A kernel-based first stage in linear instrumental variable models

Thomas Vigié[*†]

January 9, 2022

Abstract

In a linear model with an endogenous variable, I propose an estimator that uses a kernel estimator in the first stage regression to build predictions of the endogenous variable. The estimator is consistent, asymptotically normal and also includes the two-stage least squares (2SLS) estimator as a special case.

I study the effect of the bandwidth used in the first stage on the mean squared error of the estimator in a simulation exercise, and compare it to some existing estimators such as the OLS and 2SLS estimators. It competes with them in all settings, and in particular outperforms the 2SLS estimator when endogeneity is not severe and/or the relationship between the endogenous variable and the instruments is nonlinear. I apply the estimator on the data set used in Chalfin (2015), and find evidence that corroborates the OLS estimates for some types of crime, but also the 2SLS ones for other types.

Keywords: instrumental variable, kernel estimators, linear models.

1 Introduction

Linear models are prevalent in the field of applied econometrics, and often characterized by the endogeneity of the variable of interest. In this setting, the most used estimator is the two stage least squares estimator (2SLS) that considers instruments correlated with the endogenous variable and that are exogenous. The estimator regresses the endogenous variable on the instruments (and possibly the other exogenous variables of the model) to construct projected values of the endogenous variable. Those values are then used to estimate the equation of interest. This estimator is consistent but biased and can feature a high variance. It is still the go-to estimator however (see Young, 2017), despite the appearance of a large body of literature on alternative estimation procedures. Some of them do not consider the problem as a two stage one, such as k-class estimators[1] (Theil, 1958), and Antoine & Lavergne (2014)'s Weighted Minimum Distance estimator.[2]

Two stage problems essentially involves two sub problems: the first stage is a prediction problem, whereas the second stage is an estimation one. A linear projection in the first stage, such as what the 2SLS estimator does, is one way of dealing with the first stage prediction problem. As an alternative, Belloni *et al.* (2012) consider a large set of instruments and use the LASSO estimator in the first stage.

This paper proposes to estimate the first stage regression with a kernel estimator, and use the resulting fitted values to estimate the parameter of interest. The flexibility of nonparametric estimators makes it a candidate of choice to construct predictions using instruments. I show the consistency and asymptotic normality of the estimator which is asymptotically efficient in the sense that its asymptotic variance reaches the efficiency bound characterized by Chamberlain (1987). Moreover, Lemma 1 shows that using a local linear

[*]8888 University Drive, V5A 1S6, Burnaby, Canada. Email address: tvigie@sfu.ca.

[†]Department of Economics, Simon Fraser University. I would like to thank Professor Bertille Antoine, Professor Chris Muris and Ricardo Meilman Lomaz Cohn for our numerous fruitful meetings and conversations, their patience and invaluable feedback. I gratefully acknowledge the financial support brought by the Peter Kennedy Memorial Graduate Fellowship for the academic year 2019-2020.

[1]k-class estimators include the 2SLS, OLS and LIML estimators as special cases. In the 2SLS case, one can also obtain it via two separate regressions due to the symmetry and idempotence of projection matrices.

[2]Their estimator is based on the first conditional moment (3) in Assumption 1.

estimator with a sufficiently large bandwidth in the first stage coincides with the 2SLS estimator, making it a special case. Thus, one should expect kernels to provide advantageous properties over the traditionally used estimators through the choice of the bandwidth. Simulation evidence shows it is the case in particular when the relationship between the endogenous variable and the instrument is nonlinear, and when the degree of endogeneity is low. I consider the data set used in Chalfin (2015), and apply the estimator to find evidence that corroborates the OLS estimates in some cases, and the 2SLS ones in other cases.

This paper resonates with the literature on efficiency properties of instrumental variable estimators based on moment conditions, studied in Jorgenson & Laffont (1974) and Chamberlain (1987) where a bound for the asymptotic variance is defined and the optimal instruments leading to that variance are characterized. Newey (1990) proposes to estimate those optimal instruments via series approximation and nearest neighbors (in the homoskedastic case), whereas Linton (2002) considers using kernel estimators.

More generally, estimators including preliminary estimated parameters are the object of the literature about semiparametric two step estimators, defined through moment conditions that include a nuisance parameter needing to be preliminary estimated. Pakes & Pollard (1989) and Pakes & Olley (1995) show consistency and asymptotic normality of such estimators, applied to the estimation of production functions in Olley & Pakes (1996) and Levinsohn & Petrin (2003). Some results can also be found in Newey & McFadden (1994). More recently, Cattaneo & Jansson (2018) derive a distributional approximation for kernel-based semiparametric estimators which include a bias term that they propose to correct for using the bootstrap.

The paper is organized as follows: section 2 presents the model and the estimator, section 3 displays the asymptotic results, section 4 is a simulation study followed by section 5 that goes over Chalfin (2015) and section 6 concludes. The proofs of the theorems are displayed in the appendix.

2 Model and estimator:

The model of interest is introduced in this section. I consider a linear model where the researcher is interested in the coefficient associated with an endogenous variable $x_{1,i}$.

