NONPARAMETRIC INFERENCE FOR DISTRIBUTIONAL TREATMENT EFFECTS IN INSTRUMENTAL VARIABLE MODELS

By Kwonsang Lee*, Bhaswar B. Bhattacharya*, Jing Qin† and Dylan S. Small*

*University of Pennsylvania** and *National Institute of Allergy and Infectious Diseases*†

In observational studies, the causal effect of a treatment on the distribution of outcomes is of interest beyond the average treatment effect. Instrumental variable methods allow for causal inference by controlling for unmeasured confounding. The existing nonparametric method for estimating the effect of the treatment on the distribution of outcomes for compliers has several drawbacks, such as producing estimates that violate the non-decreasing and non-negative properties of cumulative distribution functions. In this paper, we propose a novel nonparametric composite likelihood approach, referred to as the *binomial likelihood* (BL) method, which overcomes the limitations of the previous techniques and utilizes the advantage of likelihood methods. We show the consistency of the *maximum binomial likelihood* (MBL) estimators and derive their asymptotic distributions. Next, we develop a computationally efficient algorithm for computing the MBL estimates by combining the expectation-maximization (EM) and the pool-adjacent-violators algorithms (PAVA). Moreover, the BL method can be used to construct a binomial likelihood-ratio test (BLRT) for the null hypothesis of no distributional treatment effect. Asymptotic expansion of the BLRT test is derived and the performance of the BL method is demonstrated in simulation studies. Finally, we apply our method to a study of the effect of Vietnam veteran status on the distribution of civilian annual earnings.

1. Introduction. Randomized experiments are the gold standard for assessing the effect of a treatment but often it is not practical or ethical to randomly assign a treatment itself. However, in some settings, an encouragement to take the treatment can be randomized (Holland, 1988). In other settings, no randomization is possible but there may be a "natural experiment" such that some people are encouraged to receive the treatment compared to others in a way that is effectively random (Angrist and Krueger, 2001). For both of these settings, the instrumental variable (IV) method can be used to estimate the causal effect of a treatment (Holland, 1988; Angrist, Imbens and Rubin, 1996). The IV method is a method that controls for unmeasured confounders to make causal inferences about the effect of a treatment. An IV is informally a variable that affects the treatment

MSC 2010 subject classifications: Primary 62G05, 62G20; secondary 62G30

Keywords and phrases: Causal inference, likelihood ratio test, nonparametric binomial likelihood, order statistics, pool-adjacent-violators algorithm

but is independent of unmeasured confounders and only affects the outcome through affecting the treatment (see Section 2.1 for a more precise definition). Under a monotonicity assumption that the encouraging level of the IV never causes someone not to take the treatment, the IV method identifies the treatment effect for compliers, those subjects who would take the treatment if they received the encouraging level of the IV but would not take the treatment if they did not receive the encouraging level (Angrist, Imbens and Rubin, 1996). For several discussions of the IV method, see Abadie (2003), Angrist, Imbens and Rubin (1996), Baiocchi, Cheng and Small (2014), Brookhart and Schneeweiss (2007), Cheng, Qin and Zhang (2009), Hernan and Robins (2006), Ogburn, Rotnitzky and Robins (2015) and Tan (2006).

Much of the literature on the treatment effect in instrumental variable models has focused on estimating the average treatment effect for compliers. However, understanding the effect of the treatment on the whole distribution of outcomes for the compliers, *the distributional treatment effect* for compliers, is important for optimal individual decision-making and for social welfare comparisons. Optimal individual decision-making requires computing the expected utility of the treatments which requires knowing the whole distribution of the outcomes under the treatments being compared rather than just the average outcomes when the utility function is nonlinear (Karni, 2009). Social welfare comparisons require integration of utility functions under the distribution of the outcome (say income), which again requires knowing the effect of the treatment on the whole distribution of outcomes (Abadie, 2002; Atkinson, 1970).

Abadie (2002) developed a nonparametric method for estimating the effect of treatment on the cumulative distribution functions (CDFs) of the outcomes based on expanding the conventional IV approach described in Imbens and Angrist (1994). These are essentially *plug-in estimates* of the CDFs of the compliers. However, these estimates are not 'proper' CDFs, because they violate the monotonicity or the non-negativity conditions of CDFs. In this paper, we develop a new nonparametric likelihood-based approach for estimating the CDFs of the compliers, that enforces the estimated CDFs to be non-decreasing and non-negative, and then construct a likelihood-ratio type test statistic for the null hypothesis of no distributional treatment effect.

1.1. *Summary of results.* Nonparametric likelihood methods have been shown to have appealing properties in many settings such as providing nonparametric inferences that inherit some of the attractive properties of parametric likelihood (for example, automatic determination of the shape of confidence regions) and straightforward interpretation of side information expressed through constraints (Owen, 2001). However, the usual empirical likelihood approach does not work for the IV model because there are infinitely many solutions that maximize the likelihood. In this case, the usual nonparametric likelihood method fails to produce a meaningful estimator (Geman and Hwang, 1982). We illustrate this problem with an example in Section 2.3.

To circumvent this problem, we propose a novel nonparametric likelihood approach, which builds on the fact that the plug-in estimates identify the CDFs for the compliers at any given point based on a binomially distributed random variable, which counts the number of outcomes that are less than or equal to the value at the given point. Our nonparametric likelihood multiplies together the pieces of the likelihood contributed by these binomial random variables. This is a composite or "pseudo" likelihood rather than a true likelihood because the binomial random variables are actually dependent, but are treated as independent in the composite likelihood. Composite likelihood has been found useful in a range of areas including problems in geo-statistics, spatial extremes, space-time models, clustered data, longitudinal data, time series and statistical genetics, see Lindsay (1988), Heagerty and Lele (1998), Varin, Reid and Firth (2011) and Larribe and Fearnhead (2011). We call this composite nonparametric likelihood method, the *binomial likelihood* (BL) method because it maximizes the average of the likelihood of the binomial random variables at each point across all observation points. The *maximum binomial likelihood* (MBL) estimates are obtained by maximizing the BL under the monotonicity and non-negativity constraints. We derive the asymptotic properties of the MBL estimates, develop efficient algorithms for computing them, construct new tests for detecting distributional effects, and evaluate their performances on real and synthetic data sets. The results obtained are summarized below.

(1) We show that the MBL estimates are consistent and derive their asymptotic distribution. In fact, the plug-in estimates and the MBL estimates asymptotically have the same mean squared error (Theorem 3.2), and, hence, have the same limiting distribution. This makes the BL method useful both in theory and practice: it gives 'proper' estimates of the CDFs of compliers (satisfying the monotonicity and non-decreasing conditions), while preserving the desirable asymptotic properties of the plug-in estimates.
(2) We develop a computationally efficient algorithm for finding the MBL estimates (Section 4) by combining the expectation-maximization (EM) and pool-adjacent-violators algorithms (PAVA). The performance of the BL method is demonstrated in simulation studies, which shows that the MBL estimators perform better than other estimators, particularly when an IV is weak (weakly associated with the treatment).
(3) The BL method can be used to construct a binomial likelihood-ratio test (BLRT) for the null hypothesis of no distributional treatment effect. We derive an asymptotic expansion of the BLRT test (Theorem 6.1), relating it to the well-known Anderson-Darling two-sample test statistic (Pettitt, 1976). Using this asymptotic expansion, we derive a bootstrap procedure for implementing the BLRT test, and compare it with previous methods. Our simulations show that the BLRT is much more powerful in finite samples than tests which do not use the structure of the IV model (for example, the two-sample Kolmogorov-Smirnov test of Abadie (2002)).
(4) We obtain the MBL estimates from a study of the effect of being a Vietnam War veteran on future earnings, as studied by Angrist (1990) and Abadie (2002). To

control for possible unmeasured confounders, Angrist (1990) proposed to use the Vietnam era draft lottery as an IV. During the Vietnam era, men of military service age were randomly assigned a draft lottery number and this draft lottery number was used for prioritizing men for induction into the military. A low draft lottery number makes a man more likely to serve, although some men with a low draft lottery number did not have to serve (for example, if the person had an educational deferment or failed a physical or mental aptitude test) and some men with a high draft lottery number chose to voluntarily serve. Using the draft lottery number as an IV, we make inferences about the distributional effect of Vietnam era military service on civilian earnings in the late 1970s and early 1980s.

1.2. *Organization.* The rest of the article is organized as follows. Basic notation and assumptions of the IV model are discussed in Section 2. We also discuss the existing plug-in approach and the difficulty in applying the usual empirical likelihood to IV models. In Section 3, we introduce the BL method and derive the asymptotic properties of the MBL estimators. In Section 4, we develop an efficient algorithm for computing the MBL estimates, and in Section 5, illustrate their performance in simulations. In Section 6, we describe the BLRT for detecting distribution treatment effects, study its asymptotic properties, and compare its performance with other methods in numerical experiments. The MBL method and the BLRT test are applied to the veterans data in Section 7. Section 8 concludes our findings. Proofs of the results are given in Appendix.

2. Framework and review. In this section the framework of instrumental variable (IV) model is introduced and the existing methods are briefly reviewed. Notation and identification assumptions are discussed in Section 2.1. The existing approach of Abadie (2002) for estimating the distributional treatment effect in IV models is reviewed, and the shortcomings of this method are addressed, in Section 2.2. The drawback of the usual empirical likelihood approach for the IV models is discussed in Section 2.3. This, together with the discussion in Section 2.2, motivates our new approach of constructing the nonparametric binomial likelihood that will be introduced in Section 3.

2.1. *Notation and assumptions.* Let the subjects in the observational study be indexed by $[n] := \{1, 2, \ldots, n\}$. For $a \in [n]$, Z_a denotes the binary instrumental variable, D_a the indicator variable for whether the subject a receives the treatment or not, and Y_a the outcome variable, which, for this paper, will be assumed to be continuous. Using the potential outcome framework (Neyman, 1990; Rubin, 1974), define $D_a(0)$ as the value that D_a would be if Z_a were to be set to 0, and $D_a(1)$ as the value that D_a would be if Z_a were to be set to 1. Similarly, $Y_a(z, d)$ for $(z, d) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, is the value that the outcome Y_a would be if $Z_a = z$ and $D_a = d$. For each subject $a \in [n]$, the analyst can only observe one of the two potential values $D_a(0)$ and $D_a(1)$, and one of the four potential

TABLE 1
Compliance classes by the potential outcomes $D_1(0)$ and $D_1(1)$

	$D_1(0) = 0$	$D_1(0) = 1$
$D_1(1) = 0$	Never-takers	Defiers
$D_1(1) = 1$	Compliers	Always-takers

values $Y_a(0,0), Y_a(0,1), Y_a(1,0), Y_a(1,1)$. The observed treatment D_a is

$$D_a = Z_a D_a(1) + (1 - Z_a) D_a(0).$$

Similarly, the observed outcome Y_a can be expressed as $Y_a = Z_a D_a \cdot Y_a(1,1) + Z_a(1 - D_a) \cdot Y_a(1,0) + (1 - Z_a) D_a \cdot Y_a(0,1) + (1 - Z_a)(1 - D_a) \cdot Y_a(0,0)$. A subject's *compliance class* is determined by the combination of the potential treatment values $D_a(0)$ and $D_a(1)$, which is denoted by S_a: $S_a =$ *always-taker* (*at*) if $D_a(0) = D_a(1) = 1$; $S_a =$ *never-taker* (*nt*) if $D_a(0) = D_a(1) = 0$; $S_a =$ *complier* (*co*) if $D_a(0) = 0, D_a(1) = 1$; and $S_a =$ *defier* (*de*) if $D_a(0) = 1, D_a(1) = 0$. This is summarized in Table 1.

For the rest of this article, the following standard identifying conditions are assumed. The implications of these conditions are briefly explained in the paragraph below, see Angrist, Imbens and Rubin (1996) for more details on these assumptions.

ASSUMPTION 1. Hereafter, the following identification conditions will be imposed on the IV model:

(1) *Stable Unit Treatment Value Assumption* (SUTVA) (Rubin, 1986): The outcome (treatment) for the individual $a \in [n]$ is not affected by the values of the treatment or instrument (instrument) for other individuals and the outcome (treatment) does not depend on the way the treatment or instrument (instrument) is administered.
(2) *The instrumental variable Z_a is independent of the potential outcomes $Y_a(z,d)$ and potential treatment $D_a(z)$.*

$$Z_a \perp\!\!\!\perp (Y_a(0,0), Y_a(0,1), Y_a(1,0), Y_a(1,1), D_a(0), D_a(1))$$

(3) *Nonzero average causal effect of Z_a on D_a*: $\mathbb{P}(D_a(1) = 1) > \mathbb{P}(D_a(0) = 1)$.
(4) *Monotonicity*: $D_a(1) \geq D_a(0)$.
(5) *Exclusion restriction*: $Y_a(0,d) = Y_a(1,d)$, for $d = 0$ or 1.

Assumption 1 enables the causal effect of the treatment for the subpopulation of the compliers to be identified. The SUTVA allows us to use the notation $Y_a(z,d)$ (or $D_a(z)$), which means that the outcome (treatment) for individual a is not affected by the values of the treatment and instrument (instrument) for other individuals. Condition (2) will be satisfied if Z_a is randomized. Condition (3) requires Z_a to have some effect on the average probability of treatment. Condition (4), the monotonicity assumption, means that the possibility of $D_a(0) = 1$, $D_a(1) = 0$ is excluded, that is, there are no defiers (see Table 1). Condition (5) assures that any effect of Z_a on Y_a must be through an effect of Z_a on D_a. Under this assumption, the potential outcome can be written as $Y_a(d)$, instead of $Y_a(z,d)$.

2.1.1. *The outcome CDFs of the compliance classes.* Define the outcome CDFs of compliers without treatment, never-takers, compliers with treatment, and always-takers respectively:

$$\begin{aligned}
F_{co}^{(0)}(t) &= \mathbb{E}(\mathbf{1}\{Y_1(0) \le t\}|D_1(1) = 1, D_1(0) = 0),\\
F_{nt}(t) &= \mathbb{E}(\mathbf{1}\{Y_1(0) \le t\}|D_1(1) = 0, D_1(0) = 0),\\
F_{co}^{(1)}(t) &= \mathbb{E}(\mathbf{1}\{Y_1(1) \le t\}|D_1(1) = 1, D_1(0) = 0),\\
F_{at}(t) &= \mathbb{E}(\mathbf{1}\{Y_1(1) \le t\}|D_1(1) = 1, D_1(0) = 1).
\end{aligned} \tag{2.1}$$

Denote $\boldsymbol{F}(t) = (F_{co}^{(0)}(t), F_{nt}(t), F_{co}^{(1)}(t), F_{at}(t))$, the vector of the above CDFs. Under Assumption 1, these distributions are identified such as

$$\begin{aligned}
F_{co}^{(0)}(t) =& \mathbb{P}(Y_1 \le t|Z_1 = 0, S_1 = co),\\
F_{nt}(t) =& \mathbb{P}(Y_1 \le t|Z_1 = 0, S_1 = nt),\\
F_{co}^{(1)}(t) =& \mathbb{P}(Y_1 \le t|Z_1 = 1, S_1 = co),\\
F_{at}(t) =& \mathbb{P}(Y_1 \le t|Z_1 = 1, S_1 = at).
\end{aligned} \tag{2.2}$$

Next, for $u, v \in \{0, 1\}$, define

$$F_{uv}(t) = \mathbb{P}(Y_1 \le t|Z_1 = u, D_1 = v). \tag{2.3}$$

Note that

$$\begin{aligned}
F_{00}(t) &= \mathbb{P}(Y_1 \le t|Z_1 = 0, D_1 = 0)\\
&= \frac{\mathbb{P}(Y_1 \le t, Z_1 = 0, D_1 = 0, S_1 = co) + \mathbb{P}(Y_1 \le t, Z_1 = 0, D_1 = 0, S_1 = nt)}{\mathbb{P}(Z_1 = 0, D_1 = 0)}\\
&= \lambda_0 \mathbb{P}(Y_1 \le t|Z_1 = 0, S_1 = co) + (1 - \lambda_0)\mathbb{P}(Y_1 \le t|Z_1 = 0, S_1 = nt)\\
&= \lambda_0 F_{co}^{(0)}(t) + (1 - \lambda_0)F_{nt}(t).
\end{aligned} \tag{2.4}$$

where $\lambda_0 = \mathbb{P}(S_1 = co|Z_1 = 0, D_1 = 0)$. Similarly, it follows that $F_{01}(t) = F_{at}(t)$, $F_{10}(t) = F_{nt}(t)$, and

$$F_{11}(t) = \lambda_1 F_{co}^{(1)}(t) + (1 - \lambda_1)F_{at}(t). \tag{2.5}$$

where $\lambda_1 = \mathbb{P}(S_1 = co|Z_1 = 1, D_1 = 1)$.

Next, consider the (unknown) proportions of compliance classes $\boldsymbol{\phi} = (\phi_{nt}, \phi_{at})$, where

$$\phi_{at} = \mathbb{P}(S_1 = at),\ \phi_{nt} = \mathbb{P}(S_1 = nt),\ \phi_{co} = \mathbb{P}(S_1 = co), \tag{2.6}$$

with $\phi_{co} + \phi_{at} + \phi_{nt} = 1$. For $u, v \in \{0, 1\}$, define $n_{uv} := \sum_{a=1}^{n} \mathbf{1}\{Z_a = u, D_a = v\}$. Then $\lim_{n\to\infty} n_{uv}/n \to \mathbb{P}(Z_1 = u, D_1 = v) := \eta_{uv}$, for all $u, v \in \{0, 1\}$. From Assumption 1,

$$\eta_{00} = \phi_0(\phi_{co} + \phi_{nt}),\ \eta_{01} = \phi_0\phi_{at},\ \eta_{10} = \phi_1\phi_{nt},\ \eta_{11} = \phi_1(\phi_{co} + \phi_{at}) \tag{2.7}$$

where $\phi_0 = \mathbb{P}(Z_1 = 0)$ and $\phi_1 = 1 - \phi_0 = \mathbb{P}(Z_1 = 1)$. Moreover,

$$\lambda_0 = \frac{\phi_{co}}{\phi_{co} + \phi_{nt}}, \quad \lambda_1 = \frac{\phi_{co}}{\phi_{co} + \phi_{at}}. \tag{2.8}$$

Finally, let $H = \eta_{00}F_{00} + \eta_{01}F_{01} + \eta_{10}F_{10} + \eta_{11}F_{11}$, be the mixture distribution of the $\{F_{uv}\}_{u,v\in\{0,1\}}$. Note that in the IV model, data from the outcome variable $Y_1, Y_2, \ldots, Y_n$ are i.i.d. from H.

2.1.2. *Parameter space.* As discussed above, the IV model has three sets of parameters $(\boldsymbol{F}, \boldsymbol{\phi}, \phi_1)$: the vector of outcome CDFs $\boldsymbol{F}(t) = (F^{(0)}_{co}(t), F_{nt}(t), F^{(1)}_{co}(t), F_{at}(t))$ (defined in (2.1)), the proportion of the compliance classes $\boldsymbol{\phi}$ (2.6), and $\phi_1 = \mathbb{P}(Z_1 = 1)$.

In this section, we define the various spaces associated with these parameters. To this end, denote by $\mathbb{R}^{\mathbb{R}}$ and $[0,1]^{\mathbb{R}}$ the sets of all functions from $\mathbb{R} \to \mathbb{R}$ and $\mathbb{R} \to [0,1]$, respectively. Let the set of all non-decreasing functions from $\mathbb{R} \to [0,1]$ be $\mathbb{I}([0,1]^{\mathbb{R}})$, and the set of distribution functions from $\mathbb{R} \to [0,1]$ be $\mathcal{P}([0,1]^{\mathbb{R}})$. Define the *unrestricted* parameter space

$$\boldsymbol{\vartheta} = \left\{(\theta^{(0)}_{co}, \theta_{nt}, \theta^{(1)}_{co}, \theta_{at}) : \theta^{(0)}_{co}, \theta_{nt}, \theta^{(1)}_{co}, \theta_{at} \in \mathbb{R}^{\mathbb{R}}\right\}. \tag{2.9}$$

The *restricted* parameter space is the subset of $\boldsymbol{\vartheta}$ where each $\theta^{(0)}_{co}, \theta_{nt}, \theta^{(1)}_{co}, \theta_{at}$ is a distribution function. Formally,

$$\boldsymbol{\vartheta}_+ = \left\{(\theta^{(0)}_{co}, \theta_{nt}, \theta^{(1)}_{co}, \theta_{at}) : \theta^{(0)}_{co}, \theta_{nt}, \theta^{(1)}_{co}, \theta_{at} \in \mathcal{P}([0,1]^{\mathbb{R}})\right\}. \tag{2.10}$$

For the parameters $\boldsymbol{\phi} = (\phi_{at}, \phi_{nt})$, with $\phi_{co} = 1 - \phi_{at} - \phi_{nt}$, the unrestricted parameter space is $\mathbb{R}^2$, and the restricted parameter space is $[0,1]^2_+ := \{(x,y) \in [0,1]^2 : 0 \le x+y \le 1\}$. Finally, for the parameter $\phi_1 = \mathbb{P}(Z_1 = 1)$, the parameter space is $[0,1]$. Therefore, the complete parameter spaces for the IV model are the following:

- the *unrestricted parameter space* is $\boldsymbol{\vartheta} \times \mathbb{R}^2 \times [0,1]$,
- the *restricted parameter space* is $\boldsymbol{\vartheta}_+ \times [0,1]^2_+ \times [0,1]$.

2.2. *Review of the existing nonparametric method.* In this section we recall the existing method of estimating $(\boldsymbol{F}, \boldsymbol{\phi}, \phi_1)$. To begin with, note that ϕ_0 and ϕ_1 can be easily estimated from the sample proportions as $\breve{\phi}_0 = \frac{n_{00}+n_{01}}{n}$ and $\breve{\phi}_1 = 1 - \breve{\phi}_0$. Similarly, $\eta_{00}, \eta_{01}, \eta_{10}, \eta_{11}$, and hence $\phi_{co}, \phi_{at}, \phi_{nt}$, can be estimated directly from the sample proportions as follows:

$$\breve{\boldsymbol{\phi}} = (\breve{\phi}_{at}, \breve{\phi}_{nt})' = \left(\frac{n_{01}}{n_{00} + n_{01}}, \frac{n_{10}}{n_{10} + n_{11}}\right)', \tag{2.11}$$

and $\breve{\phi}_{co} := 1 - \breve{\phi}_{at} - \breve{\phi}_{nt}$. These estimators will be referred to as the *plug-in estimators* for the compliance classes. The plug-in estimators are consistent and asymptotically Gaussian around their true values, however, in finite samples, it has the following drawback:

– The plug-in estimator $\breve{\phi}_{co}$ can be negative in the sample (whenever $n_{01}n_{10} \leq n_{00}n_{11}$). To make it non-negative, we can truncate it to zero whenever $\breve{\phi}_{co}$ is negative, but, then $\breve{\phi}_{co}$, $\breve{\phi}_{at}$, $\breve{\phi}_{at}$ do not add up to 1. Our BL method, proposed in Section 3, addresses this issue, by providing the MBL estimators $\hat{\phi}_{co}$, $\hat{\phi}_{at}$, $\hat{\phi}_{at}$ which are non-negative and add up to 1.

To estimate the outcome CDFs of the compliance classes, Abadie (2002) used the following formulas for the CDFs of the potential outcome for compliers under treatment and control, which holds under Assumptions 1:

$$F_{co}^{(1)}(t) = \frac{\mathbb{E}(\mathbf{1}\{Y_1 \leq t\}D_1|Z_1 = 1) - \mathbb{E}(\mathbf{1}\{Y_1 \leq t\}D_1|Z_1 = 0)}{\mathbb{E}(D_1|Z_1 = 1) - \mathbb{E}(D_1|Z_1 = 0)}, \tag{2.12}$$

and

$$F_{co}^{(0)}(t) = \frac{\mathbb{E}(\mathbf{1}\{Y_1 \leq t\}(1 - D_1)|Z_1 = 1) - \mathbb{E}(\mathbf{1}\{Y_1 \leq t\}(1 - D_1)|Z_1 = 0)}{\mathbb{E}((1 - D_1)|Z_1 = 1) - \mathbb{E}((1 - D_1)|Z_1 = 0)}. \tag{2.13}$$

Abadie (2002) proposed substituting the sample means for the expectation in (2.12) and (2.13), which gives

$$\breve{F}_{co}^{(0)}(t) = \frac{(\breve{\phi}_{co} + \breve{\phi}_{nt})\overline{F}_{00}(t) - \breve{\phi}_{nt}\overline{F}_{10}(t)}{\breve{\phi}_{co}}, \quad \breve{F}_{co}^{(1)}(t) = \frac{(\breve{\phi}_{co} + \breve{\phi}_{at})\overline{F}_{11}(t) - \breve{\phi}_{at}\overline{F}_{01}(t)}{\breve{\phi}_{co}}, \tag{2.14}$$

where $\breve{\phi}_{co}, \breve{\phi}_{nt}, \breve{\phi}_{at}$ are the plug-in estimators of the proportions of the compliance classes (2.11), and, for $u, v \in \{0, 1\}$,

$$\overline{F}_{uv}(t) = \frac{1}{n_{uv}} \sum_{a=1}^{n} \mathbf{1}\{Z_a = u, D_a = v, Y_a \leq t\}, \tag{2.15}$$

is the empirical distribution of (2.3) based on the observed data. The plug-in estimators in (2.14) can be written as

$$\breve{F}_{co}^{(0)}(t) = \frac{\overline{F}_{00}(t) - (1 - \breve{\lambda}_0)\overline{F}_{10}(t)}{\breve{\lambda}_0}, \quad \breve{F}_{co}^{(1)}(t) = \frac{\overline{F}_{11}(t) - (1 - \breve{\lambda}_1)\overline{F}_{01}(t)}{\breve{\lambda}_1},$$

where $\breve{\lambda}_0 = \breve{\phi}_{co}/(\breve{\phi}_{co} + \breve{\phi}_{nt})$ and $\breve{\lambda}_1 = \breve{\phi}_{co}/(\breve{\phi}_{co} + \breve{\phi}_{at})$ are the plug-in estimators of λ_0 and λ_1 respectively. Other CDFs such as F_{nt} and F_{at} are equal to the empirical distributions $\breve{F}_{at} := \overline{F}_{01}$ and $\breve{F}_{nt} := \overline{F}_{10}$. Finally, we create a vector of those nonparametric plug-in estimators of the outcome CDFs,

$$\breve{\boldsymbol{F}}(t) := (\breve{F}_{co}^{(0)}(t), \breve{F}_{nt}(t), \breve{F}_{co}^{(1)}(t), \breve{F}_{at}(t))'. \tag{2.16}$$

There are three problems with the nonparametric plug-in estimators which this paper seeks to improve:

TABLE 2
The structure of the data for the single consent design

Z_a	D_a	Compliance Classes	Distribution	Count
0	0	Never-Takers/Compliers without treatment	$\lambda_0 F_{co}^{(0)} + (1-\lambda_0)F_{nt}$	n_{00}
1	0	Never-Takers	F_{nt}	n_{10}
1	1	Compliers with treatment	$F_{co}^{(1)}$	n_{11}

(1) The plug-in estimators $\breve{F}_{co}^{(1)}(t)$ and $\breve{F}_{co}^{(0)}(t)$ always violate the non-decreasing condition of CDFs.
(2) The plug-in estimators may produce estimates which are outside of the interval [0,1]. This is called the *violation of non-negativeness*.
(3) Finally, the plug-in estimators can be highly unstable in the weak instrument setting (meaning that the IV is only weakly associated with the treatment so that there are a small proportion of compliers) because the denominators of both equations (2.12) and (2.13) depend on the proportion of compliers in the entire population.

The three problems arise at the same time when the IV is weak or the sample size is relatively small. The maximum binomial likelihood (MBL) method proposed in Section 3 overcomes these issues and has the appealing properties of likelihood methods.

2.3. *Failure of usual nonparametric likelihood methods for the IV model.* Usual nonparametric likelihood methods, when they are applied to the IV model, are to maximize the likelihood under Assumptions 1 and the assumption that the data is independent and identically distributed. No further assumptions about the distribution of the data are needed (see Owen (2001) for a general discussion of nonparametric maximum likelihood). To see that usual nonparametric likelihood methods do not work well for the IV model, we consider the single consent design where the treatment is only available for encouraged subjects. Consequently, there are only never-takers and compliers (Zelen, 1979), see Table 2. We also assume that the proportion of compliers given $Z_a = 0$, $\mathbb{P}(S_a = co|Z_a = 0, D_a = 0) = \lambda_0$, and the probability of being assigned to no encouragement, $\mathbb{P}(Z_a = 0) = \phi_0$ are known. We will focus on estimating F_{nt} and $F_{co}^{(0)}$. The outcome distribution of compliers with treatment $F_{co}^{(1)}$ can be identified from the data $Y_a|Z_a = 1, D_a = 1$. Moreover, since the data $Y_a|Z_a = 1, D_a = 1$ provide no information about the never-takers' distribution or the compliers without treatment distribution, we will ignore this part of the data in what follows. As we ignore the data with $Z_a = 1, D_a = 1$, $Z_a = 1$ automatically implies $Z_a = 1, D_a = 0$.

Nonparametric maximum likelihood puts point masses at all observed values, $Y_a = y_a$ (Owen, 2001). Let $\mathrm{d}F_{nt}(y_a) = p_a$ and $\mathrm{d}F_{co}^{(0)}(y_a) = q_a$. Then, the log-likelihood function (up to an additive constant) is given by

$$\ell(\boldsymbol{p}, \boldsymbol{q}) = \sum_{a=1}^{n} \left[\mathbf{1}\{Z_a = 1\} \log p_a + \mathbf{1}\{Z_a = 0\} \log(\lambda_0 q_a + (1-\lambda_0)p_a)\right].$$

where $\boldsymbol{p} = (p_1, \dots, p_n), \boldsymbol{q} = (q_1, \dots, q_n)$ and $n = n_{00} + n_{10}$. Now, we need to solve the following optimization problem;

$$\text{maximize} \quad \ell(\boldsymbol{p}, \boldsymbol{q}) \quad \text{subject to} \quad \sum_{a=1}^{n} p_a = \sum_{a=1}^{n} q_a = 1.$$

It is easy to see that the solution for this optimization problem is any set of (p_a, q_a), $a \in [n] := \{1, 2, \dots, n\}$ that satisfies the following condition:

$$p_a = \frac{1}{(1-\lambda_0)(n_{10}+n_{00})} \text{ when } Z_a = 1, \quad \lambda_0 q_a + (1-\lambda_0) p_a = \frac{1}{n_{10}+n_{00}} \text{ when } Z_a = 0. \tag{2.17}$$

Observe that there are infinitely many pairs of $(\boldsymbol{p}, \boldsymbol{q})$ which satisfy this condition: there are $2n$ parameters to estimate, but there are only n equations in (2.17). To illustrate the difficulties that arise, we consider a simple numerical example.

Example 1. Suppose we have two observations (0.99, 1.99) with $Z_a = 1, D_a = 0$ (from F_{nt}), and three observations (1, 2, 5) with $Z_a = 0, D_a = 0$ (from the mixture of $\lambda_0 F_{co}^{(0)} + (1-\lambda_0) F_{nt}$ with the known $\lambda_0 = 1/3$). Since we assume $\lambda_0 = 1/3$, we roughly expect that two of the three subjects in the subpopulation $Z_a = D_a = 0$ are never-takers and one is a complier. As the observations (0.99, 1.99) in the subpopulation $Z_a = 1, D_a = 0$ are from F_{nt}, it seems natural to estimate that of the observations (1, 2, 5) in the subpopulation $Z_a = D_a = 0$, 1 and 2 are never-takers (since they are close to our never-takers' outcomes 0.99 and 1.99) and the observation 5 is a complier (that is, $F_{co}^{(0)}$ puts probability 1 on the point mass 5). In other words, $\mathrm{d}F_{nt} = (0.25, 0.25, 0.25, 0.25, 0)$ and $\mathrm{d}F_{co}^{(0)} = (0, 0, 0, 0, 1)$ on (0.99, 1, 1.99, 2, 5) can be natural estimates. However, these estimates cannot be obtained from equation (2.17) and in fact, one of infinitely many solutions to equation (2.17) is $\mathrm{d}F_{nt} = (0.3, 0, 0.3, 0.1, 0.3)$ and $\mathrm{d}F_{co}^{(0)} = (0, 0.6, 0, 0.4, 0)$ on (0.99, 1, 1.99, 2, 5), which are much different from the natural estimates. This results from the fact that the usual nonparametric maximum likelihood approach ignores the closeness between the observed data points. The only notion of closeness it considers between two distributions is their probability of giving *exactly* the same value.

