3-2009

Efficient Nonparametric Estimation of Causal Effects in Randomized Trials With Noncompliance

Jing Cheng

Dylan Small
University of Pennsylvania

Zhiqiang Tan

Thomas R. Ten Have
University of Pennsylvania

Recommended Citation

Cheng, J., Small, D., Tan, Z., & Ten Have, T. R. (2009). Efficient Nonparametric Estimation of Causal Effects in Randomized Trials With Noncompliance. *Biometrika, 96* (1), 19-36. http://dx.doi.org/10.1093/biomet/asn056

Efficient Nonparametric Estimation of Causal Effects in Randomized Trials With Noncompliance

Abstract

Causal approaches based on the potential outcome framework provide a useful tool for addressing noncompliance problems in randomized trials. We propose a new estimator of causal treatment effects in randomized clinical trials with noncompliance. We use the empirical likelihood approach to construct a profile random sieve likelihood and take into account the mixture structure in outcome distributions, so that our estimator is robust to parametric distribution assumptions and provides substantial finite-sample efficiency gains over the standard instrumental variable estimator. Our estimator is asymptotically equivalent to the standard instrumental variable estimator, and it can be applied to outcome variables with a continuous, ordinal or binary scale. We apply our method to data from a randomized trial of an intervention to improve the treatment of depression among depressed elderly patients in primary care practices.

Keywords

causal effect, efficient nonparametric estimation, empirical likelihood, instrumental variable, noncompliance, randomized trial

Disciplines

Biostatistics | Statistics and Probability

Efficient Nonparametric Estimation of Causal Effects in Randomized Trials with Noncompliance

BY Jing Cheng
Division of Biostatistics, University of Florida College of Medicine, Gainesville, Florida 32610, U.S.A.
jcheng@biostat.ufl.edu

Dylan S. Small
Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.
dsmall@wharton.upenn.edu

Zhiqiang Tan
Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, U.S.A
ztan@stat.rutgers.edu

AND Thomas R. Ten Have
Division of Biostatistics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, U.S.A.
ttenhave@upenn.edu

SUMMARY

Causal approaches based on the potential outcome framework provide a useful tool for addressing noncompliance problems in randomized trials. We propose a new estimator of causal treatment effects in randomized clinical trials with noncompliance. We use the empirical likelihood approach to construct a profile random sieve likelihood and take into account the mixture structure in outcome distributions, so that our estimator is robust to parametric distribution assumptions and provides substantial finite-sample efficiency gains over the standard instrumental variable estimator. Our estimator is asymptotically equivalent to the standard instrumental variable estimator, and it can be applied to outcome variables with a continuous, ordinal or binary scale. We apply our method to data from a randomized trial of an intervention to improve the treatment of depression among depressed elderly patients in primary care practices.

Some key words: Causal effect; Efficient nonparametric estimation; Empirical likelihood; Noncompliance; Randomized trials.

1. Introduction

When there is noncompliance in randomized trials, there is often interest in estimating the causal effect of actually receiving the treatment compared to receiving the control. Knowledge of this effect is useful for predicting the impact of the treatment in a setting for which compliance patterns might differ from the randomized trial and for scientific understanding of the treatment (Sommer & Zeger, 1991; Sheiner & Rubin, 1995; Small et al., 2006; Cheng & Small, 2006).

Note that intention-to-treat analysis is not suitable for estimating the causal effect of actually receiving the treatment when there is noncompliance because it estimates the effect of assignment to the treatment group. An as-treated analysis seeks to estimate the causal effect of receiving the treatment but is biased if compliers are not comparable to noncompliers. Imbens & Angrist (1994) and Angrist et al. (1996) show that the causal effect of actually receiving the treatment for the subgroup of subjects who would receive the treatment if assigned to the treatment group and would receive the control if assigned to the

control group, called the complier average causal effect or the local average treatment effect in the econometrics literature (Imbens & Angrist, 1994), is nonparametrically identified under certain, often plausible, assumptions that do not require compliers and noncompliers to be comparable. These assumptions, henceforth referred to as the instrumental variable assumptions, are discussed in §2. The complier average causal effect can be consistently estimated under the instrumental variable assumptions by the standard two-stage least-squares instrumental variables estimator. Imbens & Rubin (1997a, b) demonstrate that, under the assumptions, the standard instrumental variable estimator is an inefficient estimator of the complier average causal effect because it does not make full use of the mixture structure of the outcome distributions of the four observed groups defined by the cross classification of the randomization and treatment received; see §2.4 for further discussion. Imbens & Rubin (1997b) present three new alternatives to the standard IV estimator. One is based on a normal approximation and two are based on multinomial approximations to the outcome distributions in the four groups. In a simulation study with normally distributed outcomes, Imbens & Rubin (1997b) show that all three alternative estimators are more efficient than the standard IV estimator. However, the estimator that is based on a normal approximation to the outcome distributions can have substantial bias when the outcomes are not normal; this is demonstrated in §4. The estimators based on multinomial approximations to the outcome distributions are in principle nonparametric. However, a systematic approach for choosing the multinomial approximations is needed.

Multinomial approximations to the outcome distributions are a type of sieve. A sieve is a sequence of approximations $\{F_n\}$ to a space F of distributions such that $F_n \to F$ as $n \to \infty$ (Grenander, 1981). Maximizing the likelihood over a sieve rather than the whole parameter space often leads to desirable statistical properties, especially when the underlying parameter space is large (Shen & Wong, 1994). However, the construction of sieves is not an easy task. One approach to constructing sieves is to use a random approximation $\hat{F}_n$ that depends on the data, a random sieve. The empirical likelihood approach (Owen, 1991) is based on an easily constructed random sieve (Shen et al., 1999). In this paper, we use the empirical

likelihood approach to construct an efficient estimator for the complier average causal effect.

2. Notation, Assumptions and Review of Established Estimators

2.1 Notation

We consider a two-arm randomized trial with N subjects, n_0 of whom are randomly assigned to the control group. We use letters with and without star to denote vectors and scalars respectively. Let R_* be the N-dimensional vector of randomization assignments for all subjects, with individual element $R_i = r \in \{0, 1\}$ according to whether subject i is assigned active treatment, $R_i = 1$, or control, $R_i = 0$. We let A_*^{r*} be the N-dimensional vector of potential treatment receiveds under the vector of randomization assignment r_* with individual element $A_i^{r*} = a \in \{0, 1\}$ according to whether subject i would take the control or treatment under randomization assignment r_*. We let $Y_*^{r*,a*}$ be the vector of potential responses under randomization assignment r_* and treatment receiveds a_*, with individual element $Y_i^{r*,a*}$ being the potential response for subject i with the vectors of randomization assignments r_* and treatment receiveds a_*. The sets of $\{Y_i^{r*,a*} | r_* \in \{0, 1\}^N, a_* \in \{0, 1\}^N\}$ and $\{A_i^{r*} | r_* \in \{0, 1\}^N\}$ are 'potential' responses and treatment receiveds in the sense that we can only observe one member of each set. The observed outcome and treatment received variables for subject i are $Y_i^{R*,A_*^{R*}} \equiv Y_i$ and $A_i^{R*} \equiv A_i$ respectively.

2.2. Assumptions

We make similar assumptions to those in Angrist et al. (1996).

Assumption 1: *Stable unit treatment value assumption* (Rubin, 1980). (i). If $r = r'$, then $A_i^{r*} = A_i^{r'*}$ for subject i. (ii). If $r = r'$ and $a = a'$, then $Y_i^{r*,a*} = Y_i^{r'*,a'*}$ for subject i. This assumption allows us to write $Y_i^{r*,a*}, A_i^{r*}$ as $Y_i^{r,a}, A_i^{r}$.

Assumption 2: *Random assignment.* This assumption implies independence between assignment and pretreatment variables including potential outcomes and treatment receiveds.

Assumption 3: *Random sampling.* We assume that the N subjects in the trial are independent and identically distributed draws from a superpopulation; that is, $Y_i^{r,a}$ and A_i^{r}, $i = 1, \ldots, N$, are independent and identically distributed with the same distribution as the random vector consisting of $Y^{r,a}$ and A^r.

Assumption 4: *Mean exclusion restriction.* We assume that $E(Y^{r,a}) = E(Y^{r',a})$ for all r, r', a; that is, the randomization assignment affects the mean of the observed outcome only through its effect on treatment received. Note that the mean exclusion restriction is weaker than the unit level exclusion restriction of Angrist et al. (1996), who assume that $Y_i^{r,a} = Y_i^{r',a}$ for all r, r', a. However, we think that in most applications in which the weaker mean exclusion restriction is plausible, the stronger unit-level exclusion restriction is also plausible and so we primarily use the weaker mean exclusion restriction because it is easier to work with this assumption.

Assumption 5: *Nonzero average causal effect of R on A.*

Assumption 6: *Monotonicity.* We assume that $\text{pr}(A^1 \geq A^0) = 1$. This assumption says that there is no one who would receive the opposite treatment of his or her assignment under both assignment to treatment and to control.

2.3 Compliance classes

A subject in a two-arm trial can be classified into one of four compliance classes:

$$C_i = \begin{cases} 0 \text{ (never-taker)} & \text{if } (A_i^0, A_i^1) = (0,0) \\ 1 \text{ (complier)} & \text{if } (A_i^0, A_i^1) = (0,1) \\ 2 \text{ (always-taker)} & \text{if } (A_i^0, A_i^1) = (1,1) \\ 3 \text{ (defier)} & \text{if } (A_i^0, A_i^1) = (1,0). \end{cases}$$

In practice, we can observe only one of A_i^0 and A_i^1, so that a subject's compliance status is not observed directly in a trial, but it can be partially identified based on treatment assignment and observed treatment-received; see Table 1. Note that the monotonicity assumption rules out the existence of defiers. For single consent design trials (Zelen, 1979), which have the property that the control group cannot access the treatment, that is, $\text{pr}(A^0 = 0) = 1$, the presence of always-takers and defiers is ruled out.

2.4 Major established estimators

Under Assumptions $1 - 6$, the compliers are the only subgroup for which a randomized trial provides information about the causal effect of receiving treatment (Angrist et al., 1996).

