Principal stratification in the Twilight Zone: Weakly separated components in finite mixture models[*]

Avi Feller
UC Berkeley

Evan Greif
Harvard University

Luke Miratrix
Harvard University

Natesh Pillai
Harvard University

This version: February 23, 2016

Abstract

Principal stratification is a widely used framework for addressing post-randomization complications in a principled way. 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 the 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 are unknown. Even though the MLE is asymptotically efficient, we show through extensive simulations that the MLE has undesirable properties in practice. In particular, when mixture components are only weakly separated, we observe "pile up", in which the MLE estimates the component means to be equal, even though they are not. We first show that parametric convergence can break down in certain situations. We then derive a simple moment estimator that displays key features of the MLE and use this estimator to approximate the finite sample behavior of the MLE. Finally, we propose a method to generate valid confidence sets via inverting a sequence of tests, and explore the case in which the component variances are unknown. Throughout, we illustrate the main ideas through an application of principal stratification to the evaluation of JOBS II, a job training program.

[*] *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 Bickel, Alex D'Amour, Peng Ding, Fabrizia Mealli, Christian Robert, Don Rubin, Dylan Small, Weixin Yao, and members of the Spencer group for helpful comments and discussion, as well as seminar participants at the 2015 Atlantic Causal Inference Conference and at the 2015 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

Principal stratification is a widely used framework for addressing post-randomization complications in a principled way (Frangakis and Rubin, 2002). Typically, the goal is to estimate a causal effect within a partially latent subgroup known as a principal stratum. While there are many possible ways to estimate these principal causal effects, by far 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.

At the same time, it has long been conventional wisdom that finite mixture models can yield pathological estimates (Redner and Walker, 1984). As Larry Wasserman remarked, finite mixture models are the "Twilight Zone of Statistics" (Wasserman, 2012). Our motivation for this paper is to understand how the pathological features of finite mixture models affect inference for component-specific means, which is a central challenge in principal stratification and is also of general scientific interest.

Despite the wealth of literature on finite mixture models, there is relatively little research on the behavior of standard estimators like the Maximum Likelihood Estimate (MLE) in settings typical of principal stratification, namely when substantial separation between mixture components is unlikely. We take a first step by carefully studying a simple but fundamental model: a two-component homoskedastic Gaussian mixture model,

$$Y_i \overset{\text{iid}}{\sim} \pi\mathcal{N}(\mu_0, \sigma^2) + (1-\pi)\mathcal{N}(\mu_1, \sigma^2), \tag{1.1}$$

in which only the component means, $\{\mu_0, \mu_1\}$, are unknown. That is, we assume that the mixing proportion, π, is known and that $0 < \pi < \frac{1}{2}$.[1] We also assume that the within component variance, σ^2, is known and equal between components (we later relax these restrictions). As we show, identification for μ_0 and μ_1 is immediate and unambiguous. Moreover, when the number of components is known (i.e., $\mu_0 \neq \mu_1$), the usual regularity conditions apply (Everitt and Hand, 1981). In other words, the MLE should be well-behaved in this simple example.[2] Unfortunately, this does not bear out in practice: even when the model is correctly specified and when $\mu_0 \neq \mu_1$, the MLE behaves poorly in many realistic settings (Day, 1969; Hosmer Jr, 1973; Redner and Walker, 1984).

1.1 Our contribution

In this paper, we (1) document two pathologies for the MLE of component-specific means that arise when components are not well separated; (2) use moment equations to build intuition for these pathologies and to characterize their behavior as a function of mixture parameters; and (3) propose a method to construct confidence sets via test inversion. In general, we present results for the standard mixture case and then explore additional complications that arise in the context of principal stratification models.

First, the main inferential challenge for the mixture model in Equation 1.1 is estimating the difference in component means, $\Delta \equiv \mu_0 - \mu_1$. There are two key issues. The first issue is known as *pile up*, which occurs when the likelihood surface for Δ is unimodal and centered at zero, despite the fact that $\Delta \neq 0$. Even with

[1] Similarly to Tan and Chang (1972), we assume that $\pi < 1/2$ so that the mixture is identified and to avoid degeneracy of the third cumulant of the mixture distribution when $\pi = 1/2$. Apart from the restriction that $\pi \neq 1/2$, this is completely general, since we can simply switch the component labels.

[2] Note that this is distinct from understanding the convergence of EM and related algorithms, as we are considering the performance of the MLE rather than our ability to find it. Balakrishnan et al. (2014) give a recent discussion.

a bimodal likelihood, the same underlying pathology that causes pile up can also lead to severe bias in the MLE. The second issue is the classic problem of choosing the global mode in a bimodal likelihood (McLachlan and Peel, 2004). This problem is particularly pernicious in the settings we explore here. In our simulations, the usual heuristic of choosing the mode with the higher likelihood is typically no better than a coin flip, often yielding the opposite sign from the truth, $\text{sgn}\left(\widehat{\Delta}^{\text{mle}}\right) \neq \text{sgn}(\Delta)$.

From a theoretical perspective, complications arise when Δ is non-zero but still "close enough" to zero (zero is the point at which the regularity conditions no longer hold). Hence, Δ is still point identified but identification is weak. 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). Chen et al. (2014) previously explored issues of weak identification in finite mixtures, though with a very different goal. In the spirit of this broad literature, we also show that "standard" asymptotic results for finite mixtures can break down in certain cases.

To understand the behavior of the MLE in finite samples, we must investigate the likelihood equations directly. As these are analytically intractable, we develop intuition by deriving a simple method of moment estimator that captures key features of the MLE. We connect the pathologies that we observe to the difficulty of estimating non-linear functions of the sample variance and skewness. We then derive analytic formulas for the probability that each pathology will occur in a given setting and conduct simulations to show that the MLE and moment estimator display these behaviors at similar rates. Overall, these diagnostic formulas are analogous to design or power calculations, giving a sense of the behavior of estimates prior to actually running the analysis.

There are many possible solutions to counter the pathologies outlined above. Following common practice in other settings with weak identification, we suggest one based on test inversion, namely generating simulation-based p-values for a grid of values of Δ and then inverting these tests to form confidence sets. This approach is known as a *grid bootstrap* in time series models (Andrews, 1993; Hansen, 1999; Mikusheva, 2007) and is closely related to the parametric bootstrap for the Likelihood Ratio Test statistic in finite mixture models (McLachlan, 1987; Chen et al., 2014). When the model is correctly specified, this approach yields exact p-values up to Monte Carlo error. This approach works well in the settings we explore here.

Finally, we discuss the case when the within-component variance, σ^2, is unknown, both for the equal- and unequal-variance case. We show that the performance of the MLE is even worse in this setting, with substantial bias for realistic values of Δ.

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 Zhang et al. (2008), Richardson et al. (2011), and Frumento et al. (2012). Following Mattei et al. (2013), we illustrate the key concepts through the evaluation of JOBS II, a job training program that has served as a popular example in the causal inference literature (see, for example, Vinokur et al., 1995; Jo and Stuart, 2009).

1.2 Related literature on finite mixtures

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

First, many early researchers used simulation to study the finite sample performance of moment and maximum likelihood estimators for finite mixtures (Day, 1969; Tan and Chang, 1972; Hosmer Jr, 1973). Redner and Walker (1984) give an extensive review. Interestingly, we are aware of very few simulation studies on the performance of these estimators published since Redner and Walker (1984), despite dramatic changes in computational power in the intervening thirty years. Frumento et al. (2016) offer an important recent exception in the context of principal stratification models. Consistent with our results, the authors find poor coverage for standard confidence intervals computed via the MLE plus information- or bootstrap-based standard errors. See also Chung et al. (2004).

Second, there has been extensive research on the asymptotic behavior of finite mixtures models. The standard result is that, with a known number of components and under general regularity conditions, mixture parameters have $\sqrt{n}$-convergence (Redner and Walker, 1984; Chen, 1995). Chen (1995), however, shows that parametric convergence can break down in certain cases (see also Heinrich and Kahn, 2015). Similar issues arise in the asymptotic distribution of the Likelihood Ratio Test (LRT) statistic for testing the number of components in finite mixtures (McLachlan and Peel, 2004). McLachlan (1987) proposes a parametric bootstrap to resolve these issues in settings including the two-component Gaussian mixture model. Chen et al. (2014) propose a similar method when the parameters are *near* but not at a singularity in the parameter space. In their setting, however, the goal is to estimate the mixing proportions, π, although the broad argument is quite similar to what we discuss here. The asymptotic behavior of the LRT with known mixing proportion, π, is addressed in Quinn et al. (1987), Goffinet et al. (1992) and Polymenis and Titterington (1999). See also Aitkin and Rubin (1985).

Finally, the *sign error* issue we address is exactly the well-studied question of choosing the "correct" mode in a multi-modal likelihood. McLachlan and Peel (2004) and Blatt and Hero (2007) give reviews. In virtually all cases, the standard recommendation is to choose the mode with the highest value of the likelihood. Some exceptions include Gan and Jiang (1999) and Biernacki (2005), who introduce tests that leverage different methods for computing the score function; Mercatanti (2013), who suggests a method that incorporates moment estimates; and Frumento et al. (2012), who propose a scaled log-likelihood ratio statistic to compare different modes.

1.3 Bayesian inference for finite mixtures

Finite mixture models are an important topic in Bayesian statistics, in part because mixtures fit naturally into the Bayesian paradigm (Frühwirth-Schnatter, 2006). This approach offers some distinct advantages over likelihood-based inference.[3] 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.[4]

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. Our results provide a framework for assessing how strong that prior information must be to avoid the pitfalls that we document, for example, by extending the grid bootstrap to incorporate a prior.

1.4 Paper plan

The paper proceeds as follows. In Section 2, we give an overview of finite mixture models and their applications to causal inference. In Section 3, we describe finite sample properties of the MLE and give results when the separation between μ_1 and μ_0 goes to zero asymptotically. In Section 4, we derive the method of moments estimator, characterize the pathologies, and connect these results to the MLE. In Section 5, we propose methods for obtaining confidence intervals for the component means via test inversion. In Section 6, we apply these methods to the JOBS II example. In Section 7, we relax the assumption of known variance. We conclude with some thoughts on possible next steps.

Finally, the supplementary materials address several points that go beyond the main text. First, we give additional discussion on applying these results to the principal stratification. In particular, we discuss the role of covariates, which are widely used as part of principal stratification models (e.g., Jo, 2002; Zhang et al., 2009). In general, we find that simply adding covariates without also incorporating additional restrictions (such as conditional independence assumptions) can increase the probability of obtaining a pathological result. Second, we address the performance of both the case-resampling bootstrap and Bayesian methods in this setting, finding poor performance overall. Finally, we give a much more detailed proof of the main theorem.

[3] 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).

[4] 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."

2 Finite mixtures in causal inference

2.1 Setup

To illustrate the role of finite mixtures in causal inference, we begin with the canonical example of a randomized experiment with noncompliance. We set up the problem using the potential outcomes framework of Neyman (1923) and Rubin (1974). We observe N individuals who are randomly assigned to a treatment group, $T_i = 1$, or control group, $T_i = 0$. 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.

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.

In practice, there is often noncompliance with treatment assignment (Angrist et al., 1996). 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. The estimands of interest are the ITT effects for Compliers and Never Takers:

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

in which $\mu_{c1}, \mu_{c0}, \mu_{n1}$, and μ_{n0} represent the outcome means for Complier assigned to treatment, Compliers assigned to control, Never Takers assigned to treatment, and Never Takers assigned to control, respectively.

In the case of one-sided noncompliance, we observe stratum membership for individuals assigned to treatment. Therefore, 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 treatment:

$$Y_i^{\text{obs}} \mid Z_i = 0 \sim \pi f_{\text{c}0}(y_i) + (1 - \pi) f_{\text{n}0}(y_i), \tag{2.1}$$

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 $\mathrm{ITT_n} = 0$, or equivalently, $\mu_{\mathrm{n}1} = \mu_{\mathrm{n}0}$. With this assumption, we can then estimate $\mathrm{ITT_c}$ with the usual instrumental variables approach (Angrist et al., 1996). Without this assumption, however, we need to impose other structure on the problem to achieve point identification (see, e.g., Zhang and Rubin, 2003).

2.2 Model-based principal stratification

One increasingly common option is to invoke parametric assumptions for the $f_{u0}(y)$. In a seminal paper, Imbens and Rubin (1997) outlined a model-based framework for instrumental variable models, proposing a parametric model for the outcome distribution conditional on stratum membership and treatment assignment, such as $f_{uz}(y_i) = \mathcal{N}(\mu_{uz}, \sigma^2_{uz})$, with u representing stratum membership and z representing treatment assignment. 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.[5] For simplicity, we focus on applications of principal stratification in which the researcher assumes that certain principal strata do not exist (i.e., monotonicity) and it is possible to directly estimate the distribution of principal strata. For example, in the case of two-sided noncompliance, the "no Defiers" assumption makes this possible (Angrist et al., 1996). While the more general case is important, inference is also more challenging. See Page et al. (2015) for a brief discussion of these assumptions.

For our example of one-sided noncompliance, we can write the observed data likelihood with Normal component distributions as:

$$\begin{aligned}\mathcal{L}_{\mathrm{obs}}(\theta) = &\\ &\prod_{i:\ Z_i=1,\ D_i^{\mathrm{obs}}=1} \pi\mathcal{N}(y_i \mid \mu_{\mathrm{c}1}, \sigma^2_{\mathrm{c}1}) \ \times \\ &\prod_{i:\ Z_i=1,\ D_i^{\mathrm{obs}}=0} (1-\pi)\mathcal{N}(y_i \mid \mu_{\mathrm{n}1}, \sigma^2_{\mathrm{n}1}) \ \times \\ &\prod_{i:\ Z_i=0} \left[\pi\mathcal{N}(y_i \mid \mu_{\mathrm{c}0}, \sigma^2_{\mathrm{c}0}) + (1-\pi)\mathcal{N}(y_i \mid \mu_{\mathrm{n}0}, \sigma^2_{\mathrm{n}0})\right],\end{aligned}$$

where θ represents the vector of parameters and $\mathcal{N}(y_i \mid \mu, \sigma^2\cdot)$ is the Normal density with mean μ and variance σ^2. The observed data likelihood for individuals assigned to treatment ($Z_i = 1$) immediately factors into the likelihood for the Compliers and the likelihood for the Never Takers. We can directly estimate the component parameters under treatment, $\mu_{\mathrm{c}1}$, $\mu_{\mathrm{n}1}$, $\sigma^2_{\mathrm{c}1}$, and $\sigma^2_{\mathrm{n}1}$. We can also directly estimate π among individuals assigned to treatment. Therefore, we are essentially left with a two-component Normal mixture model with known π among those individuals assigned to control.[6] Note that the mixture portion of this likelihood is in general unbounded due to the possibility of either σ^2_{u1} or σ^2_{u0} being close to 0. However, the whole likelihood is bounded with probability one due to the effect of the non-mixture terms.[7]

[5] 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).

[6] 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 treat π as being estimated directly from the treatment group.

[7] We are assuming that at least one observation is made in each compliance stratum of the treatment group.

To estimate ITT_c and ITT_n, we need to estimate $\mu_{\mathrm{c}0}$ and $\mu_{\mathrm{n}0}$. This means that the component-specific variances, $\sigma^2_{\mathrm{c}0}$ and $\sigma^2_{\mathrm{n}0}$, are nuisance parameters. We initially assume that the component-specific variances are constant across components: $\sigma^2 = \sigma^2_{\mathrm{c}1} = \sigma^2_{\mathrm{c}0} = \sigma^2_{\mathrm{n}1} = \sigma^2_{\mathrm{n}0}$ (see, for example, Gallop et al., 2009). Since we can directly estimate $\sigma^2_{\mathrm{c}1}$ and $\sigma^2_{\mathrm{n}1}$, this means that we no longer need to estimate the component-specific variances in the finite mixture model and can treat these parameters as known.[8] We relax this assumption in Section 7.

2.3 JOBS II

For a running example, we use 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). We focus on a subset of $N = 410$ high risk individuals, with $N_1 = 278$ randomly assigned to treatment and $N_0 = 132$ randomly assigned to control. For illustration, we estimate the impact of the program on (log) depression score six months after randomization.

