, $\pi = 0.3$, and $\sigma = 1$.

G Proofs

In this appendix, we provide detailed proofs for the key asymptotic results in Section 2. We first start with the proof regarding convergence rates of $\widehat{\delta}_n^{\text{mle}}$ and $\left|\widehat{\delta}_n^{\text{mle}}\right|$ under the asymmetric and symmetric setting of model (2.1).

G.1 PROOF OF THEOREM 2.1

Throughout this proof, for the ease of presentation, we denote

$$g(x, \delta) := \pi\phi(x, -\delta) + (1 - \pi)\phi(x, c\delta),$$

for any $\delta \in \Theta$ where $\{\phi(x, \delta)\}$ denotes the family of Gaussian distribution with location parameter δ and scale is fixed to be 1. Additionally, we also remind that $c = \pi/(1 - \pi)$, with this quantity thus being a known constant. To streamline the argument, we divide the proof into two parts. In Section G.1.1, we provide the proof for the upper bounds of the convergence rate of MLE. Then, in Section G.1.2, we present the proof for the lower bounds.

G.1.1 Proof for upper bounds

The proof technique for the upper bounds utilizes the strategy of comparing the convergence rate of density estimation to that of parameter estimation in mixture models, which had been employed successfully in the previous work (Chen, 1995; Nguyen, 2013; Ho and Nguyen, 2016; Heinrich and Kahn, 2018).

Convergence rate of density estimation The convergence rate of density estimation in Gaussian mixture models had been studied rigorously in the literature (Ghosal and van der Vaart, 2001). Regarding our model (2.1), we have the following result regarding the convergence rate of $g(x, \widehat{\delta}_n^{\text{mle}})$ to $g(x, \delta_n)$ under Hellinger metric.

Proposition 3. *Under the setting of model (2.1), the following holds*

$$\sup_{\delta_n \in \Theta} \mathbb{E}_{\delta_n}\left(h\left(g(x, \widehat{\delta}_n^{mle}), g(x, \delta_n)\right)\right) \lesssim \left(\frac{\log n}{n}\right)^{1/2},$$

where Θ is a bounded (growing) parameter space. Here, $\mathbb{E}_{\delta_n}$ denotes the expectation taken with respect to product measure with mixture density of $Y_1, \ldots, Y_n$ under the model (2.1).

The proof of the above result follows from a standard application of Theorem 7.4 in (van de Geer, 2000); therefore, it is omitted.

From density estimation to parameter estimation Equipped with $(\log n/n)^{1/2}$ rate of density estimation in Proposition (3), to achieve the convergence rates of $\widehat{\delta}_n^{\text{mle}}$ and $\left|\widehat{\delta}_n^{\text{mle}}\right|$

under the asymmetric and symmetric setting of model (2.1), it is sufficient to demonstrate the following result:

Lemma G.1. *Given $\pi \in (0, 1/2]$ and $\Theta = [-1, 1]$, the following holds*

(a) (Asymmetric regime) When $\pi \in (0, 1/2)$, then

$$\inf_{\delta^{(1)},\delta^{(2)}\in\Theta} h\left(g(x,\delta^{(1)}), g(x,\delta^{(2)})\right) / \left|\delta^{(1)} - \delta^{(2)}\right|^3 > 0.$$

(b) (Symmetric regime) When $\pi = 1/2$, then

$$\inf_{\delta^{(1)},\delta^{(2)}\in\Theta} h\left(g(x,\delta^{(1)}), g(x,\delta^{(2)})\right) / \left||\delta^{(1)}| - |\delta^{(2)}|\right|^2 > 0.$$

Proof. (a) Due to the basic inequality between total variational distance and Hellinger distance $h \geq V$, it suffices to prove that

$$\inf_{\delta^{(1)},\delta^{(2)}\in\Theta} V\left(g(x,\delta^{(1)}), g(x,\delta^{(2)})\right) / |\delta^{(1)} - \delta^{(2)}|^3 > 0. \tag{G.1}$$

Assume that the conclusion of (G.1) does not hold. It implies that we can find two sequences $\left\{\delta_n^{(1)}\right\}$ and $\left\{\delta_n^{(2)}\right\}$ such that $V(g(x,\delta_n^{(1)}), g(x,\delta_n^{(2)}))/|\delta_n^{(1)} - \delta_n^{(2)}|^3 \to 0$ as $n \to \infty$. For the simplicity of the presentation, we only the consider the most challenging setting of sequences $\left\{\delta_n^{(1)}\right\}$ and $\left\{\delta_n^{(2)}\right\}$ when $\delta_n^{(1)} \to 0$, $\delta_n^{(2)} \to 0$ as $n \to \infty$. The proof for other possibilities of these sequences can be argued in the similar fashion. Now, we have two distinct cases regarding the convergence of $\delta_n^{(1)}$ and $\delta_n^{(2)}$.

Case a.1: $\delta_n^{(1)}/\delta_n^{(2)} \not\to 1$ as $n \to \infty$ (Here, the limit can be thought as that of some subsequence of $\delta_n^{(1)}/\delta_n^{(2)}$. However, we replace this subsequence by the whole sequence of $\delta_n^{(1)}/\delta_n^{(2)}$ for the simplicity of the presentation). Under this case, we divide our argument into several steps.

Step 1 - Taylor expansion Now, the following equality holds

$$\frac{g(x,\delta_n^{(1)}) - g(x,\delta_n^{(2)})}{|\delta_n^{(1)} - \delta_n^{(2)}|^3} = \frac{\pi(\phi(x,-\delta_n^{(1)}) - \phi(x,-\delta_n^{(2)}))}{|\delta_n^{(1)} - \delta_n^{(2)}|^3} + \frac{(1-\pi)(\phi(x,c\delta_n^{(1)}) - \phi(x,c\delta_n^{(2)}))}{|\delta_n^{(1)} - \delta_n^{(2)}|^3}.$$

Invoking Taylor expansion up to the third order, we obtain that

$$\begin{aligned}
\phi(x,-\delta_n^{(1)})-\phi(x,-\delta_n^{(2)}) &= \sum_{\alpha=1}^{3}\frac{(\delta_n^{(2)}-\delta_n^{(1)})^\alpha}{\alpha!}\frac{\partial^\alpha\phi}{\partial\delta^\alpha}(x,-\delta_n^{(2)})+R_1(x),\\
\phi(x,c\delta_n^{(1)})-\phi(x,c\delta_n^{(2)}) &= \sum_{\alpha=1}^{3}\frac{c^\alpha(\delta_n^{(1)}-\delta_n^{(2)})^\alpha}{\alpha!}\frac{\partial^\alpha\phi}{\partial\delta^\alpha}(x,c\delta_n^{(2)})+R_2(x)\\
&= \sum_{\alpha=1}^{3}\frac{c^\alpha(\delta_n^{(1)}-\delta_n^{(2)})^\alpha}{\alpha!}\Bigg(\sum_{\tau=0}^{3-\alpha}\frac{(c+1)^\tau(\delta_n^{(2)})^\tau}{\tau!}\frac{\partial^{\alpha+\tau}\phi}{\partial\delta^{\alpha+\tau}}(x,-\delta_n^{(2)})\\
&\qquad +R_{2,\alpha}(x)\Bigg)+R_2(x),
\end{aligned}$$

where $R_1(x)$, $R_2(x)$ are respectively the Taylor remainders up to the third order from performing Taylor expansion around $-\delta_n^{(2)}$ and $c\delta_n^{(2)}$ while $R_{2,\alpha}$ are Taylor remainders up to the order $3-\alpha$ from performing Taylor expansion around $-\delta_n^2$ in $\dfrac{\partial^\alpha\phi}{\partial\delta^\alpha}(x,c\delta_n^{(2)})$ as $1\leq\alpha\leq 3$. Here, the Taylor remainders $R_1(x)$ and $R_2(x)$ satisfy

$$\max\{\|R_1(x)\|_\infty,\|R_2(x)\|_\infty\}=O\left(|\delta_n^{(1)}-\delta_n^{(2)}|^{3+\gamma}\right), \tag{G.2}$$

where $\gamma>0$ is some positive constant. It implies that $R_1(x)/|\delta_n^{(1)}-\delta_n^{(2)}|^3\to 0$ and $R_2(x)/|\delta_n^{(1)}-\delta_n^{(2)}|^3\to 0$ for all $x\in\mathbb{R}$. Similarly, $\|R_{2,\alpha}(x)\|_\infty=O(|\delta_n^{(2)}|^{3-\alpha+\gamma})$ as $1\leq\alpha\leq 3$. As $\delta_n^{(1)}/\delta_n^{(2)}\not\to 1$, we have $|\delta_n^{(2)}|/|\delta_n^{(1)}-\delta_n^{(2)}|\not\to+\infty$. Therefore, we have $|\delta_n^{(2)}|^{r-\alpha+\gamma}/|\delta_n^{(1)}-\delta_n^{(1)}|^{r-\alpha}\to 0$ as $n\to\infty$, which eventually leads to

$$(\delta_n^{(1)}-\delta_n^{(2)})^\alpha\|R_{2,\alpha}(x)\|_\infty/|\delta_n^{(1)}-\delta_n^{(2)}|^3\to 0 \tag{G.3}$$

for all $1\leq\alpha\leq 3$. Governed by the previous results, the following representation holds

$$\begin{aligned}
\frac{g(x,\delta_n^{(1)})-g(x,\delta_n^{(2)})}{|\delta_n^{(1)}-\delta_n^{(2)}|^3} &= \frac{\pi\Bigg(\sum_{\alpha=1}^{3}\dfrac{(\delta_n^{(2)}-\delta_n^{(1)})^\alpha}{\alpha!}\dfrac{\partial^\alpha\phi}{\partial\delta^\alpha}(x,-\delta_n^{(2)})+R_1(x)\Bigg)}{|\delta_n^{(1)}-\delta_n^{(2)}|^3}\\
&\quad+\frac{(1-\pi)\Bigg(\sum_{\alpha=1}^{3}\dfrac{c^\alpha(\delta_n^{(1)}-\delta_n^{(2)})^\alpha}{\alpha!}\Bigg(\sum_{\tau=0}^{3-\alpha}\dfrac{(c+1)^\tau(\delta_n^{(2)})^\tau}{\tau!}\dfrac{\partial^{\alpha+\tau}\phi}{\partial\delta^{\alpha+\tau}}(x,-\delta_n^{(2)})+R_{2,\alpha}(x)\Bigg)+R_2(x)\Bigg)}{|\delta_n^{(1)}-\delta_n^{(2)}|^3}\\
&:= \frac{\sum_{\alpha=1}^{3}A_{n,\alpha}\dfrac{\partial^\alpha\phi}{\partial\delta^\alpha}(x,-\delta_n^{(2)})+R(x)}{|\delta_n^{(1)}-\delta_n^{(2)}|^3},
\end{aligned} \tag{G.4}$$

where $R(x) = \pi R_1(x) + (1-\pi)\sum_{\alpha=1}^{3} \frac{c^\alpha(\delta_n^{(1)} - \delta_n^{(2)})^\alpha}{\alpha!} R_{2,\alpha}(x) + (1-\pi)R_2(x)$ for all $x \in \mathbb{R}$. Invoking the bounds with Taylor remainders $R_1(x), R_2(x)$, and $R_{2,\alpha}(x)$ in (G.2), (G.3), we have $\|R(x)\|_\infty / |\delta_n^{(1)} - \delta_n^{(2)}|^3 \to 0$ as $n \to \infty$.