For always-takers and never-takers, assignment to treatment has no effect on treatment received. The complier average causal effect, $E(Y^1 - Y^0|C = 1)$, can be thought of as the causal effect of receiving treatment for the subpopulation of compliers because, for compliers, assignment of treatment agrees with receipt of treatment. Angrist et al. (1996) show that, under Assumptions 1-6, the complier average causal effect is

$$\text{CACE} = \frac{E(Y|R = 1) - E(Y|R = 0)}{E(A|R = 1) - E(A|R = 0)}, \tag{1}$$

which is the intention-to-treat effect divided by the proportion of compliers. The standard instrumental variable estimator is the sample analogue of (1),

$$\hat{\text{CACE}}_{\text{S}} = \frac{\hat{E}(Y|R = 1) - \hat{E}(Y|R = 0)}{\hat{E}(A|R = 1) - \hat{E}(A|R = 0)}, \tag{2}$$

where the $\hat{E}$'s denote sample means; (2) is sometimes called the Wald estimator.

The standard instrumental variable estimator does not take full advantage of the mixture structure of the outcomes of the four observed groups in Table 1, as we will discuss in §3.1. Imbens & Rubin (1997a,b) present two approaches of using mixture modeling to estimate the complier average causal effect. One approach assumes a parametric distribution, such as normal, for the outcomes for each compliance class group under each randomization assignment. The complier average causal effect is then estimated by maximum likelihood for this model using the EM algorithm. This estimator provides considerable efficiency gains over the standard instrumental variable estimator when the parametric assumptions hold; see Table 4. However, when the parametric assumptions are wrong, this estimator can be inconsistent whereas the standard instrumental variable estimator is consistent; see Table 4 for finite-sample results.

Imbens and Rubin's other approach to using mixture modeling to estimate the complier average causal effect is to approximate the density of the outcome distribution for each compliance class under each randomization group as a piecewise constant function, and then estimate the complier average causal effect by maximum likelihood. This approach is

in principle nonparametric as the number of constant pieces in each density function can be increased with the sample size. However, Imbens & Rubin (1997b) do not provide a systematic approach for choosing the number of and locations of the pieces. We develop a systematic easily implementable approach for doing this using empirical likelihood in the next section.

3. Estimation through Empirical Likelihood Approach

3.1 Motivation and description of empirical likelihood approach

We first motivate and describe our method for single consent design trials, where the presence of always-takers and defiers is ruled out. Table 2 shows the relationship between observed (R, A) groups and latent compliance classes for a single consent design trial. The complier average causal effect can be re-expressed under Assumptions 1-6 as follows:

$$\text{CACE} = \mu^{c1} - \mu^{c0} = \mu^{c1} - \frac{\mu^{R=0} - (1-\pi_c)\mu^n}{\pi_c} = E(Y|R=1, A=1) - \frac{E(Y|R=0) - \{1 - \text{pr}(A=1|R=1)\}E(Y|R=1, A=0)}{\text{pr}(A=1|R=1)} \quad (3)$$

where μ^{c1}, μ^{c0}, μ^n and $\mu^{R=0}$ denote the mean potential outcomes of the compliers under treatment, compliers under control, never takers and the whole population of subjects when assigned to the control respectively; and π_c denotes the proportion of compliers. The standard instrumental variable estimator estimates the complier average causal effect by substituting the method of moments estimates from the sample for $E(Y|R=1, A=1)$, $E(Y|R=0)$, $\text{pr}(A=1|R=1)$ and $E(Y|R=1, A=0)$ into (3). However, as noted by Imbens & Rubin (1997b), there are restrictions on the joint density of (Y, R, A) that are not taken into account by the method of moments that can be useful for estimating $E(Y|R=0)$, $\text{pr}(A=1|R=1)$ and $E(Y|R=1, A=0)$. To be specific, Assumptions 1-6 imply the following restrictions.

Restriction 1. The distribution of $Y|R=0$ is a mixture of the outcome distribution of the never-takers under $R=0$ and the outcome distribution of the compliers under $R=0$.

Restriction 2. The mixing proportion π_c for $Y|R=0$ equals $\text{pr}(A=1|R=1)$ as a consequence of Assumption 2.

Restriction 3. The mean of the never-takers under $R = 0$ is equal to the mean of the never-takers under $R = 1$, which equals $E(Y|R = 1, A = 0)$, as a consequence of Assumption 4.

The sample mean of $Y|R = 0$ uses only the information in those of $Y_1, \ldots, Y_N$ for which $R_i = 0$ to estimate $E(Y|R = 0)$, but Restrictions 1-3 imply that there is additional information in those of $Y_1, \ldots, Y_N$ for which $R_i = 1$. Similarly, the sample proportion of $A = 1|R = 1$ uses only the information in those of $A_1, \ldots, A_N$ for which $R_i = 1$ to estimate $\text{pr}(A = 1|R = 1)$ but Restrictions 1-3 imply that there is additional information in those of $A_1, \ldots, A_N$ for which $R_i = 0$. A body of work has shown that supplementing a sample from a distribution that is a mixture of two components with samples from one or both of the components alone provides additional information for estimating aspects of the mixture distribution; see for example Hall & Titterington (1984), Lancaster & Imbens (1996) and Qin (1999). Here, the sample of $Y_1, \ldots, Y_N$ for which $R_i = 1, A_i = 0$ provides information about the never-taker component of the mixture $Y|R = 0$ and the sample of $A_1, \ldots, A_N$ for which $R_i = 1$ provides information about the mixing proportion in the mixture $Y|R = 0$. We now illustrate how this information is useful in a setting with a binary outcome in which

$$\pi_c = 0.5, \mu^n = 0.2, \mu^{c1} = 0.8, \mu^{c0} = 0.9, N = 40, n_0 = 20. \tag{4}$$

The following is a plausible sample in this setting: $\#(Y_i = 1, A_i = 1, R_i = 1) = 8$, $\#(Y_i = 0, A_i = 1, R_i = 1) = 2$, $\#(Y_i = 1, A_i = 0, R_i = 1) = 2$, $\#(Y_i = 0, A_i = 0, R_i = 1) = 8$, $\#(Y_i = 1, A_i = 0, R_i = 0) = 13$ and $\#(Y_i = 0, A_i = 0, R_i = 0) = 7$; the p-value for a χ^2 test of whether or not this sample comes from the distribution (4) is 0.37. Note that, for this sample, the method of moments estimates of the quantities in (3), namely $\hat{E}(Y|R = 1, A = 1) = 0.8$, $\hat{E}(Y|R = 0) = 0.65$, $\hat{\text{pr}}(A = 1|R = 1) = 0.5$, $\hat{E}(Y|R = 1, A = 0) = 0.2$, violate Restrictions 1-3. Figure 1 plots the profile log-likelihood for this sample under the probability model given by Assumptions 1-6 with binary outcomes. The maximum likelihood estimator of CACE, which takes into account the mixture structure of the outcomes given by Restrictions

1-3, has a noticeably higher likelihood than the standard instrumental variable estimator, which ignores some of the restrictions. The maximum likelihood estimator's property of taking into full account the mixture structure leads to substantially better estimates; in 1000 simulations from model (4), the mean squared error of the maximum likelihood estimator was 0.048 compared to 0.156 for standard instrumental variable estimator.

To take account of the mixture structure of the outcomes given by Restrictions 1-3 for more general distributions of outcomes in a nonparametric way, we use the empirical likelihood approach. The empirical likelihood for a parameter such as the complier average causal effect is the nonparametric profile likelihood for the parameter. Maximum empirical likelihood estimators have good properties for a wide class of semiparametric problems; see Owen (2001) and Qin & Lawless (1994) for discussion.

Without loss of generality, we arrange the subjects so that $R_1 = \ldots = R_{n_0} = 0$ and $R_{n_0+1} = \ldots = R_N = 1$; thus, $(Y_1, A_1), \ldots, (Y_{n_0}, A_{n_0})$ is a random sample from the population of $Y^{r=0,a=A^{r=0}}, A^{r=0}$ and $(Y_{n_0+1}, A_{n_0+1}), \ldots, (Y_N, A_N)$ is a random sample from the population of $Y^{r=1,a=A^{r=1}}, A^{r=1}$. The empirical likelihood L_E of the parameters $(\pi_c, \mu^n, \mu^{c1}, \mu^{c0})$ is:

$$L_E(\pi_c, \mu^n, \mu^{c1}, \mu^{c0}) = \max \left(\prod_{i=1}^{n_0} q_i \right) \left(\prod_{i=n_0+1}^{N} q_i \right), \tag{5}$$

subject to

$$\sum_{i=1}^{n_0} q_i = 1, \sum_{i=n_0+1}^{N} q_i = 1, q_i \geq 0, i = 1, \ldots N, \tag{6}$$

$$\sum_{i=n_0+1}^{N} q_i A_i = \pi_c, \sum_{i=n_0+1}^{N} q_i Y_i A_i = \mu^{c1}\pi_c, \sum_{i=n_0+1}^{N} q_i Y_i (1 - A_i) = \mu^n(1 - \pi_c), \tag{7}$$

There exist p_i^{c0}, p_i^n, $i = 1, \ldots, n_0$ such that

$$\pi_c p_i^{c0} + (1 - \pi_c) p_i^n = q_i, \tag{8}$$

$$\sum_{i=1}^{n_0} p_i^{c0} = \sum_{i=1}^{n_0} p_i^n = 1, p_i^{c0}, p_i^n \geq 0, i = 1, \ldots, n_0, \tag{9}$$

$$\sum_{i=1}^{n_0} p_i^n (Y_i - \mu^n) = 0, \tag{10}$$

$$\sum_{i=1}^{n_0} p_i^{c0} (Y_i - \mu^{c0}) = 0. \tag{11}$$

Note that throughout our paper, we will follow Owen (2001, Ch. 2.3) and regard tied data values Y_i, Y_j as representing distinct outcomes in the empirical likelihood as this simplifies calculations and does not affect inferences. The p_i^{c0} and p_i^n in (8)-(11) represent the population probabilities that a complier assigned to the control and a never-taker assigned to the control have the same outcome as subject i respectively. The conditions (8)-(11) involving the p_i^{c0} and p_i^n encode the restrictions on the distribution of $Y|R=0$ that come from it being a mixture of the compliers and never-takers under Assumptions 1-6, see Restrictions 1-3. The maximum empirical likelihood estimate of $(\pi_c, \mu^n, \mu^{c1}, \mu^{c0})$ is $\arg\max_{\pi_c,\mu^n,\mu^{c1},\mu^{c0}} L_E(\pi_c, \mu^n, \mu^{c1}, \mu^{c0})$. To ease the computational burden of computing the maximum empirical likelihood estimate, we do not maximize over μ^n, but instead use the method of moments estimator $\hat{\mu}^n = \sum_{i=1}^N Y_i R_i(1-A_i) / \sum_{i=1}^N R_i(1-A_i)$ and maximize $L_E(\pi_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0})$ over $(\pi_c, \mu^{c1}, \mu^{c0})$. In model (4), this approximate maximum empirical likelihood estimator of the complier average causal effect performed almost as well as the maximum empirical likelihood estimator; its mean squared error was 0.051 compared to 0.048 for the maximum empirical likelihood estimator. We now present an algorithm for finding the approximate maximum empirical likelihood estimate.