For settings in which usual nonparametric likelihood methods do not produce a unique estimate, Bickel et al. (1993) discuss three modifications: the method of sieves, the method of regularization (most commonly, penalized maximum likelihood) and regularized maximum likelihood method. However, Bickel et al. (1993) point out that all the three methods bring in additional choices that have to be made. For example, for the method of regularization, the tuning parameter and the penalty functional must be specified. The selection of the tuning parameter requires additional procedures such as cross-validation. In this paper, we would like to consider a nonparametric maximum likelihood method which takes advantage of the identifiability without introducing choice parameters. Our proposed approach automatically selects locations where it puts point masses when estimating the distribution.

3. The binomial likelihood method. In this section we describe the MBL estimation method in IV models and study its properties: The method is described in Section 3.1 starting with introducing the concept of the binomial likelihood. Then, the asymptotic properties of the MBL estimators are discussed in Section 3.2. The performance of the MBL estimation method is evaluated with simulations in Section 5.

3.1. *Binomial likelihood in IV models.* Define $\boldsymbol{\theta} : \mathbb{R} \to [0,1]^4$ such that

$$\boldsymbol{\theta}(t) = (\theta_{co}^{(0)}(t), \theta_{nt}(t), \theta_{co}^{(1)}(t), \theta_{at}(t)),$$

where $\theta_{co}^{(0)}, \theta_{nt}, \theta_{co}^{(1)}, \theta_{at} : \mathbb{R} \to [0,1]$ are functional variables representing the outcome CDFs of compliers without treatment, never-takers, compliers with treatment, and always-takers, respectively. Moreover, for $u, v \in \{0,1\}$, replacing F_{uv} by θ_{uv} in (2.3) to emphasize the fact that it is a variable,

$$\theta_{uv}(t) = \mathbb{P}(Y_1 \le t | Z_1 = u, D_1 = v). \tag{3.1}$$

Next, define $\boldsymbol{\chi} = (\chi_{nt}, \chi_{at})$ the variables corresponding to the proportions of never-takers and always-takers. (The proportion of compliers is $\chi_{co} = 1 - \chi_{nt} - \chi_{at}$). Finally, denote by υ the variable for the proportion of individuals receiving the instrument, that is, $\mathbb{P}(Z_1 = 1)$.

Denote the data $\mathcal{D}_n := ((Y_1, Z_1, D_1), (Y_2, Z_2, D_2), \ldots, (Y_n, Z_n, D_n))'$. For $u, v \in \{0,1\}$ denote the event $K_{uv}^a := \{Z_a = u, D_a = v\}$. The probability $\mathbb{P}(K_{uv}^a)$ can be easily computed in terms of the variables $(\boldsymbol{\chi}, \upsilon)$, as shown in (A.1). Then, given the data and $u, v \in \{0,1\}$, we can define a *two-point binomial likelihood function* for the data points (Y_a, Y_b) as follows:

$$\begin{aligned} L_{Y_a,Y_b}^{(u,v)}(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon | \mathcal{D}_n) &= \begin{cases} \mathbb{P}(Y_a \le Y_b) & \text{on } K_{uv}^a \text{ and } \{Y_a \le Y_b\} \\ \mathbb{P}(Y_a > Y_b) & \text{on } K_{uv}^a \text{ and } \{Y_a > Y_b\} \\ 1 & \text{otherwise} \end{cases} \\ &= \begin{cases} \mathbb{P}(K_{uv}^a) \cdot \theta_{uv}(Y_b) & \text{on } K_{uv}^a \text{ and } \{Y_a \le Y_b\} \\ \mathbb{P}(K_{uv}^a) \cdot (1 - \theta_{uv}(Y_b)) & \text{on } K_{uv}^a \text{ and } \{Y_a > Y_b\} \\ 1 & \text{otherwise} \end{cases}, \end{aligned} \tag{3.2}$$

where the last step uses (3.1) above. One natural way to obtain a composite likelihood for the IV model, is to combine (3.2) over $u, v \in \{0,1\}$ and all pairs $a, b \in [n] := \{1, 2, \ldots, n\}$. This essentially gives us the binomial likelihood, which we formally describe below:

- For every $b \in [n]$, we combine (3.2) over $u, v \in \{0,1\}$ and $a \in [n]$ to get the *likelihood function at the point* Y_b as follows:

$$L_{Y_b}(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon | \mathcal{D}_n) := \prod_{a=1}^{n} \prod_{u,v \in \{0,1\}} L_{Y_a,Y_b}^{(u,v)}(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon | \mathcal{D}_n). \tag{3.3}$$

– Note that the contributions of the one-point likelihoods are, obviously, not independent over $b \in [n]$. Nevertheless, pretending they are independent, we can combine the one-point likelihoods over $b \in [n]$, and obtain a 'quasi-likelihood' for IV model. However, to avoid dealing with potentially vanishing probabilities in the boundary, instead of taking the product over all $b \in [n]$, we take the product a over truncated set. To this end, fix $0 < \kappa < 1/2$, and define the *binomial likelihood (BL) function for the IV model* as:

$$L_\kappa(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon | \mathcal{D}_n) = \prod_{b \in I_\kappa} L_{Y_{(b)}}(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon | \mathcal{D}_n), \tag{3.4}$$

where $I_\kappa := [\lceil n\kappa \rceil, \lceil n(1-\kappa) \rceil]$ and $Y_{(r)}$ is the r-th order statistic of the set $\{Y_1, Y_2, \ldots, Y_n\}$, for $r \in [n]$.

REMARK 3.1. Note that the one-point log-likelihood functions (3.3) blow up for the extreme order statistics. To avoid technicalities arising from this, we define the BL function (3.4) over the central order statistics, that is, for $b \in I_\kappa$. Throughout the paper, κ will be any small fixed constant, and the asymptotics will be in the regime where the sample size n grows to infinity, keeping κ fixed. This estimates the CDFs of the compliance classes accurately on the bulk of the support of H, the distribution function of the outcome variable (see Theorem 3.2 for details). Hereafter, we omit dependence on κ in the BL functions, for notational brevity.

Given the binomial likelihood function (3.4), we obtain the estimates of $(\boldsymbol{\theta}, \boldsymbol{\phi}, \phi_1)$ by maximizing it over their corresponding parameters spaces:

DEFINITION 3.1. The *maximum binomial likelihood* (MBL) estimator of $(\boldsymbol{\theta}, \boldsymbol{\phi}, \phi_1)$ is defined as:

$$(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}, \hat{\phi}_1) := \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_+, \boldsymbol{\chi} \in [0,1]_+^2, \varphi \in [0,1]} L(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon | \mathcal{D}_n), \tag{3.5}$$

where $\boldsymbol{\vartheta}_+$ and $[0,1]_+$ are the restricted parameter spaces as defined in Section 2.1.2. Note that (3.5) only defines $\hat{\boldsymbol{F}} = (\hat{F}_{co}^{(0)}, \hat{F}_{nt}, \hat{F}_{co}^{(1)}, \hat{F}_{at})'$ at the knots $\{Y_{(b)}\}_{b \in I_\kappa}$. To ensure (3.5) is well-defined, we extend $\hat{\boldsymbol{F}}$ between the knots by coordinate-wise right-continuous interpolation, and extrapolating beyond the knots by to 0 and 1.

The intuitive rationale behind choosing the BL for estimating $(\boldsymbol{F}, \boldsymbol{\phi})$ is the following: The nonparametric plug-in method uses equations (2.12) and (2.13) to estimate the CDFs for compliers with and without treatment. The plug-in method identifies the CDFs for the compliers at any given point using information on whether outcomes are less than or equal to that point or not. The advantage of the binomial likelihood over the plug-in method is that it ties together the information from equations (2.12) and (2.13), by averaging the likelihoods of these binomial random variables at each point (the one point likelihood-functions) across all observed data points. In fact, maximizing the BL function over the unrestricted parameter space $\boldsymbol{\vartheta} \times \mathbb{R}^2$ (recall (2.9)) gives the plug-in estimates $(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$ (see Lemma A.1 in Appendix A.2). Therefore, to get estimates of $(\boldsymbol{F}, \boldsymbol{\phi})$ satisfying the non-negative and non-decreasing constraints, it is natural to maximize the BL function over the restricted space, as in Definition 3.1.

The full expression of the BL function is long and unwieldy. However, we can re-write this in a compact and instructive form, by grouping and re-arranging the terms. Define the following quantities:

$$T_{00}^{(n)}(\theta_{co}^{(0)}, \theta_{nt}, \boldsymbol{\chi}) = \frac{1}{n} \sum_{b \in I_\kappa} \frac{n_{00}}{n} \left\{ \log(1 - \chi_{at}) + J(\overline{F}_{00}(Y_{(b)}), \frac{\chi_{co}}{\chi_{co} + \chi_{nt}} \theta_{co}^{(0)}(Y_{(b)}) + \frac{\chi_{nt}}{\chi_{co} + \chi_{nt}} \theta_{nt}(Y_{(b)})) \right\},$$

$$T_{10}^{(n)}(\theta_{nt}, \boldsymbol{\chi}) = \frac{1}{n} \sum_{b \in I_\kappa} \frac{n_{10}}{n} \left\{ \log \chi_{nt} + J(\overline{F}_{10}(Y_{(b)}), \theta_{nt}(Y_{(b)})) \right\},$$

$$T_{01}^{(n)}(\theta_{at}, \boldsymbol{\chi}) = \frac{1}{n} \sum_{b \in I_\kappa} \frac{n_{01}}{n} \left\{ \log \chi_{at} + J(\overline{F}_{01}(Y_{(b)}), \theta_{at}(Y_{(b)})) \right\},$$

$$T_{11}^{(n)}(\theta_{co}^{(1)}, \theta_{at}, \boldsymbol{\chi}) = \frac{1}{n} \sum_{b \in I_\kappa} \frac{n_{11}}{n} \left\{ \log(1 - \chi_{nt}) + J(\overline{F}_{11}(Y_{(b)}), \frac{\chi_{co}}{\chi_{co} + \chi_{at}} \theta_{co}^{(1)}(Y_{(b)}) + \frac{\chi_{at}}{\chi_{co} + \chi_{at}} \theta_{at}(Y_{(b)})) \right\},$$

where the function $J(x, y) := x \log y + (1 - x) \log(1 - y)$. Finally, define

$$\mathbb{M}_n(\boldsymbol{\theta}, \boldsymbol{\chi}) = T_{00}^{(n)}(\theta_{co}^{(0)}, \theta_{nt}, \boldsymbol{\chi}) + T_{10}^{(n)}(\theta_{nt}, \boldsymbol{\chi}) + T_{01}^{(n)}(\theta_{at}, \boldsymbol{\chi}) + T_{11}^{(n)}(\theta_{co}^{(1)}, \theta_{at}, \boldsymbol{\chi}). \tag{3.6}$$

With these definitions, we have the following proposition, which follows by direct substitution. The proof is given in Appendix A.2.

PROPOSITION 3.1. *Let $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}, \hat{\phi}_1)$ be as defined in (3.5). Then*

$$\hat{\phi}_1 = \frac{n_{10} + n_{11}}{n}. \tag{3.7}$$

Moreover,

$$(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_+, \boldsymbol{\chi} \in [0,1]_+^2} \mathbb{M}_n(\boldsymbol{\theta}, \boldsymbol{\chi}), \tag{3.8}$$

where $\mathbb{M}_n$ is as defined above in (3.6).

This proposition shows that the MBL estimate of ϕ_1 is the proportion of individuals with instrument (that is, $Z_a = 1$) in the observed sample. Furthermore, the MBL estimates of $\boldsymbol{F}$ and $\boldsymbol{\phi}$ can be obtained by maximizing the function $\mathbb{M}_n$ (defined above in (3.6)). Hereafter, we refer to function $\mathbb{M}_n$ as the *sample binomial log-likelihood* or the *sample objective function.*

3.2. *Theoretical results.* In this section we discuss the asymptotic properties of the MBL estimators $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}})$, and how they compare with the plug-in estimators $(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$ (recall (2.14) and (2.11)). We have the following assumption:

ASSUMPTION 2. Let $\boldsymbol{\phi} = (\phi_{nt}, \phi_{at})$ be as defined (2.6) and F_{uv}, for $u, v \in \{0, 1\}$, be as in (2.4) and (2.5). Throughout the paper we assume the following:

(a) The proportion vector $\boldsymbol{\phi}$ belongs to the interior of the parameter space $[0, 1]^2_+$.
(b) The CDFs $\{F_{uv}\}_{u,v\in\{0,1\}}$ are continuous, strictly increasing, and have the same support.
(c) There exists constants $0 < C_1 \leq C_2 < \infty$ such that

$$C_1|s - t| \leq |F_{uv}(s) - F_{uv}(t)| \leq C_2|s - t|,$$

for all $K \subset \mathbb{R}$ compact, $s, t \in K$, and $u, v \in \{0, 1\}$.

In particular, Assumption 2 holds whenever the functions $\{F_{uv}\}_{u,v\in\{0,1\}}$ are differentiable and the derivatives are uniformly bounded above and below, that is, $C_1 \leq F'_{uv}(t) \leq C_2$, for all $t \in K$, and $K \subset \mathbb{R}$ compact. Under this assumption we show that the MBL estimators and the plug-in estimators have mean squared error converging to zero, after rescaling by $\sqrt{n}$. Recall that $H := \sum_{u,v\in\{0,1\}} \eta_{uv} F_{uv}$ is the true population outcome distribution of the outcome variable Y, and define $J_\kappa = [H^{-1}(\kappa), H^{-1}(1 - \kappa)]$.

THEOREM 3.2. *Fix $0 < \kappa < 1/2$. Then the MBL estimators $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}})$ (3.5) and the plug-in estimators $(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$ (recall (2.14) and (2.11)), satisfy*

$$\frac{1}{n} \sum_{b \in I_\kappa} ||\sqrt{n}\{\hat{\boldsymbol{F}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b)})\}||_2^2 = o_P(1), \tag{3.9}$$

and

$$\int_{J_\kappa} ||\sqrt{n}\{\hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t)\}||_2^2 \mathrm{d}H = o_P(1), \tag{3.10}$$

where the $o_P(1)$ term goes to zero as $n \to \infty$. Moreover, $||\sqrt{n}(\hat{\boldsymbol{\phi}} - \breve{\boldsymbol{\phi}})||_2^2 = o_P(1)$.

The proof of Theorem 3.2 is given in Appendix B. The main technical step in the proof is to show that the difference between the values of the objective function $\mathbb{M}_n$ evaluated at the MBL estimators $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}})$ and at the plug-in estimators $(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$ is small, that is,

$$\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) = o_P(1/n). \tag{3.11}$$

To show this, we consider $\tilde{\boldsymbol{F}} := (\tilde{F}^{(0)}_{co}(t), \tilde{F}_{at}(t), \tilde{F}_{nt}(t), \tilde{F}^{(0)}_{co}(t))$ that is the least-square projection of $\breve{\boldsymbol{F}}$ onto the space of distributions functions (see Definition B.1). This can be computed using the PAVA algorithm (Appendix E.1), and the closed form expression for $\tilde{\boldsymbol{F}}$ is well-known, see, for instance, Robertson, Wright and Dykstra (1988). Then using standard concentration inequalities, we can show that

$$\sum_{b\in I_\kappa}\left(\tilde{F}_{co}^{(u)}(Y_{(b)}) - \breve{F}_{co}^{(u)}(Y_{(b)})\right)^2 = o_P(n^{-\frac{1}{4}}) \tag{3.12}$$

for $u \in \{0,1\}$; this result might be of independent interest. The result of (3.12), together with the definition of $\mathbb{M}_n$ implies (3.11) (Proposition B.1). The proof of Theorem 3.2 can then be completed by a Taylor-series approximation of the objective function $\mathbb{M}_n$ (Lemma B.5). However, since the number of parameters grow with the sample size and the proportion of the compliance classes are unknown, the remainder terms have to be carefully analyzed.

As the plug-in estimators $(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$ are consistent estimators of $(\boldsymbol{F}, \boldsymbol{\phi})$, Theorem 3.2 implies that the MBL estimators are also consistent, that is,

$$\int_{J_\kappa} ||\hat{\boldsymbol{F}}(t) - \boldsymbol{F}(t)||_2^2 \mathrm{d}H = o_P(1).$$

Moreover, the plug-in estimators and the MBL estimators have the same limiting distribution, which can be derived using the Brownian bridge approximation of the empirical distribution functions. This is derived in Corollary B.2 in Appendix B.3.

REMARK 3.2. The proof of Theorem 3.2 can be easily modified to show finite dimensional convergence, that is, for every $s \geq 1$ and given $t_1 < t_2 \cdots < t_s$,

$$(||\sqrt{n}(\hat{\boldsymbol{F}}(t_a) - \breve{\boldsymbol{F}}(t_a))||_2^2)_{a\in[s]} = o_P(1).$$

This would imply that the finite dimensional distributions of the plug-in estimate process $\sqrt{n}(\breve{\boldsymbol{F}}(t) - \boldsymbol{F}(t))$ and the MBL estimate process $\sqrt{n}(\hat{\boldsymbol{F}}(t) - \boldsymbol{F}(t))$ are asymptotically the same. We present this result in terms of the mean-squared error as in (3.10), because it emerges naturally from the asymptotics of the BL function, and can be directly applied to the analysis of the binomial likelihood ratio test that is introduced in Section 6.

4. Computation of the MBL estimate. There are no closed form solutions to the MBL estimators. However, they can be computed efficiently by using a combination of the expectation-maximization (EM) algorithm and the pool-adjacent-violator algorithm (PAVA). The procedure begins by computing the *complete data binomial likelihood* which encodes the information of the unobserved compliance class membership. Then the complete data binomial likelihood is maximized using the EM algorithm. However, since the true parameters $\boldsymbol{F}(Y_{(b)})_{b\in I_\kappa}$ need to satisfy the non-decreasing condition, the maximization step in the EM algorithm is implemented using the PAVA algorithm, which enforces the monotonicity condition. We call this combined algorithm the EM-PAVA algorithm.

To begin with, we introduce the *complete data* $\overline{\mathcal{D}}_n$ including the compliance class membership $\boldsymbol{S}$, $\overline{\mathcal{D}}_n := ((Y_1, Z_1, D_1, S_1), (Y_2, Z_2, D_2, S_2), \ldots, (Y_n, Z_n, D_n, S_n))'$. The variable S_a is co if a is a complier. Similarly, $S_a = nt$ if a is a never-taker and $S_a = at$ if the subject a is an always-taker.

Note that if Z_a and S_a are known, then D_a is definitely determined, for example, if $Z_a = 0$ and $S_a = co$, then $D_a = 0$. For $u \in \{0, 1\}$, denote the event $K^a_{u,s} := \{Z_a = u, S_a = s\}$, where $s \in \{co, at, nt\}$. As in (3.2), given the complete data, we can define a *two-point complete likelihood function* for the data points (Y_a, Y_b) as follows:

$$\overline{L}^{(u,s)}_{Y_a,Y_b}(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon|\overline{\mathcal{D}}_n) = \begin{cases} \mathbb{P}(Y_a \le Y_b) & \text{on } K^a_{u,s} \text{ and } \{Y_a \le Y_b\} \\ \mathbb{P}(Y_a > Y_b) & \text{on } K^a_{u,s} \text{ and } \{Y_a > Y_b\} \\ 1 & \text{otherwise} \end{cases}$$
$$= \begin{cases} \mathbb{P}(K^a_{u,s}) \cdot \mathbb{P}(Y_a \le Y_{(b)}|K^a_{u,s}) & \text{on } K^a_{u,s} \text{ and } \{Y_a \le Y_b\} \\ \mathbb{P}(K^a_{u,s}) \cdot \mathbb{P}(Y_a > Y_{(b)}|K^a_{u,s}) & \text{on } K^a_{u,s} \text{ and } \{Y_a > Y_b\} \\ 1 & \text{otherwise.} \end{cases} \tag{4.1}$$

As in (3.4), this can be combined over a, b, u and s to get the *complete binomial likelihood.*

$$\overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon|\overline{\mathcal{D}}_n) := \prod_{b \in I_\kappa} \prod_{a=1}^{n} \prod_{s \in \{co,at,nt\}} \prod_{u \in \{0,1\}} \overline{L}^{(u,s)}_{Y_a,Y_b}(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon|\overline{\mathcal{D}}_n). \tag{4.2}$$

It is easy to see that

$$\mathbb{P}(Y_a \le t|K^a_{u,co}) = \theta^{(u)}_{co},\ \mathbb{P}(Y_a \le t|K^a_{u,nt}) = \theta_{nt},\ \mathbb{P}(Y_a \le t|K^a_{u,at}) = \theta_{at}, \tag{4.3}$$

and

$$\mathbb{P}(K^a_{u,s}) = \begin{cases} \upsilon^u(1-\upsilon)^{1-u}\chi_{co} & \text{if } s = co \\ \upsilon^u(1-\upsilon)^{1-u}\chi_{at} & \text{if } s = at \\ \upsilon^u(1-\upsilon)^{1-u}\chi_{nt} & \text{if } s = nt. \end{cases} \tag{4.4}$$

Using this (4.2) can be computed as function of $\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon$, and, as in the proof of Proposition 3.1 (see Appendix A.2), the dependence on υ in the complete data likelihood is separable, that is,

$$\log \overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon|\overline{\mathcal{D}}_n) = \ell(\upsilon) + \log \overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi}|\overline{\mathcal{D}}_n) \tag{4.5}$$

where $\ell(\upsilon) = |I_\kappa| \{(n_{00} + n_{01}) \log(1-\upsilon) + (n_{10} + n_{11}) \log \upsilon\}$, and $\log \overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi}|\overline{\mathcal{D}}_n)$ does not depend on υ. Hereafter, we will refer to $\log \overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi}|\overline{\mathcal{D}}_n)$ as the *complete data binomial log-likelihood*. This can now be used to devise a EM-type algorithm for computing the MBL estimates: Initialize $\boldsymbol{\theta}_{(0)}, \boldsymbol{\chi}_{(0)}$ to some random values that lie in the parameter space $\vartheta_+ \times [0,1]^2_+$. Given the estimates $\boldsymbol{\theta}_{(m)}, \boldsymbol{\chi}_{(m)}$ at the m-th step of the iteration, the $(m+1)$-th step is as follows:

– *(Expectation Step)* The expectation step of the $(m+1)$-th iteration of the algorithm is as follows: We denote the outputs of the m-th iteration as

$$\hat{\boldsymbol{\theta}}_{(m)} = (\boldsymbol{\theta}^{(0)}_{co,(m)}, \boldsymbol{\theta}_{nt,(m)}, \boldsymbol{\theta}^{(1)}_{co,(m)}, \boldsymbol{\theta}_{at,(m)})$$

and $\hat{\boldsymbol{\chi}}_{(m)} = (\chi_{nt,(m)}, \chi_{at,(m)})$. Given these outputs and the observed data $\mathcal{D}_n = ((Y_1, Z_1, D_1), (Y_2, Z_2, D_2), \ldots, (Y_n, Z_n, D_n))'$, the expected complete binomial log-likelihood given $\mathcal{D}_n$ is

$$Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}) := \mathbb{E}_{\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}} \left(\log \overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi}|\overline{\mathcal{D}}_n) \,|\, \mathcal{D}_n\right). \tag{4.6}$$

This can be easily calculated (see Lemma E.1 in Appendix E.2.1 for details).

– *(Maximization Step)* To begin with, define

$$(\breve{\boldsymbol{\theta}}_{(m+1)}, \hat{\boldsymbol{\chi}}_{(m+1)}) = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}, \boldsymbol{\chi} \in \mathbb{R}^2} Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}). \tag{4.7}$$

Note that

$$\breve{\boldsymbol{\theta}}_{(m+1)} = \begin{pmatrix} \breve{\boldsymbol{\theta}}^{(0)}_{co,(m+1)} \\ \breve{\boldsymbol{\theta}}_{nt,(m+1)} \\ \breve{\boldsymbol{\theta}}^{(1)}_{co,(m+1)} \\ \breve{\boldsymbol{\theta}}_{at,(m+1)}. \end{pmatrix}$$

where $\breve{\boldsymbol{\theta}}^{(0)}_{co,(m+1)} = (\breve{\theta}^{(0)}_{co,(m+1)}(Y_{(b)}))_{b \in I_\kappa}$, and similarly for $\breve{\boldsymbol{\theta}}_{nt,(m+1)}$, $\breve{\boldsymbol{\theta}}^{(1)}_{co,(m+1)}$, and $\breve{\boldsymbol{\theta}}_{at,(m+1)}$.

Observe that $(\breve{\boldsymbol{\theta}}_{(m+1)}, \hat{\boldsymbol{\chi}}_{(m+1)})$ is the unrestricted maximizer of $Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)})$. Hence, it can easily computed by solving the first-order conditions by taking the gradient of the $Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\boldsymbol{\theta}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)})$ with respect to $(\boldsymbol{\theta}(Y_{(b)}))_{b \in I_\kappa}$ and $\boldsymbol{\chi}$, and equating it to zero. The estimates can be found in Lemma E.3. Moreover, Lemma E.3 also shows that $\hat{\boldsymbol{\chi}}^{(m+1)} = (\hat{\chi}^{(m+1)}_{nt}, \hat{\chi}^{(m+1)}_{at})'$ is actually in the restricted space $[0,1]^2_+$, that is, $\hat{\chi}^{(m+1)}_{nt}, \hat{\chi}^{(m+1)}_{at} \in [0,1]$ and $0 \leq \chi^{(m+1)}_{nt} + \chi^{(m+1)}_{at} \leq 1$. This implies

$$\begin{aligned}(\breve{\boldsymbol{\theta}}_{(m+1)}, \hat{\boldsymbol{\chi}}_{(m+1)}) &= \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}, \boldsymbol{\chi} \in [0,1]^2} Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}) \\ &= \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}, \boldsymbol{\chi} \in [0,1]^2_+} Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}). \end{aligned} \tag{4.8}$$

However, in general, $\breve{\boldsymbol{\theta}}_{(m+1)} \notin \boldsymbol{\vartheta}_+$, because $\breve{\boldsymbol{\theta}}_{(m+1)}$ may not satisfy the non-decreasing condition of distribution functions. To ensure the monotonicity constraint we apply the PAVA algorithm (refer to Appendix E.1 for more details on the PAVA algorithm) to the estimate $\breve{\boldsymbol{\theta}}^{(m)}$ as follows:[1]

$$\hat{\boldsymbol{\theta}}_{(m+1)} := \mathrm{PAVA}_{\boldsymbol{w}} \begin{pmatrix} \breve{\boldsymbol{\theta}}^{(0)}_{co,(m+1)} \\ \breve{\boldsymbol{\theta}}_{nt,(m+1)} \\ \breve{\boldsymbol{\theta}}^{(1)}_{co,(m+1)} \\ \breve{\boldsymbol{\theta}}_{at,(m+1)} \end{pmatrix} \tag{4.9}$$

where the PAVA operation is applied coordinate-wise and the weight vector is $\boldsymbol{w}_{(m+1)} = (\boldsymbol{w}^{(0)}_{co,(m+1)}, \boldsymbol{w}_{nt,(m+1)}, \boldsymbol{w}^{(1)}_{co,(m+1)}, \boldsymbol{w}_{at,(m+1)})'$ with weights as defined in (E.7) in Appendix E.2.2.

The following result shows that the output of the PAVA algorithm with the weights as above, indeed maximize the expected complete data log-likelihood under the non-decreasing condition:

PROPOSITION 4.1. *Let $\breve{\boldsymbol{\theta}}_{(m)}$ be as defined in (4.7). Then*

$$(\hat{\boldsymbol{\theta}}_{(m+1)}, \hat{\boldsymbol{\chi}}_{(m+1)}) = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_+, \boldsymbol{\chi} \in [0,1]^2_+} Q_m(\boldsymbol{\theta}, \boldsymbol{\chi} | \hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}),$$

where $\hat{\boldsymbol{\theta}}_{(m+1)}$ is defined above in (4.9).

The proof of the above result is given in Appendix E.2.2. To summarize, the maximization step can be completed by using the following two step procedure: (1) obtain $(\breve{\boldsymbol{\theta}}_{(m+1)}, \hat{\boldsymbol{\chi}}_{(m+1)})$ by maximizing $Q_m(\boldsymbol{\theta}, \boldsymbol{\chi} | \hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)})$ over the unrestricted parameter spaces and (2) obtain $(\hat{\boldsymbol{\theta}}_{(m+1)}, \hat{\boldsymbol{\chi}}_{(m+1)})$ by applying the PAVA to $(\breve{\boldsymbol{\theta}}_{(m+1)}, \hat{\boldsymbol{\chi}}_{(m+1)})$.

The entire expectation and maximization steps are repeated until the values converge.

5. Simulation. In this section we conduct a simulation study to evaluate the performance of our proposed MBL estimation method. We consider three different methods to estimate the outcome CDFs for compliers; the MBL method, the plug-in estimation method described in Section 2.2 and a parametric normal mixture method of Imbens and Rubin (1997). The parametric normal mixture model assumes that all outcome distributions for compliance classes have normal distributions. Then, using the EM algorithm, it estimates the means and the variances of the outcome distributions. Specifically, we consider two simulation scenarios: (1) normal mixture models and (2) gamma mixture models.

Moreover, in each scenario, we consider two more factors that can affect the performance of these three methods. First, we consider whether there is any effect of the treatment for compliers, that is, whether the outcome distributions of $F^{(0)}_{co}$ and $F^{(1)}_{co}$ are the same or not

[1]Recall that PAVA algorithm takes input a vector $\boldsymbol{u} = (u_1, \ldots, u_n)'$ and an a weight vector $\boldsymbol{w} = (w_1, \ldots, w_n)'$, and returns another vector $\mathrm{PAVA}_{\boldsymbol{w}}(\boldsymbol{u}) := (\hat{u}_1, \ldots, \hat{u}_n)'$ which minimizes $\sum_{i=1}^n w_i(u_i - v_i)^2$, under the constraint that $v_1 \leq v_2 \leq \cdots \leq v_n$.

TABLE 3

Normal Mixture. The average performance comparison between the MBL method, the plug-in method and the parametric normal mixture method when the true distributions are normal; L_2 dist. means the average discrepancy from the true CDF.

			MBL		Plug-in		Parametric	
Causal effect	IV	Z	L_2 dist.	SE	L_2 dist.	SE	L_2 dist.	SE
No	Strong	$Z=0$	0.0030	0.0028	0.0031	0.0029	0.0016	0.0023
No	Strong	$Z=1$	0.0030	0.0028	0.0031	0.0029	0.0016	0.0023
Some	Strong	$Z=0$	0.0030	0.0030	0.0031	0.0031	0.0017	0.0026
Some	Strong	$Z=1$	0.0030	0.0028	0.0031	0.0029	0.0017	0.0025
No	Weak	$Z=0$	0.0287	0.0307	0.0951	0.7838	0.0182	0.0287
No	Weak	$Z=1$	0.0274	0.0312	0.0934	0.8526	0.0188	0.0312
Some	Weak	$Z=0$	0.0277	0.0288	0.0764	0.5993	0.0176	0.0269
Some	Weak	$Z=1$	0.0301	0.0343	0.0773	0.3546	0.0184	0.0272

the same. Second, we consider whether the IV is strong or weak. The strength of an IV is how strongly the IV is associated with the treatment. One common definition of a *weak* IV is that the first stage F-statistic when the treatment is regressed on the IV is less than 10 (Stock, Wright and Yogo, 2002). We consider a *strong* IV setting where the proportions of subpopulations (co, nt, at) is (1/3, 1/3, 1/3) (average first stage F-statistic≈124) and a weak IV setting where proportions of $(co, nt, at) = (0.10, 0.45, 0.45)$ (average first stage F-statistic≈10).

