Relaxing monotonicity in the identification of local average treatment effects

Martin Huber and Giovanni Mellace

February 15, 2013

University of St. Gallen, Dept. of Economics

Abstract: In heterogeneous treatment effect models with endogeneity, the identification of the local average treatment effect (LATE) typically relies on an instrument that satisfies two conditions: (i) joint independence of the potential post-instrument variables and the instrument and (ii) monotonicity of the treatment in the instrument, see Imbens and Angrist (1994). We show that identification is still feasible when replacing monotonicity by a strictly weaker local monotonicity condition. We demonstrate that the latter allows identifying the LATEs on the (i) compliers (whose treatment reacts to the instrument in the intended way), (ii) defiers (who react counter-intuitively), and (iii) both populations jointly. Furthermore, (i) and (iii) coincides with standard LATE if monotonicity holds. We also present an application to the quarter of birth instrument of Angrist and Krueger (1991).

Keywords: instrumental variable, treatment effects, LATE, local monotonicity.

JEL classification: C14, C21, C26.

We have benefited from comments by Alberto Abadie, Joshua Angrist, Matias Cattaneo, Markus Frölich, Stefan Hoderlein, Guido Imbens, Toru Kitagawa, Frank Kleibergen, Tobias Klein, Michael Lechner, Arthur Lewbel, Enno Mammen, Blaise Melly, and seminar/conference participants at the University of St. Gallen (Oct 2011), Brown University (Nov 2011), the Austrian-Swiss labor meeting in Stuben (Jan 2012), Harvard (Feb 2012), MIT (Mar 2012), the Spring Meeting of Young Economists in Mannheim (Apr 2012), Boston College (Apr 2012), the North American Summer Meeting in Evanston (June-July 2012), the European Meeting of the Econometric Society in Malaga (Aug 2012), the European Association of Labour Economists meeting in Bonn (Sept 2012) and the Italian Congress of Econometrics and Empirical Economics in Genoa (Jan 2013). Martin Huber gratefully acknowledges financial support from the Swiss National Science

Foundation grant PBSGP1_138770. Addresses for correspondence: Martin Huber (martin.huber@unisg.ch) and Giovanni Mellace (giovanni.mellace@unisg.ch), SEW, University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland.

1 Introduction

In many economic evaluation problems, causal inference is complicated by endogeneity, implying that the explanatory or treatment variable of interest is correlated with unobserved factors that also affect the outcome. E.g., when estimating the returns to education, the schooling choice is plausibly influenced by unobserved ability (see for instance Card, 1999) which itself most likely has an impact on the earnings outcome. Due to the endogenous treatment selection (also known as selection on unobservables) the earnings effect of education is confounded with the unobserved terms. In the presence of endogeneity, identification relies on the availability of an instrumental variable (IV) that generates exogenous variation in the treatment.

In heterogeneous treatment effect models with a binary treatment, an instrument is conventionally required to satisfy two assumptions. Firstly, it must be independent of the joint distribution of potential treatment states and potential outcomes, which excludes direct effects on the latter. A framework where this is commonly assumed are experiments with randomized treatment assignment, where the assignment serves as instrument for actually receiving the treatment. E.g., in experimental labor market program evaluations such as the Job Training Partnership Act (JTPA) analyzed by Abadie, Angrist, and Imbens (2002), randomization ensures that the assignment is independent of any variables related with the treatment and/or the outcome. Furthermore, direct effects of the mere assignment (rather than the actual treatment) on the outcome are ruled out. Secondly, the treatment state has to vary with the instrument in a weakly monotonic manner. E.g., assignment should weakly increase actual program participation of all individuals in the population, i.e., *globally*.

Under these assumptions, Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) show that the local average treatment effect (LATE) on the subpopulation of compliers, whose treatment states react on the instrument in the intended way, is identified. This is feasible because monotonicity rules out the existence of defiers, who react counter-intuitively to the instrument, e.g., by participating in a program if not being assigned to it and not

participating under assignment. The non-existence of defiers implies the identification of the potential outcome distributions (including the means) of the compliers under treatment and non-treatment, see Imbens and Rubin (1997). The difference in mean potential outcomes is equivalent to the well known Wald formula that represents the LATE as ratio of the intention to treat effect and the share of compliers.

The contribution of this paper is to show that LATEs can still be identified and $\sqrt{n}$-consistently estimated when relying on a condition that is strictly weaker than *global* monotonicity, while maintaining joint independence. We will refer to this condition as "*local* monotonicity" (LM). Crudely speaking and in contrast to (global) monotonicity, LM allows for the existence of both compliers and defiers, but requires that they do not occur at the same time at any support point of the outcome conditional on a particular treatment state. I.e., monotonicity is assumed to hold locally in subregions of the potential outcome distribution (and may switch the sign across subregions), rather than over the entire support.

More specifically (and assuming a binary instrument), LM excludes the possibility that a subject is a defier if the joint density of her observed outcome and being treated conditional on receiving the instrument is larger than the respective joint density conditional on not receiving the instrument. In fact, under the independence of the instrument, this order of the joint densities is a sufficient condition for the existence of compliers (as outlined in Section 2). By ruling out defiers in such regions by LM, the potential outcomes of the compliers are locally identified. Conversely, in support regions in which the joint density of the outcome and the treatment when not receiving the instrument dominates the joint density when receiving the instrument, defiers necessarily exist by independence and LM rules out compliers to identify the potential outcomes of the defiers. Equivalent results hold for the joint densities under non-treatment. Therefore, we demonstrate that LM is sufficient for the identification of the potential outcome distributions of the compliers and the defiers in either treatment state. Furthermore, it immediately follows that (global) monotonicity is a special case of LM, because the former requires that the joint densities are nested, see Kitagawa (2009).

As defiers are no longer assumed away, one striking improvement of generalizing monotonicity to LM is that we do not only identify the (i) LATE on the compliers, but also the LATEs (ii) on the defiers and (iii) on the joint population of compliers and defiers. Furthermore, the existence and proportion of defiers (and any other subpopulation) can be verified in the data to judge the relevance of (ii) and (iii). It will also be shown that (i) and (iii) coincide with the standard LATE with monotonicity if the defiers do not exist. However, if the defiers' proportion is larger than zero, (i), (ii), and (iii) generally differ and standard LATE is inconsistent due to incorrectly invoking monotonicity. Finally, our discussion also reveals that our set of assumptions can be partially tested in the data. In fact, a necessary (but not sufficient) condition is the satisfaction of a particular scale constraint which has also been considered by Kitagawa (2009) and is based on the intuition that the proportion of any subpopulation must be equal across treatment states. If the latter is violated, point identification of LATEs is lost, while partial identification in the spirit of Manski (1990) might still be a worthwhile alternative, see for instance Huber and Mellace (2010).

Apart from the present work, comparably few studies have considered deviations from monotonicity and their implications for LATE identification. Small and Tan (2007) weaken (individual-level) monotonicity to stochastic monotonicity. The latter requires that conditional on a set of unobservables which satisfy particular assumptions (e.g., independence of the instrument), the probability of being a complier is weakly larger than being a defier. Conversely to this paper, Small and Tan (2007) do not propose any novel approach to identify the LATE, but show that the Wald estimator, albeit biased, retains some desirable properties (such as giving the correct sign of the effect) in the limit. Klein (2010) develops methods to assess the sensitivity of the LATE to random departures from monotonicity and to approximate the bias under particular conditions. In contrast, our framework allows for LATE identification under non-random violations, given that LM is satisfied. Finally, de Chaisemartin and D'Haultfoeuilley (2012) represent monotonicity by a latent index model, see Vytlacil (2002), in which they relax the conventional rank invariance in the unobserved terms to rank similarity,

see Chernozhukov and Hansen (2005). This amounts to assuming that the unobserved terms affecting the treatment (e.g. taste) may be a function of the instrument (which allows for the existence of defiers), but have the same distribution with and without instrument conditional on the potential outcomes. de Chaisemartin and D'Haultfoeuilley (2012) show that in this case the probability limit of the Wald estimator identifies a causal effect on a specific mixture of subpopulations. In contrast, our assumptions allow identifying the LATE on subpopulations that are well defined in terms of their treatment response to the instrument.[1]

As a practical illustration of our methods, we present an application to the U.S. census data considered by Angrist and Krueger (1991) to estimate the returns to education for males by using the birth quarter as instrument for education. Arguably, among students entering school in the same year, those who are born in an earlier quarter can drop out after less years of completed education at the age when compulsory schooling ends than those born later (in particular after the end of the academic year). This suggests that education is monotonically increasing in the quarter of birth. However, the postponement of school entry due to redshirting or unobserved school policies as discussed in Barua and Lang (2009) and Klein (2010) may reverse the relation of education and quarter of birth for some individuals and thus, violate monotonicity. We therefore invoke LM instead and indeed we find statistically significant shares of both compliers and defiers.

The remainder of this paper is organized as follows. Section 2 discusses the assumptions and the identification of the LATEs on the compliers, defiers, and both populations jointly as well as the differences/connections to the standard LATE framework with monotonicity. Section 3 considers identification in the presence of bounded non-binary instruments. Section 4 proposes $\sqrt{n}$-consistent and asymptotically normal estimators of the LATEs. An empirical application to data from Angrist and Krueger (1991) is presented in Section 5. Section 6 concludes.

[1] A further study not assuming monotonicity is Gautier and Hoderlein (2012), who in contrast to this paper consider a continuous instrument and put more structure on the selection into treatment, by assuming that the treatment is a function of a linear index with a random coefficient on the instrument. Clearly, defiers may occur when the random coefficient is negative. Finally, Torgovitsky (2011) considers a nonseparable model with a continuous endogenous treatment. He shows that if the first-stage relationship between the instrument and the treatment is strictly monotone in unobservables, identification is obtained even if the instrument is discrete. However, monotonicity of the treatment in the instrument need not be assumed.

2 Assumptions and identification

Suppose that we are interested in the average effect of a binary treatment $D \in \{1, 0\}$ (e.g., participation in a training program) on an outcome Y (e.g., labor market success such as earnings) evaluated at some point in time after the treatment. Under endogeneity, the effect of D is confounded with unobserved factors that affect both the treatment and the outcome. Identification of treatment effects generally requires an instrument, denoted by Z, that is correlated with the treatment but does not have a direct effect on the outcome (i.e., any impact other than through the treatment). In this section, we will consider the case of a binary instrument ($Z \in \{0, 1\}$) such as randomized treatment assignment, whereas Section 3 discusses the case of bounded non-binary instruments. Denote by $D(z)$ the potential treatment state for instrument $Z = z$, and by $Y(d)$ the potential outcome for treatment $D = d$ (see for instance Rubin, 1974, for a discussion of the potential outcome notation). For each subject, only one potential outcome is observed, because $Y = D \cdot Y(1) + (1 - D) \cdot Y(0)$.

Table 1: Types

Type T	D(1)	D(0)	Notion
a	1	1	Always takers
c	1	0	Compliers
d	0	1	Defiers
n	0	0	Never takers

As discussed in Angrist, Imbens, and Rubin (1996) and summarized in Table 1, the population can be categorized into four types (denoted by $T \in \{a, c, d, n\}$), depending on how the treatment state changes with the instrument. The compliers react on the instrument in the intended way by taking the treatment when $Z = 1$ and abstaining from it when $Z = 0$. For the remaining three types, $D(z) \neq z$ for either $Z = 1$, or $Z = 0$, or both: The always takers are always treated irrespective of the instrument state, the never takers are never treated, and the defiers only take the treatment when $Z = 0$. It is obvious that we cannot directly observe the type any observation belongs to as either $D(1)$ or $D(0)$ remains unknown due to the fact that the actual treatment is

$D = Z \cdot D(1) + (1 - Z) \cdot D(0)$. This implies that any observation i with a particular combination of the treatment and the instrument may belong to one of two types, see Table 2. Assuming an i.i.d. framework, we will show that the potential outcome distributions of the compliers and the defiers may nevertheless be identified under conditions that are weaker than the standard LATE assumptions of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996).

Table 2: Observed subgroups and types

Observed values of Z and D	Potential types T
$\{i : Z_i = 1, D_i = 1\}$	observation i belongs either to a or to c
$\{i : Z_i = 1, D_i = 0\}$	observation i belongs either to d or to n
$\{i : Z_i = 0, D_i = 1\}$	observation i belongs either to a or to d
$\{i : Z_i = 0, D_i = 0\}$	observation i belongs either to c or to n

To characterize the identification problem, we introduce further notation that heavily borrows from Kitagawa (2009) who, in contrast to this paper, considers partial identification of the average treatment effect (ATE) on the entire population. In a first step, we define shorthand expressions for the observed joint densities of the outcome and the treatment conditional on the instrument:

$$p_1(y) = f(y, D = 1|Z = 1), \tag{1}$$

$$p_0(y) = f(y, D = 0|Z = 1), \tag{2}$$

$$q_1(y) = f(y, D = 1|Z = 0), \tag{3}$$

$$q_0(y) = f(y, D = 0|Z = 0). \tag{4}$$

I.e., $p_d(y)$ ($q_d(y)$) represents the joint density of $Y = y$ and $D = d$ given $Z = 1$ ($Z = 0$). Furthermore, denote by $\mathcal{Y}$ the support of Y and let $f(y(d))$ and $f(y(d), T = t)$ denote the density of the potential outcome and the joint density of the potential outcome and the type, respectively,

with $d \in \{0,1\}$, $t \in \{a,c,d,n\}$, and $y \in \mathcal{Y}$. By Table 2, we have that for all $y \in \mathcal{Y}$,

$$
\begin{aligned}
p_1(y) &= f(y(1), T = c|Z = 1) + f(y(1), T = a|Z = 1), && (5)\\
q_1(y) &= f(y(1), T = d|Z = 0) + f(y(1), T = a|Z = 0), && (6)\\
p_0(y) &= f(y(0), T = d|Z = 1) + f(y(0), T = n|Z = 1), && (7)\\
q_0(y) &= f(y(0), T = c|Z = 0) + f(y(0), T = n|Z = 0), && (8)\\
f(y(1)|Z = 1) - p_1(y) &= f(y(1), T = d|Z = 1) + f(y(1), T = n|Z = 1), && (9)\\
f(y(1)|Z = 0) - q_1(y) &= f(y(1), T = c|Z = 0) + f(y(1), T = n|Z = 0), && (10)\\
f(y(0)|Z = 1) - p_0(y) &= f(y(0), T = c|Z = 1) + f(y(0), T = a|Z = 1), && (11)\\
f(y(0)|Z = 0) - q_0(y) &= f(y(0), T = d|Z = 0) + f(y(0), T = a|Z = 0). && (12)
\end{aligned}
$$

Equations (5) to (8) make immediate use of the fact that any joint density $p_d(y)$, $q_d(y)$ observed in the data is constituted by the potential outcomes (given Z) of two different types. Equations (9) to (12) come from the law of total probability, implying that $f(y(d)|Z = z) = \sum_{t\in\{a,c,d,n\}} f(y(d), T = t|Z = z)$.

