Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 139 (2007) 35–75
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

Nonparametric IV estimation of local average treatment effects with covariates

Markus Frölich[a,b,*,1]

[a]University College London, UK
[b]Universität St.Gallen, Bodanstrasse 8, SIAW, 9000 St. Gallen, Switzerland

Available online 4 August 2006

Abstract

In this paper nonparametric instrumental variable estimation of local average treatment effects (LATE) is extended to incorporate covariates. Estimation of LATE is appealing since identification relies on much weaker assumptions than the identification of average treatment effects in other nonparametric instrumental variable models. Including covariates in the estimation of LATE is necessary when the instrumental variable itself is confounded, such that the IV assumptions are valid only conditional on covariates. Previous approaches to handle covariates in the estimation of LATE relied on parametric or semiparametric methods. In this paper, a nonparametric estimator for the estimation of LATE with covariates is suggested that is root-n asymptotically normal and efficient.

JEL classification: C13; C14

Keywords: Instrumental variables; LATE; Evaluation; Treatment effect; Matching; Unobserved heterogeneity

1. Introduction

In this paper the concept of the local average treatment effect (LATE) is extended to incorporate covariates and a $\sqrt{n}$ consistent nonparametric instrumental variables estimator is proposed.

*Corresponding author at: Universität St.Gallen, Bodanstrasse 8, SIAW, 9000 St. Gallen, Switzerland. Tel.: 41 71 224 2329; fax:41 71 224 2298.
E-mail address: markus.froelich@unisg.ch.
URL: http://www.siaw.unisg.ch/froelich.
[1]The author is also affiliated with the Institute for the Study of Labor (IZA), Bonn.

0304-4076/$ - see front matter
doi:10.1016/j.jeconom.2006.06.004

Instrumental variables regression is a fundamental approach to causal reasoning in econometrics. In many applications one wants to uncover the *causal* relationship between a variable D and an outcome variable Y, where the variable D is itself endogenous. For example, if D is years of schooling and Y is wages, it is of interest to learn by how much wages increase due to an additional year of schooling. In another example, if D represents participation in a training programme and Y is subsequent employment status, it is of interest how the employment probability is affected by participation in the training programme. In these examples, one would like to know how D causally affects Y, i.e. how an *exogenous* variation in D would change the variable Y. Since the variable D is endogenous, a regression of Y on D does not uncover a causal relationship. Nevertheless, if a variable Z exists that affects only D but not Y, then an exogenous variation in Z induces an exogenous variation in D and thus overcomes the endogeneity of D. Such a variable Z is called an instrumental variable and has been exploited in numerous studies to identify the effects of D on Y.

If the values of the instrumental variable Z are assigned completely at random, any variation in Z is exogenous and thus generates an unconfounded variation in D to identify the relationship between D and Y. For example, Hearst et al. (1986) and Angrist (1990) use the Vietnam era conscription lottery as an instrument to identify the effects of mandatory military conscription on subsequent civilian mortality and earnings. Imbens et al. (2001) use 'winning a prize in the lottery' as an instrument to identify the effects of unearned income on subsequent labour supply, earnings and consumption behaviour. In both examples the instrument is randomly assigned (by a lottery).

However, in many applications the instrument Z is not randomly assigned and may be confounded with D or Y. For example, college proximity may be used as an instrument to identify the returns to schooling, noting that living close to a college during childhood may induce some children to go to college but is unlikely to directly affect the wages earned in their adulthood (Card, 1995). Nevertheless, whether living close to or far from a college is not random but an active choice by their parents. Their choice, however, might itself be related to characteristics that affect their children's subsequent wages directly. Parental education is another example of an instrumental variable that is often used to identify the returns to schooling. It may appear reasonable to assume that parental schooling itself has no direct impact on their children's wages. Nevertheless, it is likely to be correlated with parents' profession, family income and wealth, which may directly affect the wage prospects of their offspring. In these cases, Z may only be a proper instrumental variable after conditioning on some covariates X.

Conventional approaches to accommodate covariates X in instrumental variables estimation (such as two-stage least squares) proceed by specifying functional form restrictions on the conditional expectation functions of Y and D. Recently, *nonparametric* identification and estimation in instrumental variable models, avoiding such delicate functional form assumptions, has received a lot of interest.[2] However, their approaches still impose identifying assumptions which may not be satisfied in many applications. Most models suppose *additive separability* in the error term, which amounts to assuming that, conditional on X, the relationship between D and Y is identical for each individual up to an intercept. In other words, a *constant treatment effect* for individuals with the same value of X is assumed. Additively separable models, thus, rule out unobserved heterogeneity and

[2]See Blundell and Powell (2003), Chesher (2003, 2005), Das (2005), Darolles et al. (2001), Florens (2003), Florens et al. (2002), Imbens and Newey (2003), Newey et al. (1999) and Newey and Powell (2003).

therefore may not be appropriate in many applications. Non-separable models permit unobserved heterogeneity in the treatment effects. Identification, however, often relies on an instrument that is sufficiently powerful to move the value of D_i (for any individual i) over the entire support of the variable D, see Blundell and Powell (2003), Florens et al. (2002) and Imbens and Newey (2003).[3] Yet, such powerful instruments are often not available. In this case, the relationship between D and Y can only be uncovered for the subpopulation that reacts on changes of the instrument Z.

This is the concept of the LATE of Imbens and Angrist (1994). The LATE is the mean effect on Y of a change in D for the subpopulation of compliers, where the compliers are all individuals whose value of D would change if the instrument Z were modified exogenously. In spite of its appealing properties, however, the LATE-concept has not been fully extended to accommodate covariates X. Previous approaches to incorporate covariates for estimating LATE, e.g. Abadie (2003), Angrist et al. (2000), Hirano et al. (2000), Yau and Little (2001), resorted to parametric or semiparametric estimation approaches.

In this paper nonparametric identification and estimation of LATE with covariates is analysed. A nonparametric estimator in the form of a ratio of two matching estimators is shown to be $\sqrt{n}$-consistent and efficient. Thus, using nonparametric regression to accommodate covariates X in the estimation of LATE does not give rise to a curse of dimensionality.

Section 2 discusses nonparametric estimation of the LATE with covariates. Section 3 proposes a nonparametric estimator and derives its properties. Section 4 discusses propensity score versions of the LATE estimator. Section 5 examines extensions to non-binary treatments and non-binary instruments. Section 6 concludes.

2. Instrumental variables regression and nonparametric LATE

Consider the triangular model

$$Y_i = \varphi(D_i, Z_i, X_i, \varepsilon_i), \tag{1}$$

$$D_i = \zeta(Z_i, X_i, v_i), \tag{2}$$

where Y is an outcome variable and D is an endogenous regressor. We are interested in estimating the causal effect of D on Y. The variable Z will be referred to as an instrumental variable. ε and v are unobserved variables. X is a vector of observed covariates. It will be assumed that conditional on X certain instrumental variables assumptions (specified below) are satisfied. Hence X contains all variables (if any) that are necessary to make these assumptions hold.

For a more compact notation, define

$$Y^d_{i,z} \equiv \varphi(d, z, X_i, \varepsilon_i),$$

$$D_{i,z} \equiv \zeta(z, X_i, v_i),$$

and notice that $Y^{D_i}_{i,Z_i} = Y_i$ and $D_{i,Z_i} = D_i$ are the observed values of Y and D.

[3]This is similar to the identification-at-infinity argument in selection models. These models largely rely on some kind of monotonicity assumption in the relationship determining the endogenous regressor. Chesher (2003) assumes monotonicity in the outcome equation and in the relationship determining the endogenous regressor. He studies identification of systems of equations and identification under local exclusion restrictions. Chesher (2005) extends this to discrete endogenous regressors. Chernozhukov and Hansen (2005) rely on monotonicity in the outcome equation.

The causal interpretation of φ is as follows: for unit i the variables Z_i, ε_i, v_i and X_i are determined by nature and D_i and Y_i are determined according to (2) and (1), respectively. If the variable D_i were manipulated by an external intervention to take the value d (without affecting the values of Z_i, ε_i, v_i or X_i), the outcome $\varphi(d, Z_i, X_i, \varepsilon_i)$ would be observed. For example, if $D \in \{0, 1\}$ is union status and Y is wages, the *potential wages* for person i are the wages that person i would receive if union membership were set for this person by some external intervention to 0 or to 1, respectively. Analogously, if D is years of schooling, the potential wages for person i are the wages that person i would receive if years of schooling were set externally to different levels. Differences between potential outcomes are also called *causal effects* (Rubin, 1974). This causal interpretation of φ requires that Y is not a cause to any of the other variables and that D is a cause only to Y, see Pearl (2000).[4]

For causal inference, we are interested in estimating the average structural function (Blundell and Powell, 2003)

$$\mathrm{E}[Y^d] = \int \varphi(d, Z, X, \varepsilon)\, \mathrm{d}F_{Z,X,\varepsilon}$$

or the average treatment effect

$$\mathrm{E}[Y^{d'} - Y^d] = \int (\varphi(d', Z, X, \varepsilon) - \varphi(d, Z, X, \varepsilon))\, \mathrm{d}F_{Z,X,\varepsilon}$$

or the marginal average treatment effect $\partial \mathrm{E}[Y^d]/\partial d$ if D is continuous.

Nonparametric instrumental variable estimation of the relationship between an endogenous variable D and an outcome variable Y is often analysed in an *additively separable* model where it is assumed that[5]

$$Y_i = \varphi(D_i, X_i) + \varepsilon_i.$$

However, the assumption of an additively separable structure is often inadequate, since it heavily restricts the permitted heterogeneity among persons: the causal effect of setting D_i to d_1 versus setting D_i to d_2 is supposed to be identical for all persons with the same x value. This amounts to assuming that the effect of union membership or of an additional year of schooling is identical for all persons with the same X characteristics. This *constant treatment effect assumption* (conditional on X) is in many situations rather implausible (Heckman, 1997). For example, the return to schooling may interact with unobserved ability.

Identification in non-separable models is analysed by Blundell and Powell (2003), Chesher (2003, 2005), Florens et al. (2002) and Imbens and Newey (2003). Suppose that the instrumental variable Z has no direct effect on Y conditional on X and ε

$$Y_i = \varphi(D_i, X_i, \varepsilon_i)$$

and that the selection equation $\zeta(Z_i, X_i, v_i)$ is either additive in Z_i and v_i (Florens et al., 2002) or strictly monotone in v_i (Imbens and Newey, 2003). By imposing a normalization on v_i, such as being $\text{uniformly}_{[0,1]}$ distributed, v_i can be identified from D_i, Z_i

[4]If D is also a cause to X, in the sense that changing D would imply also a change in X, only the direct effect of D on Y would be recovered with the following identification strategy, but not the total effect (i.e. including the effect of D on Y that is channelled through X).

[5]See Blundell and Powell (2003), Das (2005), Darolles et al. (2001), Florens (2003), Newey et al. (1999) and Newey and Powell (2003).

and X_i.[6] If one assumes that the endogeneity of D_i (i.e. the difference between $\mathrm{E}[Y^d]$ and $\mathrm{E}[Y^d|D=d]$) is generated entirely through the error term v_i, the endogeneity can be controlled for by conditioning on v

$$\mathrm{E}[\varphi(d,X,\varepsilon)|X,V,D=d]=\mathrm{E}[\varphi(d,X,\varepsilon)|X,V]. \tag{3}$$

With this assumption, the average structural function can be identified by

$$\begin{aligned}\int \mathrm{E}[Y|X,V,D=d]\,\mathrm{d}F_{X,V} &= \int \mathrm{E}[\varphi(D,X,\varepsilon)|X,V,D=d]\,\mathrm{d}F_{X,V}\\ &= \int \mathrm{E}[\varphi(d,X,\varepsilon)|X,V]\,\mathrm{d}F_{X,V}=\mathrm{E}[Y^d].\end{aligned}$$

However, a central condition for identification is that the conditional expectation $\mathrm{E}[Y|X,V,D=d]$ is defined at every v in the support of V (conditional on X). This requires that the support of the distribution of V given $D=d$ is the same as the support of the marginal distribution of V (conditional on X). For this condition to be satisfied, either the distribution of V needs to be restricted, or the instrument Z_i must be sufficiently powerful to move the regressor D_i for every unit i to any value d where $\mathrm{E}[Y^d]$ shall be estimated (Imbens and Newey, 2003).

However, in many applications the instruments available are not so powerful. It is often highly unreasonable to assume that *all* units i could be induced, through a modification of the instrument Z_i, to change D_i to a particular value. Consider the situation where D and Z are binary. The above assumption would require that all units switch D from 0 to 1 or vice versa if Z is changed from 0 to 1. If this assumption does not hold, the relationship between D and Y can no longer be analysed for the full population, but only for the *subpopulation* which reacts to the instrumental variable. This is the concept of the LATE, which restricts the analysis to the subpopulation of units that could be induced to change D through a variation in the instrumental variable.

The LATE has been introduced by Imbens and Angrist (1994) and further been analysed in several papers.[7] Most of the discussion on LATE focuses on the case where Z is a proper instrumental variable without conditioning on X covariates. Often, however, the instrumental variable assumptions are only satisfied conditional on covariates. Identification of LATE with covariates X is discussed in Angrist and Imbens (1995), Heckman and Vytlacil (1999), Abadie (2003) and Imbens (2001). Covariates are usually included in the estimation of LATE via parametric or semiparametric estimation approaches.[8] However, *nonparametric* estimation of LATE with confounding covariates X has not been analysed so far, which is the aim of this paper.

[6] If $\zeta(Z_i,X_i,v_i)$ is monotone in v_i, then $v_i=F_{D|Z,X}(D_i,Z_i,X_i)$, see Imbens and Newey (2003). If $\zeta(Z_i,X_i,v_i)=\xi(Z_i,X_i)+v_i$ is additive, then $v_i=D_i-\mathrm{E}[D|Z_i,X_i]$, see Blundell and Powell (2003) and Florens et al. (2002).

[7] See Angrist and Imbens (1995), Angrist et al. (1996), Imbens and Rubin (1997), Heckman and Vytlacil (1999), Abadie (2003) and Imbens (2001), among others.

[8] Angrist et al. (2000) and Yau and Little (2001) incorporate covariates by assuming that they enter linearly in the conditional expectation functions. Hirano et al. (2000) suggest to model the probability of being an always-taker, never-taker or a complier given covariates X by a trinomial logistic distribution and to model the outcome distributions separately for these types. Abadie (2003) initially introduces covariates in a nonparametric way but proposes parametric and semiparametric estimation methods to avoid the curse of dimensionality of nonparametric regression.

Consider first the case where the endogenous regressor $D \in \{0,1\}$ and the instrument $Z \in \{0,1\}$ are both binary. For example, D could be attending/not attending college and Z could be living close to or far from a college. (Extensions to non-binary D and non-binary Z are discussed in Section 5.)

According to the reaction of D on an external intervention on Z in Eq. (2), the units i can be distinguished into different types: For some units, D would remain unchanged if Z were changed from 0 to 1, whereas for others D would change. With D and Z binary, four different types $\tau \in \{n, c, d, a\}$ are possible:

$\tau_i = n$	if $D_{i,0} = 0$ and $D_{i,1} = 0$	Never-taker
$\tau_i = c$	if $D_{i,0} = 0$ and $D_{i,1} = 1$	Complier
$\tau_i = d$	if $D_{i,0} = 1$ and $D_{i,1} = 0$	Defier
$\tau_i = a$	if $D_{i,0} = 1$ and $D_{i,1} = 1$	Always-taker

Since the units of type always-taker and of type never-taker cannot be induced to change D through a variation in the instrumental variable, the impact of D on Y can at most be ascertained for the subpopulations of compliers and defiers. Under certain assumptions given below, the LATE γ for the subpopulation of compliers

$$\gamma = \mathrm{E}[Y^1 - Y^0 | \tau = c] = \int (\varphi(1, Z, X, \varepsilon) - \varphi(0, Z, X, \varepsilon)) \, \mathrm{d}F_{Z,X,\varepsilon|\tau=c} \tag{4}$$

is identified. This is the effect for the largest subpopulation for which an effect can be identified without parametric or constant treatment effect assumptions.

Define also the treatment effect for the compliers with characteristics x as

$$\gamma(x) = \mathrm{E}[Y^1 - Y^0 | X = x, \tau = c].$$

To identify the LATE γ, consider the model (1) and (2) and assume that:

Assumption 1 (*Monotonicity*). The subpopulation of defiers has probability measure zero:

$$\mathrm{P}(D_{i,0} > D_{i,1}) = 0.$$

Assumption 2 (*Existence of compliers*). The subpopulation of compliers has positive probability:

$$\mathrm{P}(D_{i,0} < D_{i,1}) > 0.$$

Assumption 3 (*Unconfounded type*). The relative size of the subpopulations always-takers, never-takers and compliers is independent of the instrument: for all $x \in Supp(X)$

$$\mathrm{P}(\tau_i = t | X_i = x, Z_i = 0) = \mathrm{P}(\tau_i = t | X_i = x, Z_i = 1) \quad \text{for } t \in \{a, n, c\}.$$

Assumption 4 (*Mean exclusion restriction*). The potential outcomes are mean independent of the instrumental variable Z in each subpopulation: for all $x \in Supp(X)$

$$\mathrm{E}[Y^0_{i,Z_i} | X_i = x, Z_i = 0, \tau_i = t] = \mathrm{E}[Y^0_{i,Z_i} | X_i = x, Z_i = 1, \tau_i = t] \quad \text{for } t \in \{n, c\},$$

$$\mathrm{E}[Y^1_{i,Z_i} | X_i = x, Z_i = 0, \tau_i = t] = \mathrm{E}[Y^1_{i,Z_i} | X_i = x, Z_i = 1, \tau_i = t] \quad \text{for } t \in \{a, c\}.$$

Assumption 5 (*Common support*). The support of X is identical in both subpopulations:

$$Supp(X|Z = 1) = Supp(X|Z = 0).$$

An equivalent representation of the common support condition is that $0<\pi(x)<1\ \forall x$ with $f_x(x)>0$, where $\pi(x)=\mathrm{P}(Z=1|X=x)$.

Assumptions 1 and 2 rule out the existence of subpopulations that are affected by the instrument in an opposite direction. Monotonicity ensures that the effect of Z on D has the same direction for all units. For college proximity as an instrument to identifying the returns to attending college, monotonicity requires that any child which would not have attended college if living close to a college, would also not have done so if living far from a college. The existence assumption requires that the college attendance decision depends, for at least some children, on the proximity to the nearest college.

Assumption 3 allows to identify the effect of Z on D and to estimate the fraction of compliers. Without conditioning on covariates X this assumption may often be invalidated because of selection effects. For example, parents who would like to see their children attending college but fear that they might not want to attend if living too far away, might decide to reside closer to a college. In this case, the subpopulation living close to a college would contain a higher fraction of compliers than those living far away. Validity of Assumption 3 requires that the vector X contains all variables that affect the choice of residence Z as well as the type τ (which is determined by $D_{i,0}$ and $D_{i,1}$).

Assumption 4 rules out a direct effect of Z on Y. Conditional on X, any effect of Z should be channelled through D such that the potential outcomes are not correlated with the instrument. Without conditioning on X, this assumption may often be invalid, for example, if college proximity itself has a direct effect on the child's wages in its later career or if the families who decided to reside close to a college are different from those who decided to live far from a college.

Assumption 5 requires that for any value of X (in its support) both values of the instrument Z can be observed.

Theorem 1 (*Identification of LATE*). *Under Assumptions* 1–5 *and supposing that* $\mathrm{E}[Y]<\infty$, *the LATE* γ *for the subpopulation of compliers is nonparametrically identified as*

$$\gamma=\mathrm{E}[Y^1-Y^0|\tau=c]=\frac{\int(\mathrm{E}[Y|X=x,Z=1]-\mathrm{E}[Y|X=x,Z=0])f_x(x)\,\mathrm{d}x}{\int(\mathrm{E}[D|X=x,Z=1]-\mathrm{E}[D|X=x,Z=0])f_x(x)\,\mathrm{d}x}. \tag{5}$$

To prove Theorem 1 note that the size of the complier-subpopulation with characteristics x is identified as

$$\mathrm{P}(\tau=c|X=x)=\mathrm{E}[D|X=x,Z=1]-\mathrm{E}[D|X=x,Z=0], \tag{6}$$

and that for all x with $\mathrm{P}(\tau=c|X=x)>0$ the treatment effect $\gamma(x)$ is identified as

$$\gamma(x)=\mathrm{E}[Y^1_{i,Z_i}-Y^0_{i,Z_i}|X_i=x,\tau_i=c]=\frac{\mathrm{E}[Y|X=x,Z=1]-\mathrm{E}[Y|X=x,Z=0]}{\mathrm{E}[D|X=x,Z=1]-\mathrm{E}[D|X=x,Z=0]}, \tag{7}$$

see Appendix.