An important complication in this study is that only 55% of those individuals assigned to treatment actually enrolled in the program. Therefore, we are not only interested in the overall ITT, but also in the ITT for Compliers and the ITT for Never Takers. Note that we do not invoke the exclusion restriction for Never Takers; in other words, we want to estimate ITT_n rather than assume that $\mathrm{ITT}_\mathrm{n} = 0$. To do so, we follow Mattei et al. (2013) and assume the outcome distribution is Normal for each U and Z combination. As an intermediate step, we are interested in $\Delta = \mu_{\mathrm{c}0} - \mu_{\mathrm{n}0}$ for this model.

Figure 1 shows the distribution of the MLE of Δ for 1000 fake data sets generated from a two-component homoskedastic Normal mixture model with $N = 132$, $\pi = 0.55$, and $\sigma = 1$ (so that all estimates are in effect size units).[9] In Figures 1a and 1b, the assumed values of Δ are 0.5 and 1.0 standard deviations, respectively, which are quite large differences in the context of JOBS II. Clearly the sampling distribution of the MLE is markedly non-Normal, showing strong bimodality in addition to a large spike around zero.

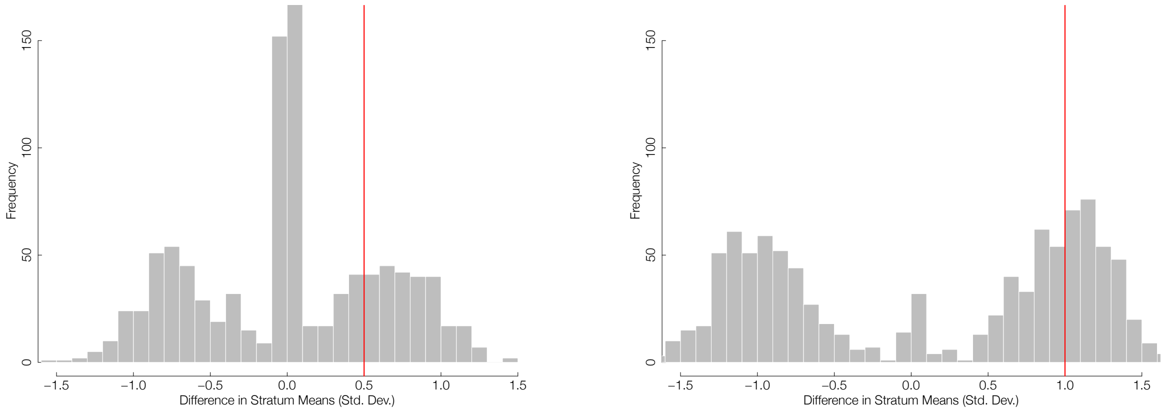
3 Likelihood inference in finite mixture models

3.1 Asymptotic results

We begin by documenting that standard asymptotic results can break down when components are not well separated. In particular, standard results state that, with a known number of components and under general regularity conditions, mixture parameters have $\sqrt{n}$-convergence (Redner and Walker, 1984; Chen, 1995). Based on examples like those in Figure 1, these results do not necessarily "kick in" for settings where the difference in component means is small relative to the sample size. To explain this disconnect, we consider the case in which Y_i are distributed as in Equation 1.1 but the value of the component means μ_1 and μ_0 vary with n, becoming $\mu_{1,n}$ and $\mu_{0,n}$. Correspondingly, the separation of means becomes

[8] As with the mixing proportion, there is a very small amount of information about the overall σ^2 in the finite mixture model. Again, we ignore this complication.

[9] Note that this does not fit into the parameterization of Equation 1.1, but that all the same results hold for $\pi > 1/2$ and $\Delta = \mu_1 - \mu_0$. We switch the parameterization to match the JOBS II data set and Equation 2.1. Also, the only unknowns in the likelihood are the component means; all other parameters are assumed known and fixed at the correct values. Finally, we calculate the MLE directly, rather than via EM (Dempster et al., 1977).



(a) Assumed Difference in Means is 0.5 SD

(b) Assumed Difference in Means is 1.0 SD

Figure 1: Distribution of the MLE for Δ for 1000 fake data sets generated from a two-component homoskedastic Normal mixture model with $N = 132$, $\pi = 0.55$, and $\sigma = 1$.

$$\Delta_n = \mu_{0,n} - \mu_{1,n}. \tag{3.1}$$

We prove that if $\Delta_n = o(n^{-1/4})$ in a class of models similar to those considered by Chen (1995), then $|\widehat{\Delta}_n - \Delta_n| = O_p(n^{-1/4})$ and $|\widehat{\Delta}_n - \Delta_n| \neq o_p(n^{-1/4})$, with $\widehat{\Delta}_n$ denoting the maximum-likelihood estimator of Δ_n. That is, if the separation of $\mu_{1,n}$ and $\mu_{0,n}$ disappears quickly enough, the maximum-likelihood estimator's parametric rate of convergence is sacrificed.

Theorem 3.1. *Let $Y_{i,n}$ for $i \in \{1, \ldots, n\}$ and $n \in \mathbb{N}$ be drawn independently from the model*

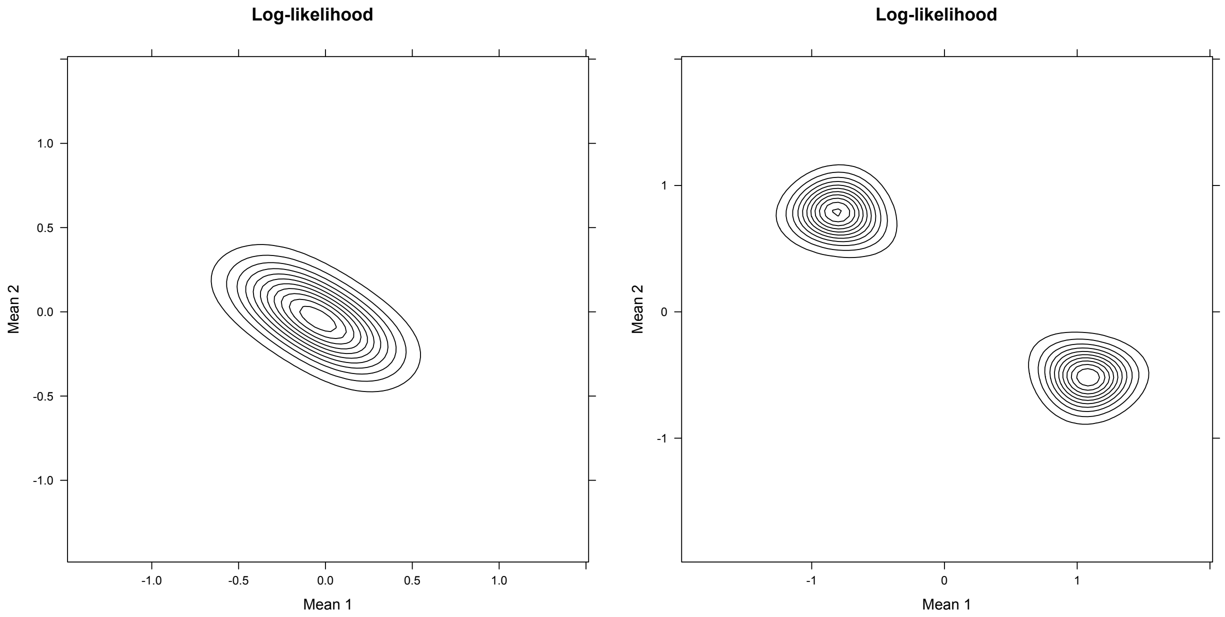
$$Y_{i,n} \sim \pi\mathcal{N}\left(-\Delta_n, \sigma^2\right) + (1-\pi)\mathcal{N}\left(\frac{\pi}{1-\pi}\Delta_n, \sigma^2\right), \tag{3.2}$$

with $\Delta_n = o(n^{-1/4})$. Then $|\widehat{\Delta}_n - \Delta_n| = O_p(n^{-1/4})$ and $|\widehat{\Delta}_n - \Delta_n| \neq o_p(n^{-1/4})$.

The proof of this result, which follows Chen (1995), is given in the Appendix. The key theoretical aspect of the above model is that the Fisher Information of Δ_n in each $y_{i,n}$, $I(\Delta_n)$, is 0 when $\Delta_n = 0$. Thus, the Fisher Information of Δ_n contained in the entire sample $\{y_{1,n}, \ldots, y_{n,n}\}$ is $nI(\Delta_n)$, and there are two competing limits: $n \to \infty$ and $I(\Delta_n) \to 0$.

3.2 Pathologies of the MLE

The above result shows that the usual parametric asymptotic rate of convergence of the MLE can break down in a certain class of models. Recognizing that we should not necessarily expect parametric convergence, we now turn to characterizing the finite sample behavior of the MLE. The three-part sampling distribution in Figure 1 is a good illustration of the key inferential issues we face. We refer to the large point mass at zero, such as in Figure 1a, as *pile up*. For these simulated data sets, $\widehat{\Delta}^{\text{mle}} \approx 0$ even though $\Delta \neq 0$. This generally occurs when the likelihood surface is unimodal, as shown in Figure 2a. We refer to those MLEs with an



(a) Example log-likelihood for $\Delta = 0.2$

(b) Example log-likelihood for $\Delta = 1.0$

Figure 2: Examples of unimodal and bimodal log-likelihoods generated from the two-component Gaussian mixture

opposite sign from the truth as a *sign error*, in which $\text{sgn}(\widehat{\Delta}^{\text{mle}}) \neq \text{sgn}(\Delta)$. For this problem, the mixture likelihood surface is generally bimodal (see Figure 2b). A sign error occurs when the global MLE, obtained by choosing the mode with the higher likelihood (McLachlan and Peel, 2004), chooses the wrong mode.

These pathologies are present across extensive simulations. Specifically, we set parameters to $\pi \in \{0.2, 0.325, 0.45\}$, for $N \in \{50, 100, 200, 400, 500, 1000, 2000, 5000\}$ and $\Delta \in \{0.25, 0.5, 0.75, 1\}$. For each set of parameters, we simulate 1000 fake data sets from the Gaussian mixture model in Equation 1.1. For each data set, we then calculate the MLE, compute standard errors via the Hessian of the log-likelihood evaluated at the MLE (i.e., the observed Fisher information), and generate nominal confidence intervals via the point estimate ± 1.96 standard errors. We then calculate the average bias and coverage rates at each setting of the parameter values, as well as the empirical probabilities of pile up and sign error.

Figure 3 shows the empirical bias and coverage, respectively, for $\widehat{\Delta}^{\text{mle}}$. The left column shows simulation results when the true $\Delta = 0.25$ and the right column shows results when the true $\Delta = 0.75$. For both cases, $\pi = 0.325$ and $\sigma^2 = 1$. The first row shows the average estimate for $\widehat{\Delta}^{\text{mle}}$ across simulations, which is essentially 0 when $\Delta = 0.25$ and slowly climbs from 0 when $\Delta = 0.75$. The second row shows the empirical coverage of the 95% confidence intervals for Δ. Strikingly, this coverage deteriorates as Δ increases, although it begins to climb again for $\Delta = 0.75$. The third row shows the empirical probability of each pathology (where "correct" corresponds to an MLE that is neither pile up nor the wrong sign). For $\Delta = 0.25$, the probability of pile up is shockingly high, and is the likeliest outcome for sample sizes for sample sizes less than $N = 200$. Even with sample sizes as high as $N = 5000$, the estimated sign is as likely to be negative as positive. These pathologies help to explain the seemingly strange pattern with bias and coverage. For $\Delta = 0.25$, for example, coverage gets worse because confidence intervals get tighter around a point estimate that has the wrong sign.

Finally, while we focus on the behavior of the MLE itself, many factors contribute to poor coverage. Following Rubin and Thayer (1983) and Frumento et al. (2016), we note that in many settings the log-likelihood at the MLE is not well-approximated by a quadratic, which partly explains the simulation results (see also McLachlan and Peel, 2004). In the supplementary materials, we also show that bootstrap-based standard errors do not resolve these issues.

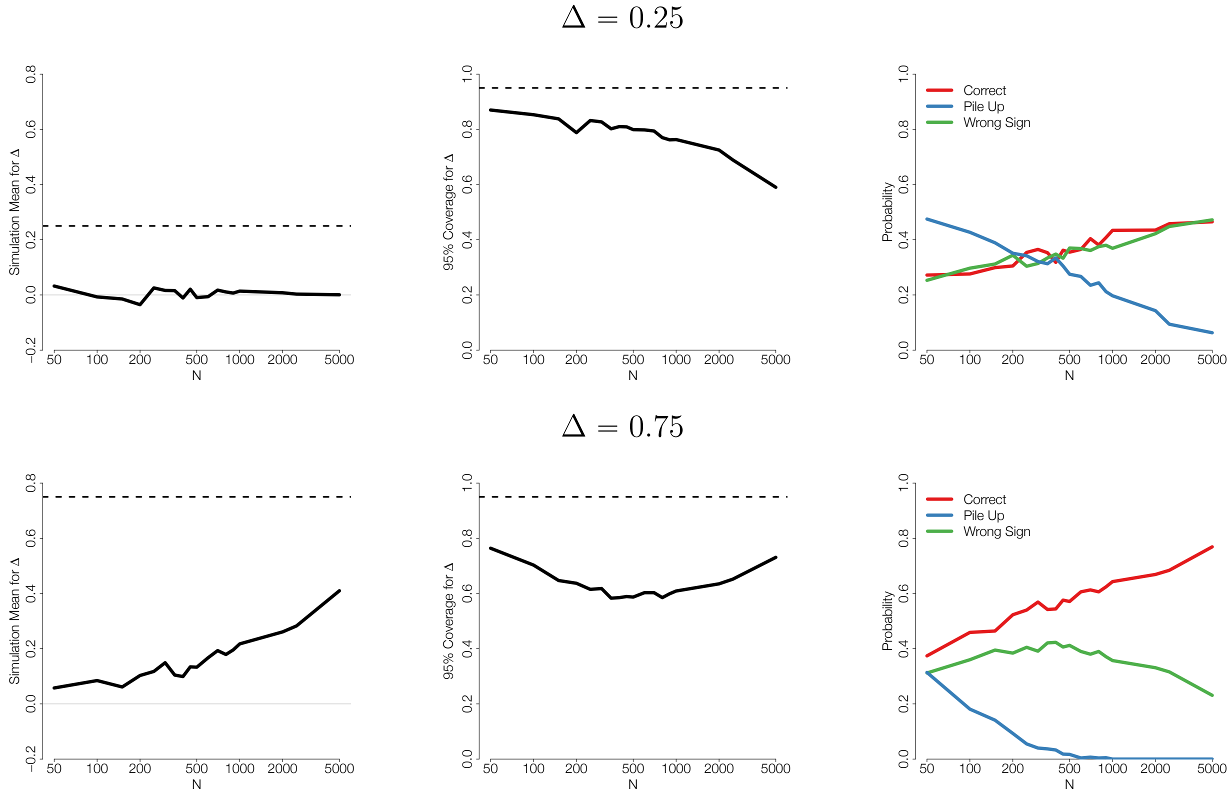


Figure 3: Simulation results for the MLE of Δ. The dotted lines represent the true values.

4 Method of Moments

4.1 Point estimation and confidence intervals

To understand this poor behavior for the MLE, we would ideally investigate the likelihood equations directly. Unfortunately, the likelihood equations are intractable and the mixture likelihood surface itself is notoriously complicated (Lindsay, 1983). We therefore develop a simple method of moments estimate, $\widehat{\Delta}^{\text{mom}}$, that captures the essential features of the MLE in the settings we consider.[10] Our key point—which has been made repeatedly before (e.g., Kiefer, 1978)—is that $\widehat{\Delta}^{\text{mle}}$ and $\widehat{\Delta}^{\text{mom}}$ essentially use the same sample information in practice. While the MLE incorporates an infinite number of moments and is an efficient estimator (Redner

[10] Note that moment estimators for finite mixture models can be quite complex—the goal here is to build intuition rather than to improve on existing moment estimators. See Furman and Lindsay (1994a,b). Titterington (2004) offers a review.

and Walker, 1984), in practice there is negligible information in the higher order moments, at least beyond the third or fourth moments. As a result, there is little reason to expect the MLE to fare better than the MOM estimate in settings where the latter breaks down. Similarly, by characterizing the MOM estimate, we can obtain good heuristic approximations of the MLE's behavior.