We now impose the first identifying assumption which invokes independence between Z and the joint distribution of the potential treatment states and outcomes, see Imbens and Angrist (1994):

Assumption 1:

$Z \perp (D(1), D(0), Y(1), Y(0))$ (joint independence),

where "$\perp$" denotes independence. Assumption 1 is standard in the literature on the LATE and implies the randomization of the instrument (such that it is unrelated with factors affecting the treatment and/or outcome) and the exclusion of direct effects on the outcome. It follows that not only the potential outcomes, but also the types, which are defined by the potential treatment

states, are independent of the instrument. Therefore, equations (5) to (12) simplify to

$$
\begin{aligned}
p_1(y) &= f(y(1), T = c) + f(y(1), T = a), && (13)\\
q_1(y) &= f(y(1), T = d) + f(y(1), T = a), && (14)\\
p_0(y) &= f(y(0), T = d) + f(y(0), T = n), && (15)\\
q_0(y) &= f(y(0), T = c) + f(y(0), T = n), && (16)\\
f(y(1)) - p_1(y) &= f(y(1), T = d) + f(y(1), T = n), && (17)\\
f(y(1)) - q_1(y) &= f(y(1), T = c) + f(y(1), T = n), && (18)\\
f(y(0)) - p_0(y) &= f(y(0), T = c) + f(y(0), T = a), && (19)\\
f(y(0)) - q_0(y) &= f(y(0), T = d) + f(y(0), T = a), && (20)
\end{aligned}
$$

see Kitagawa (2009) for a more detailed discussion.

To understand the implications of (13) to (20) for the identification of $f(y(d), T = t)$, consider, for instance, the always takers. (13) and (14) imply that $f(y(1), T = a)$ cannot be larger than $\min(p_1(y), q_1(y))$, under Assumption 1. Secondly, by (19) and (20), the upper bound of $f(y(0), T = a)$ is $f(y(0)) - \max(p_0(y), q_0(y))$ (which is, however, not observed because $f(y(0))$ is unknown). Similar results can be derived for all other types. E.g., $f(y(0), T = n)$ is bounded from above by $\min(p_0(y), q_0(y))$ and $f(y(1), T = n)$ by $f(y(1)) - \max(p_1(y), q_1(y))$. Furthermore, Assumption 1 also provides information on the local existence and relative importance of compliers and defiers. I.e., compliers necessarily exist locally if $p_1(y) > q_1(y)$ or $q_0(y) > p_0(y)$, respectively, because by (13) to (16) this means that compliers dominate defiers (whose proportion is at least zero): $f(y(1), T = c) > f(y(1), T = d) \geq 0$ or $f(y(0), T = c) > f(y(0), T = d) \geq 0$, respectively. In this case, $p_1(y) - q_1(y)$ or $q_0(y) - p_0(y)$, respectively, provide a lower bound on the compliers. Equivalently, $p_1(y) < q_1(y)$ or $q_0(y) < p_0(y)$ point to the local existence of defiers and their dominance over compliers, such that $q_1(y) - p_1(y)$ or $q_0(y) - p_0(y)$, respectively, bound their density of the defiers from below.

By (13) to (25), also any type proportion $\Pr(T = t)$ can be bounded. To this end, we define

the following integrals also used in Kitagawa (2009) to keep the notation tractable:

$$\delta_1 = \int_{\mathcal{Y}} \max(p_1(y), q_1(y))dy, \tag{21}$$

$$\delta_0 = \int_{\mathcal{Y}} \max(p_0(y), q_0(y))dy, \tag{22}$$

$$\lambda_1 = \int_{\mathcal{Y}} \min(p_1(y), q_1(y))dy, \tag{23}$$

$$\lambda_0 = \int_{\mathcal{Y}} \min(p_0(y), q_0(y))dy. \tag{24}$$

I.e., δ_d is the integrated density envelope of $(p_d(y), q_d(y))$, while λ_d is the inner integrated density envelope of $(p_d(y), q_d(y))$. Furthermore, note that

$$\int_{\mathcal{Y}} f(y(1), T = t)dy = \int_{\mathcal{Y}} f(y(0), T = t)dy = \Pr(T = t) \quad \forall t = \{a, c, d, n\}, \tag{25}$$

because very intuitively, the proportion of any type ($\Pr(T = t)$) is necessarily equal across the potential outcome distributions under treatment and non-treatment. This is what Kitagawa (2009) refers to as scale constraint.

Again, consider the always takers to investigate the identifying power of Assumption 1 and the scale constraint. As $f(y(1), T = a)$ and $f(y(0), T = a)$ are bounded by $\min(p_1(y), q_1(y))$ and $f(y(0)) - \max(p_0(y), q_0(y))$, respectively, it follows from (22), (23), and the scale constraint(25) that the proportion of always takers, $\Pr(T = a)$, is bounded from above by the minimum of λ_1 and $1 - \delta_0$ (with $\int_{\mathcal{Y}} f(y(1))dy = 1$). Concerning the latter, note that (19) and (20) imply that $\max(p_0(y), q_0(y)) \leq f(y(0))$. Therefore, by integrating this expression we get $\delta_0 \leq 1$, otherwise Assumption 1 would be violated. Analogously, $\Pr(T = n)$ is bounded from above by the minimum of λ_0 and $1-\delta_1$ (where $\delta_1 \leq 1$). Likewise, $\int_{\mathcal{Y}} p_1(y)dy - \lambda_1 = \Pr(D = 1|Z = 1) - \lambda_1$ is a lower bound on the proportion of compliers under treatment, because the lower bound on $f(y(1), T = c)$ is $p_1(y) - q_1(y)$ if $p_1(y) > q_1(y)$ and zero if $p_1(y) \leq q_1(y)$. The respective bound under non-treatment is $\int_{\mathcal{Y}} q_0(y)dy - \lambda_0 = \Pr(D = 0|Z = 0) - \lambda_0$, such that the maximum of the proportions under treatment and non-treatment provide a lower bound on $\Pr(T = c)$. Equivalently, the lower bound

on $\Pr(T = d)$ is obtained by the maximum of $\Pr(D = 1|Z = 0) - \lambda_1$ and $\Pr(D = 0|Z = 1) - \lambda_0$.

It is obvious that Assumption 1 only allows us to derive bounds on the densities and proportions of various types. To obtain point identification, we also impose a local monotonicity assumption (LM), which rules out that compliers and defiers exist at the same time for a given value of the potential outcomes. Put differently, it is assumed that the support of the potential outcomes of the compliers and defiers do not overlap in either treatment state.[2] However, in contrast to (global) monotonicity, neither of the two populations is ruled out completely.

Assumption 2:

Either $\Pr(D(1) \geq D(0)|y(d)) = 1$ or $\Pr(D(0) \geq D(1)|y(d)) = 1$ at every value of the potential outcomes $y(d),\ d = 0, 1$ (local monotonicity)

Assumption 2 implies that whenever compliers exist locally ($\Pr(T = c|Y(1), Y(0)) > 0$), defiers are ruled out ($\Pr(T = d|Y(1), Y(0)) = 0$) and vice versa. To understand the logic of this restriction and its interaction with Assumption 1, several remarks are worth noting. Firstly, even though Assumption 2 does not specify the direction of LM for particular values of the observed outcome and the treatment, it must hold that $D(1) \geq D(0)$ whenever $p_1(y) > q_1(y)$ under treatment and $q_0(y) > p_0(y)$ under non-treatment, otherwise Assumption 1 is violated. I.e., Assumption 1 tells us which direction of LM is consistent with the data. Likewise, $D(0) \geq D(1)$ whenever $p_1(y) < q_1(y)$ under treatment and $q_0(y) < p_0(y)$ under non-treatment. By taking a look at (13),(14) and (15),(16), respectively, we also see that $D(1) = D(0)$ if $p_1(y) = q_1(y)$ under treatment and $q_0(y) = p_0(y)$ under non-treatment, such that only always takers or never takers, respectively, exist in this case. Therefore, Assumption 1 and 2 together imply the following:

$$D_i(1) \geq D_i(0)|p_1(Y_i) \geq q_1(Y_i) \text{ and } D_i(0) \geq D_i(1)|p_1(Y_i) \leq q_1(Y_i) \quad \forall \text{ subjects } i,$$

$$D_i(1) \geq D_i(0)|q_0(Y_i) \geq p_0(Y_i) \text{ and } D_i(0) \geq D_i(1)|q_0(Y_i) \leq p_0(Y_i) \quad \forall \text{ subjects } i.$$

This requires that, for instance, any treated complier ($D_i(1) > D_i(0)|p_1(Y_i) > q_1(Y_i)$) would live

[2] We thank Joshua Angrist and Toru Kitagawa for a fruitful discussion on the interpretation of LM.

in a region satisfying $q_0(y) > p_0(y)$ was she not treated (which, however, still allows for distinct support regions of complier outcomes across treatment states).

Secondly, imposing LM and ruling out defiers when $p_1(y) > q_1(y)$ or $q_0(y) > p_0(y)$ entails point identification of the compliers' density: $f(y(1), T = c) = p_1(y) - q_1(y)$ and $f(y(0), T = c) = q_0(y) - p_0(y)$, respectively, because in this case, $q_1(y)$, $p_0(y)$ only consist of always takers and never takers, respectively. I.e., the lower bounds on $f(y(1), T = c)$,$f(y(0), T = c)$ under Assumption 1 alone coincide with the point identified densities under Assumptions 1 and 2. Equivalent arguments hold for $p_1(y) < q_1(y)$ or $q_0(y) < p_0(y)$, which imply the local existence of defiers under Assumption 1 and the point identification of their densities under both assumptions, because $p_1(y)$, $q_0(y)$ now exclusively contain always takers and never takers, respectively. It therefore also follows that

$$\Pr(D = 1|Z = 1) - \lambda_1, \quad \Pr(D = 0|Z = 0) - \lambda_0$$

and

$$\Pr(D = 1|Z = 0) - \lambda_1, \quad \Pr(D = 0|Z = 1) - \lambda_0$$

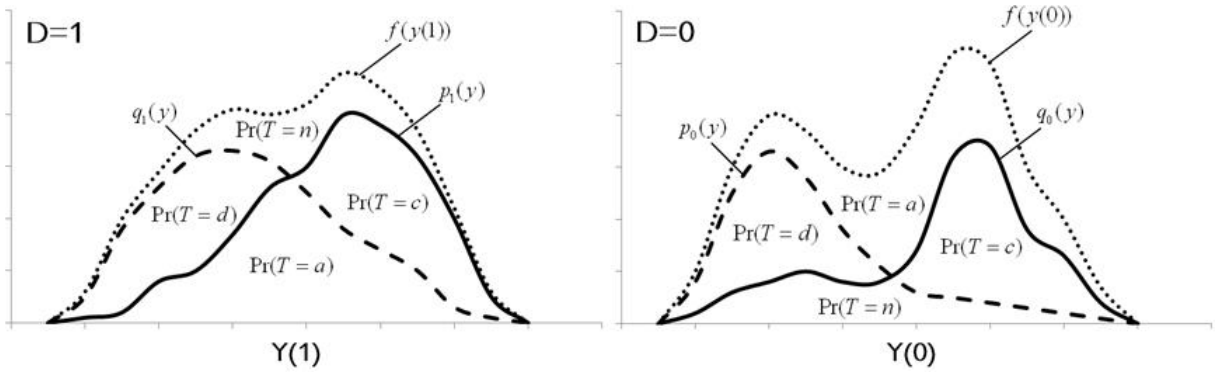
are the point identified shares under treatment and non-treatment of the compliers and defiers, respectively. Note that while $\Pr(D = 1|Z = 0) - \lambda_1 > 0$, $\Pr(D = 0|Z = 1) - \lambda_0 > 0$ imply a deviation from (global) monotonicity conditional on the satisfaction of Assumption 1 as in our framework, they point to a violation of either monotonicity or Assumption 1 (or both) if not even the independence of the instrument can be ensured. This can be used to build tests for the joint satisfaction of Assumption 1 and monotonicity in the same way as for monotonicity alone given that Assumption 1 holds.[3]

Concerning the always takers, we have already discussed that $f(y(1), T = a) = q_1(y)$ if $p_1(y) > q_1(y)$ and $f(y(1), T = a) = p_1(y)$ if $p_1(y) < q_1(y)$ such that $\min(p_1(y), q_1(y))$ is the point

[3]E.g., it is easy to see that testing the null hypothesis $H_0 : \Pr(D = 1|Z = 0) - \lambda_1 + \Pr(D = 0|Z = 1) - \lambda_0 = 0$ against the alternative $H_1 : \Pr(D = 1|Z = 0) - \lambda_1 + \Pr(D = 0|Z = 1) - \lambda_0 > 0$ jointly verifies Assumption 1 and monotonicity, with $\Pr(D = 1|Z = 0) - \lambda_1 + \Pr(D = 0|Z = 1) - \lambda_0$ being the sum of violations under treatment and non-treatment.

identified density under treatment, while it was the upper bound under Assumption 1 alone. Likewise, $\min(p_0(y), q_0(y))$ point identifies the density of the never takers under non-treatment. It follows that λ_1, $1-\delta_0$ and $1-\delta_1$, λ_0 are the respective proportions under treatment and non-treatment of the always takers and defiers. I.e., Assumptions 1 and 2 permit the identification of all type proportions and the density functions of the compliers and defiers under treatment and non-treatment. In contrast, the density function of the always takers is only identified under treatment (because $f(y(0))$ is unknown), that of the never takers only under non-treatment (because $f(y(1))$ is unknown). To visualize the results, Figure 1 displays the density functions $f(y(d)), p_d(y), q_d(y)$ as well as the locations and proportions of types defined by the intersections and differences of these densities for a hypothetical example under treatment and non-treatment.