Step 2 - Non-vanishing coefficients Assume that the coefficients $A_{n,\alpha}/|\delta_n^{(1)} - \delta_n^{(2)}|^3 \to 0$ as $n \to \infty$ for all $1 \leq \alpha \leq 3$. From the formulations of $A_{n,\alpha}$ in (G.4), we can quickly compute that $A_{n,1} = 0$ while

$$\begin{aligned}
A_{n,2} &= \frac{c}{2}(\delta_n^{(2)} - \delta_n^{(1)})(\delta_n^{(1)} + \delta_n^{(2)}), \\
A_{n,3} &= \frac{\pi(\delta_n^{(2)} - \delta_n^{(1)})^3}{3!} + (1-\pi)c(c+1)^2(\delta_n^{(2)})^2 \frac{(\delta_n^{(1)} - \delta_n^{(2)})}{2!} \\
&\quad + (1-\pi)(c+1)c^2 \frac{(\delta_n^{(1)} - \delta_n^{(2)})^2}{2!}\delta_n^{(2)} + (1-\pi)c^3 \frac{(\delta_n^{(1)} - \delta_n^{(2)})^3}{3!}.
\end{aligned}$$

As $A_{n,2}/|\delta_n^{(1)} - \delta_n^{(2)}|^3 \to 0$, it implies that $(\delta_n^{(1)} + \delta_n^{(2)})/|\delta_n^{(1)} - \delta_n^{(2)}|^2 \to 0$, which leads to $\delta_n^{(1)}/\delta_n^{(2)} \to -1$ as $n \to \infty$. Plugging this limit into $A_{n,3}/|\delta_n^{(1)} - \delta_n^{(2)}|^3 \to 0$ yields the following equation

$$\frac{8\pi}{3!} - (1-\pi)c(c+1)^2 + 2(1-\pi)c^2(c+1) - \frac{8(1-\pi)c^3}{3!} = 0,$$

which has only a unique solution $\pi = 1/2$, a contradiction to the assumption of asymmetric setting, i.e., $\pi \in (0, 1/2)$. Therefore, not all the coefficients $A_{n,\alpha}/|\delta_n^{(1)} - \delta_n^{(2)}|^3 \to 0$ when $n \to \infty$ as $1 \leq \alpha \leq 3$.

Step 3 - Fatou's argument Denote $m_n = |\delta_n^{(1)} - \delta_n^{(2)}|^3 / \max_{1 \leq \alpha \leq 3} |A_{n,\alpha}|$. Since not all the coefficients $A_{n,\alpha}/|\delta_n^{(1)} - \delta_n^{(2)}|^3 \to 0$ as $1 \leq \alpha \leq 3$, we have $m_n \not\to \infty$. Therefore, we obtain that

$$m_n \frac{g(x, \delta_n^{(1)}) - g(x, \delta_n^{(2)})}{|\delta_n^{(1)} - \delta_n^{(2)}|^3} = m_n \frac{\sum_{\alpha=1}^{3} A_{n,\alpha} \frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x, -\delta_n^{(2)}) + R(x)}{|\delta_n^{(1)} - \delta_n^{(2)}|^3} \to \sum_{\alpha=1}^{3} \beta_\alpha \frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x, 0),$$

for all x where $A_{n,\alpha}/\max_{1\leq\alpha\leq 3}|A_{n,\alpha}| \to \beta_\alpha$ as $1 \leq \alpha \leq 3$ such that at least one of β_α has absolute value to be 1. Invoking Fatou's lemma, the following holds

$$0 = \lim_{n\to\infty} \frac{m_n V\left(g(x,\delta_n^{(1)}), g(x,\delta_n^{(2)})\right)}{|\delta^{(1)} - \delta^{(2)}|^3} \geq \int \liminf_{n\to\infty} \frac{m_n \left|g(x,\delta_n^{(1)}) - g(x,\delta_n^{(2)})\right|}{|\delta_n^{(1)} - \delta_n^{(2)}|^3} dx$$
$$= \int \sum_{\alpha=1}^{3} \beta_\alpha \frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x,0) dx.$$

The above inequality leads to $\sum_{\alpha=1}^{3} \beta_\alpha \frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x,0) = 0$ for almost surely x. Nevertheless, due to the strong order identifiability of location Gaussian distribution Chen (1995), the above equation implies that $\beta_\alpha = 0$ for all $1 \leq \alpha \leq 3$, which is a contradiction. Therefore, Case a.1 cannot holds.

Case a.2: $\delta_n^{(1)}/\delta_n^{(2)} \to 1$ as $n \to \infty$. It implies that $|\delta_n^{(2)}|/|\delta_n^{(1)} - \delta_n^{(2)}| \to \infty$ as $n \to \infty$. As $V\left(g(x,\delta_n^{(1)}), g(x,\delta_n^{(2)})\right)/|\delta_n^{(1)} - \delta_n^{(2)}|^3 \to 0$, it implies that

$$V\left(g(x,\delta_n^{(1)}), g(x,\delta_n^{(2)})\right)/|\delta_n^{(1)} - \delta_n^{(2)}|^2 \to 0,$$

as $n \to \infty$ for all $x \in \mathbb{R}$. Similar to the Taylor expansion argument in Step 1 in Case a.1, by means of Taylor expansion up to the second order, we obtain that

$$\begin{aligned}
\frac{g(x,\delta_n^{(1)}) - g(x,\delta_n^{(2)})}{|\delta_n^{(1)} - \delta_n^{(2)}|^2} &= \frac{\pi(\phi(x,-\delta_n^{(1)}) - \phi(x,-\delta_n^{(2)})) + (1-\pi)(\phi(x,c\delta_n^{(1)}) - \phi(x,c\delta_n^{(2)}))}{|\delta_n^{(1)} - \delta_n^{(2)}|^2} \\
&= \frac{\pi\left(\sum_{\alpha=1}^{2} \frac{(\delta_n^{(2)} - \delta_n^{(1)})^\alpha}{\alpha!} \frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x,-\delta_n^{(2)}) + R_1'(x)\right)}{|\delta_n^{(1)} - \delta_n^{(2)}|^2} \\
&+ \frac{(1-\pi)\left(\sum_{\alpha=1}^{2} \frac{c^\alpha(\delta_n^{(1)} - \delta_n^{(2)})^\alpha}{\alpha!}\left(\sum_{\tau=0}^{2-\alpha} \frac{(c+1)^\tau(\delta_n^{(2)})^\tau}{\tau!} \frac{\partial^{\alpha+\tau}\phi}{\partial\delta^{\alpha+\tau}}(x,-\delta_n^{(2)}) + R_{2,\alpha}'(x)\right) + R_2'(x)\right)}{|\delta_n^{(1)} - \delta_n^{(2)}|^2} \\
&= \frac{\sum_{\alpha=1}^{2} A_{n,\alpha} \frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x,-\delta_n^{(2)}) + R'(x)}{|\delta_n^{(1)} - \delta_n^{(2)}|^2} \to 0,
\end{aligned}$$

where $\|R'(x)\|_\infty = O\left(|\delta_n^{(2)}|^{1+\gamma}|\delta_n^{(1)} - \delta_n^{(2)}|\right)$ for some $\gamma > 0$. By means of the calculations with $A_{n,\alpha}$ in Case a.1, we have

$$\frac{\|R'(x)\|_\infty}{|A_{n,2}|} = \frac{O\left(\left|\delta_n^{(2)}\right|^{1+\gamma}\left|\delta_n^{(1)} - \delta_n^{(2)}\right|\right)}{\left|\delta_n^{(2)} - \delta_n^{(1)}\right|\left|\delta_n^{(1)} + \delta_n^{(2)}\right|} \to 0.$$

Now, if $A_{n,\alpha}/|\delta_n^{(1)} - \delta_n^{(2)}|^2 \to 0$ for all $1 \leq \alpha \leq 2$, we have $|\delta_n^{(1)} + \delta_n^{(2)}|/|\delta_n^{(1)} - \delta_n^{(2)}| \to 0$, which implies that $\delta_n^{(1)}/\delta_n^{(2)} \to -1$, a contradiction to the assumption of Case a.2. According to the argument in Step 3 in Case a.1, by denoting $m_n' = |\delta_n^{(1)} - \delta_n^{(2)}|^2/\max_{1\leq\alpha\leq 2}|A_{n,\alpha}|$, we have $m_n' \not\to \infty$. Therefore, we have

$$m_n' \frac{g(x, \delta_n^{(1)}) - g(x, \delta_n^{(2)})}{|\delta_n^{(1)} - \delta_n^{(2)}|^2} \to \sum_{\alpha=1}^{2} \tau_\alpha \frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x, 0),$$

for all x for some coefficients τ_α such that at least one of them has absolute value to be 1. By virtue of Fatou's lemma in Step 3 in Case a.1 with $\lim_{n\to\infty} V\left(g(x, \delta_n^{(1)}), g(x, \delta_n^{(2)})\right)/|\delta_n^{(1)} - \delta_n^{(2)}|^2$, we achieve that $\sum_{\alpha=1}^{2} \tau_\alpha \frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x, 0) = 0$ for almost surely x. However, the strong identifability of location Gaussian distribution implies that $\tau_\alpha = 0$ for all $1 \leq \alpha \leq 2$, which is a contradiction. Therefore, Case a.2 cannot happen.

Combining the results from Case a.1 and Case a.2, we achieve the conclusion of (G.1). As a consequence, the conclusion of part (a) of Lemma G.1 follows.

(b) Similar to the proof strategy of part (a), to obtain the conclusion of this result, it is sufficient to demonstrate that

$$\inf_{\delta^{(1)}, \delta^{(2)} \in \delta} V\left(g(x, \delta^{(1)}), g(x, \delta^{(2)})\right)/\left||\delta^{(1)}| - |\delta^{(2)}|\right|^2 > 0. \tag{G.5}$$

Assume that the conclusion of (G.5) does not hold. It implies that we can find two sequences $\left\{\delta_n^{(1)}\right\}$ and $\left\{\delta_n^{(2)}\right\}$ such that

$$V\left(g(x, \delta_n^{(1)}), g(x, \delta_n^{(2)})\right)/\left||\delta_n^{(1)}| - |\delta_n^{(2)}|\right|^2 \to 0$$

as $n \to \infty$. Similar to the proof argument of part (a), we only consider the possibility that $\delta_n^{(1)} \to 0$ and $\delta_n^{(2)} \to 0$ as $n \to \infty$. Now, we have two different settings of $\delta_n^{(1)}$ and $\delta_n^{(2)}$.

Case b.1: $\delta_n^{(1)}/\delta_n^{(2)} \not\to 1$ as $n \to \infty$ and $\delta_n^{(1)}\delta_n^{(2)} \geq 0$ for all n (Here, the limit and the inequality can be thought as those of some subsequence of $\delta_n^{(1)}$ and $\delta_n^{(2)}$. However, we replace

this subsequence by the whole sequence of $\delta_n^{(1)}$ and $\delta_n^{(2)}$ for the simplicity of the presentation). Under that setting, we have

$$\frac{V\left(g(x,\delta_n^{(1)}),g(x,\delta_n^{(2)})\right)}{\left||\delta_n^{(1)}|-|\delta_n^{(2)}|\right|^2}=\frac{V\left(g(x,\delta_n^{(1)}),g(x,\delta_n^{(2)})\right)}{|\delta_n^{(1)}-\delta_n^{(2)}|^2}\to 0.$$

To ease the understanding, we divide our argument for Case b.1 into two separate steps.