3.2 Computation for empirical likelihood approach

To find the approximate maximum empirical likelihood estimate, we conduct a grid search over π_c, finding $\max_{\mu^{c1},\mu^{c0}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0})$ over a grid of $\tilde{\pi}_c$ from 0 to 1. As we will see below, $\arg\max_{\mu^{c1}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0})$ does not depend on μ^{c0} and $\arg\max_{\mu^{c0}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0})$ does not depend on μ^{c1}, so finding the maximizing μ^{c1} and μ^{c0} can be done separately. For finding the maximizing μ^{c1}, we note that $\arg\max_{\mu^{c1}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0})$ equals $\arg\max_{\mu^{c1}} \prod_{i:R_i=A_i=1} q_i$ subject to (i) $\sum_{i:R_i=A_i=1} q_i = \tilde{\pi}_c$, (ii) $q_i \geq 0, i = 1, \ldots, N$ and (iii) $\sum_{i:R_i=A_i=1} q_i Y_i = \tilde{\pi}_c \mu^{c1}$. By multiplying the q_i's by $1/\tilde{\pi}_c$, we see that finding $\arg\max_{\mu^{c1}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0})$ is equivalent to finding the maximum empirical likelihood estimator of the mean of the population of $Y^{1,1}|C=1$ based on the random sample $Y_1, \ldots, Y_n | A_i = 1, R_i = 1$; consequently, $\arg\max_{\mu^{c1}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0})$ is the mean of $Y_1, \ldots, Y_n | A_i = 1, R_i = 1$; see Theorem 2.1 of Owen (2001). Thus, our estimate of μ^{c1} is $\hat{\mu}^{c1} = \left(\sum_{i=1}^N R_i A_i\right)^{-1} \sum_{i=1}^N Y_i R_i A_i$. For finding

our estimate of μ^{c0}, let $(q_1^*, \ldots, q_{n_0}^*) = \arg\max_{q_1,\ldots,q_{n_0}} \prod_{i=1}^{n_0} q_i$ subject to (6) and (8)-(10) with $\mu^n = \hat{\mu}^n$, $\pi_c = \tilde{\pi}_c$. We have that

$$\arg\max_{\mu^{c0}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0}) = \frac{\sum_{i=1}^{n_0} q_i^* Y_i - (1-\tilde{\pi}_c)\hat{\mu}^n}{\tilde{\pi}_c}, \tag{12}$$

where we use the fact that, for the μ^{c0} that satisfies $\sum_{i=1}^{n_0} q_i^* Y_i = \tilde{\pi}_c \mu^{c0} + (1-\tilde{\pi}_c)\hat{\mu}^n$, the constraints (8)-(11) are satisfied for $q_1 = q_1^*, \ldots, q_{n_0} = q_{n_0}^*$. Thus, to find $\arg\max_{\mu_{c0}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0})$, we just need to find $q_1^*, \ldots, q_{n_0}^*$. To do this, we note that we can view $(q_1^*, \ldots, q_{n_0}^*)$ as the maximum likelihood estimate of the category probabilities for the sample $Y_1, \ldots, Y_{n_0}$ from an independent and identically distributed multinomial model with categories $Y_1, \ldots, Y_{n_0}$, corresponding category probabilities $q_1, \ldots, q_{n_0}$ and parameter restrictions given by (6) and (8)-(10) with $\mu^n = \hat{\mu}^n, \pi_c = \tilde{\pi}_c$. Finding the maximum likelihood estimate directly is challenging because of the complex parameter restrictions in (8)-(10). However, consider using the EM algorithm, where we regard each subject's compliance class as 'missing data.' We can reexpress the observed data likelihood $\prod_{i=1}^{n_0} q_i$ and the parameter restrictions (6) and (8)-(10) in terms of p_i^{c0} and p_i^n; see Appendix 1 for details. We can then use the EM algorithm to find the p_i^{c0} and p_i^n to maximize the observed data likelihood and then find the corresponding maximizing q_i's by (8). The complete-data likelihood is $\prod_{i:R_i=0,C_i=1} p_i^{c0} \prod_{i:R_i=0,C_i=0} p_i^n$. Since the complete data follows an exponential family distribution, the E-step has a closed form expression. The M-step involves a calculation analogous to finding the empirical likelihood for the mean (Owen, 1988); convex duality enables us to avoid maximizing over $p_i^{c0}, p_i^n, i = 1, \ldots, n_0$, and instead we maximize over a single variable. The tractability of both the E- and M- steps makes the EM algorithm with each subject's compliance class as missing data easy to use for finding $q_1^*, \ldots, q_{n_0}^*$ and hence finding $\arg\max_{\mu^{c0}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \mu^{c1}, \mu^{c0})$ by (12).

Note that, given $q_i, i = 1, \ldots, n_0$, there are typically more than one set of $p_i^{c0}, p_i^n, i = 1, \ldots, n_0$, that satisfy the constraints (8)-(10). Numerical experiments, not shown here, verify that although the EM algorithm converges to different values of p_i^{c0} and p_i^n for different sets of starting values for the p_i^{c0} and p_i^n, the corresponding q_i's to which the EM algorithm

converges are the same, as Lemma 1 shows more formally.

Lemma 1. *Regardless of the starting values for the* $p_i^{c0}, p_i^n, i = 1, \ldots, n_0$*, the sequence of estimates of* q_i *from the EM algorithm converges to the global maximum of the likelihood* $\prod_{i=1}^{n_0} q_i$ *subject to the restrictions* (6) and (8)-(10) *with* $\mu^n = \hat{\mu}^n$*,* $\pi_c = \tilde{\pi}_c$*.*

The proof of Lemma 1 is outlined in Appendix 2.

In summary, we estimate $\pi_c, \mu^n, \mu^{c1}, \mu^{c0}$ as follows; a program is available from the authors.

Step 1. We obtain $\hat{\mu}^n$ as the sample mean of $Y|R = 1, A = 0$.

Step 2. We obtain $\hat{\mu}^{c1}$ as the sample mean of $Y|R = 1, A = 1$.

Step 3. For a grid of $\tilde{\pi}_c$, we find the maximum empirical likelihood estimate of μ^{c0} given $\pi_c = \tilde{\pi}_c$, $\mu^n = \hat{\mu}^n$, $\mu^{c1} = \hat{\mu}^{c1}$ using the EM algorithm described above. Then $\hat{\pi}_c = \arg\max_{\tilde{\pi}_c} \max_{\mu^{c0}} L_E(\tilde{\pi}_c, \hat{\mu}^n, \hat{\mu}^{c1}, \mu^{c0})$ and $\hat{\mu}^{c0} = \arg\max_{\mu^{c0}} L_E(\hat{\pi}_c, \hat{\mu}^n, \hat{\mu}^{c1}, \mu^{c0})$.

Step 4. Our approximate maximum empirical likelihood estimate of the complier average causal effect is $\hat{\text{CACE}}_{\text{A}} = \hat{\mu}^{c1} - \hat{\mu}^{c0}$.

3.3 Estimation in trials in which the assigned to control group can access the treatment

Our method illustrated in §3.1 can be directly applied to more general trials under Assumptions 1-6 in which the control group can access the treatment. For such trials, we have one more compliance class, the always-takers, in addition to the compliers and never-takers, see Table 3; we denote the proportion of always takers and the mean of always takers' potential outcomes by π_a and μ^a respectively. The empirical likelihood L_E of the parameters $(\pi_c, \pi_a, \mu^n, \mu^a, \mu^{c1}, \mu^{c0})$ is the maximum likelihood for multinomial distributions $(q_1, \ldots, q_{n_0})$ on $(Y_1, A_1), \ldots (Y_{n_0}, A_{n_0})$ and $(q_{n_0+1}, \ldots, q_N)$ on $(Y_{n_0+1}, A_{n_0+1}), \ldots, (Y_N, A_N)$ that are consistent with $(\pi_c, \pi_a, \mu^n, \mu^a, \mu^{c1}, \mu^{c0})$ and the restrictions on the parameter space specified by Assumptions 1-6, namely $L_E(\pi_c, \pi_a, \mu^n, \mu^a, \mu^{c1}, \mu^{c0}) = \max\left(\prod_{i=1}^{n_0} q_i\right)\left(\prod_{i=n_0+1}^{N} q_i\right)$ subject to (i) $\sum_{i=1}^{n_0} q_i = 1, \sum_{i=n_0+1}^{N} q_i = 1$; (ii) $q_i \geq 0, i = 1, \ldots, N$; (iii) $\sum_{i:R_i=1,A_i=0} q_i = 1 - \pi_a - \pi_c$; (iv) $\sum_{i:R_i=0,A_i=1} q_i = \pi_a$; (v) $\sum_{i:R_i=1,A_i=0} q_i Y_i = \mu^n(1-\pi_a-\pi_c)$; (vi) $\sum_{i:R_i=0,A_i=1} q_i Y_i = \mu^a \pi_a$; (vii) There exist p_i^{c0}, p_i^n for the i with $R_i = 0, A_i = 0$ such that (viia) $\{\pi_c/(1-\pi_a)\}p_i^{c0} + \{(1-$