Figure 1: Graphical illustration of the identification of type locations and proportions



Thirdly, the (global) monotonicity assumption of Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996) is a special case of LM. To see this, assume that defiers do not exist globally (i.e., assume positive monotonicity, a symmetric argument holds under negative monotonicity when compliers are ruled out) such that (14) and (15) reduce to $q_1(y) = f(y(1), T = a)$ and $p_0(y) = f(y(0), T = n)$, respectively, $\forall\ y \in \mathcal{Y}$. As discussed in Kitagawa (2009), this together with (13) to (16) implies the following nested density configuration: $p_1(y) \geq q_1(y)$ and $q_0(y) \geq p_0(y)$, $\forall\ y \in \mathcal{Y}$. Then, Assumption 2 simplifies to $\Pr(D(1) \geq D(0)) = 1$ which is (global) monotonicity.

We now formally discuss our identification results. The first proposition shows that under Assumptions 1 and 2, the proportions of all types and the potential outcome distributions of the compliers, defiers, and always takers under treatment as well as those of the compliers, defiers, and never takers under non-treatment are identified.

Proposition 1. *Under Assumptions 1 and 2,*

1. $f(y(1), T = a) = \min(p_1(y), q_1(y))$,
2. $f(y(1), T = c) = p_1(y) - \min(p_1(y), q_1(y))$,
3. $f(y(1), T = d) = q_1(y) - \min(p_1(y), q_1(y))$,
4. $f(y(1), T = n) = f(y(1)) - \max(p_1(y), q_1(y))$
5. $f(y(0), T = n) = \min(p_0(y), q_0(y))$,
6. $f(y(0), T = c) = q_0(y) - \min(p_0(y), q_0(y))$,
7. $f(y(0), T = d) = p_0(y) - \min(p_0(y), q_0(y))$,
8. $f(y(0), T = a) = f(y(0)) - \max(p_0(y), q_0(y))$ *for all* $y \in \mathcal{Y}$.
9. *the type proportions are identified by*
   $\Pr(T = a) = \lambda_1$, $\Pr(T = c) = \Pr(D = 1|Z = 1) - \lambda_1$,
   $\Pr(T = d) = \Pr(D = 1|Z = 0) - \lambda_1$, *and* $\Pr(T = n) = \lambda_0$.

Proof. See Appendix A. ■

Based on Proposition 1, the LATEs (i) on the compliers, (ii) on the defiers, and (iii) on the joint population of compliers and defiers are identified. The intuition is that since $f(y(d)|T = t) = \frac{f(y(d),T=t)}{\Pr(T=t)}$, we can use the results of Proposition 1 to identify the potential outcome distributions under treatment and non-treatment given $T \in \{c, d\}$ in order to identify the LATEs. Furthermore, if defiers do not exist, (i) and (iii) coincide with the standard LATE expression $\left(\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}\right)$ under positive monotonicity, whereas (ii) and (iii) coincide with standard

LATE under negative monotonicity if compliers do not exist. These results are formally stated in Proposition 2.

Proposition 2. *Under Assumptions 1 and 2,*

1.

$$\begin{aligned} E(Y(1)-Y(0)|T=c,d) &= \frac{\int_{\mathcal{Y}} y\cdot(\max(p_1(y),q_1(y))-\min(p_1(y),q_1(y)))dy}{\Pr(D=1|Z=1)+\Pr(D=1|Z=0)-2\cdot\lambda_1} \\ &- \frac{\int_{\mathcal{Y}} y\cdot(\max(p_0(y),q_0(y))-\min(p_0(y),q_0(y)))dy}{\Pr(D=1|Z=1)+\Pr(D=1|Z=0)-2\cdot\lambda_1}. \end{aligned} \tag{26}$$

2.

$$\begin{aligned} E(Y(1)-Y(0)|T=c) &= \frac{\int_{\mathcal{Y}} y\cdot(p_1(y)-\min(p_1(y),q_1(y)))dy}{\Pr(D=1|Z=1)-\lambda_1} \\ &- \frac{\int_{\mathcal{Y}} y\cdot(q_0(y)-\min(p_0(y),q_0(y)))dy}{\Pr(D=1|Z=1)-\lambda_1}. \end{aligned} \tag{27}$$

3.

$$\begin{aligned} E(Y(1)-Y(0)|T=d) &= \frac{\int_{\mathcal{Y}} y\cdot(q_1(y)-\min(p_1(y),q_1(y)))dy}{\Pr(D=1|Z=0)-\lambda_1} \\ &- \frac{\int_{\mathcal{Y}} y\cdot(p_0(y)-\min(p_0(y),q_0(y)))dy}{\Pr(D=1|Z=0)-\lambda_1}. \end{aligned} \tag{28}$$

4. If $\Pr(T=d)=0$ *and* $\Pr(T=c)>0$*, (26) is equivalent to* $E(Y(1)-Y(0)|T=c)=\frac{E(Y|Z=1)-E(Y|Z=0)}{E(D|Z=1)-E(D|Z=0)}$*, whereas* $E(Y(1)-Y(0)|T=d)$ *is not defined.*

5. If $\Pr(T=c)=0$ *and* $\Pr(T=d)>0$*, (26) is equivalent to* $E(Y(1)-Y(0)|T=d)=\frac{E(Y|Z=0)-E(Y|Z=1)}{E(D|Z=0)-E(D|Z=1)}$*, whereas* $E(Y(1)-Y(0)|T=c)$ *is not defined.*

Proof. See Appendix A. ■

Our discussion has shown that if the instrument satisfies Assumption 1, LATE identification does not necessarily rely on global monotonicity. In fact, the LATEs considered are equivalent

to the LATE under monotonicity if the latter assumption is indeed satisfied, but can also be identified under the weaker LM, which is partially testable. Moreover, if Assumption 2 does not hold, neither does monotonicity, such that in terms of identification there appear to be no gains when relying on standard LATE assumptions rather than the ones proposed in this section. However, albeit more general than monotonicity, LM may still be restrictive in applications, in particular with outcomes of limited support. E.g., for binary outcomes it requires that the potential outcomes of all compliers given a particular treatment state are either zero or one while all defier outcomes have the respective opposite value. Therefore, the plausibility of LM has to be critically judged in the empirical problem at hand.

To judge the implications of our assumptions in a structural model, consider the following two stage endogenous treatment selection model, with the first stage being characterized by a random coefficient model:[4]

$$\begin{aligned} Y_i &= \varphi(D_i, \epsilon_i), \\ D_i &= I\{\gamma_0 + \gamma_i Z_i + \nu_i > 0\}. \end{aligned} \tag{29}$$

$I\{\cdot\}$ is the indicator function which is equal to one if its argument holds true and zero otherwise. φ is a general function and ϵ_i, ν_i denote the unobservables in the outcome and treatment equation and may be arbitrarily correlated. γ_0, γ_i denote the constant term and the random coefficient on the instrument, respectively. Our assumptions require that whenever $p_1(Y_i) \geq q_1(Y_i)$ or $q_0(Y_i) \geq p_0(Y_i)$, respectively, $\gamma_i \geq 0$ such that $D_i(1) = I(\gamma_0 + \gamma_i Z_i + \nu_i > 0) \geq D_i(0) = I(\gamma_0 + \nu_i > 0)$, which locally rules out defiers. For $p_1(Y_i) \leq q_1(Y_i)$ or $q_0(Y_i) \leq p_0(Y_i)$, respectively, it must hold that $\gamma_i \leq 0$ such that $D_i(0) = I(\gamma_0 + \nu_i > 0) \geq D_i(1) = I(\gamma_0 + \gamma_i Z_i + \nu_i > 0)$. Note that global monotonicity would restrict γ_i to be either weakly positive or weakly negative of any i, while Assumption 2 restricts γ_i only locally.

To give an idea about possible set ups in which LM holds while monotonicity does not, we

[4]We are indebted to Joshua Angrist for making valuable suggestions concerning potential models that fit our framework.

provide two parametric examples that put further structure on the equations in (29). Firstly, assume that the outcome equation is characterized by the following model:

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 D_i \epsilon_i + \epsilon_i, \tag{30}$$

where α_0 is the constant and α_1, α_2 are the coefficients on the treatment and its interaction (capturing individual effect heterogeneity) and ϵ_i is assumed to have finite first and second moments. In this case, $Y_i(0) = \alpha_0 + \epsilon_i$, $Y_i(1) = \alpha_0 + \alpha_1 + (1 + \alpha_2)\epsilon_i$ and the individual treatment effect is $\alpha_1 + \alpha_2 \epsilon_i$. Moreover, assume that the coefficient on Z in the first stage is a deterministic function of ϵ_i:

$$\gamma_i = \beta_0 + \rho \epsilon_i, \tag{31}$$

where β_0 is a constant and ρ the coefficient on the error in the structural equation. For $\rho > 0$, it follows that $D_i(1) \geq D_i(0)$ for all $\epsilon_i \geq 0$ and $D_i(1) \leq D_i(0)$ for all $\epsilon_i \leq 0$, while the contrary holds for $\rho < 0$. As Y_i is a monotonic function of ϵ_i (unless α_2 is exactly -1 and $D_i = 1$), the outcomes of the compliers and defiers do not overlap conditional on the treatment state so that LM is satisfied.

Secondly, we consider an extension of our set up to a Roy (1951)-type model, which implies that the probability of treatment increases with the gains it creates. To this end, we maintain the previous outcome equation (30), but modify the first stage:

$$D_i \;=\; I\{Y_i(1) - Y_i(0) + \gamma_i Z_i + \nu_i > 0\} = I\{\beta_0 + \alpha_1 + (\alpha_2 + \rho Z_i)\epsilon_i + \nu_i > 0\}, \tag{32}$$

where the individual level treatment effect e.g., the returns to education or training, now influences the selection into treatment. In this case ν_i, if different from zero, may be interpreted as individual costs, disutility, or utility of the treatment not reflected by the treatment effect per se. The instrument Z exogenously shifts participation, but the direction depends again on ϵ_i as specified

in (31). The expression left of the equality follows from substituting $Y_i(1) - Y_i(0)$ by $\alpha_1 + \alpha_2\epsilon_i$ and using (31). Again, this model implies a non-overlapping support of the potential outcomes of compliers and defiers due to γ_i being a deterministic function of ϵ_i and Y_i being monotonic in ϵ_i.

Restrictions of the kind used in our examples are of course not innocuous and raise the question whether there exist empirical problems in which they seem realistic. A potentially interesting application appears to be the estimation of the returns to education based on quarter of birth instruments, see Angrist and Krueger (1991). As already discussed in Section 1, receiving a particular level of education (D) might not be monotonic in the quarter of birth (Z) due to postponing school entry (e.g., redshirting). Now assume that ϵ reflects unobserved innate ability. Then, LM is satisfied if the wage (Y)[5] is a positive function of ability and if it is ability that also determines postponement, i.e., if high ability children are admitted to school in the same year while low ability children are held back for another year ($\rho > 0$). In this case, compliers are situated on the upper part of the wage distribution conditional on a particular level of education and defiers on the lower part. This motivates the empirical application to the Angrist and Krueger (1991) data presented in Section 5, which also assesses the plausibility of the assumptions after visually inspecting the estimated outcome distributions of compliers and defiers.

We conclude this section by discussing a testable necessary, albeit not sufficient condition for Assumptions 1 and 2, namely the satisfaction of the scale constraint (25). E.g., the proportion of always takers must be equal under treatment and non-treatment, i.e., $\lambda_1 = 1 - \delta_0$, and the same applies to the never takers, $\lambda_0 = 1 - \delta_1$, or any other subpopulation. Interestingly, one constraint implies the other, which follows from Lemma A.2 in Kitagawa (2009):

$$\delta_1 + \delta_0 + \lambda_1 + \lambda_0 = 2,$$

because the four elements add up to the sum of the integrals of $p_1(y)$, $q_1(y)$, $p_0(y)$, and $q_0(y)$ (with

[5] As wage is non-negative, note that the models described before can easily be specified such that they generate a non-negative outcome, e.g., by letting α_0 in (30) be sufficiently large and letting ϵ_i be of bounded support such that $\alpha_0 + \alpha_1 D_i + \alpha_2 D_i\epsilon_i + \epsilon_i \geq 0$ always holds.

$\int_{\mathcal{Y}}(p_1(y)+p_0(y))dy = \int_{\mathcal{Y}}(q_1(y)+q_0(y))dy = 1$). Rearranging the terms yields $(1-\delta_0)-\lambda_1 = \lambda_0-(1-\delta_1)$. Therefore, it suffices to test one scale constraint, e.g., $\lambda_1 = 1-\delta_0$, as its satisfaction also entails the validity of the remaining three. However, if $\lambda_1 \neq 1-\delta_0$ (or equivalently $\lambda_0 \neq 1-\delta_1$), point identification of LATEs is generally not feasible, at least without imposing further restrictive assumptions. It therefore appears advisable to test this implication in empirical applications and we do so in our estimations presented in Section 5. We nevertheless need to bear in mind that our assumptions may be violated even if the scale constraint is not rejected.