To define $\widehat{\Delta}^{\text{mom}}$, we first let κ_k be the kth cumulant of the observed mixture distribution (Tan and Chang, 1972). Then the first two mixture cumulants, the mean and variance, are:

$$\begin{aligned}\kappa_1 &= \pi\mu_0 + (1-\pi)\mu_1,\\ \kappa_2 &= \sigma^2 + \pi(1-\pi)(\mu_1-\mu_0)^2,\end{aligned}$$

where σ^2 is the within-component variance and $\pi(1-\pi)(\mu_1-\mu_0)^2$ is the between-component variance. Substituting in $\Delta \equiv \mu_1 - \mu_0$ yields

$$\kappa_1 = \mu_1 + \pi\Delta, \tag{4.1}$$

$$\kappa_2 = \sigma^2 + \pi(1-\pi)\Delta^2. \tag{4.2}$$

Since π and σ^2 are known, we can immediately estimate Δ^2 using the observed mixture variance:

$$\widehat{\Delta^2} = \frac{\widehat{\kappa}_2 - \sigma^2}{\pi(1-\pi)}.$$

In words, $\widehat{\Delta^2}$ is the scaled difference between the total mixture variance and the within-component variance. The estimate of $\widehat{\kappa}_2$ is the usual unbiased estimate of the sample variance. Note that we do not constrain the numerator to be non-negative in order to highlight the pathological issues that arise in the MLE.

If we know the sign of Δ, then we take the square root to obtain a moment estimate for Δ. However, if the sign of Δ is unknown, we must estimate it. We therefore need one more moment equation. A natural choice is the third cumulant of the observed mixture:

$$\kappa_3 = \pi(1-\pi)(1-2\pi)\Delta^3. \tag{4.3}$$

Since π is assumed to be known, $\widehat{\text{sgn}(\Delta)} = \text{sgn}(\widehat{\kappa}_3)$. We do not consider the knife-edge case of $\pi = 1/2$, in which there is no information in the data about the ordering of the components.

This yields the following moment estimator for Δ, with $\pi < \frac{1}{2}$:

$$\widehat{\Delta}^{\text{mom}} = \text{sgn}(\widehat{\kappa}_3)\sqrt{\frac{\widehat{\kappa}_2 - \sigma^2}{\pi(1-\pi)}}. \tag{4.4}$$

Note that the third moment contains some information about the magnitude of Δ, in addition to information about its sign. Therefore, in this simple case, Δ is over-identified and we can use the Generalized Method of Moments (GMM; Hansen, 1982) or the Moment Generating Function method of Quandt and Ramsey (1978) to combine the moment equations. However, we use this simple estimate to approximate the MLE, and so we do not need to focus on these more complex approaches.

Accounting for uncertainty in $\widehat{\Delta}^{\text{mom}}$ is conceptually straightforward. In the case of the two-component

Gaussian mixture model, Tan and Chang (1972) derive the sampling distribution of the first three cumulants up to $O(1/n^2)$:

$$\begin{pmatrix} \widehat{\kappa}_1 \\ \widehat{\kappa}_2 \\ \widehat{\kappa}_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \kappa_1 \\ \kappa_2 \\ \kappa_3 \end{pmatrix}, \frac{1}{n} \begin{pmatrix} \kappa_2 & \kappa_3 & c^{\pi}_{13}\Delta^4 \\ & c^{\pi}_{22}\Delta^4 + 2\kappa_2^2 & c^{\pi}_{32}\Delta^5 + 6\kappa_2\kappa_3 \\ & & c^{\pi}_{33a}\Delta^6 + c^{\pi}_{33b}\kappa_2\Delta^4 + 6\kappa_2^3 \end{pmatrix} \right), \tag{4.5}$$

where c^{π}_{rc} are constants that depend on π.[11] We can then apply the Delta method to Equation 4.4 to obtain approximate confidence intervals for $\widehat{\Delta}^{\text{mom}}$.

In practice, however, $\widehat{\Delta}^{\text{mom}}$ is not a smooth function of $\widehat{\kappa}_2$ and $\widehat{\kappa}_3$. To make matters worse, estimates of higher order moments are extremely noisy: the variance of the sample variance depends on the fourth moment, and the variance of the sample skewness depends on the sixth moment. As a result, even if the true value of Δ is far from zero, due to sampling variability, the observed sample moments could yield estimates at or near critical points in the mapping from the sample moments to $\widehat{\Delta}^{\text{mom}}$.

We can now immediately see our two pathologies:

- **Pile up.** If the estimated overall variance is less than the assumed within-group variance, $\widehat{\kappa}_2 < \sigma^2$, then $\widehat{\Delta^2} < 0$ and $\widehat{\Delta}^{\text{mom}}$ is undefined. Heuristically, the MLE in these settings is $\widehat{\Delta}^{\text{mle}} = 0$ and generally corresponds to a unimodal likelihood. In other words, the MLE implicitly restricts the estimate of Δ^2 to be nonnegative.

- **Sign error.** If the estimated sign of the skewness does not equal the true sign of the skewness, $\text{sgn}(\widehat{\kappa}_3) \neq \text{sgn}(\kappa_3)$, then $\text{sgn}\left(\widehat{\Delta}^{\text{mom}}\right) \neq \text{sgn}(\Delta)$. Heuristically, this corresponds to the case when the higher mode in a bimodal likelihood does not, in fact, correspond to the global mode.

4.2 Assessing pathologies in practice

Using the sampling distributions from Equation 4.5, we can calculate the approximate probability of each pathology occurring for any given Δ, π, n, and σ^2. For the case where $\Delta > 0$ and $\pi < \frac{1}{2}$, the marginal probabilities of the two pathologies are:

$$\mathbb{P}(\text{pile up}) = \mathbb{P}(\widehat{\kappa}_2 < \sigma^2) \approx \Phi\left(\frac{-\sqrt{n}\pi(1-\pi)\Delta^2}{\sqrt{c_\pi\Delta^4 + 2\kappa_2^2}} \right), \tag{4.6}$$

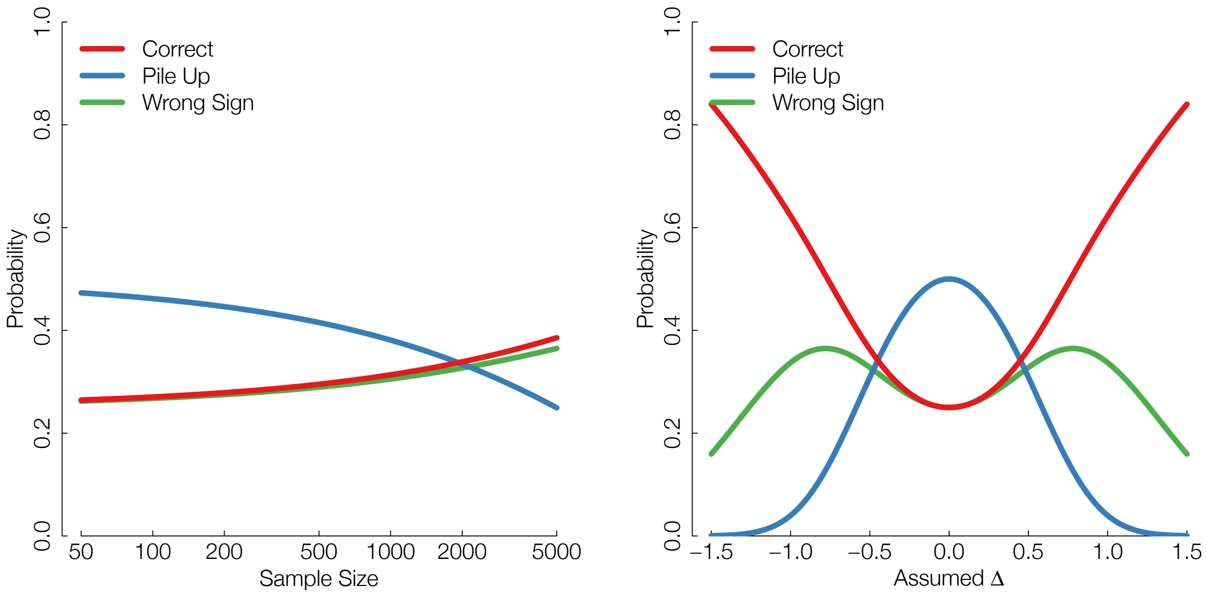
$$\mathbb{P}(\text{sign error}) = \mathbb{P}(\text{sgn}(\widehat{\kappa}_3) \neq \text{sgn}(\kappa_3)) \approx \Phi\left(\frac{-\sqrt{n}\pi(1-\pi)(1-2\pi)\Delta^2}{\sqrt{c''_\pi\Delta^6 + c''_\pi\kappa_2\Delta^4 + 6\kappa_2^2}} \right). \tag{4.7}$$

We use simulation to assess the accuracy of these approximations; see the supplementary materials for additional information. For pile up, the moment approximations are uniformly excellent, with over 95% agreement across simulation values. For sign error, the approximations are quite good for moderate π, but are less informative with smaller N and more extreme π.

Figure F.7a shows the joint approximate probabilities of pile up and sign error, calculated via the joint distribution from Equation 4.5, with $\pi = 0.325$, $\Delta = 0.25$, and varying sample size. Unsurprisingly, as the

[11] $c^{\pi}_{13} = 1 - 6\pi(1-\pi)$, $c^{\pi}_{22} = \pi(1-\pi)(1-6\pi(1-\pi))$, $c^{\pi}_{32} = \pi(1-\pi)(1-2\pi)(1-12\pi(1-\pi))$, $c^{\pi}_{33a} = \pi(1-\pi)(1-30\pi(1-\pi) + 120\pi^2(1-\pi)^2) + 9\pi^2(1-\pi)^2(1-2\pi)^2$, $c^{\pi}_{33b} = 9\pi(1-\pi)(1-6\pi(1-\pi))$.

sample size and Δ increase, the chance of a pathological result indeed decreases. However, these pathologies are hardly "small sample" issues—for $\Delta = 0.25$, which would be quite large in many social science applications, pile up remains the likeliest outcome even with sample sizes in the thousands. For $\Delta = 0.75$ (not shown) there is only a 70% chance of a correct estimate for $N = 2000$. Figure F.7b shows results for a moderate sample size of $N = 200$. In this case, if the mixture means are 0.5 standard deviations apart, a given estimate is just as likely to be a pile up, have the wrong sign, or be correct. These formulas can be used in a manner analogous to design or power calculations, as researchers typically know N and π prior to estimating a mixture model. Graphs such as Figure F.7b, can immediately show whether pile up and sign error are meaningful issues for plausible values of Δ.



(a) $\Delta = 0.25$

(b) $N = 200$

Figure 4: Normal approximations for the probability of each pathology for given sample size and separation of means. π is assumed to be 0.325. The "correct" outcome is defined as one minus the probability of pile up or sign error.

Finally, in smaller samples, we might be concerned about the Normal approximation to the sampling distributions for $\widehat{\kappa}_2$ and $\widehat{\kappa}_3$. In this case, it is convenient to use a case-resampling bootstrap to approximate these distributions. While this does not yield analytical formulas for the probability of each pathology, it is straightforward to estimate both $\mathbb{P}\{\widehat{\kappa}_2 < \sigma^2\}$ and $\mathbb{P}\{\widehat{\kappa}_3 < 0\}$, without needing to rely on an asymptotic approximation. Note that this is a standard application of the case-resampling bootstrap for which all the usual guarantees hold—rather than using the bootstrap for uncertainty on mixture component means.

4.3 Bias due to pile up

If $\widehat{\Delta^2} < 0$, it is clear that $\widehat{\Delta}^{\text{mom}}$ is undefined. Less obvious—but no less important—is that $\widehat{\Delta}^{\text{mom}}$ is severely biased if $\widehat{\Delta^2} > 0$ but the probability of pile up is large. Intuitively, this occurs because we are implicitly conditioning on the fact that $\widehat{\Delta^2} > 0$ when $\widehat{\Delta}^{\text{mom}}$ is well-defined; we are therefore estimating the mean of a

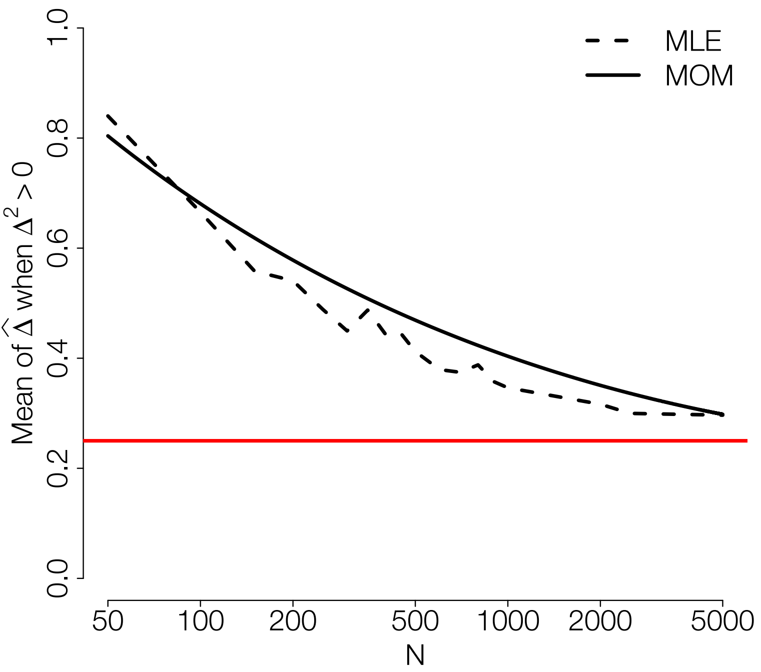


Figure 5: Mean for $\widehat{\Delta}^{\text{mle}}$ (simulated) and $\widehat{\Delta}^{\text{mom}}$ (calculated) when $\widehat{\kappa}_2 > \sigma^2$. The true value is $\Delta = 0.25$

truncated Normal. We see the same behavior with the MLE; the mode of the bimodal likelihood is biased away from zero because we are implicitly conditioning on the fact that the observed likelihood surface is indeed bimodal rather than unimodal.

Figure 5 shows the bias of $\widehat{\Delta}^{\text{mle}}$ in simulation studies for which there are no pathologies and for which the true $\Delta = 0.25$ and $\pi = 0.325$. The bias is substantial throughout, and for sample sizes $N < 1000$, the bias is larger in magnitude than Δ. As with the overall MLE, we can characterize this behavior using the MOM estimate. In particular, we use LOTUS and numerical integration to obtain the expected value of the truncated distribution, $\mathbb{E}[\widehat{\Delta}^{\text{mom}} \mid \widehat{\kappa}_2 > \sigma^2]$. As shown in Figure 5, this is a good approximation to the bias calculated via simulation.

It is useful to note that the bias of the MLE in this setting is closely related to the bias induced by introducing identifiability constraints, 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 $\Delta = 0$.

5 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 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 (Imbens and Rubin, 2015).

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). In the supplementary material, we explore the effects of mis-specification on the resulting confidence sets. Unsurprisingly, we find that the approach works well under mild mis-specification (i.e., generating data via t_{50} rather than Normal) but otherwise performs poorly. Note that 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 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. Since the additional details are not central to our argument, we discuss inference for the broader principal stratification model in the supplementary materials.

5.1 Overview of grid bootstrap

To conduct a grid bootstrap, we first need a grid. Define $\boldsymbol{\Delta} = \{\Delta_0, \Delta_1, \dots, \Delta_n\}$ 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 $\Delta_j \in \boldsymbol{\Delta}$:

$$H_0 : \Delta = \Delta_j \text{ vs. } H_1 : \Delta \neq \Delta_j. \tag{5.1}$$

For convenience we first center the data (we return to this in the next section). 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 $\Delta_j \in \boldsymbol{\Delta}$
    - Calculate the observed test statistic, $t_j^{\text{obs}} = t(\mathbf{y}^{\text{obs}}, \Delta_j)$.
    - 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^{obs} 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.