$\gamma(x)$ identifies the treatment effect for the compliers with characteristics $X=x$. More interesting, however, would often be an estimate of the average treatment effect for the subpopulation of *all compliers*, which is the largest subpopulation for which a treatment effect is identified.

If Assumptions 1–4 are valid without conditioning on X, the LATE γ would be identified by $(\mathrm{E}[Y|Z=1]-\mathrm{E}[Y|Z=0])/(\mathrm{E}[D|Z=1]-\mathrm{E}[D|Z=0])$. If conditioning on X is

necessary, the conditional effects $\gamma(x)$ need to be weighted by the distribution of x in the all-compliers subpopulation to obtain the average treatment effect γ for all compliers:

$$\gamma = \mathrm{E}[Y^1_{i,Z_i} - Y^0_{i,Z_i} | \tau_i = c] = \int \gamma(x)\,\mathrm{d}F_{x|\tau=c},$$

where $F_{x|\tau=c}$ denotes the distribution function of X in the subpopulation of all compliers.[9]

Since the subpopulation of compliers is not identified, $\mathrm{d}F_{x|\tau=c}$ is also not identified. However, by Bayes' theorem $\mathrm{d}F_{x|\tau=c} = (\mathrm{P}(\tau = c|X = x)/\mathrm{P}(\tau = c))\,\mathrm{d}F_x$, which gives

$$\gamma = \int \gamma(x) \cdot \frac{\mathrm{P}(\tau = c|X = x)}{\mathrm{P}(\tau = c)}\,\mathrm{d}F_x.$$

Inserting (7) and noting that the fraction of compliers corresponds to (6) gives

$$\gamma = \frac{\int (\mathrm{E}[Y|X = x, Z = 1] - \mathrm{E}[Y|X = x, Z = 0]) f_x(x)\,\mathrm{d}x}{\mathrm{P}(\tau = c)}.$$

With $\mathrm{P}(\tau = c) = \int \mathrm{P}(\tau = c|X = x)\,\mathrm{d}F_x$ and using (6), the LATE on all compliers is given by (5). By Assumption 5 the conditional expectations are identified in the $Z = 1$ and 0 subpopulations.

3. Nonparametric conditional LATE estimation

In this section nonparametric estimation of the LATE γ is discussed. Define the conditional mean functions $m_z(x) = \mathrm{E}[Y|X = x, Z = z]$ and $\mu_z(x) = \mathrm{E}[D|X = x, Z = z]$ and let $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ be corresponding nonparametric regression estimators thereof. A nonparametric imputation estimator of γ is

$$\frac{\sum_i (\hat{m}_1(X_i) - \hat{m}_0(X_i))}{\sum_i (\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))},$$

where the expected values $\mathrm{E}[Y|X, Z]$ and $\mathrm{E}[D|X, Z]$ are imputed for each observation X_i.

Using the observed values Y_i and D_i as estimates of $\mathrm{E}[Y_i|X_i, Z = z]$ and $\mathrm{E}[D_i|X_i, Z = z]$, whenever $z = Z_i$, gives the conditional LATE estimator $\hat{\gamma}$ as

$$\hat{\gamma} = \frac{\sum_{i:Z_i=1} (Y_i - \hat{m}_0(X_i)) - \sum_{i:Z_i=0} (Y_i - \hat{m}_1(X_i))}{\sum_{i:Z_i=1} (D_i - \hat{\mu}_0(X_i)) - \sum_{i:Z_i=0} (D_i - \hat{\mu}_1(X_i))}. \tag{8}$$

The estimator $\hat{\gamma}$ corresponds to a ratio of two *matching estimators*, which are frequently used in treatment evaluation to estimate average treatment effects when the endogeneity of the regressor D can be completely controlled for by observed covariates (Angrist and Krueger, 1999; Heckman et al., 1999).

To analyse the properties of $\hat{\gamma}$, it is useful to first derive the semiparametric efficiency bound for the estimation of the LATE γ. In the following theorems it is supposed that X is continuous. X could easily be allowed to contain both continuous and discrete regressors or only discrete regressors, at the expense of a more complex notation.

[9] An alternative approach would be to weight $\gamma(x)$ by the distribution $\mathrm{d}F_x$, as has been suggested for example by Das (2005) and others. Unless $\gamma(x)$ is constant or $\mathrm{P}(\tau = c|X) = \mathrm{P}(\tau = c)$, the weighting $\int \gamma(x)\,\mathrm{d}F_x = \int ((\mathrm{E}[Y|X = x, Z = 1] - \mathrm{E}[Y|X = x, Z = 0])/(\mathrm{E}[D|X = x, Z = 1] - \mathrm{E}[D|X = x, Z = 0])) f_x(x)\,\mathrm{d}x$ would represent a different causal parameter. In addition, estimation of $\int \gamma(x)\,\mathrm{d}F_x$ might be more difficult in small samples than of (5) if the effect of Z on D (i.e. the denominator) is small.

Theorem 2 (*Efficiency bound*). *The semiparametric variance bound for γ is*

$$\mathscr{V} = \frac{1}{\Gamma^2}\mathrm{E}\left[\frac{\sigma^2_{Y_1}(X) - 2\gamma\sigma^2_{Y_1D_1}(X) + \gamma^2\sigma^2_{D_1}(X)}{\pi(X)} + \frac{\sigma^2_{Y_0}(X) - 2\gamma\sigma^2_{Y_0D_0}(X) + \gamma^2\sigma^2_{D_0}(X)}{1-\pi(X)}\right] + \frac{1}{\Gamma^2}\mathrm{E}[(m_1(X) - m_0(X) - \gamma\mu_1(X) + \gamma\mu_0(X))^2], \tag{9}$$

where $\Gamma = \int(\mu_1(x) - \mu_0(x))f_x(x)\,\mathrm{d}x$ is the denominator in (5), *$\pi(x) = \mathrm{P}(Z = 1|X = x)$ is the probability that Z takes the value one given X, and $\sigma^2_{Y_z}(x)$, $\sigma^2_{D_z}(x)$ and $\sigma^2_{Y_zD_z}(x)$ are the conditional variances and covariances in the $Z = z$ subpopulation*: $\sigma^2_{Y_1}(X) = Var[Y|X = x, Z = 1]$, *and $\sigma^2_{Y_1D_1}(x) = Cov[Y, D|X = x, Z = 1]$ etc.* (*Proof in Appendix.*)

In the following Theorem 3, the asymptotic distribution of $\hat{\gamma}$ is derived and conditions on the preliminary estimators $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ are given for obtaining efficiency. In Theorems 4 and 5 it is shown that local polynomial regression as well as series estimation can yield efficient estimates.

Theorem 3 (*Asymptotic normality of $\hat{\gamma}$*). *Suppose that*

(i) *the LATE γ is identified,*
(ii) $\{(Y_i, D_i, Z_i, X_i)\}_{i=1}^n$ *are iid,* $Z_i \in \{0, 1\}$ *and* $X_i \in \Re^k$,
(iii) $\mathrm{E}[Y^2] < \infty$,
(iv) *the nonparametric estimator $\hat{m}_1$ of $m_1(x) = \mathrm{E}[Y|X = x, Z = 1]$ is asymptotically linear*

$$\hat{m}_1(x) - m_1(x) = \frac{1}{n_1}\sum\nolimits_{j:Z_j=1}\xi^m_1(Y_j, X_j, x) + b^m_1(x) + R^m_1(x),$$

where $n_z = \sum_{i=1}^n 1(Z_i = z)$, with the properties:

(A) $\mathrm{E}[\xi^m_1(Y_j, X_j, X)|X = x, Z_j = 1] = 0$;
(B) $\mathrm{E}[\xi^m_1(Y_j, X_j, X_i)^2|Z_j = 1, Z_i = 0] = o(n)$;
(C) $\frac{1}{\sqrt{n_0}}\sum_{i:Z_i=0} b^m_1(X_i) = \mathrm{o_p}(1)$;
(D) $\frac{1}{\sqrt{n_0}}\sum_{i:Z_i=0} R^m_1(X_i) = \mathrm{o_p}(1)$;
(E) $$\mathrm{E}[\xi^m_1(Y_j, X_j, X_i)|\, Y_j, X_j, Z_j = 1, Z_i = 0] = (Y_j - m_1(X_j))\frac{f_{x|z=0}(X_j)}{f_{x|z=1}(X_j)} + \mathrm{o_p}(1),$$

and analogously for $\hat{m}_0$, $\hat{\mu}_1$, $\hat{\mu}_0$. Then the estimator (8) *of the LATE γ is asymptotically normally distributed*

$$\sqrt{n}(\hat{\gamma} - \gamma) \to \mathrm{N}(0, \mathscr{V}), \tag{10}$$

where $\mathscr{V}$ is given by (9).

Condition (i) requires that the LATE γ is identified by the Assumptions 1–5 discussed previously. Condition (ii) supposes random sampling. This condition readily could be relaxed to allow for stratified sampling on X and Z,[10] as long as the population density function $f_x(x)$ can be recovered, for example through known sampling weights.

Condition (iv) gives conditions on the nonparametric estimators $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$. The term $\xi^m_z(Y_j, X_j, x)$ is the mean-zero local influence function of $\hat{m}_z(x)$ and captures its

[10]For example, iid sampling within $(Y, D, X)|Z = 1$ and $(Y, D, X)|Z = 0$.

variance $b_z^m(x)$ is the local bias of $\hat{m}_z(x)$, and $R_z^m(x)$ is a residual term. Condition (B) requires that the variance of ξ_z^m does not grow too fast. Condition (C) constrains the local bias term to be of order $o_p(n^{-(1/2)})$. This assumption could be relaxed to permit a local bias term of order $O_p(n^{-1/2})$, but this would introduce an asymptotic bias term in (10). The condition (E) is not necessary for $\sqrt{n}$-asymptotic normality, but is imposed to achieve efficiency.

It remains to find nonparametric estimators $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ that satisfy the conditions of Theorem 3. Potential estimators are kernel and local polynomial regression, nearest-neighbour regression and series estimators. Theorem 4 shows that local polynomial regression, which include Nadaraya–Watson and local linear regression, satisfies the conditions of Theorem 3 and thus leads to $\sqrt{n}$-asymptotically normal and efficient estimation of LATE.

Theorem 4 (*Efficiency of local polynomial regression*). *Consider the assumptions*:

(1) $m_1(x)$, $m_0(x)$, $\mu_1(x)$ *and* $\mu_0(x)$ *are* $\bar{p}$*-times continuously differentiable with* $\bar{p}$th *derivative Hölder continuous*,[11] *where* $\bar{p} > k$ *and* $k = \dim(X)$,
(2) *the bandwidth sequence* h_{n_1} *satisfies* $n_1 h_{n_1}^k / \ln n_1 \to \infty$ *and* $n_1 h_{n_1}^{2\bar{p}} \to 0$,
(3) *the Kernel function K is symmetric, compact and Lipschitz continuous*,[12]
(4) *the density* $f_{x|z=1}$ *of* X is $\bar{p}$*-times continuously differentiable with its* $\bar{p}$*th derivative Hölder continuous*,
(5a) *the Kernel function K has moments of order* 1 *through* $\bar{p}$ *equal to zero*,
(5b) *the Kernel function K has moments of order* 1 *through* $\bar{p} - 1$ *equal to zero*,
(6) $m_1(x)$, $m_0(x)$, $\mu_1(x)$ and $\mu_0(x)$ *are estimated at an interior point of Supp*$(X|Z = 1)$,[13]

Therem 4a. *If Assumptions* 1–4 *and* 5a *are satisfied, local polynomial regression of order* $\bar{p}$ *satisfies the conditions of Theorem* 3.

Theorem 4b. *If Assumptions* 1–4 *and* 5b *and* 6 *are satisfied, local polynomial regression of order* $0 \leqslant p < \bar{p}$ *satisfies the conditions of Theorem* 3.

Nearest-neighbour regression is not covered by Theorem 4. Nearest-neighbour regression may lead to efficient estimation of LATE if the number of neighbours increases with sample size at an appropriate rate. For a fixed number of neighbours, however, the estimator would be inefficient. Abadie and Imbens (2006) analyse the asymptotic properties of matching estimators with a fixed number of matches (i.e. using nearest-neighbour regression with a fixed number of neighbours). They show that these matching estimators are inefficient and have bias terms of stochastic order $n^{-1/k}$. Consequently, they are asymptotically biased for $k = 2$ and are not $\sqrt{n}$-consistent if X contains more than two continuous regressors. Since the LATE γ corresponds to a ratio of two matching estimators, both the numerator and the denominator of (5) would be estimated inefficiently or not $\sqrt{n}$-consistently if matching estimators with a fixed number of matches were used.

[11] Hölder continuity of a function $\zeta(x)$ at x_0 means that there exist $\alpha \in (0, 1]$ and $C > 0$ such that $|\zeta(x) - \zeta(x_0)| \leqslant C \cdot \|x - x_0\|^\alpha$ for all x.

[12] Lipschitz continuous means Hölder continuous of order $\alpha = 1$.

[13] This requires a trimming function to trim observations in the boundary.

As an alternative to local polynomial regression, $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ could also be estimated by nonparametric series regression. Under the assumptions of Theorem 5, the LATE estimator based on series regression estimates of $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$ is $\sqrt{n}$-asymptotically normal and efficient. Let $p^K(x) = ((p_{1K}(x), \ldots, p_{KK}(x))'$ be a vector of basis functions, where K is the (nonrandom) number of series terms included, which grows to infinity with increasing sample size. Let $p^K = (p^K(X_1), \ldots, p^K(X_n))'$ be the matrix for the data observations. A series estimator of, for example, $\mathrm{E}[Y|X = X_i]$ is $\hat{\mathrm{E}}[Y|X = X_i] = p^K(X_i)'\hat{\eta}$ where the coefficient vector is $\hat{\eta} = (p^{K\prime}p^K)^- p^{K\prime}(Y_1, \ldots, Y_n)'$. For the estimation of the conditional LATE, series estimators of the four conditional mean functions $m_1(x)$, $m_0(x)$, $\mu_1(x)$ and $\mu_0(x)$ are needed. These can be obtained by separate regressions in the $Z = 0$ and $Z = 1$ subpopulations, respectively. Alternatively, by noting that $\mathrm{E}[Y|X, Z = 1] = \mathrm{E}[YZ|X]/\mathrm{E}[Z|X]$, one can estimate m_1 by using series estimators of $\mathrm{E}[YZ|X]$ and $\mathrm{E}[Z|X]$, and analogously for m_0, μ_1 and μ_0. The proof is based on the latter approach, where the five conditional expectations $\mathrm{E}[YZ|X]$, $\mathrm{E}[Y(1-Z)|X]$, $\mathrm{E}[DZ|X]$, $\mathrm{E}[D(1-Z)|X]$ and $\mathrm{E}[Z|X]$ are estimated by series regression, using orthogonal power series. For convenience of notation it is assumed that the same basis functions are used in all five series regressions.

Theorem 5 (*Efficiency of series estimation*). *Suppose that*:

(i) *the LATE* γ *is identified*,
(ii) $\{(Y_i, D_i, Z_i, X_i)\}_{i=1}^n$ *are iid with* $Z_i \in \{0, 1\}$ *and* $X_i \in \Re^k$,
(iii) $m_1(x)$, $m_0(x)$, $\mu_1(x)$, $\mu_0(x)$ *are bounded on the support of* X *and* $\pi(x)$ *is bounded away from zero and from one on the support of* X,
(iv) $\sigma^2_{Y_1}(x), \sigma^2_{Y_0}(x), \sigma^2_{D_1}(x), \sigma^2_{D_0}(x)$ *are bounded on the support of* X,
(v) $\pi(x)$, $m_1(x)$, $m_0(x)$, $\mu_1(x)$, $\mu_0(x)$ *are continuously differentiable of order* s, *with* $s > 3k$ *where* k *is the dimension of* X
(vi) X *is continuously distributed with support equal to a Cartesian product of compact intervals*,
(vii) *the density of* X *is bounded away from zero on its support*,
(viii) *the series terms* $p_{kK}(x)$ *are products of polynomials that are orthonormal with respect to the uniform weight, with* $K = n^v$ *for some* $1/(4(s/k - 1)) < v < \frac{1}{7}$.

Then the conditional LATE estimator $\hat{\gamma}_{\text{series}}$ *based on series estimates of* $\mathrm{E}[YZ|X]$, $\mathrm{E}[Y(1-Z)|X]$, $\mathrm{E}[DZ|X]$, $\mathrm{E}[D(1-Z)|X]$ *and* $\mathrm{E}[Z|X]$ *is asymptotically normal*

$$\sqrt{n}(\hat{\gamma}_{\text{series}} - \gamma) \longrightarrow \mathrm{N}(0, \mathscr{V}),$$

where $\mathscr{V}$ *is given by* (9).

4. LATE with propensity score

The conditional LATE estimator $\hat{\gamma}$ corresponds to a ratio of two *matching* estimators. In the treatment evaluation literature, two alternative approaches to matching on X are widely used: matching on the propensity score and weighting by the propensity score, where the propensity score is the (one-dimensional) treatment participation probability. Matching and weighting by the propensity score have the advantage that they do not require high-dimensional nonparametric regression of $m_z(x)$ and $\mu_z(x)$.

Propensity score based estimators can also be constructed for the estimation of γ, with $\pi(x) = \mathrm{P}(Z = 1|X = x)$ as the propensity score. By noting that $\mathrm{E}[YZ/\pi(X)] = \int (1/\pi(x))\,\mathrm{E}[YZ|X = x]f_x(x)\,\mathrm{d}x = \int m_1(x)f_x(x)\,\mathrm{d}x$, the conditional LATE can be written as

$$\gamma = \frac{\mathrm{E}[YZ/\pi(X) - Y(1-Z)/(1-\pi(X))]}{\mathrm{E}[DZ/\pi(X) - D(1-Z)/(1-\pi(X))]} \tag{11}$$

and estimated by the *propensity score weighting estimator*

$$\hat{\gamma}_{\pi w} = \frac{\sum_i (Y_i Z_i/\pi(X_i) - Y_i(1-Z_i)/(1-\pi(X_i)))}{\sum_i (D_i Z_i/\pi(X_i) - D_i(1-Z_i)/(1-\pi(X_i)))}. \tag{12}$$

To derive the propensity score matching estimator, note that the conditional LATE can also be written as

$$\gamma = \frac{\int (m_{\pi 1}(\rho) - m_{\pi 0}(\rho)) \cdot f_\pi(\rho)\,\mathrm{d}\rho}{\int (\mu_{\pi 1}(\rho) - \mu_{\pi 0}(\rho)) \cdot f_\pi(\rho)\,\mathrm{d}\rho}, \tag{13}$$

where $m_{\pi 1}(\rho) = \mathrm{E}[Y|\pi(X) = \rho, Z = 1]$ and $\mu_{\pi 1}(\rho) = \mathrm{E}[D|\pi(X) = \rho, Z = 1]$ are the conditional means given the propensity score and f_π is the density function of $\pi(x)$ in the population. The *propensity score matching estimator* is

$$\hat{\gamma}_{\pi m} = \frac{\sum_i (\hat{m}_{\pi 1}(\pi_i) - \hat{m}_{\pi 0}(\pi_i))}{\sum_i (\hat{\mu}_{\pi 1}(\pi_i) - \hat{\mu}_{\pi 0}(\pi_i))}, \tag{14}$$

where $\pi_i = \pi(X_i)$. An advantage of propensity score matching $\hat{\gamma}_{\pi m}$ over $\hat{\gamma}$ is that it requires only *one-dimensional* nonparametric regression.