3 Non-binary instruments

This section discusses the identification of LATEs in the presence of a multi-valued instrument with bounded support.[6] Under (global) monotonicity, Frölich (2007) shows that if the support of Z is bounded such that $Z \in [z_{\min}, z_{\max}]$, where $z_{\min}$ and $z_{\max}$ are finite upper and lower bounds, it is possible to define and identify LATEs on the compliers with respect to any two distinct subsets of the support of Z. The proportion of compliers in general varies depending on the choice of subsets and is maximized when choosing the endpoints $z_{\min}$, $z_{\max}$. In our framework which allows for compliers and defiers, this result no longer holds in general without specifying LM more tightly. To see this, let z and $z' \in [z_{\min}, z_{\max}]$ denote two subsets such that $z \neq z'$. Define $\tilde{Z}$ as

$$\tilde{Z} = \begin{cases} 1 & \text{if } Z \in z \\ 0 & \text{if } Z \in z' \end{cases}. \tag{33}$$

As an example, consider the case that the instrument can take three values, e.g. $Z \in \{0,1,2\}$, such that instead of Assumption 1 we invoke the following independence assumption:

Assumption 1a:

$Z \perp (D(2), D(1), D(0), Y(1), Y(0))$.

[6] We thank Toru Kitagawa for very helpful comments concerning the case of non-binary instruments.

Without imposing any form of monotonicity, there now exist eight types according to $D(2), D(1), D(0)$, see Table 3. Positive monotonicity rules out types 3, 5, 6, and 7 such that only always takers (type 1), never takers (type 8) and compliers when switching the instrument from 0 to 1 (type 2) or from 1 to 2 (type 4) exist. In this framework, one could possibly think of five different definitions of z, z': (i) $z = \{0\}, z' = \{1\}$, (ii) $z = \{1\}, z' = \{2\}$, (iii) $z = \{0\}, z' = \{2\}$, (iv) $z = \{0, 1\}, z' = \{2\}$, (v) $z = \{0\}, z' = \{1, 2\}$. (iii) maximizes the complier proportion, namely the joint proportion of types 2 and 4. This is the case because it may induce individuals to react on the treatment that are otherwise always or never takers when the instrument has less asymptotic power, i.e., operates over a smaller support, such as in (i), which only covers type 2, and in (ii), which covers type 4. In contrast, (iv) and (v) may be chosen to maximize finite sample power, because these set ups encounter at least as many observations as (iii), at the cost of a weakly lower complier proportion.

Table 3: Types for $Z \in \{0, 1, 2\}$

Type T	D(2)	D(1)	D(0)
1	1	1	1
2	1	1	0
3	1	0	1
4	1	0	0
5	0	1	1
6	0	1	0
7	0	0	1
8	0	0	0

Identification becomes more complicated if we abandon (global) monotonicity. Without further restrictions, all eight types may exist, out of which two are pure compliers (types 2 and 4), two are pure defiers (types 5 and 7) and two even switch from compliance to defiance (type 6) or vice versa (type 3). Clearly, if LM is imposed w.r.t. $D(1), D(0)$ only, which allows identifying LATEs within (i), or w.r.t. $D(2), D(1)$ only, which allows identifying LATEs within (ii), identification of LATEs in (iii) to (v) is generally not feasible. The reason is that the densities of compliers and defiers across (i) and (ii) may net each other out when coarsening the values of the instrument as in (iii) and (iv) or when considering endpoints only as in (v). I.e., some $y(1), y(0)$ might be inhabited by compliers in (i) and defiers in (ii) or vice versa such that any definition of

z, z' not consisting of neighboring support points in Z does in general not identify LATEs. One possibility to establish identification is to assume that LM holds over all values in the support of the instrument.

Assumption 2a:

Either $\Pr(D(2) \geq D(1) \geq D(0)|y(d)) = 1$ or $\Pr(D(0) \geq D(1) \geq D(2)|y(d)) = 1$ at every value of the potential outcomes $y(d),\ d = 0, 1$.

Assumption 2a rules out types 3 and 6 globally, implying that no individuals switch their treatment state in opposite directions for distinct pairs of instrument values. Furthermore, either defying types 5 and 7 or complying types 2 and 4 must not exist locally for any $y(d)$, meaning that over the entire range of instrument values, the support of defiers and compliers never overlaps. Under Assumptions 1a and 2a, the LATEs on types 2, 4, 5, and 7 are identified. I.e., (i) identifies the LATEs on types 2 and 7 and (ii) those on types 4 and 5. Analogously to the set up under global monotonicity, (iii) now maximizes both the proportions of compliers and defiers by identifying the LATEs on the types 2 and 4 jointly as well as on 5 and 7 jointly.

4 Estimation

This section discusses estimation based on the sample analogs of (26), (27) and (28). Throughout the exposition, we will assume that $\Pr(T = c) = \Pr(D = 1|Z = 1) - \lambda_1$ and $\Pr(T = d) = \Pr(D = 1|Z = 0) - \lambda_1$ are bounded away from zero, otherwise issues similar to the weak instrument problem in standard IV models would arise that are likely to invalidate the asymptotic properties presented below. While these issues are clearly interesting to look at, they are beyond the scope of this paper. Note that if the outcome is discrete, all elements of the identification results outlined in Proposition 2 can be estimated at a parametric rate, including the densities for which estimates (denoted by $\hat{p}$, $\hat{q}$) can be obtained using indicator functions for the values of Y: $\hat{p}_d(y) = \frac{1}{\sum Z_i} \sum (Z_i \cdot I\{Y_i = y, D_i = d\})$ and $\hat{q}_d(y) = \frac{1}{\sum 1-Z_i} \sum ((1 - Z_i) \cdot I\{Y_i = y, D_i = d\})$. In what follows we will discuss under which conditions estimators are $\sqrt{n}$-consistent and asymptotically

normal for the more complicated case that Y is continuous with the densities being estimated at a slower rate than $\sqrt{n}$. As outlined in Appendix B, the estimators are characterized by a semiparametric two step GMM procedure and belong to the class of MINPIN estimators, a general class of semiparametric two step M-estimators introduced in Andrews (1994a). By applying his results it follows that the subsequent estimators of $E(Y(1)-Y(0)|T=c,d)$, $E(Y(1)-Y(0)|T=c)$, and $E(Y(1)-Y(0)|T=d)$ (denoted by $\hat{\mu}_{c,d}$, $\hat{\mu}_c$, and $\hat{\mu}_d$) have the desired properties given that the (plug-in) first step estimators $\hat{f}(Y_i|D=d,Z=z)$ satisfy particular conditions explained further below and in Appendix B.[7]

$$
\begin{aligned}
\hat{\mu}_{c,d} &= \frac{\sum_{i=1}^n Y_i \cdot [I\{\hat{p}_1(Y_i) \geq \hat{q}_1(Y_i)\} \cdot (\hat{p}_1(Y_i) - \hat{q}_1(Y_i)) + I\{\hat{p}_1(Y_i) \leq \hat{q}_1(Y_i)\} \cdot (\hat{q}_1(Y_i) - \hat{p}_1(Y_i))]}{\hat{P}_{1|1} + \hat{P}_{1|0} - 2 \cdot \hat{\lambda}_1} \\
&- \frac{\sum_{i=1}^n Y_i \cdot [I\{\hat{p}_0(Y_i) \geq \hat{q}_0(Y_i)\} \cdot (\hat{p}_0(Y_i) - \hat{q}_0(Y_i)) + I\{\hat{p}_0(Y_i) \leq \hat{q}_0(Y_i)\} \cdot (\hat{q}_0(Y_i) - \hat{p}_0(Y_i))]}{\hat{P}_{1|1} + \hat{P}_{1|0} - 2 \cdot \hat{\lambda}_1}, \\
\hat{\mu}_c &= \frac{\sum_{i=1}^n Y_i \cdot I\{\hat{p}_1(Y_i) \geq \hat{q}_1(Y_i)\} \cdot (\hat{p}_1(Y_i) - \hat{q}_1(Y_i))}{\hat{P}_{1|1} - \hat{\lambda}_1} \\
&- \frac{\sum_{i=1}^n Y_i \cdot I\{\hat{p}_0(Y_i) \leq \hat{q}_0(Y_i)\} \cdot (\hat{q}_0(Y_i) - \hat{p}_0(Y_i))}{\hat{P}_{1|1} - \hat{\lambda}_1}, \\
\hat{\mu}_d &= \frac{\sum_{i=1}^n Y_i \cdot I\{\hat{p}_1(Y_i) \leq \hat{q}_1(Y_i)\} \cdot (\hat{q}_1(Y_i) - \hat{p}_1(Y_i))}{\hat{P}_{1|0} - \cdot\hat{\lambda}_1} \\
&- \frac{\sum_{i=1}^n Y_i \cdot I\{\hat{p}_0(Y_i) \geq \hat{q}_0(Y_i)\} \cdot (\hat{p}_0(Y_i) - \hat{q}_0(Y_i))}{\hat{P}_{1|0} - \cdot\hat{\lambda}_1},
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{P}_{1|1} &= \frac{\sum_{i=1}^n D_i \cdot Z_i}{\sum_{i=1}^n Z_i}, \quad \hat{P}_{1|0} = \frac{\sum_{i=1}^n D_i \cdot (1-Z_i)}{\sum_{i=1}^n (1-Z_i)}, \\
\hat{\lambda}_1 &= \sum_{i=1}^n I\{\hat{p}_1(Y_i) \leq \hat{q}_1(Y_i)\} \cdot \hat{p}_1(Y_i) + I\{\hat{p}_1(Y_i) > \hat{q}_1(Y_i)\} \cdot \hat{q}_1(Y_i), \\
\hat{p}_1(Y_i) &= \hat{P}_{1|1} \cdot \hat{f}(Y_i|D=1,Z=1), \quad \hat{q}_1(Y_i) = \hat{P}_{1|0} \cdot \hat{f}(Y_i|D=1,Z=0), \\
\hat{p}_0(Y_i) &= (1-\hat{P}_{1|1}) \cdot \hat{f}(Y_i|D=0,Z=1), \quad \hat{q}_0(Y_i) = (1-\hat{P}_{1|0}) \cdot \hat{f}(Y_i|D=0,Z=0), \\
&\text{and} \quad \hat{f}(Y_i|D=d,Z=z) \text{ is a non-parametric preliminary estimator of } f(Y_i|D=d,Z=z).
\end{aligned}
$$

[7] Instead of evaluating the densities at the empirical data points, asymptotically equivalent estimators can be obtained by estimating the densities at an equidistant grid of values between the empirical lower and upper bound of the outcome support and taking the sample analogs of Proposition 2. The asymptotic properties can be derived in a similar manner and are not discussed here.

Andrews (1994a) proposes a set of assumptions under which the class of MINPIN estimators is $\sqrt{n}$-consistent and asymptotically normal. The assumptions include standard regularity conditions (e.g., boundedness of the parameter space of the second step objects) and an orthogonality condition between first step objects (the densities) and second step parameters (the LATEs, λ_1, and $\Pr(D = d|Z = z)$). The latter ensures that the first step estimators of the densities $f(Y_i|D = d, Z = z)$ do not affect the asymptotic variances of the LATEs, which requires that they convergence uniformly (rather than pointwise) sufficiently fast, i.e., at least at rate $n^{-\frac{1}{4}}$.[8] As an example, Andrews (1995) discusses conditions under which nonparametric multidimensional kernel estimators satisfy this property in an estimation problem with a particular form of weak temporal dependence. Here, the first step problem is econometrically less challenging because we only need to estimate one-dimensional densities in an i.d.d. framework. Uniform almost sure convergence (which implies convergence in probability) can be easily established if the support of Y is unbounded, see for instance Theorem 1.4 in Li and Racine (2007). If the support is bounded, the bias of (local constant) kernel density estimators is potentially large close to the boundaries and of lower order than in the interior. To obtain uniform convergence in this case, one may either use specific boundary kernels designed to overcome this problem, or a local linear density estimator (instead of local constant estimation) where the bias is of the same order at the boundaries as in the interior, see for instance Jones (1993), or adaptive bandwidth methods for boundaries as discussed in Dai and Sperlich (2010).[9] Concerning the required rate of convergence of $n^{-\frac{1}{4}}$, the latter is easily obtained in the univariate case, where the fastest possible rate is $n^{-\frac{2}{5}}$.

A further important assumption in Andrews (1994a) is the smoothness of the expectations of the moment functions. Note that the presence of indicator functions (such as for example $I\{\hat{p}_1(Y_i) > \hat{q}_1(Y_i)\}$) in the LATE estimators does not allow imposing such smoothness conditions at the unit level as for example discussed in Newey (1994). However, as the estimators contain averages of these indicator functions, the smoothness condition of Andrews (1994a) is satisfied

[8] MINPIN estimators that do not satisfy the orthogonality condition may nevertheless be asymptotically normal. However, the variance term of the second step estimators will then also be a function of the first step estimation, see Andrews (1991).

[9] We refer to Cheng, Fan, and Marron (1997) and Dai and Sperlich (2010) for a comprehensive review of the literature on boundary corrections in kernel density estimation

in our case. This gives rise to a stochastic equicontinuity assumption on the empirical processes involved, which is a further requirement of asymptotic normality. Proposition 3 states that under Assumption E in Appendix B, which adapts the assumptions of Andrews (1994a) to our framework, the LATE estimators are $\sqrt{n}$-consistent and asymptotically normal.

Proposition 3. *Assume that* $\Pr(T = c)$ *and* $\Pr(T = c)$ *are bounded away from zero,* $p_d(y)$, $q_d(y)$ *with* $d \in \{1, 0\}$ *cross a finite number of times in* $\mathcal{Y}$*, and Assumption E in Appendix B is satisfied. It follows that*

1. $\hat{\mu}_{c,d} \xrightarrow{p} E(Y(1) - Y(0)|T = c, d)$
2. $\hat{\mu}_c \xrightarrow{p} E(Y(1) - Y(0)|T = c)$
3. $\hat{\mu}_d \xrightarrow{p} E(Y(1) - Y(0)|T = d)$
4. $\hat{\lambda}_1 \xrightarrow{p} \Pr(T = a)$
5. $\sqrt{n} \cdot (\hat{\mu}_{c,d} - E(Y(1) - Y(0)|T = c, d)) \xrightarrow{d} \mathcal{N}(0, V_{c,d})$
6. $\sqrt{n} \cdot (\hat{\mu}_c - E(Y(1) - Y(0)|T = c)) \xrightarrow{d} \mathcal{N}(0, V_c)$
7. $\sqrt{n} \cdot (\hat{\mu}_d - E(Y(1) - Y(0)|T = d)) \xrightarrow{d} \mathcal{N}(0, V_d)$
8. $\sqrt{n} \cdot (\hat{\lambda}_1 - \Pr(T = a)) \xrightarrow{d} \mathcal{N}(0, V_{\lambda_1})$

with $V_{c,d}$, V_c, V_d, *and* V_{λ_1} *as given in Appendix C.*

Proof. See Appendix B. ■

Chen, Linton, and Keilegom (2003) generalize the results of Andrews (1994a) to a wider class of estimators (including the MINPIN class) and discuss the assumptions needed for the validity of bootstrap based inference. We show in Appendix B that $V_{c,d}$, V_c, V_d, and V_{λ_1} can be consistently estimated by bootstrapping under only slightly stronger assumptions than required for estimation per se. There, we also discuss in greater detail the assumptions of Andrews (1994a) and Chen, Linton, and Keilegom (2003) and the conditions under which they are satisfied in our setting.