$\pi_a - \pi_c)/(1-\pi_a)\}p_i^n = q_i$, (viib) $\sum p_i^{c0} = \sum p_i^n = 1$, (viic) $p_i^{c0}, p_i^n \geq 0$, (viid) $\sum p_i^n(y_i - \mu^n) = 0$ and (viie) $\sum p_i^{c0}(Y_i - \mu^{c0}) = 0$; and (viii) There exist p_i^{c1}, p_i^a for the i with $R_i = 1, A_i = 1$ such that (viiia) $\{\pi_c/(\pi_c + \pi_a)\}p_i^{c1} + \{\pi_a/(\pi_c + \pi_a)\}p_i^a = q_i$, (viiib) $\sum p_i^{c1} = \sum p_i^a = 1$; (viiic) $p_i^{c1}, p_i^a \geq 0$; (viiid) $\sum p_i^a(Y_i - \mu^a) = 0$ and (viiie) $\sum p_i^{c1}(Y_i - \mu^{c1}) = 0$. As with the single consent design, rather than finding the maximum empirical likelihood estimate of $(\pi_c, \pi_a, \mu^n, \mu^a, \mu^{c1}, \mu^{c0})$, we find the approximate maximum empirical likelihood estimate by setting μ^n equal to the sample mean of $Y|R = 1, A = 0$, corresponding to the known never-takers in the sample, and μ^a equal to the sample mean of $Y|R = 0, A = 1$, corresponding to the known always-takers in the sample, and then maximizing the empirical likelihood over $(\pi_c, \pi_a, \mu^{c1}, \mu^{c0})$. This can be done by using the EM algorithm for estimating μ^{c0} in the $Y|R = 0, A = 0$ sample as in §3.2, and an analogous EM algorithm for estimating μ^{c1} in the $Y|R = 1, A = 1$ sample. The details are provided in a technical report available from the authors.

4 Simulation Studies

We compare our approximate maximum empirical likelihood estimator with the standard instrumental variable estimator and Imbens and Rubin's parametric estimator, considering single consent design trials as discussed in §3.1. We set $\pi_c = 0.5$ and compare the three estimators under different outcome distributions and under sample sizes of $N = 100$ and $N = 500$ with $\text{pr}(R = 1) = 0.5$. The outcome distributions we consider are Normal, gamma, and lognormal distributions. For each outcome distribution, we set $\mu^{c1} = 2$, $\mu^{c0} = 1$, so that the CACE $= \mu^{c1} - \mu^{c0} = 1$. The variances are fixed at 1.

Before explaining our settings for μ^n, we discuss the impact of the distance between μ^n and μ^{c0} on the efficiency of the approximate maximum empirical likelihood estimator relative to standard instrumental variable estimator. The distance between μ^n and μ^{c0} is a measure of the separation between the distributions of the compliers and never-takers under the control. To see the impact of the distance between μ^n and μ^{c0}, we consider under what conditions the approximate maximum empirical likelihood and standard instrumental variable estimators are equal. Standard instrumental variable estimator estimates the complier average causal

effect by substituting method of moments estimates into (3). The approximate maximum empirical likelihood estimator estimates the complier average causal effect by substituting maximum empirical likelihood estimates into (3) conditional on $E(Y|R=1,A=0)$ being set equal to its method of moments estimate. The approximate maximum empirical likelihood estimator equals the standard instrumental variable estimator if the method of moments estimates of $\text{pr}(A=1|R=1)$ and $E(Y|R=1,A=0)$, denoted by $\hat{\text{pr}}(A=1|R=1)$ and $\hat{\mu}^n$ respectively, satisfy (8)-(10) with $q_i = 1/n_0$ for $i = 1,\ldots,n_0$. This will happen if and only if $\hat{\mu}^n$ is between the trimmed mean of $Y|R=0$ over the 0 to $\{1-\hat{\text{pr}}(A=1|R=1)\}$ quantiles and the trimmed mean of $Y|R=0$ over the $\hat{\text{pr}}(A=1|R=1)$ to 1 quantiles. It is more likely that $\hat{\mu}^n$ will escape these bounds when the distributions of the compliers and the never-takers are more separated. When $\hat{\mu}^n$ does escape these bounds, we expect that the approximate maximum empirical likelihood estimator will provide a better estimate than standard instrumental variable estimator because the approximate maximum empirical likelihood estimator is taking better account of the mixture structure of outcomes implied by Assumptions 1-6. Thus, we expect that the approximate maximum empirical likelihood estimator will gain more efficiency over standard instrumental variable estimator when the distance between μ^n and μ^{c0} is greater, because then the distributions of the compliers and never-takers under the control are more separated.

To see the effect of the separation between the compliers and never-takers under the control, we chose two sets of values for μ^{c0} and μ^n such that the distributions of the compliers and never-takers under the control are well separated under one set of values but are close to each other under another set of values. In setting N^1, the distributions of $Y_i^{1,1}|C_i = 1, Y_i^{0,0}|C_i = 1$ and $Y_i^{0,0} = Y_i^{1,0}|C_i = 0$ are Normal with (mean, variance) combinations $(2,1),(1,1)$ and $(3,1)$, respectively. In setting G^1 and LN^1, the distributions are gamma and lognormal, respectively. Settings N^2, G^2 and LN^2 differ only in that the (mean, variance) combination of $Y_i^{0,0} = Y_i^{1,0}|C_i = 0$ is $(1.5,1)$.

For each setting, we present summary results over 1000 replications with sample sizes of 100 and 500. Table 4 shows the bias and mean squared error from the three different

estimators for the complier average causal effect for the different settings considered. Table 4 shows the following features.

First, the parametric estimator based on the normality assumption is unbiased and more efficient than standard instrumental variable and approximate maximum empirical likelihood estimators under the true normal distributions, but shows biases of $23\% - 40\%$ and is less efficient than other two estimators under nonnormal distributions.

Secondly, both the approximate maximum empirical likelihood and standard instrumental variable estimators have low bias for all settings considered. The approximate maximum empirical likelihood estimator has bias below 5% when the distributions of the never-takers and the compliers under the control are close to each other. When the distributions of the never-takers and compliers under the control are well separated and the sample size is 100, the approximate maximum empirical likelihood estimator has a bias of about 10% but this bias drops to below 5% when the sample size increases to 500.

Thirdly, the approximate maximum empirical likelihood estimator is more efficient than standard instrumental variable estimator for all settings considered. The gain in mean squared error is more substantial when the distributions of never-takers and compliers under the control are well separated, as expected from the discussion above. The gain in mean squared error is as large as 56%. The gain is generally smaller with a sample size of 500 rather than 100. In additional simulations not presented in Table 4, we found that there is still a gain in mean squared error with the approximate maximum empirical likelihood estimator over standard instrumental variable estimator with a sample size of 1000.

We also did a simulation study for the setting of §3.3 in which the assigned to control group can access the treatment. The results are not presented, but are available from the authors. The pattern of results is similar to that for the single consent design trials.

5 Asymptotic Properties

In §4, we showed that the approximate maximum empirical likelihood estimator gains over standard instrumental variable estimator in a range of finite-sample situations, with larger gains when the compliers and never-takers' outcome distributions under the control

are more separated. The standard instrumental variable estimator is based on estimating the distribution of (Y, A, R) by the empirical distribution of (Y, A, R); the method of moments estimators on which standard instrumental variable estimator is based are the moments of the empirical distribution. The source of the approximate maximum empirical likelihood estimator's gain over standard instrumental variable estimator is that the empirical distribution of (Y, A, R) might not satisfy the restrictions given by Assumptions 1-6. The approximate maximum empirical likelihood estimator takes these restrictions into account to provide a better estimate of the distribution of (Y, A, R) than the empirical distribution. However, unless the distribution of (Y, A, R) is 'at the boundary' of the restrictions given by Assumptions 1-6, the empirical distribution of (Y, A, R) should satisfy the restrictions with probability converging to 1 as the sample size $N \to \infty$. Consequently, the approximate maximum empirical likelihood estimator will be asymptotically equivalent to the standard instrumental variable estimator. We establish this result in Theorem 1 under condition (13) below. Condition (13) specifies that the distribution of (Y, A, R) is not 'at the boundary' of the restriction that the $Y|R = 0$ is a mixture of the compliers and never-takers under the control in the sense that the distributions of the compliers and never-takers under the control overlap at least minimally. In condition (13) below, we let F^{c0} and F^{n0} denote the cumulative distribution functions of potential outcomes under the control for compliers and never-takers respectively, and we let $G = \pi_c F^{c0} + (1 - \pi_c)F^{n0}$ denote the cumulative distribution function of potential outcomes under the control. The condition is

$$\begin{aligned} \frac{1}{1-\pi_c}\int_{-\infty}^{G^{-1}(1-\pi_c)} z dG(z) &< \textstyle\int_{-\infty}^{\infty} z dF^{n0}(z) = \mu^n, \\ \mu^n = \int_{-\infty}^{\infty} z dF^{n0}(z) &< \tfrac{1}{1-\pi_c}\textstyle\int_{G^{-1}(\pi_c)}^{\infty} z dG(z). \end{aligned} \tag{13}$$

Condition (13) says that the trimmed mean of the π_n-smallest part of the mixture of never-takers and compliers is strictly less than the mean of the never-takers and that the trimmed mean of the π_n-largest part of the mixture of never-takers and compliers is strictly greater than the mean of the never-takers. Under condition (13), we have

Theorem 1. *Consider a single consent design. Suppose* (i) (13) *holds,* (ii) $0 < \pi_c < 1$ *and* (iii) $n_0/N = d, 0 < d < 1$. *Then,* $\text{pr}(\hat{\text{CACE}}_{\text{A}} = \hat{\text{CACE}}_{\text{S}}) \rightarrow 1$ *as* $N \rightarrow \infty$.