Step 1 - Taylor expansion By means of Taylor expansion up to the second order as that of Case a.2 in the proof of part (a), we obtain that

$$\frac{g(x,\delta_n^{(1)})-g(x,\delta_n^{(2)})}{|\delta_n^{(1)}-\delta_n^{(2)}|^2}=\frac{\sum\limits_{\alpha=1}^{2}A_{n,\alpha}\dfrac{\partial^\alpha\phi}{\partial\delta^\alpha}(x,-\delta_n^{(2)})+R'(x)}{\left|\delta_n^{(1)}-\delta_n^{(2)}\right|^2}\to 0,$$

where $R'(x)$ is a combination of Taylor remainders such that

$$\|R'(x)\|_\infty=O\left(\left|\delta_n^{(2)}\right|^{1+\gamma}\left|\delta_n^{(1)}-\delta_n^{(2)}\right|\right),$$

for some positive constant γ and $A_{n,\alpha}$ are defined as in that in Case a.2 when $\pi=1/2$. Since $\delta_n^{(1)}/\delta_n^{(2)}\not\to 1$, we have $|\delta_n^{(2)}|/|\delta_n^{(1)}-\delta_n^{(2)}|\not\to\infty$. Therefore, it leads to

$$\|R(x)\|_\infty/|\delta_n^{(1)}-\delta_n^{(2)}|\to 0$$

as $n\to\infty$.

Step 2 - Non-vanishing coefficients and Fatou's argument Assume that $A_{n,\alpha}/|\delta_n^{(1)}-\delta_n^{(2)}|^2\to 0$ for all $1\leq\alpha\leq 2$. From the formulation of $A_{n,2}$, we have

$$(\delta_n^{(1)}+\delta_n^{(2)})/|\delta_n^{(1)}-\delta_n^{(2)}|\to 0.$$

It implies that $\delta_n^{(1)}/\delta_n^{(2)}\to -1$ as $n\to\infty$, which is a contradiction to the condition that $\delta_n^{(1)}\delta_n^{(2)}\geq 0$. Therefore, not all of the coefficients of $A_{n,\alpha}/|\delta_n^{(1)}-\delta_n^{(2)}|^2$ go to 0. From here, by means of the Fatou's argument in Step 3 of Case a.1, we achieve the conclusion that Case b.1 cannot hold.

Case b.2 $\delta_n^{(1)}/\delta_n^{(2)} \not\to 1$ and $\delta_n^{(1)}\delta_n^{(2)} < 0$ for all n. Under that setting, we have

$$\frac{V\left(g(x,\delta_n^{(1)}), g(x,\delta_n^{(2)})\right)}{\left||\delta_n^{(1)}| - |\delta_n^{(2)}|\right|^2} = \frac{V\left(g(x,\delta_n^{(1)}), g(x,\delta_n^{(2)})\right)}{\left|\delta_n^{(1)} + \delta_n^{(2)}\right|^2} \to 0.$$

We also divide the argument of Case b.2 into two main key steps.

Step 1 - Taylor expansion By means of Taylor expansion up to the second order, we obtain

$$\begin{aligned}
\frac{g(x,\delta_n^{(1)}) - g(x,\delta_n^{(2)})}{|\delta_n^{(1)} + \delta_n^{(2)}|^2} &= \frac{\frac{1}{2}(\phi(x,-\delta_n^{(1)}) - \phi(x,\delta_n^{(2)})) + \frac{1}{2}(\phi(x,\delta_n^{(1)}) - \phi(x,-\delta_n^{(2)}))}{|\delta_n^{(1)} + \delta_n^{(2)}|^2} \\
&= \frac{\frac{1}{2}\left(\sum_{\alpha=1}^{2} \frac{(-\delta_n^{(2)} - \delta_n^{(1)})^\alpha}{\alpha!}\frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x,\delta_n^{(2)}) + R_1''(x)\right)}{|\delta_n^{(1)} + \delta_n^{(2)}|^2} \\
&+ \frac{\frac{1}{2}\left(\sum_{\alpha=1}^{2} \frac{(\delta_n^{(1)} + \delta_n^{(2)})^\alpha}{\alpha!}\left(\sum_{\tau=0}^{2-\alpha} \frac{2^\tau(-\delta_n^{(2)})^\tau}{\tau!}\frac{\partial^{\alpha+\tau}\phi}{\partial\delta^{\alpha+\tau}}(x,\delta_n^{(2)}) + R_{2,\alpha}''(x)\right) + R_2''(x)\right)}{|\delta_n^{(1)} + \delta_n^{(2)}|^2} \\
&:= \frac{\sum_{\alpha=1}^{2} B_{n,\alpha}\frac{\partial^\alpha \phi}{\partial\delta^\alpha}(x,\delta_n^{(2)}) + R''(x)}{|\delta_n^{(1)} + \delta_n^{(2)}|^2} \to 0,
\end{aligned}$$

where $R''(x)$ is the combination of Taylor remainders such that

$$\|R''(x)\|_\infty = O\left(|\delta_n^{(2)}|^{1+\gamma}|\delta_n^{(1)} + \delta_n^{(2)}|\right),$$

which implies that $\|R''(x)\|_\infty/|\delta_n^{(1)} + \delta_n^{(2)}|^2 \to 0$ as $n \to \infty$.

Step 2 - Non-vanishing coefficients and Fatou's argument Assume that $B_{n,\alpha}/|\delta_n^{(1)} + \delta_n^{(2)}|^2 \to 0$ for all $1 \leq \alpha \leq 2$. Direct computation with $B_{n,2}$ implies that

$$(\delta_n^{(1)} - \delta_n^{(2)})/|\delta_n^{(1)} + \delta_n^{(2)}| \to 0$$

as $n \to \infty$. It leads to $\delta_n^{(1)}/\delta_n^{(2)} \to 1$, which is a contradiction to the assumption that $\delta_n^{(1)}\delta_n^{(2)} < 0$. From here, the Fatou's argument in Step 3 of Case a.1, we also obtain the conclusion that Case b.2 does not hold.

Case b.3 $\delta_n^{(1)}/\delta_n^{(2)} \to 1$ as $n \to \infty$. This implies that $\delta_n^{(1)}\delta_n^{(2)} > 0$ when n is sufficiently large. From here, the proof argument of this case is similar to that of Case a.2 in part (a), which also yields the contradiction.

As a consequence, we achieve the conclusion of part (b) of the lemma. □

G.1.2 Proof for lower bounds

(a) Based on the proof technique of Theorem 3.2 in Heinrich and Kahn (2018), to achieve the conclusion with the lower bound of part (a) of the theorem, it is sufficient to demonstrate that

$$\inf_{\delta^{(1)},\delta^{(2)}\in\Theta_{1,n}} h\left(g(x,\delta^{(1)}), g(x,\delta^{(2)})\right) / \left|\delta^{(1)} - \delta^{(2)}\right|^r = 0 \tag{G.6}$$

for any $1 \leq r < 3$. We divide the proof argument for the above result into several key steps.

Step 1 - Constructing sequences In fact, we construct two sequences $\left\{\delta_n^{(1)}\right\}$ and $\left\{\delta_n^{(2)}\right\}$ such that $\delta_n^{(1)} = -\delta_n^{(2)}$ for all $n \geq 1$ and $\delta_n^{(1)} \to 0$ as $n \to \infty$. For any fixed $r < 3$, by means of Taylor expansion up to the second order as that in Step 1 of Case a.1 in part (a) of Theorem 2.1 (cf. Equation (G.4)), the following holds

$$g(x,\delta_n^{(1)}) - g(x,\delta_n^{(2)}) = \sum_{\alpha=1}^{2} A_{n,\alpha} \frac{\partial^\alpha \phi}{\partial \delta^\alpha}(x, -\delta_n^{(2)}) + R(x),$$

where $R(x)$ is a combination of Taylor remainders where its detail formulation is postponed to later discussion. Additionally, the formulations of $A_{n,\alpha}$ satisfy $A_{n,1} = 0$ and

$$A_{n,2} = \frac{c}{2}(\delta_n^{(2)} - \delta_n^{(1)})(\delta_n^{(1)} + \delta_n^{(2)}) = 0.$$

Step 2 - Hellinger bound and Taylor remainders Equipped with the above results, we have

$$\begin{aligned}
\frac{h^2\left(g(x,\delta_n^{(1)}), g(x,\delta_n^{(2)})\right)}{\left|\delta_n^{(1)} - \delta_n^{(2)}\right|^{2r}} &= \int \frac{\left(g(x,\delta_n^{(1)}) - g(x,\delta_n^{(2)})\right)^2}{2^r \left|\delta_n^{(2)}\right|^{2r} \left(\sqrt{g(x,\delta_n^{(1)})} + \sqrt{g(x,\delta_n^{(2)})}\right)^2} dx \\
&= \int \frac{(R(x))^2}{2^r \left|\delta_n^{(2)}\right|^{2r} \left(\sqrt{g(x,\delta_n^{(1)})} + \sqrt{g(x,\delta_n^{(2)})}\right)^2} dx.
\end{aligned}$$

To validate that the above term goes to 0, we will need to investigate the concrete formulation of $R(x)$. In particular, the formulation of $R(x)$ is

$$R(x) = \pi R_1(x) + (1-\pi)\sum_{\alpha=1}^{2} \frac{c^\alpha(\delta_n^{(1)} - \delta_n^{(2)})^\alpha}{\alpha!} R_{2,\alpha}(x) + (1-\pi)R_2(x),$$

where the formulations of Taylor remainders $R_1(x)$, $R_{2,\alpha}(x)$, and $R_2(x)$ are as follows

$$R_1(x) = \frac{3\left(\delta_n^{(2)} - \delta_n^{(1)}\right)^3}{3!} \int_0^1 (1-t)^2 \frac{\partial^3 \phi}{\partial \delta^3}\left(x, -\delta_n^{(2)} + t\left(\delta_n^{(2)} - \delta_n^{(1)}\right)\right) dt,$$

$$R_2(x) = \frac{3c^3\left(\delta_n^{(1)} - \delta_n^{(2)}\right)^3}{3!} \int_0^1 (1-t)^2 \frac{\partial^3 \phi}{\partial \delta^3}\left(x, c\delta_n^{(2)} + t\left(c\delta_n^{(1)} - c\delta_n^{(2)}\right)\right) dt,$$

$$R_{2,\alpha}(x) = \frac{(3-\alpha)(c+1)^{3-\alpha}\left(\delta_n^{(2)}\right)^{3-\alpha}}{(3-\alpha)!\alpha!} \int_0^1 (1-t)^{2-\alpha} \frac{\partial^3 \phi}{\partial \delta^3}\left(x, -\delta_n^{(2)} + t(c+1)\delta_n^{(2)}\right) dt$$

for any $1 \leq \alpha \leq 2$.