5 Empirical application

This section provides an application to the 1980 U.S. census data analyzed by Angrist and Krueger (1991), which contain 329,509 males born between 1930 and 1939. Angrist and Krueger (1991) assess the effect of education on wages by using the quarter of birth as instrument to control for potential endogeneity (e.g., due to unobserved ability) between the treatment and the outcome. The idea is that the quarter of birth instrument affects education through age-related schooling regulations. As documented in Angrist and Krueger (1992), state-specific rules require that a child must have attained the first grade admission age, which is six years in most cases, at a particular date during the year. Because schooling is compulsory until the age of 16 in most states, see Appendix 2 in Angrist and Krueger (1991), students who are born early in the year are in 10th grade when turning 16. As the school year usually starts in September and ends in July, these students have nine years of completed education if they decide to quit education as soon as possible. In contrast, students born after the end of the academic year but still entering school in the same year they turn six will have ten years of completed education at age 16. This suggests education to be monotonically increasing in the quarter of birth.

However, the quarter of birth instrument is far from being undisputed. E.g., Bound, Jaeger, and Baker (1995) challenge the validity of the exclusion restriction and present empirical results that point to systematic patterns in seasonality of birth (for instance w.r.t. performance in school, health, and family income) which may imply a direct association with the outcome. For this reason, we will only consider quarters two and three in our analysis, i.e., the warmer seasons of the year. We acknowledge that this may not completely dissipate all concerns about seasonality, but nevertheless assume that Assumption 1 is satisfied for the subsample born in the second or third quarter of the year.

Secondly, a crucial question for standard IV estimation is whether positive monotonicity holds for all individuals. This appears unlikely in the light of strategic school entry behavior as documented by Barua and Lang (2009), which may entail deviations from the schooling

regulations. The authors present empirical evidence of redshirting based on 1960 U.S. census data, implying that many parents did not enroll their children at the earliest permissible entry age but postponed school entry. This occurred particularly often when born late in the year. Likewise, Klein (2010) acknowledges that postponement may be also induced by schools, which are generally not obliged to admit all children who turn six before the state-wide cutoff date. I.e., some school districts can choose not to accept applicants that turn six late in the year and thus delay entry for one year. As discussed in Klein (2010), both redshirting and school policies may reverse the relation of education and the instrument for some individuals. Because children with postponement are close to seven when entering school and will just have started 10th grade when turning 16, some of them may decide to drop out immediately, with only nine years of completed education. In contrast, students born earlier will be at an advanced stage of the 10th grade when turning 16 and might therefore decide to complete the grade, thus having at least 10 years of completed education. For these individuals, compulsory schooling decreases in the quarter of birth and therefore violates monotonicity.

The implausibility of monotonicity motivates the use of our weaker LM, while the exclusion restriction will be maintained. As already mentioned, we confine our analysis to those males born in the second or third quarter. The instrument Z is equal to zero if born in the second quarter and equal to one if born in the third quarter. Our treatment D is a binary indicator that is equal to zero if receiving high school education or less (i.e., up to 12 years of education) and one if obtaining at least some higher education (i.e., 13 years or more). I.e., we are interested in the returns to having at least some college education. According to our definition, roughly 60% (40%) of our sample receive lower (higher) education. As in Angrist and Krueger (1991), the outcome variable Y is the log weekly wage. To investigate the incidence of compliance and defiance over birth cohorts, we perform the analysis separately for cohorts 1930-32 (53,527 observations) at the lower end and for 1937-39 (48,794 observations) at the upper end of the data window. This also helps tackling one concern raised in Angrist and Krueger (1991), namely that IV estimates are likely to be downward biased when pooling all cohorts because the effect of age on wages is not

taken into account.

Estimation is based on the sample analogs of Proposition 2. The densities $p_d(y), q_d(y)$ are estimated by kernel density estimation within subgroups defined by the treatment and the instrument and evaluated on an equidistant grid of 1000 values between the empirical lower and upper bounds of the outcome support. The Silverman (1986) rule of thumb is used for bandwidth choice.[10] Concerning inference, we bootstrap the parameters of interest 1999 times to approximate their distributions. This allows us to compute p-values by assessing the rank of the estimates in their respective re-centered bootstrap distributions. We use two-sided hypothesis tests, see for instance equation (6) of MacKinnon (2006), to obtain the p-values of the scale constraint and the LATE estimates and one-sided tests for the type proportions (the theoretical lower bound of which is zero).

Table 4 presents the results separately for the oldest and youngest three cohorts. The first column gives the estimate of $1 - \delta_0 - \lambda_1$, which tests the scale constraint and is necessarily zero if Assumptions 1 and 2 are satisfied. With p-values of 0.77 and 0.37, respectively, the data do not provide evidence for their violation. The second and third columns contain the estimates of the proportions of the compliers and defiers, respectively. Interestingly, the proportion of compliers is significantly positive at the 5% level only in the 1930-32 sample and that of the defiers only in the 1937-39 sample, which is also reflected by the dominance of compliers (defiers) in the older (younger) cohorts. This points to considerable variation in compliance and defiance across cohorts which is not captured when pooling all cohorts and/or invoking global monotonicity, which would average out complier and defier proportions. E.g., running a first stage OLS regression of D on a constant and Z in the 1937-39 sample yields a slope coefficient of -0.0013. The latter is asymptotically equivalent to $\Pr(T = c) - \Pr(T = d)$ and would correspond to the estimated share of individuals whose treatment is affected by the instrument had we assumed monotonicity. As the first stage estimate is insignificant, we would incorrectly conclude that the shares of compliers and/or defiers are not statistically different from zero due to averaging them out in the first stage,

[10] Undersmoothing by taking 0.75 times the bandwidth suggested by the Silverman (1986) rule of thumb does not crucially change the results.

while in fact both proportions are significant at the 10 % level.

The existence of defiers is accounted for by the methods based on Proposition 2 for the estimation of the LATEs on the compliers (LATE_c), defiers (LATE_d), and the joint population ($\text{LATE}_{c,d}$), but ignored by the Wald estimator (LATE_{mon}). Taking a look at the older cohorts, we see that all effects are rather imprecise such that their values should not be taken at face value. In the younger cohorts, LATE_c and $\text{LATE}_{c,d}$ are significantly positive at the 5 % level, while LATE_d is only borderline significant, but of similar magnitude as the former two parameters. This suggests a substantial increase in the wages of compliers and defiers when obtaining higher education. In contrast, LATE(mon.) is negative due to the negative first stage and insignificant. This demonstrates that the estimates are not robust to incorrectly invoking (global) monotonicity when in fact both compliers and defiers are likely to exist.

Table 4: LATEs on log weekly wage by youngest and oldest cohorts

	$1-\delta_0-\lambda_1$	$\Pr(T=c)$	$\Pr(T=d)$	LATE_c	LATE_d	$\text{LATE}_{c,d}$	LATE(mon.)
Cohorts 1930-32							
Estimate	0.0011	0.0137	0.0031	0.1595	-2.5824	-0.3507	1.0050
P-value	(0.7784)	(0.0025)	(0.4187)	(0.4442)	(0.4392)	(0.8734)	(0.2391)
Cohorts 1937-39							
Estimate	-0.0005	0.0077	0.0090	1.2721	1.0070	1.1290	-0.8752
P-value	(0.3922)	(0.0800)	(0.0125)	(0.0120)	(0.1021)	(0.0320)	(0.3792)

Note: P-values are based on 1999 bootstrap draws.

The results indicate that monotonicity is violated and therefore, the Wald estimator is inconsistent. However, the question is whether LM is a plausible alternative when interpreting our results. Even though the scale constraint is not rejected, the assumption that the potential wages of compliers and defiers do not overlap may still be violated. If there exists a reasonable theory on which outcome regions are inhabited by compliers and defiers, then visually inspecting the support of both groups appears to be a useful plausibility check. E.g., as discussed in Section 2, one might assume that conditional on education, individuals who earn low wages have a low level of innate ability which also induced them to postpone schooling. This suggests that defiers are concentrated in the lower part of the wage distribution and compliers in the upper part given the treatment state. We may verify our presumption by plotting the estimated densities of the

compliers and defiers.

Figure 2: Est. of $f(y(1), T = c)$, $f(y(1), T = d)$ and $f(y(0), T = c)$, $f(y(0), T = d)$

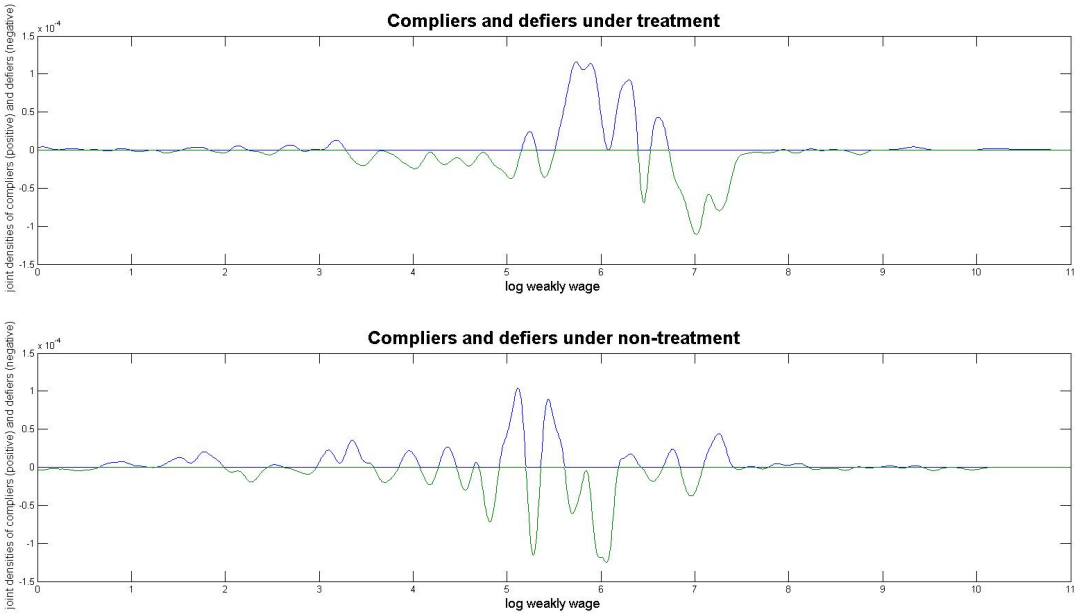


Figure 2 displays the estimates of the joint densities of the outcome values and being a complier or defier, respectively, under treatment ($f(y(1), T = c)$, $f(y(1), T = d)$, upper graph) and non-treatment ($f(y(0), T = c)$, $f(y(0), T = d)$, lower graph) for the youngest cohorts (1937-39). Positive densities represent compliers, negative ones defiers. We see that our theory is not supported by the data. In either treatment state, several switches between compliance and defiance occur over the support of the outcome that appear hard to justify by any reasonable economic model. Although LM may not be convincing in the problem considered, our method at the very least allows testing the even stronger standard LATE assumptions which are obviously violated due to the occurrence of defiers. I.e., the data do provide us with information about the validity of monotonicity assumptions, a fact that has so far been ignored in the applied IV literature.

6 Conclusion

We have demonstrated that local average treatment effects (LATEs) are identified under strictly weaker conditions than the standard assumptions invoked in the literature, see Imbens and Angrist (1994) and Angrist, Imbens, and Rubin (1996). Under the joint independence of the instrument and the potential treatment states/outcomes, (global) monotonicity of the treatment in the instrument may be weakened to local monotonicity (LM). This brings the improvement that defiers need no longer be assumed away such that LATEs on the defiers as well as on the joint population of defiers and compliers are identified for the first time in addition to the effect on the compliers. Furthermore, our set of assumptions can be partially verified in the data. Even though improving on monotonicity, it nevertheless has to be acknowledged that LM is by no means innocuous, because it puts restrictions on the potential outcomes of latent subgroups. Specifically, it requires that the outcome distributions of defiers and compliers do not overlap in either treatment state. It nevertheless appears to be the weakest possible assumption under which (along with the independence of the instrument) point identification is feasible in the heterogeneous treatment effect model with endogeneity.

As an empirical illustration, we have applied our methods to U.S. census data previously analyzed by Angrist and Krueger (1991) to estimate the returns to higher education for males by using the birth quarter as instrument for education. We have documented the presence of both compliers and defiers, who specifically in later cohorts are of similar magnitude as the compliers. Our results show that the LATE estimates are not robust to ignoring defiers, which illustrates the inconsistency of the Wald estimator. While the methods proposed in this paper can be used to test the validity of standard LATE assumptions, one needs to bear in mind that also LM imposes strong restrictions on the data, which are only partly testable. A visual inspection of the outcome distributions of compliers and defiers in either treatment state may help to judge the plausibility of LM.

A Proofs of the identification results

Proof of Proposition 1.