The proof of Theorem 1 is in Appendix 2.

In spite of the asymptotic equivalence result in Theorem 1, the simulation study in §4 showed that the approximate maximum empirical likelihood estimator can provide substantial gains in practical situations. The gains provided by the approximate maximum empirical likelihood estimator are analogous to the gains provided in estimating a population mean in the knowledge of restrictions on the range of the mean. For example, consider estimating the mean μ of a normal distribution $N(\mu, \sigma^2)$ based on a random sample $Y_1, \ldots, Y_N$ when it is known that μ is less than or equal to an upper bound μ_U. If μ is reasonably close to μ_U, then the maximum likelihood estimate will gain substantially over the sample mean, the maximum likelihood estimate if μ is unrestricted, for many sample sizes. However, as long as μ is less than μ_U by any amount, the estimators are equivalent asymptotically because, for large enough N, the sample mean is less than μ_U with high probability.

6 Application to Depression Study

In this section, we apply our method to analyze a randomized trial of an intervention to improve treatment of depression among depressed elderly patients in primary care practices (Bruce et al., 2004). The encouragement intervention was that a depression care specialist collaborated with the patient's primary care physician to facilitate adherence to a depression treatment strategy and provide education and assessment to the patient. The control was usual care. The study involved 539 depressed patients in 20 primary care practices at three sites followed for six visits: baseline, 4, 8, 12, 18 and 24 months. Each practice was randomized to either intervention, treatment, or usual care, control. For illustrative purposes, we ignore the fact that the trial was a group randomized trial and treat it as a completely randomized trial; for analyses that account for the group randomization, see Small et al. (2007). Compliance with the intervention was categorized as a binary variable, whether or not a patient had seen a depression care specialist in the prior four months of

follow-up. Patients in practices randomized to the usual-care group did not have access to the depression specialist, so there are only compliers and never-takers in this trial. To see the effects of estimators under different situations, we analyze two outcomes. One is the patients' Hamilton depression scores measured at 4 months, which take integer values between 0 and 50. A lower value of the outcome means less depression. Another outcome of analysis is the composite anti-depression scores among males at one site measured at 12 months. This is an integer-valued score from 0 to 4 that indicates how much the patient is being treated for depression. A score of 3 or 4 is considered adequate treatment for depression while 1 or 2 means the patient is being treated in some way, but not a what is considered an adequate dose.

Table 5 shows the three estimates of the complier average causal effect for the Hamilton and composite anti-depression scores described above. The percentile bootstrap with 1000 resamples was used to compute approximate 95% confidence intervals. We first consider the Hamilton score at 4 months; see the second column of Table 5. The scores were observed for 517 subjects and 92.7% of these subjects that were assigned to treatment complied with the treatment. All the complier average causal effect estimates are negative and the 95% confidence intervals do not include zero, indicating that the intervention has a significant beneficial effect on depression compared to usual care. Comparing the three estimation methods, we first note from the histograms of the Hamilton outcome in Fig. 2 (a)-(c) that the Hamilton scores for the never-takers and compliers under the treatment are far from normally distributed, suggesting that the parametric estimator based on the normality assumption is probably a biased estimator. The standard instrumental variable estimator and the approximate maximum empirical likelihood estimator provide very similar point estimates and similar 95% confidence intervals; see below for more explanation of this similarity. We now consider the outcome of the composite anti-depression scores among males at the site at 12 months, given in the third column of Table 5. The scores were observed for 37 subjects and 75% of these subjects who were assigned to treatment complied with the treatment. The approximate maximum empirical likelihood and standard instrumental variable com-

plier average causal effect estimates show a significant beneficial effect of the intervention on treating depression while the parametric normal estimate does not show a significant effect. As for the Hamilton score, the histograms of the composite anti-depression outcomes in Fig. 2 (d)-(f) show that the composite anti-depression scores from the never-takers and compliers under the control are far from normally distributed, suggesting that the parametric estimator based on the normality assumption is a biased estimator. Unlike for the Hamilton score, for the complier average causal effect of the intervention on the composite anti-depression score, the approximate maximum empirical likelihood estimate has a substantially narrower 95% confidence interval than standard instrumental variable estimate.

The greater gain in efficiency of the approximate maximum empirical likelihood estimate compared to standard instrumental variable estimate for the composite anti-depression study rather than the Hamilton study is related to three factors. First, the sample size in the $R = 0$ group is smaller for the composite anti-depression study, making it more likely that the empirical distribution of (Y, A, R) will deviate from the restrictions implied by Assumptions 1-6. Secondly, the compliance rate among the subjects assigned to treatment is higher for the Hamilton study, 93%, than the composite anti-depression study, 75%, providing less scope in the Hamilton study for the extra information about Assumptions 1-6 used by the approximate maximum empirical likelihood estimator to have an impact. Thirdly, the separation between the never-takers' and compliers' outcome distributions in the control group is greater for the composite anti-depression than for the Hamilton; if we use the estimates of μ^n and μ^{c0} obtained by substituting method of moments estimates into the population expressions for these quantities in (3), the estimated absolute standardized difference between the never-takers' and compliers' means in the control group is 2.34 for the composite anti-depression compared to 0.72 for the Hamilton. As we have shown in our simulation studies, the approximate maximum empirical likelihood estimator will have a larger gain in efficiency over standard instrumental variable estimator when the distributions of the never-takers and compliers in the control group are more separated.

7 Discussion

Our method can be extended to observational studies in which a variable R which encourages, $R = 1$, or does not encourage, $R = 0$, a subject to take the treatment is not randomly assigned but is 'as good as randomly assigned', that is, ignorable, conditional on some covariates; such studies are discussed in Abadie (2003) and examples are given in Table 1 of Angrist & Krueger (2001). Suppose we replace Assumption 2 with Assumption $2'$ that the encouragment variable R is independent of $Y^{1,1}, Y^{1,0}, Y^{0,1}, Y^{0,0}, A^0, A^1$ conditional on a subject's covariate vector X and that the encouragement variables of different subjects are independent. Also, suppose we expand Assumption 3 to Assumption $3'$ that $X_i, Y_i^{1,1}, Y_i^{1,0}, Y_i^{0,1}, Y_i^{0,0}, A_i^0, A_i^1$ are independent and identically distributed draws from a superpopulation and expand Assumption 4 to condition on covariates, i.e, let Assumption $4'$ be that $E(Y^{r,a}|X) = E(Y^{r',a}|X)$ for all r, r', a, X. Furthermore, for a single consent design, suppose we consider linear models for the expected potential outcomes in a compliance class given the covariates and a logistic model for compliance given the covariates, i.e., $E(Y^{1,1}|C = 1, X) = X'\beta^{c1}$, $E(Y^{0,0}|C = 1, X) = X'\beta^{c0}$, $E(Y^{1,0}|C = 0, X) = E(Y^{0,0}|C = 0, X) = X'\beta^{n}$ and $\text{pr}(C = 1|X) = \text{expit}(X'\alpha)$, where $\text{expit}(z) = e^z/(1 + e^z)$. We include an intercept in the covariate vector X and let p denote the dimension of X. Under this model, the complier average causal effect for compliers with covariate vector X is $X'\beta^{c1} - X'\beta^{c0}$. Under Assumptions 1, $2'$, $3'$, $4'$, 5 and 6 and the above models for the outcomes and compliance probabilities, we have that the empirical likelihood of α, β^{c1}, β^{c0} and β^n is $L_E(\alpha, \beta^n, \beta^{c1}, \beta^{c0}) = \max_{q_1,\ldots,q_N} \prod_{i=1}^N q_i$ subject to (i) $\sum_{i=1}^{n_0} q_i = 1, \sum_{i=n_0+1}^N q_i = 1$; (ii) $q_i \geq 0$, $i = 1, \ldots, N$; (iii) $\sum_{i=n_0+1}^N q_i X_{ij}\{A_i - \text{expit}(X_i'\alpha)\} = 0$, $j = 1, \ldots, p$; (iv) $\sum_{i=n_0+1}^N q_i A_i X_{ij}(Y_i - X_i'\beta^{c1}) = 0$, $j = 1, \ldots, p$; (v) $\sum_{i=n_0+1}^N q_i(1 - A_i)X_{ij}(Y_i - X_i'\beta^{n}) = 0$, $j = 1, \ldots, p$; (vi) there exist t_i^{c0}, t_i^n, $i = 1, \ldots, n_0$ such that (via) $t_i^{c0} + t_i^n = q_i$; (vib) $t_i^{c0}, t_i^n \geq 0$; (vic) $\sum_{i=1}^{n_0} t_i^{c0} + \sum_{i=1}^{n_0} t_i^n = 1$; (vid) $\sum_{i=1}^{n_0} t_i^{c0} X_{ij}\{1 - \text{expit}(X_i'\alpha)\} + \sum_{i=1}^{n_0} t_i^n X_{ij}\{-\text{expit}(X_i'\alpha)\} = 0$; (vie) $\sum_{i=1}^{n_0} t_i^n X_{ij}(Y_i - X_i'\beta^n) = 0$, $j = 1, \ldots, p$; and (vif) $\sum_{i=1}^{n_0} t_i^{c0} X_{ij}(Y_i - X_i'\beta^{c0}) = 0$, $j = 1, \ldots, p$. Here the t_i^{c0}, t_i^n, respectively represent the population probabilities that a subject assigned to the control has the same outcome and covariates as subject i and is a complier, never-taker respectively. The above expression for the empirical likelihood builds

on Owen's (2001, Ch. 4) discussion of empirical likelihood for regression models. As in our method of §3, we can compute the approximate maximum empirical likelihood estimate by estimating β^n using the $R = 1, A = 0$ sample and maximizing the empirical likelihood over α, β^{c1} and β^{c0} given $\beta^n = \hat{\beta}^n$.