5.2 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 $\widehat{\kappa}_2$ and $\widehat{\kappa}_3$.[12] Equation 4.5 suggests a natural combination of the estimated cumulants:

$$t_\kappa(\mathbf{y}, \Delta_j) = (d_2, d_3)\text{Var}(\kappa_2, \kappa_3)^{-1}(d_2, d_3)^T, \tag{5.2}$$

where $d_k = \widehat{\kappa}_k - \kappa_k$, and we use the assumed null of $\Delta = \Delta_j$ to obtain (κ_2, κ_3) and $\text{Var}(\kappa_2, \kappa_3)$. In practice, the Normal approximation in Equation 4.5 is excellent, even for modest sample sizes (say $N > 100$). This implies:

$$t_\kappa(\mathbf{y}, \Delta_j) \overset{a}{\sim} \chi_2^2.$$

We can therefore obtain a p-value via a Wald test at each grid point, which is fast 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 to obtain a joint test statistic as in Equation 5.2:

$$t_\kappa(\mathbf{y}, \Delta_j) = (d_1, d_2, d_3)\text{Var}(\kappa_1, \kappa_2, \kappa_3)^{-1}(d_1, d_2, d_3)^T \sim \chi_3^2. \tag{5.3}$$

As above, we can obtain p-values via the grid bootstrap rather than via the χ^2 distribution. Figure 6 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 A.1 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.

6 Application to JOBS II

We now return to our example of JOBS II. As described in Section 2, we focus on the subset of $N_1 = 278$ randomly assigned to treatment and $N_0 = 132$ randomly assigned to control. In the treatment group,

[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 $\widehat{\kappa}_2$ or $\widehat{\kappa}_3$.

[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.

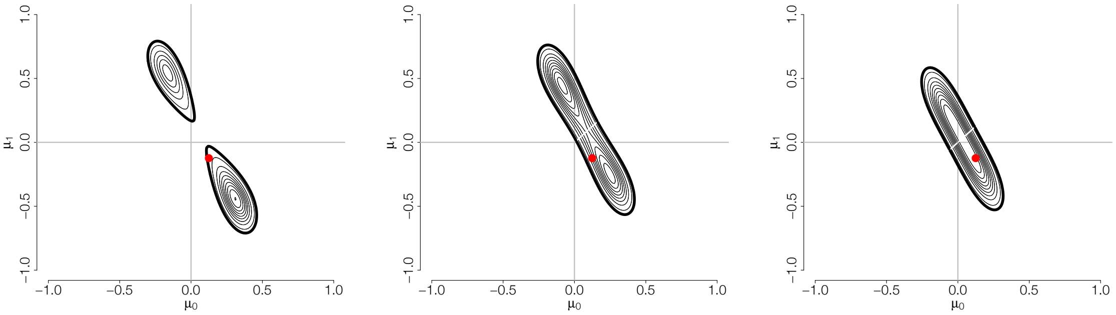


Figure 6: Three examples of the distribution of Wald test p-values from Equation 5.3. Simulated data are from Equation 1.1 with $N = 1000$, $\pi = 0.325$, $\sigma^2 = 1$, $\mu_0 = \frac{1}{8}$, $\mu_1 = -\frac{1}{8}$. The dark line shows the cutoff for $p = 0.05$. The red dot shows the true value. Note that the Wald test is undefined when $\mu_0 = \mu_1$.

Table 1: Summary statistics for observed groups in JOBS II

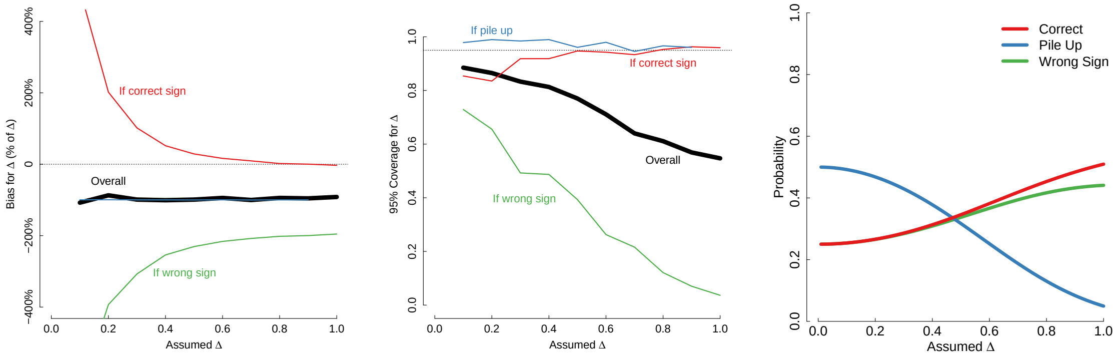
Z	**D$^{\text{obs}}$**	**Observed Mean**	**Observed SD**	**Possible Principal Strata**
1	1	-0.16	1.03	Compliers
1	0	0.05	0.96	Never Takers
0	0	0.14	0.98	Compliers and Never Takers

$1 - \pi = 45$ percent of individuals do not enroll in the program. The primary outcome is log depression score six months after randomization. For convenience, we standardize the outcome by subtracting off the grand mean and dividing by $\widehat{\sigma}_1 = \sqrt{\pi\widehat{\sigma}^2_{\text{n}1} + (1-\pi)\widehat{\sigma}^2_{\text{c}1}}$, the estimated within-component standard deviation under treatment. Table 1 shows summary statistics for the three observed groups for the standardized outcome. 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 $\widehat{\sigma}_{\text{c}1}$ and $\widehat{\sigma}_{\text{n}1}$ are quite close, which is consistent with our equal variance assumption (we relax this assumption below).

To assess the performance of the MLE in this setting, we simulate data with the same parameters as in the JOBS II example, $N = 132$, $\pi = 0.55$, and $\sigma^2 = 1$. Figures 7a and 7b show the bias and 95% coverage for the MLE for these parameters and assumed values of Δ. Figure 7c shows the probability of each pathology for these parameters. Note that that the results are unchanged if we consider negative values of Δ.

The pattern is striking. For values of $\Delta < 0.5$, the most likely $\widehat{\Delta}^{\text{mle}}$ is 0. That is, the likeliest outcome is a unimodal likelihood with the mode centered at $\Delta = 0$. Unsurprisingly, the MLE has poor bias and coverage properties for reasonable values of Δ. In particular, we see that the coverage of $\widehat{\Delta}^{\text{mle}}$ decreases as the assumed value of Δ increases from 0.1 to 2.0. This is due to the increasing probability of a sign error occurring and the poor conditional coverage of $\widehat{\Delta}^{\text{mle}}$ conditional on making a sign error. These suggest that, at best, we should interpret the MLE in this example with caution.

Figure 8 shows the outcome distribution and bootstrap sampling distributions for $\widehat{\kappa}_2$ and $\widehat{\kappa}_3$. The observed standard deviation for the control group is $\widehat{\kappa}_2 = 0.98$, and, based on the bootstrap, $\mathbb{P}(\widehat{\kappa}_2 < 1) = 0.65$. This suggests that there is little information in the second moment about the magnitude of Δ. The observed third sample moment for the control group is $\widehat{\kappa}_3 = 0.17$, with a bootstrap probability $\mathbb{P}(\widehat{\kappa}_3 < 0) = 0.12$. This suggests that the outcome distribution is sufficiently skewed such that the discontinuity, and hence the



(a) Bias for $\widehat{\Delta}^{\text{mle}}$ (b) 95% Coverage for $\widehat{\Delta}^{\text{mle}}$ (c) Probability of pathologies

Figure 7: Simulation results for the bias and coverage of the MLE for $N = 132$, and $\pi = 0.55$, $\sigma^2 = 1$, and assumed values of Δ. The thick black line shows the overall bias and coverage. The thinner lines show the bias and coverage results conditional on pathology.

a sign error, is not a major concern.

We fit a simple Normal mixture model to the data from the control group and find $\widehat{\Delta}^{\text{mle}} = 0.00$, which is consistent with the univariate results in Mattei et al. (2013).[14] We also estimate a 95% confidence interval of $[-0.9, 0.9]$. Given the evidence of pile up, our analysis suggests that we should interpret this point estimate with caution. Figure 9a shows the p-values from the grid bootstrap and from the Wald tests for Δ, which are nearly identical to each other. Like the MLE, this distribution is centered at $\Delta = 0$, with a 95% confidence set of $[-1.27, 1.27]$, roughly 40% wider than the corresponding nominal confidence interval. Figure 9b shows the corresponding confidence set for both $\mu_{\text{c}0}$ and $\mu_{\text{n}0}$, which is also centered at $\mu_{\text{c}0} = \mu_{\text{n}0}$.

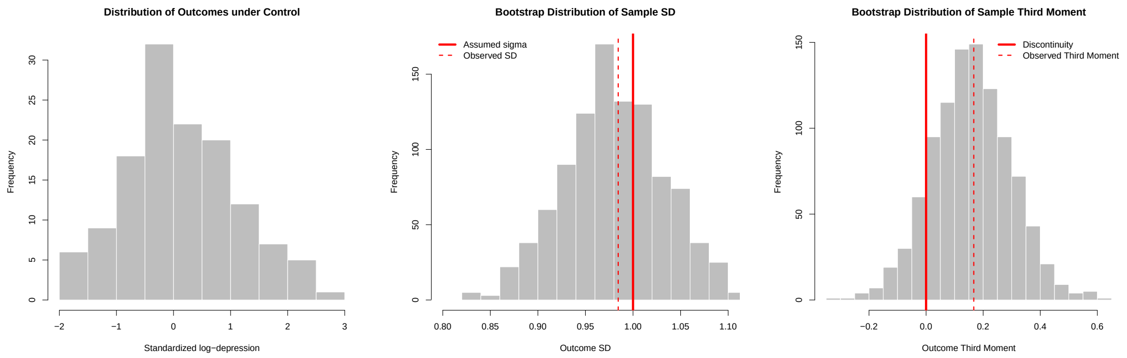
Finally, we can obtain confidence sets for ITT_{c} and ITT_{n}. First, the 97.5% confidence intervals for the outcomes under treatment are $\text{CS}_{0.025}(\mu_{\text{c}1}) = [-0.35, 0.03]$ and $\text{CS}_{0.025}(\mu_{\text{n}1}) = [-0.14, 0.24]$. Therefore, the confidence sets for the treatment effects are $\text{CS}_{0.05}(\text{ITT}_{\text{c}}) = [-1.21, 0.61]$ and $\text{CS}_{0.05}(\text{ITT}_{\text{n}}) = [-1.14, 0.95]$. These are considerably wider than the corresponding MLE confidence intervals, $\text{CS}^{\text{mle}}_{0.05}(\text{ITT}_{\text{c}}) = [-0.88, 0.27]$ and $\text{CS}^{\text{mle}}_{0.05}(\text{ITT}_{\text{n}}) = [-0.59, 0.42]$.

7 Extension: Unknown variance

We now return to the assumption that the component variances are equal, which might be unrealistic in practice (e.g., Gallop et al., 2009). There are two ways to relax this assumption. Consider the general two-component Gaussian mixture model:

$$Y_i \overset{\text{iid}}{\sim} \pi \mathcal{N}(\mu_0, \sigma_0^2) + (1 - \pi)\mathcal{N}(\mu_1, \sigma_1^2).$$

[14] 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 $\widehat{\mu}_{\text{c}1} = 1.96$ and $\widehat{\mu}_{\text{n}1} = 2.08$ on the depression scale. The treatment effect point estimates are $\widehat{\text{ITT}}_{\text{c}} = -0.206$ and $\widehat{\text{ITT}}_{\text{n}} = -0.084$, which imply $\widehat{\mu}_{\text{c}0} = 1.96 + 0.206 = 2.166$ and $\widehat{\mu}_{\text{n}0} = 2.08 + 0.084 = 2.164$. 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.

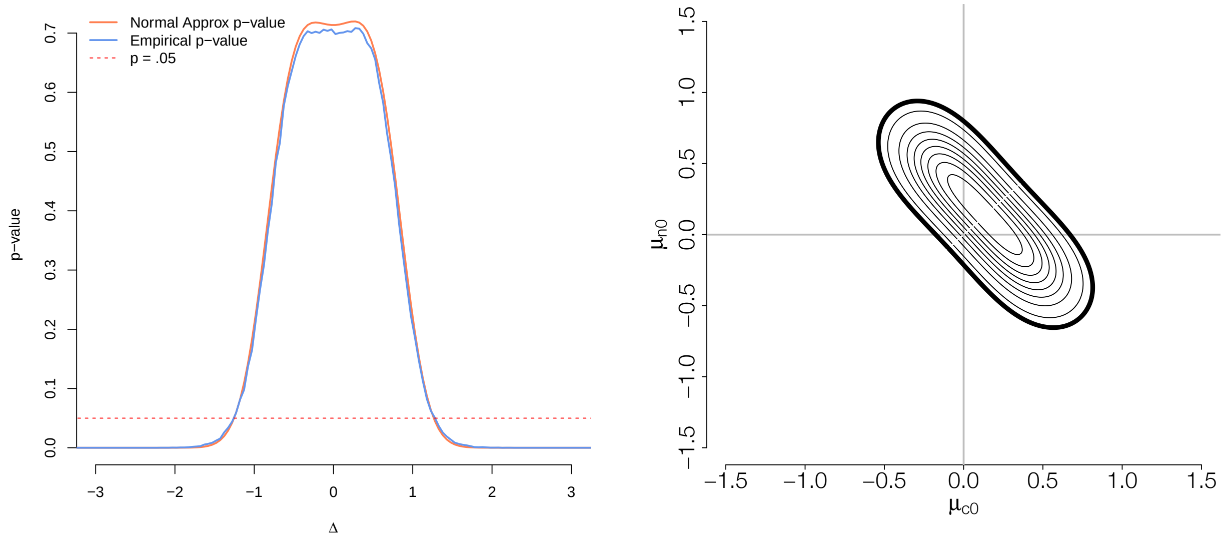


(a) Outcome dist.

(b) Bootstrap dist. for $\widehat{\kappa}_2$

(c) Bootstrap dist. for $\widehat{\kappa}_3$

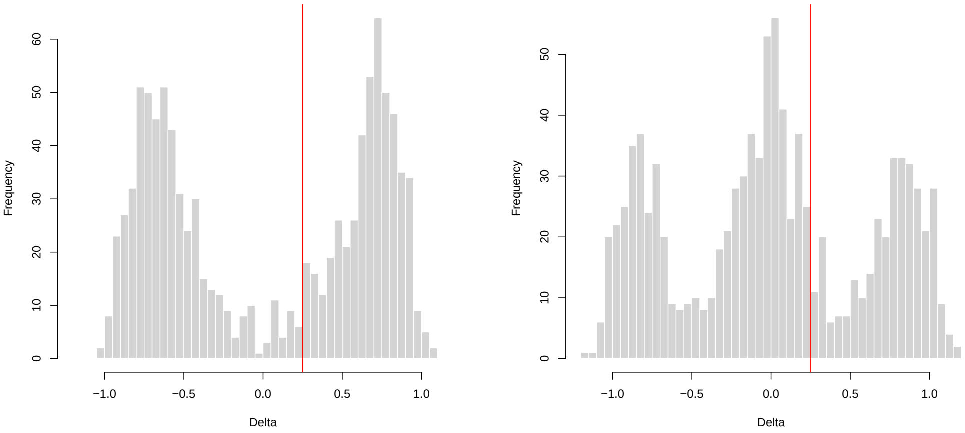
Figure 8: Overview of sample moments for individuals assigned to the control group in JOBS II.



(a) p-values for Δ

(b) p-values for μ_{c0} and μ_{n0}

Figure 9: p-values for parameters in the control group of JOBS II; p-values for Δ use both a grid bootstrap and Wald test; p-values for μ_{c0} and μ_{n0} use a Wald test only. The dark line indicates $p = 0.05$.



(a) Unknown, equal variance

(b) Unknown, unequal variance

Figure 10: Distribution of $\widehat{\Delta}^{\text{mle}}$ for $\Delta = 0.25$, $N = 5000$, $\pi = 0.325$, and $\sigma^2 = 1$. The red vertical line indicates the true value.

In the equal-variance case, we assume that $\sigma^2 = \sigma_0^2 = \sigma_1^2$ and that σ^2 is unknown. In the unequal-variance case, we assume that σ_0^2 and σ_1^2 are both unknown and are not constrained to be equal.