To prove Eq. (13), note that by the law of iterated expectations

$$\begin{aligned}
\mathrm{E}\left[\frac{YZ}{\pi(X)}\right] &= \mathrm{E}\mathrm{E}\left[\frac{YZ}{\pi(X)}|\pi(X) = \rho\right] \\
&= \mathrm{E}\left[\frac{1}{\rho}\mathrm{E}[Y|\pi(X) = \rho, Z = 1]\mathrm{P}[Z = 1|\pi(X) = \rho]\right] \\
&= \mathrm{E}[\mathrm{E}[Y|\pi(X) = \rho, Z = 1]] = \int m_{\pi 1}(\rho)f_\pi(\rho)\,\mathrm{d}\rho.
\end{aligned}$$

Since furthermore, $\mathrm{E}[YZ/\pi(X)] = \int m_1(x)f_x(x)\,\mathrm{d}x$, as shown above, it follows that $\int m_1(x)f_x(x)\,\mathrm{d}x = \int m_{\pi 1}(\rho)f_\pi(\rho)\,\mathrm{d}\rho$. With analogous results for $m_{\pi 0}$, $\mu_{\pi 1}$ and $\mu_{\pi 0}$, Eq. (13) follows.

In many applications, the propensity score $\pi(x)$ might be unknown and needs to be estimated. Even when the propensity score is known, using an estimated propensity score in the estimator might be worthwhile. Whether knowledge of the true propensity score affects efficient estimation has been analysed in several articles for matching estimators. Analogous results can be derived here for the estimation of LATE. Theorem 6 shows that the efficiency bound for the estimator $\hat{\gamma}$ is not affected by knowledge of the propensity score. This result is similar to Hahn (1998).

Theorem 6 (*Efficiency bound with knowledge of propensity score*). *Suppose that the propensity score $\pi(x)$ is known. The semiparametric variance bound for γ is $\mathscr{V}$, where $\mathscr{V}$* is given by (9).

For the propensity score weighting estimator, noting the similarities to Hirano et al. (2003), I conjecture that weighting with an appropriately estimated propensity score would be efficient, whereas weighting with the true propensity score would not be. A proof, however, is beyond the scope of this paper.

For the propensity score matching estimator, to assess efficiency consider first the situation where the propensity score $\pi(x)$ is known. Provided that the conditions of Theorems 3 or 5 are satisfied with X replaced by $\pi(X)$, the asymptotic variance of propensity score matching with respect to the known propensity score π is

$$V_{\pi m} = \frac{1}{\Gamma^2}\mathrm{E}\left[\frac{\sigma^2_{\pi Y_1}(\pi) - 2\gamma\sigma^2_{\pi Y_1 D_1}(\pi) + \gamma^2\sigma^2_{\pi D_1}(\pi)}{\pi} + \frac{\sigma^2_{\pi Y_0}(\pi) - 2\gamma\sigma^2_{\pi Y_0 D_0}(\pi) + \gamma^2\sigma^2_{\pi D_0}(\pi)}{1-\pi}\right]$$
$$+ \frac{1}{\Gamma^2}\mathrm{E}\Big[\big(m_{\pi 1}(\pi) - m_{\pi 0}(\pi) - \gamma\mu_{\pi 1}(\pi) + \gamma\mu_{\pi 0}(\pi)\big)^2\Big], \tag{15}$$

where $\sigma^2_{\pi Y_1}(\rho) = Var[Y|\pi(X)=\rho, Z=1]$ and $\sigma^2_{\pi Y_1 D_1}(\rho) = Cov[Y, D|\pi(X)=\rho, Z=1]$. The following theorem shows that this variance is generally larger than $\mathscr{V}$. If the propensity score is unknown, its estimation would add even further to the variance of the propensity score matching estimator.[14] Therefore, matching on the propensity score is generally inefficient.

Theorem 7 (*Inefficiency of propensity score matching*). *The difference between the asymptotic variance when matching on the known propensity score and the asymptotic variance when matching on X is non-negative and given by*

$$V_{\pi m} - \mathscr{V} = \frac{1}{\Gamma^2}\mathrm{E}\left[\frac{1-\pi}{\pi}Var\big(m_1(X) - \gamma\mu_1(X)|\pi\big) + \frac{\pi}{1-\pi}Var\big(m_0(X) - \gamma\mu_0(X)|\pi\big)\right]$$
$$+ \frac{2}{\Gamma^2}\mathrm{E}[Cov(m_1(X) - \gamma\mu_1(X), m_0(X) - \gamma\mu_0(X)|\pi)] \geqslant 0.$$

Generally, $V_{\pi m} - \mathscr{V}$ is strictly positive unless the support of the propensity score $\pi(x)$ contains only values where both variances are zero or where $\sqrt{(Var(m_1(X) - \gamma\mu_1(X)|\pi))/(Var(m_0(X) - \gamma\mu_0(X)|\pi))}$
$-\pi(1 + \sqrt{(Var(m_1(X) - \gamma\mu_1(X)|\pi))/(Var(m_0(X) - \gamma\mu_0(X)|\pi))}) = 0.$

5. Non-binary treatments and non-binary instruments

Whereas the previous two sections considered the estimation of LATE with D and Z binary, extensions to non-binary instrumental variables Z and non-binary endogenous regressors D are considered here. It will be seen that the formula (5) still applies to the estimation of the LATE, such that the results derived in Theorems 2–7 largely carry through to non-binary instrumental variables and non-binary endogenous regressors.

Examine first the situation with a binary regressor D and a single non-binary instrument Z, which has bounded support $Supp(Z) = [z_{\min}, z_{\max}]$. Obviously, a LATE could be defined with respect to any two distinct values of Z. However, this would yield a multitude of pair-wise treatment effects, each of them referring to a different population. Instead of estimating many pair-wise effects, one would often prefer to estimate the average treatment

[14] Unless a joint estimator of π, $m_{\pi 1}$, $m_{\pi 0}$, $\mu_{\pi 1}$ and $\mu_{\pi 0}$ with smaller variance can be constructed.

effect in the *largest* subpopulation for which an effect can be identified, which is the subpopulation of all individuals who react to the instrument.

Define the subpopulation of compliers as all individuals with $D_{i,z_{\min}} = 0$ and $D_{i,z_{\max}} = 1$. The compliers comprise all individuals who switch from $D = 0$ to 1 at some point when the instrument Z is increased from $z_{\min}$ to $z_{\max}$. The value of z which triggers the switch can be different for different individuals. If monotonicity holds with respect to any two values z and z', each individual switches D at most once. The following assumptions are extensions of Assumptions 1–5 to a non-binary instrument.

Assumption 1 (*Monotonicity*). The effect of Z on D is monotonous

$$\mathrm{P}(D_{i,z} > D_{i,z'}) = 0 \quad \text{for any values} \quad z, z' \text{ with } z_{\min} \leqslant z < z' \leqslant z_{\max}.$$

Assumption 2 (*Existence of compliers*). The subpopulation of compliers has positive probability

$$\mathrm{P}(\tau = c) > 0 \quad \text{where } \tau_i = c \text{ if } D_{i,z_{\min}} < D_{i,z_{\max}}.$$

Assumption 3 (*Unconfounded type*). For any two values $z, z' \in Supp(Z)$, any $d, d' \in \{0,1\}$ and for all $x \in Supp(X)$

$$\mathrm{P}(D_{i,z} = d, D_{i,z'} = d' | X_i = x, Z_i = z) = \mathrm{P}(D_{i,z} = d, D_{i,z'} = d' | X_i = x).$$

Assumption 4 (*Mean exclusion restriction*). For any two values $z, z' \in Supp(Z)$, any $d, d' \in \{0,1\}$ and for all $x \in Supp(X)$

$$\mathrm{E}[Y^d_{i,Z_i} | X_i = x, D_{i,z} = d, D_{i,z'} = d', Z_i = z] = \mathrm{E}[Y^d_{i,Z_i} | X_i = x, D_{i,z} = d, D_{i,z'} = d'].$$

Assumption 5 (*Common support*). The support of X is identical for $z_{\min}$ and $z_{\max}$

$$Supp(X|Z = z_{\min}) = Supp(X|Z = z_{\max}) = Supp(X).$$

Theorem 8 (*LATE with non-binary instrument*). *Suppose that D is binary, the instrument Z has bounded support* $Supp(Z) = [z_{\min}, z_{\max}]$ *and Assumptions* 1–5' *are satisfied. The LATE for the subpopulation of compliers is nonparametrically identified as*

$$\mathrm{E}[Y^1 - Y^0 | \tau = c] = \frac{\int (\mathrm{E}[Y|X = x, Z = z_{\max}] - \mathrm{E}[Y|X = x, Z = z_{\min}]) f_x(x)\,\mathrm{d}x}{\int (\mathrm{E}[D|X = x, Z = z_{\max}] - \mathrm{E}[D|X = x, Z = z_{\min}]) f_x(x)\,\mathrm{d}x}. \tag{16}$$

This formula is analogous to (5) with $Z = 0$ and $Z = 1$ replaced with the endpoints of the support of Z. If Z is discrete, the results of Theorems 2–7 can be applied. For a continuous instrument, however, $\sqrt{n}$-consistency usually cannot be achieved, unless it is mixed continuous-discrete with mass points at $z_{\min}$ and $z_{\max}$.

From (16) a bias-variance trade-off in the estimation of the LATE with non-binary Z becomes visible. Although (16) incorporates the proper weighting of the different complier subgroups and leads to an unbiased estimator of LATE, only observations with Z_i equal (or close) to $z_{\min}$ or $z_{\max}$ are used for estimation. Observations with Z_i between the endpoints $z_{\min}$ and $z_{\max}$ are neglected, which might lead to a large variance. Variance could be reduced, at the expense of a larger bias, by weighting the complier subgroups differently or by choosing larger bandwidth values for the estimators $\hat{m}_z(x)$ and $\hat{\mu}_z(x)$. A detailed analysis is left for future research.

Now consider the situation with multiple instrumental variables, i.e. Z being vector valued. Since the different instrumental variables act through their effect on D, the

different components of Z can be summarized conveniently by using $p(z,x) = \mathrm{P}(D = 1|X = x, Z = z)$ as instrument. If D follows an index structure in the sense that D_i depends on Z_i only via $p(Z_i, X_i)$,[15] and Assumption 1–5' are satisfied with respect to $p(z,x)$, the LATE is identified as

$$\mathrm{E}[Y^1 - Y^0|\tau = c] = \frac{\int(\mathrm{E}[Y|X = x, p(Z,X) = \bar{p}_x] - \mathrm{E}[Y|X = x, p(Z,X) = \underline{p}_x])f_x(x)\,\mathrm{d}x}{\int(\mathrm{E}[D|X = x, p(Z,X) = \bar{p}_x] - \mathrm{E}[D|X = x, p(Z,X) = \underline{p}_x])f_x(x)\,\mathrm{d}x}, \quad (17)$$

where $\bar{p}_x = \max_z p(z,x)$ and $\underline{p}_x = \min_z p(z,x)$. Again, this formula is analogous to (5).[16] The two groups of observations on which estimation is based are those with $p(z,x) = \bar{p}_x$ and those with $p(z,x) = \underline{p}_x$. Exact knowledge of $p(z,x)$ is not needed for estimation; it is sufficient to identify the set of observations for which $p(Z,X)$ is highest and lowest, respectively. For example, if Z contains two binary instrumental variables (Z_1, Z_2) which, for any value of X, both have a positive effect on D, then the observations with $Z_1 = Z_2 = 0$ and those with $Z_1 = Z_2 = 1$ represent the endpoints of the support of $p(Z,X)$ given X and are used for the estimation.

Finally, consider the situation where the *endogenous regressor D is non-binary*. Suppose that $D \in \{0, \ldots, K\}$ is discrete and that the instrument Z is binary. With D taking many different values, the *compliance intensity* can differ among units. Some units might be induced to change from $D_i = d$ to $D_i = d + 1$ as a reaction on changing Z_i from 0 to 1. Other units might change, for example, from $D_i = d'$ to $D_i = d' + 2$. Suppose D is years of schooling and Z an instrument that influences the schooling decision. If Z were changed exogenously, some individuals might respond by increasing school attendance by an additional year. Other individuals might increase school attendance even by two or three years. Furthermore, even if Z were set to zero for all individuals, they would attend different years of schooling. Hence a change in Z induces a variety of different reactions in D, which cannot be disentangled. Only a weighted average of these effects can be identified. According to their reaction on a change in Z from 0 to 1, the population can be partitioned into the types $c_{0,0}, c_{0,1}, \ldots, c_{K,K}$ where

$$\tau_i = c_{k,l} \quad \text{if } D_{i,0} = k \text{ and } D_{i,1} = l. \quad (18)$$

Assuming monotonicity, the defier-types $c_{k,l}$ for $k > l$ do not exist. The types $c_{k,k}$ represent those units that do not react on a change in Z (these are the always-takers and the never-takers in the setup where D is binary). The types $c_{k,l}$ for $k < l$ are the compliers, which comply by increasing D_i from k to l. These compliers comply at different base levels k and with different intensities $l - k$. In the returns to schooling example, $\mathrm{E}[Y^{k+1} - Y^k|X, \tau = c_{k,k+1}]$ measures the return to one additional year of schooling for the $c_{k,k+1}$ subpopulations. $\mathrm{E}[Y^{k+2} - Y^k|X, \tau = c_{k,k+2}]$ measures the return to two additional years of schooling, which can be interpreted as twice the average return of one additional year. Similarly, $\mathrm{E}[Y^{k+3} - Y^k|X, \tau = c_{k,k+3}]$ is three times the average return to one additional year. Hence the effective weight contribution of the $c_{k,l}$ subpopulation to the measurement of the return to one additional year of schooling is $(l - k) \cdot \mathrm{P}(\tau = c_{k,l})$. Accordingly, the weighted average treatment effect $\gamma_w(x)$ for all compliers with

[15] I.e. $D_{i,z} = D_{i,z'}$ if $p(z, X_i) = p(z', X_i)$. In other words, D_i does not change if Z_i is varied within a set where $p(\cdot, X_i)$ remains constant. See e.g. Heckman and Vytlacil, 2001. If the index structure is not satisfied, a more involved monotonicity condition would be needed.

[16] The denominator of (17) can also be written as $\int(\bar{p}_x - \underline{p}_x)f_x(x)\,\mathrm{d}x$.

characteristics x can be defined as[17]

$$\gamma_w(X) = \frac{\sum_k^K \sum_{l>k}^K \mathrm{E}[Y^l - Y^k | X, \tau = c_{k,l}] \cdot \mathrm{P}(\tau = c_{k,l}|X)}{\sum_k^K \sum_{l>k}^K (l-k) \cdot \mathrm{P}(\tau = c_{k,l}|X)}, \tag{19}$$

where $\gamma_w(x)$ is the effect of the induced treatment change, averaged over the different complier groups and normalized by the intensity of compliance.

To obtain the weighted average effect for the subpopulation of all compliers (i.e. all subpopulations $c_{k,l}$ with $k<l$), one would need to weight $\gamma_w(x)$ by the distribution of X in the complier subpopulation

$$\int \gamma_w(x)\, \mathrm{d}F_{x|\text{complier}},$$

where $F_{x|\text{complier}}$ is the distribution of X in the all-compliers subpopulation.

Unfortunately, the distribution of X in the all-compliers subpopulation is not identified if D takes more than two different values. In particular, the size of the all-compliers subpopulation is no longer identified by the distribution of D and Z.[18] Nevertheless, if one defines the all-compliers subpopulation in terms of *compliance intensity units*, the distribution of X is identified. In the intensity-weighted complier subpopulation, each complier is weighted by its compliance intensity. In the case where $D \in \{0,1,2\}$, the subpopulation $c_{0,2}$ receives twice the weight of the subpopulation $c_{0,1}$. In the years-of-schooling example, the subpopulation $c_{0,2}$ complies with two additional years of schooling. If the returns to a year of schooling are the same for each year of schooling, an individual who complies with two additional years can be thought of as an observation that measures twice the effect of one additional year of schooling. Or, in other words, as two (correlated) measurements of the return to a year of schooling. Unless these two measurements are perfectly correlated, the individual who complies with two additional years contributes more to the estimation of the return to schooling than an individual who complies with only one additional year. Consequently, the individuals who comply with more than one year should receive a higher weight when averaging the return to schooling over the distribution of X. If each individual is weighted by its number of additional years, the weighted distribution function of X in the all-compliers subpopulation, in the case where $D \in \{0,1,2\}$, is

$$f^w_{x|\text{complier}} = \frac{f_{x|\tau=c_{0,1}} \mathrm{P}(\tau = c_{0,1}) + f_{x|\tau=c_{1,2}} \mathrm{P}(\tau = c_{1,2}) + 2 f_{x|\tau=c_{0,2}} \mathrm{P}(\tau = c_{0,2})}{\mathrm{P}(\tau = c_{0,1}) + \mathrm{P}(\tau = c_{1,2}) + 2\mathrm{P}(\tau = c_{0,2})}$$

or in the general case

$$f^w_{x|\text{complier}} = \frac{\sum_k^K \sum_{l>k}^K (l-k) \cdot f_{x|\tau=c_{k,l}} \mathrm{P}(\tau = c_{k,l})}{\sum_k^K \sum_{l>k}^K (l-k) \cdot \mathrm{P}(\tau = c_{k,l})}. \tag{20}$$

[17]The presentation in Angrist and Imbens (1995) looks different from the definition of γ_w used here, as they present the effect in terms of overlapping subpopulations. Nevertheless, both definitions are equivalent.

[18]Consider the following example: for D taking values in $\{0,1,2\}$, the population can be partitioned in the subpopulations: $\{c_{00}, c_{01}, c_{02}, c_{11}, c_{12}, c_{22}\}$ with the all-compliers subpopulation consisting of $\{c_{01}, c_{02}, c_{12}\}$. The two partitions $\{c_{00}, c_{01}, c_{02}, c_{11}, c_{12}, c_{22}\} = \{0.1, 0.1, 0.3, 0.3, 0.1, 0.1\}$ and $\{0.1, 0.2, 0.2, 0.2, 0.2, 0.1\}$ generate the same distribution of D given Z: $\mathrm{P}(D=0|Z=0) = 0.5$, $\mathrm{P}(D=1|Z=0) = 0.4$, $\mathrm{P}(D=2|Z=0) = 0.1$, $\mathrm{P}(D=0|Z=1) = 0.1$, $\mathrm{P}(D=1|Z=1) = 0.4$, $\mathrm{P}(D=2|Z=1) = 0.5$. However, the size of the all-compliers subpopulation is different for the two partitions (0.5 and 0.6, respectively). Hence the size of the all-compliers subpopulation is not identified from the observable variables.

With respect to this weighted distribution function, the weighted LATE is identified, again by a formula analogous to (5).

Theorem 9 (*LATE with non-binary endogenous regressor*). *Suppose that D is discrete with bounded support, the instrument Z is binary and* Assumptions 1, 2, 5 *are satisfied as well as* Assumptions 3 *and* 4 *with respect to all types* $t \in \{c_{k,l} : k \leqslant l\}$, *defined in* (18). *The weighted LATE for the subpopulation of compliers is nonparametrically identified as*

$$\gamma_w = \int \gamma_w(x) \cdot f^w_{x|\text{complier}}(x)\,\mathrm{d}x \tag{21}$$

$$= \frac{\int (\mathrm{E}[Y|X=x,Z=1] - \mathrm{E}[Y|X=x,Z=0]) \cdot f_x(x)\,\mathrm{d}x}{\int (\mathrm{E}[D|X=x,Z=1] - \mathrm{E}[D|X=x,Z=0]) \cdot f_x(x)\,\mathrm{d}x}. \tag{22}$$

6. Conclusions

In this paper nonparametric instrumental variables estimation of LATE has been extended to accommodate confounding covariates X. A nonparametric LATE estimator has been proposed and its asymptotic properties have been derived.

Identification of LATE requires weaker assumptions than many alternative nonparametric instrumental variable models. Identification and estimation of LATE has usually been discussed without covariates. However, when the instrument is not randomly assigned, the instrumental variable assumptions are often only valid conditional on a vector of confounding covariates X. Usually covariates X have been included via parametric modelling. The LATE estimator proposed in this paper, however, incorporates covariates X in a fully *nonparametric* way. This estimator corresponds to a ratio of two matching estimators and it is $\sqrt{n}$-consistent, asymptotically normal and efficient.