Assumption 1. *The random variables satisfy the following conditions*

$$y_i = a + x_{1,i}b + x_{2,i}'c + u_i, \tag{1}$$
$$x_{1,i} = g_0(z_i) + v_i, \tag{2}$$
$$\mathbb{E}[u_i|z_i, x_{2,i}] = \mathbb{E}[v_i|z_i, x_{2,i}] = 0. \tag{3}$$

where $x_{1,i}$ is of dimension 1, $x_{2,i}$ is of dimension d, and y_i and z_i are univariate.

For convenience, I rewrite the first equation as $y_i = x_i'\beta_0 + u_i$ where $x_i = \begin{pmatrix} 1 & x_{1,i} & x_{2,i}' \end{pmatrix}'$ and $\beta_0 = (a \;\; b \;\; c')'$.

Note that the conditional mean independence $\mathbb{E}[u_i|z_i] = 0$ and $\mathbb{E}[u_i|x_{2,i}] = 0$ can be replaced by the more specific assumptions $\mathbb{E}[g_0(z_i)u_i] = 0$ and $\mathbb{E}[x_{2,i}u_i] = 0$ and the same results apply throughout the paper. Besides, this setup can accommodate for more endogenous variables if similar equations as (2) and (3) are defined for each endogenous variable and associated error terms.

Let f_Z denote the probability density function of z_i, $\mathcal{Z}$ the support of z_i, $i = 1, ..., n$ and let $\mathbb{V}[u_i|z_i, x_{2,i}] \equiv \sigma^2(z_i, x_{2,i}) < \infty$. It will be useful to denote the first and second derivatives of a function f by f' and f''. If A is a matrix or a vector, A' denotes its transpose.

Assumption 2. *An i.i.d. random sample of size n is available:* $\{(y_i, x_i, z_i), i = 1, ..., n\}$.

Let X be the $n \times (d+2)$ matrix that stacks the x_i' vertically. Let

$$Y \equiv \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X \equiv \begin{pmatrix} 1 & x_{1,1} & x'_{2,1} \\ 1 & x_{1,2} & x'_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x'_{2,n} \end{pmatrix},$$

$$\hat{X} \equiv \begin{pmatrix} 1 & \hat{g}(z_1) & x'_{2,1} \\ 1 & \hat{g}(z_2) & x'_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & \hat{g}(z_n) & x'_{2,n} \end{pmatrix},$$

$$x_{0,i} \equiv \begin{pmatrix} 1 & g_0(z_i) & x'_{2,i} \end{pmatrix}',$$

where $\hat{g}(z_i), i = 1, ..., n$ is a kernel estimator of $g_0(z_i)$. Expression (3) in Assumption 1 implies that $\mathbb{E}[x_{0,i}u_i] = 0$. The estimator is then defined as:

$$\hat{\beta} \equiv \left(\hat{X}'X\right)^{-1} \hat{X}'Y, \tag{4}$$

and I call it the kernel instrumental variable (k-IV) estimator. In its essence, it is a just-identified IV estimator using $\hat{g}(z_i)$ as the instrument. In the remainder of the paper, I consider two different kernel estimators of $g_0(z_i)$, that nevertheless share the same properties in terms of bandwidth behavior: the Nadaraya-Watson kernel estimator (also called local constant estimator) and the local linear estimator. Those estimators are linear estimators and take the matrix form $\hat{X} \equiv LX$ with $L \equiv \{l_j(z_i)\}_{i=1,...,n;j=1,...,n}$. The entries of L correspond to weights using z_i as the explanatory variable and a bandwidth h. Let the kernel function be denoted by K. The weights for the local constant and local linear estimators are respectively defined as follows:

$$l_j^{lc}(z) \equiv \frac{K\left(\frac{z_j - z}{h}\right)}{\sum_{s=1}^{n} K\left(\frac{z_s - z}{h}\right)}, \tag{5}$$

$$l_j^{ll}(z) \equiv \frac{K\left(\frac{z_j - z}{h}\right)\left(\sum_{i=1}^{n} K\left(\frac{z_i - z}{h}\right)(z_i - z)^2 - (z_j - z)\sum_{i=1}^{n} K\left(\frac{z_i - z}{h}\right)(z_i - z)\right)}{\sum_{s=1}^{n} K\left(\frac{z_s - z}{h}\right)\left(\sum_{i=1}^{n} K\left(\frac{z_i - z}{h}\right)(z_i - z)^2 - (z_s - z)\sum_{i=1}^{n} K\left(\frac{z_i - z}{h}\right)(z_i - z)\right)}. \tag{6}$$

Hence:

$$\hat{g}(z_i) = \sum_{j=1}^{n} l_j(z_i)\, x_{1,j}, \tag{7}$$

where l_j can either be l_j^{lc} or l_j^{ll}. While the local constant estimator can be viewed as a weighted local average, the local linear estimator at the data point z_i can be viewed as the intercept of a weighted OLS regression where the instrument is centered around z_i. The form of $\hat{\beta}$ in (4) is not new : if $\hat{X} = P_Z X$, $\hat{\beta}$ is the 2SLS estimator and if $\hat{X} = X$, it is the OLS estimator.[3] In this paper, $\hat{X} = LX$ as defined above, leading to the k-IV estimator in (4). I now move to its asymptotic properties.

[3] $\hat{X}$ can also be obtained through other regression methods, such as the LASSO estimator as in Belloni *et al.* (2012).