Simulation studies show that the performance of the MLE with unknown component variance continues to be poor. Interestingly, the sampling distribution of the MLE for Δ shows a different pattern from the known variance case; see Figure 10 for an example with $\Delta = 0.25$ and $N = 5000$. For the MLE with unknown, equal variance, there is no pile up. However, the bias is much more substantial than for the case with known, equal variance—the two modes are centered at around ± 0.75, roughly three times larger than the truth. For the MLE with unknown, different variances, the distribution of the MLE has a distinct trimodal distribution. The middle mode is centered at zero; the other two modes are similarly biased as in the case with unknown, equal variance. This pattern appears to persist across a broad range of parameter values.

We can gain intuition for the MLE with unknown, equal variance by deriving the method of moments analog. With known σ, we use κ_2 to estimate the magnitude of Δ and use κ_3 to estimate the sign of Δ. With unknown σ, we use κ_2 to estimate the variance itself and use κ_3 to estimate both the sign and magnitude of Δ:

$$\widehat{\Delta}^{\text{mom},\kappa_3} = \operatorname{sgn}\left(\widehat{\kappa}_3\right)\left(\frac{|\widehat{\kappa}_3|}{\pi(1-\pi)(1-2\pi)}\right)^{1/3}.$$

Since this estimator is well-defined for all values of $\widehat{\kappa}_3$, pile up is not a concern. However, the absolute value is still a discontinuity in the mapping from the sample distribution to the statistic. Therefore, we implicitly condition on $\widehat{\kappa}_3 > 0$, which leads to bias, since we are taking the mean of a truncated distribution.

Figure 11 shows the bias from the simulation studies for $\widehat{\Delta}^{\text{mle}}$ for cases with known and unknown σ, and with $\Delta = 0.25$ and $\pi = 0.325$. The bias with unknown variance is much more substantial than with known variance. As for the case with known variance, we can use LOTUS and numerical integration to

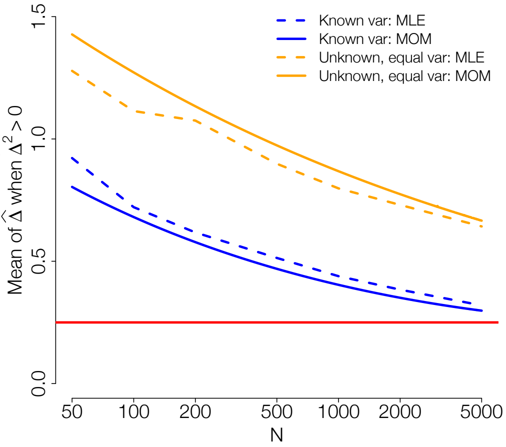


Figure 11: Mean for $\widehat{\Delta}^{\text{mle}}$ (simulated) and $\widehat{\Delta}^{\text{mom}}$ (calculated) when estimators are well-defined. The true value is $\Delta = 0.25$.

obtain the expected value of the truncated distribution, $\mathbb{E}[\widehat{\Delta}^{\text{mom},\kappa_3} \mid \widehat{\kappa}_3 > 0]$. This also appears to be a good approximation to the bias calculated via simulation.

Finally, we do not discuss the moment estimator for unknown, unequal variance. This uses the first four cumulants and does not have a form that gives any useful insight.

8 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.

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 follow Grilli and Mealli (2008) and recommend that researchers first compute nonparametric bounds (see also Zhang and Rubin, 2003; Lee, 2009).

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, such as in Figure 4. Analogous to design analysis, these calculations can provide practical guidance on whether mixture modeling will yield useful inference. In addition, confidence sets generated via test inversion are a meaningful check on the model-based results.

More optimistically, we agree that incorporating multiple outcomes, such as in Mattei et al. (2013), 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). Alternatively, prior information (e.g., Hirano et al., 2000; Stein, 2013) can improve inference, even if problems with the likelihood remain. Finally, extensive model checking (e.g., Zhang et al., 2009; Frumento et al., 2012) can increase the credibility of inference with finite mixtures.

Although we do not address them in depth here, covariates play a very central role in principal stratification in practice (e.g., Jo, 2002; Zhang et al., 2009; Zigler and Belin, 2011; Ding et al., 2011). In the supplementary materials, we briefly explore whether covariates can indeed improve inference in the settings we consider here. Perhaps surprisingly, our preliminary results suggest that covariates can actually make pathological results *worse* without additional restrictions. Thus, merely adding covariates to the model is not enough—the covariates must also be coupled with additional modeling assumptions. This is an important direction 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 beyond test inversion, particularly those that more directly leverage 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 D. B. Rubin (1985). Estimation and hypothesis testing in finite mixture models. *Journal of the Royal Statistical Society, Series B 47*(1), 67–75.

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

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

Balakrishnan, S., M. J. Wainwright, and B. Yu (2014). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Arxiv* 1408.2156.

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

Berger, R. L. and D. D. Boos (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 D. A. Freedman (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.

Blatt, D. and A. O. I. Hero (2007). On tests for global maximum of the log-likelihood function. *IEEE Transactions on Information Theory 53*(7), 2510–2525.

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

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

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

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

Compiani, G. and Y. Kitamura (2013). Using Mixtures in Econometric Models: A Brief Review and Some New Results.

Day, N. E. (1969). Estimating the components of a mixture of normal distributions. *Biometrika 56*(3), 463–474.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological) 39*(1), 1–38.

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

Ding, P., Z. Geng, W. Yan, and X.-H. Zhou (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., T. E. Raghunathan, and Y. Li (2010). Bayesian inference for causal mediation effects using principal stratification with dichotomous mediators and outcomes. *Biostatistics 11*(2), 353–372.

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

Frangakis, C. E. and D. B. Rubin (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., F. Mealli, B. Pacini, and D. B. Rubin (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., F. Mealli, B. Pacini, and D. B. Rubin (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.

Furman, W. D. and B. G. Lindsay (1994a). Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Computational Statistics & Data Analysis 17*, 493–507.

Furman, W. D. and B. G. Lindsay (1994b). Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics & Data Analysis 17*, 473–492.

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

Gan, L. and J. Jiang (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*, 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/`.

Goffinet, B., P. Loisel, and B. Laurent (1992). Testing in normal mixture models when the proportions are known. *Biometrika 79*(4), 842–846.

Grazian, C. and C. Robert (2015). Jeffreys priors for mixture estimation. *arXiv*, 1511.03145.

Griffin, B. A., D. F. McCaffrey, and A. R. Morral (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 F. Mealli (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 F. Leisch (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.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica 50*(4), 1029–1054.

Heinrich, P. and J. Kahn (2015). Minimax rates for finite mixture estimation. *arXiv:1504.03506 [math.ST]*.

Henry, M., Y. Kitamura, and B. Salanie (2014). Partial identification of finite mixtures in econometric models. *Quantitative Economics 5*(1), 123–144.

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

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

Hosmer Jr, D. W. (1973). A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of sample. *Biometrics 29*(4), 761–770.

Huang, M., R. Li, and S. Wang (2013). Nonparametric Mixture of Regression Models. *Journal of the American Statistical Association 108*(503), 929–941.

Huang, M. and W. Yao (2012). Mixture of Regression Models With Varying Mixing Proportions: A Semiparametric Approach. *Journal of the American Statistical Association 107*(498), 711–724.

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

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

Jasra, A., C. C. Holmes, and D. A. Stephens (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 E. A. Stuart (2009). On the use of propensity scores in principal causal effect estimation. *Statistics in Medicine 28*(23), 2857–2875.

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

Kiefer, N. M. (1978). Estimating mixtures of Normal distributions and switching regressions: comment. *Journal of the American Statistical Association 73*(364), 744–745.

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

Lindsay, B. (1989). Moment matrices: Applications in mixtures. *Annals of Statistics 17*(2), 722–740.

Lindsay, B. G. (1983). The geometry of mixture likelihoods: a general theory. *The Annals of Statistics 11*(1), 86–94.

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

Little, R. J. and L. H. Y. Yau (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., F. Li, and F. Mealli (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 D. Peel (2004). *Finite mixture models.* John Wiley & Sons.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 36*(3), 318–324.

Mealli, F. and D. B. Rubin (2002). Discussion of "Estimation of Intervention Effects with Noncompliance: Alternative Model Specifications" by Booil Jo. *Journal of Educational and Behavioral Statistics 27*(4), 411–415.

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., F. Li, and F. Mealli (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.

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.

Nolen, T. L. and M. G. Hudgens (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.

Page, L. C., A. Feller, T. Grindal, L. Miratrix, and M. A. Somers (2015). Principal stratification: A tool for understanding variation in program effects across endogenous subgroups. *American Journal of Evaluation 36*(4), 514–531.

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

Polymenis, A. and D. M. Titterington (1999). A note on the distribution of the likelihood ratio statistic for normal mixture models with known proportions. *Journal of Statistical Computation and Simulation 64*(2), 167–175.

Quandt, R. E. and J. B. Ramsey (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association 73*(364), 730–738.

Quinn, B. G., G. J. McLachlan, and N. L. Hjort (1987). A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. *Journal of the Royal Statistical Society, Series B 49*(3), 311–314.

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

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

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.

Rubin, D. B. and D. T. Thayer (1983). More on EM for ML factor analysis, with discussion. *Psychometrika 48*, 253–257.

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

Shephard, N. G. and A. C. Harvey (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 J. H. Stock (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica 65*(3), 557–586.

Stein, N. (2013). *Advances in Empirical Bayes Modeling and Bayesian Computation*. Ph. D. thesis, Harvard University.

Tan, W. Y. and W. Chang (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., M. R. Elliott, M. Joffe, E. Zanutto, and C. Datto (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., A. Smith, and U. Makov (1985). *Statistical Analysis of Finite Mixture Distributions.* John Wiley & Sons.

Titterington, D. M. (2004). *Mixture Distributions – II.* John Wiley & Sons.

Vinokur, A. D., R. H. Price, and Y. Schul (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/`.

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

Zhang, J. L., D. B. Rubin, and F. Mealli (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 T. R. Belin (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.

A Appendix

A.1 Coverage for confidence sets via test inversion

Figure A.1 shows the coverage probabilities for 95% confidence sets based on the test inversion algorithm described in Section 5. Note that the coverage of the grid bootstrap intervals are practically exact and dramatically outperform the corresponding “naive” confidence intervals for the MLE.

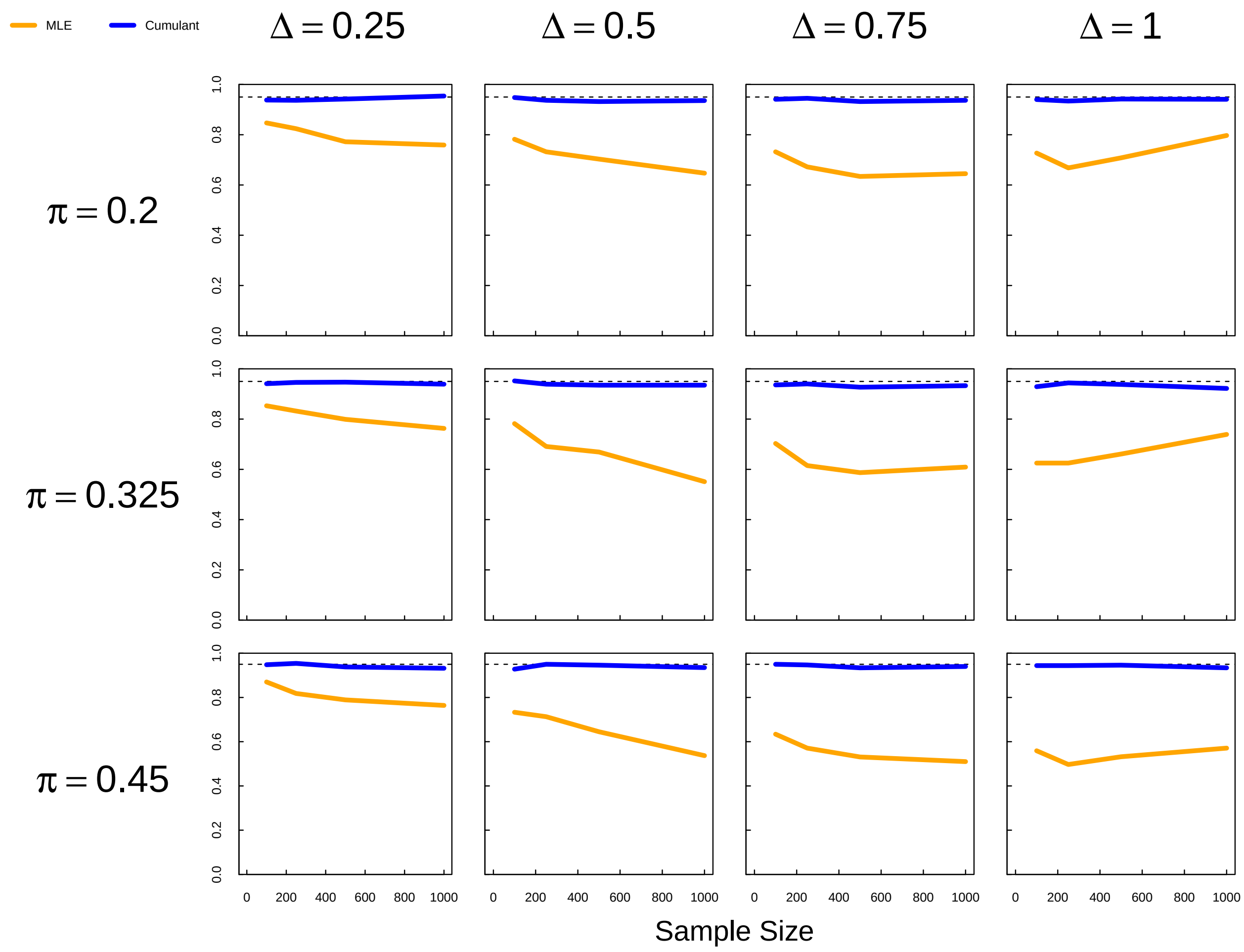


Figure A.1: Coverage for 95% confidence sets based on the test inversion algorithm described in Section 5. The results for the MLE are for the standard finite mixtures estimator, as in Figure 3.

A.2 Asymptotic Results

We prove Theorem 3.1. Our proof relies on a Taylor expansion similar to that used by Chen in the proof of Proposition 1 of Chen (1995). For a longer, more detailed proof see the Supplementary Materials.