Step 3 - Taylor remainders control Now, Holder's inequality leads to

$$R_1^2(x) \leq \frac{\left(\delta_n^{(2)} - \delta_n^{(1)}\right)^6}{4} \int_0^1 (1-t)^4 \left(\frac{\partial^3 \phi}{\partial \delta^3}\left(x, -\delta_n^{(2)} + t\left(\delta_n^{(2)} - \delta_n^{(1)}\right)\right)\right)^2 dt.$$

Due to the formulation of location Gaussian kernel with variance 1, we can check that

$$\sup_{t\in[0,1]} \int \frac{\left(\frac{\partial^3 \phi}{\partial \delta^3}\left(x, -\delta_n^{(2)} + t\left(\delta_n^{(2)} - \delta_n^{(1)}\right)\right)\right)^2}{\phi(x, -\delta_n^{(2)})} dx < \infty.$$

Equipped with the above results, the following holds

$$\begin{aligned}\int \frac{R_1^2(x)}{2^{r-1}\left|\delta_n^{(2)}\right|^{2r}\left(\sqrt{g(x,\delta_n^{(1)})} + \sqrt{g(x,\delta_n^{(2)})}\right)^2} dx &\leq \int \frac{R_1^2(x)}{2^{r-1}\left|\delta_n^{(2)}\right|^{2r}\pi\phi(x,-\delta_n^{(2)})} dx \\ &\lesssim \left|\delta_n^{(2)}\right|^{6-2r} \to 0 \end{aligned} \tag{G.7}$$

as $n \to \infty$ where the first inequality is due to the inequality $\left(\sqrt{g(x,\delta_n^{(1)})}+\sqrt{g(x,\delta_n^{(2)})}\right)^2 \geq \pi\phi(x,-\delta_n^{(2)})$. By means of the similar argument, we also obtain that

$$\begin{aligned}
\int \frac{R_2^2(x)}{2^{r-1}\left|\delta_n^{(2)}\right|^{2r}\left(\sqrt{g(x,\delta_n^{(1)})}+\sqrt{g(x,\delta_n^{(2)})}\right)^2}dx &\leq \int \frac{R_2^2(x)}{2^{r-1}\left|\delta_n^{(2)}\right|^{2r}(1-\pi)\phi(x,c\delta_n^{(2)})}dx \\
&\lesssim \left|\delta_n^{(2)}\right|^{6-2r} \to 0, \\
\int \frac{\left(\delta_n^{(1)}-\delta_n^{(2)}\right)^{\alpha} R_{2,\alpha}^2(x)}{2^{r-1}\left|\delta_n^{(2)}\right|^{2r}\left(\sqrt{g(x,\delta_n^{(1)})}+\sqrt{g(x,\delta_n^{(2)})}\right)^2}dx &\leq \int \frac{\left(\delta_n^{(1)}-\delta_n^{(2)}\right)^{\alpha} R_{2,\alpha}^2(x)}{2^{r-1}\left|\delta_n^{(2)}\right|^{2r}\pi\phi(x,-\delta_n^{(2)})}dx \\
&\lesssim \left|\delta_n^{(2)}\right|^{6-2r} \to 0. \qquad \text{(G.8)}
\end{aligned}$$

Invoking Cauchy-Schwarz's inequality, the following inequality holds

$$R^2(x) \leq 3\left((\pi R_1(x))^2+\left((1-\pi)\sum_{\alpha=1}^{2}\frac{c^{\alpha}(\delta_n^{(1)}-\delta_n^{(2)})^{\alpha}}{\alpha!}R_{2,\alpha}(x)\right)^2+((1-\pi)R_2(x))^2\right). \qquad \text{(G.9)}$$

Combining the results from (G.7), (G.8), and (G.9), we achieve that

$$\int R^2(x)/\left(2^{r-1}\left|\delta_n^{(2)}\right|^{2r}\left(\sqrt{g(x,\delta_n^{(1)})}+\sqrt{g(x,\delta_n^{(2)})}\right)^2\right)dx \to 0.$$

As a consequence, we achieve the conclusion with the lower bound of part (a) of the theorem.

(b) Similar to the proof argument of part (a), to achieve the conclusion of the lower bound of part (b), it is sufficient to demonstrate that

$$\inf_{\delta^{(1)},\delta^{(2)}\in\Theta_{2,n}} h\left(g(x,\delta^{(1)}),g(x,\delta^{(2)})\right)/\left|\left|\delta^{(1)}\right|-\left|\delta^{(2)}\right|\right|^r = 0 \qquad \text{(G.10)}$$

for any $1 \leq r < 2$. In particular, we choose two sequences $\left\{\overline{\delta}_n^{(1)}\right\}$ and $\left\{\overline{\delta}_n^{(2)}\right\}$ such that $\overline{\delta}_n^{(1)} = 2\overline{\delta}_n^{(2)}$ for all $n \geq 1$ and $\overline{\delta}_n^{(1)} \to 0$ as $n \to \infty$. For any $r < 2$, invoking Taylor expansion up to the first order as that of Case b.1 in the proof of Theorem 2.1, we have

$$g(x,\overline{\delta}_n^{(1)}) - g(x,\overline{\delta}_n^{(2)}) = \overline{R}(x),$$

where the formulation of $\overline{R}(x)$ is

$$\overline{R}(x) = \frac{1}{2}\overline{R}_1(x) + \frac{1}{2}\left(\overline{\delta}_n^{(1)} - \overline{\delta}_n^{(2)}\right)\overline{R}_{2,1}(x) + \frac{1}{2}\overline{R}_2(x).$$

Here, the detail formulations of Taylor remainders $\overline{R}_1(x)$, $\overline{R}_{2,1}(x)$, and $\overline{R}_2(x)$ are

$$\overline{R}_1(x) = \frac{2\left(\overline{\delta}_n^{(2)} - \overline{\delta}_n^{(1)}\right)^2}{2!}\int_0^1 (1-t)\frac{\partial^2\phi}{\partial\delta^2}\left(x, -\overline{\delta}_n^{(2)} + t\left(\overline{\delta}_n^{(2)} - \overline{\delta}_n^{(1)}\right)\right)dt,$$

$$\overline{R}_2(x) = \frac{2\left(\overline{\delta}_n^{(1)} - \overline{\delta}_n^{(2)}\right)^2}{2!}\int_0^1 (1-t)\frac{\partial^2\phi}{\partial\delta^2}\left(x, \overline{\delta}_n^{(2)} + t\left(\overline{\delta}_n^{(1)} - \overline{\delta}_n^{(2)}\right)\right)dt,$$

$$\overline{R}_{2,1}(x) = 2\overline{\delta}_n^{(2)}\int_0^1 \frac{\partial^2\phi}{\partial\delta^2}\left(x, -\overline{\delta}_n^{(2)} + 2t\overline{\delta}_n^{(2)}\right)dt.$$

With the choice that $\overline{\delta}_n^{(1)} = 2\overline{\delta}_n^{(2)} \to 0$ and the same argument as Step 3 in part (a), we can argue that

$$\int \overline{R}^2(x) \Big/ \left(2^{r-1}\left|\overline{\delta}_n^{(2)}\right|^{2r}\left(\sqrt{g(x,\overline{\delta}_n^{(1)})} + \sqrt{g(x,\overline{\delta}_n^{(2)})}\right)^2\right) \to 0$$

as $n \to \infty$. Therefore, for any $1 \leq r < 2$, we achieve

$$h\left(g(x,\overline{\delta}_n^{(1)}), g(x,\overline{\delta}_n^{(2)})\right) \Big/ \left|\left|\overline{\delta}_n^{(1)}\right| - \left|\overline{\delta}_n^{(2)}\right|\right|^r \to 0.$$

As a consequence, we achieve the conclusion of part (b) of the theorem.

G.2 PROOF OF THEOREM 2.2

For the sake of presentation, we denote $v := \sigma^2$ and $g(x,\delta,v) := \pi f(x,-\delta,v) + (1-\pi)f(x,c\delta,v)$ for all $\delta \in \Theta, \sigma \in \Omega$ where $f(x,\delta,v)$ is the density of location-scale Gaussian distribution with location δ and scale v. For the simplicity of the proof argument, we only focus on the proof for the upper bounds of the theorem. The proof for the lower bounds can be argued similarly as that of the lower bounds in Theorem 2.1 in Section G.1.2.

(a) By means of the proof argument with the upper bound of Theorem 2.1, in order to achieve the upper bound of part (a), it is sufficient to demonstrate that

$$\inf_{\substack{\delta^{(1)},\delta^{(2)}\in\Theta \\ v^{(1)},v^{(2)}\in\Omega}} \frac{V\left(g(x,\delta^{(1)},v^{(1)}), g(x,\delta^{(2)},v^{(2)})\right)}{|\delta^{(1)} - \delta^{(2)}|^3 + |v^{(1)} - v^{(2)}|^{3/2}} > 0, \tag{G.11}$$

where $\Theta = [-1, 1]$ and Ω is a bounded set containing $\overline{\sigma}$. Assume that the above inequality does not hold. It implies that we can find sequences $\left\{\delta_n^{(1)}\right\}$, $\left\{\delta_n^{(2)}\right\}$, $\left\{v_n^{(1)}\right\}$, and $\left\{v_n^{(2)}\right\}$ such that

$$\frac{V\left(g(x, \delta_n^{(1)}, v_n^{(1)}), g(x, \delta_n^{(2)}, v_n^{(2)})\right)}{|\delta_n^{(1)} - \delta_n^{(2)}|^3 + |v_n^{(1)} - v_n^{(2)}|^{3/2}} \to 0$$

as $n \to \infty$. To simplify the presentation, we only consider the most challenging setting $\delta_n^{(1)} \to 0, \delta_n^{(2)} \to 0$, $v_n^{(1)} \to v_0$, $v_n^{(2)} \to v_0$ for some $v_0 \in \Omega$. Additionally, we denote

$$D_n = |\delta_n^{(1)} - \delta_n^{(2)}|^3 + |v_n^{(1)} - v_n^{(2)}|^{3/2}.$$

Now, we consider the following settings with $\delta_n^{(1)}$ and $\delta_n^{(2)}$.

Case a.1: $\delta_n^{(1)}/\delta_n^{(2)} \not\to 1$ as $n \to \infty$. Similar to the structure of the proof of Theorem 2.1, we also divide the proof argument of this case into two key steps.