3 Asymptotic properties

3.1 Consistency

Consistency of the k-IV estimator is proven using a uniform convergence result on the nonparametric estimator $\hat{g}(z_i)$. Let S be a compact subset of $\mathbb{R}$ that includes $z_i, i = 1, ..., n$ and excludes the boundaries of $\mathcal{Z}$. The following assumption ensures that the kernel estimator used in the first stage converges uniformly to the true conditional expectation function g_0.

Assumption 3. *Assume,* $\forall i$:

(i) $f_Z(z_i)$ *is differentiable almost surely over* S, $g_0(z_i)$ *is twice differentiable almost surely over* S, $|m(z_i) - m(z)| \leq C|z_i - z|$ $\forall i$ *for some* $C > 0$, *where* $m = g_0''$ *and* f_Z'.

(ii) $\mathbb{E}\left[v_i^2|z\right] \equiv \sigma_v^2(z)$ *is a continuous function, and* $\inf_{\{z \in S\}} f_Z(z) \geq \delta > 0$.

(iii) The kernel K *is symmetric, bounded and has compact support (i.e., for scalars* u, u', $\exists c > 0$ *such that* $K(u) = 0$ *for* $|u| \geq c$*). Assume* $\left||u|^l K(u) - |u'|^l K(u')\right| \leq C_2 |u - u'|$ $\forall 0 \leq l \leq 3$.

Assumption 3(i) is a Lipschitz condition assumption on the derivatives of the conditional expectation of $x_{1,i}$ given z_i and of the density of z_i, whereas (ii) ensures the variance of the error term v_i is continuous and the density of z_i is positive on a compact subset of $\mathbb{R}$. Condition (iii) is a condition on the type of kernel being used.

In order to obtain asymptotic normality of the estimator, I need to define

$$\begin{aligned}
\Omega_0 &\equiv \mathbb{V}\left[x_{0,i}u_i\right] \\
&= \mathbb{V}\left[\mathbb{E}\left(x_{0,i}u_i|z_i, x_{2,i}\right)\right] + \mathbb{E}\left[\mathbb{V}\left(x_{0,i}u_i|z_i, x_{2,i}\right)\right] \\
&= 0 + \mathbb{E}\left[\mathbb{V}\left(x_{0,i}u_i|z_i, x_{2,i}\right)\right] \\
&= \mathbb{E}\left[x_{0,i}\mathbb{V}\left(u_i|z_i, x_{2,i}\right)x_{0,i}'\right] \\
&= \mathbb{E}\left[x_{0,i}\sigma^2\left(z_i, x_{2,i}\right)x_{0,i}'\right]. \qquad (8)
\end{aligned}$$

If errors are homoskedastic, i.e. if $\mathbb{V}\left[u_i|z_i, x_{2,i}\right] = \sigma^2 < \infty$ then $\Omega_0 = \mathbb{E}\left[x_{0,i}\sigma^2 x_{0,i}'\right] = \sigma^2\mathbb{E}\left[x_{0,i}x_{0,i}'\right]$.

Assumption 4. $\mathbb{E}\left[x_{0,i}x_{0,i}'\right]$ *is nonsingular and* Ω_0 *is a finite positive definite matrix.*

The first part of Assumption 4 is a regularity assumption for the identification of β_0 through the moment condition $\mathbb{E}\left[u_i x_{0,i}\right] = 0$. It is needed so that $\hat{X}'X$ in (4) converges in probability to a well defined matrix after dividing by the sample size. In words, it says that there is no collinearity between the variables in $x_{0,i}$. In particular, $g_0(z_i)$ cannot be expressed as a linear combination of the variables in $x_{2,i}$. The second part is needed to invoke a central limit theorem when showing asymptotic normality. The nonsingularity of $\mathbb{E}\left[x_{0,i}x_{0,i}'\right]$ and finiteness of $\sigma^2(z_i, x_{2,i})$ imply $\mathbb{E}\left[x_{0,i}x_{0,i}'\right] < \infty$, i.e. the second moments of all the variables in x_i as well as $g_0(z_i)$ are finite. It thus implies that their first moment in absolute value is also finite, a property I will be using in conjunction with uniform convergence of the kernel estimator. The condition that Ω_0 is positive definite implies $\sigma^2(z_i, x_{2,i})$ is also finite, and $\mathbb{E}\left[\left|u_i^2\right|\right]$ is finite as $\mathbb{E}\left[\left|u_i^2\right|\right] = \mathbb{E}\left[u_i^2\right] = \mathbb{E}\left[\mathbb{E}\left[u_i^2|z_i, x_{2,i}\right]\right] = \mathbb{E}\left[\sigma^2(z_i, x_{2,i})\right]$. The proof involves the uniform convergence of the kernel estimators via the probability that the kernel estimator is at most at a distance of $\varepsilon > 0$ from $g_0(z_i)$ which tends to one. The term $|\hat{g}(z_i) - g_0(z_i)|$ is then multiplied by $|u_i|$ and the absolute value of each component of x_i and summed over, so the right hand side of the inequalities has terms such as $\frac{1}{n}\sum_{i=1}^n |u_i|\varepsilon$. Those terms need to be controlled as $n \to \infty$, which finiteness of the moment in absolute value helps achieve. The first theorem of the paper is a consistency result.