When deriving the approximate maximum empirical likelihood estimator, we have assumed the weak exclusion restriction that the never-takers', always takers', respectively, mean is the same under assignment to treatment and control, rather than the strong exclusion restriction that the never-takers', always- takers', respectively entire outcome distribution is the same under assignment to treatment and control. In most situations in which the weak exclusion restriction is plausible, we think that the strong exclusion restriction will also be plausible. We are currently adapting our approach to situations in which the strong exclusion restriction is plausible by enabling the empirical likelihood approach to use more equality constraints for aspects of the never-takers and always-takers under $R = 0$ and $R = 1$ distributions respectively than just equality of means.

ACKNOWLEDGEMENT

We thank the associate editor, a referee and Professor D.M. Titterington for their helpful comments and suggestions.

APPENDIX 1

Details of the EM algorithm

Reexpressing the observed data likelihood $\prod_{i=1}^{n_0} q_i$ and the parameter restrictions (6) and (8)-(10) in terms of p_i^{c0}, p_i^n, we have that the observed data likelihood, with $\pi_c = \tilde{\pi}_c$, $\mu^n = \hat{\mu}^n$, is $\prod_{i=1}^{n_0}\{\tilde{\pi}_c p_i^{c0} + (1 - \tilde{\pi}_c)p_i^n\}$, with parameter restrictions

$$\sum_{i=1}^{n_0} p_i^{c0} = \sum_{i=1}^{n_0} p_i^n = 1, p_i^{c0} \geq 0, p_i^n \geq 0, i = 1, \ldots, n_0, \sum_{i=1}^{n_0} p_i^n (Y_i - \hat{\mu}^n) = 0. \qquad \text{(A1)}$$

where $q_i = \tilde{\pi}_c p_i^{c0} + (1 - \tilde{\pi}_c)p_i^n$ and $\mu^{c0} = \sum_{i=1}^{n_0} p_i^{c0} Y_i$. Note that, if $\hat{\mu}^n$ is such that there is no p_i^{c0}, p_i^n that satisfies (A1), then our approximate maximum empirical likelihood estimator does not exist; in this case we can modify the approximate maximum empirical likelihood

estimator to use $\hat{\mu}^n$ as the closest point to $\sum_{i:R_i=1,A_i=0} Y_i/\#\{R_i = 1, A_i = 0\}$, the usual estimate of μ^n for the approximate maximum empirical likelihood estimator, such that there exists p_i^{c0}, p_i^n, $i = 1, \ldots, n_0$, that satisfy (A1). If we view each subject's compliance class as missing data, the complete data likelihood is $\prod_{i:R_i=0,C_i=1} p_i^{c0} \prod_{i:R_i=0,C_i=0} p_i^n$.

E-step. The expectation of the complete data log-likelihood conditional on the observed data and the parameter estimates $p_i^{c0^{(k-1)}}$ and $p_i^{n^{(k-1)}}$ at the $(k-1)^{\text{th}}$ step is

$$\begin{aligned} Q^{(k)} &= E(\sum_{i=1}^{n_0}[C_i(\log p_i^{c0} + \log \tilde{\pi}_c) + (1 - C_i)\{\log p_i^n + \log(1 - \tilde{\pi}_c)\}|Y_1, \ldots, Y_{n_0}, p_i^{c0^{(k-1)}}, p_i^{n^{(k-1)}}] \\ &= \sum_{i=1}^{n_0}[W_i^{(k)}(\log p_i^{c0} + \log \tilde{\pi}_c) + (1 - W_i^{(k)})\{\log p_i^n + \log(1 - \tilde{\pi}_c)\}] \end{aligned}$$

where $W_i^{(k)} = \text{pr}^{(k-1)}(C_i = 1|Y_i, R_i = 0, A_i = 0) = \tilde{\pi}_c p_i^{c0^{(k-1)}}/\{\tilde{\pi}_c p_i^{c0^{(k-1)}} + (1 - \tilde{\pi}_c) p_i^{n^{(k-1)}}\}$.

M-step. We wish to maximize $Q^{(k)}$ over p_i^{c0}, p_i^n subject to (A1) with $\mu^n = \hat{\mu}^n, \pi_c = \tilde{\pi}_c$. We do this by conducting a grid search over $\mu^{c0} = \sum_{i=1}^{n_0} p_i^{c0} Y_i$. We now discuss maximizing $Q^{(k)}$ given $\mu^{c0} = \tilde{\mu}^{c0}$. We will denote the maximizing values of p_i^{c0}, p_i^n for $\mu^{c0} = \tilde{\mu}^{c0}$ by $\tilde{p}_i^{c0}, \tilde{p}_i^n$. Note that $\tilde{\mu}^{c0}$ is a possible value of μ^{c0} if and only if

$$\{p_i^{c0}, i = 1, \ldots, n_0 | \sum_i p_i^{c0} = 1, p_i^{c0} \geq 0, \sum_i p_i^{c0}(Y_i - \mu^{c0}) = 0\} \text{ is not empty.} \tag{A2}$$

For such a $\tilde{\mu}^{c0}$, maximizing $Q^{(k)}$ via Lagrange multipliers subject to (A1) and $\mu^{c0} = \tilde{\mu}^{c0}$ gives

$$\tilde{p}_i^{c0} = \frac{W_i^{(k)}}{(\sum_i W_i^{(k)})\{1 + \tilde{t}^c(Y_i - \tilde{\mu}^{c0})\}}, \qquad \tilde{p}_i^n = \frac{1 - W_i^{(k)}}{\{\sum_i (1 - W_i^{(k)})\}\{1 + \tilde{t}^n(Y_i - \hat{\mu}^n)\}}$$

where $\tilde{t}^c$ and $\tilde{t}^n$ can be determined in terms of $\tilde{\mu}^{c0}$ and $\hat{\mu}^n$ by

$$0 = \sum_i \tilde{p}_i^{c0}(Y_i - \tilde{\mu}^{c0}) = \sum_i \frac{W_i^{(k)}(Y_i - \tilde{\mu}^{c0})}{(\sum_i W_i^{(k)})\{1 + \tilde{t}^c(Y_i - \tilde{\mu}^{c0})\}} \tag{A3}$$

$$0 = \sum_i \tilde{p}_i^n(Y_i - \hat{\mu}^n) = \sum_i \frac{(1 - W_i^{(k)})(Y_i - \hat{\mu}^n)}{\{\sum_i (1 - W_i^{(k)})\}\{1 + \tilde{t}^n(Y_i - \hat{\mu}^n)\}} \tag{A4}$$

The rightmost expressions in (A3) and (A4) are monotonically decreasing in $\tilde{t}^c$ and $\tilde{t}^n$ re-

spectively, so that a safeguarded zero-finding algorithm, such as Brent's method, can be used. Starting points for the zero finding algorithm can be found by noting that, since $0 \leq \tilde{p}_i^{c0}, \tilde{p}_i^n \leq 1$,

$$\tilde{t}^c \in \left(\frac{1 - \frac{W_i^{(k)}}{\sum_i W_i^{(k)}}}{\tilde{\mu}^{c0} - Y_{(n_0)}}, \frac{1 - \frac{W_i^{(k)}}{\sum_i W_i^{(k)}}}{\tilde{\mu}^{c0} - Y_{(1)}}\right) , \; \tilde{t}^n \in \left(\frac{1 - \frac{(1-W_i^{(k)})}{\sum_i (1-W_i^{(k)})}}{\hat{\mu}^n - Y_{(n_0)}}, \frac{1 - \frac{(1-W_i^{(k)})}{\sum_i (1-W_i^{(k)})}}{\hat{\mu}^n - Y_{(1)}}\right)$$

where $Y_{(n_0)} = \max(Y_i | R_i = 0)$ and $Y_{(1)} = \min(Y_i | R_i = 0)$. The kth-step parameter estimates $p_i^{c0^{(k)}}, p_i^{n^{(k)}}$, $i = 1, \ldots, n_0$, are the $\tilde{p}_i^{c0}, \tilde{p}_i^n$ that correspond to the $\tilde{\mu}^{c0}$ that maximizes $Q^{(k)}$ over the grid of $\tilde{\mu}^{c0}$ considered. Note that we can avoid the need to consider the constraint (A2) by replacing the logarithm function with the pseudo-logarithm function of Owen (2001, p. 62) in the definition of $Q^{(k)}$.

Appendix 2

Proofs

Outline proof of Lemma 1. The complete proof is provided in a technical report available from the authors. Here we outline the steps in the proof.

Step 1. We show that maximizing $\prod_{i=1}^{n_0} q_i$ subject to (6) and (8)-(10) with $\mu^n = \hat{\mu}^n, \pi_c = \tilde{\pi}_c$ is a convex optimization problem so that there is a unique global maximum.

Step 2. Our problem involves maximization over a constrained parameter space. Nettleton (1999) shows that, under regularity assumptions, the EM algorithm converges to either (a) a stationary point or (b) a boundary point of the constrained parameter space at which the likelihood function can be increased only by moving in a direction outside the parameter space. For an unconstrained parameter space, under regularity assumptions, the EM algorithm converges only to points of type (a) (Wu, 1983). We show that, even though our parameter space is constrained, under regularity assumptions, the EM algorithm converges only to points of type (a) for our problem.

Step 3. We combine the results in Steps 1 and 2 with results about EM for unconstrained problems of Wu (1983) and Dempster et al. (1977) to prove the lemma.