Step 1 - Taylor expansion Under this setting, by means of Taylor expansion up to the third order, we obtain that

$$\begin{aligned}
&\frac{g(x, \delta_n^{(1)}, v_n^{(1)}) - g(x, \delta_n^{(2)}, v_n^{(2)})}{D_n} \qquad \text{(G.12)}\\
&= \frac{\pi\Big(f(x, -\delta_n^{(1)}, v_n^{(1)}) - f(x, -\delta_n^{(2)}, v_n^{(2)})\Big) + (1-\pi)\Big(f(x, c\delta_n^{(1)}, v_n^{(1)}) - f(x, c\delta_n^{(2)}, v_n^{(2)})\Big)}{D_n}\\
&= \frac{\pi\Big(\sum\limits_{|\alpha|\leq 3} \frac{(\delta_n^{(2)} - \delta_n^{(1)})^{\alpha_1}(v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\alpha_1!\alpha_2!} \frac{\partial^{|\alpha|} f}{\partial \delta^{\alpha_1} \partial v^{\alpha_2}}(x, -\delta_n^{(2)}, v_n^{(2)}) + R_1(x)\Big)}{D_n}\\
&+ \frac{(1-\pi)\Big(\sum\limits_{|\alpha|\leq 3} \frac{c^{\alpha_1}(\delta_n^{(1)} - \delta_n^{(2)})^{\alpha_1}(v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\alpha_1!\alpha_2!} \frac{\partial^{|\alpha|} f}{\partial \delta^{\alpha_1} \partial v^{\alpha_2}}(x, c\delta_n^{(2)}, v_n^{(2)}) + R_2(x)\Big)}{D_n}\\
&= \frac{\pi\Big(\sum\limits_{|\alpha|\leq 3} \frac{1}{2^{\alpha_2}} \frac{(\delta_n^{(2)} - \delta_n^{(1)})^{\alpha_1}(v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\alpha_1!\alpha_2!} \frac{\partial^{\alpha_1+2\alpha_2} f}{\partial \delta^{\alpha_1+2\alpha_2}}(x, -\delta_n^{(2)}, v_n^{(2)}) + R_1(x)\Big)}{D_n}\\
&+ \frac{(1-\pi)\Big(\sum\limits_{|\alpha|\leq 3} \frac{1}{2^{\alpha_2}} \frac{c^{\alpha_1}(\delta_n^{(1)} - \delta_n^{(2)})^{\alpha_1}(v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\alpha_1!\alpha_2!} \frac{\partial^{\alpha_1+2\alpha_2} f}{\partial \delta^{\alpha_1+2\alpha_2}}(x, c\delta_n^{(2)}, v_n^{(2)}) + R_2(x)\Big)}{D_n},
\end{aligned}$$

where the last equality is due to the PDE structure of location-scale Gaussian distribution, which is given by

$$\frac{\partial^2 f}{\partial \delta^2}(x, \delta, \sigma) = 2\frac{\partial f}{\partial \sigma^2}(x, \delta, \sigma).$$

Additionally, $R_1(x)$ and $R_2(x)$ are Taylor remainders that satisfy the following inequality

$$\max\{\|R_1(x)\|_\infty, \|R_2(x)\|_\infty\} = O\left(|\delta_n^{(1)} - \delta_n^{(2)}|^{3+\gamma} + |v_n^{(1)} - v_n^{(2)}|^{3+\gamma}\right)$$

for some $\gamma > 0$. It implies that $R_1(x)/D_n \to 0$ and $R_2(x)/D_n \to 0$ for all x as $n \to \infty$. Now, by means of Taylor expansion up to the third order, we further have

$$\frac{\partial^{\alpha_1+2\alpha_2} f}{\partial \delta^{\alpha_1+2\alpha_2}}(x, c\delta_n^{(2)}, v_n^{(2)}) = \sum_{\tau=0}^{3-|\alpha|} \frac{(c+1)^\tau (\delta_n^{(2)})^\tau}{\tau!} \frac{\partial^{\alpha_1+2\alpha_2+\tau} f}{\partial \delta^{\alpha_1+2\alpha_2+\tau}}(x, -\delta_n^{(2)}, v_n^{(2)}) + R_{2,\alpha}(x) \quad \text{(G.13)}$$

for each $\alpha = (\alpha_1, \alpha_2)$ such that $1 \leq |\alpha| \leq 3$. Here, $R_{2,\alpha}(x)$ is a Taylor remainder that satisfies $\|R_{2,\alpha}(x)\|_\infty = O\left(|\delta_n^{(2)}|^{3-|\alpha|+\gamma}\right)$ for all α. By plugging equations (G.13) into (G.12), the following holds

$$\begin{aligned}
&\frac{g(x, \delta_n^{(1)}, v_n^{(1)}) - g(x, \delta_n^{(2)}, v_n^{(2)})}{D_n} \\
&= \frac{\pi\left(\sum_{|\alpha|\leq 3} \frac{1}{2^{\alpha_2}} \frac{(\delta_n^{(2)} - \delta_n^{(1)})^{\alpha_1} (v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\alpha_1! \alpha_2!} \frac{\partial^{\alpha_1+2\alpha_2} f}{\partial \delta^{\alpha_1+2\alpha_2}}(x, -\delta_n^{(2)}, v_n^{(2)})\right)}{D_n} \\
&+ \frac{(1-\pi)\left(\sum_{|\alpha|\leq 3} \sum_{\tau=0}^{3-|\alpha|} \frac{1}{2^{\alpha_2}} \frac{c^{\alpha_1}(c+1)^\tau (\delta_n^{(2)})^\tau (\delta_n^{(1)} - \delta_n^{(2)})^{\alpha_1} (v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\tau! \alpha_1! \alpha_2!} \frac{\partial^{\alpha_1+2\alpha_2} f}{\partial \delta^{\alpha_1+2\alpha_2}}(x, -\delta_n^{(2)}, v_n^{(2)})\right)}{D_n} \\
&+ \frac{\pi R_1(x) + (1-\pi) R_2(x) + \sum_{|\alpha|\leq 3} \frac{1}{2^{\alpha_2}} \frac{c^{\alpha_1} (\delta_n^{(1)} - \delta_n^{(2)})^{\alpha_1} (v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\alpha_1! \alpha_2!} R_{2,\alpha}(x)}{D_n} \\
&= \frac{\sum_{l=1}^{6} A_{n,l} \frac{\partial^l f}{\partial \delta^l}(x, -\delta_n^{(2)}, v_n^{(2)}) + R(x)}{D_n},
\end{aligned}$$

where the detail formulations of $A_{n,l}$ and $R(x)$ are as follows

$$\begin{aligned} A_{n,l} &= \pi \sum_{\alpha_1,\alpha_2} \frac{1}{2^{\alpha_2}} \frac{(\delta_n^{(2)} - \delta_n^{(1)})^{\alpha_1} (v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\alpha_1! \alpha_2!} \\ &\quad + (1-\pi) \sum_{\alpha_1,\alpha_2,\tau} \frac{1}{2^{\alpha_2}} \frac{c^{\alpha_1}(c+1)^{\tau} (\delta_n^{(2)})^{\tau} (\delta_n^{(1)} - \delta_n^{(2)})^{\alpha_1} (v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\tau! \alpha_1! \alpha_2!}, \\ R(x) &= \pi R_1(x) + (1-\pi) R_2(x) + \sum_{|\alpha| \leq 3} \frac{1}{2^{\alpha_2}} \frac{c^{\alpha_1} (\delta_n^{(1)} - \delta_n^{(2)})^{\alpha_1} (v_n^{(1)} - v_n^{(2)})^{\alpha_2}}{\alpha_1! \alpha_2!} R_{2,\alpha}(x) \end{aligned}$$

for any $1 \leq l \leq 6$ and $x \in \mathbb{R}$. Here, the ranges of α_1, α_2 in the first sum of $A_{n,l}$ satisfy $\alpha_1 + 2\alpha_2 = l$, $1 \leq |\alpha| \leq 3$ while the ranges of α_1, α_2, τ in the second sum of $A_{n,l}$ satisfy $\alpha_1 + 2\alpha_2 + \tau = l$, $0 \leq \tau \leq 3 - |\alpha|$, and $1 \leq |\alpha| \leq 3$. According to the hypothesis $\delta_n^{(1)}/\delta_n^{(2)} \not\to 1$, we have

$$|\delta_n^{(2)}|/|\delta_n^{(1)} - \delta_n^{(2)}| \not\to \infty.$$

Therefore, we have

$$\frac{|\delta_n^{(1)} - \delta_n^{(2)}|^{\alpha_1} |v_n^{(1)} - v_n^{(2)}|^{\alpha_2} \|R_{2,\alpha}(x)\|_{\infty}}{D_n} = \frac{O\left(|\delta_n^{(1)} - \delta_n^{(2)}|^{\alpha_1} |v_n^{(1)} - v_n^{(2)}|^{\alpha_2} |\delta_n^{(2)}|^{3-|\alpha|+\gamma}\right)}{D_n} \to 0.$$

As a consequence, we have $\|R(x)\|_{\infty}/D_n \to 0$ as $n \to \infty$.

Step 2 - Non-vanishing coefficients and Fatou's argument Assume that all the coefficients $A_{n,l}/D_n \to 0$ for all $1 \leq l \leq 6$ as $n \to \infty$. We denote the following key term

$$\overline{M}_n := \max\left\{|\delta_n^{(1)} - \delta_n^{(2)}|, |v_n^{(1)} - v_n^{(2)}|^{1/2}\right\}.$$

As $|\delta_n^{(2)}|/|\delta_n^{(1)} - \delta_n^{(2)}| \not\to \infty$, we also have $|\delta_n^{(2)}|/M_n \not\to \infty$. Now, we denote $\delta_n^{(2)}/M_n \to x$, $(\delta_n^{(2)} - \delta_n^{(1)})/M_n \to y$, and $(v_n^{(1)} - v_n^{(2)})/M_n^2 \to z$ as $n \to \infty$. From the definition of $\overline{M}_n$, at least one among y and z is different from 0. By dividing both the numerator and the denominator of $A_{n,l}/D_n$ by $\overline{M}_n^l$ as $1 \leq l \leq 3$, as $n \to \infty$, we have the following system of polynomial equations

$$\begin{aligned} cy^2 + z - 2cxy &= 0, \\ \frac{\pi(1-2\pi)}{3!(1-\pi)^2} y^3 + \frac{1}{2} xz + \frac{c^2}{2} xy^2 - \frac{\pi}{2(1-\pi)^2} x^2 y &= 0. \end{aligned}$$

The above system of polynomial equations leads to $\pi(1-2\pi)y(y^2 - 3xy + 3x^2) = 0$, which only holds when $y = 0$. Therefore, it leads to $z = 0$, which is a contradiction. It implies that

not all the coefficients $A_{n,l}/D_n \to 0$ as $n \to \infty$. Denote $m_n = D_n / \max_{1\le l\le 6} |A_{n,l}|$. According to the previous result, we have $m_n \not\to \infty$. Now, we have that

$$m_n \frac{g(x, \delta_n^{(1)}, v_n^{(1)}) - g(x, \delta_n^{(2)}, v_n^{(2)})}{D_n} \to \sum_{l=1}^{6} \tau_l \frac{\partial^l f}{\partial \delta^l}(x, 0, v_0)$$

for some coefficients τ_l such that not all of them are 0. Similar to the proof argument of Theorem 2.1, by invoking Fatou's lemma with $V\left(g(x, \delta_n^{(1)}, v_n^{(1)}), g(x, \delta_n^{(2)}, v_n^{(2)})\right)/D_n \to 0$, the following equation holds

$$\sum_{l=1}^{6} \tau_l \frac{\partial^l f}{\partial \delta^l}(x, 0, v_0) = 0$$

for almost surely x. However, due to the linear independence of $\left\{\frac{\partial^l f}{\partial \delta^l}(x, 0, v_0)\right\}$, we have $\tau_l = 0$ for all $1 \le l \le 6$, which is a contradiction. Therefore, Case a.1 does not hold.