Theorem 1. *(consistency) If Assumptions 1, 2, 3, and 4 are satisfied, and if the bandwidth* $h \overset{n\to\infty}{\to} 0$, $\ln(n)/nh \overset{n\to\infty}{\to} 0$ *then, for* $\hat{\beta}$ *defined in (4):*

$$\hat{\beta} - \beta_0 = o_{\mathbb{P}}(1).$$

This result uses different assumptions than Pakes & Olley (1995) who make assumptions about the empirical moment function used to estimate the parameter of interest, its derivatives and the rates of convergence of the nonparametric components. In particular, the rate of decay of $n^{-1/4}$ for the bandwidth is assumed for the kernel estimators to converge uniformly to the functional they estimate. This rate is calculated in Bierens (1987) among other references. In Theorem 1, I choose to take the approach developed in Masry (1996) and briefly exposed in Li & Racine (2007). Hansen (2008) derives similar results under different conditions. In particular, his results allow for kernels with non compact supports to be used, and uniform convergence is proven over expanding sets and unbounded sets, as opposed to Masry (1996) who considers compact sets.

Kernel estimators can also accommodate more than one regressor. The k-IV estimator remains just-identified, and the estimate $\hat{g}(z_i)$ is potentially subject to the curse of dimensionality. The extension to a vector of instruments is straightforward however, in the sense that the proofs are using the same arguments as in the single instrument case. Subsection 7.5 presents that setup and the corresponding results.

The next lemma shows that the 2SLS estimator is a special case of the k-IV estimator when making the bandwidth go to infinity for a fixed n.

Lemma 1. *Under Assumptions 1 and 2, as $h \to \infty$ for a given n, if the first stage regressions for $\hat{\beta}$ and $\hat{\beta}_{2SLS}$ both include the same regressors, the k-IV estimator defined in (4) using the local linear estimator converges to the 2SLS estimator, i.e.:*

$$\hat{\beta} \overset{h\to\infty}{\to} \hat{\beta}_{2SLS}.$$

The intuition behind this lemma is straightforward: when the bandwidth goes to infinity, a local linear regression becomes global (as the same weight $K(0) < \infty$ is assigned to each observation), and the resulting estimator of $g_0(z_i)$ is the fitted value of an OLS regression evaluated at z_i. It implies that $\hat{X} = P_Z X$ where $P_Z \equiv Z(Z'Z)^{-1}Z'$. The final estimator is then $(X'P_Z X)^{-1} X'P_Z Y = \hat{\beta}_{2SLS}$.[4] If one considers multiple instruments, then the result of Lemma 1 holds for $\min_{1\leq s\leq q}\{h_s\} \to \infty$. It is common use to include the exogenous variables $x_{2,i}$ in the first stage regression for the 2SLS estimator, whereas the first stage regression for $\hat{\beta}$ consists in estimating $g_0(z_i)$, not $g_0(z_i, x'_{2,i})$. So in the presence of exogenous variables, the equivalence between the estimators is established if $g_0(z_i, x'_{2,i})$ is estimated with large bandwidths. Besides, including $x_{2,i}$ in the estimation of g_0 would expose one to the curse of dimensionality, where a high amount of observations is needed to accurately estimate g_0. In section 4, the equivalence between the two estimators can be observed for values of the bandwidth of 100 in three out of four designs. The 2SLS estimator being a special case, it suggests that there is room for improvement through the choice of the bandwidth. In the next subsection, I show asymptotic normality of the k-IV estimator.

3.2 Asymptotic normality

In this section, I establish asymptotic normality of the k-IV estimator both for the heteroskedastic and homoskedastic case.

Theorem 2. *Under Assumptions 1, 2, 3, 4, if $h \overset{n\to\infty}{\to} 0$ and $\ln(n)/nh \overset{n\to\infty}{\to} 0$ then, with $\hat{\beta}$ defined in (4):*

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \overset{d}{\to} \mathcal{N}(0, \Sigma), \tag{9}$$

where $\Sigma \equiv \left(\mathbb{E}\left[x_{0,i}x'_{0,i}\right]\right)^{-1} \Omega_0 \left(\mathbb{E}\left[x_{0,i}x'_{0,i}\right]\right)^{-1}$. If, in addition, $\mathbb{V}[u_i|z_i, x_{2,i}] = \sigma^2 < \infty \ \forall i = 1, ..., n$, then:

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \overset{d}{\to} \mathcal{N}\left(0, \sigma^2\left(\mathbb{E}\left[x_{0,i}x'_{0,i}\right]\right)^{-1}\right). \tag{10}$$

[4] It also suggests that $\hat{\beta}$ defined in (4) is consistent if $h \to \infty$ by a sequential asymptotics argument (exposed briefly in Wang & Yu, 2016):

$$\lim_{n\to\infty}\left\{\lim_{h\to\infty}\hat{\beta}\right\} = \lim_{n\to\infty}\hat{\beta}_{2SLS} = \beta_0,$$

a different consistency result in the case where $h \to \infty$ faster than $n \to \infty$.