Proof of Theorem 1. Let $Z_1, \ldots, Z_{n_0}$ denote the $Y|R=0$ sample, and let $\hat{\pi}_c^{R=1}$ equal the method of moments estimate of π_c based on the $R=1$ sample, $\hat{\pi}_c^{R=1} = \#\{R_i = 1, A_i = 1\}/(N-n_0)$. Note that, if there exist p_i^{c0}, p_i^n that satisfy (i) $\hat{\pi}_c^{R=1} p_i^{c0} + (1-\hat{\pi}_c^{R=1}) p_i^n = 1/n_0$, (ii) $\sum_{i=1}^{n_0} p_i^n (Z_i - \hat{\mu}^n) = 0$, (iii) $\sum_{i=1}^{n_0} p_i^{c0} = \sum_{i=1}^{n_0} p_i^n = 1$ and (iv) $p_i^{c_0}, p_i^n \geq 0$, then the approximate maximum empirical likelihood estimator equals the standard instrumental variable estimator and the maximizing values of q_i are $q_i = 1/n_0, i = 1, \ldots, n_0$. By considering the minimum and maximum values of $\sum_{i=1}^{n_0} p_i^n Z_i$ subject to (i), (iii) and (iv) above, we have that there exist p_i^{c0}, p_i^n that satisfy (i)-(iv) if and only if $\hat{\mu}^n \in [\mu_l(N), \mu_u(N)]$, where

$$\begin{aligned}
\mu_l(N) &= \sum_{i=1}^{\lfloor k_{n_0} \rfloor} Z_{(i)} \frac{1}{k_{n_0}} + Z_{(\lfloor k_{n_0} \rfloor + 1)} \frac{k_{n_0} - \lfloor k_{n_0} \rfloor}{k_{n_0}}, \\
\mu_u(N) &= \sum_{i=n_0 - \lfloor k_{n_0} \rfloor + 1}^{n_0} Z_{(i)} \frac{1}{k_{n_0}} + Z_{(n_0 - \lfloor k_{n_0} \rfloor)} \frac{k_{n_0} - \lfloor k_{n_0} \rfloor}{k_{n_0}},
\end{aligned}$$

$k_{n_0} = n_0(1 - \hat{\pi}_c^{R=1})$ and $\lfloor k \rfloor$ is the greatest integer less than or equal to k. Let $\tilde{\mu}_l(N)$ and $\tilde{\mu}_u(N)$ be the trimmed sample means of $Z_1, \ldots, Z_{n_0}$ trimmed to the $[0, 1-\pi_c]$ quantiles and $[\pi_c, 1]$ quantiles respectively; that is,

$$\tilde{\mu}_l(N) = \sum_{i=1}^{\lfloor n_0(1-\pi_c) \rfloor} Z_{(i)} \frac{1}{1-\pi_c} + Z_{(\lfloor n_0(1-\pi_c) \rfloor + 1)} \frac{n_0(1-\pi_c) - \lfloor n_0(1-\pi_c) \rfloor}{n_0(1-\pi_c)}.$$

Then, letting G denote the cumulative distribution function of the potential outcomes under the control, we have that, as $N \to \infty$, in probability,

$$\begin{aligned}
\tilde{\mu}_l(N) &\to \frac{1}{1-\pi_c} \int_{-\infty}^{G^{-1}(1-\pi_c)} z dG(z) = \mu_l^\infty, \\
\tilde{\mu}_u(N) &\to \int_{G^{-1}(\pi_c)}^{\infty} z dG(z) = \mu_u^\infty,
\end{aligned}$$

by the properties of trimmed means (Shao, 2003, Ch. 5). Now we show that $\mu_l(N) \to \mu_l^\infty$ in probability and $\mu_u(N) \to \mu_u^\infty$ in probability by showing that $|\mu_l(N) - \tilde{\mu}_l(N)| \to 0$ in probability and $|\mu_u(N) - \tilde{\mu}_u(N)| \to 0$ in probability as $N \to \infty$. We have

$$|\mu_l(N) - \tilde{\mu}_l(N)| \leq |s| \max\left(|Z_{(\lceil n_0(1-\pi_c)+n_0 s+1 \rceil)}|, |Z_{(\lfloor n_0(1-\pi_c)-n_0 s-1 \rfloor)}|\right) \tag{A5}$$

where $s = |(1-\hat{\pi}_c^{R=1})^{-1} - (1-\pi_c)^{-1}|$ and $\lceil k \rceil$ is the least integer greater than or equal to k. The first term on the right hand side of (A5) converges in probability to 0 as $N \to \infty$ and the second term converges in probability to a number less than or equal to $\max(|G^{-1}(1-\pi_c+a)|, |G^{-1}(1-\pi_c-a)|)$ for any number $a > 0$, for this, note that $n_0 = dN \to \infty$ as $N \to \infty$ since $d > 0$. This shows that the right-hand side, and hence the left hand side, of (A5) converges in probability to 0 as $N \to \infty$. Similarly,

$$|\mu_u(N) - \tilde{\mu}_u(N)| \leq |s| \max\left(|Z_{(\lceil n_0\pi_c+n_0 s+1 \rceil)}|, |Z_{(\lfloor n_0\pi_c-n_0 s-1 \rfloor)}|\right) \to 0 \text{ in probability.}$$

Thus, we conclude that $\mu_l(N) \to \mu_l^\infty$ in probability and $\mu_u(N) \to \mu_u^\infty$ in probability. By assumption (13) that the distributions of compliers and never-takers overlap, we have that $\mu_l^\infty < \mu^n$ and $\mu_u^\infty > \mu^n$. Combining the facts that $\mu_l^\infty < \mu^n < \mu_u^\infty$, $\mu_l(N) \to \mu_l^\infty$ in probability and $\mu_u(N) \to \mu_u^\infty$ in probability with the fact that $\hat{\mu}^n \to \mu^n$ in probability, by the law of large numbers, because $N - n_0 = (1-d)N \to \infty$ as $N \to \infty$, we conclude that $\text{pr}\{\mu_l(N) < \hat{\mu}^n < \mu_u(N)\} \to 1$ as $N \to \infty$. Thus, $\text{pr}(\hat{\text{CACE}}_\text{A} = \hat{\text{CACE}}_\text{S}) \to 1$ as $N \to \infty$.

REFERENCES

ABADIE, A. (2003). Semiparametric instrumental variable estimation of treatment response models. *J. Economet.* **113**, 231-63.

ANGRIST, J.D., IMBENS, G.W. & RUBIN, D.B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444-55.

ANGRIST, J.D. & KRUEGER, A.B. (2001). Instrumental variables and the search for identification. *J. Econ. Persp.* **15**, 1-17.

BOYD, S. & VANDENBERGHE, L. (2004). *Convex Optimization.* Cambridge: Cambridge University Press.

BRUCE, M., TEN HAVE, T., REYNOLDS, C., KATZ, I., SCHULBERG, H., MULSANT, B., BROWN, G., MCAVAY, G., PEARSON, J. & ALEXOPOULOS, G. (2004). Reduc-

ing suicidal ideation and depressive symptoms in depressed older primary care patients: a randomized controlled trial. *J. Am. Med. Assoc.* **291**, 1081-91.

CHENG, J. & SMALL, D. (2006). Bounds on causal effects in three-arm trials with noncompliance. *J. R. Statist. Soc. B* **68**, 815-36.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc. B* **39**, 1-38.

GRENANDER, U. (1981). *Abstract Inference*. New York: Wiley.

HALL, P. and TITTERINGTON, D.M. (1984). Efficient nonparametric estimation of mixture proportions. *J. R. Statist. Soc. B* **46**, 465-73.

HOLLAND, P.W. (1986). Statistics and causal inference. *J. Am. Statist. Assoc.* **81**, 945-60.

IMBENS, G.W. & ANGRIST, J.D. (1994). Identification and estimation of local average treatment effects. *Econometrica* **62**, 467-76.

IMBENS, G.W. & RUBIN, D.B. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *Ann. Statist.* **25**, 305-27.

IMBENS, G.W. & RUBIN, D.B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Rev. Econ. Stud.* **64**, 555-74.

LANCASTER, T. & IMBENS, G.W. (1996). Case control studies with contaminated controls. *J. of Econom.* **71**, 145-60.

NETTLETON, D. (1999). Convergence properties of the EM algorithm in constrained parameter spaces. *Can. J. Statist.* **27**, 639-48.

OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75**, 237-49.

OWEN, A.B. (2001). *Empirical Likelihood*. Boca Raton, FL: Chapman & Hall/CRC.

QIN, J. (1999). Empirical likelihood based confidence intervals for mixture proportions. *Ann. Statist.* **27**, 1368-84.

QIN, J. & LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22**, 300-25.

RUBIN, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.* **66**, 688-701.

RUBIN, D.B. (1980). Comment on a paper by D. Basu. *J. Am. Statist. Assoc.* **75**, 591-3.

SHAO, J. (2003). *Mathematical Statistics,* 2nd ed. New York: Springer.

SHEINER, L.B. & RUBIN, D.B. (1995). Intention-to-treat analysis and the goals of clinical trials. *Clin. Pharmacol. & Therap.* **57**, 6-15.

SHEN, X. & WONG, W.H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580-615.

SHEN, X., SHI, J. & WONG, W.H. (1999). Random sieve likelihood and general regression models. *J. Am. Statist. Assoc.* **94**, 835-46.

SMALL, D.S., TEN HAVE, T.R., JOFFE, M.M. & CHENG, J. (2006). Random effects logistic models for analyzing efficacy of a longitudinal randomized treatment with nonadherence. *Statist. Med.* **25**, 1981-2007.

SMALL, D.S., TEN HAVE, T.R. & ROSENBAUM, P.R. (2008). Randomization inference in a group-randomized trial of treatments for depression: covariate adjustment, noncompliance and quantile effects. *J. Am. Statist. Assoc.* **103**, 271-9.

SOMMER, A. & ZEGER, S.L. (1991). On estimating efficacy from clinical trials. *Statist. Med.* **10**, 45-52.

TANNER, M. & WONG, W. (1987). The calculation of posterior distributions by data augmentation (with Discussion). *J. Am. Statist. Assoc.* **82**, 528-50.

WU, C.F.J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95-103.

ZELEN, M. (1979). A new design for randomized clinical trials. *N. Engl. J. Med.* **300**, 1242-5.