Case a.2: $\delta_n^{(1)}/\delta_n^{(2)} \to 1$ as $n \to \infty$. It implies that $|\delta_n^{(2)}|/|\delta_n^{(1)} - \delta_n^{(2)}| \to \infty$. Similar to Case a.2 in the proof of Theorem 2.1, the main challenge with that setting is that $R(x)/D_n$ does not converge to 0; therefore, we cannot hinge upon the previous argument in Case a.1 to argue the contradiction with this case. To be able to deal with that problem, we will demonstrate two key properties under that setting: $\max_{1\le l\le 6} \{|A_{n,l}|\}/D_n \not\to 0$ and $\|R(x)\|_\infty / \max_{1\le l\le 6} |A_{n,l}| \to 0$. Indeed, we have the following possibilities regarding $\delta_n^{(1)}, \delta_n^{(2)}, v_n^{(1)}$, and $v_n^{(2)}$.

Case a.2.1: $|v_n^{(1)} - v_n^{(2)}| / \left\{|\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(1)} + \delta_n^{(2)}|\right\} \to \infty$. Assume by the contrary that the following term $\max_{1\le l\le 6} \{|A_{n,l}|\}/D_n \to 0$. From the formulation of $A_{n,2}$, we have

$$|A_{n,2}| = \frac{1}{2}\left|(v_n^{(1)} - v_n^{(2)}) - c(\delta_n^{(2)} - \delta_n^{(1)})(\delta_n^{(2)} + \delta_n^{(1)})\right| \gtrsim |v_n^{(1)} - v_n^{(2)}|,$$

as n is sufficiently large due to the assumption of Case a.2.1. Since $A_{n,2}/D_n \to 0$, it implies that $(v_n^{(1)} - v_n^{(2)})/D_n \to 0$. Therefore, it leads to $(\delta_n^{(1)} - \delta_n^{(2)})(\delta_n^{(2)} + \delta_n^{(1)})/D_n \to 0$. As $|\delta_n^{(2)}|/|\delta_n^{(1)} - \delta_n^{(2)}| \to \infty$, the previous limit implies that $|\delta_n^{(1)} - \delta_n^{(2)}|^2/D_n \to 0$. These results mean that

$$1 = \frac{|v_n^{(1)} - v_n^{(2)}|^{3/2} + |\delta_n^{(1)} - \delta_n^{(2)}|^3}{D_n} \to 0,$$

which is a contradiction. Therefore, we have $\max_{1\le l\le 6}\{|A_{n,l}|\}/D_n \not\to 0$. Now, for any $1 \le |\alpha| \le 3$, as n is sufficiently large, we have

$$\frac{|\delta_n^{(1)}-\delta_n^{(2)}|^{\alpha_1}|v_n^{(1)}-v_n^{(2)}|^{\alpha_2}\|R_{2,\alpha}(x)\|_\infty}{\max_{1\le l\le 6}\{|A_{n,l}|\}} \le \frac{O(|\delta_n^{(1)}-\delta_n^{(2)}|^{\alpha_1}|v_n^{(1)}-v_n^{(2)}|^{\alpha_2}|\delta_n^{(2)}|^{3-|\alpha|+\gamma})}{|v_n^{(1)}-v_n^{(2)}|} \to 0.$$

Hence, we achieve that $\|R(x)\|_\infty/\max_{1\le l\le 6}\{|A_{n,l}|\} \to 0$ for all $x \in \mathbb{R}$.

Case a.2.2: $|v_n^{(1)}-v_n^{(2)}|/\left\{|\delta_n^{(1)}-\delta_n^{(2)}||\delta_n^{(1)}+\delta_n^{(2)}|\right\} \to \bar{c} \neq c$. Under that assumption, we have

$$|A_{n,2}| = \frac{1}{2}\left|(v_n^{(1)}-v_n^{(2)}) - c(\delta_n^{(2)}-\delta_n^{(1)})(\delta_n^{(2)}+\delta_n^{(1)})\right| \gtrsim |\delta_n^{(1)}-\delta_n^{(2)}||\delta_n^{(1)}+\delta_n^{(2)}|$$

when n is sufficiently large. If we have $\max_{1\le l\le 6}\{|A_{n,l}|\}/D_n \to 0$, then $|A_{n,2}|/D_n$ leads to both $(v_n^{(1)}-v_n^{(2)})/D_n \to 0$ and $\left(\delta_n^{(1)}-\delta_n^{(2)}\right)\left(\delta_n^{(1)}+\delta_n^{(2)}\right)/D_n \to 0$, which does not hold according to the argument of Case a.2.1. Therefore, $\max_{1\le l\le 6}\{|A_{n,l}|\}/D_n \not\to 0$. On the other hand, for any $1 \le |\alpha| \le 3$, as n is sufficiently large, we have

$$\begin{aligned}\frac{|\delta_n^{(1)}-\delta_n^{(2)}|^{\alpha_1}|v_n^{(1)}-v_n^{(2)}|^{\alpha_2}\|R_{2,\alpha}(x)\|_\infty}{\max_{1\le l\le 6}\{|A_{n,l}|\}} &\le \frac{O\left(|\delta_n^{(1)}-\delta_n^{(2)}|^{\alpha_1}|v_n^{(1)}-v_n^{(2)}|^{\alpha_2}|\delta_n^{(2)}|^{3-|\alpha|+\gamma}\right)}{|\delta_n^{(1)}-\delta_n^{(2)}||\delta_n^{(1)}+\delta_n^{(2)}|} \\ &= \frac{O\left(|\delta_n^{(1)}-\delta_n^{(2)}|^{|\alpha|}|\delta_n^{(2)}|^{3-\alpha_1+\gamma}\right)}{|\delta_n^{(1)}-\delta_n^{(2)}||\delta_n^{(1)}+\delta_n^{(2)}|} \to 0.\end{aligned}$$

Hence, we achieve that $\|R(x)\|_\infty/\max_{1\le l\le 6}\{|A_{n,l}|\} \to 0$ for all $x \in \mathbb{R}$.

Case a.2.3: $|v_n^{(1)}-v_n^{(2)}|/\left\{|\delta_n^{(1)}-\delta_n^{(2)}||\delta_n^{(1)}+\delta_n^{(2)}|\right\} \to c$. Without loss of generality, we assume that $(v_n^{(1)}-v_n^{(2)})/(\delta_n^{(1)}-\delta_n^{(2)})(\delta_n^{(1)}+\delta_n^{(2)}) \to c$ as the argument when this ratio goes to $-c$ is similar. Under this assumption, we have

$$\frac{|A_{n,3}|}{|\delta_n^{(1)}-\delta_n^{(2)}||\delta_n^{(1)}+\delta_n^{(2)}||\delta_n^{(2)}|} \to \left|\frac{c}{2} - \frac{(1-\pi)c(c+1)^2}{4}\right| > 0.$$

Therefore, as n is sufficiently large, we have $|A_{n,3}| \gtrsim |\delta_n^{(1)}-\delta_n^{(2)}||\delta_n^{(1)}+\delta_n^{(2)}||\delta_n^{(2)}|$. If we have $\max_{1\le l\le 6}\{|A_{n,l}|\}/D_n \to 0$, then $|A_{n,3}|/D_n \to 0$ leads to $|\delta_n^{(1)}-\delta_n^{(2)}||\delta_n^{(1)}+\delta_n^{(2)}||\delta_n^{(2)}|/D_n \to 0$.

Therefore, the following holds

$$|v_n^{(1)} - v_n^{(2)}|^{3/2} / \left\{ |\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(1)} + \delta_n^{(2)}||\delta_n^{(2)}| \right\} \to \infty,$$

which means $|v_n^{(1)} - v_n^{(2)}|/|\delta_n^{(2)}|^2 \to \infty$ — a contradiction to the assumption of Case a.2.3. Hence, $\max\limits_{1\leq l\leq 6} \{|A_{n,l}|\} / D_n \not\to 0$. On the other hand, for any $1 \leq |\alpha| \leq 3$, as n is sufficiently large, we have

$$\begin{aligned} \frac{|\delta_n^{(1)} - \delta_n^{(2)}|^{\alpha_1} |v_n^{(1)} - v_n^{(2)}|^{\alpha_2} \|R_{2,\alpha}(x)\|_\infty}{\max\limits_{1\leq l\leq 6} \{|A_{n,l}|\}} &\leq \frac{O\left(|\delta_n^{(1)} - \delta_n^{(2)}|^{\alpha_1} |v_n^{(1)} - v_n^{(2)}|^{\alpha_2} |\delta_n^{(2)}|^{3-|\alpha|+\gamma}\right)}{|\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(1)} + \delta_n^{(2)}||\delta_n^{(2)}|} \\ &= \frac{O\left(|\delta_n^{(1)} - \delta_n^{(2)}|^{|\alpha|} |\delta_n^{(2)}|^{3-\alpha_1+\gamma}\right)}{|\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(1)} + \delta_n^{(2)}||\delta_n^{(2)}|} \to 0. \end{aligned}$$

Thus, we obtain that $\|R(x)\|_\infty / \max\limits_{1\leq l\leq 6} \{|A_{n,l}|\} \to 0$ for all $x \in \mathbb{R}$.

Governed by the results from Case a.2.1, Case a.2.2, and Case a.2.3, we finally achieve that $\max\limits_{1\leq l\leq 6} \{|A_{n,l}|\} / D_n \not\to 0$ and $\|R(x)\|_\infty / \max\limits_{1\leq l\leq 6} |A_{n,l}| \to 0$. Denote $m_n' = D_n / \max\limits_{1\leq l\leq 6} \{|A_{n,l}|\}$. Then, we will have $m_n' \not\to \infty$. Thus, the following limit holds

$$m_n' \frac{g(x, \delta_n^{(1)}, v_n^{(1)}) - g(x, \delta_n^{(2)}, v_n^{(2)})}{D_n} \to \sum_{l=1}^{6} \tau_l' \frac{\partial^l f}{\partial \delta^l}(x, 0, v_0),$$

for some coefficients τ_l' such that not all of them are 0. By means of Fatou's lemma with the ratio $V\left(g(x, \delta_n^{(1)}, v_n^{(1)}), g(x, \delta_n^{(2)}, v_n^{(2)})\right) / D_n \to 0$, we obtain that

$$\sum_{l=1}^{6} \tau_l' \frac{\partial^l f}{\partial \delta^l}(x, 0, v_0) = 0.$$

However, due to the linear independence of $\left\{ \dfrac{\partial^l f}{\partial \delta^l}(x, 0, v_0) \right\}$, we will have $\tau_l' = 0$ for all $1 \leq l \leq 6$, which is a contradiction. Therefore, Case a.2 does not hold. As a consequence, we achieve the conclusion with the upper bound of part (a) of the theorem.

(b) Similar to the proof argument of part (a), it is sufficient to demonstrate that

$$\inf_{\substack{\delta^{(1)}, \delta^{(2)} \in \Theta \\ v^{(1)}, v^{(2)} \in \Omega}} \frac{V\left(g(x, \delta^{(1)}, v^{(1)}), g(x, \delta^{(2)}, v^{(2)})\right)}{\left||\delta^{(1)}| - |\delta^{(2)}|\right|^4 + |v^{(1)} - v^{(2)}|^2} > 0,$$

where $\Theta = [-1, 1]$ and Ω is a bounded set containing $\overline{\sigma}$. Assume that the above inequality

does not hold. It implies that we can find sequences $\left\{\delta_n^{(1)}\right\}$, $\left\{\delta_n^{(2)}\right\}$, $\left\{v_n^{(1)}\right\}$, and $\left\{v_n^{(2)}\right\}$ such that

$$\frac{V\left(g(x,\delta_n^{(1)},v_n^{(1)}),g(x,\delta_n^{(2)},v_n^{(2)})\right)}{\left|\left|\delta_n^{(1)}\right|-\left|\delta_n^{(2)}\right|\right|^4+|v_n^{(1)}-v_n^{(2)}|^2}\to 0$$

as $n\to\infty$. Similar the proof argument of part (a), we only consider the most challenging setting $\delta_n^{(1)}\to 0, \delta_n^{(2)}\to 0$, $v_n^{(1)}\to v_0$, $v_n^{(2)}\to v_0$ for some $v_0\in\Omega$. For the convenience of presentation, we denote

$$\overline{D}_n=\left|\left|\delta_n^{(1)}\right|-\left|\delta_n^{(2)}\right|\right|^4+|v_n^{(1)}-v_n^{(2)}|^2.$$

Now, we have three settings with $\delta_n^{(1)}$ and $\delta_n^{(2)}$ in the proof of part (b).