Similarly to Theorem 1, $h \to 0$ and $nh \to \infty$ are not enough for convergence of the kernel estimator to be uniform. On top of that, one needs the bandwidth to not decrease too fast, in which case the variance of the kernel estimator does not settle fast enough for the estimator to be contained in a uniform band for each $z_i \in \mathcal{Z}$. This asymptotic distribution resonates with Newey (1990)'s for which the order series are required to go to infinity. If multiple instruments are considered, then one needs to assume $\max_{1 \leq s \leq q} \{h_s\} \to 0$, $h_1 h_2 ... h_q \overset{n \to \infty}{\to} 0$, and $\ln(n) / (n h_1 h_2 ... h_q) \to 0$ as $n \to \infty$ as mentioned in Theorem 4.

The asymptotic variance in both the homoskedastic and heteroskedastic case is independent of the estimator used for g_0, as mentioned in Newey (1994). Note that in the homoskedastic case, the asymptotic variance coincides with the lower bound derived by Chamberlain (1987) for estimators based on $\mathbb{E}[u_i | z_i, x_{2,i}] = 0$. The bound is defined as

$$\Lambda_0^* \equiv \left(\mathbb{E}\left[D_0'(w_i) \Sigma_0^{-1}(w_i) D_0(w_i) \right] \right)^{-1},$$

where $w_i \equiv (z_i, x_{2,i})$, $D_0'(w_i) \equiv \mathbb{E}\left[\frac{\partial u_i}{\partial \beta'} | w_i \right]$ and $\Sigma_0(w_i) \equiv \mathbb{E}\left[u_i^2 | w_i \right]$. Given the model described in Assumption 1, $D_0'(w_i) = -\mathbb{E}[x_i | w_i] = -x_{0,i}$ and $\Sigma_0(w_i) = \sigma^2$ if errors are homoskedastic. Hence, in this case, estimating $g_0(z_i)$ via kernel estimators amounts to estimating optimal instruments as in Linton (2002). In the heteroskedastic case, the asymptotic variance formula from (9) corresponds to the bound derived by Chamberlain (1987) for estimators based on the unconditional moments $\mathbb{E}[u_i] = \mathbb{E}[g_0(z_i) u_i] = \mathbb{E}[x_{2,i} u_i] = 0$ implied by the conditional ones $\mathbb{E}[u_i | w_i] = 0$. This estimator is thus a significant improvement over the 2SLS estimator in terms of asymptotic variance. The next theorem proposes a consistent estimator of the asymptotic variance in (9) and (10).

Theorem 3. *Under the same conditions as Theorem 2, and the assumptions that* $\mathbb{E}\left[|g_0(z_i)| u_i^2 \right] < \infty$ *and* $\mathbb{E}\left[|x_{2,i}| u_i^2 \right] < \infty$, *the following holds:*

$$\left(\frac{1}{n} \sum_{i=1}^n \hat{x}_i \hat{x}_i' \right)^{-1} \frac{1}{n} \sum_{i=1}^n \hat{x}_i \hat{u}_i^2 \hat{x}_i' \left(\frac{1}{n} \sum_{i=1}^n \hat{x}_i \hat{x}_i' \right)^{-1} \overset{\mathbb{P}}{\to} \left(\mathbb{E}\left[x_{0,i} x_{0,i}' \right] \right)^{-1} \Omega_0 \left(\mathbb{E}\left[x_{0,i} x_{0,i}' \right] \right)^{-1}, \tag{11}$$

$$\hat{\sigma}^2 \left(\frac{1}{n} \sum_{i=1}^n \hat{x}_i x_i' \right)^{-1} \overset{\mathbb{P}}{\to} \sigma^2 \left(\mathbb{E}\left[x_{0,i} x_{0,i}' \right] \right)^{-1}, \tag{12}$$

where $\hat{x}_i \equiv \left(1 \;\; \hat{g}(z_i) \;\; x_{2,i}' \right)$, $\hat{u}_i \equiv y_i - x_i' \hat{\beta}$ *and* $\hat{\sigma}^2 \equiv \frac{1}{n - (d+2)} \sum_{i=1}^n \hat{u}_i^2$.

I now turn to a Monte Carlo experiment to analyze the behavior of the k-IV estimator compared to some classical ones.

4 Simulation study

4.1 Estimation

I study the performance of the k-IV estimator compared to some traditionally used in the literature. Throughout the study, the data generating processes differ in the relationship between the endogenous variable and the instrument. There are three levels of endogeneity according to the covariance between u_i and v_i that I denote σ_{uv} : low ($\sigma_{uv} = 0.1$), moderate ($\sigma_{uv} = 0.5$), high ($\sigma_{uv} = 0.9$), and two sample sizes: $n = \{100, 1000\}$. Three k-IV estimators are considered: $\hat{\beta}_{lc}$ and $\hat{\beta}_{ll}$ are the k-IV estimators defined in (4) with either the local constant or local linear estimator, computed for different values of the bandwidth h whereas the estimator denoted $\hat{\beta}_{NP}$ is the k-IV estimator obtained from using a local linear estimator with a bandwidth selected through least squares cross validation in the first stage.[5] This estimator is easily computed as bandwidth

[5] Define the cross validation criterion

$$CV(h) \equiv \frac{1}{n} \sum_{i=1}^n (x_i - \hat{g}_{-i}(z_i))^2$$

where $\hat{g}_{-i}(z_i)$ is the kernel estimator of $g(z_i)$ using all of the observations but observation i. This criterion is minimized to select the optimal bandwidth. It is implemented in the np package in **R** that is used to compute $\hat{\beta}_{NP}$.