1. Assumption 2 together with (13) and (14) implies that $f(y(1), T = a) = q_1(y)$ if $p_1(y) \geq q_1(y)$ and $f(y(1), T = a) = p_1(y)$ if $p_1(y) \leq q_1(y)$, therefore $f(y(1), T = a) = \min(p_1(y), q_1(y))$.
2. This follows immediately by substituting $f(y(1), T = a) = \min(p_1(y), q_1(y))$ into (13).
3. This follows immediately by substituting $f(y(1), T = a) = \min(p_1(y), q_1(y))$ into (14).
4. Assumption 2 together with (17) and (18) implies that $f(y(1), T = n) = f(y(1)) - p_1(y)$ if $p_1(y) \geq q_1(y)$ and $f(y(1), T = n) = f(y(1)) - q_1(y)$ if $p_1(y) \leq q_1(y)$, therefore $f(y(1), T = n) = f(y(1)) - \max(p_1(y), q_1(y))$.
5. Assumption 2 together with (15) and (16) implies that $f(y(0), T = n) = q_0(y)$ if $q_0(y) \geq p_0(y)$ and $f(y(0), T = n) = p_0(y)$ if $q_0(y) \leq p_0(y)$, therefore $f(y(0), T = n) = \min(p_0(y), q_0(y))$.
6. This follows immediately by substituting $f(y(0), T = n) = \min(p_0(y), q_0(y))$ into (16).
7. This follows immediately by substituting $f(y(0), T = n) = \min(p_0(y), q_0(y))$ into (15).
8. Assumption 2 together with (19) and (20) implies that $f(y(0), T = a) = f(y(0)) - q_0(y)$ if $q_0(y) \geq p_0(y)$ and $f(y(0), T = a) = f(y(0)) - p_0(y)$ if $q_0(y) \leq p_0(y)$, therefore $f(y(0), T = a) = f(y(0)) - \max(p_0(y), q_0(y))$.
9. Integration of $f(y(1), T = a) = \min(p_1(y), q_1(y))$ immediately gives $\Pr(T = a) = \lambda_1$ and integration of $f(y(0), T = n) = \min(p_0(y), q_0(y))$ gives $\Pr(T = n) = \lambda_0$. Concerning the remaining type proportions note that the integrals over (13) and (14) give $\Pr(T = c) = \Pr(D = 1|Z = 1) - \Pr(T = a)$ and $\Pr(T = d) = \Pr(D = 1|Z = 0) - \Pr(T = a)$. It follows that $\Pr(T = c) = \Pr(D = 1|Z = 1) - \lambda_1$ and $\Pr(T = d) = \Pr(D = 1|Z = 0) - \lambda_1$.

■

Proof of Proposition 2.

1. It suffices to show that $f(y(1)|T = d, c) = \frac{\max(p_1(y), q_1(y)) - \min(p_1(y), q_1(y))}{\Pr(D=1|Z=1) + \Pr(D=1|Z=0) - 2 \cdot \lambda_1}$ since a symmetric argument can be used to demonstrate that $f(y(0)|T = d, c) = \frac{\max(p_0(y), q_0(y)) - \min(p_0(y), q_0(y))}{\Pr(D=1|Z=1) + \Pr(D=1|Z=0) - 2 \cdot \lambda_1}$. From Proposition 1, it follows that

$$\begin{aligned} f(y(1), T = c, d) &= f(y(1), T = c) + f(y(1), T = d), \\ &= p_1(y) + q_1(y) - 2 \cdot \min(p_1(y), q_1(y)), \\ &= \max(p_1(y), q_1(y)) - \min(p_1(y), q_1(y)). \end{aligned}$$

Therefore,

$$\begin{aligned} f(y(1)|T=c,d) &= \frac{f(y(1),T=c,d)}{\Pr(T=c,d)}, \\ &= \frac{f(y(1),T=c,d)}{\pi_c+\pi_d}, \\ &= \frac{\max(p_1(y),q_1(y))-\min(p_1(y),q_1(y))}{\Pr(D=1|Z=1)+\Pr(D=1|Z=0)-2\cdot\lambda_1}, \end{aligned}$$

which ends this part of the proof.

2. Similarly to the proof of point 1 we will just show that $f(y(1)|T=c) = \frac{p_1(y)-\min(p_1(y),q_1(y))}{\Pr(D=1|Z=1)-\lambda_1}$. From Proposition 1 it follows that

$$\begin{aligned} f(y(1)|T=c,d) &= \frac{f(y(1),T=c)}{\Pr(T=c)}, \\ &= \frac{p_1(y)-\min(p_1(y),q_1(y))}{\Pr(D=1|Z=1)-\lambda_1}, \end{aligned}$$

which ends this part of the proof.

3. The proof of this point is symmetric to the one of point 2 and is therefore omitted.

4. Under positive monotonicity,

$$\begin{aligned} \max(p_1(y),q_1(y)) &= p_1(y), \\ \max(p_0(y),q_0(y)) &= q_0(y), \\ \min(p_1(y),q_1(y)) &= q_1(y), \\ \min(p_0(y),q_0(y)) &= p_0(y), \\ \lambda_1 &= \Pr(D=1|Z=0). \end{aligned}$$

Therefore, (26) simplifies to

$$\begin{aligned} E(Y(1)-Y(0)|T=c,d) &= \frac{\int_{\mathcal{Y}} y\cdot(p_1(y)-q_1(y))dy}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}, \\ &- \frac{\int_{\mathcal{Y}} y\cdot(q_0(y)-p_0(y))dy}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}, \\ &= \frac{\int_{\mathcal{Y}} y\cdot p_1(y)dy-\int_{\mathcal{Y}} y\cdot q_1(y)dy}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}, \\ &- \frac{\int_{\mathcal{Y}} y\cdot q_0(y)dy-\int_{\mathcal{Y}} y\cdot p_0(y)dy}{\Pr(D=1|Z=1)-\Pr(D=1|Z=0)}. \end{aligned}$$

Considering $\int_{\mathcal{Y}} y \cdot p_1(y)dy$, it is easy to see that

$$\begin{aligned}
\int_{\mathcal{Y}} y \cdot p_1(y)dy &= \int_{\mathcal{Y}} y \cdot f(Y, D=1|Z=1)dy, \\
&= \int_{\mathcal{Y}} y \cdot f(Y|Z=1, D=1) \cdot \Pr(D=1|Z=1)dy, \\
&= \Pr(D=1|Z=1) \cdot \int_{\mathcal{Y}} y \cdot f(Y|Z=1, D=1)dy, \\
&= \Pr(D=1|Z=1) \cdot E(Y|Z=1, D=1).
\end{aligned}$$

In a similar way it can be shown that

$$\begin{aligned}
\int_{\mathcal{Y}} y \cdot q_1(y)dy &= \Pr(D=1|Z=0) \cdot E(Y|Z=0, D=1), \\
\int_{\mathcal{Y}} y \cdot q_0(y)dy &= \Pr(D=0|Z=0) \cdot E(Y|Z=0, D=0), \\
\int_{\mathcal{Y}} y \cdot p_0(y)dy &= \Pr(D=0|Z=1) \cdot E(Y|Z=1, D=0).
\end{aligned}$$

Therefore,

$$\begin{aligned}
E(Y(1)-Y(0)|T=c,d) &= \frac{\Pr(D=1|Z=1) \cdot E(Y|Z=1, D=1)}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)} \\
&+ \frac{\Pr(D=0|Z=1) \cdot E(Y|Z=1, D=0)}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)} \\
&- \frac{\Pr(D=1|Z=0) \cdot E(Y|Z=0, D=1)}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)} \\
&- \frac{\Pr(D=0|Z=0) \cdot E(Y|Z=0, D=0)}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)} \\
&= \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)},
\end{aligned}$$

which is the Wald formula. It is easy to see that also (27) gives the same result:

$$\begin{aligned}
E(Y(1)-Y(0)|T=c) &= \frac{\int_{\mathcal{Y}} y \cdot (p_1(y) - q_1(y))dy}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)} \\
&- \frac{\int_{\mathcal{Y}} y \cdot (q_0(y) - p_0(y))dy}{\Pr(D=1|Z=1) - \Pr(D=1|Z=0)}, \\
&= E(Y(1)-Y(0)|T=c,d).
\end{aligned}$$

Finally, since the denominator of (28) is zero, this parameter is not defined.

5. The proof of this point is symmetric to the one of point 4 and is therefore omitted.

■

B Proof of the asymptotic properties of the estimators

In what follows we will just derive the asymptotic properties of $\hat{\mu}_c$, the estimator of $E(Y(1)-Y(0)|T=c)$, because the asymptotic properties of $\hat{\mu}_{c,d}$ and $\hat{\mu}_d$ can be obtained in an equivalent way. The proof is based on the fact that $\hat{\mu}_c$ and $\hat{\lambda}_1$ are the unique solutions of a two step semiparametric GMM optimization problem and belong to the class of MINPIN estimators as defined in Andrews (1994a). Consistency and asymptotic normality are shown by applying Theorem A-1 and 2 therein.

We start by introducing some notation. Define $W=(Y,D,Z)$ to be the joint distribution of the variables. Denote by Θ the finite dimensional parameter set (we assume $\Theta \subset \mathbb{R}^4$) and by $\mathcal{T}$ the infinite dimensional parameter set of the first step. We assume $\mathcal{T}$ to be a pseudo-metric space with pseudo-metric ρ. The true values of the unknown parameters θ and τ are denoted by θ_0 and τ_0, respectively. In estimation problem, the finite dimensional parameter vector is given by

$$\theta = \begin{pmatrix} \mu_c \\ \lambda_1 \\ P_{1|1} \\ P_{1|0} \end{pmatrix} = \begin{pmatrix} E(Y(1)-Y(0)|T=c) \\ \lambda_1 \\ \Pr(D=1|Z=1) \\ \Pr(D=1|Z=0) \end{pmatrix},$$

and the infinite dimensional parameter vector is

$$\tau(W_i) = \begin{pmatrix} \tau_1(W_i) \\ \tau_2(W_i) \\ \tau_3(W_i) \\ \tau_4(W_i) \end{pmatrix} = \begin{pmatrix} f(Y_i|D=1,Z=1) \\ f(Y_i|D=1,Z=0) \\ f(Y_i|D=0,Z=1) \\ f(Y_i|D=0,Z=0) \end{pmatrix}.$$

Let $\bar{m}_n(\theta,\tau) = \frac{\sum m(W_i,\theta,\tau(W_i))}{n}$ be a non-random measurable vector-valued function $\Theta \times \mathcal{T} \mapsto \mathbb{R}^4,\ \Theta \subset \mathbb{R}^4$, where

$$m(W_i,\theta,\tau(W_i)) = \begin{pmatrix} n\cdot\gamma_c(W_i) - \mu_c\cdot(P_{1|1}-\lambda_1) \\ n\cdot\gamma_{\lambda_1}(W_i) - \lambda_1 \\ (D_i - P_{1|1})\cdot Z_i \\ (D_i - P_{1|0})\cdot(1-Z_i) \end{pmatrix},$$

with

$$\begin{aligned} \gamma_c(W_i) &= [Y_i \cdot I\{P_{1|1}\cdot\tau_1(W_i) \geq P_{1|0}\cdot\tau_2(Y_i)\}\cdot(P_{1|1}\cdot\tau_1(W_i) - P_{1|0}\cdot\tau_2(W_i))] \\ &- [Y_i \cdot I\{(1-P_{1|0})\cdot\tau_4(W_i) \geq (1-P_{1|1})\cdot\tau_3(W_i)\}\cdot((1-P_{1|0})\cdot\tau_4(W_i) - (1-P_{1|1})\cdot\tau_3(W_i))], \end{aligned}$$

and

$$\gamma_{\lambda_1}(W_i) \quad = \quad I\{P_{1|0} \cdot \tau_2(W_i) \geq P_{1|1} \cdot \tau_1(Y_i)\} \cdot P_{1|1} \cdot \tau_1(W_i) + I\{P_{1|1} \cdot \tau_1(W_i) \geq P_{1|0} \cdot \tau_2(Y_i)\} \cdot P_{1|0} \cdot \tau_2(W_i).$$

Note that the first moment condition is the difference between n times the sample counterpart of the numerator of μ_c and its population equivalent ($\sum_{i=1}^{n} \mu_c \cdot (P_{1|1} - \lambda_1) = n \cdot \mu_c \cdot (P_{1|1} - \lambda_1)$).

Given a preliminary estimator $\hat{\tau}$, the estimator $\hat{\theta}$ solves the minimization problem

$$\min_{\theta \in \Theta} \bar{m}_n(\theta, \hat{\tau})' \bar{m}_n(\theta, \hat{\tau}).$$

To show that our estimator belongs to the class of MINPIN estimators, first consider the definition of a MINPIN estimator given in Andrews (1994a):

Definition. *A sequence of MINPIN estimators $\{\hat{\theta}\}$ is any sequence of random variables that satisfies*

$$d(\bar{m}_n(\hat{\theta}, \hat{\tau}), \hat{\kappa}) = \inf_{\theta \in \Theta} d(\bar{m}_n(\theta, \hat{\tau}), \hat{\kappa}) \quad w.p. \to 1,$$

where $\hat{\kappa}$ is similarly to τ a preliminary and possibly infinite dimensional estimator. Usually either $d(\cdot, \kappa) = \bar{m}'\kappa\bar{m}/2$, where κ are weights, or κ does not exist as in our just identified case. Therefore, $\hat{\theta}$ is a MINPIN estimator with $d(m, \kappa) = \bar{m}'\bar{m}/2$.