Case b.1: $\delta_n^{(1)}/\delta_n^{(2)}\not\to 1$ as $n\to\infty$ and $\delta_n^{(1)}\delta_n^{(2)}\geq 0$ for all n. Under this case, we have

$$\overline{D}_n=|\delta_n^{(1)}-\delta_n^{(2)}|^4+|v_n^{(1)}-v_n^{(2)}|^2.$$

To facilitate the proof argument of this case, we also divide it into two key steps.

Step 1 - Taylor expansion Using the similar argument as that of part (a), by means of Taylor expansion up to the fourth order, we get the following representation

$$\frac{g(x,\delta_n^{(1)},v_n^{(1)})-g(x,\delta_n^{(2)},v_n^{(2)})}{\overline{D}_n}=\frac{\sum_{l=1}^{8}B_{n,l}\dfrac{\partial^l f}{\partial\delta^l}(x,-\delta_n^{(2)},v_n^{(2)})+\overline{R}(x)}{\overline{D}_n},$$

where the formulations of $B_{n,l}$ and $\overline{R}(x)$ are as follows

$$\begin{aligned}
B_{n,l} &= \frac{1}{2}\sum_{\alpha_1,\alpha_2}\frac{1}{2^{\alpha_2}}\frac{(\delta_n^{(2)}-\delta_n^{(1)})^{\alpha_1}(v_n^{(1)}-v_n^{(2)})^{\alpha_2}}{\alpha_1!\alpha_2!}\\
&\qquad +\frac{1}{2}\sum_{\alpha_1,\alpha_2,\tau}\frac{1}{2^{\alpha_2}}\frac{2^\tau(\delta_n^{(2)})^\tau(\delta_n^{(1)}-\delta_n^{(2)})^{\alpha_1}(v_n^{(1)}-v_n^{(2)})^{\alpha_2}}{\tau!\alpha_1!\alpha_2!},\\
\overline{R}(x) &= \frac{1}{2}\overline{R}_1(x)+\frac{1}{2}\overline{R}_2(x)+\sum_{|\alpha|\leq 4}\frac{1}{2^{\alpha_2}}\frac{(\delta_n^{(1)}-\delta_n^{(2)})^{\alpha_1}(v_n^{(1)}-v_n^{(2)})^{\alpha_2}}{\alpha_1!\alpha_2!}\overline{R}_{2,\alpha}(x).
\end{aligned}$$

Here, the ranges of α_1,α_2 in the first sum of $B_{n,l}$ satisfy $\alpha_1+2\alpha_2=l$, $1\leq|\alpha|\leq 4$ while the ranges of α_1,α_2,τ in the second sum of $B_{n,l}$ satisfy $\alpha_1+2\alpha_2+\tau=l$, $0\leq\tau\leq 4-|\alpha|$,

and $1 \leq |\alpha| \leq 4$. Additionally, $\overline{R}_1(x)$ is a Taylor remainder from expanding $f(x, -\delta_n^{(1)}, v_n^{(1)})$ around $f(x, -\delta_n^{(2)}, v_n^{(2)})$ up to the fourth order, $\overline{R}_2(x)$ is Taylor remainder from expanding $f(x, c\delta_n^{(1)}, v_n^{(1)})$ around $f(x, c\delta_n^{(2)}, v_n^{(2)})$ up to the fourth order, and $\overline{R}_{2,\alpha}(x)$ is Taylor remainder from expanding $\dfrac{\partial^{\alpha_1+2\alpha_2} f}{\partial \delta^{\alpha_1+2\alpha_2}}(x, c\delta_n^{(2)}, v_n^{(2)})$ around $\dfrac{\partial^{\alpha_1+2\alpha_2} f}{\partial \delta^{\alpha_1+2\alpha_2}}(x, -\delta_n^{(2)}, v_n^{(2)})$ up to the order $4-|\alpha|$. Similar to the argument of Case a.1, the assumption of Case b.1 is sufficient to guarantee that $\overline{R}(x)/\overline{D}_n \to 0$.

Step 2 - Non-vanishing coefficients and Fatou's argument Assume that all the coefficients $B_{n,l}/\overline{D}_n \to 0$ for all $1 \leq l \leq 8$ as $n \to \infty$. Remind from part (a) that we denote

$$\overline{M}_n := \max\left\{|\delta_n^{(1)} - \delta_n^{(2)}|, |v_n^{(1)} - v_n^{(2)}|^{1/2}\right\}.$$

Additionally, we also denote $\delta_n^{(2)}/\overline{M}_n \to x$, $(\delta_n^{(2)} - \delta_n^{(1)})/\overline{M}_n \to y$, and $(v_n^{(1)} - v_n^{(2)})/\overline{M}_n^2 \to z$ as $n \to \infty$ where at least one from y and z is different from 0. Due to the assumption that $\delta_n^{(1)}\delta_n^{(2)} \geq 0$, we have $x(x-y) \geq 0$. Now, by dividing both the numerator and the denominator of $B_{n,l}/D_n$ by $\overline{M}_n^l$ as $1 \leq l \leq 4$, as $n \to \infty$, we have the following system of polynomial equations

$$\begin{aligned} y^2 + z - 2xy &= 0, \\ \frac{y^4}{4!} + \frac{y^2 z}{4} + \frac{z^2}{8} - \frac{xyz}{2} + \frac{x^2 z}{2} - \frac{xy^3}{6} + \frac{x^2 y^2}{2} - \frac{2x^3 y}{3} &= 0. \end{aligned}$$

When $x = 0$, the above system of polynomial equations leads to $y = z = 0$, which is a contradiction with the assumption that at least one of y, z is different from 0. When $x \neq 0$, the above system of polynomial equations leads to $y^3 - 4xy^2 + 6x^2 y - 4x^3 = 0$, which leads to $y = 2x$ — a contradiction to the condition $x(x-y) \geq 0$ and $x \neq 0$. Therefore, not all of the coefficients $B_{n,l}/\overline{D}_n \to 0$ as $n \to \infty$. From here, using the same proof argument as that of Case a.1 in part (a), we achieve the conclusion that Case b.1 cannot hold.

Case b.2: $\delta_n^{(1)}/\delta_n^{(2)} \not\to 1$ as $n \to \infty$ and $\delta_n^{(1)}\delta_n^{(2)} < 0$ for all n. Under this case, we have

$$\overline{D}_n = |\delta_n^{(1)} + \delta_n^{(2)}|^4 + |v_n^{(1)} - v_n^{(2)}|^2.$$

By means of Taylor expansion up to the fourth order, we obtain the following representation

$$\begin{aligned}
&\frac{g(x,\delta_n^{(1)},v_n^{(1)})-g(x,\delta_n^{(2)},v_n^{(2)})}{\overline{D}_n}\\
&= \frac{\frac{1}{2}(f(x,-\delta_n^{(1)},v_n^{(1)})-f(x,\delta_n^{(2)},v_n^{(2)}))+\frac{1}{2}(f(x,\delta_n^{(1)},v_n^{(1)})-f(x,-\delta_n^{(2)},v_n^{(2)}))}{\overline{D}_n}\\
&= \frac{\sum_{\alpha=1}^{8} C_{n,l}\frac{\partial^l f}{\partial\delta^l}(x,\delta_n^{(2)},v_n^{(2)})+\widetilde{R}(x)}{\overline{D}_n},
\end{aligned}$$

where the formulations of $C_{n,l}$ and $\overline{R}_1(x)$ are as follows

$$\begin{aligned}
C_{n,l} &= \frac{1}{2}\sum_{\alpha_1,\alpha_2}\frac{1}{2^{\alpha_2}}\frac{(-\delta_n^{(2)}-\delta_n^{(1)})^{\alpha_1}(v_n^{(1)}-v_n^{(2)})^{\alpha_2}}{\alpha_1!\alpha_2!}\\
&\quad+\frac{1}{2}\sum_{\alpha_1,\alpha_2,\tau}\frac{1}{2^{\alpha_2}}\frac{2^{\tau}(-\delta_n^{(2)})^{\tau}(\delta_n^{(1)}+\delta_n^{(2)})^{\alpha_1}(v_n^{(1)}-v_n^{(2)})^{\alpha_2}}{\tau!\alpha_1!\alpha_2!},\\
\widetilde{R}(x) &= \frac{1}{2}\widetilde{R}_1(x)+\frac{1}{2}\widetilde{R}_2(x)+\sum_{|\alpha|\le 4}\frac{1}{2^{\alpha_2}}\frac{c^{\alpha_1}(\delta_n^{(1)}+\delta_n^{(2)})^{\alpha_1}(v_n^{(1)}-v_n^{(2)})^{\alpha_2}}{\alpha_1!\alpha_2!}\widetilde{R}_{2,\alpha}(x).
\end{aligned}$$

Here, the ranges of α_1,α_2 in the first sum of $C_{n,l}$ satisfy $\alpha_1+2\alpha_2=l$, $1\le|\alpha|\le 4$ while the ranges of α_1,α_2,τ in the second sum of $C_{n,l}$ satisfy $\alpha_1+2\alpha_2+\tau=l$, $0\le\tau\le 4-|\alpha|$, and $1\le|\alpha|\le 4$. Additionally, $\widetilde{R}_1(x)$ is a Taylor remainder from expanding $f(x,-\delta_n^{(1)},v_n^{(1)})$ around $f(x,\delta_n^{(2)},v_n^{(2)})$ up to the fourth order, $\widetilde{R}_2(x)$ is a Taylor remainder from expanding $f(x,\delta_n^{(1)},v_n^{(1)})$ around $f(x,-\delta_n^{(2)},v_n^{(2)})$ up to the fourth order, and $\widetilde{R}_{2,\alpha}(x)$ is a Taylor remainder from expanding $\frac{\partial^{\alpha_1+2\alpha_2}f}{\partial\delta^{\alpha_1+2\alpha_2}}(x,-\delta_n^{(2)},v_n^{(2)})$ around $\frac{\partial^{\alpha_1+2\alpha_2}f}{\partial\delta^{\alpha_1+2\alpha_2}}(x,\delta_n^{(2)},v_n^{(2)})$ up to the order $4-|\alpha|$. Due to the assumption of Case b.2, we can check that $\|\widetilde{R}(x)\|_\infty/\overline{D}_n\to 0$ as $n\to\infty$.