selection via cross validation has been widely studied and implemented in the majority of statistical software. Hence, it is relevant to compare its performances with those of $\hat{\beta}_{ll}$ to see how optimal this selection criterion is. The other estimators I consider are: the OLS estimator, the 2SLS estimator, the Weighted Minimum Distance-Fuller augmented (WMDF) estimator from Antoine & Lavergne (2014) labelled $\hat{\beta}_{WMDF}$, and the estimators of Donald & Newey (2001) and Lee & Shin (2018), respectively labelled $\hat{\beta}_{DN}$ and $\hat{\beta}_{CSA}$.[6] Throughout the experiment, the random variables are generated as follows:

$$\begin{aligned}
y_i &= \beta_0 x_i + \beta_1 w_{i,1} + \beta_2 w_{i,2} + u_i, \\
z_i &\sim \mathcal{N}(0,1), \\
w_{i,j} &\sim \mathcal{N}(0,1), \; j = 1,2, \\
\begin{pmatrix} u_i \\ v_i \end{pmatrix} &\sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{uv} \\ \sigma_{uv} & 1 \end{pmatrix} \right), \\
\sigma_{uv} &\in \{0.1,\, 0.5,\, 0.9\}, \\
h &\in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 10, 100\}.
\end{aligned}$$

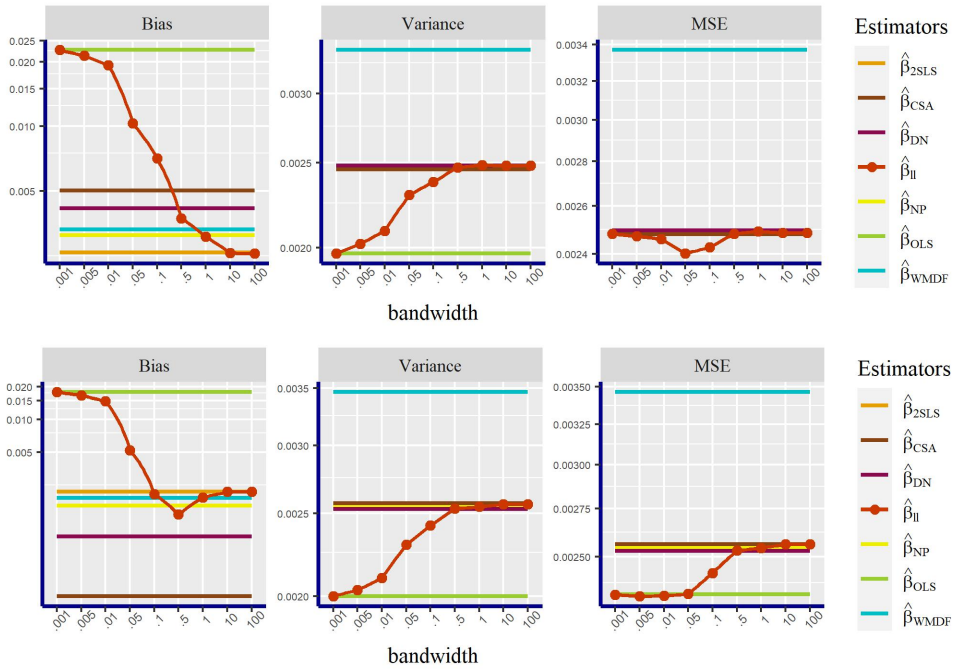
The parameter of interest is β_0, set to one in all of the simulations. I consider four designs for the first stage relationship:

Design 1	$x_i = 2z_i + v_i$
Design 2	$x_i = \ln(\lvert z_i \rvert) + v_i$
Design 3	$x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$
Design 4	$x_i = \phi(z_i) + v_i$
Design 5	$x_i = \mathbb{1}\{z_i + v_i > 0\},\; z_i \sim \mathcal{N}(-0.5, 1)$

where $\phi()$ is the normal density function. The first design is a benchmark where one can expect the 2SLS estimator to perform well while the other designs involve nonlinearities in the z_i. The last design is encountered in the program evaluation literature where the assignment variable (x_i here) is a binary variable but compliance is not perfect. It depends on an underlying variable (z_i here) so the probability of compliance is first estimated, and used to estimate the average treatment effect. Figures 1 through 10 show the performance of the different estimators in terms of bias, variance and MSE over 1,000 simulation rounds. The graphs for moderate levels of endogeneity are displayed in the appendix as they show the same behavior as either low or high endogeneity levels. The optimal bandwidth chosen through cross validation for $\hat{\beta}_{NP}$ leads to a higher bias than $\hat{\beta}_{2SLS}$ but its variance is only slightly smaller than it. The optimal bandwidth is nevertheless close to being optimal in a MSE sense for all degrees of endogeneity and sample sizes considered. In all designs, $\hat{\beta}_{ll}$ coincides with $\hat{\beta}_{2SLS}$ for bandwidths of 100 and higher, illustrating Lemma 1. On the other hand, for small bandwidths, both $\hat{\beta}_{ll}$ and $\hat{\beta}_{lc}$ are close to $\hat{\beta}_{OLS}$ as when estimating $g_0(z_i)$, a small bandwidth will lead the kernel estimators to heavily discount the observations z_j, $j \neq i$, making the weight on x_i close to one and the final estimate $\hat{g}(z_i)$ very close to x_i.