Acknowledgements

I am grateful for discussions and comments to Joshua Angrist, Manuel Arellano, Richard Blundell, Bernd Fitzenberger, Per Johansson, Michael Lechner, Costas Meghir, Hidehiko Ichimura, Jeff Smith, the editors and two anonymous referees. The paper was presented at the Econometric Society European Winter Meeting (Budapest, 2002), the EC2 meeting in London (December, 2003) and at seminars in Dublin, Konstanz and München. This research was supported by the Swiss National Science Foundation (project NSF 4043-058311), the Grundlagenforschungsfonds HSG (project G02110112) and the Marie Curie Individual Fellowship MEIF-CT-2004-006873.

Appendix A

A.1. Derivation of the treatment effect $\gamma(x)$

It is shown that Assumptions 1–4 identify the LATE $\gamma(x)$ for all x with $\mathrm{P}(\tau = c|X = x) > 0$ according to (7). By partitioning the population into the subpopulations always-takers, never-takers, compliers and defiers, the expected value of Y given X and Z can be

written as

$$\begin{aligned}
\mathrm{E}[Y_i|X_i=x,Z_i=z] &= \mathrm{E}[Y^{D_i}_{i,Z_i}|X_i=x,Z_i=z,\tau_i=n]\cdot\mathrm{P}(\tau_i=n|X_i=x,Z_i=z)\\
&\quad+\mathrm{E}[Y^{D_i}_{i,Z_i}|X_i=x,Z_i=z,\tau_i=c]\cdot\mathrm{P}(\tau_i=c|X_i=x,Z_i=z)\\
&\quad+\mathrm{E}[Y^{D_i}_{i,Z_i}|X_i=x,Z_i=z,\tau_i=d]\cdot\mathrm{P}(\tau_i=d|X_i=x,Z_i=z)\\
&\quad+\mathrm{E}[Y^{D_i}_{i,Z_i}|X_i=x,Z_i=z,\tau_i=a]\cdot\mathrm{P}(\tau_i=a|X_i=x,Z_i=z)\\
&= \mathrm{E}[Y^{0}_{i,Z_i}|X_i=x,Z_i=z,\tau_i=n]\cdot\mathrm{P}(\tau_i=n|X_i=x)\\
&\quad+\mathrm{E}[Y^{D_i}_{i,Z_i}|X_i=x,Z_i=z,\tau_i=c]\cdot\mathrm{P}(\tau_i=c|X_i=x)\\
&\quad+\mathrm{E}[Y^{1}_{i,Z_i}|X_i=x,Z_i=z,\tau_i=a]\cdot\mathrm{P}(\tau_i=a|X_i=x)
\end{aligned}$$

by Assumptions 1 and 3 and the definition of the types. By the mean exclusion restriction (Assumption 4) the potential outcomes are independent of Z in the always- and in the never-taker subpopulation. Hence when taking the difference $\mathrm{E}[Y|X,Z=1]-\mathrm{E}[Y|X,Z=0]$ the respective terms for the always- and for the never-takers cancel, such that

$$\begin{aligned}
&\mathrm{E}[Y_i|X_i=x,Z_i=1]-\mathrm{E}[Y_i|X_i=x,Z_i=0]\\
&\quad=(\mathrm{E}[Y^{1}_{i,Z_i}|X_i=x,Z_i=1,\tau_i=c]-\mathrm{E}[Y^{0}_{i,Z_i}|X_i=x,Z_i=0,\tau_i=c])\cdot\mathrm{P}(\tau_i=c|X_i=x).
\end{aligned}$$

Exploiting the mean exclusion restriction for the compliers gives

$$\begin{aligned}
&\mathrm{E}[Y_i|X_i=x,Z_i=1]-\mathrm{E}[Y_i|X_i=x,Z_i=0]\\
&\quad=\mathrm{E}[Y^{1}_{i,Z_i}-Y^{0}_{i,Z_i}|X_i=x,\tau_i=c]\cdot\mathrm{P}(\tau_i=c|X_i=x).
\end{aligned}\tag{23}$$

Hence,

$$\gamma(x)=\mathrm{E}[Y^{1}_{i,Z_i}-Y^{0}_{i,Z_i}|X_i=x,\tau_i=c]=\frac{\mathrm{E}[Y_i|X_i=x,Z_i=1]-\mathrm{E}[Y_i|X_i=x,Z_i=0]}{\mathrm{P}(\tau_i=c|X_i=x)}.$$

Noticing that $\mathrm{E}[D|X,Z=0]=\mathrm{P}(D=1|X,Z=0)=\mathrm{P}(\tau=a|X)+\mathrm{P}(\tau=d|X)$ and $\mathrm{E}[D|X,Z=1]=\mathrm{P}(D=1|X,Z=1)=\mathrm{P}(\tau=a|X)+\mathrm{P}(\tau=c|X)$, the relative size of the subpopulation of compliers is identified as

$$\mathrm{P}(\tau_i=c|X_i=x)=\mathrm{E}[D_i|X_i=x,Z_i=1]-\mathrm{E}[D_i|X_i=x,Z_i=0],$$

and it follows that the average treatment effect in the subpopulation of compliers is

$$\gamma(x)=\frac{\mathrm{E}[Y_i|X_i=x,Z_i=1]-\mathrm{E}[Y_i|X_i=x,Z_i=0]}{\mathrm{E}[D_i|X_i=x,Z_i=1]-\mathrm{E}[D_i|X_i=x,Z_i=0]}.$$

A.2. Proof of Theorem 2

Semiparametric efficiency bounds were introduced by Stein (1956) and developed by Koshevnik and Levit (1976), Pfanzagl and Wefelmeyer (1982), Begun et al. (1983) and Bickel et al. (1993). See also the survey of Newey (1990, 1994). The approach followed here is similar to Hahn (1998).

The joint density of (Y, D, Z, X) with Z binary can be written as

$$f(y,d,z,x) = f(y,d|z,x)f(z|x)f(x) = \{f_1(y,d|x)\pi(x)\}^z\{f_0(y,d|x)(1-\pi(x))\}^{1-z}f(x),$$

where $f_1(y,d|x) \equiv f(y,d|z=1,x)$ and $\pi(x) = \mathrm{P}(Z=1|X=x)$.

Consider a regular parametric submodel indexed by θ with θ_0 corresponding to the true model: $f(y,d,z,x|\theta_0) = f(y,d,z,x)$. The density $f(y,d,z,x|\theta)$ can be written as

$$f(y,d,z,x|\theta) = \{f_1(y,d|x,\theta)\pi(x,\theta)\}^z\{f_0(y,d|x,\theta)(1-\pi(x,\theta))\}^{1-z}f(x,\theta),$$

where f_1 and f_0 admit an interchange of the order of integration and differentiation

$$\int \frac{\partial f_1(y,d|x,\theta)}{\partial\theta}\,\mathrm{d}y\,\mathrm{d}d = \frac{\partial}{\partial\theta}\int f_1(y,d|x,\theta)\,\mathrm{d}y\,\mathrm{d}d = 0$$

and analogously for f_0. Sufficient conditions for permitting interchanging differentiation and integration are, for example, given by Theorem 1.3.2 of Amemiya (1985). These are $\partial f_1(y,d|x,\theta)/\partial\theta$ is continuous in $\theta \in \Theta$ and y, where Θ is an open set; $\int f_1(y,d|x,\theta)\,\mathrm{d}y\,\mathrm{d}d$ exists; and $\int |\partial f_1(y,d|x,\theta)/\partial\theta|\,\mathrm{d}y\,\mathrm{d}d < \infty$ for all $\theta \in \Theta$.

The corresponding score of $f(y,d,z,x|\theta)$ is

$$\begin{aligned} S(y,d,z,x|\theta) &= \frac{\partial \ln f(y,d,z,x|\theta)}{\partial\theta} \\ &= z\cdot\check{f}_1(y,d|x,\theta) + (1-z)\cdot\check{f}_0(y,d|x,\theta) + \frac{z-\pi(x,\theta)}{1-\pi(x,\theta)}\check{\pi}(x,\theta) + \check{f}(x,\theta), \end{aligned}$$

where $\check{f}_1(y,d|x,\theta) = \partial \ln f_1(y,d|x,\theta)/\partial\theta$, and $\check{f}_0$ analogously, and $\check{\pi}(x,\theta) = \partial \ln \pi(x,\theta)/\partial\theta$ and $\check{f}(x,\theta) = \partial \ln f(x,\theta)/\partial\theta$.

At the true value θ_0 the expectation of the score is zero. The tangent space of the model is the set of functions that are mean zero and satisfy the additive structure of the score

$$\mathfrak{I} = \{z\cdot s_1(y,d|x) + (1-z)\cdot s_0(y,d|x) + (z-\pi(x))\cdot s_\pi(x) + s_x(x)\} \tag{24}$$

for any functions s_1, s_0, s_π, s_x satisfying the mean-zero property:

$$\int s_1(y,d|x)f_1(y,d|x)\,\mathrm{d}y\,\mathrm{d}d = 0\ \forall x,$$

$$\int s_0(y,d|x)f_0(y,d|x)\,\mathrm{d}y\,\mathrm{d}d = 0\ \forall x,$$

$$\int s_x(x)f(x)\,\mathrm{d}x = 0,$$

and $s_\pi(x)$ being a square-integrable measurable function of x.

The *semiparametric variance bound* of γ is the variance of the projection on $\mathfrak{I}$ of a function $\psi(Y,D,Z,X)$ (with $\mathrm{E}[\psi] = 0$ and $\mathrm{E}[\|\psi(\cdot)\|^2] < \infty$) that satisfies for all regular parametric submodels

$$\frac{\partial\gamma(F_\theta)}{\partial\theta}_{|\theta=\theta_0} = \mathrm{E}[\psi(Y,D,Z,X)\cdot S(Y,D,Z,X)]_{|\theta=\theta_0}. \tag{25}$$

Write γ as

$$\begin{aligned}\gamma = \frac{\Delta}{\Gamma} &= \frac{\int(\mathrm{E}_\theta[Y|X,Z=1]-\mathrm{E}_\theta[Y|X,Z=0])\cdot f(x,\theta)\,\mathrm{d}x}{\int(\mathrm{E}_\theta[D|X,Z=1]-\mathrm{E}_\theta[D|X,Z=0])\cdot f(x,\theta)\,\mathrm{d}x}\\ &= \frac{\int\left(\int\int yf_1(y,d|x,\theta)\,\mathrm{d}y\,\mathrm{d}d-\int\int yf_0(y,d|x,\theta)\,\mathrm{d}y\,\mathrm{d}d\right)\cdot f(x,\theta)\,\mathrm{d}x}{\int\left(\int\int \mathrm{d}f_1(y,d|x,\theta)\,\mathrm{d}y\,\mathrm{d}d-\int\int \mathrm{d}f_0(y,d|x,\theta)\,\mathrm{d}y\,\mathrm{d}d\right)\cdot f(x,\theta)\,\mathrm{d}x}\\ &= \frac{\int\int\int yf_1(y,d|x,\theta)f(x,\theta)\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x-\int\int\int yf_0(y,d|x,\theta)f(x,\theta)\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x}{\int\int\int \mathrm{d}f_1(y,d|x,\theta)f(x,\theta)\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x-\int\int\int \mathrm{d}f_0(y,d|x,\theta)f(x,\theta)\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x}\end{aligned}$$

since $\mathrm{E}[Y|X,Z=1]=\int\int yf_1(y,d|x)\,\mathrm{d}y\,\mathrm{d}d$.[19]

Computing the pathwise derivative and evaluating it at θ_0 gives

$$\begin{aligned}\frac{\partial\gamma(F_\theta)}{\partial\theta}_{|\theta=\theta_0} &= \frac{(\partial\Delta/\partial\theta)\Gamma-\Delta(\partial\Gamma/\partial\theta)}{\Gamma^2}_{|\theta=\theta_0} = \frac{\partial\Delta/\partial\theta}{\Gamma}-\gamma\frac{\partial\Gamma/\partial\theta}{\Gamma}_{|\theta=\theta_0}\\ &= \frac{\int\int\int y(\dot f_1 f+f_1\dot f)\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x-\int\int\int y(\dot f_0 f+f_0\dot f)\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x}{\Gamma}\\ &\quad-\gamma\frac{\int\int\int d(\dot f_1 f+f_1\dot f)\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x-\int\int\int d(\dot f_0 f+f_0\dot f)\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x}{\Gamma}\\ &= \frac{\int\int\int y\{\dot f_1-\dot f_0\}f\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x}{\Gamma}-\gamma\frac{\int\int\int d\{\dot f_1-\dot f_0\}f\,\mathrm{d}y\,\mathrm{d}d\,\mathrm{d}x}{\Gamma}\\ &\quad+\frac{\int(m_1(x)-m_0(x)-\gamma\mu_1(x)+\gamma\mu_0(x))\dot f(x)\,\mathrm{d}x}{\Gamma}\end{aligned}$$

where $\dot f_1=(\partial/\partial\theta)f_1(y,d|x,\theta)_{|\theta=\theta_0}$, $\dot f_0=(\partial/\partial\theta)f_0(y,d|x,\theta)_{|\theta=\theta_0}$ and $\dot f=(\partial/\partial\theta)f(x,\theta)_{|\theta=\theta_0}$.

Choose $\psi(Y,D,Z,X)$ as

$$\begin{aligned}\psi(y,d,z,x) = \frac{z}{\Gamma}\frac{y-m_1(x)-\gamma d+\gamma\mu_1(x)}{\pi(x)}+\frac{1-z}{\Gamma}\frac{\gamma d-\gamma\mu_0(x)-y+m_0(x)}{1-\pi(x)}\\ +\frac{m_1(x)-m_0(x)-\gamma\mu_1(x)+\gamma\mu_0(x)}{\Gamma}.\end{aligned}\tag{26}$$

Note that ψ satisfies (25)

$$\frac{\partial\gamma(F_\theta)}{\partial\theta}_{|\theta=\theta_0}=\mathrm{E}[\psi(Y,D,Z,X)\cdot S(Y,D,Z,X)]_{|\theta=\theta_0}$$

and that ψ lies in the tangent space (24)[20]

$$\psi\in\mathfrak{I}.$$

[19]And analogously for $\mathrm{E}[D|X,Z=1]$, $\mathrm{E}[Y|X,Z=0]$ and $\mathrm{E}[D|X,Z=0]$.

[20]The calculations are available from the author.

Since ψ lies in the tangent space, the variance bound is the expected square of ψ:

$$\begin{aligned}
\mathrm{E}[\psi(Y,D,Z,X)^2] &= \frac{1}{\Gamma^2}\mathrm{E}\left[Z\left(\frac{Y-m_1(X)-\gamma D+\gamma\mu_1(X)}{\pi(X)}\right)^2\right] \\
&\quad+\frac{1}{\Gamma^2}\mathrm{E}\left[(1-Z)\left(\frac{\gamma D-\gamma\mu_0(X)-Y+m_0(X)}{1-\pi(X)}\right)^2\right] \\
&\quad+\frac{1}{\Gamma^2}\mathrm{E}\left[Z\frac{Y-m_1(X)-\gamma D+\gamma\mu_1(X)}{\pi(X)}\cdot(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))\right] \\
&\quad+\frac{1}{\Gamma^2}\mathrm{E}\left[(1-Z)\frac{\gamma D-\gamma\mu_0(X)-Y+m_0(X)}{1-\pi(X)}\cdot(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))\right] \\
&\quad+\frac{1}{\Gamma^2}\mathrm{E}[(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))^2] \\
&= \frac{1}{\Gamma^2}\mathrm{E}\left[\mathrm{E}[(Y-m_1(X)-\gamma D+\gamma\mu_1(X))^2|X,Z=1]\frac{\mathrm{P}(Z=1|X)}{\pi(X)^2}\right] \\
&\quad+\frac{1}{\Gamma^2}\mathrm{E}\left[\mathrm{E}[(\gamma D-\gamma\mu_0(X)-Y+m_0(X))^2|X,Z=0]\frac{\mathrm{P}(Z=0|X)}{(1-\pi(X))^2}\right] \\
&\quad+\frac{1}{\Gamma^2}\mathrm{E}[(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))^2]
\end{aligned}$$

by iterated expectations. Defining $\sigma^2_{Y_1}(x)=Var[Y|X,Z=1]$, $\sigma^2_{D_1}(x)=Var[D|X,Z=1]$ and $\sigma^2_{Y_1D_1}(x)=Cov[Y,D|X,Z=1]$ and analogously $\sigma^2_{Y_0}(x)$, $\sigma^2_{D_0}(x)$, $\sigma^2_{Y_0D_0}(x)$ for $Z=0$, gives the asymptotic variance bound:

$$\begin{aligned}
&\mathrm{E}\left[\frac{\sigma^2_{Y_1}(X)-2\gamma\sigma^2_{Y_1D_1}(X)+\gamma^2\sigma^2_{D_1}(X)}{\Gamma^2\pi(X)}+\frac{\sigma^2_{Y_0}(X)-2\gamma\sigma^2_{Y_0D_0}(X)+\gamma^2\sigma^2_{D_0}(X)}{\Gamma^2(1-\pi(X))}\right] \\
&\quad+\mathrm{E}\left[\frac{(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))^2}{\Gamma^2}\right].
\end{aligned}$$

A.3. Proof of Theorem 3

Let $\hat\gamma=\hat\Delta/\hat\Gamma$ denote the estimator (8) of the average treatment effect on the compliers (5)

$$\gamma=\frac{\Delta}{\Gamma}=\frac{\int(m_1(x)-m_0(x))f_x(x)\,\mathrm{d}x}{\int(\mu_1(x)-\mu_0(x))f_x(x)\,\mathrm{d}x}.$$

To derive the asymptotic distribution of $\hat\gamma$, note that $\hat\gamma-\gamma$ can be written as

$$(\hat\gamma-\gamma)=\frac{\hat\Delta}{\hat\Gamma}-\frac{\Delta}{\Gamma}=\left(\frac{\hat\Delta-\Delta}{\Gamma}-\gamma\frac{\hat\Gamma-\Gamma}{\Gamma}\right)\cdot\left(1-\frac{\hat\Gamma-\Gamma}{\hat\Gamma}\right). \tag{27}$$

The derivation proceeds in two steps. First it is shown that the last term $(1-(\hat\Gamma-\Gamma)/\hat\Gamma)$ in (27) is $1+\mathrm{o_p}(1)$. Hence the first-order behaviour of $\hat\gamma-\gamma$ is determined by the term $(\hat\Delta-\Delta)/\Gamma-\gamma(\hat\Gamma-\Gamma)/\Gamma$ in

$$\hat\gamma-\gamma=\left(\frac{\hat\Delta-\Delta}{\Gamma}-\gamma\frac{\hat\Gamma-\Gamma}{\Gamma}\right)\cdot\big(1+\mathrm{o_p}(1)\big). \tag{28}$$

In the second step the asymptotic distribution of this first-order term is derived.