Every simulation is repeated for 10,000 times with the sample size $n = 1000$, and the performance of estimating the true CDFs is compared in terms of the *average discrepancy*, where measurement of the discrepancy between the estimated and the true CDFs is defined by L_2 distance, that is, if the true CDF is F and our estimated CDF is $\hat{F}$, the L_2 distance is

$$L_2(F, \hat{F}) = \int (F(x) - \hat{F}(x))^2 \mathrm{d}F(x). \tag{5.1}$$

EXAMPLE 2. (Normal Mixture) For the case that there is no causal effect of Z_1 on Y_1 for the compliers we consider

$$F_{co}^{(0)} = F_{co}^{(1)} \sim N(0, 4^2),\ F_{nt} \sim N(2, 4^2),\ F_{at} \sim N(-2, 4^2),$$

and for the case that there is some effect of Z_1 on Y_1 we consider

$$F_{co}^{(0)} \sim N(1, 4^2),\ F_{co}^{(1)} \sim N(-1, 4^2),\ F_{nt} \sim N(2, 4^2), F_{at} \sim N(-2, 4^2).$$

Table 3 shows simulation results for the normal mixture case. It is shown that the average performance of the MBL method is better than that of the plug-in estimation method in all

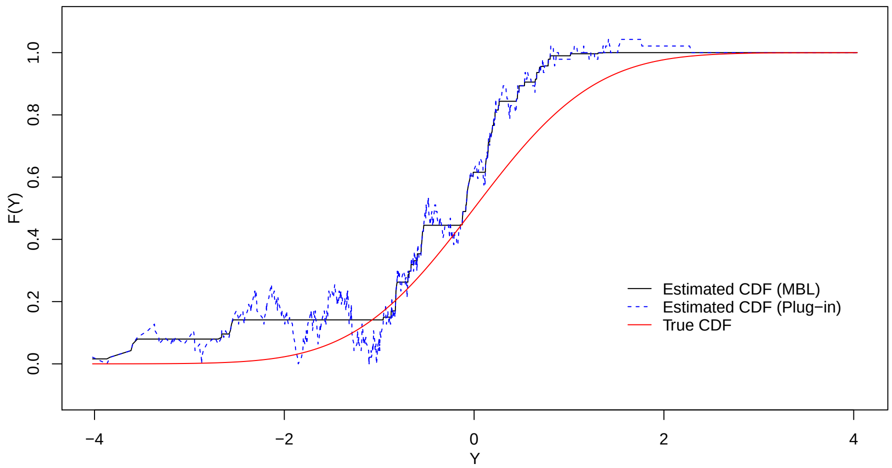

Fig 1: A simulation result for the MBL estimate and the plug-in estimate when the IV is weak.

cases, and it is not much different from the average performance of the parametric normal mixture method that assumes the correct parametric distributions. All three method are greatly affected by the IV being weak. In particular, the plug-in method is more sensitive to the IV being weak; the plug-in estimation method and the MBL method perform similarly when the IV is strong but the MBL method is much better than the plug-in estimation method when the IV is weak. Figure 1 shows the comparison between the plug-in estimate and the MBL estimate when the IV is weak for a randomly selected simulation. Even when $n = 1000$ is comparatively large, the plug-in estimate is significantly fluctuating. The BL estimate, the staircase-like solid curve, is much smoother and closer to the true CDF than the plug-in estimate, the dashed curve. Moreover, the plug-in estimate violates the non-decreasing condition for a CDF while the MBL estimate respects the non-decreasing condition.

The above example shows that, under the normal assumption, the normal parametric mixture method is, as expected, the best method of the three methods. However, the assumption of normality is a strong assumption. If normality does not hold, then the normal parametric method is no longer the best, as illustrated in the following example:

Example 3. (Gamma Mixture) Let $\Gamma(\alpha, \beta)$ be a Gamma distribution with shape α and rate β. For the case that there is no causal effect of Z_1 on Y_1 for the compliers we consider

$$F_{co}^{(0)} = F_{co}^{(1)} \sim \Gamma(1.2, 1),\ F_{nt} \sim N(1, 1),\ F_{at} \sim N(1.4, 1),$$

TABLE 4
Gamma Mixture case. The average performance comparison between the MBL method, the plug-in estimation method and the parametric normal mixture method when the true distributions are non-normal; L_2 dist. means the average discrepancy from the true CDF.

			MBL		Plug-in		Parametric	
Causal effect	IV	Z	L_2 dist.	SE	L_2 dist.	SE	L_2 dist.	SE
No	Strong	$Z=0$	0.0030	0.0029	0.0031	0.0030	0.0054	0.0128
No	Strong	$Z=1$	0.0029	0.0029	0.0030	0.0030	0.0061	0.0172
Some	Strong	$Z=0$	0.0029	0.0027	0.0030	0.0028	0.0148	0.0443
Some	Strong	$Z=1$	0.0028	0.0027	0.0029	0.0027	0.0198	0.0608
No	Weak	$Z=0$	0.0290	0.0310	0.1048	0.9679	0.0888	0.1239
No	Weak	$Z=1$	0.0271	0.0312	0.1116	1.6740	0.0753	0.1266
Some	Weak	$Z=0$	0.0280	0.0294	0.0523	0.1292	0.0711	0.1173
Some	Weak	$Z=1$	0.0271	0.0285	0.0508	0.1039	0.1047	0.1371

and for the case that there is some effect of Z_1 on Y_1 we consider that

$$F_{co}^{(0)} \sim \Gamma(1.1, 1),\ F_{co}^{(1)} \sim \Gamma(1.3, 1),\ F_{nt} \sim N(1, 1),\ F_{at} \sim N(1.4, 1).$$

The results are summarized in Table 4. It shows that the MBL method is the dominant method in all scenarios considered when normality is not satisfied. Though the plug-in estimation method has a similar performance in the strong IV setting, it is much worse than the MBL method in the weak IV setting as in the normal mixture model setting. Also, the parametric normal mixture method has significantly increased average discrepancies with large standard errors.

The above examples show that MBL method is robust to any distribution assumption and is the least sensitive to the IV being weak.

6. Testing for distributional treatment effect. A central question in many observational studies is to understand if the treatment has any effect on the distribution of an outcome. Under the IV assumptions, this corresponds to testing whether the treatment has any effect on the outcome distribution for compliers. The null hypothesis can be formulated by as follows:

$$H_0 : F_{co}^{(0)}(t) = F_{co}^{(1)}(t),\ \text{for all } t \in \mathbb{R}. \tag{6.1}$$

The MBL method can be used to construct a likelihood-ratio type test statistic in a similar way that the maximum likelihood estimation method can be extended to constructing a likelihood ratio test. We take the difference in two binomial log-likelihood values; one is obtained with the constraint $F_{co}^{(0)} = F_{co}^{(1)}$ and the other is obtained without this constraint. This gives a new test for detecting the distributional treatment effect (6.1), and hereafter, we call it the *binomial likelihood ratio test* (BLRT), which is described in Section 6.2.

Before introducing the BLRT, we will quickly review the existing method for testing (6.1) from Abadie (2002).

6.1. *Existing approach.* The existing approach for testing (6.1) is based on the following observation, which is obtained by taking the difference of (2.12) and (2.13):

$$F_{co}^{(0)}(t) - F_{co}^{(1)}(t) = K \cdot (F_0(t) - F_1(t)), \tag{6.2}$$

where $F_0(t) = \mathbb{P}(Y_1 \leq t | Z_1 = 0)$, $F_1(t) = \mathbb{P}(Y_1 \leq t | Z_1 = 1)$, and $K = 1/(\mathbb{E}(D_1|Z_1 = 1) - \mathbb{E}(D_1|Z_1 = 0)) < \infty$.

This implies that the null hypothesis of no distributional treatment effect is the same as the testing $F_0 = F_1$. Note that the distributions F_0 and F_1 can be estimated directly by their corresponding empirical distribution functions:

$$\overline{F}_0(t) = \frac{\sum_{b=1}^n \mathbf{1}\{Y_b \leq t, Z_b = 0\}}{\sum_{b=1}^n \mathbf{1}\{Z_b = 0\}} \quad \text{and} \quad \overline{F}_1(t) = \frac{\sum_{b=1}^n \mathbf{1}\{Y_b \leq t, Z_b = 1\}}{\sum_{b=1}^n \mathbf{1}\{Z_b = 1\}}.$$

Therefore, Abadie (2002) proposed using the well-known Kolmogorov-Smirnov test statistic

$$T_{\mathrm{KS}} := \sup_{t \in \mathbb{R}} |\overline{F}_0(t) - \overline{F}_1(t)|,$$

for testing (6.1). In other words, the test proposed by Abadie (2002) is the KS test of whether the distribution of the $Z_1 = 0$ group is the same as the distribution of the $Z_1 = 1$ group. The KS test does not make any use of the structure of the IV model. Also, it does not require accurate estimation of the two distributions $F_{co}^{(0)}$ and $F_{co}^{(1)}$, which often leads to a comparatively low power when sample size is not large. Furthermore, when an IV is weak (that is, the proportion of compliers ϕ_{co} is small), the power of the KS test gets worse because the KS test reduces the difference between $F_{co}^{(0)}$ and $F_{co}^{(1)}$ by a factor $1/K$ from (6.2).

6.2. *Binomial likelihood ratio test.* The MBL estimators of the true CDFs $\boldsymbol{F}$, can be obtained by maximizing the binomial log-likelihood $\mathbb{M}_n$ over the restricted parameter space as introduced in Section 3. The same maximizing scheme can be used to estimate the outcome CDFs under the null by imposing an additional restriction of $F_{co}^{(0)} = F_{co}^{(1)}$ over the *restricted null parameter space* that is defined as

$$\boldsymbol{\vartheta}_{+,0} = \left\{(\theta_{co}, \theta_{nt}, \theta_{at}) : \theta_{co}, \theta_{nt}, \theta_{at} \in \mathcal{P}([0,1]^{\mathbb{R}})\right\}, \tag{6.3}$$

where $\mathcal{P}([0,1]^{\mathbb{R}})$ is the set of distribution functions from $\mathbb{R} \to [0,1]$.

Now, the BLRT statistic is obtained by taking the difference of the binomial log-likelihoods under the null and the alternative:

$$T_n := \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_+} \mathbb{M}_n(\boldsymbol{\theta}, \breve{\boldsymbol{\phi}}) - \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_{+,0}} \mathbb{M}_n(\boldsymbol{\theta}, \breve{\boldsymbol{\phi}}), \tag{6.4}$$

where $\boldsymbol{\vartheta}_+$ and $\boldsymbol{\vartheta}_{+,0}$ are the restricted parameter space the restricted null parameter space, defined in (2.10) and (6.3), respectively. Note that, unlike in Section 3, we estimate the proportion of the compliance classes $\boldsymbol{\phi}$ by the plug-in estimates $\breve{\boldsymbol{\phi}}$, instead of maximizing the binomial log-likelihood for the parameter $\boldsymbol{\phi}$. Both the approaches have very similar finite sample performances, since $\breve{\boldsymbol{\phi}} \in [0,1]^2_+$ with probability 1, and using $\breve{\boldsymbol{\phi}}$ simplifies the asymptotic analysis of the BLRT.

THEOREM 6.1. *Let T_n be the BLRT statistic as defined in (6.4). Then, for $0 < \kappa < 1/2$ fixed,*

$$T_n = \frac{1}{n} \sum_{b \in I_\kappa} \frac{(\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(1)}(Y_{(b)}))^2}{\sum_{u,v \in \{0,1\}} \frac{1}{\breve{\lambda}_{uv}^2} \frac{n}{n_{uv}} \overline{F}_{uv}(Y_{(b)})(1 - \overline{F}_{uv}(Y_{(b)}))} + o_P(1), \tag{6.5}$$

where $\breve{\lambda}_{00} = \breve{\lambda}_0$, $\breve{\lambda}_{01} = \frac{\breve{\lambda}_1}{1-\breve{\lambda}_1}$, $\breve{\lambda}_{10} = \frac{\breve{\lambda}_0}{1-\breve{\lambda}_0}$, and $\breve{\lambda}_{11} = \breve{\lambda}_1$.

The above theorem gives an asymptotically equivalent representation of the BLRT statistic, which can be used to devise a bootstrap strategy for simulating the null distribution of BLRT statistic and constructing the rejection region. To this end, define

$$\hat{\boldsymbol{\psi}} = (\hat{\psi}_{co}, \hat{\psi}_{nt}, \hat{\psi}_{at})' := \arg \max_{\boldsymbol{\psi} \in \boldsymbol{\vartheta}_{+,0}} \mathbb{M}_n(\boldsymbol{\psi}, \breve{\boldsymbol{\phi}}).$$

Let B be the number of bootstrap samples. For $1 \leq t \leq B$, we have the following steps:

- Fix $\boldsymbol{Z}^{(t)} = \boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)'$ and sample the compliance class membership $\boldsymbol{S}^{(t)} = (S_1^{(t)}, S_2^{(t)}, \ldots, S_n^{(t)})'$ based on the estimated proportions $\breve{\boldsymbol{\phi}}$. Then, determine $\boldsymbol{D}^{(t)} = (D_1^{(t)}, D_2^{(t)}, \ldots, D_n^{(t)})'$ based on $\boldsymbol{Z}^{(t)}$ and $\boldsymbol{S}^{(t)}$. For example, if $Z_a^{(t)} = 1$ and $S_a^{(t)} = nt$, then $D_a^{(t)} = 0$.
- From the MBL estimate $\hat{\boldsymbol{\psi}}$ under the null, take a sample of $\boldsymbol{Y}^{(t)} = (Y_1^{(t)}, Y_2^{(t)}, \ldots, Y_n^{(t)})'$ based on $\boldsymbol{D}^{(t)}$ and $\boldsymbol{S}^{(t)}$. For example, if $D_a^{(t)} = 0$ and $S_a^{(t)} = nt$, then $Y_a^{(t)}$ is a random sample from the estimated distribution function $\hat{\psi}_{nt}$.
- Repeat the two steps above for $1 \leq t \leq B$. Then use the bootstrapped samples $\{(\boldsymbol{Z}^{(t)}, \boldsymbol{D}^{(t)}, \boldsymbol{Y}^{(t)})\}_{1 \leq t \leq B}$ for simulating the asymptotic null distribution of T_n using the expression on the RHS of (6.5).

The statistic on the RHS of (6.5) takes the form a two-sample Anderson-Darling test (Pettitt, 1976) based on the plug-in estimators $\breve{F}_{co}^{(0)}$ and $\breve{F}_{co}^{(1)}$, conditioning on $(\boldsymbol{Z}, \boldsymbol{D})$ and truncated by the interval $[\kappa, 1-\kappa]$.[2] We elaborate on this connection in the following remark:

[2]Given i.i.d. samples from two distributions F and G of sizes n and m respectively, the two-sample Anderson-Darling test (Pettitt, 1976) for testing the null hypothesis $H_0 : F = G$, is $A_{mn} = \int \frac{(F_m(t) - G_n(t))^2}{H_N(t)(1-H_N(t))} \mathrm{d}H_N(t)$, where F_m and G_n are the empirical distributions of F and G, respectively, and $H_N = \frac{m}{N} F_m + \frac{n}{N} G_n$.

Remark 6.1. Recall that $H = \eta_{00}F_{00} + \eta_{01}F_{01} + \eta_{10}F_{10} + \eta_{11}F_{11}$. Now, referring to (D.15), we have

$$n\,\mathrm{Var}(\breve{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(1)}(t)|(\boldsymbol{Z},\boldsymbol{D})) = \sum_{u,v\in\{0,1\}} \frac{1}{\breve{\lambda}_{uv}^2}\frac{n}{n_{uv}}F_{uv}(t)(1-F_{uv}(t)).$$

Then by the Riemann sum approximation, (6.5) can be re-written as

$$T_n = \int_{J_\kappa} \frac{(\breve{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(1)})^2}{\mathrm{Var}(\breve{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(0)}(t)|(\boldsymbol{Z},\boldsymbol{D}))}\mathrm{d}H(t) + o_P(1), \tag{6.6}$$

Now, let $\Delta_n(t) := \mathbb{E}(\breve{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(1)}(t)|(\boldsymbol{Z},\boldsymbol{D}))$, the conditional mean of the numerator in (6.6). Recalling the definitions of the plug-in estimators $\breve{F}_{co}^{(0)}$ and $\breve{F}_{co}^{(1)}$ from (2.14), and noting that $\mathbb{E}(\overline{F}_{uv}(t)|(\boldsymbol{Z},\boldsymbol{D})) = F_{uv}(t)$, for $u,v\in\{0,1\}$, it follows that

$$\begin{aligned}\Delta_n(t) &= \frac{1}{\breve{\lambda}_0}F_{00}(t) - \frac{1-\breve{\lambda}_0}{\breve{\lambda}_0}F_{10}(t) + \frac{1-\breve{\lambda}_1}{\breve{\lambda}_1}F_{01}(t) - \frac{1}{\breve{\lambda}_1}F_{11}(t)\\ &= \frac{\lambda_0}{\breve{\lambda}_0}F_{co}^{(0)}(t) + \left(\frac{\breve{\lambda}_0 - \lambda_0}{\breve{\lambda}_0}\right)F_{nt}(t) + \left(\frac{\lambda_1 - \breve{\lambda}_1}{\breve{\lambda}_1}\right)F_{at}(t) - \frac{\lambda_1}{\breve{\lambda}_1}F_{co}^{(1)}(t).\end{aligned} \tag{6.7}$$

Now, since $\sqrt{n}(\breve{\lambda}_0 - \lambda_0, \breve{\lambda}_1 - \lambda_1)$ has a limiting normal distribution, an application of the Delta theorem shows that $\sqrt{n}\Delta_n(t)$ also converges to a normal distribution. Using this combined with Theorem 6.1 and Lemma B.7, it is possible to derive the limiting distribution of the BLRT. However, in general, the limiting distribution is quite complicated, and for constructing the rejection region, it is more instructive to deal to with (6.5). However, in some special cases, the asymptotic null distribution of the test statistic can be simplified:

- Assume that λ_0 and λ_1 are known, that is, $\breve{\lambda}_0 = \lambda_0$ and $\breve{\lambda}_1 = \lambda_1$. Then, under the null $F_{co}^{(0)}(t) = F_{co}^{(1)}(t) = F_{co}(t)$, (6.7) gives $\mathbb{E}(\breve{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(1)}(t)|(\boldsymbol{Z},\boldsymbol{D})) = 0$. Then from Lemma B.7, $\sqrt{n}(\breve{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(1)}(t)) \Rightarrow G(t)$, where

$$G(t) = \frac{1}{\lambda_{00}}\frac{B_{00}(F_{00}(t))}{\sqrt{\eta_{00}}} - \frac{1}{\lambda_{10}}\frac{B_{10}(F_{10}(t))}{\sqrt{\eta_{10}}} + \frac{1}{\lambda_{01}}\frac{B_{01}(F_{01}(t))}{\sqrt{\eta_{01}}} - \frac{1}{\lambda_{11}}\frac{B_{11}(F_{11}(t))}{\sqrt{\eta_{11}}},$$

  for independent Brownian bridges $B_{00}, B_{01}, B_{10}, B_{11}$, and $\lambda_{00} = \lambda_0$, $\lambda_{01} = \frac{\lambda_1}{1-\lambda_1}$, $\lambda_{10} = \frac{\lambda_0}{1-\lambda_0}$, and $\lambda_{11} = \lambda_1$. Then by Donsker's invariance, under the null hypothesis H_0,

$$T_n \xrightarrow{D} \int_{J_\kappa} \frac{G(t)^2}{\mathrm{Var}(G(t))}\mathrm{d}H(t). \tag{6.8}$$

- Assume that $F_{uv} = F$, for all $u, v \in \{0,1\}$. In this case, (6.8) simplifies even further: Chasing definitions it is easy to check that $\sum_{u,v} \frac{1}{\lambda_{uv}^2 \eta_{uv}} = \frac{1}{\phi_{co}^2 \phi_0(1-\phi_0)}$. Then $G(t) \Rightarrow \frac{1}{\phi_{co}\sqrt{\phi_0(1-\phi_0)}} B(F(t))$, where B is the standard Brownian bridge. This gives, under H_0,

$$T_n \xrightarrow{D} \int_{F^{-1}(\kappa)}^{F^{-1}(1-\kappa)} \frac{B(F(t))^2}{\mathrm{Var}(F(t))} \mathrm{d}F(t)$$
$$= \int_{F^{-1}(\kappa)}^{F^{-1}(1-\kappa)} \frac{B(F(t))^2}{F(t)(1-F(t))} \mathrm{d}F(t)$$
$$= \int_{\kappa}^{1-\kappa} \frac{B(t)^2}{t(1-t)} \mathrm{d}t, \tag{6.9}$$

which is the limiting distribution of the two-sample Anderson-Darling test (Pettitt, 1976) truncated to the interval $[\kappa, 1-\kappa]$.

6.3. *Simulation.* To assess the performance of the proposed BLRT, we compare it to the KS test in Abadie (2002) in a simulation study. Furthermore, the BLRT is compared with the approximated version of the BLRT and the Anderson-Darling (AD) test. The approximated version uses the RHS in (6.5) while the BLRT uses the statistic T_n (6.4). The distributions of never-takers and always-takers were fixed as $F_{nt} \sim N(-1,1)$ and $F_{at} \sim N(1,1)$. Two simulation settings were considered. First, $F_{co}^{(0)}$ and $F_{co}^{(1)}$ have normal distributions with different means when the proportion of compliers ϕ_c is 1/3 (i.e., a strong IV). Second, the same scenario is considered except for the proportion $\phi_c = 0.1$ (i.e., a weak IV). To obtain p-values, the resampling method described in the previous section is considered. For each simulation data, $B = 2000$ resamples were used to approximate the null distribution and obtain the p-value. Size and power of the tests were computed by repeating the experiment over 2000 simulated datasets .

Table 5 shows simulated size and power of the four considered tests. The first row of the table shows that the simulated sizes of the tests are approximately equal to the nominal significance level $\alpha = 0.05$, which implies that all the tests found the correct size. For power comparisons, there are three main findings. First, the exact BLRT is the most powerful in every simulation scenario. When an IV is weak such as $\phi_c = 0.1$, the BLRT significantly outperforms the KS test. Second, the approximated BLRT, KS and AD tests have similar performances although the approximated BLRT has slightly better power than the other two tests have. Finally, as n increases, the performance gap between the exact BLRT and the approximated BLRT gets narrower. Figure 2 represents the simulation results for $n = 300$ in the two settings graphically. The left panel represents the plot of the power versus the size of the additive effect of treatment when $\phi_c = 1/3$. It is shown that the BLRT detects the additive treatment effect better than the other tests. The gain of the BLRT test over the KS test is even greater when ϕ_c is smaller (that is, a weaker IV) as shown

TABLE 5

Size and power of test with a significance level 0.05.

	n	μ	IV	$N(-\mu,1)$ vs. $N(\mu,1)$ BLRT	BLRT Approx.	KS	AD
Size	300	0	Strong	0.051	0.049	0.048	0.050
Power	300	0.1		0.083	0.071	0.073	0.069
	300	0.2		0.184	0.152	0.148	0.148
	300	0.3		0.350	0.305	0.280	0.294
	300	0.4		0.511	0.456	0.430	0.446
	300	0.5		0.705	0.649	0.615	0.631
	300	0.6		0.830	0.790	0.764	0.780
	300	0.7		0.927	0.898	0.867	0.888
	300	0.8		0.965	0.945	0.929	0.938
	300	0.9		0.983	0.971	0.962	0.969
	300	1.0		0.992	0.987	0.980	0.985
	1000	0.2		0.430	0.407	0.397	0.398
	2000	0.2		0.728	0.719	0.701	0.713
Size	300	0	Weak	0.044	0.048	0.048	0.047
Power	300	0.1		0.058	0.053	0.050	0.054
	300	0.2		0.079	0.060	0.066	0.062
	300	0.3		0.098	0.073	0.075	0.072
	300	0.4		0.121	0.086	0.086	0.084
	300	0.5		0.160	0.116	0.108	0.108
	300	0.6		0.182	0.136	0.129	0.128
	300	0.7		0.215	0.172	0.162	0.166
	300	0.8		0.267	0.211	0.197	0.204
	300	0.9		0.280	0.230	0.212	0.216
	300	1.0		0.324	0.270	0.236	0.260

in the right-hand panel. In summary, in the simulation setting considered, the BLRT test dominates the KS test.

7. The effect of Vietnam era military service on future earnings. We consider the sample of 11,637 white men, born in 1950-1953, from the March Current Population Surveys of 1979 and 1981-1985 as in Angrist (1990). An indicator of draft-eligibility based on the Vietnam draft lottery outcome (men with lottery number below the ceiling are referred to here as "draft-eligible") is the instrumental variable. Also, veteran status is the non-randomized treatment variable and annual earning (in 1978 dollars) is the outcome variable. These three variables are available for every individual in the sample. More details are provided in Angrist (1990).

Figure 3 shows the estimated CDFs of veterans and non-veterans for compliers. The left plot is the plug-in method described in Section 2.2 and the right one is derived from the MBL method. In addition to the estimates of the CDFs, 95% confidence bands are included

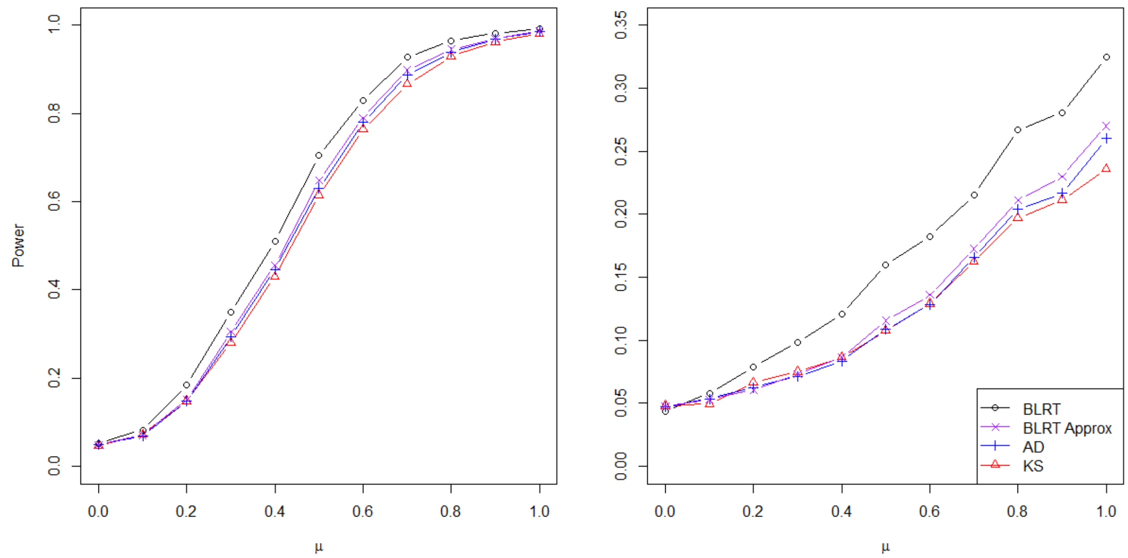

Fig 2: Power of the BLRT, approximated BLRT, KS and AD tests. Power is calculated given a significance level $\alpha = 0.05$. Left: strong IV and Right: weak IV.

in the right plot (Appendix F). In both of the plots, we see that veterans' estimated CDF is almost always above the non-veterans' estimated CDF. The gap between the two CDFs is wide at a small earning range but decreases until earnings of $35,000. The MBL method improves two features of the plug-in method in this example. The MBL estimates for compliers do not violate the non-decreasing and non-negative conditions. The improvement leads to much smoother CDFs. From satisfying the non-decreasing condition, an additional useful feature of the MBL method is obtained. There is a unique value of estimated earnings corresponding to a specific quantile level. This feature can be useful for those who want to estimate the treatment effect at a certain quantile level using the estimated CDFs. A unique estimate might not be obtained in the plug-in method because of the fluctuation of the CDF - if there are multiple values that correspond to the same quantile level, then we cannot acquire the corresponding quantile to estimate the causal treatment effect for compliers of the quantile level. For instance, it is derived from the MBL method that the effect of veteran status is estimated to have a negative impact of $2,514 on earnings for compliers when comparing the medians of veterans versus non-veterans. However, from the plug-in method, there are two earnings values that correspond to the value 0.5 of veteran CDF, making it unclear how to compare the medians of the two CDFs.

Using the BLRT described in Section 6, we conduct the hypothesis test of no distributional treatment effect, $F_{co}^{(0)} = F_{co}^{(1)}$. Both the proposed BLRT statistic T_n and the KS statistic T_{KS} are considered. Table 6 reports p-values for the two tests of equality. From

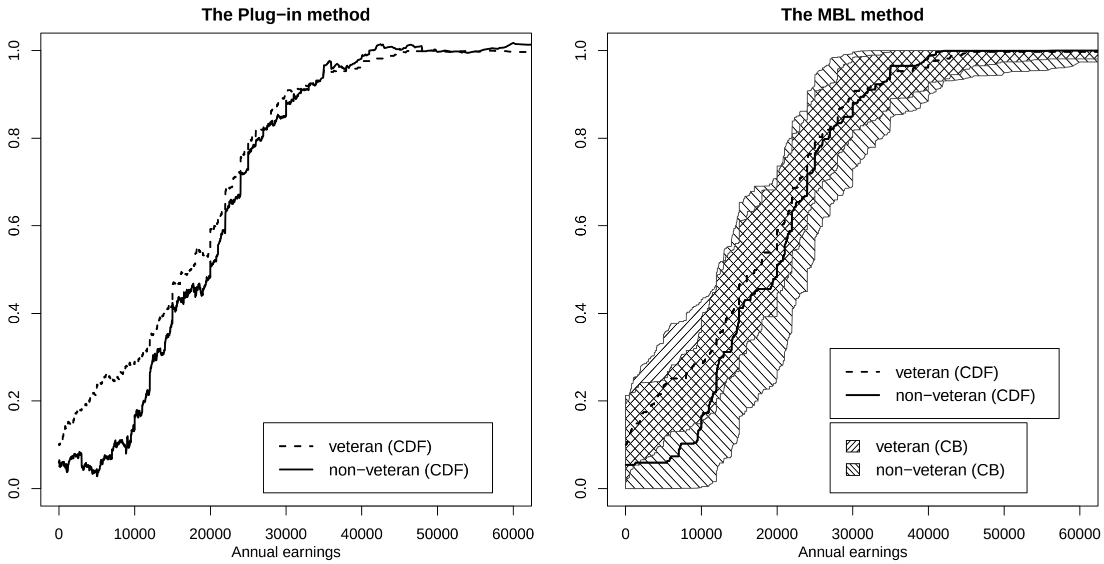

Fig 3: Estimated CDFs of Potential Earnings for Compliers: the CDFs from the plug-in method and the CDFs with 95% confidence bands (CB) from the MBL method.

TABLE 6
Tests on distributional effect of veteran status on annual earnings. BLRT represents the binomial likelihood ratio test and KS represents the Kolmogorov-Smirnov test.

	BLRT	KS
p-value	0.2485	0.1035

the resampling scheme described in Section 6, the p-value of the BLRT is computed by 2000 times bootstrap resampling ($B = 2000$). For both tests, we cannot reject equality of distributions at a significance level $\alpha = 0.05$.

8. Summary. In this paper, we introduced the notion of binomial likelihood, obtained by integrating the individual likelihoods at all observations, for estimating the distributional treatment effect for the compliers. The MBL estimator maintains the non-decreasingness and non-negativeness properties of the distribution functions, which have not been achieved for nonparametric IV estimation before, while preserving the desirable large-sample properties of the plug-in estimates. We also showed that the strength of the MBL method over existing methods is particularly pronounced in the weak IV setting. The MBL estimators can be computed efficiently using a combination of the EM algorithm and pool-adjacent-violator algorithm (PAVA).

Furthermore, we proposed a binomial likelihood ratio test (BLRT) for testing the equality of the compliers' distributions in IV models, and derived its large sample properties. It is shown with simulations that the BLRT is the most powerful to detect distributional changes, among the considered tests. The dominance of the BLRT is emphasized in finite samples with a weak IV. In large samples, the approximated BLRT is an useful alternative.

Acknowledgement: BBB thanks Abhishek Chakrabortty and Shirshendu Ganguly for helpful discussions.

References.