Table 1: The relationship between observed groups and latent compliance classes

R_i	A_i		C_i	
1	1	1 (Complier)	or	2 (Always-taker)
1	0	0 (Never-taker)	or	3 (Defier)
0	0	0 (Never-taker)	or	1 (Complier)
0	1	2 (Always-taker)	or	3 (Defier)

Table 2: The relationship between observed groups and latent compliance classes in single consent design trials

R_i	A_i		C_i	
1	1	1 (Complier)		
1	0			0 (Never-taker)
0	0	1 (Complier)	or	0 (Never-taker)

Table 3: The relationship between observed groups and latent compliance classes under Assumptions $1-6$

R_i	A_i		C_i	
1	1	1 (Complier)	or	2 (Always-taker)
1	0			0 (Never-taker)
0	0	1 (Complier)	or	0 (Never-taker)
0	1			2 (Always-taker)

Table 4: Estimates of the CACE with true value 1 in single-consent treatment trials

Distn.	N	Bias			Mean squared error		
		Std. IV	AMELE	Parametric	Std. IV	AMELE	Parametric
N^1	100	0.0178	−0.1141	−0.0240	0.3482	0.2003	0.1649
	500	0.0202	−0.0016	−0.0054	0.0679	0.0515	0.0294
N^2	100	0.0150	0.0105	−0.0053	0.1682	0.1604	0.1311
	500	0.0019	0.0020	−0.0062	0.0214	0.0211	0.0186
G^1	100	0.0429	−0.0981	0.2851	0.3697	0.1945	0.2424
	500	−0.0060	−0.0212	0.3963	0.0637	0.0529	0.1907
G^2	100	0.0088	−0.0048	0.3390	0.1957	0.1726	0.2311
	500	0.0235	0.0232	0.3765	0.0454	0.0450	0.1561
LN^1	100	0.0173	−0.1364	0.2299	0.2277	0.1008	0.1897
	500	0.0177	−0.0266	0.3666	0.0411	0.0235	0.1568
LN^2	100	−0.0007	−0.0137	0.2813	0.0670	0.0563	0.1593
	500	0.0126	0.0129	0.2627	0.0120	0.0117	0.0814

Distn., distributions; Std. IV, standard instrumental variable estimate;
AMELE, approximate maximum empirical likelihood estimate

Table 5: Results from the depression study

Estimator	Hamilton score estimate (95% CI)	Composite anti-depression score estimate (95% CI)
Std. IV	$-2.55(-4.13, -0.97)$	$1.86(0.76, 3.14)$
AMELE	$-2.54(-4.12, -0.97)$	$1.60(0.73, 2.40)$
Parametric	$-2.82(-4.39, -1.16)$	$1.41(-0.66, 2.47)$

CI, confidence interval; standard IV, standard instrumental variable estimate;
AMELE, approximate maximum empirical likelihood estimate

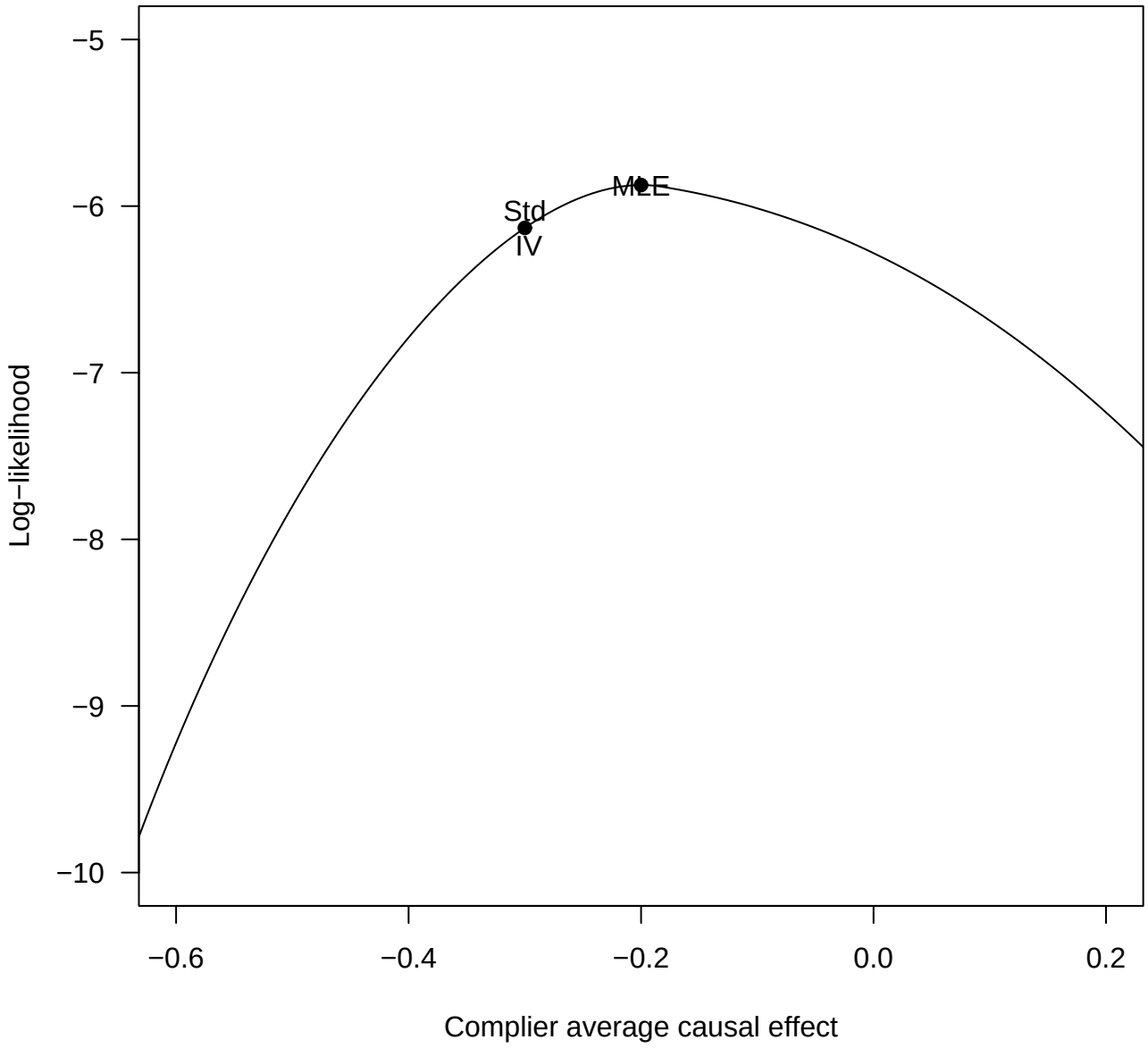

Figure 1: Profile log-likelihood for the maximum likelihood estimator and standard instrumental variable estimator of the complier average causal effect for the sample described in §3.1

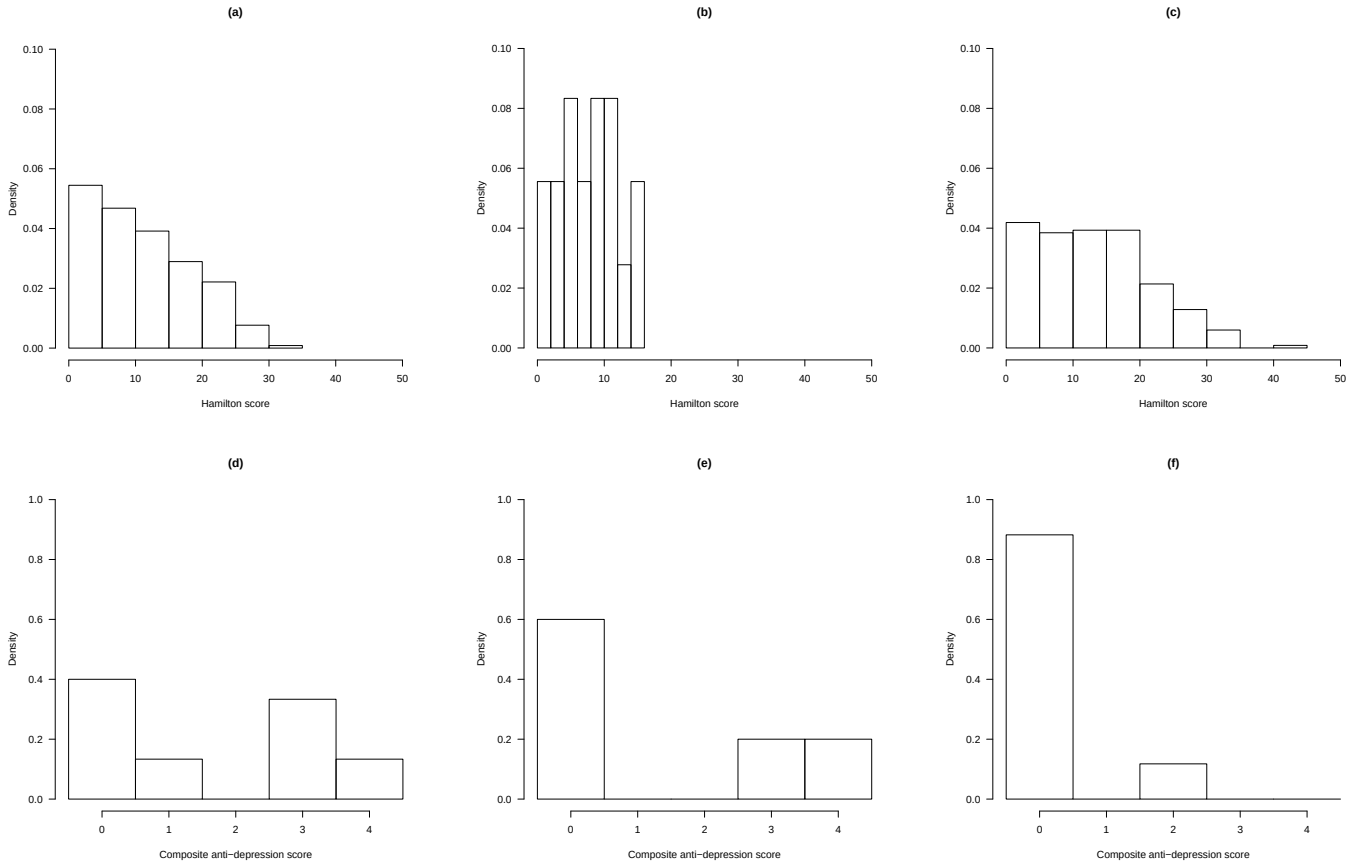

Figure 2: Depression study. Histograms of (a)-(c) the Hamilton score and (d)-(f) the composite anti-depression score for (a),(d) the $R = 1, A = 1$ group; (b), (e) the $R = 1, A = 0$ group; (c), (f) the $R = 0$ group