In a preliminary step the term $\hat{\Delta} - \Delta$ is analysed. (The derivations for $\hat{\Gamma} - \Gamma$ are analogous.) Write $\hat{\Delta} - \Delta$ as

$$\begin{aligned}\hat{\Delta} - \Delta &= \frac{1}{n}\left(\sum_{i:Z_i=1} Y_i + \sum_{i:Z_i=0} \hat{m}_1(X_i) - \sum_{i:Z_i=0} Y_i - \sum_{i:Z_i=1} \hat{m}_0(X_i)\right) - \mathrm{E}[m_1(X) - m_0(X)] \\ &= \frac{1}{n}\left(\sum_{i:Z_i=1} (Y_i - m_1(X_i)) + \sum_{i:Z_i=0} (\hat{m}_1(X_i) - m_1(X_i))\right) \\ &\quad - \frac{1}{n}\left(\sum_{i:Z_i=0} (Y_i - m_0(X_i)) + \sum_{i:Z_i=1} (\hat{m}_0(X_i) - m_0(X_i))\right) \\ &\quad + \frac{1}{n}\sum_i (m_1(X_i) - m_0(X_i)) - \mathrm{E}[m_1(X) - m_0(X)],\end{aligned}$$

and introducing the asymptotic linear representation of the nonparametric estimators $\hat{m}_z$

$$\begin{aligned}\hat{\Delta} - \Delta &= \frac{1}{n}\sum_{i:Z_i=1} (Y_i - m_1(X_i)) - \frac{1}{n}\sum_{i:Z_i=0} (Y_i - m_0(X_i)) \\ &\quad + \frac{n_0}{n}\frac{1}{n_0 n_1}\sum_{i:Z_i=0}\sum_{j:Z_j=1} \xi_1^m(Y_j, X_j, X_i) + \frac{1}{n}\sum_{i:Z_i=0} [b_1^m(X_i) + R_1^m(X_i)] \\ &\quad - \frac{n_1}{n}\frac{1}{n_0 n_1}\sum_{i:Z_i=1}\sum_{j:Z_j=0} \xi_0^m(Y_j, X_j, X_i) - \frac{1}{n}\sum_{i:Z_i=1} [b_0^m(X_i) + R_0^m(X_i)] \\ &\quad + \frac{1}{n}\sum_i (m_1(X_i) - m_0(X_i)) - \mathrm{E}[m_1(X) - m_0(X)].\end{aligned}$$

The terms $(1/n_0 n_1)\sum_{i:Z_i=0}\sum_{j:Z_j=1}\xi^m$ represent mean-zero two-sample U-statistics, to which a projection theorem can be applied:

$$\begin{aligned}\frac{1}{n_0 n_1}\sum_{i:Z_i=0}\sum_{j:Z_j=1} \xi_1^m(Y_j, X_j, X_i) &= \frac{1}{n_0}\sum_{i:Z_i=0} \mathrm{E}[\xi_1^m(Y_1, X_1, X_2)|X_2 = X_i] \\ &\quad + \frac{1}{n_1}\sum_{j:Z_j=1} \mathrm{E}[\xi_1^m(Y_1, X_1, X_2)|Y_1 = Y_j, X_1 = X_j] + \mathrm{o_p}(n^{-(1/2)}) \\ &= \frac{1}{n_1}\sum_{j:Z_j=1} \mathrm{E}[\xi_1^m(Y_j, X_j, X)|Y_j, X_j] + \mathrm{o_p}(n^{-(1/2)}),\end{aligned}$$

where the latter equality follows from condition (A) of Theorem 3. Application of the projection theorem requires that $\mathrm{E}[\|\xi_1^m(Y_j, X_j, X_i)\|^2] = \mathrm{o}(n)$, see Hoeffding (1948), Serfling (1980) or Heckman et al. (1998), which is satisfied from condition (B) of Theorem 3.

With this projection theorem it follows that

$$\begin{aligned}\hat{\Delta} - \Delta &= \frac{1}{n}\sum_{i:Z_i=1} (Y_i - m_1(X_i)) - \frac{1}{n}\sum_{i:Z_i=0} (Y_i - m_0(X_i)) \\ &\quad + \frac{n_0}{nn_1}\sum_{j:Z_j=1} \mathrm{E}[\xi_1^m(Y_j, X_j, X)|Y_j, X_j] - \frac{n_1}{nn_0}\sum_{j:Z_j=0} \mathrm{E}[\xi_0^m(Y_j, X_j, X)|Y_j, X_j] \\ &\quad + \frac{1}{n}\sum_i (m_1(X_i) - m_0(X_i)) - \mathrm{E}[m_1(X) - m_0(X)] + \mathrm{o_p}(n^{-1/2}),\end{aligned} \tag{29}$$

where it has also been taken into account that the average bias and residual terms $(1/n)\sum b^m(X_i)$ and $(1/n)\sum R^m(X_i)$ are $o_p(n^{-1/2})$ by the conditions (C) and (D) of Theorem 3.

By a weak law of large numbers, sample means converge to their expectations and thus $\hat{\Delta}-\Delta = o_p(1)$ and analogously $\hat{\Gamma}-\Gamma = o_p(1)$. Hence

$$\frac{\hat{\Gamma}-\Gamma}{\hat{\Gamma}} = \left(1+\frac{\hat{\Gamma}-\Gamma}{\Gamma}\right)^{-1}\cdot\frac{\hat{\Gamma}-\Gamma}{\Gamma} = O_p(1)\cdot o_p(1) = o_p(1),$$

because $(1+o_p(1))^{-1} = O_p(1)$ (Van der Vaart, 1998, p. 13). This implies (28).

Hence the leading term in (28) is $((\hat{\Delta}-\Delta)/\Gamma - \gamma(\hat{\Gamma}-\Gamma)/\Gamma)$. The approximation to $\sqrt{n}(\hat{\gamma}-\gamma)$ up to first order is thus

$$\sqrt{n}(\hat{\gamma}-\gamma) = \sqrt{n}\left(\frac{\hat{\Delta}-\Delta}{\Gamma} - \gamma\frac{\hat{\Gamma}-\Gamma}{\Gamma}\right).$$

Inserting the expression (29) and the analogous expression for $\hat{\Gamma}-\Gamma$ gives

$$\sqrt{n}(\hat{\gamma}-\gamma) = \frac{\sqrt{n}}{\Gamma}\frac{1}{n}\sum_i \Xi_i$$

up to first order, where

$$\begin{aligned}\Xi_i = {} & Z_i(Y_i - m_1(X_i)) - (1-Z_i)(Y_i - m_0(X_i)) + m_1(X_i) - m_0(X_i)\\ & + \frac{n_0}{n_1}Z_i\mathrm{E}[\xi_1^m(Y_i,X_i,X)|Y_i,X_i] - \frac{n_1}{n_0}(1-Z_i)\mathrm{E}[\xi_0^m(Y_i,X_i,X)|Y_i,X_i]\\ & - \gamma Z_i(D_i-\mu_1(X_i)) + \gamma(1-Z_i)(D_i-\mu_0(X_i)) - \gamma(\mu_1(X_i)-\mu_0(X_i))\\ & - \frac{\gamma n_0}{n_1}Z_i\mathrm{E}[\xi_1^d(Y_i,X_i,X)|Y_i,X_i] + \frac{\gamma n_1}{n_0}(1-Z_i)\mathrm{E}[\xi_0^d(Y_i,X_i,X)|Y_i,X_i]\\ & - \mathrm{E}[m_1(X)-m_0(X)] + \gamma\mathrm{E}[\mu_1(X)-\mu_0(X)].\end{aligned}$$

Note that the last line is zero, because $-\mathrm{E}[m_1(X)-m_0(X)] + \gamma\mathrm{E}[\mu_1(X)-\mu_0(X)] = -\Delta + \gamma\Gamma = -\Delta + (\Delta/\Gamma)\Gamma = 0$. By condition (E) the influence functions ξ^m and ξ^d have the form

$$\begin{aligned}\mathrm{E}[\xi_1^m(Y_j,X_j,X_i)|Y_j,X_j,Z_j=1,Z_i=0] &= (Y_j-m_1(X_j))\frac{f_{x|z=0}(X_j)}{f_{x|z=1}(X_j)} + o_p(1)\\ &= (Y_j-m_1(X_j))\frac{\mathrm{P}(Z=1)}{\mathrm{P}(Z=0)}\frac{1-\pi(X_j)}{\pi(X_j)} + o_p(1),\end{aligned}$$

where $\pi(x) = \mathrm{P}(Z=1|X=x)$ is the probability that Z takes the value one given characteristics X. With $n_0/n_1 = \mathrm{P}(Z=0)/\mathrm{P}(Z=1) + o_p(1)$ it follows that

$$\begin{aligned}\Xi_i = {} & Z_i\left(Y_i - m_1(X_i) + (Y_i - m_1(X_i))\frac{1-\pi(X_i)}{\pi(X_i)}\right)\\ & - (1-Z_i)\left(Y_i - m_0(X_i) + (Y_i - m_0(X_i))\frac{\pi(X_i)}{1-\pi(X_i)}\right)\\ & - \gamma Z_i\left(D_i - \mu_1(X_i) + (D_i-\mu_1(X_i))\frac{1-\pi(X_i)}{\pi(X_i)}\right)\end{aligned}$$

$$+\gamma(1-Z_i)\left(D_i-\mu_0(X_i)+(D_i-\mu_0(X_i))\frac{\pi(X_i)}{1-\pi(X_i)}\right)$$
$$+m_1(X_i)-m_0(X_i)-\gamma(\mu_1(X_i)-\mu_0(X_i))$$

plus an $o_p(1)$ term. Collecting terms gives

$$\Xi_i = Z_i\left(\frac{(Y_i-m_1(X_i))-\gamma(D_i-\mu_1(X_i))}{\pi(X_i)}\right)+(1-Z_i)\left(\frac{\gamma(D_i-\mu_0(X_i))-(Y_i-m_0(X_i))}{1-\pi(X_i)}\right)$$
$$+m_1(X_i)-m_0(X_i)-\gamma(\mu_1(X_i)-\mu_0(X_i)).$$

Computing the variance of Ξ_i gives

$$\begin{aligned}
&Var(\Xi_i)\\
&=\mathrm{E}\left[Z_i^2\left(\frac{(Y_i-m_1(X_i))-\gamma(D_i-\mu_1(X_i))}{\pi(X_i)}\right)^2\right]\\
&\quad+\mathrm{E}\left[(1-Z_i)^2\left(\frac{\gamma(D_i-\mu_0(X_i))-(Y_i-m_0(X_i))}{1-\pi(X_i)}\right)^2\right]\\
&\quad+\mathrm{E}[(m_1(X_i)-m_0(X_i)-\gamma(\mu_1(X_i)-\mu_0(X_i)))^2]\\
&\quad+\mathrm{E}\left[Z_i\left(\frac{(Y_i-m_1(X_i))-\gamma(D_i-\mu_1(X_i))}{\pi(X_i)}\right)(1-Z_i)\left(\frac{\gamma(D_i-\mu_0(X_i))-(Y_i-m_0(X_i))}{1-\pi(X_i)}\right)\right]\\
&\quad+\mathrm{E}\left[Z_i\left(\frac{(Y_i-m_1(X_i))-\gamma(D_i-\mu_1(X_i))}{\pi(X_i)}\right)(m_1(X_i)-m_0(X_i)-\gamma(\mu_1(X_i)-\mu_0(X_i)))\right]\\
&\quad+\mathrm{E}\left[(1-Z_i)\left(\frac{\gamma(D_i-\mu_0(X_i))-(Y_i-m_0(X_i))}{1-\pi(X_i)}\right)(m_1(X_i)-m_0(X_i)-\gamma(\mu_1(X_i)-\mu_0(X_i)))\right],
\end{aligned}$$

where the fourth term is zero because $Z(1-Z)=0$ and the fifth and sixth terms are zero conditional on X and Z. Since $Z^2=Z$, it follows

$$\begin{aligned}
Var(\Xi)&=\mathrm{E}\left[Z\frac{[(Y-m_1(X))-\gamma(D-\mu_1(X))]^2}{\pi(X)^2}\right]\\
&\quad+\mathrm{E}\left[(1-Z)\frac{[\gamma(D-\mu_0(X))-(Y-m_0(X))]^2}{(1-\pi(X))^2}\right]\\
&\quad+\mathrm{E}[(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))^2]\\
&=\mathrm{E}\left[\mathrm{E}\left[\mathrm{E}[[(Y-m_1(X))-\gamma(D-\mu_1(X))]^2|X,Z=1]\cdot\frac{\mathrm{P}(Z=1|X)}{\pi(X)^2}|X\right]\right]\\
&\quad+\mathrm{E}\left[\mathrm{E}\left[\mathrm{E}[[\gamma(D-\mu_0(X))-(Y-m_0(X))]^2|X,Z=0]\cdot\frac{\mathrm{P}(Z=0|X)}{(1-\pi(X))^2}|X\right]\right]\\
&\quad+\mathrm{E}[(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))^2]\\
&=\mathrm{E}\left[\frac{\sigma^2_{Y_1}(X)-2\gamma\sigma^2_{Y_1D_1}(X)+\gamma^2\sigma^2_{D_1}(X)}{\pi(X)}+\frac{\sigma^2_{Y_0}(X)-2\gamma\sigma^2_{Y_0D_0}(X)+\gamma^2\sigma^2_{D_0}(X)}{1-\pi(X)}\right]\\
&\quad+\mathrm{E}[(m_1(X)-m_0(X)-\gamma\mu_1(X)+\gamma\mu_0(X))^2]
\end{aligned}$$

by iterated expectations, where $\sigma^2_{Y_1}(x) = Var[Y|X, Z=1] = \mathrm{E}[(Y - m_1(X))^2|X=x, Z=1]$, and $\sigma^2_{Y_1D_1}(x) = Cov[Y, D|X, Z=1]$, and $\sigma^2_{D_1}(x)$, $\sigma^2_{Y_0}(x)$, $\sigma^2_{D_0}(x)$, $\sigma^2_{Y_0D_0}(x)$ defined analogously.

Applying the Lindberg–Levy central limit theorem to

$$\sqrt{n}(\hat{\gamma} - \gamma) = \frac{\sqrt{n}}{\Gamma}\frac{1}{n}\sum_i \Xi_i$$

gives that the estimator $\hat{\gamma}$ is root-n asymptotically normal

$$\sqrt{n}(\hat{\gamma} - \gamma) \to \mathrm{N}(0, \mathscr{V})$$

with asymptotic variance

$$\mathscr{V} = \frac{1}{\Gamma^2}\mathrm{E}\left[\frac{\sigma^2_{Y_1}(X) - 2\gamma\sigma^2_{Y_1D_1}(X) + \gamma^2\sigma^2_{D_1}(X)}{\pi(X)} + \frac{\sigma^2_{Y_0}(X) - 2\gamma\sigma^2_{Y_0D_0}(X) + \gamma^2\sigma^2_{D_0}(X)}{1-\pi(X)}\right] + \frac{1}{\Gamma^2}\mathrm{E}[(m_1(X) - m_0(X) - \gamma\mu_1(X) + \gamma\mu_0(X))^2].$$

A.4. Proof of Theorem 4

Consider the local polynomial regression estimator of order p. By Heckman et al. (1998, Theorem 3), if $p = \bar{p}$ and Assumptions 1–3 of (Theorem 4) hold, the local polynomial regression estimator $\hat{m}_1$ is asymptotically linear[21]

$$\hat{m}_1(x) - m_1(x) = \frac{1}{n_1}\sum_{j:Z_j=1} \xi^m_1(Y_j, X_j, x) + b^m_1(x) + R^m_1(x), \tag{30}$$

with $b^m_1(x) = \mathrm{o}(h^{\bar{p}}_{n_1})$ and $(1/\sqrt{n_1})\sum_{j:Z_j=1} R^m_1(x) = \mathrm{o_p}(1)$. Alternatively, if $0 \leqslant p < \bar{p}$ and Assumptions 1–4, 5a and 6 of (Theorem 4) hold, $\hat{m}_1$ is asymptotically linear with $b^m_1(x) = \mathrm{O}(h^{\bar{p}}_{n_1})$.

To specify the influence function, let $x^q = x_1^{q_1}\cdots x_k^{q_k}/(q_1!\cdots q_k!)$ for a row vector $x = (x_1, \ldots, x_k)$ and a multi-index $q = (q_1, \ldots, q_k)$ of non-negative integers. Define a row vector $x^{Q(s)} = (x^q)_{q_1+\cdots+q_k=s}$. This vector of length $(s+k-1)!/s!(k-1)!$ contains all polynomials x^q of degree s. By Heckman et al. (1998, Theorem 3), the influence function is for Nadaraya–Watson regression ($p = 0$)

$$\xi^m_1(Y_j, X_j, x) = (Y_j - m_1(X_j)) \cdot \mathscr{K}_0^{-1} K\left(\frac{X_j - x}{h_{n_1}}\right)$$

and for $p > 0$

$$\xi^m_1(Y_j, X_j, x) = (Y_j - m_1(X_j)) \cdot (1, 0, \ldots, 0)\begin{pmatrix} \mathscr{K}_0 & \mathscr{K}_1 \\ \mathscr{K}_1' & \mathscr{K}_2 \end{pmatrix}^{-1} \times \left\{1, \left(\frac{X_j - x}{h_{n_1}}\right)^{Q(1)}, \ldots, \left(\frac{X_j - x}{h_{n_1}}\right)^{Q(p)}\right\}' K\left(\frac{X_j - x}{h_{n_1}}\right),$$

[21] The derivations for $\hat{m}_0(x)$, $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$ are analogous.

where

$$\mathscr{K}_0 = \kappa_{0,0},$$

$$\mathscr{K}_1 = (\kappa_{0,1}, \ldots, \kappa_{0,p}),$$

$$\mathscr{K}_2 = \begin{pmatrix} \kappa_{1,1} & \cdots & \kappa_{1,p} \\ \vdots & \ddots & \vdots \\ \kappa_{p,1} & \cdots & \kappa_{p,p} \end{pmatrix},$$

and

$$\kappa_{s,t} = \mathrm{E}\left[\left(\frac{X_j - x}{h_{n_1}}\right)^{Q(s)'} \left(\frac{X_j - x}{h_{n_1}}\right)^{Q(t)} K\left(\frac{X_j - x}{h_{n_1}}\right) \middle| Z_j = 1\right]$$

is a matrix of dimension $(s+k-1)!/s!(k-1)! \times (s+k-1)!/s!(k-1)!$.

The following proof is for $p>0$. The derivations for Nadaraya–Watson regression ($p=0$) are analogously. Before analyzing ξ_1^m further, it is useful to derive some properties of the matrix $\kappa_{s,t}$. Let $X_{j,1}$ to $X_{j,k}$ refer to the k elements of the vector X_j. The matrix $\kappa_{s,t}$ can be written as

$$\begin{aligned}\kappa_{s,t} &= \int \cdots \int \left(\frac{X_j - x}{h_{n_1}}\right)^{Q(s)'} \left(\frac{X_j - x}{h_{n_1}}\right)^{Q(t)} K\left(\frac{X_j - x}{h_{n_1}}\right) f_{x|z=1}(X_j)\,\mathrm{d}X_{j,1} \cdots \mathrm{d}X_{j,k} \\ &= h_{n_1}^k \int \cdots \int u^{Q(s)'} u^{Q(t)} K(u) f_{x|z=1}(x + uh_{n_1})\,\mathrm{d}u_1 \cdots \mathrm{d}u_k\end{aligned}$$

with the change in variables $u = (X_j - x)/h_{n_1}$.

Examine first the 1×1 matrix $\kappa_{0,0}$. By a Taylor series expansion, $f_{x|z=1}(x + uh_{n_1})$ can be written as $f_{x|z=1}(x) + \mathrm{O}(h_{n_1})$ which gives:

$$\begin{aligned}\mathscr{K}_0 = \kappa_{0,0} &= h_{n_1}^k \int \cdots \int K(u) f_{x|z=1}(x + uh_{n_1})\,\mathrm{d}u_1 \cdots \mathrm{d}u_k \\ &= h_{n_1}^k \int \cdots \int K(u) f_{x|z=1}(x)\,\mathrm{d}u_1 \cdots \mathrm{d}u_k + \mathrm{O}(h_{n_1}^{k+1}) = h_{n_1}^k f_{x|z=1}(x)\lambda_0 + \mathrm{O}(h_{n_1}^{k+1}),\end{aligned} \tag{31}$$

where

$$\lambda_0 = \int \cdots \int K(u)\,\mathrm{d}u_1 \cdots \mathrm{d}u_k.$$

Now consider the matrix $\kappa_{s,t}$. Suppose for the moment that the first $c \geqslant 1$ moments of the kernel function are zero. If $s + t > c$, then

$$\kappa_{s,t} = h_{n_1}^k f_{x|z=1}(x) \int \cdots \int u^{Q(s)'} u^{Q(t)} K(u)\,\mathrm{d}u_1 \cdots \mathrm{d}u_k + \mathrm{O}(h_{n_1}^{k+1}).$$

Otherwise, if $s + t \leqslant c$, the first terms of a Taylor series expansion vanish. Supposing that the first $c + 1 - s - t$ derivatives of $f_{x|z=1}$ exist, the elements of the matrix $\kappa_{s,t}$ are of the order $\mathrm{O}(h_{n_1}^{c+1-s-t}) \cdot h_{n_1}^k$. Taken together, the elements of the matrix $\kappa_{s,t}$ are of the order

$$\mathrm{O}(h_{n_1}^{\max(c+1-s-t,0)}) \cdot h_{n_1}^k.$$

Hence, the largest order of any element of the vector $\mathscr{K}_1$ is

$$\mathscr{K}_1 = \mathrm{O}(h_{n_1}^{\max(c+1-p,0)}) \cdot h_{n_1}^k,$$

whereas the largest order of any element of the matrix $\mathscr{K}_2$ is

$$\mathscr{K}_2 = \mathrm{O}(h_{n_1}^{\max(c+1-2p,0)}) \cdot h_{n_1}^k.$$

Hence, as long as $c \geqslant p$, the largest element of the vector $\mathscr{K}_1$ is of order $h_{n_1}^{k+1}$ or smaller. In addition, it is of smaller order than the largest element of $\mathscr{K}_2$. This gives the following important intermediate result:

$$\text{If } c \geqslant p \quad \text{then } \mathscr{K}_1 = \mathrm{O}(h_{n_1}^{k+1}) \text{ and } \mathscr{K}_1 \mathscr{K}_2^{-1} = \mathrm{O}(h_{n_1}) \tag{32}$$

or of even smaller order. By Assumption 5a and b, respectively, of Theorem 4, $c \geqslant p$ and (32) is satisfied.