Proof. We consider the case in which $\sigma^2 = 1$. All results derived below will be applicable to the homoeskedastic case by scaling the $Y_{i,n}$ by σ. Our proof relies on Taylor expansions of the sequence of log-likelihoods

$$l_n(\Delta_n) = p(Y_{1,n}, \ldots, Y_{n,n} \mid \Delta_n).$$

Let $c = \pi/(1-\pi)$ and $f(y, \mu)$ be the Gaussian kernel with mean μ and variance 1 evaluated at y. The first derivative of the log-likelihood for Δ_n is

$$l_n'(\Delta_n) = \sum_{i=1}^{n} \frac{-\pi f'(y_{i,n}, -\Delta) + c(1-\pi)f'(y_{i,n}, c\Delta)}{\pi f(y_{i,n}, -\Delta) + (1-\pi)f(c\Delta)}. \tag{A.1}$$

Note that the requirement on c for $l_n'(0) = 0$ is $c = \frac{\pi}{1-\pi}$. Thus, $l_n'(0) = 0$ regardless of the observed data. The second order derivative of the log-likelihood is

$$l_n''(\Delta_n) = \sum_{i=1}^{n} \frac{\pi f''(y_{i,n}, -\Delta_n) + c^2(1-\pi)f''(y_{i,n}, c\Delta_n)}{\pi f(y_{i,n}, -\Delta_n) + (1-\pi)f(c\Delta_n)} - \left(\frac{-\pi f'(y_{i,n}, -\Delta_n) + c(1-\pi)f'(y_{i,n}, c\Delta_n)}{\pi f(y_{i,n}, -\Delta_n) + (1-\pi)f(c\Delta_n)}\right)^2. \tag{A.2}$$

The second term in Equation G.7 vanishes when $\Delta_n = 0$, leaving

$$l_n''(0) = \frac{(\pi + c^2(1-\pi))f''(y_{i,n}, 0)}{f(y_{i,n}, 0)}, \tag{A.3}$$

which has expectation 0 when $\Delta_n = 0$. Thus, the Fisher Information of Δ_n when $\Delta_n = 0$ is 0. To be concise, we state the third and fourth derivatives of the likelihood evaluated at $\Delta_n = 0$,

$$l_n'''(0) = \sum_{i=1}^{n} \frac{(-\pi + c^3(1-\pi))f'''(y_{i,n}, 0)}{f(y_{i,n}, 0)} \tag{A.4}$$

and

$$l_n^{(4)}(0) = \sum_{i=1}^{n} \frac{(\pi + c^4(1-\pi))f^{(4)}(y_{i,n}, 0)}{f(y_{i,n}, 0)} - 3(\pi + c^2(1-\pi))^2 \left(\frac{f''(y_{i,n}, 0)}{f(y_{i,n}, 0)}\right)^2. \tag{A.5}$$

We define $A_{i,n}$, $B_{i,n}$, and $C_{i,n}$ analogously to Chen (1995):

$$A_{i,n} = \frac{f''(Y_{i,n}, 0)}{f(Y_{i,n}, 0)}, \quad B_{i,n} = \frac{f^{(3)}(Y_{i,n}, 0)}{f(Y_{i,n}, 0)}, \quad C_{i,n} = \frac{f^{(4)}(Y_{i,n}, 0)}{f(Y_{i,n}, 0)}. \tag{A.6}$$

We compute these terms explicitly for the Normal kernel with variance 1:

$$\begin{aligned} A_{i,n} &= Y_{i,n}^2 - 1, \\ B_{i,n} &= -3Y_{i,n} + Y_{i,n}^3, \\ C_{i,n} &= 3 - 6Y_{i,n}^2 + Y_{i,n}^4. \end{aligned}$$

Using the moments of $Y_{i,n}$ and the fact that $Y_{1,n}, \ldots, Y_{n,n}$ are i.i.d.we have that

$$\begin{aligned} \sum_{i=1}^{n} A_{i,n} &= O_p\left(\max\{n\Delta_n^2, n^{1/2}\}\right), \\ \sum_{i=1}^{n} B_{i,n} &= O_p\left(\max\{n\Delta_n^3, n^{1/2}\}\right), \\ \sum_{i=1}^{n} C_{i,n} &= O_p\left(\max\{n\Delta_n^4, n^{1/2}\}\right). \end{aligned}$$

We can now Taylor expand $l_n(\Delta)$:

$$l_n(\Delta_n) = l_n(0) + a\Delta_n^2 \sum_{i=1}^n A_{i,n} + b\Delta_n^3 \sum_{i=1}^n B_{i,n} + c\Delta_n^4 \sum_{i=1}^n A_{i,n}^2 + O_p(n^{1/2}\Delta_n^4), \tag{A.7}$$

where a, b, and c are constants. Taking the derivative and removing the root at 0,[15] we arrive at

$$\widehat{\Delta}_n = \left[-3b\sum_{i=1}^n B_{i,n} \pm \left(\left(3b\sum_{i=1}^n B_i \right)^2 - 32ac\sum_{i=1}^n A_{i,n} \sum_{i=1}^n A_{i,n}^2 \right)^{1/2} \right] \left[8c\sum_{i=1}^n A_{i,n}^2 \right]^{-1} (1 + o_p(1)) = O_p(n^{-1/4}), \tag{A.8}$$

since $\sum_{i=1}^n A_{i,n}^2 = O_p(n)$. Further, for $\widehat{\Delta}_n$ to be $o_p(n^{-1/4})$, $\sum_{i=1}^n A_{i,n} = o_p(n^{1/2})$, which is impossible as $A_{1,n}, \ldots, A_{n,n}$ are i.i.d. with non-zero variance. It follows that if $\Delta_n = o(n^{-1/4})$ we have

$$\left| \widehat{\Delta}_n - \Delta_n \right| = O_p(n^{-1/4})$$

and

$$\left| \widehat{\Delta}_n - \Delta_n \right| \neq o_p(n^{-1/4}).$$

□

[15] This root has probability < 1 of being a maximum, a result that comes from its dependence on the sign of $\sum_{i=1}^n A_{i,n}$.

Supplemental Materials

B Validating the Normal approximations

We present figures testing the correspondence of the method of moment indicators to their corresponding pathologies. Figure B.2 compares the incidence of pile up and $\hat{\kappa}_2 < \sigma^2$ 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 ($1\{\hat{\kappa}_2 < \sigma^2\}$) 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 method of moments indicator provides an excellent estimator for whether pile up has occured in the sample.

Figure B.3 shows the corresponding plots for the method of moments indicator of sign error ($1\{\text{sgn}(\Delta) \neq \text{sgn}(\hat{\kappa}_3)\}$). Here, due to the extra noise in κ_3, the correspondence is much less sharp. The discrepancies are most noticeable when π is close to 0 and Δ is small. However, as we see in the main text, we can still leverage the joint distribution of $(\hat{\kappa}_2, \hat{\kappa}_3)$ to derive confidence sets for Δ in practice.

C Additional discussion for principal stratification models

C.1 Incorporating covariates

In practice, causal inference researchers typically have access to a vector of pre-treatment covariates. Zhang et al. (2009), for example, argue that covariates play two main roles in principal stratification. First, they "generally improve the precision of parameter estimates because they improve the prediction of the missing potential outcomes." That is, covariates are predictive of Y, U, or both. Second, "covariates generally make assumptions more plausible, because they are conditional rather than marginal." That is, the assumption that component densities follow a Normal distribution is often more reasonable conditional on many covariates than unconditionally.[16]

We agree that the relevant modeling assumptions are often more reasonable conditionally. However, when components are only weakly separated, we find that incorporating covariates in an unrestricted way not only fails to improve precision in general, but actually increases the probability of obtaining a pathological result. Arguably, covariates are most useful in this setting when combined with conditional-independence-type assumptions, as we discuss below (Jo, 2002; Mealli and Rubin, 2002).

C.1.1 Covariates without restrictions

To illustrate the key concepts, we explore the case with a single binary covariate, X. As in the no-covariates case, the immediate goal is to estimate the overall difference in means, Δ. The idea is to improve precision by estimating the component means conditional on X, Δ_x, and then combining. Without restrictions, this simply doubles the number of equations and unknowns we must estimate.

$$
\begin{aligned}
Y_i \mid X_i = 1 &\overset{\text{iid}}{\sim} \pi_1 \mathcal{N}(\mu_{0|1}, \sigma^2) + (1 - \pi_1)\mathcal{N}(\mu_{1|1}, \sigma^2) \\
Y_i \mid X_i = 0 &\overset{\text{iid}}{\sim} \pi_0 \mathcal{N}(\mu_{0|0}, \sigma^2) + (1 - \pi_0)\mathcal{N}(\mu_{1|0}, \sigma^2)
\end{aligned}
$$

[16] There have been extensive discussions on the use of covariates in principal stratification models. See, for example, Little and Yau (1998), Hirano et al. (2000), Jo (2002); Jo and Stuart (2009), Zhang et al. (2009), Ding et al. (2011), Richardson et al. (2011), and Zigler and Belin (2011).

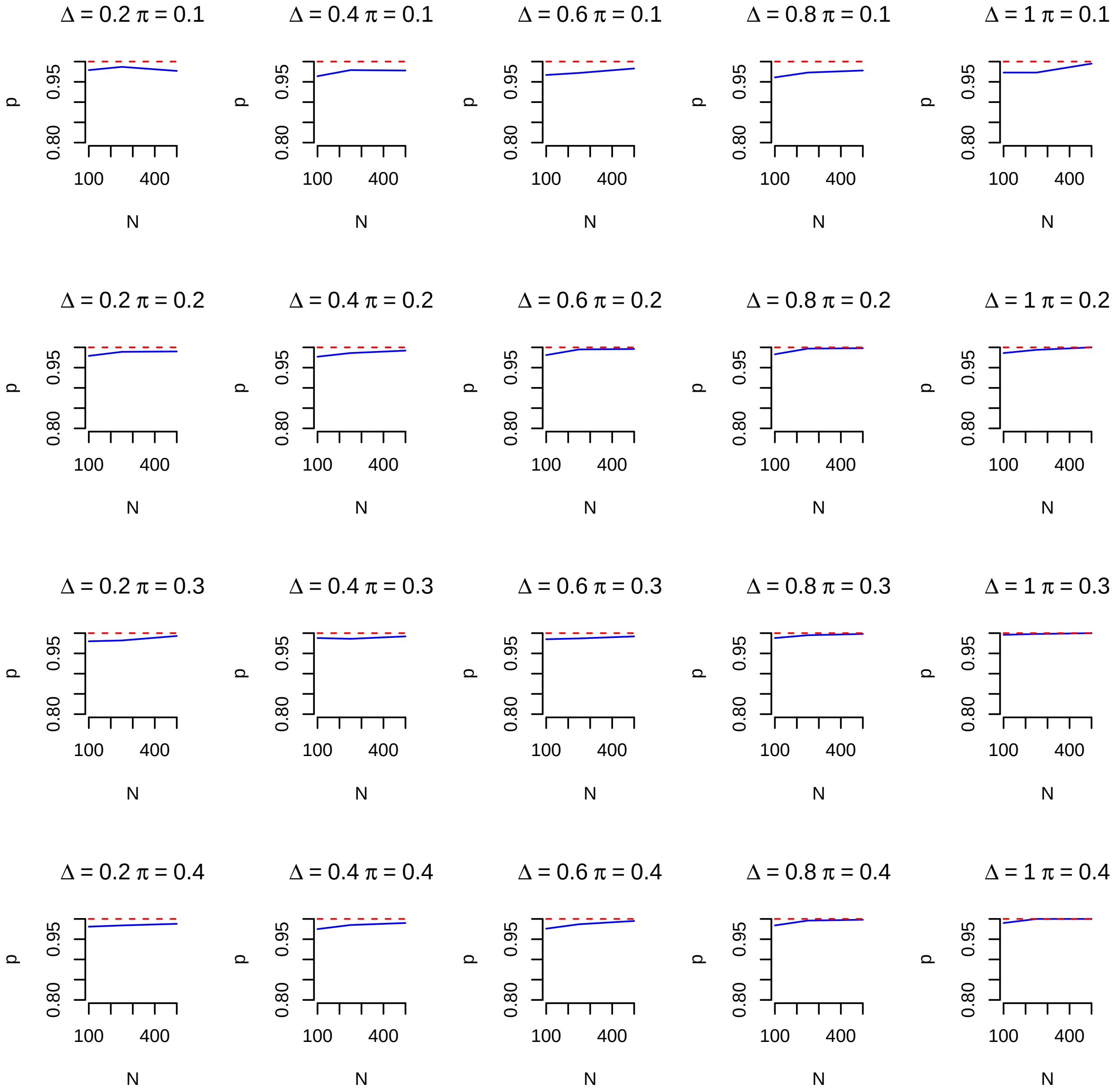


Figure B.2: Probability that the method of moments indicator of pile up ($1\{\hat{\kappa}_2 < \sigma^2\}$) agrees with whether or not pile up was observed in simulation. The dotted red line indicates 1 on the y-axis, while the blue line indicates the average agreement probability over 1000 simulated data sets.

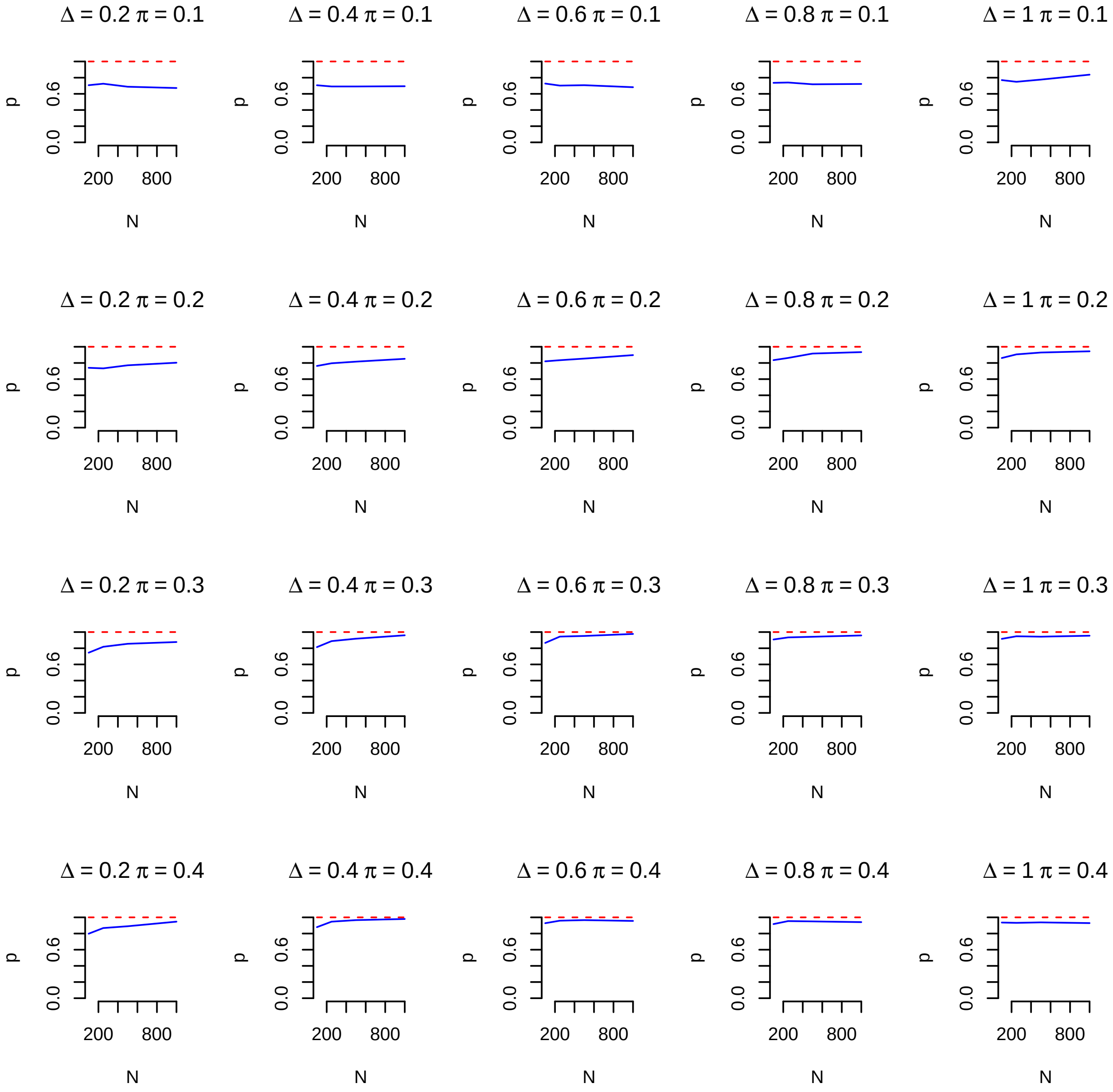


Figure B.3: Probability that the method of moments indicator of wrong sign $(1\{\mathrm{sgn}(\hat{\kappa}_3) \neq \mathrm{sgn}(\Delta)\})$ agrees with whether or not the wrong sign pathology was observed in simulation. The dotted red line indicates 1 on the y-axis, while the blue line indicates the average agreement probability over 1000 simulated data sets.

where $\pi_0, \pi_1 \in (0, 1/2)$. Letting $\Delta_1 = \mu_{0|1} - \mu_{1|1}$ and $\Delta_0 = \mu_{0|0} - \mu_{1|0}$, the moment equations become

$$\begin{aligned} \mathbb{E}[Y_i \mid X_i = 1] &= \mu_{1|1} + \pi_1 \Delta_1, && \text{(C.1)} \\ \mathbb{E}[Y_i \mid X_i = 0] &= \mu_{1|0} + \pi_0 \Delta_0, && \text{(C.2)} \\ \mathrm{Var}[Y_i \mid X_i = 1] &= \sigma^2 + \pi_1(1 - \pi_1)\Delta_1^2, && \text{(C.3)} \\ \mathrm{Var}[Y_i \mid X_i = 0] &= \sigma^2 + \pi_0(1 - \pi_1)\Delta_0^2. && \text{(C.4)} \end{aligned}$$

We can estimate Δ_1 via equations C.1 and C.3, and Δ_0 via equations C.2 and C.4. If we define $p = \mathbb{P}\{X_i = 1\}$, the population proportion that $X_i = 1$, then we can write the overall Δ as $\Delta = p\Delta_1 + (1 - p)\Delta_0$. The adjusted estimator for overall Δ is then:

$$\widehat{\Delta}^{\text{adj}} = p\widehat{\Delta}_1 + (1 - p)\widehat{\Delta}_0. \tag{C.5}$$

Unsurprisingly, the MLE for the conditional difference in means, Δ_0 and Δ_1, exhibits the same pathologies as the MLE for the overall Δ. We are simply fitting the same Gaussian mixture model twice on two subsets of the data. In each case, the smaller sample size will tend to make the probability of a pathology conditional on X higher than the unconditional probability. Table 2 gives an example for an $N_0 = N_1 = 500$, $\pi_0 = 0.45, \pi_1 = 0.15, \Delta_0 = 1.0, \Delta_1 = 0.5$, and $\sigma^2 = 1$. The table shows the probability of each of the nine pathology cases, with the rows representing the pathology status of $\widehat{\Delta}_0$, the columns representing the pathology status of $\widehat{\Delta}_1$. In this setup, there is only a 31% probability that $\widehat{\Delta}_0$ and $\widehat{\Delta}_1$ are non-zero and the correct sign, half the probability for the unconditional case.