Abadie, A. (2002). Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models. *J. Amer. Statist. Assoc.* **97** 284-292.

Abadie, A. (2003). Semiparametric Instrumental Variable Estimation of Treatment Response Models. *J. Econometrics* **113** 231-263.

Angrist, J. D. (1990). Lifetime Earnins and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *Amer. Econ. Rev.* **80** 313-336.

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996). Idenrification of Causal Effects Using Instrumental Variables. *J. Amer. Statist. Assoc.* **91** 444-455.

Angrist, J. D. and Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *J. Econ. Perspect.* **15** 69-85.

Atkinson, A. B. (1970). On the Measuremnt of Inequality. *J. Econom. Theory* **2** 244-263.

Baiocchi, M., Cheng, J. and Small, D. S. (2014). Instrumental Variable Methods for Causal Inference. *Stat. Med.* **33** 2297-2340.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). *Statistical Inference Under Order Restrictions.* John Wiley & Sons, New York.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A. and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins University Press Baltimore.

Brookhart, M. A. and Schneeweiss, S. (2007). Preference-Based Instrumental Variable Methods for the Estimation of Treatment Effects: Assessing Validity and Interpreting Results. *Int. J. Biostat.* **3** Article 14.

Buja, A. and Rolke, W. (2005). Calibration for Simultaneity: (Re)Sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data. Unpublished Manuscript.

Cheng, J., Qin, J. and Zhang, B. (2009). Semiparametric Estimation and Inference for Distributional and Genenral Treatment Effects. *J. R. Stat. Soc. Ser. B.* **71** 881-904.

de Leeuw, J., Hornik, K. and Mair, P. (2009). Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods. *J. Stat. Softw.* **32** 1-24.

Geman, S. and Hwang, C.-R. (1982). Nonparametric Maximum Likelihood Estimation by the Method of Sieves. *Ann. Statist.* **10** 401-414.

Heagerty, P. J. and Lele, S. R. (1998). A composite likelihood approach to binary spatial data. *J. Amer. Statist. Assoc.* **93** 1099–1111.

Hernan, M. A. and Robins, J. M. (2006). Instruments for Causal Inference: An Epidemiologist's Dream? *Epidemiology* **17** 360-372.

Holland, P. W. (1988). Causal Inference, Path Analysis, and Recursive Structural Equations Models. *Sociol. Methodol.* **18** 449-484.

Holst, L. (1980). On the Lengths of the Pieces of a Stick Broken at Random. *J. Appl. Probab.* **17** 623–634.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica* **62** 467-475.

Imbens, G. W. and Rubin, D. B. (1997). Estimating Outcome Distributions for Compliers in Instrumental Variables Models. *Rev. Econ. Stud.* **64** 555-574.

Karni, E. (2009). A Theory of Medical Decision Making under Uncertainty. *J. Risk Uncertain.* **39** 1-16.

Larribe, F. and Fearnhead, P. (2011). On composite likelihoods in statistical genetics. *Statist. Sinica* **21** 43.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemp. Math.* **80** 221–39.

Ma, Z., Foster, D. and Stine, R. (2015). Adaptive Monotone Shrinkage for Regression. *ArXiv e-prints.*

Neyman, J. (1990). On the application of probability theory to agricultural experiments. *Statist. Sci.* **5** 465–472.

Ogburn, E. L., Rotnitzky, A. and Robins, J. M. (2015). Doubly Robust Estmation of the Local Average Treatment Effect Curve. *J. R. Stat. Soc. Ser. B.* **77** 373-396.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC Press.

Pettitt, A. N. (1976). A Two-Sample Anderson-Darling Rank Statistic. *Biometrika* **Vol. 63** 161–168.

Robertson, T., Wright, F. T. and Dykstra, R. L. (1988). *Order Restricted Statistical Inference* **Wiley Series in Probability and Statistics**. John Wiley & Sons Inc.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66** 688.

Rubin, D. B. (1986). Statistics and Causal Inference: Comment: Which Ifs Have Causal Answers. *J. Amer. Statist. Assoc.* **81** 961-962.

Sen, P. K. (1959). On the moments of sample quantiles. *Calcutta Statist. Assoc. Bull.* **Vol. 9** 1–19.

Stock, J. H., Wright, J. H. and Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *J. Bus. Econom. Statist.* **20** 518-529.

Tan, Z. (2006). Regression and Weighting Methods for Causal Inference Using Instrumental Variables. *J. Amer. Statist. Assoc.* **101** 1607-1618.

Varin, C., Reid, N. and Firth, D. (2011). An overview of composite likelihood methods. *Statist. Sinica* **21** 5–42.

Walker, A. M. (1968). A Note on the Asymptotic Distribution of Sample Quantiles. *J. R. Stat. Soc. Ser. B.* **Vol. 30 (3)** 570–575.

Zelen, M. (1979). A New Design for Randomized Clinical Trials. *New Engl. J. Med.* **300** 1242-1245.

APPENDIX A: PRELIMINARIES

In this section we collect some basic inequalities and properties of the objective function $\mathbb{M}_n$ (3.6).

A.1. Basic inequalities. We begin with a few preliminary observations which will be used later in our proofs:

OBSERVATION A.1. *Fix $a \in [0,1]$. Then for every $x \in [0,1]$, $J(a,x) = a\log x + (1-a)\log(1-x) \le a\log a + (1-a)\log(1-a) := I(a)$.*

PROOF. The inequality is trivially satisfied when $a \in \{0,1\}$. Therefore, assume that $a \in (0,1)$, and define a random variable W which takes values $\frac{x}{a}$ and $\frac{1-x}{1-a}$ with probabilities a and $1-a$, respectively. Note that $\mathbb{E}W = 1$. Then by Jensen's inequality, $\mathbb{E}(\log W) = a\log\frac{x}{a} + (1-a)\log\frac{1-x}{1-a} \le \log \mathbb{E}W = 0$, which completes the proof of the result. □

OBSERVATION A.2. *Fix $a \in [0,1]$. Then for every $x \in [0,1]$, $a\log\frac{x}{a} + (1-a)\log\frac{1-x}{1-a} \le -\frac{1}{2}(x-a)^2$.*

PROOF. For a given $a \in (0,1)$, let $f_a(x) = a\log\frac{x}{a} + (1-a)\log\frac{1-x}{1-a}$. By a second order Taylor expansion around the point a, $f_a(x) = \frac{1}{2}(x-a)^2 f_a''(\gamma_{x,a})$ where $\gamma_{x,a} \in [x\wedge a, x \vee a]$[3] and $f_a''(x) = -\frac{a}{x^2} - \frac{1-a}{(1-x)^2}$. Note that, for $a \in (0,1)$ fixed, the function $f_a''(x)$ is convex. It is easy to check that the minimum is attained at $x_0 = (\frac{a}{1-a})^{\frac{1}{3}}$, and $f_a''(x) \ge f_a''(x_0) > 1$. This implies, $f_a(x) \le -\frac{1}{2}(x-a)^2$. □

OBSERVATION A.3. *Fix $0 < t < 1$. Suppose $Y_1, Y_2, \dots, Y_n$ are i.i.d. samples from the distribution $H = \eta_{00}F_{00} + \eta_{01}F_{01} + \eta_{10}F_{10} + \eta_{11}F_{11}$. Then, for $u, v \in \{0,1\}$*

$$F_{uv}(Y_{\lceil nt \rceil}) \xrightarrow{P} H_{uv}^{-1}(t),$$

where $H_{uv}(t) = \eta_{00}(F_{00}\circ F_{uv}^{-1})(t) + \eta_{01}(F_{01}\circ F_{uv}^{-1})(t) + \eta_{10}(F_{10}\circ F_{uv}^{-1})(t) + \eta_{11}(F_{11}\circ F_{uv}^{-1})(t)$.

PROOF. Without of generality, take $u = 0$ and $v = 0$. Then the distribution of

$$W_1 := F_{00}(Y_1),\ W_2 := F_{00}(Y_2), \dots,\ W_n := F_{00}(Y_n)$$

are i.i.d. samples with distribution function $H_{00}(t) = \eta_{00}t + \eta_{01}(F_{01}\circ F_{00}^{-1})(t) + \eta_{10}(F_{10}\circ F_{00}^{-1})(t) + \eta_{11}(F_{11}\circ F_{00}^{-1})(t)$. This implies, for any $0 < t < 1$, $F_{00}(Y_{(\lceil nt \rceil)}) = W_{(\lceil nt \rceil)} \xrightarrow{P} H_{00}^{-1}(t)$, where the last step uses the convergence of sample quantiles to the corresponding population quantiles (Walker, 1968, Theorem 1). □

[3]For $a, b \in \mathbb{R}$, define $a \wedge b = \min\{a,b\}$ and $a \vee b = \max\{a,b\}$.

A.2. Understanding the objective functions. In this section we discuss properties of the sample objective function $\mathbb{M}_n$ (3.6). We begin with the proof of Proposition 3.1.

A.2.1. *Proof of Proposition 3.1.* To begin recall the definition of $K^a_{uv} = \{Z_a = u, D_a = v\}$. By (2.7), for all $a \in [n]$,

$$\mathbb{P}(K^a_{uv}) := \eta_{uv} = \begin{cases} (1-\upsilon)(1-\chi_{at}) & \text{for} \quad u=0, v=0, \\ (1-\upsilon)\chi_{at} & \text{for} \quad u=0, v=1, \\ \upsilon\chi_{nt} & \text{for} \quad u=1, v=0, \\ \upsilon(1-\chi_{nt}) & \text{for} \quad u=1, v=1. \end{cases} \tag{A.1}$$

Next, define the two-point functions $S_{uv}(s,t)$, for $u, v \in \{0,1\}$, as follows:

$$\begin{aligned} S_{00}(s,t) &:= \mathbf{1}\{K^a_{00}\}\log\{1-\chi_{at}\} \\ &\quad + \mathbf{1}\{K^a_{00}, s \le t\}\log\left\{\frac{1-\chi_{nt}-\chi_{at}}{1-\chi_{at}}\theta^{(0)}_{co}(t) + \frac{\chi_{nt}}{1-\chi_{at}}\theta_{nt}(t)\right\}, \\ &\quad + \mathbf{1}\{K^a_{00}, s > t\}\log\left\{\frac{1-\chi_{nt}-\chi_{at}}{1-\chi_{at}}(1-\theta^{(0)}_{co}(t)) + \frac{\chi_{nt}}{1-\chi_{at}}(1-\theta_{nt}(t))\right\}, \\ S_{01}(s,t) &:= \mathbf{1}\{K^a_{01}\}\log\chi_{at} + \mathbf{1}\{K^a_{01}, s \le t\}\log\theta_{at}(t) + \mathbf{1}\{K^a_{01}, s > t\}\log(1-\theta_{at}(t)), \\ S_{10}(s,t) &:= \mathbf{1}\{K^a_{10}\}\log\chi_{nt} + \mathbf{1}\{K^a_{10}, s \le t\}\log\theta_{nt}(t) + \mathbf{1}\{K^a_{10}, s > t\}\log(1-\theta_{nt}(t)), \\ S_{11}(s,t) &:= \mathbf{1}\{K^a_{11}\}\log\{1-\chi_{nt}\} \\ &\quad + \mathbf{1}\{K^a_{11}, s \le t\}\log\left\{\frac{1-\chi_{nt}-\chi_{at}}{1-\chi_{nt}}\theta^{(1)}_{co}(t) + \frac{\chi_{at}}{1-\chi_{nt}}\theta_{at}(t)\right\} \\ &\quad + \mathbf{1}\{K^a_{11}, s > t\}\log\left\{\frac{1-\chi_{nt}-\chi_{at}}{1-\chi_{nt}}(1-\theta^{(1)}_{co}(t)) + \frac{\chi_{at}}{1-\chi_{nt}}(1-\theta_{at}(t))\right\}. \end{aligned} \tag{A.2}$$

With (A.1) and (A.2), the binomial log-likelihood (3.4) can be re-written as follows:

$$\frac{1}{n^2}\log L(\boldsymbol{\theta}, \boldsymbol{\chi}, \upsilon|\mathcal{D}) = \ell(\upsilon) + \frac{1}{n^2}\sum_{b\in I_\kappa}\sum_{a\in[n]}\sum_{u,v\in\{0,1\}} S_{uv}(Y_a, Y_{(b)}), \tag{A.3}$$

where $\ell(\upsilon) = \frac{1}{n^2}|I_\kappa|\{(n_{00}+n_{01})\log(1-\upsilon) + (n_{10}+n_{11})\log\upsilon\}$. Note that the second-term in the RHS of (A.3) above does not depend on υ. Hence,

$$\hat{\phi}_1 := \arg\max_{\upsilon\in[0,1]} \ell(\upsilon) = \frac{n_{10}+n_{11}}{n},$$

showing (3.7).

It remains to show (3.8). To this end, it suffices to show that

$$\mathbb{M}_n(\boldsymbol{\theta}, \boldsymbol{\chi}) = \frac{1}{n^2}\sum_{b\in I_\kappa}\sum_{a\in[n]}\sum_{u,v\in\{0,1\}} S_{uv}(Y_a, Y_{(b)}). \tag{A.4}$$

To this end, define $\theta_{00}(t) = \frac{1-\chi_{nt}-\chi_{at}}{1-\chi_{at}}\theta_{co}^{(0)}(t) + \frac{\chi_{nt}}{1-\chi_{at}}\theta_{nt}(t)$. Then

$$\begin{aligned}
&\frac{1}{n}\sum_{a=1}^{n} S_{00}(Y_a, Y_{(b)}) \\
&= \frac{n_{00}}{n}\Big[\log\{1-\chi_{at}\} + \overline{F}_{00}(Y_{(b)})\log\theta_{00}(Y_{(b)}) + (1-\overline{F}_{00}(Y_{(b)}))\log(1-\theta_{00}(Y_{(b)}))\Big] \\
&= \frac{n_{00}}{n}\Big[\log\{1-\chi_{at}\} + J(\overline{F}_{00}(Y_{(b)}), \theta_{00}(Y_{(b)}))\Big], \qquad \text{(A.5)}
\end{aligned}$$

where $J(x, y) := x\log y + (1-x)\log(1-y)$. Similarly, denoting $\theta_{11}(t) := \frac{1-\chi_{nt}-\chi_{at}}{1-\chi_{nt}}\theta_{co}^{(1)}(t) + \frac{\chi_{at}}{1-\chi_{nt}}\theta_{at}(t)$ gives,

$$\begin{aligned}
&\frac{1}{n}\sum_{a=1}^{n} S_{11}(Y_a, Y_{(b)}) \\
&= \frac{n_{11}}{n}\Big[\log\{1-\chi_{nt}\} + \overline{F}_{11}(Y_{(b)})\log\theta_{11}(Y_{(b)}) + (1-\overline{F}_{11}(Y_{(b)}))\log(1-\theta_{11}(Y_{(b)}))\Big] \\
&= \frac{n_{11}}{n}\Big[\log\{1-\chi_{nt}\} + J(\overline{F}_{11}(Y_{(b)}), \theta_{11}(Y_{(b)}))\Big]. \qquad \text{(A.6)}
\end{aligned}$$

Similarly,

$$\frac{1}{n}\sum_{a=1}^{n} S_{01}(Y_a, Y_{(b)}) = \frac{n_{01}}{n}\Big[\log\chi_{at} + J(\overline{F}_{01}(Y_{(b)}), \theta_{at}(Y_{(b)}))\Big], \qquad \text{(A.7)}$$

and

$$\frac{1}{n}\sum_{a=1}^{n} S_{10}(Y_a, Y_{(b)}) = \frac{n_{10}}{n}\Big[\log\chi_{nt} + J(\overline{F}_{10}(Y_{(b)}), \theta_{nt}(Y_{(b)}))\Big]. \qquad \text{(A.8)}$$

Averaging (A.5), (A.6), and (A.7), and (A.8) over $b \in I_\kappa$, gives (A.4), which implies (3.8).

A.2.2. *Unrestricted maximization of the sample objective function.* Here we show that maximizing the sample objective function $\mathbb{M}_n$ (recall (3.6)) over the unrestricted parameter space gives the plug-in estimates (2.11) and (2.14), justifying the choice of $\mathbb{M}_n$ as an approximate surrogate for the actual likelihood. As a consequence, it follows that the plug-in and the MBL estimates of the compliance classes are equal with probability 1.

LEMMA A.1. *Let $\hat{\boldsymbol{\phi}}$ be the MBL estimator and $\breve{\boldsymbol{\phi}}$ the plug-in estimator (2.11). Then*

$$\arg\max_{\boldsymbol{\theta}\in\boldsymbol{\vartheta}, \boldsymbol{\chi}\in\mathbb{R}^2} \mathbb{M}_n(\boldsymbol{\theta}, \boldsymbol{\chi}) = (\breve{\boldsymbol{G}}, \breve{\boldsymbol{\phi}}), \qquad \text{(A.9)}$$

where $\breve{\boldsymbol{G}} \in \breve{\mathcal{F}} := \{\breve{\boldsymbol{G}} \in \boldsymbol{\vartheta} : \breve{\boldsymbol{G}}(Y_{(b)}) = \breve{\boldsymbol{F}}(Y_{(b)})$ for $b \in I_\kappa\}$.

PROOF. Define $n_{00} + n_{01} = n_0$ and $n_{10} + n_{11} = n_1$. Then by Observation A.1,

$$\mathbb{M}(\boldsymbol{\theta}, \boldsymbol{\chi}) \leq \frac{1}{n} \sum_{b \in I_\kappa} \left\{ \frac{n_0}{n} I\left(\frac{n_{10}}{n_1}\right) + \frac{n_1}{n} I\left(\frac{n_{01}}{n_0}\right) + \sum_{u,v \in \{0,1\}} \frac{n_{uv}}{n} I(\overline{F}_{uv}(Y_{(b)})) \right\}.$$

Moreover, the equality is attained when

$$\begin{pmatrix} \theta_{00}(Y_{(b)}) \\ \theta_{10}(Y_{(b)}) \\ \theta_{01}(Y_{(b)}) \\ \theta_{11}(Y_{(b)}) \end{pmatrix} = \begin{pmatrix} \overline{F}_{00}(Y_{(b)}) \\ \overline{F}_{10}(Y_{(b)}) \\ \overline{F}_{01}(Y_{(b)}) \\ \overline{F}_{11}(Y_{(b)}) \end{pmatrix}$$

for $b \in I_\kappa$ and $(\chi_{nt}, \chi_{at}) = (n_{10}/n_1, n_{01}/n_0) = (\breve{\phi}_{nt}, \breve{\phi}_{at})$. This implies that the equality is attained when $(\chi_{nt}, \chi_{at}) = (\breve{\phi}_{nt}, \breve{\phi}_{at})$ and $\theta_{at}(Y_{(b)}) = \overline{F}_{01}(Y_{(b)}) = \breve{F}_{at}(Y_{(b)})$, $\theta_{nt}(Y_{(b)}) = \overline{F}_{10}(Y_{(b)}) = \breve{F}_{at}(Y_{(b)})$ and

$$\theta_{co}^{(0)}(Y_{(b)}) = \frac{\overline{F}_{00}(Y_{(b)}) - \frac{\breve{\phi}_{nt}}{\breve{\phi}_{co}+\breve{\phi}_{nt}} \overline{F}_{10}(Y_{(b)})}{\frac{\breve{\phi}_{co}}{\breve{\phi}_{co}+\breve{\phi}_{nt}}} = \breve{F}_{co}^{(0)}(Y_{(b)}) \tag{A.10}$$

$$\theta_{co}^{(1)}(Y_{(b)}) = \frac{\overline{F}_{11}(Y_{(b)}) - \frac{\breve{\phi}_{at}}{\breve{\phi}_{co}+\breve{\phi}_{at}} \overline{F}_{01}(Y_{(b)})}{\frac{\breve{\phi}_{co}}{\breve{\phi}_{co}+\breve{\phi}_{at}}} = \breve{F}_{co}^{(1)}(Y_{(b)}) \tag{A.11}$$

for $b \in I_\kappa$. This completes the proof of (A.9). □

APPENDIX B: PROOF OF THEOREM 3.2

In this section we prove Theorem 3.2, which shows that the plug-in estimator $\breve{\boldsymbol{F}}$ and the MBL estimator $\hat{\boldsymbol{F}}$ have the same limiting distribution.

B.1. Comparing the (sample) objective functions. From Lemma A.1, we know that $\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) \geq 0$. One of the main steps towards the proof of (3.9), is show that this difference is small, more precisely,

$$\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) = o_P(n^{-\frac{5}{4}}). \tag{B.1}$$

To this end, we have the following definition:

DEFINITION B.1. Define $\tilde{\boldsymbol{F}} = (\tilde{F}_{co}^{(0)}, \tilde{F}_{nt}, \tilde{F}_{co}^{(1)}, \tilde{F}_{at})' \in \mathcal{P}([0,1]^{\mathbb{R}})^4$, as follows:

$$\tilde{F}_{nt} := \breve{F}_{10} = \overline{F}_{10} \quad \textit{and} \quad \tilde{F}_{at} := \breve{F}_{01} = \overline{F}_{01},$$

are the corresponding empirical distribution functions, and

$$\tilde{F}_{co}^{(0)} := \arg \min_{\theta \in \mathcal{P}([0,1]^{\mathbb{R}})} \sum_{b \in I_\kappa} (\breve{F}_{co}^{(0)}(Y_{(b)}) - \theta(Y_{(b)}))^2,$$
$$\tilde{F}_{co}^{(1)} := \arg \min_{\theta \in \mathcal{P}([0,1]^{\mathbb{R}})} \sum_{b \in I_\kappa} (\breve{F}_{co}^{(1)}(Y_{(b)}) - \theta(Y_{(b)}))^2. \quad \text{(B.2)}$$

Note that this only defines $\tilde{F}_{co}^{(0)}$ and $\tilde{F}_{co}^{(1)}$ at the knots $\{Y_{(b)}\}_{b \in I_\kappa}$. To ensure (B.2) is well-defined we extend $\tilde{F}_{co}^{(0)}$ and $\tilde{F}_{co}^{(1)}$ between the knots by right-continuous interpolation, and extrapolate it beyond the knots to 0 and 1. Moreover, define

$$\tilde{F}_{00} := \breve{\lambda}_0 \tilde{F}_{co}^{(0)} + (1 - \breve{\lambda}_0)\tilde{F}_{nt},$$
$$\tilde{F}_{11} := \breve{\lambda}_1 \tilde{F}_{co}^{(1)} + (1 - \breve{\lambda}_1)\tilde{F}_{at}, \quad \text{(B.3)}$$

and $\tilde{F}_{10} := \tilde{F}_{nt}$ and $\tilde{F}_{01} := \tilde{F}_{at}$.

Note, since $\boldsymbol{\phi}$ is in the interior of $[0,1]_+^2$, which is an open subset of $\mathbb{R}^2$, there exists an $\varepsilon > 0$ such that $B(\boldsymbol{\phi}, \varepsilon) = \{\boldsymbol{\chi} \in \mathbb{R}^2 : ||\boldsymbol{\chi} - \boldsymbol{\phi}||_2 < \varepsilon\} \subset [0,1]_+^2$. Moreover, there exists $n \geq N(\varepsilon, \delta)$ such that $\mathbb{P}(\breve{\boldsymbol{\phi}} \notin B(\boldsymbol{\phi}, \varepsilon)) < \delta$. Therefore, $\mathbb{P}(\breve{\boldsymbol{\phi}}$ not in the interior of $[0,1]_+^2) \leq \mathbb{P}(\breve{\boldsymbol{\phi}} \notin B(\boldsymbol{\phi}, \varepsilon)) < \delta$. To this end, let

$$\mathscr{B}_1 = \left\{\breve{\boldsymbol{\phi}} \text{ is in the interior of } [0,1]_+^2\right\} \bigcap \left\{\breve{\boldsymbol{F}} \text{ is coordinate-wise in the interior of } \mathbb{R}^{[0,1]})\right\}.$$

From the discussion above it is clear that the $\mathbb{P}(\mathscr{B}_1^c) \to 0$. Therefore, for the remainder of this section all events will be on the set $\mathscr{B}_1$.

To begin with this gives,

$$\mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) \geq \mathbb{M}_n(\tilde{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}),$$

since $(\tilde{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) \in \boldsymbol{\vartheta}_+ \times [0,1]_+^2$ and $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}})$ maximizes $\mathbb{M}_n$ over $\boldsymbol{\vartheta}_+ \times [0,1]_+^2$, by definition. Therefore, to show (B.1) it suffices to prove that

$$\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\tilde{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) = o_P(n^{-\frac{5}{4}}). \quad \text{(B.4)}$$

Observation B.1. *Let $\breve{\boldsymbol{F}}$ be the plug-in estimator (2.14) and $\tilde{\boldsymbol{F}}$ as defined in Definition B.1. Then*

$$\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\tilde{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) = O_P(1) \left[\frac{1}{n} \sum_{u \in \{0,1\}} \sum_{b \in I_\kappa} (\breve{F}_{co}^{(u)}(Y_{(b)}) - \breve{F}_{co}^{(u)}(Y_{(b)}))^2\right],$$

whenever $|\tilde{F}_{uu}(Y_{(\lceil n\kappa \rceil)}) - \overline{F}_{uu}(Y_{(\lceil n\kappa \rceil)})| = o_P(1)$ *and* $|\tilde{F}_{uu}(Y_{(\lceil n(1-\kappa) \rceil)}) - \overline{F}_{uu}(Y_{(\lceil n(1-\kappa) \rceil)})| = o_P(1)$.

PROOF. Recall the definition of (3.6). Then

$$
\begin{aligned}
&\mathbb{M}_n(\breve{\boldsymbol{F}},\breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\tilde{\boldsymbol{F}},\breve{\boldsymbol{\phi}}) \\
&\qquad = \frac{1}{n}\sum_{u,v\in\{0,1\}} \frac{n_{uv}}{n}\big(J(\overline{F}_{uv},\breve{F}_{uv}) - J(\overline{F}_{uv},\tilde{F}_{uv})\big) \\
&\qquad = \frac{1}{n}\sum_{u\in\{0,1\}} \frac{n_{uu}}{n}\big(J(\overline{F}_{uu},\breve{F}_{uu}) - J(\overline{F}_{uu},\tilde{F}_{uu})\big) \qquad (\text{since } \tilde{F}_{10}=\breve{F}_{10} \text{ and } \tilde{F}_{01}=\breve{F}_{01}) \\
&\qquad = \frac{1}{n}\sum_{u\in\{0,1\}} \frac{n_{uu}}{n}\Big(J(\overline{F}_{uu},\overline{F}_{uu}) - J(\overline{F}_{uu},\tilde{F}_{uu})\Big) \\
&\qquad = \frac{1}{n}\sum_{u\in\{0,1\}} \frac{n_{uu}}{n}\sum_{b\in I_\kappa} T_u(Y_{(b)}), \qquad\qquad \text{(B.5)}
\end{aligned}
$$

where

$$T_u(Y_{(b)}) := \overline{F}_{uu}(Y_{(b)})\log\frac{\overline{F}_{uu}(Y_{(b)})}{\tilde{F}_{uu}(Y_{(b)})} + (1-\overline{F}_{uu}(Y_{(b)}))\log\frac{1-\overline{F}_{uu}(Y_{(b)})}{1-\tilde{F}_{uu}(Y_{(b)})}.$$

Now, a two-term Taylor expansion of the function $f_a(x) := -a\log\frac{x}{a} - (1-a)\log\frac{1-x}{1-a}$, at $x=a$, gives

$$T_u(Y_{(b)}) = \frac{(\tilde{F}_{uu}(Y_{(b)}) - \overline{F}_{uu}(Y_{(b)}))^2}{2}\left\{\frac{\overline{F}_{uu}(Y_{(b)})}{(\omega_{uu}(Y_{(b)}))^2} - \frac{1-\overline{F}_{uu}(Y_{(b)})}{(1-\omega_{uu}(Y_{(b)}))^2}\right\}, \qquad \text{(B.6)}$$

and $\omega_{uu}(Y_{(b)}) \in [\overline{F}_{uu}(Y_{(b)}) \wedge \tilde{F}_{uu}(Y_{(b)}), \tilde{F}_{uu}(Y_{(b)}) \vee \overline{F}_{uu}(Y_{(b)})]$.

Note that $\omega_{uu}(Y_{(b)}) \geq \overline{F}_{uu}(Y_{(\lceil n\kappa\rceil)}) \wedge \tilde{F}_{uu}(Y_{(\lceil n\kappa\rceil)})$ and $\overline{F}_{uu}(Y_{(b)}) \leq \overline{F}_{uu}(Y_{(\lceil n(1-\kappa)\rceil)})$. Therefore,

$$\frac{\overline{F}_{uu}(Y_{(b)})}{(\omega_{uu}(Y_{(b)}))^2} \leq \frac{\overline{F}_{uu}(Y_{(\lceil n(1-\kappa)\rceil)})}{\overline{F}_{uu}(Y_{(n\kappa)}) \wedge \tilde{F}_{uu}(Y_{(n\kappa)})} = O_P(1),$$

since $\overline{F}_{uu}(Y_{(\lceil n\kappa\rceil)}) \xrightarrow{P} H_{uu}^{-1}(\kappa)$, $\overline{F}_{uu}(Y_{(\lceil n(1-\kappa)\rceil)}) \xrightarrow{P} H_{uu}^{-1}(1-\kappa)$ using Observation A.3, and $|\tilde{F}_{uu}(Y_{(\lceil n\kappa\rceil)}) - \overline{F}_{uu}(Y_{(\lceil n\kappa\rceil)})| = o_P(1)$ by assumption. Similarly,

$$\frac{1-\overline{F}_{uu}(Y_{(b)})}{(1-\omega_{uu}(Y_{(b)}))^2} = O_P(1).$$

Therefore,

$$\sum_{b\in I_\kappa}|T_u(Y_{(b)})| \leq \sum_{b\in I_\kappa}\frac{(\tilde{F}_{uu}(Y_{(b)}) - \overline{F}_{uu}(Y_{(b)}))^2}{2}\left\{\frac{\overline{F}_{uu}(Y_{(b)})}{(\omega_{uu}(Y_{(b)}))^2} + \frac{1-\overline{F}_{uu}(Y_{(b)})}{(1-\omega_{uu}(Y_{(b)}))^2}\right\}$$

$$= O_P(1)\sum_{b\in I_\kappa}(\tilde{F}_{uu}(Y_{(b)}) - \overline{F}_{uu}(Y_{(b)}))^2$$
$$= O_P(1)\sum_{b\in I_\kappa}(\tilde{F}_{co}^{(u)}(Y_{(b)}) - \breve{F}_{co}^{(u)}(Y_{(b)}))^2, \tag{B.7}$$

where the last step uses $\tilde{F}_{00} = \breve{\lambda}_0\tilde{F}_{co}^{(0)} + (1-\breve{\lambda}_0)\tilde{F}_{nt}$, $\overline{F}_{00} = \breve{\lambda}_0\breve{F}_{co}^{(0)} + (1-\breve{\lambda}_0)\breve{F}_{nt}$, and $\tilde{F}_{nt} = \breve{F}_{nt}$, and similarly for $\tilde{F}_{11}$ and $\overline{F}_{11}$. □

Therefore, to show that $\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\tilde{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) = o_P(1/n)$, it suffices to show that

$$\sum_{b\in I_\kappa}(\tilde{F}_{co}^{(u)}(Y_{(b)}) - \breve{F}_{co}^{(u)}(Y_{(b)}))^2 = o_P(1),$$

for $u \in \{0,1\}$. This is the content of the following proposition, gives an error rate of $o_P(n^{-\frac{1}{4}})$.

Proposition B.1. *For $u \in \{0,1\}$,*

$$\sum_{b\in I_\kappa}\left(\tilde{F}_{co}^{(u)}(Y_{(b)}) - \breve{F}_{co}^{(u)}(Y_{(b)})\right)^2 = o_P(n^{-\frac{1}{4}}). \tag{B.8}$$

This implies, $\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) = o_P(n^{-\frac{5}{4}})$.