With these preliminaries, the influence function ξ_1^m can be examined further. Writing ε_j for $(Y_j - m_1(X_j))$ and using partitioned inverses, ξ_1^m can be written as

$$\begin{aligned}\xi_1^m(Y_j, X_j, x) = {} & \frac{\varepsilon_j}{\mathscr{K}_0 - \mathscr{K}_1 \mathscr{K}_2^{-1} \mathscr{K}_1'} \cdot \left(1, -\mathscr{K}_1 \mathscr{K}_2^{-1}\right) \\ & \times \left\{1, \left(\frac{X_j - x}{h_{n_1}}\right)^{Q(1)}, \ldots, \left(\frac{X_j - x}{h_{n_1}}\right)^{Q(p)}\right\}' K\left(\frac{X_j - x}{h_{n_1}}\right).\end{aligned}$$

Inserting the expression (31) for $\mathscr{K}_0$ and noting from (32) that $\mathscr{K}_1 \mathscr{K}_2^{-1} \mathscr{K}_1'$ and $\mathscr{K}_1 \mathscr{K}_2^{-1}$ are of lower order, it follows:

$$\begin{aligned}\xi_1^m(Y_j, X_j, x) = {} & \frac{\varepsilon_j}{h_{n_1}^k f_{x|z=1}(x)\lambda_0 + \mathrm{O}(h_{n_1}^{k+1})} \\ & \times \left(1 + \mathrm{O}(h_{n_1}) \left\{\left(\frac{X_j - x}{h_{n_1}}\right)^{Q(1)}, \ldots, \left(\frac{X_j - x}{h_{n_1}}\right)^{Q(p)}\right\} \iota\right) \cdot K\left(\frac{X_j - x}{h_{n_1}}\right),\end{aligned}$$

where ι is a column vector of ones.

For this expression, the condition (iv) of Theorem 3 needs to be verified. The conditions (A), (C) and (D) follow directly from Heckman et al. (1998, Theorem 3). It remains to verify conditions (B) and (E). Verify first condition (E):

$$\begin{aligned}& \mathrm{E}\left[\xi_1^m(Y_j, X_j, X_i) \middle| Y_j, X_j, Z_j = 1, Z_i = 0\right] \\ & \quad = \int \cdots \int \xi_1^m(Y_j, X_j, X_i) f_{x|z=0}(X_i)\, \mathrm{d}X_{i,1} \cdots \mathrm{d}X_{i,k} \\ & \quad = \int \cdots \int \frac{\varepsilon_j}{h_{n_1}^k f_{x|z=1}(X_i)\lambda_0 + \mathrm{O}_p(h_{n_1}^{k+1})} K\left(\frac{X_j - X_i}{h_{n_1}}\right) \\ & \qquad \times \left(1 + \mathrm{O}_\mathrm{p}(h_{n_1}) \left\{\left(\frac{X_j - X_i}{h_{n_1}}\right)^{Q(1)}, \ldots, \left(\frac{X_j - X_i}{h_{n_1}}\right)^{Q(p)}\right\} \iota\right) f_{x|z=0}(X_i)\, \mathrm{d}X_{i,1} \cdots \mathrm{d}X_{i,k}.\end{aligned}$$

With the change in variables $u = (X_i - X_j)/h_{n_1}$ and a Taylor series expansion $f_{x|z=0}(X_j + uh_{n_1}) = f_{x|z=0}(X_j) + \mathrm{O_p}(h_{n_1})$ and analogously for $f_{x|z=1}$, it follows:

$$
\begin{aligned}
&= \varepsilon_j \int \cdots \int \frac{f_{x|z=0}(X_j + uh_{n_1})}{f_{x|z=1}(X_j + uh_{n_1})\lambda_0 + \mathrm{O_p}(h_{n_1})} \\
&\quad \times \big(1 + \mathrm{O_p}(h_{n_1})\{(-u)^{Q(1)}, \ldots, (-u)^{Q(p)}\}\iota\big) \cdot K(-u)\,\mathrm{d}u_1 \cdots \mathrm{d}u_k \\
&\quad = \varepsilon_j \frac{f_{x|z=0}(X_j) + \mathrm{O_p}(h_{n_1})}{f_{x|z=1}(X_j)\lambda_0 + \mathrm{O_p}(h_{n_1})} \cdot \int \cdots \int (1 + \mathrm{O_p}(h_{n_1}) \\
&\qquad \times \{(-u)^{Q(1)}, \ldots, (-u)^{Q(p)}\}\iota) \cdot K(u)\,\mathrm{d}u_1 \cdots \mathrm{d}u_k \\
&\quad = \varepsilon_j \frac{f_{x|z=0}(X_j) + \mathrm{O_p}(h_{n_1})}{f_{x|z=1}(X_j)\lambda_0 + \mathrm{O_p}(h_{n_1})} \left\{ \lambda_0 + \mathrm{O_p}(h_{n_1}) \int \cdots \int \right. \\
&\qquad \left. \times \{(-u)^{Q(1)}, \ldots, (-u)^{Q(p)}\}\iota \cdot K(u)\,\mathrm{d}u_1 \cdots \mathrm{d}u_k \right\} \\
&\quad = \varepsilon_j \frac{f_{x|z=0}(X_j) + \mathrm{O_p}(h_{n_1})}{f_{x|z=1}(X_j)\lambda_0 + \mathrm{O_p}(h_{n_1})} \{\lambda_0 + \mathrm{O_p}(h_{n_1})\} \\
&\quad = (Y_j - m_1(X_j)) \frac{f_{x|z=0}(X_j)}{f_{x|z=1}(X_j)} + \mathrm{O_p}(h_{n_1}).
\end{aligned}
$$

Condition (B) is a rather weak condition and can be verified using a similar argument.

$$
\begin{aligned}
&\mathrm{E}[\xi_1^m (Y_j, X_j, X_i)^2 | Z_j = 1, Z_i = 0] \\
&\quad = \int \cdots \int \frac{\sigma^2_{Y_1}(X_j)}{h_{n_1}^{2k} f^2_{x|z=1}(X_i)\lambda_0^2 + \mathrm{O_p}(h_{n_1}^{2k+1})} K\left(\frac{X_j - X_i}{h_{n_1}}\right)^2 \\
&\qquad \times \left(1 + \mathrm{O_p}(h_{n_1}) \left\{ \left(\frac{X_j - X_i}{h_{n_1}}\right)^{Q(1)}, \ldots, \left(\frac{X_j - X_i}{h_{n_1}}\right)^{Q(p)} \right\} \iota \right)^2 \\
&\qquad \times f_{x|z=0}(X_i)\,\mathrm{d}X_{i,1} \cdots \mathrm{d}X_{i,k}.
\end{aligned}
$$

With the change in variables $u = (X_i - X_j)/h_{n_1}$ and a Taylor series expansion $f_{x|z=0}(X_j + uh_{n_1}) = f_{x|z=0}(X_j) + \mathrm{O_p}(h_{n_1})$ and analogously for $f_{x|z=1}$ it follows:

$$
\begin{aligned}
&= \frac{\sigma^2_{Y_1}(X_j)}{h_{n_1}^k \lambda_0^2} \int \cdots \int \frac{f_{x|z=0}(X_j + uh_{n_1}) K(u)^2}{f^2_{x|z=1}(X_j) + \mathrm{O_p}(h_{n_1})} \\
&\quad \times (1 + \mathrm{O_p}(h_{n_1})\{(-u)^{Q(1)}, \ldots, (-u)^{Q(p)}\}\iota)^2\,\mathrm{d}u_1 \cdots \mathrm{d}u_k \\
&\qquad = \frac{\sigma^2_{Y_1}(X_j) f_{x|z=0}(X_j)}{h_{n_1}^k \lambda_0^2 f^2_{x|z=1}(X_j)} \int \cdots \int K(u)^2\,\mathrm{d}u_1 \cdots \mathrm{d}u_k + \mathrm{O_p}(h_{n_1}) = \mathrm{O_p}(1).
\end{aligned}
$$

A.5. Proof of Theorem 5

The proof is based on Theorem 6.1 of Newey (1994) and on Newey (1995). To apply Newey's theorem it is useful to express the conditional LATE estimator using Newey's notation. The conditional LATE estimator can be written as an *m*-estimator

which solves

$$\frac{1}{n}\sum\{\hat{m}_1(X_i) - \hat{m}_0(X_i) - \hat{\gamma}\hat{\mu}_1(X_i) + \hat{\gamma}\hat{\mu}_0(X_i)\} = 0$$

based on the moment condition:

$$\mathrm{E}[\mathrm{E}[Y|X, Z=1] - \mathrm{E}[Y|X, Z=0] - \gamma\mathrm{E}[D|X, Z=1] + \gamma\mathrm{E}[D|X, Z=0]] = 0.$$

Since $\mathrm{E}[YZ|X] = \mathrm{E}[Y|X, Z=1]\cdot\mathrm{E}[Z|X]$ for Z binary, the moment condition can also be expressed as

$$\mathrm{E}\left[\frac{\mathrm{E}[YZ|X]}{\mathrm{E}[Z|X]} - \frac{\mathrm{E}[Y(1-Z)|X]}{\mathrm{E}[1-Z|X]} - \gamma\frac{\mathrm{E}[DZ|X]}{\mathrm{E}[Z|X]} + \gamma\frac{\mathrm{E}[D(1-Z)|X]}{\mathrm{E}[1-Z|X]}\right] = 0.$$

To use Newey's notation, write this moment condition as

$$\mathrm{E}[m(X_i, \gamma, h)] = 0,$$

where the moment function m is

$$m(x_i, \gamma, h) = \frac{h_1(x_i)}{h_5(x_i)} - \frac{h_2(x_i)}{1-h_5(x_i)} - \gamma\frac{h_3(x_i)}{h_5(x_i)} + \gamma\frac{h_4(x_i)}{1-h_5(x_i)} \tag{33}$$

with $h(x) = (h_1(x), \ldots, h_5(x))'$ and the true conditional expectation functions:

$$h_1(x) = \mathrm{E}[YZ|X = x],$$

$$h_2(x) = \mathrm{E}[Y(1-Z)|X = x],$$

$$h_3(x) = \mathrm{E}[DZ|X = x],$$

$$h_4(x) = \mathrm{E}[D(1-Z)|X = x],$$

$$h_5(x) = \mathrm{E}[Z|X = x].$$

In the following, $h = h(x)$ refers to the vector of true conditional expectation functions, whereas $\tilde{h} = \tilde{h}(x)$ refers to any other function from an appropriate space that includes h.

Define the residuals $u_1 = YZ - \mathrm{E}[YZ|X]$ and $u_2 = Y(1-Z) - \mathrm{E}[Y(1-Z)|X]$ and $u_3 = DZ - \mathrm{E}[DZ|X]$ and $u_4 = D(1-Z) - \mathrm{E}[D(1-Z)|X]$ and $u_5 = Z - \mathrm{E}[Z|X]$. Let $u = (u_1, \ldots, u_5)'$ be the vector of residuals.

Let $\hat{h}(x) = \hat{\eta}' p^K(x)$ be a series estimator of $h(x)$, where $\hat{\eta}$ is a $K \times 5$ matrix of coefficients. $\hat{\eta}$ is computed as

$$\hat{\eta} = \left(p^{K\prime}p^K\right)^{-} p^{K\prime}\mathbf{Y},$$

where $\mathbf{Y}$ is the $n \times 5$ data matrix with the ith row being

$$(Y_iZ_i,\ Y_i(1-Z_i),\ D_iZ_i,\ D_i(1-Z_i),\ Z_i).$$

In the following conditions, $|.|$ refers to the Euclidean norm and $|.|_d$ refers to the supremum Sobolev norm. The Sobolev norm for a vector of functions $\varsigma(x)$ is defined as

$$|\varsigma|_d = \max_{|q|\leqslant d}\sup_{x\in Supp(X)}|\partial^q\varsigma(x)|$$

with $|\varsigma|_d$ equal to infinity if $\partial^q\varsigma(x)$ does not exist for some $|q| \leqslant d$. Here, $q = (q_1, \ldots, q_k)'$ is a vector of non-negative integers and $\partial^q\varsigma(x) = \partial^{|q|}\varsigma(x)/\partial x_1^{q_1}\cdots\partial x_k^{q_k}$ is the partial derivative.

Define the magnitude of the basis functions $p^K(x)$ by

$$\zeta_d(K) = \sup_{|q|=d, x\in Supp(X)} |\partial^q p^K(x)|.$$

To apply the Theorem 6.1 of Newey (1994), the following conditions, corresponding to his Assumptions 5.4–5.6 and 6.1–6.6, need to be satisfied. In the conditions, γ and h always refer to the *true* value of LATE and the true conditional expectation functions, whereas $\tilde{\gamma}$ and $\tilde{h}$ stand for any arbitrary values and functions from an appropriate set, respectively.

Condition 1. There are ε, $b(x)$, $\tilde{b}(x)>0$ and a compact subset $\mathfrak{B}$ of $\mathfrak{R}$ with $\gamma \in \mathfrak{B}$ such that for all $\tilde{\gamma} \in \mathfrak{B}$:

(i) $m(x,\tilde{\gamma},h)$ is continuous at $\tilde{\gamma}$ with probability one;
(ii) $|m(x,\tilde{\gamma},h)|_d \leqslant b(x)$;
(iii) $|m(x,\tilde{\gamma},\tilde{h}) - m(x,\tilde{\gamma},h)|_d \leqslant \tilde{b}(x)\left(|\tilde{h}-h|_d\right)^{\varepsilon}$.

Condition 2. $\mathrm{E}[m(x,\tilde{\gamma},h)] = 0$ has a unique solution on $\mathfrak{B}$ at γ.

Condition 3. (i) The solution γ is an interior point of $\mathfrak{B}$.
(ii) There is a neighbourhood of γ such that for all $|\tilde{h}-h|_d<\varepsilon$ the functional $m(x,\tilde{\gamma},\tilde{h})$ is differentiable in $\tilde{\gamma}$ on this neighbourhood,
(iii) $\mathrm{E}[\partial m(x,\tilde{\gamma},h)/\partial\tilde{\gamma}]_{|\tilde{\gamma}=\gamma} \neq 0$,
(iv) Condition 1 is also satisfied for $m(x,\tilde{\gamma},\tilde{h})$ replaced by $\partial m(x,\tilde{\gamma},\tilde{h})/\partial\tilde{\gamma}$,
(v) $\mathrm{E}[|m(x,\gamma,h)|_d^2] < \infty$.

Condition 4. $\mathrm{E}[|u|^2|X]$ is bounded, where $u = (u_1,\ldots,u_5)'$.

Condition 5. (i) The smallest eigenvalue of $\mathrm{E}[p^K(X)p^K(X)']$ is bounded away from zero uniformly in K,
(ii) $p^K(x)$ is a subvector of $p^{K+1}(x)$ for all K,
(iii) For each K there exists a nonzero $\bar{\eta}$ such that $\bar{\eta}'p^K(x)$ is a nonzero constant on the support of X.

Condition 6. For a nonnegative integer d, if $|h|_d$ is finite then there are constants C, $\alpha_d>0$ such that for all K there is η with $|h(x)-\eta'p^K(x)|_d \leqslant CK^{-\alpha_d}$.

Condition 7. (i) There is a function $D(x,\tilde{h})$ that is linear in $\tilde{h}$ such that for all $\tilde{h}$ with $|\tilde{h}-h|_d$ small enough, $|m(x,\gamma,\tilde{h}) - m(x,\gamma,h) - D(x,\tilde{h}-h)|_d \leqslant b(x)|\tilde{h}-h|_d^2$.
(ii) $\mathrm{E}[b(X)]\zeta_d(K)(\sqrt{\frac{K}{n}} + K^{-\alpha}) \longrightarrow 0$ and $\mathrm{E}[b(X)]\sqrt{n}\zeta_d(K)^2(K/n + K^{-2\alpha}) \longrightarrow 0$.

Condition 8. There is $b(x)$ and $d>0$ with $\mathrm{E}[b(X)^2]<\infty$ such that

(i) $|D(x,\tilde{h})| \leqslant b(x)|\tilde{h}|_d$,
(ii) $K^{-2\alpha} \longrightarrow 0$,
(iii) $(\sum_{l=1}^K |p_{lK}|_d^2)^{1/2}(\sqrt{K/n} + K^{-\alpha}) \longrightarrow 0$.

Condition 9. There is $\delta(x)$ such that for all $\tilde{h}$

(i) $\mathrm{E}[D(x,\tilde{h})] = \mathrm{E}[\delta(x)'\tilde{h}]$
(ii) For each K there are $K\times 5$ matrices η_K and v_K such that $n\cdot\mathrm{E}[|\delta(X) - v_K'p^K(X)\ |^2]$

$\mathrm{E}[|h(X)-\eta_K' p^K(X)|^2]\longrightarrow 0$ and $K(\zeta_0(K)^4/n)\longrightarrow 0$ and $\zeta_0(K)^2\mathrm{E}[|h(X)-\eta_K' p^K(X)|^2]\longrightarrow 0$ and $\mathrm{E}[|\delta(X)-v_K' p^K(X)|^2]\longrightarrow 0$.

Conditions 1–3 correspond to Assumption 5.4–5.6 of Newey (1994). The Conditions 4–9 correspond to Assumptions 6.1–6.6 of Newey (1994), with the exception of Condition 7(i), which corresponds to his Assumption 5.1(i). (This is a weaker version of his Assumption 6.4(i), which he invokes to show consistency of the asymptotic variance estimator.)

By Theorem 6.1 of Newey (1994), it follows that $\sqrt{n}(\hat{\gamma}-\gamma)\xrightarrow{\mathrm{d}}\mathrm{N}(0,V)$ with

$$V=\frac{1}{\Gamma^2}\mathrm{E}[\{m(X_i,\gamma,h)+\delta(X_i)'u_i\}^2], \tag{34}$$

where u are the residuals defined above and $\delta(x)$ is

$$\delta(x)=\begin{pmatrix} \dfrac{1}{h_5(x)} \\ -\dfrac{1}{1-h_5(x)} \\ -\dfrac{\gamma}{h_5(x)} \\ \dfrac{\gamma}{1-h_5(x)} \\ -\dfrac{h_1(x)}{h_5^2(x)}-\dfrac{h_2(x)}{(1-h_5(x))^2}+\gamma\dfrac{h_3(x)}{h_5^2(x)}+\gamma\dfrac{h_4(x)}{(1-h_5(x))^2} \end{pmatrix}. \tag{35}$$

It is shown below that the above assumptions are satisfied for this $\delta(x)$.