[6] All the estimators are defined in Appendix 7.6.1.

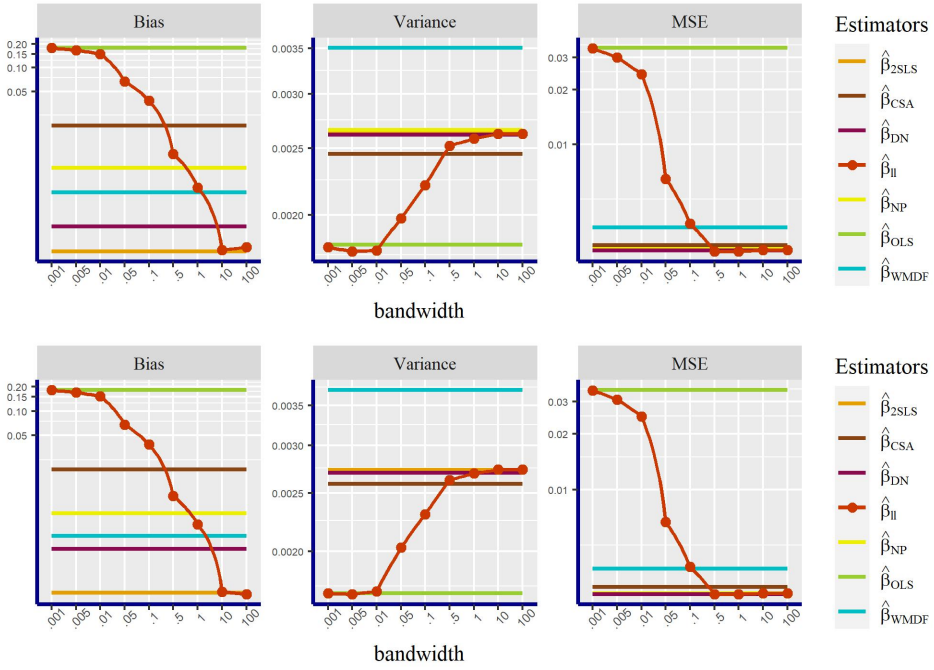
Figure 1: Design 1, low endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. The estimator using the local constant kernel estimator ($\hat{\beta}_{lc}$) is omitted for scaling purposes. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = 2z_i + v_i$.

The same pattern is observed for a high endogeneity level: the OLS estimator has the highest bias but the lowest variance, as opposed to the 2SLS estimator whose bias is the smallest, but variance the greatest (after the WMDF estimator). $\hat{\beta}_{NP}$ and $\hat{\beta}_{2SLS}$ are very close in terms of MSE, although one dominates in terms of variance and the other in terms of bias.

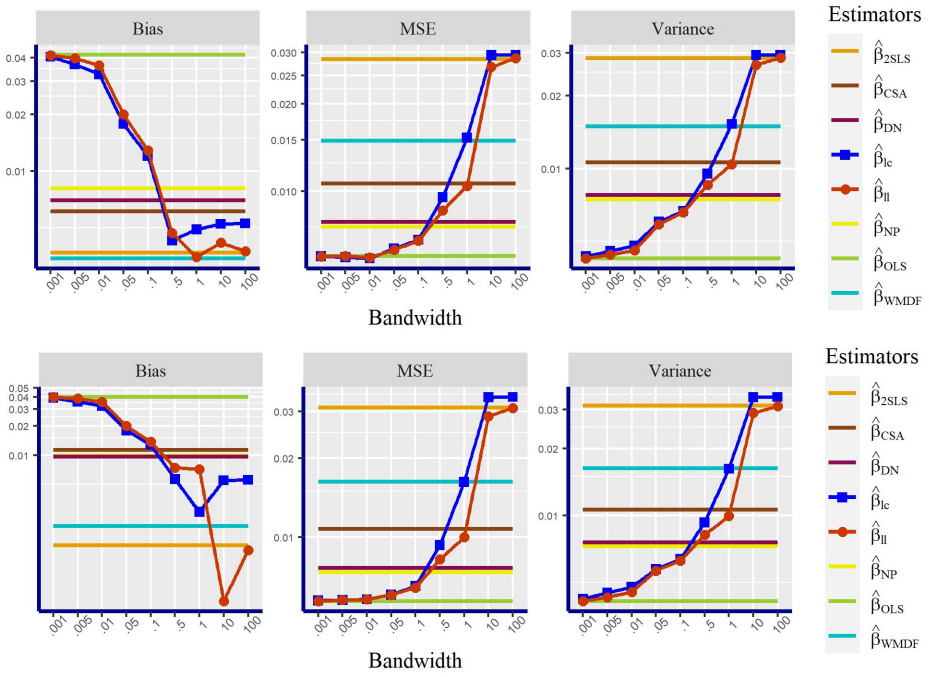
Figure 2: Design 1, high endogeneity, $n = 100$ (top), $n = 1,000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. The estimator using the local constant kernel estimator ($\hat{\beta}_{lc}$) is omitted for scaling purposes. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = 2z_i + v_i$.