B.1.1. *Proof of Proposition B.1.* To begin with, note that (B.8) implies

$$|\tilde{F}_{uu}(Y_{(\lceil n\kappa\rceil)}) - \overline{F}_{uu}(Y_{(\lceil n\kappa\rceil)})| = o_P(1), \quad |\tilde{F}_{uu}(Y_{(\lceil n(1-\kappa)\rceil)}) - \overline{F}_{uu}(Y_{(\lceil n(1-\kappa)\rceil)})| = o_P(1).$$

Therefore, using Observation B.1 gives $\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\tilde{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) = o_P(n^{-\frac{5}{4}})$, which implies $\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) = o_P(n^{-\frac{5}{4}})$, from (B.4). Therefore, the rest of this section is devoted to the proof of (B.8).

We will prove the result for $u = 0$. The other case for $u = 1$ can be done similarly. To begin with note that on the event $\mathscr{B}_1$, $|\breve{F}_{co}^{(0)}(t)| \in [0,1]$ for all $t \in \mathbb{R}$, and, therefore, maximum in (B.2) can be taken over $\mathbb{I}([0,1]^{\mathbb{R}})$, the set of increasing functions from $\mathbb{R} \to [0,1]$. Then the well-know result of Robertson, Wright and Dykstra (1988) gives

$$\tilde{F}_{co}^{(0)}(Y_{(b)}) = \min_{\ell\geq b}\max_{k\leq b}\frac{\sum_{i=k}^{\ell}\breve{F}_{co}^{(0)}(Y_{(i)})}{\ell-k+1} \tag{B.9}$$

Now, define $f(k,\ell) := \frac{\sum_{i=k}^{\ell}\breve{F}_{co}^{(0)}(Y_{(i)})}{\ell-k+1} - \breve{F}_{co}^{(0)}(Y_{(b)})$. We will use the following lemma.

LEMMA B.1. *Let $b \in I_\kappa$. For any function $f : [n] \times [n] \to \mathbb{R}$, the following inequality holds*

$$\left(\min_{\ell \geq b} \max_{k \leq b} f(k,\ell)\right)^2 \leq \left(\max_{k \leq b} f_+(k,b)\right)^2 + \left(\max_{\ell \geq b} f_-(b,\ell)\right)^2, \tag{B.10}$$

where $f_+ := \max(f, 0)$ and $f_- := (-f)_+$.

PROOF. We have

$$\min_{\ell \geq b} \max_{k \leq b} f(k,\ell) \leq \max_{k \leq b} f(k,b) \leq \max_{k \leq b} f_+(k,b).$$

Similarly,

$$\min_{\ell \geq b} \max_{k \leq b} f(k,\ell) \geq \min_{\ell \geq b} f(b,\ell) = -\max_{\ell \geq b} -f(b,\ell) \geq -\max_{\ell \geq b} f_-(b,\ell)$$

Therefore, the lemma follows. □

To prove (B.8), we have to show $\mathbb{P}_{\mathscr{B}_1}\left(\sum_{b \in I_\kappa} (\tilde{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(b)}))^2 > \delta n^{-\frac{1}{4}}\right) = o(1)$, for any $\delta > 0$, where $\mathbb{P}_{\mathscr{B}_1}(\cdot) = \mathbb{P}(\cdot \cap \mathscr{B}_1)$. To begin with, by the triangle inequality and Lemma B.1, we get

$$\begin{aligned}
&\mathbb{P}_{\mathscr{B}_1}\left(\sum_{b \in I_\kappa} \left(\tilde{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(b)})\right)^2 > \delta n^{-\frac{1}{4}}\right) \\
&\leq \sum_{b \in I_\kappa} \mathbb{P}_{\mathscr{B}_1}\left(\left(\tilde{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(b)})\right)^2 > \delta n^{-\frac{5}{4}}\right) \\
&= \sum_{b \in I_\kappa} \mathbb{P}_{\mathscr{B}_1}\left(\left(\min_{\ell \geq b} \max_{k \leq b} f(k,\ell)\right)^2 > \delta n^{-\frac{5}{4}}\right) \\
&\leq \sum_{b \in I_\kappa} \mathbb{P}_{\mathscr{B}_1}\left(\left(\max_{k \leq b} f_+(k,b)\right)^2 > \frac{\delta n^{-\frac{5}{4}}}{2}\right) + \sum_{b \in I_\kappa} \mathbb{P}_{\mathscr{B}_1}\left(\left(\max_{\ell \geq b} f_-(b,\ell)\right)^2 > \frac{\delta n^{-\frac{5}{4}}}{2}\right) \\
&\leq \sum_{b \in I_\kappa} \mathbb{P}_{\mathscr{B}_1}\left(\max_{k \leq b} f_+(k,b) > \delta_0 n^{-\frac{5}{8}}\right) + \sum_{b \in I_\kappa} \mathbb{P}_{\mathscr{B}_1}\left(\max_{\ell \geq b} f_-(b,\ell) > \delta_0 n^{-\frac{5}{8}}\right),
\end{aligned} \tag{B.11}$$

where $\delta_0 = \sqrt{\delta/2}$.

LEMMA B.2. *Let $\varepsilon = \delta_0 / n^{5/8}$. Then for any $b \in I_\kappa$,*

$$\mathbb{P}_{\mathscr{B}_1}\left(\max_{\ell \in [b, \lceil n(1-\kappa) \rceil]} f_-(b,\ell) \geq \varepsilon\right) \leq \sum_{\ell=b}^{\lceil n(1-\kappa) \rceil} \mathbb{P}_{\mathscr{B}_1}\left(\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(\ell)}) \geq \varepsilon\right). \tag{B.12}$$

Similarly,

$$\mathbb{P}_{\mathscr{B}_1}\left(\max_{k\in[\lceil n\kappa\rceil,b]} f_+(k,b)\geq\varepsilon\right)\leq\sum_{k=\lceil n\kappa\rceil}^{b}\mathbb{P}_{\mathscr{B}_1}\left(\breve{F}_{co}^{(0)}(Y_{(k)})-\breve{F}_{co}^{(0)}(Y_{(b)})\geq\varepsilon\right). \tag{B.13}$$

Proof. Note that

$$f_-(b,\ell)=\max(0,\breve{F}_{co}^{(0)}(Y_{(b)})-\frac{\breve{F}_{co}^{(0)}(Y_{(b)})+\cdots+\breve{F}_{co}^{(0)}(Y_{(\ell)})}{\ell-b+1})=\max\left(0,\frac{\sum_{i=b}^{\ell}(\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(i)}))}{\ell-b+1}\right).$$

Then

$$\begin{aligned}
&\mathbb{P}_{\mathscr{B}_1}\left(\max_{\ell\in[b,\lceil n(1-\kappa)\rceil]} f_-(b,\ell)\leq\varepsilon\right)\\
&\mathbb{P}_{\mathscr{B}_1}\left(\max_{\ell\in[b,\lceil n(1-\kappa)\rceil]}\frac{\sum_{i=b}^{\ell}(\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(i)}))}{\ell-b+1}\leq\varepsilon\right)\\
&=\mathbb{P}_{\mathscr{B}_1}\left(\frac{1}{2}\sum_{i=b}^{b+1}(\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(i)}))\leq\varepsilon,\ldots,\frac{\sum_{i=b}^{n}\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(i)}))}{n-b+1}\leq\varepsilon\right)\\
&\geq\mathbb{P}_{\mathscr{B}_1}\left(\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(b+1)})\leq 2\varepsilon,\ldots,\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(\lceil n(1-\kappa)\rceil)}\leq\varepsilon\right)\\
&\geq\mathbb{P}_{\mathscr{B}_1}\left(\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(b+1)})\leq\varepsilon,\ldots,\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(\lceil n(1-\kappa)\rceil)})\leq\varepsilon\right)\\
&=\mathbb{P}_{\mathscr{B}_1}\left(\max_{\ell\in[b,\lceil n(1-\kappa)\rceil]}(\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(\ell)})\leq\varepsilon\right).
\end{aligned}$$

The above inequality and a union bound implies (B.12).

Similarly,

$$\begin{aligned}
&\mathbb{P}_{\mathscr{B}_1}\left(\max_{k\in[\lceil n\kappa\rceil,b]} f_+(k,b)\leq\varepsilon\right)\\
&\mathbb{P}_{\mathscr{B}_1}\left(\max_{k\in[\lceil n\kappa\rceil,b]}\frac{\sum_{i=k}^{b}(\breve{F}_{co}^{(0)}(Y_{(k)})-\breve{F}_{co}^{(0)}(Y_{(b)}))}{b-k+1}\leq\varepsilon\right)\\
&=\mathbb{P}_{\mathscr{B}_1}\left(\frac{1}{2}\sum_{i=b-1}^{b}(\breve{F}_{co}^{(0)}(Y_{(i)})-\breve{F}_{co}^{(0)}(Y_{(b)}))\leq\varepsilon,\ldots,\frac{\sum_{i=\lceil n\kappa\rceil}^{b}\breve{F}_{co}^{(0)}(Y_{(b)})-\breve{F}_{co}^{(0)}(Y_{(i)}))}{b}\leq\varepsilon\right)\\
&\geq\mathbb{P}_{\mathscr{B}_1}\left(\breve{F}_{co}^{(0)}(Y_{(b-1)})-\breve{F}_{co}^{(0)}(Y_{(b)})\leq 2\varepsilon,\ldots,\breve{F}_{co}^{(0)}(Y_{(\lceil n\kappa\rceil)})-\breve{F}_{co}^{(0)}(Y_{(b)}\leq\varepsilon\right)\\
&=\mathbb{P}_{\mathscr{B}_1}\left(\max_{k\in[\lceil n\kappa\rceil,b]}(\breve{F}_{co}^{(0)}(Y_{(k)})-\breve{F}_{co}^{(0)}(Y_{(b)})\leq\varepsilon\right).
\end{aligned}$$

This implies (B.13) by a union bound. □

In light of (B.11) and above lemma, to prove Proposition B.1, we need bounds on the upper tail of the difference $\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(\ell)})$, for $\ell \geq b$ and lower tail of $\breve{F}_{co}^{(0)}(Y_{(k)}) - \breve{F}_{co}^{(0)}(Y_{(b)})$, for $k \leq b$. To this end, let $\breve{\eta}_{uv} = n_{uv}/n$, for $\{u, v\} \in \{0, 1\}$, and $\breve{\eta}_{-} = \min_{u,v} \breve{\eta}_{uv}$.

LEMMA B.3. *Let $\varepsilon = \delta_0/n^{5/8}$. Then for any $b \in I_\kappa$, such that*

$$\sum_{\ell=b}^{\lceil n(1-\kappa)\rceil} \mathbb{P}_{\mathscr{B}_1}\left(\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(\ell)}) \geq \varepsilon | (\boldsymbol{Z}, \boldsymbol{D})\right) \leq O(1/n^3) + 2n \exp\left\{-\frac{\breve{\eta}_{-}^{3/2}\breve{\lambda}_0^3\delta_0^2\sqrt{n}}{48\log n}\right\}, \tag{B.14}$$

where the constant in the $O(1/n^3)$-term is non-random. Similarly,

$$\sum_{k=\lceil n\kappa\rceil}^{b} \mathbb{P}_{\mathscr{B}_1}\left(\breve{F}_{co}^{(1)}(Y_{(k)}) - \breve{F}_{co}^{(1)}(Y_{(b)}) \geq \varepsilon | (\boldsymbol{Z}, \boldsymbol{D})\right) \leq O(1/n^3) + 2n \exp\left\{-\frac{\breve{\eta}_{-}^{3/2}\breve{\lambda}_0^3\delta_0^2\sqrt{n}}{48\log n}\right\}. \tag{B.15}$$

The proof of the above lemma is given below. Using this, the proof of Proposition B.1 can be easily completed as follows: Note that

$$n^2 \exp\left\{-\frac{\breve{\eta}_{-}^{3/2}\breve{\lambda}_0^3\delta_0^2\sqrt{n}}{48\log n}\right\} \xrightarrow{D} 0, \quad \mathbb{E}\left(n^2 \exp\left\{-\frac{\breve{\eta}_{-}^{3/2}\breve{\lambda}_0^3\delta_0^2\sqrt{n}}{48\log n}\right\}\right) \to 0,$$

by the dominated convergence theorem. Then first taking expectation over $\boldsymbol{Z}, \boldsymbol{D}$ gives

$$\sum_{\ell=b}^{\lceil n(1-\kappa)\rceil} \mathbb{P}_{\mathscr{B}_1}\left(\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(\ell)}) \geq \varepsilon\right) \leq O(1/n^3) + 2n\mathbb{E} \exp\left\{-\frac{\breve{\eta}_{-}^{3/2}\breve{\lambda}_0^3\delta_0^2\sqrt{n}}{48\log n}\right\}. \tag{B.16}$$

Therefore, from (B.13),

$$\mathbb{P}_{\mathscr{B}_1}\left(\max_{k\in[\lceil n\kappa\rceil, b]} f_{+}(k, b) > \delta_0 n^{-\frac{5}{8}}\right) \leq O(1/n^2) + 2n^2\mathbb{E} \exp\left\{-\frac{\breve{\eta}_{-}^{3/2}\breve{\lambda}_0^3\delta_0^2\sqrt{n}}{48\log n}\right\} \to 0. \tag{B.17}$$

Similarly, from (B.12) and (B.15), it can be shown that

$$\mathbb{P}_{\mathscr{B}_1}\left(\max_{k\in[b, \lceil n(1-\kappa)\rceil]} f_{-}(k, b) > \delta_0 n^{-\frac{5}{8}}\right) \to 0. \tag{B.18}$$

Adding (B.17) and (B.18) and using (B.11), completes the proof of (B.8), since $\mathbb{P}(\mathscr{B}_1^c) \to 0$.

Proof of Lemma B.3 Throughout the proof, all events will be conditional of $(\boldsymbol{Z}, \boldsymbol{D})$, and, for notational brevity, we will omit the conditioning event in all the expressions. Let

$$L = H^{-1}(\kappa) - 1 \quad \text{and} \quad R = H^{-1}(1-\kappa) + 1,$$

and let $\mathscr{B}_2$ be the event that $\{Y_{(\lceil n\kappa \rceil)}, Y_{(\lceil n(1-\kappa)\rceil)}, \in [L, R]\}$. Using $\mathbb{E}Y_{(\lceil n\kappa \rceil)} \to H^{-1}(\kappa)$, $\mathbb{E}(Y_{(\lceil n\kappa \rceil)} - \mathbb{E}Y_{(\lceil n\kappa \rceil)})^6 = O(1/n^3)$ (see Sen (1959)), and the Chebyshev's inequality it follows that

$$\mathbb{P}(Y_{(\lceil n\kappa \rceil)} \notin [L, R]) \leq \frac{\mathbb{E}(Y_{(\lceil n\kappa \rceil)} - \mathbb{E}Y_{(\lceil n\kappa \rceil)})^6}{(L - \mathbb{E}Y_{(\lceil n\kappa \rceil)})^6} = O(1/n^3). \tag{B.19}$$

Similarly, $\mathbb{P}(Y_{(\lceil n(1-\kappa)\rceil)} \notin [L, R]) = O(1/n^3)$, which gives $\mathbb{P}(\mathscr{B}_2^c) = O(1/n^3)$.

Now, partition $[L, R]$ in a grid $L = t_0 < t_1 < \ldots < t_M = R$, of size $\frac{4 \log n}{C_1 n}$, that is, $t_a = L + a\left(\frac{4 \log n}{C_1 n}\right)$, for $0 \leq a \leq M := \frac{C_1(R-L)n}{4 \log n}$. Define $N_a = \sum_{s=1}^n \mathbf{1}\{Y_s \in (t_a, t_{a+1}]\}$, the number of observations in the interval $(t_a, t_{a+1}]$. Let

$$\mathscr{B}_3 := \left\{\min_{0 \leq a \leq M} N_a \geq 1\right\} \cap \left\{\max_{0 \leq a \leq M} N_a \leq \frac{12 C_2}{C_1} \log n\right\}.$$

LEMMA B.4. $\mathbb{P}(\mathscr{B}_3^c) = O(1/n^3)$.

PROOF. Note that $\mathbb{P}(Y_j \notin (t_a, t_{a+1}]) = 1 - (H(t_{a+1}) - H(t_a)) \leq 1 - \frac{4 \log n}{n}$, by Assumption 2. Then

$$\mathbb{P}\left(\min_{0 \leq a \leq M} N_a = 0\right) = \sum_{a=0}^{M} \mathbb{P}(N_a = 0) \lesssim_\kappa n \left(1 - \frac{4 \log n}{n}\right)^n = O(1/n^3). \tag{B.20}$$

Next, note that $\mathbb{E}(N_a) = n(F(t_{a+1}) - F(t_a)) \in [4 \log n, \frac{4C_2}{C_1} \log n]$, by Assumption 2. Therefore, by the union bound followed by a Chernoff bound,[4] gives

$$\begin{aligned} \mathbb{P}\left(\max_{0 \leq a \leq M} N_a > \frac{12 C_2}{C_1} \log n\right) &= \sum_{a=0}^{M} \mathbb{P}\left(N_a - \mathbb{E}(N_a) > \frac{8 C_2}{C_1} \log n\right) \\ &\lesssim n e^{-\frac{4 C_2 \log n}{C_1}} = n e^{-4 \log n} = O(1/n^3). \end{aligned} \tag{B.21}$$

Combining (B.20) and (B.21) the proof of the lemma follows. □

[4]Suppose $X_1, X_2, \ldots, X_n$ are independent random variables taking values in $\{0, 1\}$. Let $X = \sum_{i=1}^n X_i$ denote their sum and let $\mu = \mathbb{E}(X)$. Then $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2 \mu}{3}}$, for $0 < \delta < 1$ and $\mathbb{P}(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta \mu}{2}}$, for $\delta \geq 1$.

Finally, let $\mathscr{B}_4 = \{|\breve{\eta}_{uv}/\eta_{uv} - 1| \leq 1 : \text{ for all } u, v \in \{0,1\}\}$, and set $\mathscr{B}_0 = \mathscr{B}_2 \cap \mathscr{B}_3 \cap \mathscr{B}_4$. From (B.19) and Lemma (B.4),

$$\mathbb{P}_{\mathscr{B}_1}(\mathscr{B}_0^c) \leq \mathbb{P}(\mathscr{B}_0^c) \leq \mathbb{P}(\mathscr{B}_2^c) + \mathbb{P}(\mathscr{B}_3^c) + \mathbb{P}(\mathscr{B}_3^4) = O(1/n^3). \tag{B.22}$$

Therefore, it suffices to consider events on $\mathscr{B} = \mathscr{B}_0 \cap \mathscr{B}_1$. Now, fix $b \in I_\kappa$. For any $\ell \geq b$ denote by $I_{j(\ell)} = (t_{j(\ell)},\ t_{j(\ell)+1}]$ the interval which contains $Y_{(\ell)}$, and $\overline{F}_{uv}((a,b]) := \overline{F}_{uv}(a) - F_{uv}(b)$, for $u, v \in \{0,1\}$. Then, by triangle inequality, on the set $\mathscr{B}$,

$$\begin{aligned}
|\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}]) - \overline{F}_{00}((Y_{(b)}, Y_{(\ell)}])| &\leq |\overline{F}_{00}(t_{j(b)}) - \overline{F}_{00}(Y_{(b)})| + |\overline{F}_{00}(t_{j(\ell)+1}) - \overline{F}_{00}(Y_{(\ell)})| \\
&= O\left(\frac{\log n}{n_{00}}\right) = O\left(\frac{\log n}{n}\right).
\end{aligned} \tag{B.23}$$

Now, take $\varepsilon = \delta_0/\sqrt{n}$. Then recalling the definition of $\breve{F}_{co}^{(0)}(Y_{(b)}) = \frac{\overline{F}_{00}(Y_{(b)}) - (1-\breve{\lambda})\overline{F}_{10}(Y_{(b)})}{\breve{\lambda}}$, and using triangle inequality gives,

$$\begin{aligned}
&\mathbb{P}_{\mathscr{B}}\left(\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(\ell)}) \geq \varepsilon\right) \\
&= \mathbb{P}_{\mathscr{B}}\left(\frac{\overline{F}_{00}((Y_{(b)}, Y_{(\ell)}])}{\breve{\lambda}_0} - \frac{(1-\breve{\lambda}_0)(\overline{F}_{01}((Y_{(b)}, Y_{(\ell)}]))}{\breve{\lambda}_0} \geq \varepsilon\right) \\
&\leq \mathbb{P}_{\mathscr{B}}\left(\frac{\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}])}{\breve{\lambda}_0} - \frac{1-\breve{\lambda}_0(\overline{F}_{01}((t_{j(b)}, t_{j(\ell)+1}]))}{\breve{\lambda}_0} \geq \frac{\varepsilon}{2}\right) \qquad \text{(by (B.23))} \\
&\leq T_1 + T_2,
\end{aligned}$$

where

$$T_1 = \mathbb{P}\left(|\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}]) - F_{00}((t_{j(b)}, t_{j(\ell)+1}])| \geq \frac{\breve{\lambda}_0(\frac{\varepsilon}{2} - F_{co}((t_{j(b)}, t_{j(\ell)+1}])}{2}\right)$$

and

$$T_2 = \mathbb{P}\left(|\overline{F}_{01}((t_{j(b)}, t_{j(\ell)+1}]) - F_{01}((t_{j(b)}, t_{j(\ell)+1}])| \geq \frac{\breve{\lambda}_0(\frac{\varepsilon}{2} - F_{co}((t_{j(b)}, t_{j(\ell)+1}])}{2(1-\breve{\lambda}_0)}\right).$$

Now, we will bound T_1. To begin with note that

$$-n_{00}\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}]) \sim \text{Bin}(n_{00}, -F_{00}((t_{j(b)}, t_{j(\ell)+1}])).$$

Moreover, by Assumption 2, $-F_{co}((t_{j(b)}, t_{j(\ell)+1}]) \geq C_1(t_{j(\ell)+1} - t_{j(b)}) \geq K(\ell - b)/n$, for some constant $K > 0$. Then for $|\ell - b| > \frac{4}{K\breve{\lambda}_0\sqrt{\breve{\eta}_{00}}}\sqrt{n}\log n$, where $\breve{\eta}_{00} = n_{00}/n$, we have

$$\mathbb{P}\left(|\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}]) - F_{00}((t_{j(b)}, t_{j(\ell)+1}])| \geq \frac{\breve{\lambda}_0(\frac{\varepsilon}{2} - F_{co}((t_{j(b)}, t_{j(\ell)+1}])}{2}\right)$$

$$
\begin{aligned}
&\leq \mathbb{P}\left(|\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}]) - F_{00}((t_{j(b)}, t_{j(\ell)+1}])| \geq -\frac{\breve{\lambda}_0 F_{co}((t_{j(b)}, t_{j(\ell)+1}])}{2}\right) \\
&\leq \mathbb{P}\left(|n_{00}\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}]) - n_{00}F_{00}((t_{j(b)}, t_{j(\ell)+1}])| \geq \frac{K\breve{\lambda}_0 n_{00}(\ell - b)}{2n}\right) \\
&\leq 2e^{-\frac{\breve{\eta}_{00}K^2\breve{\lambda}_0^2(\ell-b)^2}{2n}} = O(1/n^8),
\end{aligned}
$$

where the last step follows by the Hoeffding's inequality.

Now, suppose $|\ell - b| \leq \frac{4}{K\breve{\lambda}_0\sqrt{\breve{\eta}_{00}}}\sqrt{n}\log n$. Let $t = -\frac{\breve{\lambda}_0\varepsilon}{2F_{00}((t_{j(b)}, t_{j(\ell)+1}])}$. Then

$$
\begin{aligned}
&\mathbb{P}\left(|\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}]) - F_{00}((t_{j(b)}, t_{j(\ell)+1}])| \geq \frac{\breve{\lambda}(\varepsilon - F_{co}((t_{j(b)}, t_{j(\ell)+1}])}{2}\right) \\
&\leq \mathbb{P}\left(|\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}]) - F_{00}((t_{j(b)}, t_{j(\ell)+1}])| \geq \frac{\breve{\lambda}_0\varepsilon}{2}\right) \\
&\leq \mathbb{P}\left(|n_{00}\overline{F}_{00}((t_{j(b)}, t_{j(\ell)+1}]) - n_{00}F_{00}((t_{j(b)}, t_{j(\ell)+1}])| \geq -tn_{00}F_{00}((t_{j(b)}, t_{j(\ell)+1}])\right) \\
&\leq 2\exp\left\{\frac{t^2\breve{\eta}_{00}nF_{00}((t_{j(b)}, t_{j(\ell)+1}])}{3}\right\} \qquad \text{(by Chernoff bound)} \\
&\leq 2\exp\left\{\frac{\breve{\eta}_{00}\breve{\lambda}_0^2\delta_0^2}{12F_{00}((t_{j(b)}, t_{j(\ell)+1}])}\right\} \\
&\leq 2\exp\left\{-\frac{\breve{\eta}_{00}\breve{\lambda}_0^2\delta_0^2 n}{12K(\ell - b)}\right\} \qquad \text{(since } -F_{co}((t_{j(b)}, t_{j(\ell)+1}]) \geq K(\ell - b)/n) \\
&\leq 2\exp\left\{-\frac{\breve{\eta}_-^{3/2}\breve{\lambda}_0^3\delta_0^2\sqrt{n}}{48\log n}\right\} \qquad \text{(recall } \breve{\eta}_- = \min_{u,v}\breve{\eta}_{uv}).
\end{aligned}
$$

This implies $T_1 \leq O(1/n^8) + 2\exp\left\{-\frac{\breve{\eta}_-^{3/2}\breve{\lambda}_0^3\delta_0^2\sqrt{n}}{48\log n}\right\}$, and similarly, for T_2. These combined with $\mathbb{P}_{\mathscr{B}_1}(\mathscr{B}_0^c) = O(1/n^3)$ completes the proof of Lemma B.3. $\square$.

B.2. Completing the proof of Theorem 3.2. In this section we complete the proof of Theorem 3.2. To begin, we estimate the difference $\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$, To this end, recalling (3.6), we have

$$
\begin{aligned}
&\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) \\
&= \frac{1}{n}\left[\sum_{u,v\in\{0,1\}}\sum_{b\in I_\kappa}\frac{n_{uv}}{n}\left\{\overline{F}_{uv}(Y_{(b)})\log\frac{\overline{F}_{uv}(Y_{(b)})}{\hat{F}_{uv}(Y_{(b)})} + (1 - \overline{F}_{uv}(Y_{(b)})\log\frac{1 - \hat{F}_{uv}(Y_{(b)})}{1 - \hat{F}_{uv}(Y_{(b)})}\right\}\right.
\end{aligned}
$$

$$+ |I_\kappa| \left\{ \frac{n_{00}}{n} \log \frac{1-\breve{\phi}_{at}}{1-\hat{\phi}_{at}} + \frac{n_{01}}{n} \log \frac{\breve{\phi}_{at}}{\hat{\phi}_{at}} + \frac{n_{10}}{n} \log \frac{\breve{\phi}_{nt}}{\hat{\phi}_{nt}} + \frac{n_{11}}{n} \log \frac{1-\breve{\phi}_{nt}}{1-\hat{\phi}_{nt}} \right\} \Bigg],$$

$$= \frac{1}{n} \sum_{u,v\in\{0,1\}} \frac{n_{uv}}{n} \sum_{b\in I_\kappa} T_{uv}(Y_{(b)}) + \frac{|I_\kappa|}{n} \left[\frac{n_{00}+n_{01}}{n} R_{nt} + \frac{n_{10}+n_{11}}{n} R_{at} \right], \tag{B.24}$$

where

$$T_{uv}(Y_{(b)}) := \overline{F}_{uv}(Y_{(b)}) \log \frac{\overline{F}_{uv}(Y_{(b)})}{\hat{F}_{uv}(Y_{(b)})} + (1-\overline{F}_{uv}(Y_{(b)})) \log \frac{1-\overline{F}_{uv}(Y_{(b)})}{1-\hat{F}_{uv}(Y_{(b)})},$$

$$R_{at} = (1-\breve{\phi}_{at}) \log \frac{1-\breve{\phi}_{at}}{1-\hat{\phi}_{at}} + \breve{\phi}_{at} \log \frac{\breve{\phi}_{at}}{\hat{\phi}_{at}},$$

and

$$R_{at} = (1-\breve{\phi}_{nt}) \log \frac{1-\breve{\phi}_{nt}}{1-\hat{\phi}_{nt}} + \breve{\phi}_{nt} \log \frac{\breve{\phi}_{nt}}{\hat{\phi}_{nt}}.$$

Now, using $a \log \frac{a}{x} + (1-a) \log \frac{1-a}{1-x} \geq \frac{1}{2}(x-a)^2$ (Observation A.2) gives $T_{uv}(Y_{(b)}) \gtrsim (\hat{F}_{uv}(Y_{(b)}) - \overline{F}_{uv}(Y_{(b)}))^2$, $R_{at} \gtrsim (\hat{\phi}_{at} - \breve{\phi}_{at})^2$ and $R_{nt} \gtrsim (\hat{\phi}_{nt} - \breve{\phi}_{nt})^2$. Therefore,

$$\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) \gtrsim \frac{1}{n} \sum_{u,v\in\{0,1\}} \frac{n_{uv}}{n} \sum_{b\in I_\kappa} (\hat{F}_{uv}(Y_{(b)}) - \overline{F}_{uv}(Y_{(b)}))^2 + K_n ||\hat{\boldsymbol{\phi}} - \breve{\boldsymbol{\phi}}||_2^2,$$

for some constant $K_n \xrightarrow{P} K > 0$. Therefore, using $\mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) \leq \mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\tilde{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$, since $(\tilde{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) \in \boldsymbol{\vartheta}_+ \times [0,1]_+^2$ on the set $\mathscr{B}_1$ (where $\tilde{\boldsymbol{F}}$ is the PAVA estimate in Definition B.1), gives

$$\begin{aligned} \frac{1}{n} \sum_{u,v\in\{0,1\}} \frac{n_{uv}}{n} \sum_{b\in I_\kappa} (\hat{F}_{uv}(Y_{(b)}) - \overline{F}_{uv}(Y_{(b)}))^2 + K_n ||\hat{\boldsymbol{\phi}} - \breve{\boldsymbol{\phi}}||_2^2 &\lesssim \mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}}) \\ &\lesssim \mathbb{M}_n(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\tilde{\boldsymbol{F}}, \breve{\boldsymbol{\phi}}) \\ &= o_P(n^{-\frac{5}{4}}), \end{aligned} \tag{B.25}$$

by Proposition B.1.