With this expression $\delta(x)$, the asymptotic variance (34) can be calculated. Inserting (33) and the definition of the residuals u gives

$$\begin{aligned} &m(X,\gamma,h)+\delta(X)'u \\ &\quad=\frac{Z}{h_5(X)}\left(Y-\frac{h_1(X)}{h_5(X)}-\gamma D+\frac{\gamma h_3(X)}{h_5(X)}\right) \\ &\qquad+\frac{1-Z}{1-h_5(X)}\left(\gamma D-\frac{\gamma h_4(X)}{1-h_5(X)}-Y+\frac{h_2(X)}{1-h_5(X)}\right) \\ &\qquad+\frac{h_1(X)}{h_5(X)}-\frac{h_2(X)}{1-h_5(X)}-\frac{\gamma h_3(X)}{h_5(X)}+\frac{\gamma h_4(X)}{1-h_5(X)}. \end{aligned}$$

The asymptotic variance (34) is

$$\begin{aligned} &\frac{1}{\Gamma^2}\mathrm{E}[\{m(X_i,\gamma,h)+\delta(X_i)'u_i\}^2] \\ &\quad=\frac{1}{\Gamma^2}\mathrm{E}\left[\frac{Z^2}{h_5^2(X)}\left(Y-\frac{h_1(X)}{h_5(X)}-\gamma D+\frac{\gamma h_3(X)}{h_5(X)}\right)^2\right] \\ &\qquad+\frac{1}{\Gamma^2}\mathrm{E}\left[\frac{(1-Z)^2}{(1-h_5(X))^2}\left(\gamma D-\frac{\gamma h_4(X)}{1-h_5(X)}-Y+\frac{h_2(X)}{1-h_5(X)}\right)^2\right] \\ &\qquad+\frac{1}{\Gamma^2}\mathrm{E}\left[\left(\frac{h_1(X)}{h_5(X)}-\frac{h_2(X)}{1-h_5(X)}-\frac{\gamma h_3(X)}{h_5(X)}+\frac{\gamma h_4(X)}{1-h_5(X)}\right)^2\right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\Gamma^2}\mathrm{E}\left[\frac{\mathrm{P}(Z=1|X)}{h_5^2(X)}\mathrm{E}\left[\left(Y-\frac{h_1(X)}{h_5(X)}-\gamma D+\frac{\gamma h_3(X)}{h_5(X)}\right)^2\middle| X, Z=1\right]\right] \\
&\quad + \frac{1}{\Gamma^2}\mathrm{E}\left[\frac{\mathrm{P}(Z=0|X)}{(1-h_5(X))^2}\mathrm{E}\left[\left(\gamma D-\frac{\gamma h_4(X)}{1-h_5(X)}-Y+\frac{h_2(X)}{1-h_5(X)}\right)^2\middle| X, Z=0\right]\right] \\
&\quad + \frac{1}{\Gamma^2}\mathrm{E}\left[\left(\frac{h_1(X)}{h_5(X)}-\frac{h_2(X)}{1-h_5(X)}-\frac{\gamma h_3(X)}{h_5(X)}+\frac{\gamma h_4(X)}{1-h_5(X)}\right)^2\right] \\
&= \frac{1}{\Gamma^2}\mathrm{E}\left[\frac{\sigma^2_{Y_1}-2\gamma\sigma^2_{Y_1D_1}+\gamma^2\sigma^2_{D_1}}{h_5(X)}\right]+\frac{1}{\Gamma^2}\mathrm{E}\left[\frac{\sigma^2_{Y_0}-2\gamma\sigma^2_{Y_0D_0}+\gamma^2\sigma^2_{D_0}}{1-h_5(X)}\right] \\
&\quad + \frac{1}{\Gamma^2}\mathrm{E}\left[\left(\frac{h_1(X)}{h_5(X)}-\frac{h_2(X)}{1-h_5(X)}-\frac{\gamma h_3(X)}{h_5(X)}+\frac{\gamma h_4(X)}{1-h_5(X)}\right)^2\right],
\end{aligned}
$$

which, when returning to the notation of Section 3, corresponds to the expression (9).

It remains to verify the above conditions. Condition 1 is verified for $d=0$. Condition 1(i) is satisfied since $m(x,\tilde{\gamma},h)$ is linear in $\tilde{\gamma}$. Condition 1(ii) is satisfied for $b(x)$ a sufficiently large constant because h_1/h_5, $h_2/(1-h_5)$, h_3/h_5, $h_4/(1-h_5)$ are bounded by Assumption (iii) and because $\mathfrak{B}$ is bounded.

For condition 1(iii) and $\varepsilon=1$

$$
\begin{aligned}
&|m(x,\tilde{\gamma},\tilde{h})-m(x,\tilde{\gamma},h)|_0 \\
&\quad = \left|\frac{\tilde{h}_1}{\tilde{h}_5}-\frac{\tilde{h}_2}{1-\tilde{h}_5}-\tilde{\gamma}\frac{\tilde{h}_3}{\tilde{h}_5}+\tilde{\gamma}\frac{\tilde{h}_4}{1-\tilde{h}_5}-\left(\frac{h_1}{h_5}-\frac{h_2}{1-h_5}-\tilde{\gamma}\frac{h_3}{h_5}+\tilde{\gamma}\frac{h_4}{1-h_5}\right)\right|_0 \\
&\quad \leqslant \left|\tilde{h}-h\right|_0\cdot\tilde{b}
\end{aligned}
$$

with

$$
\tilde{b}=\left|\frac{h_5-h_1+\tilde{\gamma}h_3-\tilde{\gamma}h_5}{\tilde{h}_5h_5}\right|_0+\left|\frac{h_2+(1-h_5)-\tilde{\gamma}h_4-\tilde{\gamma}(1-h_5)}{(1-h_5)(1-\tilde{h}_5)}\right|_0,
$$

where $\tilde{b}$ is a finite constant because h_5 is bounded away from zero.

Condition 2 is satisfied by assumption. Condition 3(i) is satisfied by an appropriate choice of $\mathfrak{B}$. Condition 3(ii) is trivially satisfied since m is linear in $\tilde{\gamma}$. The derivative $\partial m(x,\tilde{\gamma},\tilde{h})/\partial\tilde{\gamma}$ is $\tilde{h}_4/(1-\tilde{h}_5)-\tilde{h}_3/\tilde{h}_5$. For condition 3(iii) note that its expected value $\mathrm{E}[\partial m(x,\tilde{\gamma},h)/\partial\tilde{\gamma}]_{|\tilde{\gamma}=\gamma}$ at the true γ and the true h is $-\Gamma$, which is nonzero by the assumption that LATE is identified. The verification of condition 3(iv) is analogous to condition 1. Condition 3(v) is satisfied by $m(x,\gamma,h)$being bounded on the support of X.

Condition 4 is satisfied by Assumption (iv).

Condition 5 is satisfied for power series by Assumptions (vi) and (vii) and by Lemma A.15 of Newey (1995), which further implies $\zeta_d(K)\leqslant CK^{1+2d}$ and $|p_{kK}|_d\leqslant CK^{1/2+2d}$.

Condition 6 is satisfied for power series for $d=0$ and $\alpha=s/k$ by e.g. Lemma A.12 of Newey (1995).

Condition 7 is verified for $d = 0$. Begin with condition 7(i) and let

$$D(x,\tilde{h}) = \frac{\tilde{h}_1(x)}{h_5(x)} - \frac{\tilde{h}_2(x)}{1-h_5(x)} - \gamma\frac{\tilde{h}_3(x)}{h_5(x)} + \gamma\frac{\tilde{h}_4(x)}{1-h_5(x)} - \tilde{h}_5(x)\left(\frac{h_1(x)}{h_5^2(x)} + \frac{h_2(x)}{(1-h_5(x))^2} - \gamma\frac{h_3(x)}{h_5^2(x)} - \gamma\frac{h_4(x)}{(1-h_5(x))^2}\right)$$

be the functional derivative. Inserting $D(x,\tilde{h})$ in condition 7(i) gives after some calculations

$$\begin{aligned}
&m(x,\gamma,\tilde{h}) - m(x,\gamma,h) - D(x,\tilde{h}-h)\\
&\quad= (\tilde{h}_5(x)-h_5(x))\left\{\frac{\gamma\tilde{h}_3(x)-\tilde{h}_1(x)}{\tilde{h}_5(x)h_5(x)} + \frac{\gamma\tilde{h}_4(x)-\tilde{h}_2(x)}{(1-h_5(x))(1-\tilde{h}_5(x))}\right\}\\
&\qquad+ (\tilde{h}_5(x)-h_5(x))\left(\frac{h_1(x)}{h_5^2(x)} + \frac{h_2(x)}{(1-h_5(x))^2} - \gamma\frac{h_3(x)}{h_5^2(x)} - \gamma\frac{h_4(x)}{(1-h_5(x))^2}\right)\\
&\quad= \frac{h_1-\gamma h_3}{\tilde{h}_5 h_5^2}(\tilde{h}_5-h_5)^2 + \frac{\gamma}{\tilde{h}_5 h_5}(\tilde{h}_5-h_5)(\tilde{h}_3-h_3) - \frac{1}{\tilde{h}_5 h_5}(\tilde{h}_5-h_5)(\tilde{h}_1-h_1)\\
&\qquad+ \frac{(\gamma h_4-h_2)(\tilde{h}_5-h_5)^2}{(1-\tilde{h}_5)(1-h_5)^2} + \gamma\frac{(\tilde{h}_5-h_5)(\tilde{h}_4-h_4)}{(1-\tilde{h}_5)(1-h_5)} - \frac{(\tilde{h}_5-h_5)(\tilde{h}_2-h_2)}{(1-\tilde{h}_5)(1-h_5)},
\end{aligned}$$

where in the last equation the dependence of h and of $\tilde{h}$ on x is suppressed. This gives

$$\begin{aligned}
&|m(x,\gamma,\tilde{h}) - m(x,\gamma,h) - D(x,\tilde{h}-h)|_0\\
&\leqslant \left|\frac{h_1-\gamma h_3}{\tilde{h}_5 h_5^2}\right|_0 |(\tilde{h}_5-h_5)^2|_0 + \left|\frac{\gamma}{\tilde{h}_5 h_5}\right|_0 |(\tilde{h}_5-h_5)(\tilde{h}_3-h_3)|_0\\
&\quad+ \left|\frac{1}{\tilde{h}_5 h_5}\right|_0 |(\tilde{h}_5-h_5)(\tilde{h}_1-h_1)|_0 + \left|\frac{\gamma h_4-h_2}{(1-\tilde{h}_5)(1-h_5)^2}\right|_0 |(\tilde{h}_5-h_5)^2|_0\\
&\quad+ \left|\frac{1}{(1-\tilde{h}_5)(1-h_5)}\right|_0 \{|\gamma||(\tilde{h}_5-h_5)(\tilde{h}_4-h_4)|_0 + |(\tilde{h}_5-h_5)(\tilde{h}_2-h_2)|_0\}\\
&\leqslant b\cdot|\tilde{h}(x)-h(x)|_0^2,
\end{aligned}$$

where $b = |(h_1-\gamma h_3)/\tilde{h}_5 h_5^2|_0 + |(\gamma h_4-h_2)/(1-\tilde{h}_5)(1-h_5)^2|_0 + (1+|\gamma|)(|1/\tilde{h}_5 h_5|_0 + |1/(1-\tilde{h}_5)(1-h_5)|_0)$ is a positive non-infinite constant b is non-infinite, because by Assumption (iii) the functions h_1/h_5, $h_2/(1-h_5)$, h_3/h_5 and $h_4/(1-h_5)$ are bounded on the support of X and h_5 is bounded away from 0 and 1. For $\tilde{h}$ sufficiently close to h, also $\tilde{h}_5$ will be bounded away from 0 and 1, which bounds the expressions $|1/\tilde{h}_5 h_5|_0$ and $|1/(1-\tilde{h}_5)(1-h_5)|_0$.

To verify Condition 7(ii) for $d = 0$, note that $b(x)$ is a positive non-infinite constant and that $\alpha = s/k$ and $\zeta_0(K)\leqslant CK$. Hence,

$$\mathrm{E}[b(X)]\zeta_0(K)\left(\sqrt{\frac{K}{n}} + K^{-\alpha}\right) = \mathrm{O}\left(\sqrt{\frac{K^3}{n}}\right) + \mathrm{O}(K^{1-s/k})$$

and

$$\mathrm{E}[b(X)]\sqrt{n}\zeta_0(K)^2\left(\frac{K}{n}+K^{-2\alpha}\right)=\mathrm{O}\left(\sqrt{\frac{K^6}{n}}\right)+\mathrm{O}\left(\sqrt{nK^{4(1-s/k)}}\right).$$

All terms have to be of order o(1). This requires that K increases at a smaller rate than $n^{1/6}$, that $s>k$ and that K increases at a larger rate than $n^{0.25/(s/k-1)}$. This is implied by Assumptions (v) and (viii) of Theorem 5.

Condition 8 is verified for $d=1$. Examine first $|D(x,\tilde{h})|$ with $D(x,\tilde{h})$ given above

$$|D(x,\tilde{h})|\leqslant|\tilde{h}(x)|_0 b(x)\leqslant|\tilde{h}(x)|_1 b(x),$$

where

$$b(x)=\left|\frac{1+\gamma}{h_5(x)}+\frac{1+\gamma}{1-h_5(x)}+\frac{h_1(x)}{h_5^2(x)}+\frac{h_2(x)}{(1-h_5(x))^2}-\gamma\frac{h_3(x)}{h_5^2(x)}-\gamma\frac{h_4(x)}{(1-h_5(x))^2}\right|.$$

Because h_1/h_5, $h_2/(1-h_5)$, h_3/h_5, $h_4/(1-h_5)$ are assumed to be bounded, $\mathrm{E}[b(X)^2]<\infty$.

Condition 8(ii) is satisfied if $s/k>0$. For Condition 8(iii) note that $|p_{kK}|_d\leqslant CK^{1/2+2d}$. Hence,

$$\left(\sum_{l=1}^{K}|p_{lK}|_d^2\right)^{1/2}\leqslant C\left(\sum_{l=1}^{K}K^{1+4d}\right)^{1/2}\leqslant C(K^{2+4d})^{1/2}=CK^{1+2d}.$$

This gives

$$\left(\sum_{l=1}^{K}|p_{lK}|_1^2\right)^{1/2}\sqrt{\frac{K}{n}}\leqslant C\sqrt{\frac{K^7}{n}}\longrightarrow 0$$

if K increases slower than $n^{1/7}$, and

$$\left(\sum_{l=1}^{K}|p_{lK}|_1^2\right)^{1/2}K^{-s/k}\leqslant CK^{3-s/k}\longrightarrow 0$$

if $s/k>3$.

Condition 9(i) is satisfied with $\delta(x)$ given by (35). For Condition 9(ii) note that by condition 6

$$|h(x)-\eta_K' p^K(x)|_0\leqslant CK^{-s/k}.$$

Since $\delta(x)$ is continuously differentiable of the same order as $h(x)$, it follows by Lemma A.12 of Newey (1995) that

$$|\delta(x)-\nu_K' p^K(x)|_0\leqslant CK^{-s/k}.$$

Because $|\cdot|_0$ is the supremum norm, it follows that

$$\begin{aligned}&n\cdot\mathrm{E}[|\delta(X)-\nu_K' p^K(X)|^2]\mathrm{E}[|h(X)-\eta_K' p^K(X)|^2]\\&\quad\leqslant n\cdot(|\delta(X)-\nu_K' p^K(X)|_0)^2(|h(X)-\eta_K' p^K(X)|_0)^2\\&\quad=\mathrm{O}(nK^{-4s/k}).\end{aligned}$$

For this term to vanish requires K to increase at a rate faster than $n^{1/4s/k}$, which is implied by the rate conditions.

Second, for the term

$$K\frac{\zeta_0(K)^4}{n} = \mathrm{O}\left(\frac{K^5}{n}\right)$$

to vanish requires K to increase slower than $n^{1/5}$, which is implied.

Third,

$$\begin{aligned}&\zeta_0(K)^2\mathrm{E}[|h(X)-\eta_K' p^K(X)|^2]\\ &\quad\leqslant \zeta_0(K)^2(|h(X)-\eta_K' p^K(X)|_0)^2\\ &\quad= \mathrm{O}(K^2K^{-2s/k}) = \mathrm{O}(K^{2(1-s/k)})\longrightarrow 0\end{aligned}$$

if $s/k>1$.

Finally,

$$\begin{aligned}&\mathrm{E}[|\delta(X)-v_K' p^K(X)|^2]\\ &\quad\leqslant (|\delta(X)-v_K' p^K(X)|_0)^2 = \mathrm{O}(K^{-2s/k})\longrightarrow 0\end{aligned}$$

if $s/k>0$.

The necessary conditions on the rate of K were that $s/k>3$ and that $K=n^v$ with $1/4(s/k-1)<v<\frac{1}{7}$.

A.6. Proof of Theorem 6

With knowledge of the propensity score, the parametric submodel becomes

$$f(y,d,z,x|\theta) = \{f_1(y,d|x,\theta)\pi(x)\}^z\{f_0(y,d|x,\theta)(1-\pi(x))\}^{1-z}f(x,\theta),$$

with the corresponding score

$$S(y,d,z,x|\theta) = \frac{\partial \ln f(y,d,z,x|\theta)}{\partial\theta} = z\cdot\check{f}_1(y,d|x,\theta)+(1-z)\cdot\check{f}_0(y,d|x,\theta)+\check{f}(x,\theta)$$

and tangent space

$$\mathfrak{I} = \{z\cdot s_1(y,d|x)+(1-z)\cdot s_0(y,d|x)+s_x(x)\}.$$

Note that the pathwise derivative $\partial\gamma(F_\theta)/\partial_{\theta|\theta=\theta_0}$ is the same as in Theorem 2. Choose $\psi(Y,D,Z,X)$ as in (26). After some tedious calculations, it can be verified first that

$$\frac{\partial\gamma(F_\theta)}{\partial\theta}_{|\theta=\theta_0} = \mathrm{E}[\psi(Y,D,Z,X)\cdot S(Y,D,Z,X)]_{|\theta=\theta_0}$$

and second that ψ lies in the tangent space of the model

$$\psi\in\mathfrak{I}.$$

Since ψ lies in the tangent space, the variance bound is the expected square of ψ. Because ψ is the same as in Theorem 2, the asymptotic variance bound is $\mathscr{V}$.

A.7. Proof of Theorem 7

Rewrite the expressions $\sigma^2_{\pi Y_1}(\pi)$ and $\sigma^2_{\pi Y_1 D_1}(\pi)$ as

$$\begin{aligned}
\sigma^2_{\pi Y_1}(\rho) &= \mathrm{E}[(Y - m_{\pi 1}(\rho))^2 | \pi(X) = \rho, Z = 1] \\
&= \mathrm{E}[\mathrm{E}[(Y - m_{\pi 1}(\rho) - m_1(X) + m_1(X))^2 | X, Z = 1] | \pi(X) = \rho, Z = 1] \\
&= \mathrm{E}[\mathrm{E}[(Y - m_1(X))^2 | X, Z = 1] | \pi(X) = \rho, Z = 1] \\
&\quad + \mathrm{E}[\mathrm{E}[(m_1(X) - m_{\pi 1}(\rho))^2 | X, Z = 1] | \pi(X) = \rho, Z = 1] \\
&\quad + \mathrm{E}[\mathrm{E}[((Y - m_1(X))(m_1(X) - m_{\pi 1}(\rho))) | X, Z = 1] | \pi(X) = \rho, Z = 1] \\
&= \int \{\sigma^2_{Y_1}(x) + (m_1(x) - m_{\pi 1}(\rho))^2\} f_{x|\pi,z}(x|\rho, 1)\, \mathrm{d}x \\
&= \mathrm{E}[\sigma^2_{Y_1}(X) + (m_1(X) - m_{\pi 1}(\rho))^2 | \pi(X) = \rho] \\
&= \mathrm{E}[\sigma^2_{Y_1}(X) | \pi(X) = \rho] + Var[m_1(X) | \pi(X) = \rho]
\end{aligned}$$

because

$$f_{x|\pi,z}(x|\rho, 1) = \frac{\mathrm{P}(Z = 1 | \pi = \rho, X = x) f_{x|\pi}(x|\rho)}{\mathrm{P}(Z = 1 | \pi = \rho)} = f_{x|\pi}(x|\rho).$$

Similarly,

$$\begin{aligned}
\sigma^2_{\pi Y_1 D_1}(\rho) &= \mathrm{E}[(Y - m_{\pi 1}(\rho))(D - \mu_{\pi 1}(\rho)) | \pi(X) = \rho, Z = 1] \\
&= \mathrm{E}[\mathrm{E}[(Y - m_1(X) + m_1(X) - m_{\pi 1}(\rho)) \\
&\quad \times (D - \mu_1(X) + \mu_1(X) - \mu_{\pi 1}(\rho)) | X, Z = 1] | \pi(X) = \rho, Z = 1] \\
&= \mathrm{E}[\sigma^2_{Y_1 D_1}(X) + (m_1(X) - m_{\pi 1}(\rho))(\mu_1(X) - \mu_{\pi 1}(\rho)) | \pi(X) = \rho, Z = 1] \\
&= \mathrm{E}[\sigma^2_{Y_1 D_1}(X) | \pi(X) = \rho] + Cov(m_1(X), \mu_1(X) | \pi(X) = \rho).
\end{aligned}$$