Table 2: Joint pathology probabilities for the case of a binary covariate, X_i.

		$X_i = 1$		
		No Pathology	Pile Up	Wrong Sign
	No Pathology	31%	18%	9%
$X_i = 0$	Pile Up	0%	0%	0%
	Wrong Sign	23%	13%	6%

Of course, in some extreme cases, conditioning on X can improve inference. For example, with sufficiently large samples, conditioning on X can be beneficial if the overall Δ is close to 0, but the conditional values of Δ_x are relatively large in magnitude. In practice, however, it is difficult to know if these conditions indeed hold.

C.1.2 Covariates with restrictions

There is an extensive literature on incorporating covariates into mixture models (e.g., Compiani and Kitamura, 2013; Henry et al., 2014). Particularly relevant for our application are mixtures of regression models, such as in Huang and Yao (2012) and Huang et al. (2013). We focus here on some proposals in the context of principal stratification. The most straightforward is to assume that X is predictive of stratum membership, S, but is not predictive of Y conditional on S. This yields a standard instrumental variable-type estimator. See Henry et al. (2014) for a recent discussion.

$$\begin{aligned} Y_i \mid X_i = 1 &\overset{\text{iid}}{\sim} \pi_1 \mathcal{N}(\mu_0, \sigma^2) + (1 - \pi_1)\mathcal{N}(\mu_1, \sigma^2) \\ Y_i \mid X_i = 0 &\overset{\text{iid}}{\sim} \pi_0 \mathcal{N}(\mu_0, \sigma^2) + (1 - \pi_0)\mathcal{N}(\mu_1, \sigma^2) \end{aligned}$$

With this assumption, $\Delta_0 = \Delta_1$, and we only need first moments to estimate Δ:

$$\begin{aligned} \mathbb{E}[Y_i \mid X_i = 1] &= \mu_0 + \pi_1 \Delta, && \text{(C.6)} \\ \mathbb{E}[Y_i \mid X_i = 0] &= \mu_0 + \pi_0 \Delta. && \text{(C.7)} \end{aligned}$$

Solving Equations C.6 and C.7 leads to

$$\widehat{\Delta}^{\text{iv}} = \frac{m_{1|1} - m_{1|0}}{\pi_1 - \pi_0}, \tag{C.8}$$

where $m_{1|x}$ denotes the first sample moment for $\{Y_i : X_i = x\}$. So long as $\pi_1 - \pi_0$ is large, $\widehat{\Delta}^{\text{iv}}$ does not exhibit the pathologies of the moment estimator without covariates. In this situation, the MLE that jointly models the mixtures conditional on $X_i = 0$ and on $X_i = 1$ is generally well-behaved. Heuristically, the MLE automatically incorporates different "identifying" sources of information. In this case, the conditional independence of X and Y is much more informative than the Normal mixture structure. Therefore, the resulting MLE is well-approximated by $\widehat{\Delta}^{\text{iv}}$ rather than by $\widehat{\Delta}^{\text{mom}}$.

In practice, it is rare to find a covariate that is both predictive of type and conditionally independent of the outcome. There are many possible relaxations that leverage the broader principal stratification structure; namely, that we can observe the relationship between X and Y in the other treatment arm. One assumption in the case of one-sided noncompliance, due to Jo (2002), is to assume that, given S, the treatment effect does not vary by X. Many similar assumptions are possible (see, for example Mealli and Rubin, 2002; Ding et al., 2011). Alternative approaches include *principal score* methods, which assume that, conditional on X, type and outcome are independent (Jo and Stuart, 2009). Importantly, such assumptions provide identifying information so that the resulting MLE does not need to rely on higher-order moments to estimate component means.

C.2 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, $\widehat{\mu}_{\text{c}1}$ and $\widehat{\mu}_{\text{n}1}$, and use the finite mixture model to estimate corresponding outcome means for Compliers and Never Takers assigned to control, $\widehat{\mu}_{\text{c}0}$ and $\widehat{\mu}_{\text{n}0}$. Our goal is inference for $\text{ITT}_{\text{c}} = \widehat{\mu}_{\text{c}1} - \widehat{\mu}_{\text{c}0}$ and $\text{ITT}_{\text{n}} = \widehat{\mu}_{\text{n}1} - \widehat{\mu}_{\text{n}0}$. While this is straightforward given estimates for $\mu_{\text{c}0}$ and $\mu_{\text{n}0}$, 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 $\mu_{\text{c}0}$ and $\mu_{\text{n}0}$, which we can project into univariate confidence sets, $\text{CS}_{\alpha/2}(\mu_{\text{c}0})$ and $\text{CS}_{\alpha/2}(\mu_{\text{n}0})$
- Directly obtain $(1-\alpha/2)100\%$ confidence intervals via the Normal distribution for $\mu_{\text{c}1}$ and $\mu_{\text{n}1}$, $\text{CS}_{\alpha/2}(\mu_{\text{c}1})$ and $\text{CS}_{\alpha/2}(\mu_{\text{n}1})$
- For ITT_{c} (repeat for ITT_{n}):
  - If $\text{CS}_{\alpha/2}(\mu_{\text{c}0})$ is not disjoint, obtain a $(1-\alpha)100\%$ confidence interval for ITT_{c}:

$$\text{CS}^{UB}_{\alpha}(\text{ITT}_{\text{c}}) = \text{CS}^{UB}_{\alpha/2}(\mu_{\text{c}1}) - \text{CS}^{LB}_{\alpha/2}(\mu_{\text{c}0})$$
$$\text{CS}^{LB}_{\alpha}(\text{ITT}_{\text{c}}) = \text{CS}^{LB}_{\alpha/2}(\mu_{\text{c}1}) - \text{CS}^{UB}_{\alpha/2}(\mu_{\text{c}0})$$

  - If $\text{CS}_{\alpha/2}(\mu_{\text{c}0})$ 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 in 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.

D 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 D.4 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,

$$\widehat{\Delta}^{\text{mom}} = \text{sgn}(\widehat{\kappa}_3)\sqrt{\frac{\widehat{\kappa}_2 - \sigma^2}{\pi(1-\pi)}}$$

is not a continuous mapping from the sample to $\widehat{\Delta}^{\text{mom}}$, with a boundary at $\kappa_2 \geqslant \sigma^2$ and a discontinuity at $\kappa_3 = 0$.[17] 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.

E Performance of grid bootstrap in misspecified models

As noted in Gelman (2011), intervals based off inverting hypothesis tests fail when the model is significantly misspecified. There are two main ways the assumptions of the Gaussian mixture examined in the paper can break down. First, the assumption of Gaussian tails can be incorrect. Second, the assumption of uni- or bi-modality can be violated. Since problems are typical of misspecified models in many problems, we examine these effects only briefly by simulating data from a mixture of t-distributions and a three-component Gaussian mixture and finding the grid bootstrap intervals for this simulated data. As shown in Figure E.5 for the case of an underlying t-distribution, we see that the coverage of these intervals deteriorates as the true distribution deviates from that which can be accurately approximated by a two-component Gaussian mixture.

In particular, as the degree of freedom of the t-distribution decreases, approaching a Cauchy, the coverage drops to 0 as the observed cumulants are extremely unlikely to have been generated from the assumed model. As the degree of freedom of the t-distribution increases, the t-distribution approaches a Gaussian, causing coverage to improve to exactness.

In the case of the three-component Gaussian mixture, Δ is no longer estimable, as it does not have meaning in context of the underlying model. Intuitively, the grid bootstrap interval will be nearly empty when the underlying distribution is trimodal and thus difficult to match with a two-component Gaussian mixture. In this case, the measure of the grid bootstrap interval could be thought of as a diagnostic of a misspecified model. However, we advise that more robust measures of model fit, such as likelihood ratio

[17] 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.

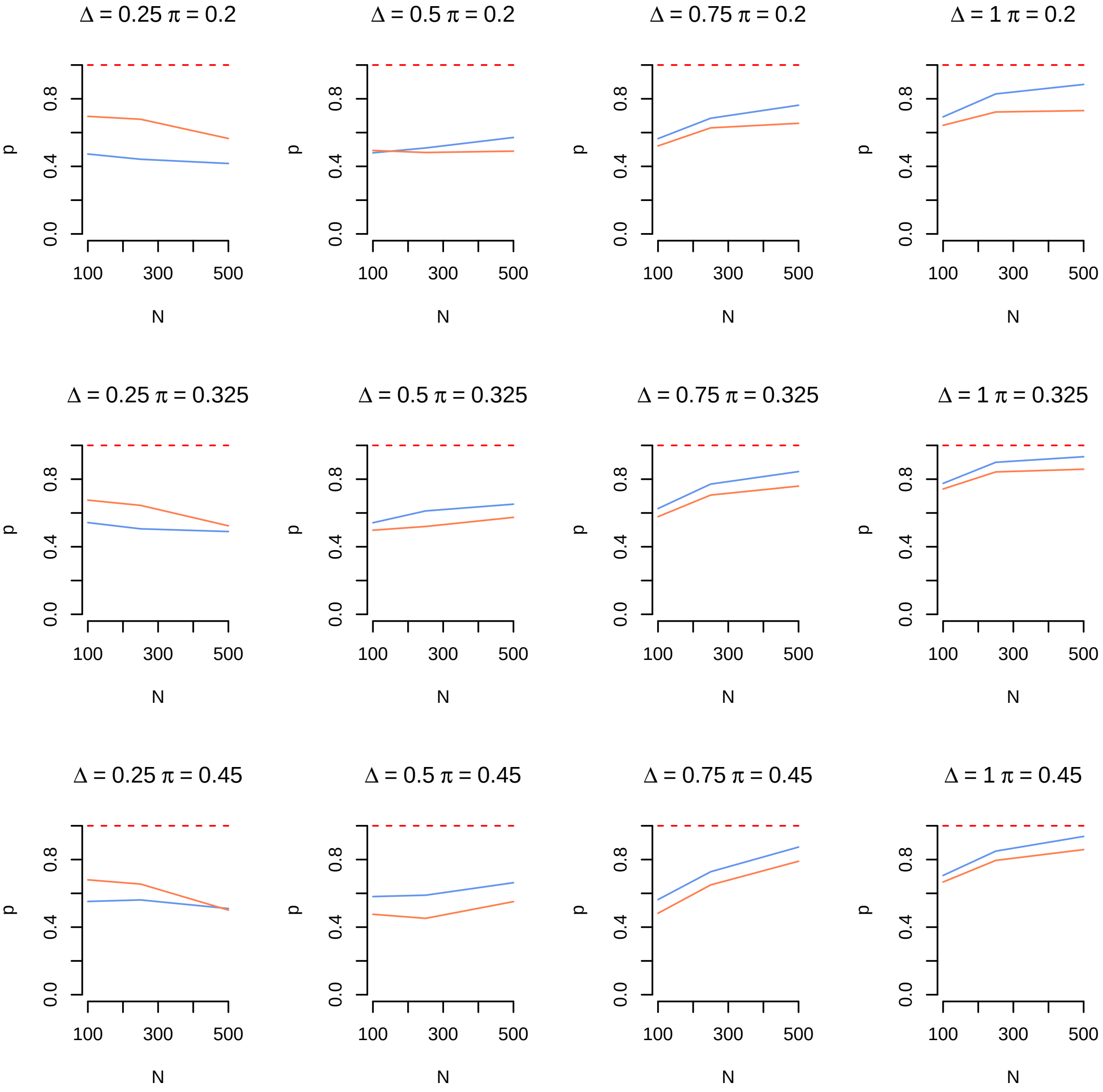


Figure D.4: Coverage probabilities for 95% confidence sets based on the case-resampling and subsampling intervals. The blue line represents the case-resampling coverage probability, while the blue line represents the subsampling coverage probability.

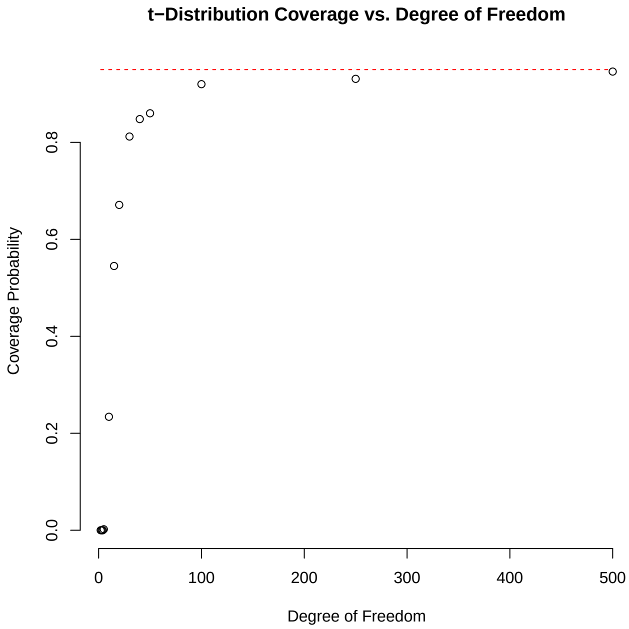


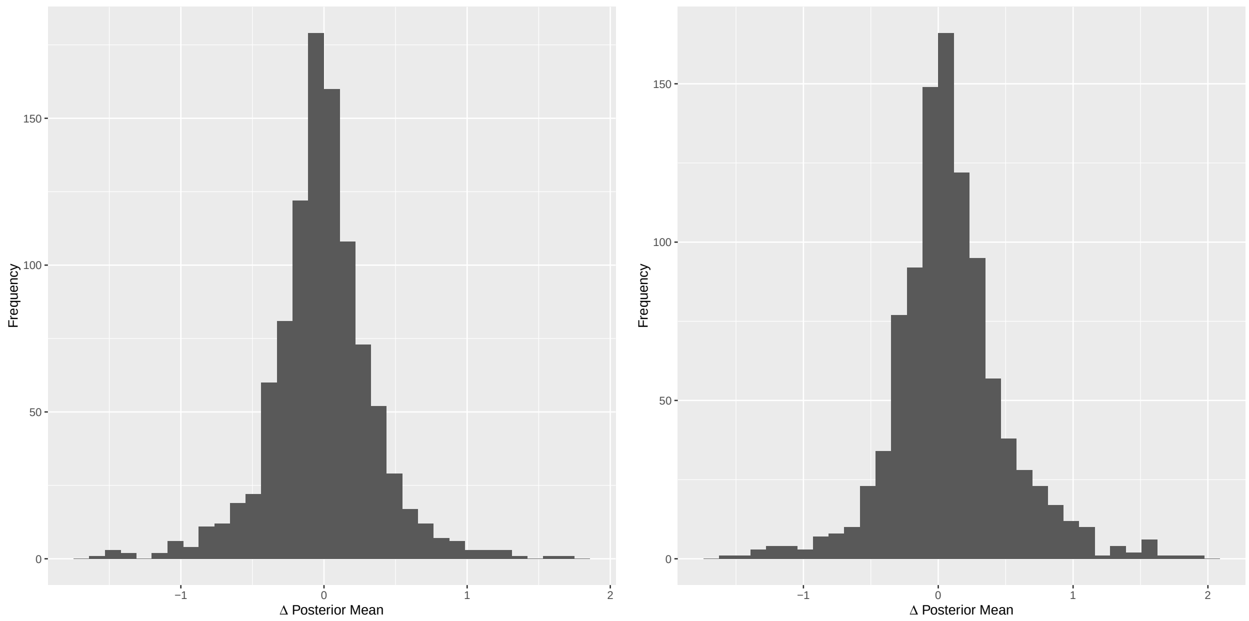
Figure E.5: Coverage of $\Delta = \mu_0 - \mu_1$ for various degrees of freedom of the (true) underlying t-distribution, averaged over 1000 Monte Carlo simulations.

tests for the number of components of the mixture (Lindsay, 1989) be used when determining whether or not to proceed with the grid bootstrap interval.