If we choose $\hat{\tau}$ such that Assumptions C and N of Andrews (1994a) are satisfied we can apply Theorem A-1 and Theorem 2 therein to show consistency and asymptotic normality of $\hat{\theta}$. For example, if the support of Y is not bounded, one might want to estimate τ by kernel density estimation:

$$\hat{\tau}(W_i) = \begin{pmatrix} \frac{1}{l_1 \cdot n \cdot \sum_{i=1}^{n} \hat{\tau}_1(W_i)} \cdot \sum_{j=1}^{n} D_i \cdot Z_i \cdot K\left(\frac{Y_i - Y_j}{l_1}\right) \\ \frac{1}{l_2 \cdot n \cdot \sum_{i=1}^{n} \hat{\tau}_2(W_i)} \cdot \sum_{j=1}^{n} D_i \cdot (1 - Z_i) \cdot K\left(\frac{Y_i - Y_j}{l_2}\right) \\ \frac{1}{l_3 \cdot n \cdot \sum_{i=1}^{n} \hat{\tau}_3(W_i)} \cdot \sum_{j=1}^{n} (1 - D_i) \cdot Z_i \cdot K\left(\frac{Y_i - Y_j}{l_3}\right) \\ \frac{1}{l_4 \cdot n \cdot \sum_{i=1}^{n} \hat{\tau}_4(W_i)} \cdot \sum_{j=1}^{n} (1 - D_i) \cdot (1 - Z_i) \cdot K\left(\frac{Y_i - Y_j}{l_4}\right) \end{pmatrix},$$

where $K(\cdot)$ is the kernel function (e.g., the Gaussian kernel) and l_1, l_2, l_3 and l_4, are bandwidth parameters that are assumed to be optimally chosen. If the outcome is of bounded support, one may use boundary kernels, local linear density estimation, or adaptive bandwidth methods to overcome the poor properties of standard (local constant) kernel density estimation at the boundaries of the support of Y, see the discussion in Section 4).

We introduce some further notation required in our Assumption E below, which adapts Assumptions C and N

of Andrews (1994a) to our framework. Let Θ_0 be a subset of Θ that contains a neighborhood around θ_0 and define

$$\begin{aligned} \upsilon_n(\tau) &= \sqrt{n} \cdot (\bar{m}_n(\theta_0, \tau) - E(\bar{m}_n(\theta_0, \tau))), \\ H &= \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} E(m_n(\theta_0, \tau_0)), \\ S &= \lim_{n\to\infty} Var(\sqrt{n} \cdot \bar{m}_n(\theta_0, \tau_0)). \end{aligned}$$

Assumption E:

1. $W_1 = (Y_1, D_1, Z_1), \ldots, W_n = (Y_n, D_n, Z_n)$ is an i.i.d. sample from the joint distribution of (Y, D, Z).
2. Θ is bounded, θ_0 lies in an interior of Θ and $E|Y|^{2+\eta} < \infty$ for some integer $\eta \geq 0$.
3. $E(m_n(\theta, \tau))$ is continuously differentiable in θ on Θ_0 and $\frac{\partial}{\partial\theta} E(m_n(\theta, \tau))$ satisfy weak law of large numbers over $\Theta \times \mathcal{T}$.
4. $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial\theta} E(m_n(\theta, \tau))$ and $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E(\bar{m}_n(\theta,\tau))}{n}$ exist uniformly over $(\theta, \tau) \in \Theta_0 \times \mathcal{T}$ and $\Theta \times \mathcal{T}$, respectively. The matrices S and H exist.
5. H is non singular and $\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E(\bar{m}_n(\theta,\tau))}{n}$ is bounded away from zero for all θ outside any given neighborhood of θ_0.
6. $f(Y|D = d, Z = z)$ is absolutely continuous with respect to Lebesgue measure for $d, z = 1, 0$.
7. $\Pr(\hat{\tau} \in \mathcal{T}) \to 1$ and $\hat{\tau} \xrightarrow{p} \tau$.
8. $\sqrt{n} \cdot E(m_n(\theta_0, \hat{\tau})) \xrightarrow{p} 0$.
9. $\upsilon_n(\tau_0) \xrightarrow{d} \mathcal{N}(0, S)$.
10. $\upsilon_n(\cdot)$ is stochastically equicontinuous at τ_0.

Assumption E(1) can be relaxed to allow for some time dependence structure in the data. The first part of Assumption E(2) is standard and ensures that a sequence of consistent estimators of θ exists. The second part of Assumption E(2) is required to obtain uniform convergence of the first step estimators and to apply the weak law of large numbers and the central limit theorem. Assumptions E(3) to E(5) hold naturally under Assumptions E(1) and E(2) for H and S given below. Assumption E(6) is required for the first step estimation. Assumption E(7) is crucial and imposes uniform convergence of $\hat{\tau}$. When Y is bounded, uniform convergence can be obtained by using boundary kernels estimators (see Section 4). Assumption E(9) is satisfied by applying a standard central limit theorem. Assumption E(10) is a smoothness condition on the empirical process $\upsilon_n(\cdot)$ and is satisfied under Assumption E(1), E(2), and E(7) and the weak law of large numbers (see Andrews, 1994a and 1994b). To see this, consider the following pseudo-metric $\rho(\tau, \tau^*) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E \|\bar{m}_n(\theta_0, \tau) - \bar{m}_n(\theta_0, \tau^*)\|$. Under E(7),

$\rho(\hat{\tau}, \tau_0) \xrightarrow{p} 0$ and under E(1) and the second part of E(2), $\upsilon_n(\hat{\tau}) - \upsilon_n(\tau_0) \xrightarrow{p} 0$ by the weak law of large numbers, which is the definition of stochastic equicontinuity given in Andrews (1994a). Finally, Assumption E(8) is an asymptotic orthogonality condition of θ and τ, which ensures that the estimation of τ does not affect the asymptotic variance of θ. As discussed in Andrews (1994a), E(8) holds under stochastic equicontinuity if $\sup_{y \in \mathcal{Y}} |\hat{\tau}(y) - \tau_0(y)| = o_p\left(n^{-\frac{1}{4}}\right)$. Since the densities in τ are univariate, this rate of convergence can be easily obtained. Otherwise, Assumption E(8) could be replaced by $\sqrt{n} \cdot E(m_n(\theta_0, \hat{\tau})) \xrightarrow{d} \mathcal{N}(0, A)$ and in that case $\sqrt{n} \cdot (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1}(S + A)(H^{-1})')$. I.e., asymptotic normality would still hold but the variance of $\hat{\theta}$ would be affected by the first step density estimation.

By applying Theorem A-1 and Theorem 2 of Andrews (1994a) under Assumption E, we have that

$$\hat{\theta} \xrightarrow{p} \theta_0 \quad \text{and} \quad \sqrt{n} \cdot (\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, H^{-1} S (H^{-1})').$$

■

Let

$$\begin{aligned}
\tilde{\mu}_c &= \mu_c \cdot \Pr(T = c) \\
\gamma_c(y) &= y \cdot (p_1(y) - \min(p_1(y), q_1(y))) - y \cdot (p_0(y) - \min(p_0(y), q_0(y))), \\
\gamma_{\lambda_1}(y) &= \min(p_1(y), q_1(y)), \\
h_{\mu_c, P_{1|1}} &= \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot p_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy - \mu_c, \\
h_{\mu_c, P_{1|0}} &= \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy - \int_{y \in \mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot p_1(y) dy - \mu_c.
\end{aligned}$$

Then, H and S are given by

$$H = \begin{pmatrix}
-\Pr(T = c) & \mu_c & h_{\mu_c, P_{1|1}} & h_{\mu_c, P_{1|0}} \\
0 & -1 & \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy & \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \\
0 & 0 & -E(Z) & 0 \\
0 & 0 & 0 & -E(1 - Z)
\end{pmatrix},$$

and

$$S = \begin{pmatrix} S_1 & S_2 \\ S_2' & S_3 \end{pmatrix},$$

where

$$S_1 = \begin{pmatrix}
\int_{\mathcal{Y}} (n \cdot \gamma_c(y) - \tilde{\mu}_c)^2 dy & \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1) \cdot (n \cdot \gamma_{\lambda_1}(y) - \lambda_1) dy \\
\int_{\mathcal{Y}} (n \cdot \gamma_c(y) - \tilde{\mu}_c) \cdot (n \cdot \gamma_{\lambda_1}(y) - \lambda_1) dy & \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1)^2 dy
\end{pmatrix},$$

$$S_2 = \begin{pmatrix} Cov\left((n \cdot \gamma_c(y) - \tilde{\mu}_c), (D - \Pr(D=1|Z=1)) \cdot Z\right) & Cov\left((n \cdot \gamma_c(y) - \tilde{\mu}_c), (D - \Pr(D=1|Z=0)) \cdot (1-Z)\right) \\ Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=1)) \cdot Z\right) & Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D=1|Z=0)) \cdot (1-Z)\right) \end{pmatrix},$$

$$S_3 = \begin{pmatrix} \Pr(D=1|Z=1) \cdot \Pr(D=0|Z=1) \cdot E(Z) & 0 \\ 0 & \Pr(D=1|Z=0) \cdot \Pr(D=0|Z=0) \cdot E(1-Z) \end{pmatrix}.$$

Finally, note that even though one can easily obtain consistent estimators of $\Omega = H^{-1}S(H^{-1})'$ (and therefore of the variances provided in the next section) by taking sample counterparts, it might be preferable to use the bootstrap instead. As it has already been pointed out in Section 4, Chen, Linton, and Keilegom (2003) provide conditions under which the variances of particular two step semiparametric estimators, including MINPIN estimators with i.i.d. observations, can be consistently estimated by bootstrapping, see Theorem B therein. It is easy to see that this is the case under a minor modification of our Assumption E. In particular one needs to replace "in probability" with "almost surely" in E(7) and use the strong rather than the weak law of large numbers. Moreover, E(6) to E(10) must hold in each bootstrap sample. Under these assumptions, $V_{c,d}$, V_c, V_d, and V_{λ_1} can be consistently estimated using the bootstrap.

C Variance formulae

Direct computation of the first and second elements of Ω gives

$$
\begin{aligned}
V_c &= \frac{\int_{\mathcal{Y}} (n\cdot\gamma_c(y)-\tilde{\mu}_c)^2 dy + 2\cdot\mu_c\cdot\int_{\mathcal{Y}}(n\cdot\gamma_c(y)-\tilde{\mu}_c)(n\cdot\gamma_{\lambda_1}(y)-\lambda_1)dy + \mu_c^2\cdot\int_{\mathcal{Y}}(n\cdot\gamma_{\lambda_1}(y)-\lambda_1)^2 dy}{(\Pr(T=c))^2} \\
&+ \frac{2\cdot h_{\mu_c,P_{1|1}}\cdot Cov\left((n\cdot\gamma_c(y)-\tilde{\mu}_c),(D-\Pr(D=1|Z=1))\cdot Z\right)}{(\Pr(T=c))^2\cdot E(Z)} \\
&+ \frac{2\cdot\mu_c\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\le q_1(y)\}\cdot p_1(y)dy\cdot Cov\left((n\cdot\gamma_c(y)-\tilde{\mu}_c),(D-\Pr(D=1|Z=1))\cdot Z\right)}{(\Pr(T=c))^2\cdot E(Z)} \\
&+ \frac{2\cdot\mu_c\cdot h_{\mu_c,P_{1|1}}\cdot Cov\left((n\cdot\gamma_{\lambda_1}(y)-\lambda_1),(D-\Pr(D=1|Z=1))\cdot Z\right)}{(\Pr(T=c))^2\cdot E(Z)} \\
&- \frac{2\cdot\mu_c^2\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\le q_1(y)\}\cdot p_1(y)dy\cdot Cov\left((n\cdot\gamma_{\lambda_1}(y)-\lambda_1),(D-\Pr(D=1|Z=1))\cdot Z\right)}{(\Pr(T=c))^2\cdot E(Z)} \\
&+ \frac{2\cdot\mu_c\cdot h_{\mu_c,P_{1|1}}\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\le q_1(y)\}\cdot p_1(y)dy\cdot\Pr(D=1|Z=1)\cdot\Pr(D=0|Z=1)\cdot E(Z)}{(\Pr(T=c))^2\cdot E(Z)} \\
&- \frac{2\cdot\mu_c^2\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\le q_1(y)\}\cdot p_1(y)dy^2\cdot\Pr(D=1|Z=1)\cdot\Pr(D=0|Z=1)\cdot E(Z)}{(\Pr(T=c))^2\cdot E(Z)} \\
&+ \frac{h_{\mu_c,P_{1|1}}^2\cdot\Pr(D=1|Z=1)\cdot\Pr(D=0|Z=1)\cdot E(Z)}{(\Pr(T=c))^2\cdot E(Z)^2} \\
&+ \frac{2\cdot h_{\mu_c,P_{1|0}}\cdot Cov\left((n\cdot\gamma_c(y)-\tilde{\mu}_c),(D-\Pr(D=1|Z=0))\cdot(1-Z)\right)}{(\Pr(T=c))^2\cdot E(1-Z)} \\
&+ \frac{2\cdot\mu_c\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\ge q_1(y)\}\cdot q_1(y)dy\cdot Cov\left((n\cdot\gamma_c(y)-\tilde{\mu}_c),(D-\Pr(D=1|Z=0))\cdot(1-Z)\right)}{(\Pr(T=c))^2\cdot E(1-Z)} \\
&+ \frac{2\cdot\mu_c\cdot h_{\mu_c,P_{1|0}}\cdot Cov\left((n\cdot\gamma_{\lambda_1}(y)-\lambda_1),(D-\Pr(D=1|Z=0))\cdot(1-Z)\right)}{(\Pr(T=c))^2\cdot E(1-Z)} \\
&- \frac{2\cdot\mu_c^2\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\ge q_1(y)\}\cdot q_1(y)dy\cdot Cov\left((n\cdot\gamma_{\lambda_1}(y)-\lambda_1),(D-\Pr(D=1|Z=0))\cdot(1-Z)\right)}{(\Pr(T=c))^2\cdot E(1-Z)} \\
&+ \frac{2\cdot\mu_c\cdot h_{\mu_c,P_{1|0}}\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\ge q_1(y)\}\cdot q_1(y)dy\cdot\Pr(D=1|Z=0)\cdot\Pr(D=0|Z=0)\cdot E(1-Z)}{(\Pr(T=c))^2\cdot E(1-Z)} \\
&- \frac{2\cdot\mu_c^2\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\ge q_1(y)\}\cdot q_1(y)dy^2\cdot\Pr(D=1|Z=1)\cdot\Pr(D=0|Z=1)\cdot E(Z)}{(\Pr(T=c))^2\cdot E(1-Z)} \\
&+ \frac{h_{\mu_c,P_{1|0}}^2\cdot\Pr(D=1|Z=0)\cdot\Pr(D=0|Z=0)\cdot E(1-Z)}{(\Pr(T=c))^2\cdot E(1-Z)^2},
\end{aligned}
$$

and

$$
\begin{aligned}
V_{\lambda_1} &= \int_{\mathcal{Y}} (n\cdot\gamma_{\lambda_1}(y)-\lambda_1)^2 dy \\
&+ \frac{E(D|Z=1,p_1(y)\le q_1(y))\cdot\Pr(D=1|Z=1)\cdot\Pr(D=0|Z=1)\cdot E(Z)}{E(Z)} \\
&- \frac{2\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\le q_1(y)\}\cdot p_1(y)dy\cdot Cov\left((n\cdot\gamma_{\lambda_1}(y)-\lambda_1),(D-\Pr(D=1|Z=1))\cdot Z\right)}{E(Z)} \\
&+ \frac{\int_{y\in\mathcal{Y}} I\{p_1(y)\ge q_1(y)\}\cdot q_1(y)dy\cdot\Pr(D=1|Z=0)\cdot\Pr(D=0|Z=0)\cdot E(1-Z)}{E(1-Z)} \\
&- \frac{2\cdot\int_{y\in\mathcal{Y}} I\{p_1(y)\ge q_1(y)\}\cdot q_1(y)dy\cdot Cov\left((n\cdot\gamma_{\lambda_1}(y)-\lambda_1),(D-\Pr(D=1|Z=0))\cdot(1-Z)\right)}{E(1-Z)}.
\end{aligned}
$$