Furthermore with

$$\begin{aligned}
&Var(m_1(X) - m_0(X) - \gamma\mu_1(X) + \gamma\mu_0(X) | \pi) \\
&\quad = \mathrm{E}[(m_1(X) - m_0(X) - \gamma\mu_1(X) + \gamma\mu_0(X))^2 | \pi] - (m_{\pi 1}(\pi) - m_{\pi 0}(\pi) - \gamma\mu_{\pi 1}(\pi) \\
&\quad + \gamma\mu_{\pi 0}(\pi))^2
\end{aligned}$$

it follows that

$$\begin{aligned}
&\mathrm{E}[(m_{\pi 1}(\pi) - m_{\pi 0}(\pi) - \gamma\mu_{\pi 1}(\pi) + \gamma\mu_{\pi 0}(\pi))^2] \\
&\quad = \mathrm{E}[(m_1(X) - m_0(X) - \gamma\mu_1(X) + \gamma\mu_0(X))^2] \\
&\qquad - \mathrm{E}[Var(m_1(X) - m_0(X) - \gamma\mu_1(X) + \gamma\mu_0(X) | \pi)].
\end{aligned}$$

By inserting these expressions in (15), the difference in the variance of propensity score matching and the variance of matching on X can be written after some calculations as

$$V_{\pi m} - V = \frac{1}{\Gamma^2} \mathrm{E}\left[\frac{1 - \pi}{\pi} V_1(\pi) + \frac{\pi}{1 - \pi} V_0(\pi) + 2C(\pi)\right],$$

where $V_1(\pi) = Var(m_1(X) - \gamma\mu_1(X) | \pi)$ and $V_0(\pi) = Var(m_0(X) - \gamma\mu_0(X) | \pi)$ and $C(\pi) = Cov((m_1(X) - \gamma\mu_1(X)), (m_0(X) - \gamma\mu_0(X)) | \pi)$. The two variance terms are necessarily non-negative, but the covariance term could be negative. It is to be shown that the

whole expression $V_{\pi m} - \mathscr{V}$ is non-negative. Note that the covariance is bounded in absolute value by the variances: $|C(\pi)| \leqslant \sqrt{V_1(\pi)V_0(\pi)}$. This gives a lower bound on $V_{\pi m} - \mathscr{V}$

$$V_{\pi m} - \mathscr{V} \geqslant \frac{1}{I^2}\mathrm{E}\left[\frac{1-\pi}{\pi}V_1(\pi) + \frac{\pi}{1-\pi}V_0(\pi) - 2\sqrt{V_1(\pi)V_0(\pi)}\right]$$
$$= \frac{1}{I^2}\mathrm{E}\left[V_0(\pi)(\sqrt{V_1(\pi)/V_0(\pi)} - \pi(1+\sqrt{V_1(\pi)/V_0(\pi)}))^2/\pi(1-\pi)\right].$$

Since all terms are non-negative it follows that $V_{\pi m} - \mathscr{V} \geqslant 0$. Generally, $V_{\pi m} - \mathscr{V}$ is strictly positive unless the support of $\pi(x)$ contains only values where the variances $V_1(\pi)$ and $V_0(\pi)$ are both zero or where $\sqrt{V_1(\pi)/V_0(\pi)} - \pi(1+\sqrt{V_1(\pi)/V_0(\pi)}) = 0$ holds. In this special case, $V_{\pi m} = \mathscr{V}$.

A.8. Proof of Theorem 8

The proof of Theorem 8 is illustrated for Z being discrete and taking values $Z \in \{0,1,2\}$. The extension to discrete Z with more than three distinct values in its support is obvious. An extension to continuous Z is also straightforward and similar to the derivations in Heckman and Vytlacil (2001). First, it is shown that the average treatment effect for the compliers with characteristics X can be expressed as

$$\mathrm{E}[Y^1 - Y^0|X, \tau = c] = \frac{\mathrm{E}[Y|X, Z = z_{\max}] - \mathrm{E}[Y|X, Z = z_{\min}]}{\mathrm{P}(\tau = c|X)}. \tag{36}$$

With Z taking values in $\{0,1,2\}$ and by monotonicity (Assumption 1′), the subpopulation of compliers consists of two groups: those who switch D when Z is increased from 0 to 1 (i.e. $D_{i,0} < D_{i,1} = D_{i,2}$) and those who switch D when Z is increased from 1 to 2 (i.e. $D_{i,0} = D_{i,1} < D_{i,2}$). By monotonicity each individual switches at most once, and therefore these two subpopulations are a partition of the complier subpopulation. The treatment effect for compliers with characteristics X is

$$\mathrm{E}[Y^1 - Y^0|X, \tau = c] = \mathrm{E}[Y^1 - Y^0|X, D_0 < D_1] \cdot \mathrm{P}(D_0 < D_1|X, \tau = c)$$
$$+ \mathrm{E}[Y^1 - Y^0|X, D_1 < D_2] \cdot \mathrm{P}(D_1 < D_2|X, \tau = c). \tag{37}$$

By Assumptions 1′, 3′ and 4′ and derivations similar to those used in Theorem 1, the treatment effect in the first subpopulation is identified as

$$\mathrm{E}[Y^1 - Y^0|X, D_0 < D_1] = \frac{\mathrm{E}[Y|X, Z = 1] - \mathrm{E}[Y|X, Z = 0]}{\mathrm{P}(D_0 < D_1|X)} \tag{38}$$

if $\mathrm{P}(D_0 < D_1|X) > 0$. The treatment effect in the second subpopulation is

$$\mathrm{E}[Y^1 - Y^0|X, D_1 < D_2] = \frac{\mathrm{E}[Y|X, Z = 2] - \mathrm{E}[Y|X, Z = 1]}{\mathrm{P}(D_1 < D_2|X)}, \tag{39}$$

if $\mathrm{P}(D_1 < D_2|X) > 0$. Note furthermore that

$$\mathrm{P}(D_0 < D_1|X, D_0 < D_2) = \frac{\mathrm{P}(D_0 < D_1|X)}{\mathrm{P}(D_0 < D_2|X)}$$

and analogously

$$\mathrm{P}(D_1 < D_2 | X, D_0 < D_2) = \frac{\mathrm{P}(D_1 < D_2 | X)}{\mathrm{P}(D_0 < D_2 | X)},$$

provided $\mathrm{P}(D_0 < D_2 | X) > 0$.

Inserting these two expressions in (37) yields

$$\begin{aligned}
\mathrm{E}[Y^1 - Y^0 | X, \tau = c] &= \frac{\mathrm{E}[Y|X, Z=1] - \mathrm{E}[Y|X, Z=0]}{\mathrm{P}(D_0 < D_2|X)} \\
&\quad + \frac{\mathrm{E}[Y|X, Z=2] - \mathrm{E}[Y|X, Z=1]}{\mathrm{P}(D_0 < D_2|X)} \\
&= \frac{\mathrm{E}[Y|X, Z=2] - \mathrm{E}[Y|X, Z=0]}{\mathrm{P}(D_0 < D_2|X)} \\
&= \frac{\mathrm{E}[Y|X, Z=z_{\max}] - \mathrm{E}[Y|X, Z=z_{\min}]}{\mathrm{P}(\tau = c|X)},
\end{aligned}$$

which equals (36).

The LATE can now be derived by an analogous reasoning as in Theorem 1 as

$$\begin{aligned}
\mathrm{E}[Y^1 - Y^0 | \tau = c] &= \int \mathrm{E}[Y^1 - Y^0 | X, \tau = c] \, \mathrm{d}F_{x|\tau=c} \\
&= \int \mathrm{E}[Y^1 - Y^0 | X, \tau = c] \frac{\mathrm{P}(\tau = c|X)}{\mathrm{P}(\tau = c)} \, \mathrm{d}F_x \\
&= \frac{\int (\mathrm{E}[Y|X=x, Z=z_{\max}] - \mathrm{E}[Y|X=x, Z=z_{\min}]) f_x(x) \, \mathrm{d}x}{\mathrm{P}(\tau = c)}.
\end{aligned}$$

To identify the denominator, note that

$$\begin{aligned}
\mathrm{E}[D|X, Z=2] &= \mathrm{P}(D_2 = 1 | X, Z = 2) \\
&= \mathrm{P}(D_2 = 1, D_0 = 0 | X, Z = 2) + \mathrm{P}(D_2 = 1, D_0 = 1 | X, Z = 2) \\
&= \mathrm{P}(D_2 = 1, D_0 = 0 | X) + \mathrm{P}(D_2 = 1, D_0 = 1 | X)
\end{aligned}$$

by Assumption 3′. Analogously

$$\begin{aligned}
\mathrm{E}[D|X, Z=0] &= \mathrm{P}(D_0 = 1 | X, Z = 0) \\
&= \mathrm{P}(D_0 = 1, D_2 = 0 | X, Z = 0) + \mathrm{P}(D_0 = 1, D_2 \\
&= 1 | X, Z = 0) \\
&= \mathrm{P}(D_0 = 1, D_2 = 0 | X) + \mathrm{P}(D_0 = 1, D_2 = 1 | X) \\
&= \mathrm{P}(D_0 = 1, D_2 = 1 | X)
\end{aligned}$$

because $\mathrm{P}(D_0 = 1, D_2 = 0 | X) = 0$ by monotonicity (Assumption 1′). This gives

$$\mathrm{E}[D|X, Z=2] - \mathrm{E}[D|X, Z=0] = \mathrm{P}(D_0 < D_2 | X) = \mathrm{P}(\tau = c|X).$$

Using $\mathrm{P}(\tau = c) = \int \mathrm{P}(\tau = c|X) \, \mathrm{d}F_x$, the LATE can be expressed as

$$\mathrm{E}[Y^1 - Y^0 | \tau = c] = \frac{\int (\mathrm{E}[Y|X=x, Z=z_{\max}] - \mathrm{E}[Y|X=x, Z=z_{\min}]) f_x(x) \, \mathrm{d}x}{\int (\mathrm{E}[D|X=x, Z=z_{\max}] - \mathrm{E}[D|X=x, Z=z_{\min}]) f_x(x) \, \mathrm{d}x}.$$

A.9. Proof of Theorem 9

In a first step, it is shown that $\gamma_w(x)$ is identified as

$$\gamma_w(X) = \frac{\mathrm{E}[Y|X, Z=1] - \mathrm{E}[Y|X, Z=0]}{\mathrm{E}[D|X, Z=1] - \mathrm{E}[D|X, Z=0]}. \tag{40}$$

To see this, note that the population is partitioned by $\tau = c_{k,l}$ for $k \leqslant l$. Hence $\sum_{k\leqslant l} \mathrm{P}(\tau = c_{k,l}|X) = 1$ and $\mathrm{E}[Y|X, Z=1] = \sum_{k\leqslant l} \mathrm{E}[Y|X, Z=1, \tau = c_{k,l}]\mathrm{P}(\tau = c_{k,l}|\, X, Z=1) = \sum_{k\leqslant l} \mathrm{E}[Y^l|X, \tau = c_{k,l}]\mathrm{P}(\tau = c_{k,l}|X)$ by the exclusion and the unconfoundedness assumption. Analogously, $\mathrm{E}[Y|X, Z=0] = \sum_{k\leqslant l} \mathrm{E}[Y^k|X, \tau = c_{k,l}]\mathrm{P}(\tau = c_{k,l}|X)$ and $\mathrm{E}[D|X, Z=1] = \sum_{k\leqslant l} l \cdot \mathrm{P}(\tau = c_{k,l}|X)$ and $\mathrm{E}[D|X, Z=0] = \sum_{k\leqslant l} k \cdot \mathrm{P}(\tau = c_{k,l}|X)$.

In a second step, it is to show that the weighted treatment effect $\gamma_w = \int \gamma_w(x) \cdot f^w_{x|\text{complier}}(x)\,\mathrm{d}x$ is given by (22). Using Bayes' theorem $f_{x|\tau=c_{k,l}} = \mathrm{P}(\tau = c_{k,l}|X) f_x / \mathrm{P}(\tau = c_{k,l})$, the weighted distribution function of X in the all-compliers subpopulation is

$$f^w_{x|\text{complier}} = \frac{\sum_k^K \sum_{l>k}^K \mathrm{P}(\tau = c_{k,l}|X) \cdot (l-k)}{\sum_k^K \sum_{l>k}^K \mathrm{P}(\tau = c_{k,l}) \cdot (l-k)} \cdot f_x, \tag{41}$$

where $f_x(x)$ is the density function of X in the full population. With this weighted distribution function the treatment effect in the subpopulation of all compliers is

$$\begin{aligned}
\gamma_w &= \int \gamma_w(x) \cdot f^w_{x|\text{complier}}(x)\,\mathrm{d}x \\
&= \int \frac{\mathrm{E}[Y|X=x, Z=1] - \mathrm{E}[Y|X=x, Z=0]}{\sum_k \sum_{l>k} \mathrm{P}(\tau = c_{k,l}) \cdot (l-k)} \cdot f_x(x)\,\mathrm{d}x \\
&= \frac{\int \mathrm{E}[Y|X=x, Z=1] - \mathrm{E}[Y|X=x, Z=0] \cdot f_x(x)\,\mathrm{d}x}{\int \sum_k \sum_{l>k} \mathrm{P}(\tau = c_{k,l}|X=x) \cdot (l-k) \cdot f_x(x)\,\mathrm{d}x} \\
&= \frac{\int (\mathrm{E}[Y|X=x, Z=1] - \mathrm{E}[Y|X=x, Z=0]) \cdot f_x(x)\,\mathrm{d}x}{\int (\mathrm{E}[D|X=x, Z=1] - \mathrm{E}[D|X=x, Z=0]) \cdot f_x(x)\,\mathrm{d}x}.
\end{aligned}$$

References

Abadie, A., 2003. Semiparametric instrumental variable estimation of treatment response models. Journal of Econometrics 113, 231–263.

Abadie, A., Imbens, G., 2006. Large sample properties of matching estimators for average treatment effects. Econometrica 74, 235–267.

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Cambridge, MA.

Angrist, J., 1990. Lifetime earnings and the vietnam era draft lottery: evidence from social security administrative records. American Economic Review 80, 313–336.

Angrist, J., Imbens, G., 1995. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. Journal of American Statistical Association 90, 431–442.

Angrist, J., Krueger, A., 1999. Empirical strategies in labor economics. In: Ashenfelter, O., Card, D. (Eds.), The Handbook of Labor Economics. North-Holland, New York, pp. 1277–1366.

Angrist, J., Imbens, G., Rubin, D., 1996. Identification of causal effects using instrumental variables. Journal of American Statistical Association 91, 444–472 (with discussion).

Angrist, J., Graddy, K., Imbens, G., 2000. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. Review of Economic Studies 67, 499–527.

Begun, J., Hall, W., Huang, W., Wellner, J., 1983. Information and asymptotic efficiency in parametric-nonparametric models. Annals of Statistics 11, 432–452.
Bickel, P., Klaassen, C.A.J., Ritov, Y., Wellner, J.A., 1993. Efficient and Adaptive Estimation for Semiparametric Models. John Hopkins University Press, Baltimore.
Blundell, R., Powell, J., 2003. Endogeneity in nonparametric and semiparametric regression models. In: Hansen, L., Dewatripont, M., Turnovsky, S.J. (Eds.), Advances in Economics and Econometrics. Cambridge University Press, Cambridge, pp. 312–357.
Card, D., 1995. Using geographic variation in college proximity to estimate the return to schooling. In: Christofides, L., Grant, E., Swidinsky, R. (Eds.), Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp. University of Toronto Press, Toronto, pp. 201–222.
Chernozhukov, V., Hansen, C., 2005. An IV model of quantile treatment effects. Econometrica 73, 245–261.
Chesher, A., 2003. Identification in nonseparable models. Econometrica 71, 1405–1441.
Chesher, A., 2005. Nonparametric identification under discrete variation. Econometrica 73, 1525–1550.
Darolles, S., Florens, J., Renault, E., 2001. Nonparametric instrumental regression. Mimeo, Toulouse.
Das, M., 2005. Instrumental variables estimators of nonparametric models with discrete endogenous regressors. Journal of Econometrics 124, 335–361.
Florens, J., 2003. Inverse problems and structural econometrics: the example of instrumental variables. In: Hansen, L., Dewatripont, M., Turnovsky, S.J. (Eds.), Advances in Economics and Econometrics. Cambridge University Press, Cambridge, pp. 284–311.
Florens, J., Heckman, J., Meghir, C., Vytlacil, E., 2002. Instrumental variables, local instrumental variables and control functions. cemmap Working Paper 15/02.
Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66, 315–331.
Hearst, N., Newman, T., Hulley, S., 1986. Delayed effects of the military draft on mortality: a randomized natural experiment. New England Journal of Medicine 314, 620–624.
Heckman, J., 1997. Instrumental variables—a study of implicit behavioral assumptions used in making program evaluations. Journal of Human Resources 32, 441–462.
Heckman, J., Vytlacil, E., 1999. Local instrumental variables and latent variable models for identifying and bounding treatment effects. Proceedings National Academic Sciences, USA Economic Sciences 96, 4730–4734.
Heckman, J., Vytlacil, E., 2001. Local instrumental variables. In: Hsiao, C., Morimune, K., Powell, J. (Eds.), Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya. Cambridge University Press, Cambridge.
Heckman, J., Ichimura, H., Todd, P., 1998. Matching as an econometric evaluation estimator. Review of Economic Studies 65, 261–294.
Heckman, J., LaLonde, R., Smith, J., 1999. The economics and econometrics of active labour market programs. In: Ashenfelter, O., Card, D. (Eds.), The Handbook of Labor Economics. North-Holland, New York, pp. 1865–2097.
Hirano, K., Imbens, G., Rubin, D., Zhou, X., 2000. Assessing the effect of an influenza vaccine in an encouragement design. Biostatistics 1, 69–88.
Hirano, K., Imbens, G., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71, 1161–1189.
Hoeffding, W., 1948. A class of statistics with asymptotically normal distribution. Annals of Mathematical Statistics 19, 293–325.
Imbens, G., 2001. Some remarks on instrumental variables. In: Lechner, M., Pfeiffer, F. (Eds.), Econometric Evaluation of Labour Market Policies. Physica/Springer, Heidelberg, pp. 17–42.
Imbens, G., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–475.
Imbens, G., Newey, W., 2003. Identification and estimation of triangular simultaneous equations models without additivity. Presented at the EC2 Conference, London, December 2003.
Imbens, G., Rubin, D., 1997. Estimating outcome distributions for compliers in instrumental variables models. Review of Economic Studies 64, 555–574.
Imbens, G., Rubin, D., Sacerdote, B., 2001. Estimating the effect of unearned income on labor earnings, savings, and consumption: evidence from a survey of lottery players. American Economic Review 91, 778–794.
Koshevnik, Y.A., Levit, B.Y., 1976. On a non-parametric analogue of the information matrix. Theory of Probability and Applications 21, 738–753.
Newey, W., 1990. Semiparametric efficiency bounds. Journal of Applied Econometrics 5, 99–135.

Newey, W., 1994. The asymptotic variance of semiparametric estimators. Econometrica 62, 1349–1382.
Newey, W., 1995. Convergence rates for series estimators. In: Maddala, G.S., Phillips, P., Srinivasan, T.N. (Eds.), Advances in Econometrics and Quantitative Economics: Essays in Honor of Professor C.R. Rao. Blackwell, Oxford, pp. 254–275.
Newey, W., Powell, J., 2003. Instrumental variable estimation of nonparametric models. Econometrica 71, 1565–1578.
Newey, W., Powell, J., Vella, F., 1999. Nonparametric estimation of triangular simultaneous equations models. Econometrica 67, 565–603.
Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge.
Pfanzagl, J., Wefelmeyer, W., 1982. Contributions to a General Asymptotic Statistical Theory. Springer, Heidelberg.
Rubin, D., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.
Serfling, R.J., 1980. Approximation Theorems of Mathematical Statistics. Wiley, New York.
Stein, C., 1956. Efficient nonparametric testing and estimation. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, vol. 1. University of California Press, Berkeley, pp. 1–2.
van der Vaart, A.W., 1998. Asymptotic Statistics. Cambridge University Press, Cambridge.
Yau, L., Little, R., 2001. Inference for the complier-average causal effect from longitudinal data subject to noncompliance and missing data, with application to a job training assessment for the unemployed. Journal of American Statistical Association 96, 1232–1244.