Assume that all the coefficients $C_{n,l}/\overline{D}_n\to 0$ for all $1\le l\le 8$ as $n\to\infty$. We denote

$$\widetilde{M}_n := \max\left\{|\delta_n^{(1)}+\delta_n^{(2)}|,|v_n^{(1)}-v_n^{(2)}|^{1/2}\right\}.$$

From the definition of $\widetilde{M}_n$, we can denote $\delta_n^{(2)}/\widetilde{M}_n\to x_1$, $(\delta_n^{(2)}+\delta_n^{(1)})/\widetilde{M}_n\to y_1$, and $(v_n^{(1)}-v_n^{(2)})/\widetilde{M}_n^2\to z_1$ as $n\to\infty$ where at least one from y_1 and z_1 is different from 0. Due to the assumption that $\delta_n^{(1)}\delta_n^{(2)}<0$, we have $x_1(y_1-x_1)\le 0$. Now, by dividing both the numerator and the denominator of $C_{n,l}/D_n$ by $\widetilde{M}_n^4$ as $1\le l\le 4$, as $n\to\infty$, we have the following

system of polynomial equations

$$y_1^2 + z_1 - 2x_1y_1 = 0,$$
$$\frac{y_1^4}{4!} + \frac{y_1^2 z_1}{4} + \frac{z_1^2}{8} - \frac{x_1 y_1 z_1}{2} + \frac{x_1^2 z_1}{2} - \frac{x_1 y_1^3}{6} + \frac{x_1^2 y_1^2}{2} - \frac{2x_1^3 y_1}{3} = 0.$$

If $x_1 = 0$, the above system leads to $y_1 = z_1 = 0$, which is a contradiction with the assumption of y_1, z_1. As $x_1 \neq 0$, the above system of polynomial equations leads to $y_1 = 2x_1$ — a contradiction to the condition $x_1(y_1 - x_1) \leq 0$ and $x_1 \neq 0$. Therefore, not all of the coefficients $C_{n,l}/\overline{D}_n \to 0$ as $n \to \infty$. From here, using the same proof argument as that of Case a.1 in part (a), we achieve the conclusion that Case b.2 cannot hold.

Case b.3: $\delta_n^{(1)}/\delta_n^{(2)} \to 1$ as $n \to \infty$. Under this assumption, we have $\delta_n^{(1)}\delta_n^{(2)} > 0$ as n is sufficiently large. Without loss of generality, we assume that $\delta_n^{(1)}\delta_n^{(2)} > 0$ for all n. Therefore, we have

$$\overline{D}_n = |\delta_n^{(1)} - \delta_n^{(2)}|^4 + |v_n^{(1)} - v_n^{(2)}|^2.$$

Remind from case b.1 that we have the following representation

$$\frac{g(x, \delta_n^{(1)}, v_n^{(1)}) - g(x, \delta_n^{(2)}, v_n^{(2)})}{\overline{D}_n} = \frac{\sum_{l=1}^{8} B_{n,l} \frac{\partial^l f}{\partial \delta^l}(x, -\delta_n^{(2)}, v_n^{(2)}) + \overline{R}(x)}{\overline{D}_n}.$$

The main challenge in Case b.3 is that $\|\overline{R}(x)\|_\infty/\overline{D}_n \not\to 0$ as $n \to \infty$. To avoid this issue, we will utilize the technique in Case a.2 of the proof of Theorem 2.2. In particular, we will demonstrate two key properties: $\|\overline{R}(x)\|_\infty/\max_{1\leq l\leq 8}|B_{n,l}| \to 0$ and $\max_{1\leq l\leq 8}|B_{n,l}|/\overline{D}_n \not\to 0$ as $n \to \infty$.

Under the settings of Case a.2.1 and Case a.2.2 in the proof of part (a), with the same argument as that in these cases, we have $|B_{n,2}|/\overline{D}_n \not\to 0$ and $\|\overline{R}(x)\|_\infty/|B_{n,2}| \to 0$. Therefore, we have $\overline{R}(x)/\max_{1\leq l\leq 8}|B_{n,l}| \to 0$ and $\max_{1\leq l\leq 8}|B_{n,l}|/\overline{D}_n \not\to 0$ under the settings of Case a.2.1 and Case a.2.2. It implies that we only need to focus on the setting that

$$|v_n^{(1)} - v_n^{(2)}|/\left\{|\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(1)} + \delta_n^{(2)}|\right\} \to 1.$$

Without loss of generality, we assume that $(v_n^{(1)} - v_n^{(2)})/\left\{(\delta_n^{(1)} - \delta_n^{(2)})(\delta_n^{(1)} + \delta_n^{(2)})\right\} \to 1$ as the argument for the setting that this ratio goes to -1 is similar. Under this setting, we can easily check that

$$|B_{n,4}|/\left\{|\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(2)}|^3\right\} \to 4.$$

Therefore, as n is sufficiently large, we have

$$|B_{n,4}| \gtrsim |\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(2)}|^3.$$

If we have $\max_{1\leq l\leq 8} |B_{n,l}|/\overline{D}_n \to 0$, then $|B_{n,4}|/\overline{D}_n \to 0$ leads to $|\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(2)}|^3/\overline{D}_n \to 0$. Therefore, the following holds

$$|v_n^{(1)} - v_n^{(2)}|^2 / \left\{|\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(2)}|^3\right\} \to \infty,$$

which means $|v_n^{(1)} - v_n^{(2)}|/|\delta_n^{(2)}|^2 \to \infty$, which is a contradiction to the assumption that $(v_n^{(1)} - v_n^{(2)})/\left\{(\delta_n^{(1)} - \delta_n^{(2)})(\delta_n^{(1)} + \delta_n^{(2)})\right\} \to 1$. Thus, we have $\max_{1\leq l\leq 8} |B_{n,l}|/\overline{D}_n \not\to 0$. On the other hand, as n is sufficiently large, we have

$$\begin{aligned}\frac{|\delta_n^{(1)} - \delta_n^{(2)}|^{\alpha_1}|v_n^{(1)} - v_n^{(2)}|^{\alpha_2}\|\overline{R}_{2,\alpha}(x)\|_\infty}{\max_{1\leq l\leq 8}\{|B_{n,l}|\}} &\leq \frac{O\left(|\delta_n^{(1)} - \delta_n^{(2)}|^{\alpha_1}|v_n^{(1)} - v_n^{(2)}|^{\alpha_2}|\delta_n^{(2)}|^{4-|\alpha|+\gamma}\right)}{|\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(2)}|^3} \\ &= \frac{O\left(|\delta_n^{(1)} - \delta_n^{(2)}|^{|\alpha|}|\delta_n^{(2)}|^{4-\alpha_1+\gamma}\right)}{|\delta_n^{(1)} - \delta_n^{(2)}||\delta_n^{(2)}|^3} \to 0.\end{aligned}$$

It implies that $\|\overline{R}(x)\|_\infty / \max_{1\leq l\leq 8}\{|B_{n,l}|\} \to 0$. From here, using the same argument as that of Case a.2.3, we obtain the contradiction, which leads to the conclusion that Case b.3 cannot hold. As a consequence, we achieve the conclusion of part (b) of the theorem.

G.3 Proof of extra results

In this appendix, we provide proof for an additional result with the non-polynomial convergence rate of MLE $\widehat{\delta}_n^{\text{mle}}$ under the known variances setting (2.1).

Proposition 4. *Under the symmetric regime of the true model (2.1), we have*

$$\sup_{\delta_n\in\Theta} \mathbb{E}_{\delta_n}\left|\widehat{\delta}_n^{mle} - \delta_n\right| \gtrsim n^{-1/r},$$

where $\Theta = [-1, 1]$. Here, $\mathbb{E}_{\delta_n}$ denotes the expectation taken with respect to product measure with mixture density of $Y_1, \ldots, Y_n$ under the model (2.1).

Proof. We divide our argument for the proof of this result into two key parts.

Part 1 - Upper bound of Hellinger distance between mixing densities in terms of their corresponding parameters To obtain the conclusion for this inequality, we first

prove the following key result

$$\inf_{\delta^{(1)},\delta^{(2)}\in\Theta} h\left(g(x,\delta^{(1)}),g(x,\delta^{(2)})\right)/\left|\delta^{(1)}-\delta^{(2)}\right|^r=0 \tag{G.14}$$

for any $r\geq 1$. In fact, we construct two sequences $\left\{\delta_n^{(1)}\right\}$ and $\left\{\delta_n^{(2)}\right\}$ such that $\delta_n^{(1)}=-\delta_n^{(2)}$ for all $n\geq 1$. Then, it is clear that $h\left(g(x,\delta_n^{(1)}),g(x,\delta_n^{(2)})\right)=0$ for all $n\geq 1$. Therefore, it is straightforward that $h\left(g(x,\delta_n^{(1)}),g(x,\delta_n^{(2)})\right)\leq\left|\delta_n^{(1)}-\delta_n^{(2)}\right|^r$ for any $r\geq 1$. As a consequence, we achieve the conclusion of (G.14).

Part 2 - Le Cam's argument for minimax lower bound Now, we follow the traditional Le Cam's argument for minimax lower bound to achieve the conclusion with non-polynomial convergence rate of $\widehat{\delta}_n^{\text{mle}}$ to δ_n (Yu, 1997). In particular, due to the result from (G.14), for any $\epsilon_n>0$ sufficiently small and any fixed $r\geq 1$, we can find $\delta_n^{(1)}$ and $\delta_n^{(2)}$ such that $\left|\delta_n^{(1)}-\delta_n^{(2)}\right|=2\epsilon_n$ and $h\left(g(x,\delta_n^{(1)}),g(x,\delta_n^{(2)})\right)\leq C\epsilon_n^r$ where C is a fixed positive constant. Invoking Lemma 1 from Yu (1997), the following inequality holds

$$\sup_{\delta_n\in\Theta}\mathbb{E}_{\delta_n}\left|\widehat{\delta}_n^{\text{mle}}-\delta_n\right|\geq\sup_{\delta_n\in\left\{\delta_n^{(1)},\delta_n^{(2)}\right\}}\mathbb{E}_{\delta_n}|\widehat{\delta}_n-\delta_n|\geq\epsilon_n\left[1-V\left(g^n\left(x,\delta_n^{(1)}\right),g^n\left(x,\delta_n^{(2)}\right)\right)\right], \tag{G.15}$$

where $g^n\left(x,\delta_n^{(1)}\right)$ denotes the density of n i.i.d. samples $Y_1,\ldots,Y_n$. By means of classical inequality between total variation distance and Hellinger distance $V\leq h$, we obtain that

$$V\left(g^n(x,\delta_n^{(1)}),g^n(x,\delta_n^{(2)})\right)\leq h\left(g^n(x,\delta_n^{(1)}),g^n(x,\delta_n^{(2)})\right)\leq\sqrt{1-(1-C^2\epsilon_n^{2r})^n}.$$

By choosing $C^2\epsilon_n^{2r}=1/n$, it is clear that

$$\epsilon_n\left[1-V\left(g^n\left(x,\delta_n^{(1)}\right),g^n\left(x,\delta_n^{(2)}\right)\right)\right]\gtrsim\epsilon_n\gtrsim n^{-1/2r}. \tag{G.16}$$

Combining the results from (G.15) and (G.16), we achieve the conclusion that

$$\sup_{\delta_n\in\Theta}\mathbb{E}_{\delta_n}\left|\widehat{\delta}_n^{\text{mle}}-\delta_n\right|\gtrsim n^{-1/r}$$

for any $r\geq 2$. □