By replacing $m(W_i, \theta, \tau(W_i))$ with

$$m_d(W_i, \theta, \tau(W_i)) = \begin{pmatrix} n \cdot \gamma_c(W_i) - \mu_d \cdot (P_{1|1} - \lambda_1) \\ n \cdot \gamma_{\lambda_1}(W_i) - \lambda_1 \\ (D_i - P_{1|1}) \cdot Z_i \\ (D_i - P_{1|0}) \cdot (1 - Z_i) \end{pmatrix},$$

it can be shown that

$$\begin{aligned}
V_d = & \frac{\int_{\mathcal{Y}} (n \cdot \gamma_d(y) - \tilde{\mu}_d)^2 dy + 2 \cdot \mu_d \cdot \int_{\mathcal{Y}} (n \cdot \gamma_d(y) - \tilde{\mu}_d)(n \cdot \gamma_{\lambda_1}(y) - \lambda_1) dy + \mu_d^2 \cdot \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1)^2 dy}{(\Pr(T = d))^2} \\
& + \frac{2 \cdot h_{\mu_d, P_{1|1}} \cdot Cov\left((n \cdot \gamma_d(y) - \tilde{\mu}_d), (D - \Pr(D = 1|Z = 1)) \cdot Z\right)}{(\Pr(T = d))^2 \cdot E(Z)} \\
& + \frac{2 \cdot \mu_d \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot Cov\left((n \cdot \gamma_d(y) - \tilde{\mu}_d), (D - \Pr(D = 1|Z = 1)) \cdot Z\right)}{(\Pr(T = d))^2 \cdot E(Z)} \\
& + \frac{2 \cdot \mu_d \cdot h_{\mu_d, P_{1|1}} \cdot Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 1)) \cdot Z\right)}{(\Pr(T = d))^2 \cdot E(Z)} \\
& - \frac{2 \cdot \mu_d^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 1)) \cdot Z\right)}{(\Pr(T = d))^2 \cdot E(Z)} \\
& + \frac{2 \cdot \mu_d \cdot h_{\mu_d, P_{1|1}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = d))^2 \cdot E(Z)} \\
& - \frac{2 \cdot \mu_d^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy^2 \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = d))^2 \cdot E(Z)} \\
& + \frac{h_{\mu_d, P_{1|1}}^2 \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = d))^2 \cdot E(Z)^2} \\
& + \frac{2 \cdot h_{\mu_d, P_{1|0}} \cdot Cov\left((n \cdot \gamma_d(y) - \tilde{\mu}_d), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z)\right)}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\
& + \frac{2 \cdot \mu_d \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot Cov\left((n \cdot \gamma_d(y) - \tilde{\mu}_d), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z)\right)}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\
& + \frac{2 \cdot \mu_d \cdot h_{\mu_d, P_{1|0}} \cdot Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z)\right)}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\
& - \frac{2 \cdot \mu_d^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z)\right)}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\
& + \frac{2 \cdot \mu_d \cdot h_{\mu_d, P_{1|0}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot \Pr(D = 1|Z = 0) \cdot \Pr(D = 0|Z = 0) \cdot E(1 - Z)}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\
& - \frac{2 \cdot \mu_d^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy^2 \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = d))^2 \cdot E(1 - Z)} \\
& + \frac{h_{\mu_d, P_{1|0}}^2 \cdot \Pr(D = 1|Z = 0) \cdot \Pr(D = 0|Z = 0) \cdot E(1 - Z)}{(\Pr(T = d))^2 \cdot E(1 - Z)^2}.
\end{aligned}$$

Similarly by replacing $m(W_i, \theta, \tau(W_i))$ with

$$m_{c,d}(W_i, \theta, \tau(W_i)) = \begin{pmatrix} n \cdot \gamma_c(W_i) - \mu_{c,d} \cdot (P_{1|1} - \lambda_1) \\ n \cdot \gamma_{\lambda_1}(W_i) - \lambda_1 \\ (D_i - P_{1|1}) \cdot Z_i \\ (D_i - P_{1|0}) \cdot (1 - Z_i) \end{pmatrix},$$

it can be shown that

$$\begin{aligned}
V_{c,d} &= \frac{\int_{\mathcal{Y}} (n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d})^2 dy + 2 \cdot \mu_{c,d} \cdot \int_{\mathcal{Y}} (n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d})(n \cdot \gamma_{\lambda_1}(y) - \lambda_1) dy + \mu_{c,d}^2 \cdot \int_{\mathcal{Y}} (n \cdot \gamma_{\lambda_1}(y) - \lambda_1)^2 dy}{(\Pr(T = c) + \Pr(T = d))^2} \\
&+ \frac{2 \cdot h_{\mu_{c,d}, P_{1|1}} \cdot Cov\left((n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d}), (D - \Pr(D = 1|Z = 1)) \cdot Z\right)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(Z)} \\
&+ \frac{2 \cdot \mu_{c,d} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot Cov\left((n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d}), (D - \Pr(D = 1|Z = 1)) \cdot Z\right)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(Z)} \\
&+ \frac{2 \cdot \mu_{c,d} \cdot h_{\mu_{c,d}, P_{1|1}} \cdot Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 1)) \cdot Z\right)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(Z)} \\
&- \frac{2 \cdot \mu_{c,d}^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 1)) \cdot Z\right)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(Z)} \\
&+ \frac{2 \cdot \mu_{c,d} \cdot h_{\mu_{c,d}, P_{1|1}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(Z)} \\
&- \frac{2 \cdot \mu_{c,d}^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \leq q_1(y)\} \cdot p_1(y) dy^2 \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(Z)} \\
&+ \frac{h_{\mu_{c,d}, P_{1|1}}^2 \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(Z)^2} \\
&+ \frac{2 \cdot h_{\mu_{c,d}, P_{1|0}} \cdot Cov\left((n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d}), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z)\right)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(1 - Z)} \\
&+ \frac{2 \cdot \mu_{c,d} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot Cov\left((n \cdot \gamma_{c,d}(y) - \tilde{\mu}_{c,d}), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z)\right)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(1 - Z)} \\
&+ \frac{2 \cdot \mu_{c,d} \cdot h_{\mu_{c,d}, P_{1|0}} \cdot Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z)\right)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(1 - Z)} \\
&- \frac{2 \cdot \mu_{c,d}^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot Cov\left((n \cdot \gamma_{\lambda_1}(y) - \lambda_1), (D - \Pr(D = 1|Z = 0)) \cdot (1 - Z)\right)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(1 - Z)} \\
&+ \frac{2 \cdot \mu_{c,d} \cdot h_{\mu_{c,d}, P_{1|0}} \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy \cdot \Pr(D = 1|Z = 0) \cdot \Pr(D = 0|Z = 0) \cdot E(1 - Z)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(1 - Z)} \\
&- \frac{2 \cdot \mu_{c,d}^2 \cdot \int_{y \in \mathcal{Y}} I\{p_1(y) \geq q_1(y)\} \cdot q_1(y) dy^2 \cdot \Pr(D = 1|Z = 1) \cdot \Pr(D = 0|Z = 1) \cdot E(Z)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(1 - Z)} \\
&+ \frac{h_{\mu_{c,d}, P_{1|0}}^2 \cdot \Pr(D = 1|Z = 0) \cdot \Pr(D = 0|Z = 0) \cdot E(1 - Z)}{(\Pr(T = c) + \Pr(T = d))^2 \cdot E(1 - Z)^2},
\end{aligned}$$

where

$$
\begin{aligned}
\tilde{\mu}_{c,d} &= \mu_{c,d} \cdot (\Pr(T=c) + \Pr(T=d)), \\
\tilde{\mu}_{d} &= \mu_{d} \cdot \Pr(T=d),
\end{aligned}
$$

$$
\begin{aligned}
\gamma_{c,d}(y) &= y \cdot (\max(p_1(y), q_1(y)) - \min(p_1(y), q_1(y))) - y \cdot (\max(p_0(y), q_0(y)) - \min(p_0(y), q_0(y))), \\
\gamma_{d}(y) &= y \cdot (q_1(y) - \min(p_1(y), q_1(y))) - y \cdot (q_0(y) - \min(p_0(y), q_0(y))),
\end{aligned}
$$

$$
\begin{aligned}
h_{\mu_{c,d},P_{1|1}} &= \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot p_1(y)dy - \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot q_1(y)dy \\
&+ \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot q_1(y)dy - \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot p_1(y)dy - \mu_{c,d}, \\
h_{\mu_{c,d},P_{1|0}} &= \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot q_1(y)dy - \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \geq q_1(y)\} \cdot p_1(y)dy \\
&+ \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot p_1(y)dy - \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot q_1(y)dy - \mu_{c,d},
\end{aligned}
$$

$$
\begin{aligned}
h_{\mu_{d},P_{1|1}} &= \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot q_1(y)dy - \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot p_1(y)dy - \mu_{d}, \\
h_{\mu_{d},P_{1|1}} &= \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot p_1(y)dy - \int_{y\in\mathcal{Y}} y \cdot I\{p_1(y) \leq q_1(y)\} \cdot q_1(y)dy - \mu_{d}.
\end{aligned}
$$

References

Abadie, A., J. Angrist, and G. W. Imbens (2002): "Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings," *Econometrica*, 70, 91–117.

Andrews, D. W. (1991): "Asymptotics for Kernel-Based Non-Orthogonal Semiparametric Estimators," *mimeo*.

——— (1994a): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica*, 62(1), 43–72.

——— (1994b): "Empirical process methods in econometrics," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. McFadden, pp. 2247–2294. Elsevier.

——— (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–586.

Angrist, J., G. Imbens, and D. Rubin (1996): "Identification of Causal Effects using Instrumental Variables," *Journal of American Statistical Association*, 91, 444–472 (with discussion).

Angrist, J., and A. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?," *Quarterly Journal of Economics*, 106, 979–1014.

——— (1992): "The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples," *Journal of the American Statistical Association*, 87, 328–336.

Barua, R., and K. Lang (2009): "School Entry, Educational Attainment, and Quarter of Birth: A Cautionary Tale of LATE," *NBER Working Paper 15236*.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association*, 90, 443–450.

Card, D. (1999): "The Causal Effect of Education on Earnings," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, pp. 1802–1863. North-Holland, Amsterdam.

Chen, X., O. Linton, and I. V. Keilegom (2003): "Estimation of Semiparametric Models When the Criterion Function Is Not Smooth," *Econometrica*, 71, 1591–1608.

Cheng, M.-Y., J. Fan, and J. S. Marron (1997): "On automatic boundary corrections," *Annals of Statistics*, 25, 1691–1708.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261.

Dai, J., and S. Sperlich (2010): "Simple and effective boundary correction for kernel densities and regression with an application to the world income and Engel curve estimation," *Computational Statistics & Data Analysis*, 54, 2487 – 2497.

de Chaisemartin, C., and X. D'Haultfoeuilley (2012): "LATE again, with defiers," *mimeo, CREST*.

Frölich, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139, 35–75.

Gautier, E., and S. Hoderlein (2012): "Estimating Treatment Effects with Random Coefficients in the Selection Equation," *mimeo, Boston College*.

Huber, M., and G. Mellace (2010): "Sharp IV bounds on average treatment effects under endogeneity and noncompliance," *University of St Gallen, Dept. of Economics Discussion Paper no. 2010-31*.

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

Imbens, G. W., and D. Rubin (1997): "Estimating outcome distributions for compliers in instrumental variables models," *Review of Economic Studies*, 64, 555–574.

Jones, M. C. (1993): "Simple boundary correction for kernel density estimation," *Statistics and Computing*, 3, 135–146.

Kitagawa, T. (2009): "Identification Region of the Potential Outcome Distribution under Instrument Independence," *CeMMAP working paper 30/09*.

Klein, T. J. (2010): "Heterogeneous treatment effects: Instrumental variables without monotonicity?," *Journal of Econometrics*, 155, 99–116.

Li, Q., and J. S. Racine (2007): *Nonparametric econometrics: theory and practice*. Princeton University Press, Princeton and Oxford.

MacKinnon, J. G. (2006): "Bootstrap Methods in Econometrics," *The Economic Record*, 82, S2–S18.

Manski, C. F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review, Papers and Proceedings*, 80, 319–323.

Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Roy, A. (1951): "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 3, 135–146.

Rubin, D. B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Silverman, B. (1986): *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Small, D. S., and Z. Tan (2007): "A Stochastic Monotonicity Assumption for the Instrumental Variables Method," *Technical report, Department of Statistics, Wharton School, University of Pennsylvania*.

Torgovitsky, A. (2011): "Identification of Nonseparable Models with General Instruments," *Working paper, Yale University.*

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331–341.