F Performance of Posterior Mean and Median

Bayesian inference for finite mixtures introduces some unique challenges for specifying priors (e.g., Grazian and Robert, 2015). Nonetheless, heuristically, inference based on the likelihood alone should be similar to inference for a posterior with a sufficiently vague prior. 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.6 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 $\Delta < 0$.

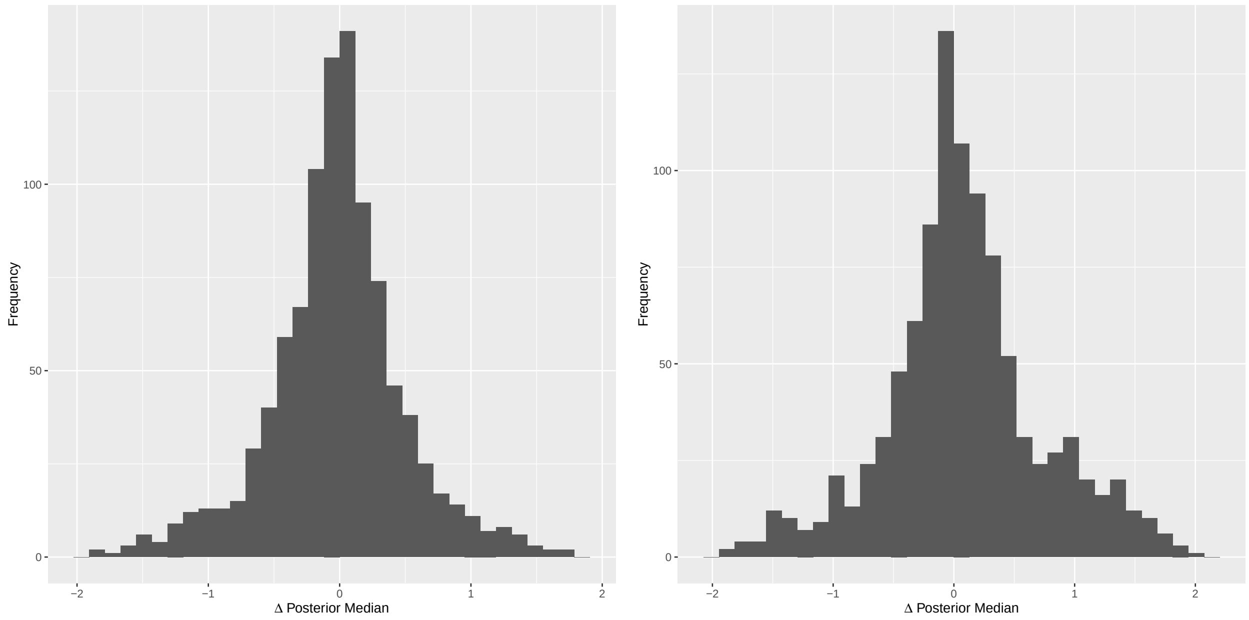
Figure F.7 shows the corresponding plot for the distribution of the posterior median of Δ. As we can see, the median also concentrates about 0 and appears unable to determine the sign of Δ.



(a) $\Delta = 0.5$

(b) $\Delta = 1$

Figure F.6: 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 $\sigma = 1$.



(a) $\Delta = 0.5$

(b) $\Delta = 1$

Figure F.7: 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 Asymptotics

The goal of this section is to illustrate the proof of Proposition 1 given in Chen (1995), and our most basic extension of that proof. We'll start by restating the definitions for $O(\cdot)$, $o(\cdot)$, $O_p(\cdot)$, and $o_p(\cdot)$.

We start with the non-probabilistic definitions of $O(\cdot)$ and $o(\cdot)$.

Definition 1. *Let $f(n)$ and $g(n)$ be functions for $n \in \mathbb{N}$. We say that $f = O(g)$, if there exists a constant $c < \infty$ such that*

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} < c.$$

Definition 2. *Let $f(n)$ and $g(n)$ be functions for $n \in \mathbb{N}$. We say that $f = o(g)$ if*

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} = 0.$$

We now state the probabilistic definitions of $O_p(\cdot)$ and $o_p(\cdot)$.

Definition 3. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables and $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of constants. We say that $X_n = O_p(a_n)$ if the collection of random variables $\{X_n/a_n\}_{n\in\mathbb{N}}$ is stochastically bounded.*

Definition 4. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables and $\{a_n\}_{n\in\mathbb{N}}$ be a sequence of constants. We say that $X_n = o_p(a_n)$ if*

$$\lim_{n\to\infty} X_n/a_n = 0$$

in probability.

G.1 Results

Here we present the two results of interest: Chen's result (Proposition 1 from his 1995 paper) and our basic extension of that result. The general setup is a two-component mixture model with generic location density $f(y, \mu)$ having mean μ. That is, we view Y as arising from

$$Y \sim \pi f(y, \mu_0) + (1 - \pi) f(y, \mu_1),$$

with $\pi \in (0, 1)$. Chen respecifies the model[18] as

$$Y \sim \pi f(y, -\Delta) + (1 - \pi) f(y, c\Delta), \tag{G.1}$$

where c is a constant. Note that

$$\mathbb{E}[Y] = -\pi\Delta + c(1 - \pi)\Delta,$$

so that the requirement that Y has mean 0 is equivalent to

$$c = \frac{\pi}{1 - \pi}.$$

The following proposition is equivalent (up to notation) to Proposition 1 of Chen (1995).

[18] Note that Chen uses h in place of Δ and specifies $c = 2$.

Proposition 1. *Let $Y_1, \ldots, Y_n$ be i.i.d. from Equation G.1 with $\pi = \frac{2}{3}$, $c = 2$, and $\Delta = 0$. That is, Y is assumed to come from the model*

$$Y \sim \frac{2}{3} f(y, -\Delta) + \frac{1}{3} f(y, 2\Delta),$$

for $\Delta = 0$. Further assume the density function $f(y, \mu)$ satisfies three regularity conditions.

1.
$$\mathbb{E}\left|\frac{f^{(i)}(Y,\mu)}{f(Y,\mu)}\right|^2 < \infty \tag{G.2}$$

   for $i \in \{2, 3, 4\}$.

2. *There exists a function $g(y)$ such that*
$$\left|\frac{f^{(4)}(Y,\mu_1)}{f(Y,\mu_1)} - \frac{f^{(4)}(Y,\mu_2)}{f(Y,\mu_2)}\right| \leqslant g(Y)|\mu_1 - \mu_2|^{\epsilon}, \tag{G.3}$$

   for some $\epsilon > 0$.

3.
$$\mathbb{E}[g^2(Y)] < \infty. \tag{G.4}$$

Then $\widehat{\Delta}_n$, the MLE of Δ estimated from $Y_1, \ldots, Y_n$, is $O_p(n^{-1/4})$.

We now describe our basic extension of Chen's result.

Proposition 2. *Let Y be as in Equation G.1 but with $\pi = \frac{2}{3}$ and $c = 2$. However, instead of letting Δ be fixed at 0, substitute $\Delta_n = o(n^{-1/4})$. That is, let $Y_{i,n}$ for $i \in \{1, \ldots, n\}$ and $n \in \mathbb{N}$ come from the model*

$$Y_{i,n} \sim \frac{2}{3}\mathcal{N}(y, -\Delta_n) + \frac{1}{3}\mathcal{N}(y, 2\Delta_n), \tag{G.5}$$

with the variance of the Normal kernels fixed at 1.[19] *Then $|\widehat{\Delta}_n - \Delta_n|$ is $O_p(n^{-1/4})$ and $|\widehat{\Delta}_n - \Delta_n|$ is not $o_p(n^{-1/4})$, where $\widehat{\Delta}_n$ is the maximum-likelihood estimator of Δ_n based on the data $Y_{1,n}, \ldots, Y_{n,n}$.*

G.2 Proofs

This section presents the proofs of the previous results. We start with an abbreviated version of Chen's proof, which relies on a Taylor expansion of the log-likelihood of Δ around $\Delta = 0$.

Proof. The first derivative of the log-likelihood for Δ, in terms of the general weight π and constant c given in Equation G.1, is

$$l'(\Delta) = \sum_{i=1}^{n} \frac{-\pi f'(y_i, -\Delta) + c(1-\pi) f'(y_i, c\Delta)}{\pi f(y_i, -\Delta) + (1-\pi) f(c\Delta)}. \tag{G.6}$$

Note that the requirement on c for $l'(0) = 0$ is equivalent to the condition for Y to have zero mean. Namely, $c = \frac{\pi}{1-\pi}$. Chen's example satisfies this requirement with $\pi = \frac{2}{3}$ and $c = 2$. Thus, $l'(\Delta) = 0$ regardless of the observed data. The second order derivative of the log-likelihood is

$$l''(\Delta) = \sum_{i=1}^{n} \frac{\pi f''(y_i, -\Delta) + c^2(1-\pi) f''(y_i, c\Delta)}{\pi f(y_i, -\Delta) + (1-\pi) f(c\Delta)} - \left(\frac{-\pi f'(y_i, -\Delta) + c(1-\pi) f'(y_i, c\Delta)}{\pi f(y_i, -\Delta) + (1-\pi) f(c\Delta)}\right)^2. \tag{G.7}$$

[19] The proof runs with the variance of the Normal kernels fixed at σ^2, but this case introduces tedious scale factors.

Importantly, the second term in Equation G.7 vanishes when $\Delta = 0$, leaving

$$l''(0) = \left(\pi + c^2(1-\pi)\right) \frac{f''(y_i, 0)}{f(y_i, 0)}, \tag{G.8}$$

which has expectation 0 when the true model has $\Delta = 0$. Thus, the Fisher Information of Δ, when $\Delta = 0$, is 0. To be concise, we simply state the third and fourth derivatives of the likelihood evaluated at $\Delta = 0$,

$$l'''(0) = \sum_{i=1}^{n} \left(-\pi + c^3(1-\pi)\right) \frac{f'''(y_i, 0)}{f(y_i, 0)} \tag{G.9}$$

and

$$l^{(4)}(0) = \sum_{i=1}^{n} \left(\pi + c^4(1-\pi)\right) \frac{f^{(4)}(y_i, 0)}{f(y_i, 0)} - 12 \left(\frac{f''(y_i, 0)}{f(y_i, 0)}\right)^2. \tag{G.10}$$

Define

$$A_i = \frac{f''(Y_i, 0)}{f(Y_i, 0)}; \quad B_i = \frac{f^{(3)}(Y_i, 0)}{f(Y_i, 0)}; \quad C_i = \frac{f^{(4)}(Y_i, 0)}{f(Y_i, 0)}. \tag{G.11}$$

Under the regularity conditions of Equations G.2, G.3, and G.4, A_i, B_i, and C_i have mean 0 and finite variance when expectation is taken with respect to the model with $\Delta = 0$. Thus, the sum of their first n terms are $O_p(n^{1/2})$. Using Taylor's theorem, $l(\Delta)$ can be written as

$$l(\Delta) = l(0) + \Delta^2 \sum_{i=1}^{n} A_i + \frac{1}{3}\Delta^3 \sum_{i=1}^{n} B_i - \frac{1}{2}\Delta^4 \sum_{i=1}^{n} A_i^2 + O_p(n^{1/2}\Delta^4), \tag{G.12}$$

where the regularity conditions in Equations G.3 and G.4 have been used to control the error term. Differentiating with respect to Δ and removing the root at $\Delta = 0$, which asymptotically has probability < 1 of being a maximum,[20] the MLE of Δ has two solutions

$$\widehat{\Delta} = \left[\sum_{i=1}^{n} B_i \pm \left(\left(\sum_{i=1}^{n} B_i\right)^2 + 16 \sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} A_i\right)^{1/2}\right] \times \left[4 \sum_{i=1}^{n} A_i^2\right]^{-1} (1 + o_p(1)). \tag{G.13}$$

Because the n-term sum of A_i^2 is $O_p(n)$, while the other n-term sums are $O_p(n^{1/2})$, this simplifies to

$$\widehat{\Delta} = \delta_0 \left(\sum_{i=1}^{n} A_i\right) \left[\sum_{i=1}^{n} A_i^2\right]^{-1/2} \left[\sum_{i=1}^{n} A_i\right]^{1/2} [1 + o_p(1)] = O_p(n^{-1/4}). \tag{G.14}$$

□

Our proof uses the same Taylor expansion as Chen's proof. However, since in our model the true model has $\Delta = \Delta_n$ rather than $\Delta = 0$, the means of A_i, B_i, and C_i will not be 0; in particular, they will depend on n. Our proof consists of showing that the n-term sums of A_i, B_i, and C_i are still $O_p(n^{1/2})$. We can then simplify to the expression in Equation G.14.

Proof. We define $Y_{i,n}$ as in Equation G.5. We then define $A_{i,n}$, $B_{i,n}$, and $C_{i,n}$ analagously to Equation G.15:

$$A_{i,n} = \frac{f''(Y_{i,n}, 0)}{f(Y_{i,n}, 0)}; \quad B_{i,n} = \frac{f^{(3)}(Y_{i,n}, 0)}{f(Y_{i,n}, 0)}; \quad C_{i,n} = \frac{f^{(4)}(Y_{i,n}, 0)}{f(Y_{i,n}, 0)}. \tag{G.15}$$

[20] This result comes from the CLT applied to $\sum_{i=1}^{n} A_i.$, as the sign of $l''(0)$ determines whether or not $\Delta = 0$ is a maximizer.

We compute these terms explicitly for the Normal kernel with variance 1, finding that

$$
\begin{aligned}
A_{i,n} &= Y_{i,n}^2 - 1 \\
B_{i,n} &= -3Y_{i,n} + Y_{i,n}^3 \\
C_{i,n} &= 3 - 6Y_{i,n}^2 + Y_{i,n}^4.
\end{aligned}
$$

Using the moments of $Y_{i,n}$ and the fact that $Y_{1,n}, \ldots, Y_{n,n}$ are i.i.d., we have that

$$
\begin{aligned}
\sum_{i=1}^{n} A_{i,n} &= O_p\left(\max\{n\Delta_n^2, n^{1/2}\}\right); \ \sum_{i=1}^{n} A_{i,n} \neq o_p\left(n^{1/2}\right) \\
\sum_{i=1}^{n} B_{i,n} &= O_p\left(\max\{n\Delta_n^3, n^{1/2}\}\right); \ \sum_{i=1}^{n} B_{i,n} \neq o_p\left(n^{1/2}\right) \\
\sum_{i=1}^{n} C_{i,n} &= O_p\left(\max\{n\Delta_n^4, n^{1/2}\}\right); \ \sum_{i=1}^{n} C_{i,n} \neq o_p\left(n^{1/2}\right).
\end{aligned}
$$

For Chen's expression in Equation G.14 to be valid, the n-term sums of $A_{i,n}, B_{i,n}$, and $C_{i,n}$ need to be $O_p(n^{1/2})$. This occurs when $\Delta_n = O(n^{-1/4})$. If the means collapse more slowly, Chen's expression is no longer valid and we have no result. If the means collapse more rapidly (that is, $\Delta_n = o(n^{-1/4})$), then parametric convergence of the maximum-likelihood estimator is lost, as the n-term sums of $A_{i,n}$, $B_{i,n}$, and $C_{i,n}$ are all still $O_p(n^{1/2})$ but Δ_n goes to 0 at a faster rate. □

A more general result relaxes the assumptions on π and Δ_n. As long as $l'(0) = 0$, the above argument will run with $\Delta_n = o(n^{-1/4})$ as the critical rate, since the constants in the Taylor expansion will change, but not the ultimate result. If $l'(0) \neq 0$, the Fisher Information of the model for $\Delta = 0$ will not necessarily be 0. Theoretically, in this case normal asymptotics should kick in and the $O_p(n^{-1/2})$ rate should be recovered.