Therefore, (B.25) implies $||\sqrt{n}(\hat{\boldsymbol{\phi}} - \breve{\boldsymbol{\phi}})||_2^2 = o_P(n^{-\frac{1}{4}})$. Moreover,

$$\sum_{b\in I_\kappa} (\hat{F}_{nt}(Y_{(b)}) - \breve{F}_{nt}(Y_{(b)}))^2 = o_P(n^{-\frac{1}{4}}), \quad \sum_{b\in I_\kappa} (\hat{F}_{at}(Y_{(b)}) - \breve{F}_{at}(Y_{(b)}))^2 = o_P(n^{-\frac{1}{4}}), \tag{B.26}$$

since $\hat{F}_{10} = \hat{F}_{nt}$, $\overline{F}_{10} = \breve{F}_{nt}$, $\hat{F}_{01} = \hat{F}_{at}$, $\overline{F}_{01} = \breve{F}_{at}$, and $n_{uv}/n \xrightarrow{P} \eta_{uv}$. Next, observe that on $\mathscr{B}_1$, $|\breve{F}_{00}(t)| \leq 1$, and

$$(\hat{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(0)}(t))^2$$

$$
\begin{aligned}
&= \left(\frac{\hat{F}_{00}(t) - (1-\hat{\lambda}_0)\hat{F}_{nt}(t)}{\hat{\lambda}_0} - \frac{\breve{F}_{00}(t) - (1-\breve{\lambda}_0)\breve{F}_{nt}(t)}{\breve{\lambda}_0} \right)^2 \\
&\lesssim \left(\frac{\hat{F}_{00}(t)}{\hat{\lambda}_0} - \frac{\breve{F}_{00}(t)}{\breve{\lambda}_0} \right)^2 + \left(\frac{(1-\hat{\lambda}_0)\hat{F}_{nt}(t)}{\hat{\lambda}_0} - \frac{(1-\breve{\lambda}_0)\breve{F}_{nt}(t)}{\breve{\lambda}_0} \right)^2 \\
&= O_P(1)(\hat{F}_{00}(t) - \breve{F}_{00}(t))^2 + O_P(1)(\hat{F}_{nt}(t) - \breve{F}_{nt}(t))^2 + \left(\frac{1}{\hat{\lambda}_0} - \frac{1}{\breve{\lambda}_0} \right)^2. \qquad \text{(B.27)}
\end{aligned}
$$

Adding (B.27) over $Y_{(b)}$, $b \in I_\kappa$, and using $\left(\frac{1}{\hat{\lambda}_0} - \frac{1}{\breve{\lambda}_0}\right)^2 = o_P(n^{-\frac{5}{4}})$, since $||\sqrt{n}(\hat{\boldsymbol{\phi}} - \breve{\boldsymbol{\phi}})||_2^2 = o_P(n^{-\frac{1}{4}})$, (B.25), and (B.26), gives $\sum_{b \in I_\kappa} (\hat{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(b)}))^2 = o_P(n^{-\frac{1}{4}})$. Similarly, we can show that $\sum_{b \in I_\kappa} (\hat{F}_{co}^{(1)}(Y_{(b)}) - \breve{F}_{co}^{(1)}(Y_{(b)}))^2 = o_P(n^{-\frac{1}{4}})$. This, together with (B.26) gives the following lemma:

Lemma B.5. *The BL estimators $(\hat{\boldsymbol{F}}, \hat{\boldsymbol{\phi}})$ and the plug-in estimators $(\breve{\boldsymbol{F}}, \breve{\boldsymbol{\phi}})$ satisfy $||\sqrt{n}(\hat{\boldsymbol{\phi}} - \breve{\boldsymbol{\phi}})||_2^2 = o_P(n^{-\frac{1}{4}})$ and*

$$
\sum_{b \in I_\kappa} ||\hat{\boldsymbol{F}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b)})||_2^2 = o_P(n^{-\frac{1}{4}}).
$$

This shows the MBL estimators and the plug-in estimators are close in average squared error with respect to the empirical distribution $\overline{H} = \sum_{u,v \in \{0,1\}} \frac{n_{uv}}{n} \overline{F}_{uv}$. To complete the proof of Theorem (3.2), we need to show that the average with respect to the empirical distribution can be replaced by the average (integral) with respect to the population distribution function $H = \sum_{u,v \in \{0,1\}} \eta_{uv} \overline{F}_{uv}$.

Lemma B.6. *The BL estimators $\hat{\boldsymbol{F}}$ and the plug-in estimators $\breve{\boldsymbol{F}}$ satisfy*

$$
\int_{J_\kappa} ||\sqrt{n}\{\hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t)\}||_2^2 \mathrm{d}H = o_P(1),
$$

where the $o_P(1)$ terms goes to zero as $n \to \infty$.

Proof. For $b \in I_\kappa$ and $Y_{(b)} \leq t < Y_{(b+1)}$, $\breve{\boldsymbol{F}}(t) = \breve{\boldsymbol{F}}(Y_{(b)})$. Moreover, $\hat{\boldsymbol{F}}(t) \leq \hat{\boldsymbol{F}}(Y_{(b+1)})$, where the inequality holds coordinate-wise. Then, for $Y_{(b)} \leq t < Y_{(b+1)}$,

$$
\begin{aligned}
||\hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t)||_2 &\leq ||\hat{\boldsymbol{F}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b)})||_2 + ||\breve{\boldsymbol{F}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b+1)})||_2 \\
&\leq ||\hat{\boldsymbol{F}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b)})||_2 + O(1/n).
\end{aligned}
$$

Therefore,

$$
\sum_{b \in I_\kappa} \int_{Y_{(b)}}^{Y_{(b+1)}} ||\sqrt{n}\{\hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t)\}||_2^2 \mathrm{d}H
$$

$$\lesssim_\kappa \Delta n \sum_{b \in I_\kappa} ||\hat{\boldsymbol{F}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b)})||_2^2 + \Delta, \tag{B.28}$$

where $\Delta := \sup_{b\in[n]}(H(Y_{(b+1)}) - H(Y_{(b)})) \overset{D}{=} \sup_{b\in[n]}(U_{(b+1)} - U_{(b)}) = O_P(\log n/n)$, by (Holst, 1980, Theorem 3.1). Then, by (B.28)

$$\begin{aligned}
&\int_{Y_{(\lceil n\kappa\rceil)}}^{Y_{(\lceil n(1-\kappa)\rceil)}} ||\sqrt{n}\{\hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t)\}||_2^2 \mathrm{d}H \\
&\qquad = \sum_{b\in I_\kappa} \int_{Y_{(b)}}^{Y_{(b+1)}} ||\sqrt{n}\{\hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t)\}||_2^2 \mathrm{d}H \\
&\qquad \leq O_P(\log n) \sum_{b\in I_\kappa} ||\{\hat{\boldsymbol{F}}(Y_{(b)}) - \breve{\boldsymbol{F}}(Y_{(b)})\}||_2^2 + o_P(1) \\
&\qquad = o_P(1),
\end{aligned}$$

where the last step follows from Lemma B.5.

To complete the proof we need to take care of the boundary effects. As before, by triangle inequality

$$\begin{aligned}
&\int_{H^{-1}(\kappa)}^{Y_{\lceil n\kappa\rceil}} ||\sqrt{n}\{\hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t)\}||_2^2 \mathrm{d}H \\
&\qquad \lesssim ||\sqrt{n}\{\hat{\boldsymbol{F}}(Y_{\lceil n\kappa\rceil}) - \breve{\boldsymbol{F}}(Y_{\lceil n\kappa\rceil})\}||_2^2 (Y_{\lceil n\kappa\rceil} - \kappa) + o_P(1) \\
&\qquad = o_P(1),
\end{aligned} \tag{B.29}$$

where the last step uses $\sqrt{n}\{\hat{\boldsymbol{F}}(Y_{\lceil n\kappa\rceil}) - \breve{\boldsymbol{F}}(Y_{\lceil n\kappa\rceil})\} = o_P(1)$ (which follows by a simple modification of the proof of Theorem 3.2) and $\sqrt{n}(Y_{\lceil n\kappa\rceil} - \kappa) = O_P(1)$. Similarly, it can be shown that

$$\int_{Y_{\lceil n(1-\kappa)\rceil}}^{H^{-1}(1-\kappa)} ||\sqrt{n}\{\hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t)\}||_2^2 \mathrm{d}H = o_P(1). \tag{B.30}$$

The proof now follows by combining (B.29) and (B.30) with (B.28). □

B.3. Limiting distribution of the MBL estimators. As the limiting distribution of the empirical distributions $\overline{F}_{uv}$ are well-known, Theorem 3.2 can be used to derive the limiting distribution of the MBL estimators $\hat{\boldsymbol{F}}$.

COROLLARY B.2. *Fix $0 < \kappa < 1/2$. Then for any continuous function $h : \mathbb{R} \to \mathbb{R}$,*

$$\int_{J_\kappa} h(t) \cdot \sqrt{n}\{\hat{\boldsymbol{F}}(t) - \mathbb{E}(\breve{\boldsymbol{F}}(t)|\boldsymbol{Z}, \boldsymbol{D})\} \mathrm{d}H \overset{D}{\to} \int_{J_\kappa} h(t) \cdot \boldsymbol{G}(t) \mathrm{d}H, \tag{B.31}$$

where

$$\boldsymbol{G}(t) := \begin{pmatrix} \frac{1}{\phi_c}\left\{\sqrt{\frac{\phi_c+\phi_n}{\phi_0}}B_{00}(F_{00}(t)) - \sqrt{\frac{\phi_n}{\phi_1}}B_{10}(F_{10}(t))\right\} \\ \frac{B_{01}(F_{01}(t))}{\sqrt{\phi_0\phi_{at}}} \\ \frac{1}{\phi_c}\left\{\sqrt{\frac{\phi_c+\phi_a}{\phi_1}}B_{11}(F_{11}(t)) - \sqrt{\frac{\phi_a}{\phi_0}}B_{01}(F_{01}(t))\right\} \\ \frac{B_{10}(F_{10}(t))}{\sqrt{\phi_0\phi_{nt}}}, \end{pmatrix},$$

and $B_{00}(\cdot), B_{01}(\cdot), B_{10}(\cdot)$, and $B_{11}(\cdot)$ are independent standard Brownian bridges, and the integrals in (B.31) *are defined coordinate-wise.*

B.3.1. *Proof of Corollary B.2.* The joint distribution of the process $\sqrt{n}(\overline{F}_{uv}(t)-F_{uv}(t))_{u,v\in\{0,1\}}$ can be easily derived from empirical process theory. To this end, let $D[0,1]$ be the space of all right-continuous functions on $[0,1]$ with left limits equipped with the supremum norm metric. A sequence of random functions $\{X_n(\cdot)\}_{n\geq 1}$ in $D[0,1]$ converges to $X(\cdot)\in D[0,1]$, denoted by $X_n(t) \overset{w}{\Rightarrow} X(t)$, if $\mathbb{E}(f(X_n)) \to \mathbb{E}f(X)$, for all bounded continuous function $f : D[0,1] \to \mathbb{R}$. Now, considering $\sqrt{n}(\overline{F}_{uv}(t) - F_{uv}(t))_{u,v\in\{0,1\}}$ as a random element of $D[0,1]^4$ equipped with the product topology, we have the following result:

LEMMA B.7. *Let B_{00}, B_{01}, B_{10}, and B_{11} be independent Brownian bridges. Then*

$$\sqrt{n}\begin{pmatrix}\overline{F}_{00}(t) - F_{00}(t) \\ \overline{F}_{01}(t) - F_{01}(t) \\ \overline{F}_{11}(t) - F_{11}(t) \\ \overline{F}_{10}(t) - F_{10}(t)\end{pmatrix} \overset{w}{\Rightarrow} \begin{pmatrix} \frac{B_{00}(F_{00}(t))}{\sqrt{\eta_{00}}} \\ \frac{B_{01}(F_{01}(t))}{\sqrt{\eta_{01}}} \\ \frac{B_{11}(F_{11}(t))}{\sqrt{\eta_{11}}} \\ \frac{B_{10}(F_{10}(t))}{\sqrt{\eta_{10}}} \end{pmatrix} \tag{B.32}$$

PROOF. Note that $\mathbb{E}(\breve{\boldsymbol{F}}(t)|\boldsymbol{Z},\boldsymbol{D})\}$ is the mean of the plug-in estimate conditional on the sigma-algebra generated by $(\boldsymbol{Z},\boldsymbol{D}) = ((Z_1,D_1),(Z_2,D_2),\ldots,(Z_n,D_n))$. Conditioned on this sigma-algebra, $\{n_{uv}\}_{u,v\in\{0,1\}}$ are fixed, and $\mathbb{E}(\overline{F}_{uv}(t)|\boldsymbol{D},\boldsymbol{Z}) = \mathbb{P}(Y_1 \leq t|Z_1 = u, D_1 = v) = F_{uv}(t)$. Moreover, if $s<t$, $\mathrm{Cov}(\overline{F}_{uv}(s),\overline{F}_{uv}(t)|\boldsymbol{D},\boldsymbol{Z}) = \frac{1}{n_{uv}}F_{uv}(s)(1-F_{uv}(t))$, and for $(u,v)\neq(u'v')$, $\mathrm{Cov}(\overline{F}_{uv}(s),\overline{F}_{u'v'}(t)|\boldsymbol{D},\boldsymbol{Z}) = 0$ since,

$$\begin{aligned} &\mathbb{E}\overline{F}_{uv}(t)\overline{F}_{u'v'}(t) \\ &= \frac{1}{n_{uv}n_{u'v'}} \sum_{\substack{a,a'\in[n] \\ Z_a=u,D_a=v,Z_{a'}=u,D_{a'}=v}} \mathbb{P}(Y_a \leq s, Y_{a'} \leq t | Z_a = u, D_a = v, Z_{a'} = u', D_{a'} = v') \\ &= F_{uv}(s)F_{u'v'}(t), \end{aligned} \tag{B.33}$$

whenever $(u,v)\neq(u'v')$.

Now, it is well-known that for each (u,v), $\sqrt{n_{uv}}(\overline{F}_{uv}(t) - F_{uv}(t)) \overset{w}{\Rightarrow} (B_{uv}(F_{uv}(t))_{\{u,v\}\in\{0,1\}}$, and therefore, $(\sqrt{n_{uv}}(\overline{F}_{uv}(t) - F_{uv}(t)))_{\{u,v\}\in\{0,1\}} \Rightarrow (B_{uv}(F_{uv}(t))_{\{u,v\}\in\{0,1\}}$. Then,

$$(\sqrt{n}(\overline{F}_{uv}(t) - F_{uv}(t)))_{\{u,v\}\in\{0,1\}} \overset{w}{\Rightarrow} (\frac{B_{uv}(F_{uv}(t))}{\sqrt{\eta_{uv}}})_{\{u,v\}\in\{0,1\}},$$

and the result follows. □

For $\boldsymbol{\chi} = (\chi_{nt}, \chi_{at}) \in \mathbb{R}^2$, define

$$C(\boldsymbol{\chi}) := \begin{pmatrix} \frac{1-\chi_{at}}{1-\chi_{nt}-\chi_{at}} & 0 & 0 & -\frac{\chi_{nt}}{1-\chi_{nt}-\chi_{at}} \\ 0 & 1 & 0 & 0 \\ 0 & -\frac{\chi_{at}}{1-\chi_{nt}-\chi_{at}} & \frac{1-\chi_{nt}}{1-\chi_{nt}-\chi_{at}} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Note that $\mathbb{E}(\breve{\boldsymbol{F}}(t)|\boldsymbol{Z},\boldsymbol{D}) = C(\breve{\boldsymbol{\phi}})(F_{00}(t), F_{01}(t), F_{11}(t), F_{10}(t))'$ and

$$\boldsymbol{G}(t) = C(\boldsymbol{\phi})\left(\frac{B_{00}(F_{00}(t))}{\sqrt{\eta_{00}}}, \frac{B_{01}(F_{01}(t))}{\sqrt{\eta_{01}}}, \frac{B_{11}(F_{11}(t))}{\sqrt{\eta_{11}}}, \frac{B_{10}(F_{10}(t))}{\sqrt{\eta_{10}}}\right)',$$

where $\boldsymbol{G}(\cdot)$ is as defined in the statement of Corollary B.2. Now, using the above lemma and the Donsker's invariance principle, and noting that $C(\breve{\boldsymbol{\phi}}) \overset{P}{\to} C(\boldsymbol{\phi})$, it follows that

$$\begin{aligned} &\int_{J_\kappa} h(t)\cdot\sqrt{n}(\breve{\boldsymbol{F}}(t) - \mathbb{E}(\breve{\boldsymbol{F}}(t)|\boldsymbol{Z},\boldsymbol{D}))\mathrm{d}H \\ &= \int_{J_\kappa} h(t)\cdot C(\breve{\boldsymbol{\phi}})\cdot\sqrt{n}\begin{pmatrix}\overline{F}_{00}(t) - F_{00}(t)\\ \overline{F}_{01}(t) - F_{01}(t)\\ \overline{F}_{11}(t) - F_{11}(t)\\ \overline{F}_{10}(t) - F_{10}(t)\end{pmatrix}\mathrm{d}H \\ &\overset{D}{\to} \int_{J_\kappa} h(t)\cdot\boldsymbol{G}(t)\mathrm{d}H, \end{aligned} \tag{B.34}$$

for any continuous function $h : \mathbb{R} \to \mathbb{R}$. This implies

$$\begin{aligned} &\int_{J_\kappa} h(t)\cdot\sqrt{n}\{\hat{\boldsymbol{F}}(t) - \mathbb{E}(\breve{\boldsymbol{F}}(t)|\boldsymbol{Z},\boldsymbol{D})\}\mathrm{d}H \\ &= \int_{J_\kappa} h(t)\cdot\sqrt{n}\{\breve{\boldsymbol{F}}(t) - \mathbb{E}(\breve{\boldsymbol{F}}(t)|\boldsymbol{Z},\boldsymbol{D})\}\mathrm{d}H + \int_{J_\kappa} h(t)\cdot\sqrt{n}\{\hat{\boldsymbol{F}}(t) - \breve{\boldsymbol{F}}(t)\}\mathrm{d}H \\ &\overset{D}{\to} \int_{J_\kappa}\boldsymbol{G}(t)\mathrm{d}H, \end{aligned} \tag{B.35}$$

using (B.34) for the first term, and second term is $o_P(1)$ by applying the Cauchy-Schwarz inequality followed by Theorem 3.2.

APPENDIX C: THE MBL ESTIMATORS UNDER THE NULL

In this section we analyze the MBL estimate of the distribution functions of the compliance classes under the null. To this end, define

$$\hat{\boldsymbol{\psi}} = (\hat{\psi}_{co}, \hat{\psi}_{nt}, \hat{\psi}_{at})' := \arg\max_{\boldsymbol{\psi} \in \boldsymbol{\vartheta}_{+,0}} \mathbb{M}_n(\boldsymbol{\psi}, \breve{\boldsymbol{\phi}}),$$

and

$$\begin{aligned} F_{00}^{\breve{\boldsymbol{\phi}}}(t) &= \breve{\lambda}_0 F_{co}(t) + (1-\breve{\lambda}_0) F_{nt}(t) \\ F_{01}^{\breve{\boldsymbol{\phi}}}(t) &= F_{at}(t), \\ F_{10}^{\breve{\boldsymbol{\phi}}}(t) &= F_{nt}(t), \\ F_{11}^{\breve{\boldsymbol{\phi}}}(t) &= \breve{\lambda}_1 F_{co}(t) + (1-\breve{\lambda}_1) F_{at}(t). \end{aligned} \tag{C.1}$$

Next, define the *population objective function*,[5] under the null H_0, as follows:

$$\mathbb{M}(\boldsymbol{\psi}, \breve{\boldsymbol{\phi}}) = T_{00}(\psi_{co}, \psi_{nt}, \breve{\boldsymbol{\phi}}) + T_{10}(\psi_{nt}, \breve{\boldsymbol{\phi}}) + T_{10}(\psi_{nt}, \breve{\boldsymbol{\phi}}) + T_{11}(\psi_{co}, \psi_{at}, \breve{\boldsymbol{\phi}}), \tag{C.2}$$

where

$$\begin{aligned} T_{00}(\psi_{co}, \psi_{nt}, \breve{\boldsymbol{\phi}}) &= \frac{1}{n} \sum_{b \in I_\kappa} \frac{n_{00}}{n} \left\{ \log\left(1 - \breve{\phi}_{at}\right) + J(F_{00}^{\breve{\boldsymbol{\phi}}}(Y_{(b)}), \breve{\lambda}_0 \psi_{co}(Y_{(b)}) + (1-\breve{\lambda}_0)\psi_{nt}(Y_{(b)})) \right\}, \\ T_{10}(\psi_{nt}, \breve{\boldsymbol{\phi}}) &= \frac{1}{n} \sum_{b \in I_\kappa} \frac{n_{10}}{n} \left\{ \log \breve{\phi}_{nt} + J(F_{10}^{\breve{\boldsymbol{\phi}}}(Y_{(b)}), \psi_{nt}(Y_{(b)})) \right\}, \\ T_{01}(\psi_{at}, \breve{\boldsymbol{\phi}}) &= \frac{1}{n} \sum_{b \in I_\kappa} \frac{n_{01}}{n} \left\{ \log \breve{\phi}_{at} + J(F_{01}^{\breve{\boldsymbol{\phi}}}(Y_{(b)}), \psi_{at}(Y_{(b)})) \right\}, \\ T_{11}(\psi_{co}, \psi_{at}, \breve{\boldsymbol{\phi}}) &= \frac{1}{n} \sum_{b \in I_\kappa} \frac{n_{11}}{n} \left\{ \log\left(1 - \breve{\phi}_{nt}\right) + J(F_{11}^{\breve{\boldsymbol{\phi}}}(Y_{(b)}), \breve{\lambda}_1 \psi_{co}(Y_{(b)}) + (1-\breve{\lambda}_1)\psi_{at}(Y_{(b)})) \right\}. \end{aligned}$$

Finally, recall the plug-in estimators of the distribution functions of the compliers:

$$\breve{F}_{co}^{(0)}(t) = \frac{\overline{F}_{00}(t) - (1-\breve{\lambda}_0)\overline{F}_{10}(t)}{\breve{\lambda}_0}, \quad \breve{F}_{co}^{(1)}(t) = \frac{\overline{F}_{11}(t) - (1-\breve{\lambda}_1)\overline{F}_{01}(t)}{\breve{\lambda}_1}.$$

The following result, as in Lemma B.5, shows how the MBL estimator under the null is related to the plug-in estimator:

[5]Note that we are slightly abusing terminology here, because the population objective function depends on the sample $\{Y_1, Y_2, \ldots, Y_n\}$. Ideally, one should define $\mathbb{M}(\cdot,\cdot)$ as an integral with respect to the population distribution function H. However, for technical reasons, it is more convenient for us to define $\mathbb{M}(\cdot,\cdot)$ with respect to the empirical measure instead.

PROPOSITION C.1. *Fix $0 < \kappa < 1$, and let I_κ be as defined in (3.4). Then, under the null H_0, the MBL estimator $\hat{\boldsymbol{\psi}} = (\hat{\psi}_{co}, \hat{\psi}_{nt}, \hat{\psi}_{at})'$ satisfies*

$$\frac{1}{n}\sum_{b\in I_\kappa}\left\|\begin{pmatrix}\sqrt{n}(\hat{\psi}_{co}(Y_{(b)}) - \breve{\tau}_{co}(Y_{(b)}))\\ \sqrt{n}(\hat{\psi}_{nt}(Y_{(b)}) - \breve{\tau}_{nt}(Y_{(b)}))\\ \sqrt{n}(\hat{\psi}_{at}(Y_{(b)}) - \breve{\tau}_{at}(Y_{(b)}))\end{pmatrix}\right\|_2^2 = o_P(1),$$

where

$$\begin{aligned}\breve{\tau}_{co}(t) &:= \frac{(\breve{C}_{01}(t) + \breve{C}_{11}(t))\breve{F}_{co}^{(0)}(t) + (\breve{C}_{10}(t) + \breve{C}_{00}(t))\breve{F}_{co}^{(1)}(t)}{\sum_{u,v\in\{0,1\}}\breve{C}_{uv}(t)}\\ \breve{\tau}_{nt}(t) &:= \overline{F}_{10}(t) + \frac{\frac{\check{\lambda}_0}{1-\check{\lambda}_0}\breve{C}_{10}(t)\left\{\breve{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(1)}(t)\right\}}{\sum_{u,v\in\{0,1\}}\breve{C}_{uv}(t)}\\ \breve{\tau}_{at}(t) &:= \overline{F}_{01}(t) + \frac{\frac{\check{\lambda}_1}{1-\check{\lambda}_1}\breve{C}_{01}(t)\left\{\breve{F}_{co}^{(1)}(t) - \breve{F}_{co}^{(0)}(t)\right\}}{\sum_{u,v\in\{0,1\}}\breve{C}_{uv}(t)},\end{aligned} \tag{C.3}$$

where $\breve{C}_{uv} = \frac{\check{\lambda}_0^2\check{\lambda}_1^2}{\check{\lambda}_{uv}^2}\cdot\frac{n}{n_{uv}}\cdot F_{uv}^{\breve{\boldsymbol{\phi}}}(Y_{(b)})(1 - F_{uv}^{\breve{\boldsymbol{\phi}}}(Y_{(b)}))$, with $\check{\lambda}_{uv}$ as defined in Theorem 6.1.

C.1. Proof of Proposition C.1. To begin with define

$$\breve{\boldsymbol{\psi}} := (\breve{\psi}_{co}, \breve{\psi}_{nt}, \breve{\psi}_{at})' := \max_{\boldsymbol{\psi}\in\boldsymbol{\vartheta}_0}\mathbb{M}_n(\boldsymbol{\psi}, \breve{\boldsymbol{\phi}}),$$

where $\boldsymbol{\vartheta}_0 = \left\{(\theta_{co}, \theta_{nt}, \theta_{at}) : \theta_{co}, \theta_{nt}, \theta_{at} \in \mathbb{R}^{\mathbb{R}}\right\}$, is the unrestricted null parameter space. In this case there is no-closed form expression of $\breve{\boldsymbol{\psi}}$. However, by the asymptotic expansion of the sample null objective function we can find an asymptotically equivalent formula for $\breve{\boldsymbol{\psi}}$.

LEMMA C.1. *Let $\breve{\boldsymbol{\psi}}$ and $\breve{\boldsymbol{\tau}} = (\breve{\tau}_{co}, \breve{\tau}_{nt}, \breve{\tau}_{at})'$ be as defined above. Then*

$$\sum_{b\in I_\kappa}||\breve{\boldsymbol{\psi}}(Y_{(b)}) - \breve{\boldsymbol{\tau}}(Y_{(b)})||_2^2 = o_P(1), \tag{C.4}$$

whenever $||\breve{\boldsymbol{\psi}}(Y_{(b)}) - \boldsymbol{F}_0(Y_{(b)})||_2 = o_P(1)$, for every $b \in I_\kappa$, where $\boldsymbol{F}_0 = (F_{co}, F_{nt}, F_{at})'$, with $F_{co}^{(0)} = F_{co}^{(1)} := F_{co}$, is the vector of true distribution functions under the null.

PROOF. Next, recall the definitions of $\mathbb{M}_n$ and $\mathbb{M}$ from (3.6) and (C.2). Then

$$\begin{aligned}&(\mathbb{M}_n - \mathbb{M})(\breve{\boldsymbol{\psi}}, \breve{\boldsymbol{\phi}}) - (\mathbb{M}_n - \mathbb{M})(\boldsymbol{F}_0, \breve{\boldsymbol{\phi}})\\ &= \frac{1}{n}\sum_{u,v\in\{0,1\}}\sum_{b\in I_\kappa}\frac{n_{uv}}{n}\frac{(\overline{F}_{uv}(Y_{(b)}) - F_{uv}^{\breve{\boldsymbol{\phi}}}(Y_{(b)}))}{F_{uv}^{\breve{\boldsymbol{\phi}}}(Y_{(b)})(1 - F_{uv}^{\breve{\boldsymbol{\phi}}}(Y_{(b)}))}(\hat{\psi}_{uv}(Y_{(b)}) - F_{uv}^{\breve{\boldsymbol{\phi}}}(Y_{(b)})) + O_P(n^{-\frac{3}{2}})\end{aligned}$$

$$= \frac{1}{n} \sum_{u,v \in \{0,1\}} \sum_{b \in I_\kappa} Q_{uv}^{\breve{\phi}}(Y_{(b)})(\hat{\psi}_{uv}(Y_{(b)}) - F_{uv}^{\breve{\phi}}(Y_{(b)})) + O_P(n^{-\frac{3}{2}}),$$

where $Q_{uv}^{\breve{\phi}}(Y_{(b)}) = \frac{n_{uv}}{n} \cdot \frac{1}{F_{uv}^{\breve{\phi}}(Y_{(b)})(1-F_{uv}^{\breve{\phi}}(Y_{(b)}))}$.

Now, under the null hypothesis, $F_{co}^{(0)} = F_{co}^{(1)} := F_{co}$, we can re-group the terms in the above sum in terms of $\hat{\psi}_{co}(Y_{(b)}) - F_{co}(Y_{(b)}), \hat{\psi}_{nt}(Y_{(b)}) - F_{nt}(Y_{(b)})$ and $\hat{\psi}_{at}(Y_{(b)}) - F_{at}(Y_{(b)})$, to get

$$\begin{aligned}(\mathbb{M}_n - \mathbb{M})(\breve{\boldsymbol{\psi}}, \breve{\boldsymbol{\phi}}) &- (\mathbb{M}_n - \mathbb{M})(\boldsymbol{F}_0, \breve{\boldsymbol{\phi}}) \\ &= \frac{1}{\sqrt{n}} \left\{ \sum_{b \in I_\kappa} (\breve{\boldsymbol{\psi}}(Y_{(b)}) - \boldsymbol{F}_0(Y_{(b)}))' \boldsymbol{Z}_n(Y_{(b)}) \right\} + O_P(n^{-\frac{3}{2}}), \end{aligned} \tag{C.5}$$

where $\boldsymbol{F}_0 = (F_{co}, F_{nt}, F_{at})'$ is the vector of true distribution functions under the null, and

$$\boldsymbol{Z}_n(t) = \frac{1}{\sqrt{n}} \begin{pmatrix} \breve{\lambda}_0 Q_{00}^{\breve{\phi}}(t)\{\overline{F}_{00}(t) - F_{00}^{\breve{\phi}}(t)\} + \breve{\lambda}_1 Q_{11}^{\breve{\phi}}(t)\{\overline{F}_{11}(t) - F_{11}^{\breve{\phi}}(t)\} \\ (1-\breve{\lambda}_0) Q_{00}^{\breve{\phi}}(t)\{\overline{F}_{00}(t) - F_{00}^{\breve{\phi}}(t)\} + Q_{10}^{\breve{\phi}}(t)\{\overline{F}_{10}(t) - F_{10}^{\breve{\phi}}(t)\} \\ (1-\breve{\lambda}_1) Q_{11}^{\breve{\phi}}(t)\{\overline{F}_{11}(t) - F_{11}^{\breve{\phi}}(t)\} + Q_{01}^{\breve{\phi}}(t)\{\overline{F}_{01}(t) - F_{01}^{\breve{\phi}}(t)\} \end{pmatrix}. \tag{C.6}$$

Next, denote by $\boldsymbol{V}_n$ the Hessian matrix of $\mathbb{M}(\boldsymbol{\psi}, \breve{\boldsymbol{\phi}})$ at the point $((\boldsymbol{F}_0(Y_{(b)}))_{b \in I_\kappa})$. Note that the Hessian matrix is block diagonal

$$\boldsymbol{V}_n = \operatorname{diag}(\boldsymbol{V}_n(Y_{(b)}))_{b \in I_\kappa}, \tag{C.7}$$

where $\boldsymbol{V}_n(\cdot) : \mathbb{R} \to \mathbb{R}^{3\times 3}$ is given by the following:

$$\begin{aligned}\boldsymbol{V}_n(t) &= \begin{pmatrix} \frac{\partial^2 \mathbb{M}(\boldsymbol{\psi},\breve{\boldsymbol{\phi}})}{\partial \psi_{co}(t)^2} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\psi},\breve{\boldsymbol{\phi}})}{\partial \psi_{co}(t)\partial \psi_{nt}(t)} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\psi},\breve{\boldsymbol{\phi}})}{\partial \psi_{co}(t)\partial \psi_{at}(t)} \\ \frac{\partial^2 \mathbb{M}(\boldsymbol{\psi},\breve{\boldsymbol{\phi}})}{\partial \psi_{nt}(t)\partial \psi_{co}(t)} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\psi},\breve{\boldsymbol{\phi}})}{\partial \psi_{nt}(t)^2} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\psi},\breve{\boldsymbol{\phi}})}{\partial \psi_{nt}(t)\partial \psi_{at}(t)} \\ \frac{\partial^2 \mathbb{M}(\boldsymbol{\psi},\breve{\boldsymbol{\phi}})}{\partial \psi_{at}(t)\partial \psi_{co}(t)} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\psi},\breve{\boldsymbol{\phi}})}{\partial \psi_{at}(t)\partial \psi_{nt}(t)} & \frac{\partial^2 \mathbb{M}(\boldsymbol{\psi},\breve{\boldsymbol{\phi}})}{\partial \psi_{at}(t)^2} \end{pmatrix} |_{\boldsymbol{\psi}=\boldsymbol{F}_0} \\ &= -\frac{1}{n} \begin{pmatrix} \breve{\lambda}_0^2 Q_{00}^{\breve{\phi}}(t) + \breve{\lambda}_1^2 Q_{11}^{\breve{\phi}}(t) & \breve{\lambda}_0(1-\breve{\lambda}_0) Q_{00}^{\breve{\phi}}(t) & \breve{\lambda}_1(1-\breve{\lambda}_1) Q_{11}^{\breve{\phi}}(t) \\ \breve{\lambda}_0(1-\breve{\lambda}_0) Q_{00}^{\breve{\phi}}(t) & (1-\breve{\lambda}_0)^2 Q_{00}^{\breve{\phi}}(t) + Q_{10}^{\breve{\phi}}(t) & 0 \\ \breve{\lambda}_1(1-\breve{\lambda}_1) Q_{11}^{\breve{\phi}}(t) & 0 & (1-\breve{\lambda}_1)^2 Q_{11}^{\breve{\phi}}(t) + Q_{01}^{\breve{\phi}}(t) \end{pmatrix}.\end{aligned}$$

Now, by a second order Taylor expansion of $\mathbb{M}(\hat{\boldsymbol{\psi}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}(\boldsymbol{F}_0, \breve{\boldsymbol{\phi}})$ around the point $((\boldsymbol{F}_0(Y_{(b)}))_{b \in I_\kappa})$ gives,

$$\mathbb{M}(\breve{\boldsymbol{\psi}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}(\boldsymbol{F}_0, \breve{\boldsymbol{\phi}}) = \frac{1}{2} \cdot \sum_{b \in I_\kappa} (\breve{\boldsymbol{\psi}}(Y_{(b)}) - \boldsymbol{F}_0(Y_{(b)}))' \boldsymbol{V}(Y_{(b)}) (\breve{\boldsymbol{\psi}}(Y_{(b)}) - \boldsymbol{F}_0(Y_{(b)})) + O_P(n^{-\frac{3}{2}}),$$

since the gradient of $\mathbb{M}(\boldsymbol{\psi}, \breve{\boldsymbol{\phi}})$ at the point $((\boldsymbol{F}_0(Y_{(b)}))_{b\in I_\kappa})$ is zero (by arguments similar to the proof of Lemma A.1). Then from (C.5)

$$\mathbb{M}_n(\breve{\boldsymbol{\psi}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\boldsymbol{F}_0, \breve{\boldsymbol{\phi}}) = \frac{1}{\sqrt{n}}\left\{\sum_{b\in I_\kappa}(\breve{\boldsymbol{\psi}}(Y_{(b)}) - \boldsymbol{F}_0(Y_{(b)}))'\boldsymbol{Z}_n(Y_{(b)})\right\}$$
$$+ \frac{1}{2}\cdot\sum_{b\in I_\kappa}(\breve{\boldsymbol{\psi}}(Y_{(b)}) - \boldsymbol{F}_0(Y_{(b)}))'\boldsymbol{V}_n(Y_{(b)})(\breve{\boldsymbol{\psi}}(Y_{(b)}) - \boldsymbol{F}_0(Y_{(b)})) + O_P(n^{-\frac{3}{2}}), \quad \text{(C.8)}$$

Similarly, replacing $\breve{\boldsymbol{\psi}}$ by $\boldsymbol{F}_0 - n^{-\frac{1}{2}}\boldsymbol{V}_n^{-1}\boldsymbol{Z}_n = \breve{\boldsymbol{\tau}}$ (by Lemma C.2 below), in (C.8), where $\boldsymbol{Z}_n = (\boldsymbol{Z}_n(Y_{(b)})')'_{b\in I_\kappa}$ and $\boldsymbol{V}_n = \mathrm{diag}(\boldsymbol{V}_n(Y_{(b)}))_{b\in I_\kappa}$, gives

$$\mathbb{M}_n(\breve{\boldsymbol{\tau}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\boldsymbol{F}, \breve{\boldsymbol{\phi}}) = -\frac{1}{2}\cdot\frac{1}{n}\sum_{b\in I}\boldsymbol{Z}_n(Y_{(b)})'\boldsymbol{V}_n(Y_{(b)})^{-1}\boldsymbol{Z}_n(Y_{(b)}) + O_P(n^{-\frac{3}{2}}), \qquad \text{(C.9)}$$

since $\frac{1}{n}\sum_{b\in I_\kappa}||\breve{\boldsymbol{\tau}}(Y_{(b)})||_2^2 = O_P(1/n)$. This implies, subtracting (C.9) from (C.8) gives,

$$\mathbb{M}_n(\breve{\boldsymbol{\psi}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\breve{\boldsymbol{\tau}}, \breve{\boldsymbol{\phi}})$$
$$= \sum_{b\in I_\kappa}(\breve{\boldsymbol{\psi}}(Y_{(b)}) - \breve{\boldsymbol{\tau}}(Y_{(b)}))'\boldsymbol{V}_n(Y_{(b)})(\breve{\boldsymbol{\tau}}(Y_{(b)}) - \hat{\boldsymbol{F}}_0(Y_{(b)})) + O_P(n^{-\frac{3}{2}}). \qquad \text{(C.10)}$$

Now, since $\mathbb{M}_n(\breve{\boldsymbol{\psi}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\breve{\boldsymbol{\tau}}, \breve{\boldsymbol{\phi}}) \geq 0$ and $\sup_{b\in I_\kappa}|| - \boldsymbol{V}(Y_{(b)})^{-1}||_\infty = O_P(n)$ (seen from Lemma C.2 below), the result in (C.4) follows.[6] □

LEMMA C.2. *For $t \in (0,1)$,*

$$\boldsymbol{V}_n^{-1}(t)\boldsymbol{Z}_n(t) = \sqrt{n}\begin{pmatrix} F_{co}(t) - \breve{\tau}_{co}(t) \\ F_{nt}(t) - \breve{\tau}_{nt}(t) \\ F_{at}(t) - \breve{\tau}_{at}(t) \end{pmatrix},$$

where $\breve{\tau}_{co}(t)$, $\breve{\tau}_{nt}(t)$, and $\breve{\tau}_{at}(t)$, are as defined in Proposition C.1.

PROOF. Recall $\boldsymbol{V}_n(t)$ from (C.7). Then by direct calculations, the inverse matrix $\boldsymbol{V}_n^{-1}(t)$ can be computed as:

$$\boldsymbol{V}_n^{-1}(t) = -\frac{n}{|\det(\boldsymbol{V}_n(t))|}\begin{pmatrix} W_{11}(t) & W_{12}(t) & W_{13}(t) \\ W_{21}(t) & W_{22}(t) & W_{23}(t) \\ W_{31}(t) & W_{32}(t) & W_{33}(t) \end{pmatrix}$$

where

$$|\det(\boldsymbol{V}_n(t))| = Q_{00}^{\breve{\phi}}(t)Q_{01}^{\breve{\phi}}(t)Q_{10}^{\breve{\phi}}(t)Q_{11}^{\breve{\phi}}(t)\cdot\left(\frac{\breve{\lambda}_1^2}{Q_{00}^{\breve{\phi}}(t)} + \frac{\breve{\lambda}_0^2(1-\breve{\lambda}_1)^2}{Q_{01}^{\breve{\phi}}(t)} + \frac{(1-\breve{\lambda}_0)^2\breve{\lambda}_1^2}{Q_{10}^{\breve{\phi}}(t)} + \frac{\breve{\lambda}_0^2}{Q_{11}^{\breve{\phi}}(t)}\right)$$

[6] For a symmetric matrix $\boldsymbol{A}$, denote by $||\boldsymbol{A}||_\infty$ the maximum eigenvalue of $\boldsymbol{A}$.

$$W_{11}(t) = \Big((1-\breve{\lambda}_0)^2 Q_{00}^{\breve{\phi}}(t) + Q_{10}^{\breve{\phi}}(t)\Big) \times \Big((1-\breve{\lambda}_1)^2 Q_{11}^{\breve{\phi}}(t) + Q_{01}^{\breve{\phi}}(t)\Big)$$
$$W_{12}(t) = -\breve{\lambda}_0(1-\breve{\lambda}_0) Q_{00}^{\breve{\phi}}(t) \times \Big((1-\breve{\lambda}_1)^2 Q_{11}^{\breve{\phi}}(t) + Q_{01}^{\breve{\phi}}(t)\Big)$$
$$W_{13}(t) = \Big((1-\breve{\lambda}_0)^2 Q_{00}^{\breve{\phi}}(t) + Q_{10}^{\breve{\phi}}(t)\Big) \times -\breve{\lambda}_1(1-\breve{\lambda}_1) Q_{11}^{\breve{\phi}}(t)$$
$$W_{21}(t) = W_{12}(t)$$
$$W_{22}(t) = \breve{\lambda}_0^2(1-\breve{\lambda}_1)^2 Q_{00}^{\breve{\phi}}(t) Q_{11}^{\breve{\phi}}(t) + \breve{\lambda}_0^2 Q_{00}^{\breve{\phi}}(t) Q_{01}^{\breve{\phi}}(t) + \breve{\lambda}_1^2 Q_{01}^{\breve{\phi}}(t) Q_{11}^{\breve{\phi}}(t)$$
$$W_{23}(t) = \breve{\lambda}_0(1-\breve{\lambda}_0)\breve{\lambda}_1(1-\breve{\lambda}_1) Q_{00}^{\breve{\phi}}(t) Q_{11}^{\breve{\phi}}(t)$$
$$W_{31}(t) = W_{13}(t)$$
$$W_{32}(t) = W_{23}(t)$$
$$W_{33}(t) = (1-\breve{\lambda}_0)^2\breve{\lambda}_1^2 Q_{00}^{\breve{\phi}}(t) Q_{11}^{\breve{\phi}}(t) + \breve{\lambda}_0^2 Q_{00}^{\breve{\phi}}(t) Q_{10}^{\breve{\phi}}(t) + \breve{\lambda}_1^2 Q_{10}^{\breve{\phi}}(t) Q_{11}^{\breve{\phi}}(t).$$

Then recall the matrix $\boldsymbol{Z}_n(t)$ from (C.6). The proof of the result follows by direct multiplication. □

The proof Proposition C.1 can now completed by arguments similar to the proof of Proposition B.1. We outline the steps below, omitting the details:

- To begin with define, $\tilde{\boldsymbol{\tau}} = (\tilde{\tau}_{co}, \tilde{\tau}_{nt}, \tilde{\tau}_{at})$, as follows:

$$\tilde{\tau}_s := \arg \min_{\theta \in \mathcal{P}([0,1]^{\mathbb{R}})} \sum_{b \in I_\kappa} (\breve{\tau}_s(Y_{(b)}) - \theta(Y_{(b)}))^2, \tag{C.11}$$

  where $s \in \{co, nt, at\}$. Then as in Proposition B.1, it can shown that

$$\sum_{b \in I_\kappa} ||\tilde{\boldsymbol{\tau}}(Y_{(b)}) - \breve{\boldsymbol{\tau}}(Y_{(b)})||_2^2 = o_P(n^{-\frac{1}{4}}).$$

  This implies $\mathbb{M}_n(\breve{\boldsymbol{\tau}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{\psi}}, \breve{\boldsymbol{\phi}}) \leq \mathbb{M}_n(\breve{\boldsymbol{\tau}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\tilde{\boldsymbol{\tau}}, \breve{\boldsymbol{\phi}}) = o_P(n^{-\frac{5}{4}})$.
- Then as in the proof of Lemma B.5 in Section B.2, it follows that $\sum_{b \in I_\kappa} ||\hat{\boldsymbol{\psi}}(Y_{(b)}) - \breve{\boldsymbol{\tau}}(Y_{(b)})||_2^2 = o_P(n^{-\frac{1}{4}})$, completing the proof of Proposition C.1.

APPENDIX D: PROOF OF THEOREM 6.1

In this section we derive the asymptotic distribution of the binomial likelihood ratio statistic (BLRT). Recall from (6.4) that the BLRT statistic

$$T_n := \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_+} \mathbb{M}_n(\boldsymbol{\theta}, \breve{\boldsymbol{\phi}}) - \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_{+,0}} \mathbb{M}_n(\boldsymbol{\theta}, \breve{\boldsymbol{\phi}}),$$

where $\breve{\boldsymbol{\phi}} = (\breve{\phi}_{nt}, \breve{\phi}_{at})$, with $\breve{\phi}_a = \frac{n_{01}}{n_{00}+n_{01}}, \breve{\phi}_n = \frac{n_{10}}{n_{10}+n_{11}}$ and $\breve{\phi}_c = 1 - \breve{\phi}_n - \breve{\phi}_a$, is the plug-in estimate of the compliance classes. Denote

$$\hat{\boldsymbol{\theta}} := \arg\min \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_+} \mathbb{M}_n(\boldsymbol{\theta}, \breve{\boldsymbol{\phi}}), \quad \hat{\boldsymbol{\psi}} := \arg\min \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}_{+,0}} \mathbb{M}_n(\boldsymbol{\psi}, \breve{\boldsymbol{\phi}}), \tag{D.1}$$

where $\hat{\boldsymbol{\theta}}(t) = (\hat{\theta}_{co}^{(0)}(t), \hat{\theta}_{nt}(t), \hat{\theta}_{co}^{(1)}(t), \hat{\theta}_{at}(t))$ and $\hat{\boldsymbol{\psi}}(t) = (\hat{\psi}_{co}(t), \hat{\psi}_{nt}(t), \hat{\psi}_{at}(t))$. Finally, recall from (2.11), the plug-in estimates of λ_0, λ_1

$$\breve{\lambda}_0 = \frac{\breve{\phi}_c}{\breve{\phi}_c + \breve{\phi}_n}, \quad \breve{\lambda}_1 = \frac{\breve{\phi}_c}{\breve{\phi}_c + \breve{\phi}_a},$$

and, define

$$\begin{aligned} \hat{\theta}_{00}(t) &= \breve{\lambda}_0 \hat{\theta}_{co}^{(0)}(t) + (1 - \breve{\lambda}_0)\hat{\theta}_{nt}(t), & \hat{\psi}_{00}(t) &= \breve{\lambda}_0 \hat{\psi}_{co}(t) + (1 - \breve{\lambda}_0)\hat{\psi}_{nt}(t) \\ \hat{\theta}_{01}(t) &= \hat{\theta}_{at}(t), & \hat{\psi}_{01}(t) &= \hat{\psi}_{at}(t) \\ \hat{\theta}_{10}(t) &= \hat{\theta}_{nt}(t), & \hat{\psi}_{10}(t) &= \hat{\psi}_{nt}(t) \\ \hat{\theta}_{11}(t) &= \breve{\lambda}_1 \hat{\theta}_{co}^{(1)}(t) + (1 - \breve{\lambda}_1)\hat{\theta}_{at}(t), & \hat{\psi}_{11}(t) &= \breve{\lambda}_1 \hat{\psi}_{co}(t) + (1 - \breve{\lambda}_1)\hat{\psi}_{at}(t), \end{aligned}$$

With the above notation, we now have the following lemma, which shows that, under the null, the restricted as well as the unrestricted MBL estimates of the true distribution functions are asymptotically close to the plug-in estimates.

LEMMA D.1. *Under the null H_0, the following holds:*

$$\sum_{b \in I_\kappa} \sum_{u,v \in \{0,1\}} \left(\hat{\theta}_{uv}(Y_{(b)}) - \overline{F}_{uv}(Y_{(b)}) \right)^2 = o_P(1), \tag{D.2}$$

and

$$\sum_{b \in I_\kappa} \sum_{u,v \in \{0,1\}} \left(\hat{\psi}_{uv}(Y_{(b)}) - \breve{\tau}_{uv}(Y_{(b)}) \right)^2 = o_P(1), \tag{D.3}$$

where $\breve{\tau}_{00}(t) = \breve{\lambda}_0 \breve{\tau}_{co}(t) + (1 - \breve{\lambda}_0)\breve{\tau}_{nt}(t)$, $\breve{\tau}_{01}(t) = \breve{\tau}_{at}(t)$, $\breve{\tau}_{10}(t) = \breve{\tau}_{nt}(t)$, and $\breve{\tau}_{11}(t) = \breve{\lambda}_1 \breve{\tau}_{co}(t) + (1 - \breve{\lambda}_1)\breve{\tau}_{at}(t)$, with $(\breve{\tau}_{co}(t), \breve{\tau}_{nt}(t), \breve{\tau}_{at}(t))$ as defined in Proposition C.1.

PROOF. The result in (D.2) can be shown by arguments similar to the proof Lemma of B.5. Recall that Lemma B.5 shows that $\sum_{b \in I_\kappa} \sum_{u,v \in \{0,1\}} (\hat{F}_{uv}(Y_{(b)}) - \overline{F}_{uv}(Y_{(b)}))^2 = o_P(1)$, where $\hat{\boldsymbol{F}}$ is the BL estimate of $\boldsymbol{F}$, the vector of true distribution functions, when the proportion of the compliance classes are estimated by maximizing the binomial likelihood function. On the other hand, $\hat{\boldsymbol{\theta}}$ is the BL estimate of $\boldsymbol{F}$, when the proportion of the compliance classes are estimated by the plug-in estimates. Nevertheless, the proof of Lemma B.5 can be repeated verbatim to show (D.2).

The result in (D.3) follows from Proposition C.1 and the definition of $\{\breve{\tau}_{uv}\}_{u,v \in \{0,1\}}$. □

Using this lemma leading term of the asymptotic expansion of the BLRT can be derived as follows:

LEMMA D.2. *Let* $\{\breve{\tau}_{uv}\}_{u,v\in\{0,1\}}$ *and* $\{\hat{\psi}_{uv}\}_{u,v\in\{0,1\}}$, *be as defined above. Then the BLRT statistic satisfies*

$$T_n = \frac{1}{n}\sum_{\{u,v\}\in\{0,1\}}\sum_{b\in I_\kappa} n_{uv}\left\{\frac{(\breve{\tau}_{uv}(Y_{(b)}) - \overline{F}_{uv}(Y_{(b)}))^2}{\overline{F}_{uv}(Y_{(b)})(1-\overline{F}_{uv}(Y_{(b)}))}\right\} + O_P(1/\sqrt{n}) \tag{D.4}$$

PROOF. Recall that the definitions of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\psi}}$ from (D.1). Then, the BLRT can be rewritten as,

$$T_n = \mathbb{M}_n(\hat{\boldsymbol{\theta}}, \breve{\boldsymbol{\phi}}) - \mathbb{M}_n(\hat{\boldsymbol{\psi}}, \breve{\boldsymbol{\phi}}) = \frac{1}{n}\sum_{\{u,v\}\in\{0,1\}}\sum_{b\in I_\kappa}(T_{uv}(Y_{(b)}|\hat{\boldsymbol{\theta}}) - T_{uv}(Y_{(b)}|\hat{\boldsymbol{\psi}}), \tag{D.5}$$

where

$$\begin{aligned}
T_{uv}(Y_{(b)}|\hat{\boldsymbol{\theta}}) =& \frac{n_{uv}}{n}\left\{\overline{F}_{uv}(Y_{(b)})\log\hat{\theta}_{uv}(Y_{(b)}) + (1-\overline{F}_{uv}(Y_{(b)}))\log(1-\hat{\theta}_{uv}(Y_{(b)}))\right\}\\
T_{uv}(Y_{(b)}|\hat{\boldsymbol{\psi}}) =& \frac{n_{uv}}{n}\left\{\overline{F}_{uv}(Y_{(b)})\log\hat{\psi}_{uv}(Y_{(b)}) + (1-\overline{F}_{uv}(Y_{(b)}))\log(1-\hat{\psi}_{uv}(Y_{(b)}))\right\}.
\end{aligned}$$

Recall the definition of the (negative) binary entropy function $I(x) = x\log x + (1-x)\log(1-x)$. Then, note that

$$\begin{aligned}
&T_{uv}(Y_{(b)}|\hat{\boldsymbol{\theta}}) - I(\overline{F}_{uv}(Y_{(b)}))\\
&= \frac{n_{uv}}{n}\left\{\overline{F}_{uv}(Y_{(b)})\log\frac{\hat{\theta}_{uv}(Y_{(b)})}{\overline{F}_{uv}(Y_{(b)})} + (1-\overline{F}_{uv}(Y_{(b)}))\log\frac{1-\hat{\theta}_{uv}(Y_{(b)})}{1-\overline{F}_{uv}(Y_{(b)})}\right\}\\
&= R_{uv}^{(b)},
\end{aligned} \tag{D.6}$$

where

$$R_{uv}^{(b)} = \frac{n_{uv}}{n}\cdot\frac{(\hat{\theta}_{uv}(Y_{(b)}) - \overline{F}_{uv}(Y_{(b)}))^2}{4}\left\{\frac{\overline{F}_{uv}(Y_{(b)})}{(\omega_{uv}(Y_{(b)}))^2} - \frac{1-\overline{F}_{uv}(Y_{(b)})}{(1-\omega_{uv}(Y_{(b)}))^2}\right\},$$

and $\omega_{uv}(Y_{(b)}) \in [\overline{F}_{uv}(Y_{(b)})\wedge\hat{\theta}_{uv}(Y_{(b)}), \hat{\theta}_{uv}(Y_{(b)})\vee\overline{F}_{uv}(Y_{(b)})]$.

Note that $\omega_{uv}(Y_{(b)}) \geq \overline{F}_{uv}(Y_{(\lceil n\kappa\rceil)})\wedge\hat{\theta}_{uv}(Y_{(\lceil n\kappa\rceil)})$ and $\overline{F}_{uv}(Y_{(b)}) \leq \overline{F}_{uv}(Y_{(\lceil n(1-\kappa)\rceil)})$. Therefore,

$$\frac{\overline{F}_{uv}(Y_{(b)})}{(\omega_{uv}(Y_{(b)}))^2} \leq \frac{\overline{F}_{uv}(Y_{(\lceil n(1-\kappa)\rceil)})}{\overline{F}_{uv}(Y_{(n\kappa)})\wedge\hat{\theta}_{uv}(Y_{(n\kappa)})} = O_P(1),$$

since $\overline{F}_{uv}(Y_{(\lceil n\kappa\rceil)}) \xrightarrow{P} H_{uv}^{-1}(\kappa)$, $\overline{F}_{uv}(Y_{(\lceil n(1-\kappa)\rceil)}) \xrightarrow{P} H_{uv}^{-1}(1-\kappa)$ using Observation A.3, and $|\hat{\theta}_{uv}(Y_{(\lceil n\kappa\rceil)}) - \overline{F}_{uv}(Y_{(\lceil n\kappa\rceil)})| = o_P(1)$ by Lemma D.1. Similarly,

$$\frac{1-\overline{F}_{uv}(Y_{(b)})}{(1-\omega_{uv}(Y_{(b)}))^2} = O_P(1).$$

Therefore,

$$\sum_{b\in I_\kappa}|R_{uv}^{(b)}| \le O_P(1)\sum_{b\in I_\kappa}|\hat{\theta}_{uv}(Y_{(b)})-\overline{F}_{uv}(Y_{(b)})|^2$$
$$\le O_P(1)\sum_{b\in I_\kappa}|\overline{F}_{uv}(Y_{(b)})-\hat{\theta}_{uv}(Y_{(b)})|^2 = o_P(1), \tag{D.7}$$

by (D.2). Therefore, by (D.6),

$$\frac{1}{n}\sum_{\{u,v\}\in\{0,1\}}\sum_{b\in I_\kappa}T_{uv}(Y_{(b)}|\hat{\boldsymbol{\theta}})-I(\overline{F}_{uv}(Y_{(b)}))=o_P(1/n). \tag{D.8}$$

Similarly, by a second order Taylor approximation,

$$T_{uv}(Y_{(b)}|\hat{\boldsymbol{\psi}})-I(\overline{F}_{uv}(Y_{(b)}))=\frac{n_{uv}}{n}\cdot\frac{1}{2}\cdot\frac{(\hat{\psi}_{uv}(Y_{(b)})-\overline{F}_{uv}(Y_{(b)}))^2}{\overline{F}_{uv}(Y_{(b)})(1-\overline{F}_{uv}(Y_{(b)}))}+W_{uv}^{(b)}, \tag{D.9}$$

where

$$W_{uv}^{(b)}=\frac{n_{uv}}{n}\cdot\frac{(\hat{\psi}_{uv}(Y_{(b)})-\overline{F}_{uv}(Y_{(b)}))^3}{6}\left\{\frac{\overline{F}_{uv}(Y_{(b)})}{(\omega_{uv}(Y_{(b)}))^3}-\frac{1-\overline{F}_{uv}(Y_{(b)})}{(1-\omega_{uv}(Y_{(b)}))^3}\right\},$$

and $\omega_{uv}(Y_{(b)})\in[\overline{F}_{uv}(Y_{(b)})\wedge\hat{\psi}_{uv}(Y_{(b)}),\hat{\psi}_{uv}(Y_{(b)})\vee\overline{F}_{uv}(Y_{(b)})]$. Now, as in (D.7),

$$\begin{aligned}
\frac{1}{n}\sum_{b\in I_\kappa}|W_{uv}^{(b)}| &\le O_P(1)\frac{1}{n}\sum_{b\in I_\kappa}|\hat{\psi}_{uv}(Y_{(b)})-\overline{F}_{uv}(Y_{(b)})|^3\\
&\le O_P(1)\frac{1}{n}\sum_{b\in I_\kappa}|\overline{F}_{uv}(Y_{(b)})-\hat{\psi}_{uv}(Y_{(b)})|^3\\
&\le O_P(1)\frac{1}{n}\sum_{b\in I_\kappa}|\overline{F}_{uv}(Y_{(b)})-\breve{\tau}_{uv}(Y_{(b)})|^3+O_P(1)\frac{1}{n}\sum_{b\in I_\kappa}|\hat{\psi}_{uv}(Y_{(b)})-\breve{\tau}_{uv}(Y_{(b)})|^3\\
&\le O_P(n^{-\frac{3}{2}})+O_P(1)\frac{1}{n}\left(\sum_{b\in I_\kappa}|\hat{\psi}_{uv}(Y_{(b)})-\breve{\tau}_{uv}(Y_{(b)})|^2\right)^{\frac{3}{2}}\\
&=o_P(1/n),
\end{aligned} \tag{D.10}$$

using $\sup_t|\overline{F}_{uv}(t)-\breve{\tau}_{uv}(t)|=O_P(1/\sqrt{n})$ for the first term, and Cauchy-Schwarz followed by (D.3) in the second term.

Therefore, combing (D.9) with (D.10) above gives,

$$\frac{1}{n}\sum_{\{u,v\}\in\{0,1\}}\sum_{b\in I_\kappa}T_{uv}(Y_{(b)}|\hat{\boldsymbol{\psi}})-I(\overline{F}_{uv}(Y_{(b)}))$$

$$= \frac{1}{n} \sum_{\{u,v\} \in \{0,1\}} \sum_{b \in I_\kappa} \frac{n_{uv}}{n} \cdot \frac{1}{2} \cdot \frac{(\hat{\psi}_{uv}(Y_{(b)}) - \overline{F}_{uv}(Y_{(b)}))^2}{\overline{F}_{uv}(Y_{(b)})(1 - \overline{F}_{uv}(Y_{(b)}))} + o_P(1/n)$$

$$= \frac{1}{n} \sum_{\{u,v\} \in \{0,1\}} \sum_{b \in I_\kappa} \frac{n_{uv}}{n} \cdot \frac{1}{2} \cdot \frac{(\breve{\tau}_{uv}(Y_{(b)}) - \overline{F}_{uv}(Y_{(b)}))^2}{\overline{F}_{uv}(Y_{(b)})(1 - \overline{F}_{uv}(Y_{(b)}))} + o_P(1/n), \tag{D.11}$$

where the last step uses triangle inequality and (D.3). Combining (D.8) and (D.11) with (D.5) the result follows. □

The proof of Theorem 6.1 can now be completed by simplifying the RHS of (D.4). To this end, we have the following observation:

OBSERVATION D.1. *Let* $\{\breve{\tau}_{uv}\}_{u,v\in\{0,1\}}$ *be as defined in Lemma D.1. Then, for* $u, v \in \{0,1\}$,

$$\breve{\tau}_{uv}(t) = \overline{F}_{uv}(t) + \breve{\lambda}_{uv} \cdot \frac{\breve{C}_{uv}(t)}{\sum_{u,v\in\{0,1\}} \breve{C}_{uv}(t)} \left(\breve{F}_{co}^{(1)}(t) - \breve{F}_{co}^{(0)}(t) \right), \tag{D.12}$$

where $\{\breve{\lambda}_{uv}\}_{u,v\in\{0,1\}}$ *and* $\{\breve{C}_{uv}\}_{u,v\in\{0,1\}}$ *are as defined in Theorem 6.1 and Proposition (C.1), respectively.*

PROOF. Recall the plug-in estimates $\breve{F}_{co}^{(0)}(t)$ and $\breve{F}_{co}^{(1)}(t)$ from (2.14). Then from (C.3) it follows that

$$\begin{aligned}
\breve{\tau}_{00}(t) &= \breve{\lambda}_0 \breve{\tau}_{co}(t) + (1 - \breve{\lambda}_0)\breve{\tau}_{nt}(t) \\
&= \overline{F}_{00}(t) + \frac{\breve{C}_{00}(t)\breve{\lambda}_0}{\sum_{u,v\in\{0,1\}} \breve{C}_{uv}(t)} \left\{ \frac{\overline{F}_{11}(t) - (1 - \breve{\lambda}_1)\overline{F}_{01}(t)}{\breve{\lambda}_1} - \frac{\overline{F}_{00}(t) - (1 - \breve{\lambda}_0)\overline{F}_{10}(t)}{\breve{\lambda}_0} \right\} \\
&= \overline{F}_{00}(t) + \breve{\lambda}_{00} \cdot \frac{\breve{C}_{00}(t)}{\sum_{u,v\in\{0,1\}} \breve{C}_{uv}(t)} \left(\breve{F}_{co}^{(1)}(t) - \breve{F}_{co}^{(0)}(t) \right).
\end{aligned}$$

The expressions for $\breve{\tau}_{01}(t)$, $\breve{\tau}_{10}(t)$, and $\breve{\tau}_{11}(t)$ can be computed similarly. □

Substituting (D.12) in the RHS of equation (D.4), the leading term of the BLRT simplifies as follows:

$$T_n = \frac{1}{n} \sum_{u,v\in\{0,1\}} \sum_{b\in I_\kappa} \left[\frac{\breve{\lambda}_{uv}\breve{C}_{uv}(Y_{(b)})\sqrt{\breve{Q}_{uv}(Y_{(b)})}}{\sum_{u,v\in\{0,1\}} \breve{C}_{uv}(Y_{(b)})} \sqrt{n}(\breve{F}_{co}^{(1)}(Y_{(b)}) - \breve{F}_{co}^{(0)}(Y_{(b)})) \right]^2 + o_P(1).$$

using $\breve{Q}_{uv}(t) = \frac{n_{uv}/n}{\overline{F}_{uv}(t)(1-\overline{F}_{uv})(t)}$. Next, recalling he definition of $\{\breve{C}_{uv}(t)\}_{u,v\in\{0,1\}}$ from Proposition C.1, note that

$$\breve{C}_{uv}(t) = \frac{\breve{\lambda}_0^2\breve{\lambda}_1^2}{\breve{\lambda}_{uv}^2} \cdot \frac{n}{n_{uv}} \cdot F_{uv}^{\breve{\phi}}(t)(1 - F_{uv}^{\breve{\phi}}(t)) = \frac{\breve{\lambda}_0^2\breve{\lambda}_1^2}{\breve{\lambda}_{uv}^2} \cdot \frac{1}{Q_{uv}^{\breve{\phi}}(t)}.$$

Therefore, the leading term of T_n can is further simplified as

$$\begin{aligned} T_n &= (\breve{\lambda}_0^2\breve{\lambda}_1^2) \cdot \frac{1}{n} \sum_{b\in I_\kappa} \frac{\sum_{u,v\in\{0,1\}} \frac{\breve{\lambda}_0^2\breve{\lambda}_1^2}{\breve{\lambda}_{uv}^2} \frac{\breve{Q}_{uv}(Y_{(b)})}{Q_{uv}^{\breve{\phi}}(Y_{(b)})^2}}{(\sum_{u,v\in\{0,1\}} \breve{C}_{uv}(Y_{(b)}))^2} (\sqrt{n}\{\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(1)}(Y_{(b)})\})^2 + o_P(1) \\ &= (\breve{\lambda}_0^2\breve{\lambda}_1^2) \cdot \frac{1}{n} \sum_{b\in I_\kappa} \frac{1}{\sum_{u,v\in\{0,1\}} \breve{C}_{uv}(Y_{(b)})} (\sqrt{n}\{\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(1)}(Y_{(b)})\})^2 + o_P(1), \end{aligned} \tag{D.13}$$

where the last step uses $\sup_{b\in I_\kappa} \left|\frac{\breve{Q}_{uv}(Y_{(b)})}{Q_{uv}^{\breve{\phi}}(Y_{(b)})} - 1\right| = o_P(1)$, since both $\breve{Q}_{uv}(t)$ and $Q_{uv}^{\breve{\phi}}(t)$ converges to $\frac{\eta_{uv}}{F_{uv}(t)(1-F_{uv}(t))} := Q_{uv}(t)$ in supremum norm. Using this and $\breve{\lambda}_0 \xrightarrow{P} \lambda_0$ and $\breve{\lambda}_1 \xrightarrow{P} \lambda_1$, the RHS of (D.13) can be further simplified as

$$\begin{aligned} T_n &= \frac{1}{n} \sum_{b\in I_\kappa} \frac{(\sqrt{n}\{\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(1)}(Y_{(b)})\})^2}{\sum_{u,v\in\{0,1\}} \frac{1}{\breve{\lambda}_{uv}^2}\frac{n}{n_{uv}} F_{uv}(Y_{(b)})(1 - F_{uv}(Y_{(b)}))} + o_P(1) \\ &= \frac{1}{n} \sum_{b\in I_\kappa} \frac{(\sqrt{n}\{\breve{F}_{co}^{(0)}(Y_{(b)}) - \breve{F}_{co}^{(1)}(Y_{(b)})\})^2}{\sum_{u,v\in\{0,1\}} \frac{1}{\breve{\lambda}_{uv}^2}\frac{n}{n_{uv}} \overline{F}_{uv}(Y_{(b)})(1 - \overline{F}_{uv}(Y_{(b)}))} + o_P(1). \end{aligned} \tag{D.14}$$

where the last step follows from $\sup_{b\in I_\kappa} |\frac{\overline{F}_{uv}(Y_{(b)})}{F_{uv}(Y_{(b)})} - 1| = o_P(1)$. This completes the proof of Theorem 6.1. □

To re-write the denominator in (D.14) as the conditional variance of the numerator, note that $n\,\mathrm{Var}(\overline{F}_{uv}(t)|(\boldsymbol{Z}, \boldsymbol{D})) = \frac{n}{n_{uv}} F_{uv}(t)(1 - F_{uv}(t))$. Therefore, by the independence of $\overline{F}_{uv}$ (conditional on $(\boldsymbol{Z}, \boldsymbol{D})$),

$$\begin{aligned} &n\,\mathrm{Var}(\breve{F}_{co}^{(0)}(t) - \breve{F}_{co}^{(1)}(t)|(\boldsymbol{Z}, \boldsymbol{D})) \\ &= \mathrm{Var}\left(\frac{\overline{F}_{00}(t) - (1 - \breve{\lambda}_0)\overline{F}_{10}(t)}{\breve{\lambda}_0} - \frac{\overline{F}_{11}(t) - (1 - \breve{\lambda}_1)\overline{F}_{01}(t)}{\breve{\lambda}_1} |(\boldsymbol{Z}, \boldsymbol{D})\right) \\ &= \sum_{u,v\in\{0,1\}} \frac{1}{\breve{\lambda}_{uv}^2}\frac{n}{n_{uv}} F_{uv}(t)(1 - F_{uv}(t)), \end{aligned} \tag{D.15}$$

recalling the definition of $\{\breve{\lambda}_{uv}\}_{u,v\in\{0,1\}}$ from Theorem 6.1. Therefore, by mimicking the proof of Lemma B.6, the RHS of (D.14) simplifies to (6.6).

APPENDIX E: PROOFS FROM SECTION 4

In this section we recall the well-known PAVA algorithm Barlow et al. (1972); de Leeuw, Hornik and Mair (2009), elaborate on the EM-PAVA algorithm, and prove Proposition 4.1.

E.1. The PAVA Algorithm. The PAVA algorithm takes input a vector $\boldsymbol{u} = (u_1, \dots, u_n)'$ and an a weight vector $\boldsymbol{w} = (w_1, \dots, w_n)'$, and returns another vector $\mathrm{PAVA}_{\boldsymbol{w}}(\boldsymbol{u}) := (\hat{u}_1, \dots, \hat{u}_n)'$ such that

$$\mathrm{PAVA}_{\boldsymbol{w}}(\boldsymbol{u}) := \arg \min_{v_1 \le v_2 \le \cdots \le v_n} \sum_{i=1}^{n} w_i (u_i - v_i)^2. \tag{E.1}$$

The weighted PAVA algorithm is as follows: To begin with set $\hat{u}_a = u_a$ for all $a \in [n]$.

Step 1. If $\hat{u}_1 \le \hat{u}_2$, move to Step 2. Otherwise, $\hat{u}_1 > \hat{u}_2$ in which case the values are updated as

$$\hat{u}_1 = \hat{u}_2 \leftarrow \frac{w_1 u_1 + w_2 u_2}{w_1 + w_2},$$

the weighted average of the original values of $\{u_1, u_2\}$. Then, move to Step 2. Note that the first step does not update the points from the third to the last, that is, $\hat{u}_a = u_a$, for $a \in [3, n]$.

Step 2. For the a-th point, compare $\hat{u}_a$ with $\hat{u}_{a+1}$. If $\hat{u}_a \le \hat{u}_{a+1}$, then $\hat{u}_a$ remains the same and the algorithm moves to the next point. If $\hat{u}_a > \hat{u}_{a+1}$, then $\hat{u}_a = \hat{u}_{a+1} \leftarrow \frac{w_a \hat{u}_a + w_{a+1} \hat{u}_{a+1}}{w_1 + w_2}$, the weighted average of $\{\hat{u}_a, \hat{u}_{a+1}\}$. Then new value is compared with $\hat{u}_{a-1}$. If the required monotonicity assumption is achieved, that is, $\hat{u}_{a-1} \le \hat{u}_a$, then the algorithm moves to the $(a+1)$-th point. Otherwise, $\hat{u}_{a-1} > \hat{u}_a$, in which case $\hat{u}_{a-1} = \hat{u}_a = \hat{u}_{a+1}$ is updated by the weighted average of $\{\hat{u}_{a-1}, \hat{u}_a, \hat{u}_{a+1}\}$. This repeated until a sequence the partial sequence $\hat{u}_1, \dots, \hat{u}_a$ is non-decreasing. Then the algorithm moves to the $(a+1)$-th point.

It is well known that the output $\mathrm{PAVA}_{\boldsymbol{w}}(\boldsymbol{u}) = (\hat{u}_1, \hat{u}_2, \dots, \hat{u}_n)$ of the above algorithm is non-decreasing and is the solution for the optimization problem (E.1). For example, suppose $\boldsymbol{u} = (3, 2, 1)$ and $\boldsymbol{w} = (1/3, 1/3, 1/3)$. Then PAVA algorithm updates $\boldsymbol{u}$ in the following order,

$$(3, 2, 1) \to (5/2, 5/2, 1) \to (5/2, 7/4, 7/4) \to (2, 2, 2).$$

For our experiments, we have used the `pava` function in the R package `Iso`, which implements weighted PAVA algorithm described.

E.2. Proofs from the EM-PAVA Algorithm. In this section we fill in the details of the EM-PAVA algorithm described in Section 4 above.

E.2.1. *The Expectation Step.* To begin recall the definition of $K_{u,S}^a = \{Z_a = u, S_a = s\}$. Then recalling (4.3) and (4.4), the complete data binomial log-likelihood (4.5) can be re-written as follows:

$$\log \overline{L}(\boldsymbol{\theta}, \boldsymbol{\chi}|\overline{\mathcal{D}}_n) = \sum_{a\in[n]} \sum_{b\in I_\kappa} \sum_{u\in\{0,1\}} \sum_{s\in\{co,nt,at\}} S_{u,s}(Y_a, Y_{(b)}), \tag{E.2}$$

where

$$\begin{aligned} S_{u,co}(Y_a, Y_{(b)}) = \mathbf{1}\{K_{u,co}^a\} \log \chi_{co} &+ \mathbf{1}\{Y_a \le Y_{(b)}, K_{u,co}^a\} \log \theta_{co}^{(u)}(Y_{(b)}) \\ &+ \mathbf{1}\{Y_a > Y_{(b)}, K_{u,co}^a\} \log(1 - \theta_{co}^{(u)}(Y_{(b)})), \\ S_{u,nt}(Y_a, Y_{(b)}) = \mathbf{1}\{K_{u,nt}^a\} \log \chi_{nt} &+ \mathbf{1}\{Y_a \le Y_{(b)}, K_{u,nt}^a\} \log \theta_{nt}(Y_{(b)}) \\ &+ \mathbf{1}\{Y_a > Y_{(b)}, K_{u,nt}^a\} \log(1 - \theta_{nt}(Y_{(b)})), \\ S_{u,at}(Y_a, Y_{(b)}) = \mathbf{1}\{K_{u,at}^a\} \log \chi_{at} &+ \mathbf{1}\{Y_a \le Y_{(b)}, K_{u,at}^a\} \log \theta_{at}(Y_{(b)}) \\ &+ \mathbf{1}\{Y_a > Y_{(b)}, K_{u,at}^a\} \log(1 - \theta_{at}(Y_{(b)})). \end{aligned}$$

The recalling (4.6) we have

$$Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}) = \sum_{b\in I_\kappa} \sum_{u\in\{0,1\}} \sum_{s\in\{co,nt,at\}} Q_{u,s}(Y_{(b)}), \tag{E.3}$$

where, for $u \in \{0,1\}$, $Q_{u,s}(Y_{(b)}) := \mathbb{E}_{\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}} \left(\sum_{a=1}^n S_{u,s}(Y_a, Y_{(b)})|\mathcal{D}_n\right)$. To compute (E.3), we need to compute the following probabilities:

$$\begin{aligned} r_0^{(m)} &:= \mathbb{P}_{\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}}(S_a = co \mid Z_a = 0, D_a = 0, Y_a \le Y_{(b)}) \\ r_1^{(m)} &:= \mathbb{P}_{\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}}(S_a = co \mid Z_a = 1, D_a = 1, Y_a \le Y_{(b)}) \\ \rho_0^{(m)} &:= \mathbb{P}_{\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}}(S_a = co \mid Z_a = 0, D_a = 0, Y_a > Y_{(b)}) \\ \rho_1^{(m)} &:= \mathbb{P}_{\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}}(S_a = co \mid Z_a = 1, D_a = 1, Y_a > Y_{(b)}). \end{aligned}$$

LEMMA E.1. *Let $r_0^{(m)}, r_1^{(m)}, \rho_0^{(m)}, \rho_0^{(m)}$ be as defined above. Then, for $u \in \{0,1\}$,*

$$\begin{aligned} Q_{u,co}(Y_{(b)}) =& n_{uu} \Bigg\{ \left(\overline{F}_{uu}(Y_{(b)}) r_u^{(m)} + (1 - \overline{F}_{uu}(Y_{(b)}))\rho_u^{(m)}\right) \log(1 - \hat{\chi}_{nt,(m)} - \hat{\chi}_{at,(m)}) \\ &+ \overline{F}_{uu}(Y_{(b)}) r_u^{(m)} \log \theta_{co}^{(u)}(Y_{(b)}) + (1 - \overline{F}_{uu}(Y_{(b)}))\rho_u^{(m)} \log(1 - \theta_{co}^{(u)}(Y_{(b)})) \Bigg\}. \end{aligned}$$

Similarly,

$$Q_{0,nt}(Y_{(b)}) = n_{00} \Bigg\{ \left(\overline{F}_{00}(Y_{(b)})(1 - r_0^{(m)}) + (1 - \overline{F}_{00}(Y_{(b)})(1 - \rho_0^{(m)})\right) \log \hat{\chi}_{nt,(m)}$$

$$+ \overline{F}_{00}(Y_{(b)})(1 - r_0^{(m)}) \log \theta_{nt}(Y_{(b)}) + (1 - \overline{F}_{00}(Y_{(b)})(1 - \rho_0^{(m)}) \log(1 - \theta_{nt}(Y_{(b)}))\Bigg\},$$

and

$$Q_{1,at}(Y_{(b)}) = n_{11}\Bigg\{ \Big(\overline{F}_{11}(Y_{(b)})(1 - r_1^{(m)}) + (1 - \overline{F}_{11}(Y_{(b)})(1 - \rho_1^{(m)})\Big) \log \hat{\chi}_{at,(m)}$$
$$+ \overline{F}_{11}(Y_{(b)})(1 - r_1^{(m)}) \log \theta_{at}(Y_{(b)}) + (1 - \overline{F}_{11}(Y_{(b)})(1 - \rho_1^{(m)}) \log(1 - \theta_{at}(Y_{(b)}))\Bigg\}.$$

Finally, $Q_{1,nt}(Y_{(b)}) = n_{10} \log \chi_{nt} + n_{10} J(\overline{F}_{10}(Y_{(b)}), \theta_{nt}(Y_{(b)}))$ *and* $Q_{0,at}(Y_{(b)}) = n_{01} \log \chi_{at} + n_{01} J(\overline{F}_{01}(Y_{(b)}), \theta_{at}(Y_{(b)}))$*, where* $J(x, y) = x \log y + (1 - x) \log(1 - y)$*.*

The proof of the above lemma is an easy consequence of Lemma E.2 below. This completes the proof of the expectation step of the EM algorithm, at the $(m + 1)$-th iteration.

LEMMA E.2. *For every integer* $m \geq 1$*,*

$$r_0^{(m)} = \frac{\hat{\chi}_{co,(m)} \hat{\theta}^{(0)}_{co,(m)}(Y_{(b)})}{\hat{\chi}_{co,(m)} \hat{\theta}^{(0)}_{co,(m)}(Y_{(b)}) + \hat{\chi}_{nt,(m)} \hat{\theta}_{nt,(m)}(Y_{(b)})},$$
$$r_1^{(m)} = \frac{\hat{\chi}_{co,(m)} \hat{\theta}^{(1)}_{co,(m)}(Y_{(b)})}{\hat{\chi}_{co,(m)} \hat{\theta}^{(1)}_{co,(m)}(Y_{(b)}) + \hat{\chi}_{at,(m)} \hat{\theta}_{at,(m)}(Y_{(b)})},$$
$$\rho_0^{(m)} = \frac{\hat{\chi}_{co,(m)}(1 - \hat{\theta}^{(0)}_{co,(m)}(Y_{(b)}))}{\hat{\chi}_{co,(m)}(1 - \hat{\theta}^{(0)}_{co,(m)}(Y_{(b)})) + \hat{\chi}_{nt,(m)}(1 - \hat{\theta}_{nt,(m)}(Y_{(b)}))},$$
$$\rho_1^{(m)} = \frac{\hat{\chi}_{co,(m)}(1 - \hat{\theta}^{(1)}_{co,(m)}(Y_{(b)}))}{\hat{\chi}_{co,(m)}(1 - \hat{\theta}^{(1)}_{co,(m)}(Y_{(b)})) + \hat{\chi}_{at,(m)}(1 - \hat{\theta}_{at,(m)}(Y_{(b)}))}.$$

where $\hat{\chi}_{co,(m)} = 1 - \hat{\chi}_{at,(m)} - \hat{\chi}_{nt,(m)}$*.*

PROOF. Throughout the proof, we denote $\mathbb{P} = \mathbb{P}_{\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}}$ for notational simplicity.
To begin with, note that

$$\begin{aligned} r_0^{(m)} &= \mathbb{P}(S_a = co \mid Z_a = 0, D_a = 0, Y_a \leq Y_{(b)}) \\ &= \frac{\mathbb{P}(S_a = co, Z_a = 0, D_a = 0, Y_a \leq Y_{(b)})}{\mathbb{P}(S_a = co, Z_a = 0, D_a = 0, Y_a \leq Y_{(b)}) + \mathbb{P}(S_a = nt, Z_a = 0, D_a = 0, Y_a \leq Y_{(b)})}. \end{aligned} \tag{E.4}$$

Now,

$$\begin{aligned}
&\frac{\mathbb{P}(S_a = co, Z_a = 0, D_a = 0, Y_a \le Y_{(b)})}{\mathbb{P}(Z_a = 0, D_a = 0)} \\
&= \mathbb{P}(Y_a \le Y_{(b)} \mid S_a = co, Z_a = 0, D_a = 0) \cdot \mathbb{P}(S_a = co \mid Z_a = 0, D_a = 0) \\
&= \frac{\hat{\chi}_{co,(m)}}{\hat{\chi}_{co,(m)} + \hat{\chi}_{nt,(m)}} \mathbb{P}(Y_a \le Y_{(b)} \mid S_a = co, Z_a = 0) \\
&= \frac{\hat{\chi}_{co,(m)}}{\hat{\chi}_{co,(m)} + \hat{\chi}_{nt,(m)}} \hat{\theta}^{(0)}_{co,(m)}(Y_{(b)}). \qquad \text{(E.5)}
\end{aligned}$$

Moreover,

$$\begin{aligned}
&\frac{\mathbb{P}(S_a = co, Z_a = 0, D_a = 0, Y_a \le Y_{(b)}) + \mathbb{P}(S_a = nt, Z_a = 0, D_a = 0, Y_a \le Y_{(b)})}{\mathbb{P}(Z_a = 0, D_a = 0)} \\
&= \mathbb{P}(Y_a \le Y_{(b)} \mid S_a = co, Z_a = 0, D_a = 0) \cdot \mathbb{P}(S_a = co \mid Z_a = 0, D_a = 0) \\
&+ \mathbb{P}(Y_a \le Y_{(b)} \mid S_a = nt, Z_a = 0, D_a = 0) \cdot \mathbb{P}(S_a = nt \mid Z_a = 0, D_a = 0) \\
&= \frac{\hat{\chi}_{co,(m)}}{\hat{\chi}_{co,(m)} + \hat{\chi}_{nt,(m)}} \hat{\theta}^{(0)}_{co,(m)}(Y_{(b)}) + \frac{\hat{\chi}_{nt,(m)}}{\hat{\chi}_{co,(m)} + \hat{\chi}_{nt,(m)}} \hat{\theta}_{nt,(m)}(Y_{(b)}). \qquad \text{(E.6)}
\end{aligned}$$

Substituting (E.5) and (E.6) in (E.4) the identity for $r_0^{(m)}$ follows. The other identities can be proved similarly. □

E.2.2. *The Maximization Step.* Recall from (4.7), that

$$(\breve{\boldsymbol{\theta}}_{(m+1)}, \hat{\boldsymbol{\chi}}_{(m+1)}) = \arg \max_{\boldsymbol{\theta} \in \boldsymbol{\vartheta}, \boldsymbol{\chi} \in \mathbb{R}^2} Q_m(\boldsymbol{\theta}, \boldsymbol{\chi} | \hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}),$$

where $\breve{\boldsymbol{\theta}}_{(m+1)}(t) = (\breve{\theta}^{(0)}_{co,(m+1)}(t), \breve{\theta}_{nt,(m+1)}(t), \theta^{(1)}_{co,(m+1)}, \breve{\theta}_{at,(m+1)})'$ and $\hat{\boldsymbol{\chi}}_{(m+1)} = (\hat{\chi}_{nt,(m+1)}, \hat{\chi}_{at,(m+1)})'$.

Lemma E.3. *Let $r_0^{(m)}, r_1^{(m)}, \rho_0^{(m)}, \rho_0^{(m)}$ be as in Lemma E.2. Then*

$$\begin{aligned}
\breve{\theta}^{(0)}_{co,(m+1)}(Y_{(b)}) &= \frac{\overline{F}_{00}(Y_{(b)}) r_0^{(m)}}{\overline{F}_{00}(Y_{(b)}) r_0^{(m)} + (1 - \overline{F}_{00}(Y_{(b)})) \rho_0^{(m)}}, \\
\breve{\theta}_{nt,(m+1)}(Y_{(b)}) &= \frac{n_{00} \overline{F}_{00}(Y_{(b)})(1 - r_0^{(m)}) + n_{10} \overline{F}_{10}(Y_{(b)})}{n_{00} \overline{F}_{00}(Y_{(b)})(1 - r_0^{(m)}) + n_{00}(1 - \overline{F}_{00}(Y_{(b)}))(1 - \rho_0^{(m)}) + n_{10}}, \\
\breve{\theta}^{(1)}_{co,(m+1)}(Y_{(b)}) &= \frac{\overline{F}_{11}(Y_{(b)}) r_1^{(m)}}{\overline{F}_{11}(Y_{(b)}) r_1^{(m)} + (1 - \overline{F}_{11}(Y_{(b)})) \rho_1^{(m)}}, \\
\breve{\theta}_{at,(m+1)}(Y_{(b)}) &= \frac{n_{11} \overline{F}_{11}(Y_{(b)})(1 - r_1^{(m)}) + n_{01} \overline{F}_{01}(Y_{(b)})}{n_{11} \overline{F}_{11}(Y_{(b)})(1 - r_1^{(m)}) + n_{11}(1 - \overline{F}_{11}(Y_{(b)}))(1 - \rho_1^{(m)}) + n_{01}};
\end{aligned}$$

and

$$\hat{\chi}_{nt,(m+1)} = \frac{1}{|I_\kappa|n}\sum_{b\in I_\kappa}\left\{n_{00}\overline{F}_{00}(Y_{(b)})(1-r_0^{(m)}) + n_{00}(1-\overline{F}_{00}(Y_{(b)}))(1-\rho_0^{(m)}) + n_{10}\right\}$$
$$\hat{\chi}_{at,(m+1)} = \frac{1}{|I_\kappa|n}\sum_{b\in I_\kappa}\left\{n_{01} + n_{11}\overline{F}_{11}(Y_{(b)})(1-r_1^{(m)}) + n_{11}(1-\overline{F}_{11}(Y_{(b)}))(1-\rho_1^{(m)})\right\}.$$

Moreover, $\hat{\boldsymbol{\chi}}_{(m+1)} = (\hat{\chi}_{nt,(m+1)}, \hat{\chi}_{at,(m+1)})' \in [0,1]^2_+$*, that is,* $\hat{\chi}_{nt,(m+1)}, \hat{\chi}_{at,(m+1)} \in (0,1)$ *and* $0 \le \chi_{nt,(m+1)} + \chi_{at,(m+1)} \le 1$*.*

Proof. This follows from Lemma E.1, by solving the first-order conditions obtained by taking the gradient of the $Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)})$ with respect to $(\boldsymbol{\theta}(Y_{(b)}))_{b\in I_\kappa}$ and $\boldsymbol{\chi}$, and equating it to zero.

To see, $\hat{\boldsymbol{\chi}}_{(m+1)} \in [0,1]^2_+$, note that $\chi_{nt,(m+1)}$ and $\chi_{at,(m+1)}$ are obtained by maximizing with respect to a, b a function of the form $x\log(a) + y\log(b) + z\log(1-a-b)$, for some non-negative quantities x, y, z. Clearly, this is maximized when $a = x/(x+y+z)$, $b = y/(x+y+z)$, which satisfy the requited constraints: $a, b \in [0,1]$ and $0 \le a+b \le 1$. □

To ensure the monotonicity constraint we apply the PAVA algorithm with the following weights to the vector $\breve{\boldsymbol{\theta}}_{(m+1)}$, which is computed in the above lemma:

$$\begin{aligned}
w^{(0)}_{co,(m+1)}(Y_{(b)}) &= n_{00}\overline{F}_{00}(Y_{(b)})r_0^{(m)} + n_{00}(1-\overline{F}_{00}(Y_{(b)})\rho_0^{(m)},\\
w_{nt,(m+1)}(Y_{(b)}) &= n_{00}\overline{F}_{00}(Y_{(b)})(1-r_0^{(m)}) + n_{00}(1-\overline{F}_{00}(Y_{(b)})(1-\rho_0^{(m)}) + n_{10},\\
w_{at,(m+1)}(Y_{(b)}) &= n_{11}\overline{F}_{11}(Y_{(b)})(1-r_1^{(m)}) + n_{11}(1-\overline{F}_{11}(Y_{(b)})(1-\rho_1^{(m)}) + n_{01},\\
w^{(1)}_{co,(m+1)}(Y_{(b)}) &= n_{11}\overline{F}_{11}(Y_{(b)})r_1^{(m)} + n_{11}(1-\overline{F}_{11}(Y_{(b)})\rho_1^{(m)}.
\end{aligned} \tag{E.7}$$

This completes the description of the EM-PAVA algorithm. Proposition 4.1, which is proved below, shows that this procedure indeed maximizes $Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)})$ over the restricted parameter space $\boldsymbol{\vartheta}_+ \times [0,1]^2_+$.

***Proof of Proposition 4.1*:** A collection of $f_1, f_2, \dots, f_n : \mathbb{R} \to \mathbb{R}$ is said to be *nice* with respect to a given weight vector $\boldsymbol{w} = (w_1, w_2, \dots, w_n)'$ if the following hold:

– there exists $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \dots, \tilde{\theta}_n)'$ such that $\bar{\theta}_{ab} := \arg\max \sum_{s=a}^b f_s(\theta)$ can be represented as the weighted average of $(\tilde{\theta}_a, \dots, \tilde{\theta}_b)$, that is,

$$\bar{\theta}_{ab} = \frac{\sum_{s=a}^b w_s\tilde{\theta}_a}{\sum_{s=a}^b w_s} \quad \forall\ a \le b,$$

– $\sum_{s=a}^b f_s(\theta)$ is strictly increasing when $\theta \le \bar{\theta}_{ab}$ and is strictly decreasing when $\theta > \bar{\theta}_{ab}$.

We will use the following well-known result about maximizing the sum of nice functions under the monotonicity constraint.

LEMMA E.4. *Ma, Foster and Stine (2015) Let $f_1, f_2, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ be collection of functions, and $\breve{z}_a = \arg\max_{z\in\mathbb{R}} f_a(z)$. If this collection of functions is nice with respect to a given weight vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)'$, then*

$$\arg\max_{z_1\leq\ldots\leq z_n} \sum_{s=1}^{n} f_s(z_s) = \mathrm{PAVA}_{\boldsymbol{w}}(\breve{z}_1, \ldots, \breve{z}_n),$$

where the PAVA algorithm uses the weight vector $\boldsymbol{w}$.

Since $\hat{\boldsymbol{\chi}}_{(m+1)} \in [0,1]^2_+$ (Lemma E.3), it suffices to show that $\hat{\boldsymbol{\theta}}_{(m+1)} \in \boldsymbol{\vartheta}_+$ and it is the restricted maximum. Note that the estimates

$$\breve{\theta}^{(0)}_{co,(m+1)}(Y_{(b)}), \breve{\theta}_{nt,(m+1)}(Y_{(b)}), \breve{\theta}^{(1)}_{co,(m+1)}(Y_{(b)}), \breve{\theta}_{at,(m+1)}(Y_{(b)}) \in [0,1],$$

for each $b \in I_\kappa$. Therefore, the PAVA estimates $\hat{\boldsymbol{\theta}}_{(m+1)} \in \boldsymbol{\vartheta}_+$. Next, to apply Lemma E.4 above, define the following four functions $f_{1b}, f_{2b}, f_{3b}, f_{4b}$:

$$\begin{aligned}
f_{1b}(\theta_{1b}) &= n_{00}\overline{F}_{00}(Y_{(b)})r_0^{(m)} \log\theta_{1b} + n_{00}(1-\overline{F}_{00}(Y_{(b)})\rho_0^{(m)} \log(1-\theta_{1b})\\
f_{2b}(\theta_{2b}) &= \left\{n_{00}\overline{F}_{00}(Y_{(b)})(1-r_0^{(m)}) + n_{10}\overline{F}_{10}(Y_{(b)})\right\}\log\theta_{2b}\\
&\quad + \left\{n_{00}(1-\overline{F}_{00}(Y_{(b)})(1-\rho_0^{(m)}) + n_{10}(1-\overline{F}_{10}(Y_{(b)}))\right\}\log(1-\theta_{2b})\\
f_{3b}(\theta_{3b}) &= n_{11}\overline{F}_{11}(Y_{(b)})r_1^{(m)} \log\theta_{3b} + n_{11}(1-\overline{F}_{11}(Y_{(b)})\rho_1^{(m)} \log(1-\theta_{3b})\\
f_{4b}(\theta_{4b}) &= \left\{n_{11}\overline{F}_{11}(Y_{(b)})(1-r_1^{(m)}) + n_{01}\overline{F}_{01}(Y_{(b)})\right\}\log\theta_{4b}\\
&\quad + \left\{n_{11}(1-\overline{F}_{11}(Y_{(b)})(1-\rho_1^{(m)}) + n_{01}(1-\overline{F}_{01}(Y_{(b)}))\right\}\log(1-\theta_{4b}).
\end{aligned}$$

where $\theta_{1b} = \theta^{(0)}_{co}(Y_{(b)}), \theta_{2b} = \theta_{nt}(Y_{(b)}), \theta_{3b} = \theta^{(1)}_{co}(Y_{(b)})$ and $\theta_{4b} = \theta_{at}(Y_{(b)})$. Then, from (E.3) and Lemma E.1, it follows that

$$Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)}) = C(\boldsymbol{\chi}) + \sum_{s=1}^{4}\sum_{b\in I_\kappa} f_{sb}(\theta_{sb}),$$

where $C(\boldsymbol{\chi})$ is a function depending only of $\boldsymbol{\chi}$. Therefore, maximizing $Q_m(\boldsymbol{\theta}, \boldsymbol{\chi}|\hat{\boldsymbol{\theta}}_{(m)}, \hat{\boldsymbol{\chi}}_{(m)})$ is equivalent to maximizing $\sum_{b\in I_\kappa} f_{sb}(\theta_{sb})$ for each s. Now, for each s, it is easy to see that the functions $f_{sb}(\theta_{sb})$, for $b \in I_\kappa$, satisfy the condition in Lemma E.4 with weights as in (E.7), and, hence the proof of Proposition 4.1 follows.

APPENDIX F: CONFIDENCE BAND CONSTRUCTION

We adopt an approach described in Buja and Rolke (2005) to construct confidence bands for the estimated CDFs. For an estimated CDF $\hat{F}(t)$ on fixed and ordered locations $(t_1, \ldots, t_n)$, this procedure finds an upper bound $u(t)$ and a lower bound $\ell(t)$ such that $F(t) \in [\ell(t), u(t)]$ with probability $100(1-\alpha)\%$ simultaneously, for all $t \in T = \{t_1, t_2, \ldots, t_n\}$. Begin by resampling the data B times and obtain estimates of the CDFs $\hat{F}_b(t)$, for $1 \leq b \leq B$ and $t \in T$. Then the following steps are implemented:

(a) Allocate the values of $\{\hat{F}_b(t_i) : 1 \leq b \leq B, 1 \leq i \leq n\}$ in a $B \times n$ matrix, where the b-th row is $(\hat{F}_b(t_1), \ldots, \hat{F}_b(t_n))$.
(b) Destroy the relationships within rows, by sorting the columns. This yields, for a fixed location of $t \in T$, a set of order statistics that are estimates of marginal quantiles. After sorting, the b-th row will contain the estimates for the $b/(B+1)$-quantiles, for $1 \leq b \leq B$. Define, $q_{s=\frac{b}{B+1}}(t_i)$ to be the (b, i)-th element of this sorted matrix, and $\ell_s(t_i) = q_s(t_i)$ and $u_s(t_i) = q_{1-s}(t_i)$.
(c) For each $\hat{F}_b$, determine the minimal parameter value $s = s_b$ such that $\ell_s(t_i) \leq \hat{F}_b(t_i) \leq u_s(t_i)$, simultaneously, for all $1 \leq i \leq n$. The bisection algorithm can find the minimal s efficiently.
(d) For the collection of parameter values $(s_b)_{b \in B}$ determine the upper $1-\alpha$ quantile. This will be the estimate $\hat{s}_\alpha$ for a band with coverage probability minimally $\geq 1-\alpha$: $[\ell(t), u(t)] = [\ell_{\hat{s}_\alpha}(t), u_{\hat{s}_\alpha}(t)]$, for $t \in T$.

K. Lee
University of Pennsylvania
433 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, Pennsylvania 19104
USA
E-mail: kwonlee@wharton.upenn.edu

B. B. Bhattacharya
University of Pennsylvania
466 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, Pennsylvania 19104
USA
E-mail: bhaswar@wharton.upenn.edu

J. Qin
Biostatistics Research Branch
National Institute of Allergy and Infectious Diseases
6700B Rockledge Drive 7609, Room 5131
Bethesda, Maryland 20892
USA
E-mail: jingqin@niaid.nih.gov

D. S. Small
University of Pennsylvania
464 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, Pennsylvania 19104
USA
E-mail: dsmall@wharton.upenn.edu