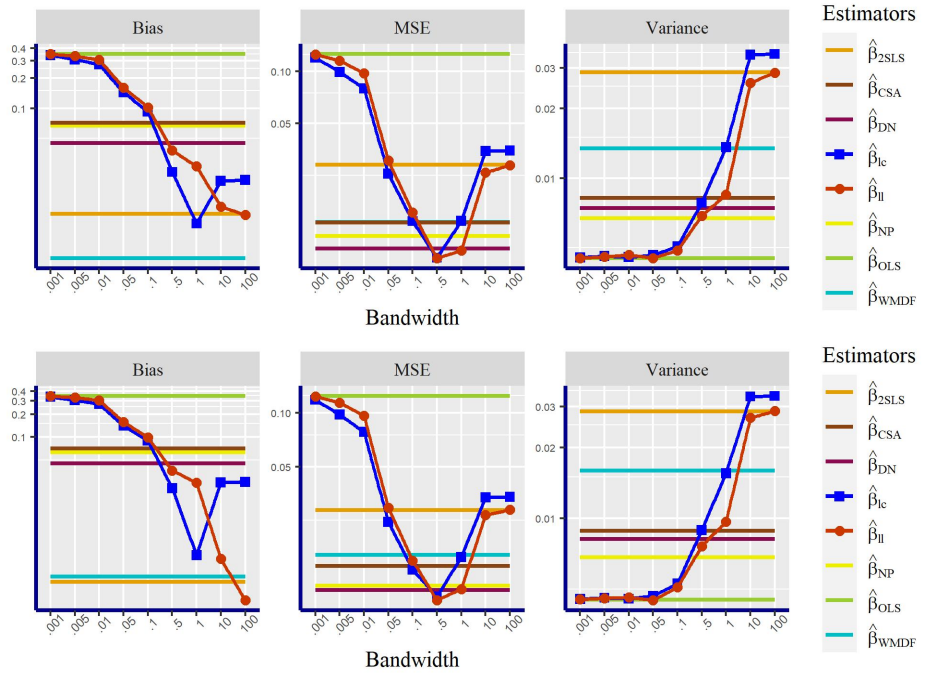
In design 2, $\hat{\beta}_{ll}$ features better properties than $\hat{\beta}_{lc}$ overall, with a few exceptions. Their behavior in terms of bias differs with a minimum reached for different bandwidths in the case of $n = 1,000$, but not in terms of variance: the higher the bandwidth, the higher the variance. The bias of $\hat{\beta}_{lc}$ and $\hat{\beta}_{ll}$ decreases with the bandwidth, and their variance increases with it, so a bias-variance trade off occurs for the MSE. Both perform roughly equivalently bias wise, but $\hat{\beta}_{ll}$ has an advantage in terms of variance. Note also that $\hat{\beta}_{NP}$ performs better than $\hat{\beta}_{2SLS}$ MSE wise in all of the designs, where the lower variance outweighs the higher bias.

Figure 3: Design 2, low endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \ln(|z_i|) + v_i$.

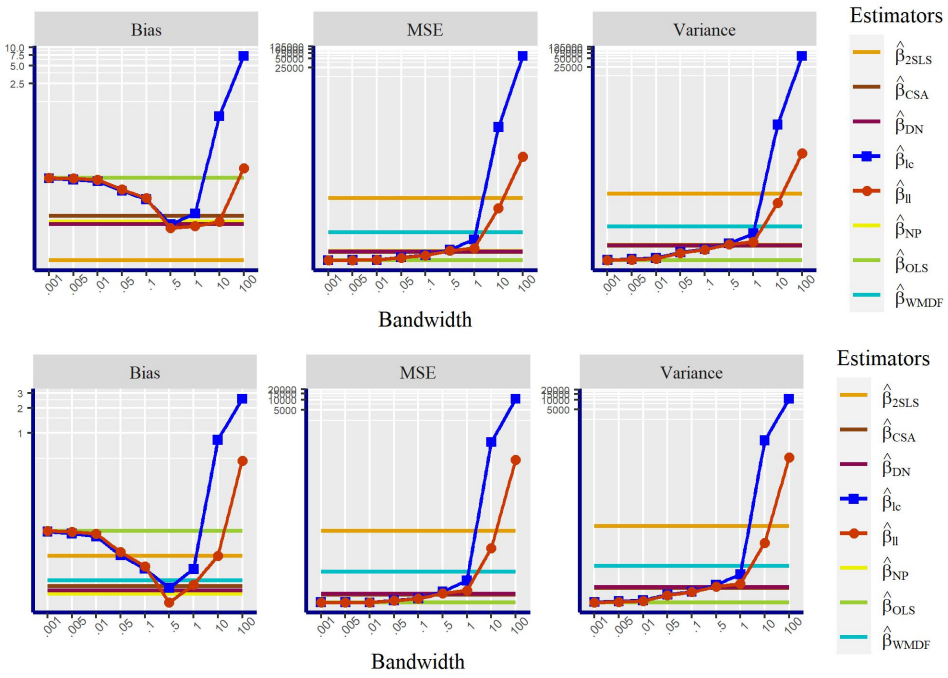
Figure 4: Design 2, high endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \ln(|z_i|) + v_i$.

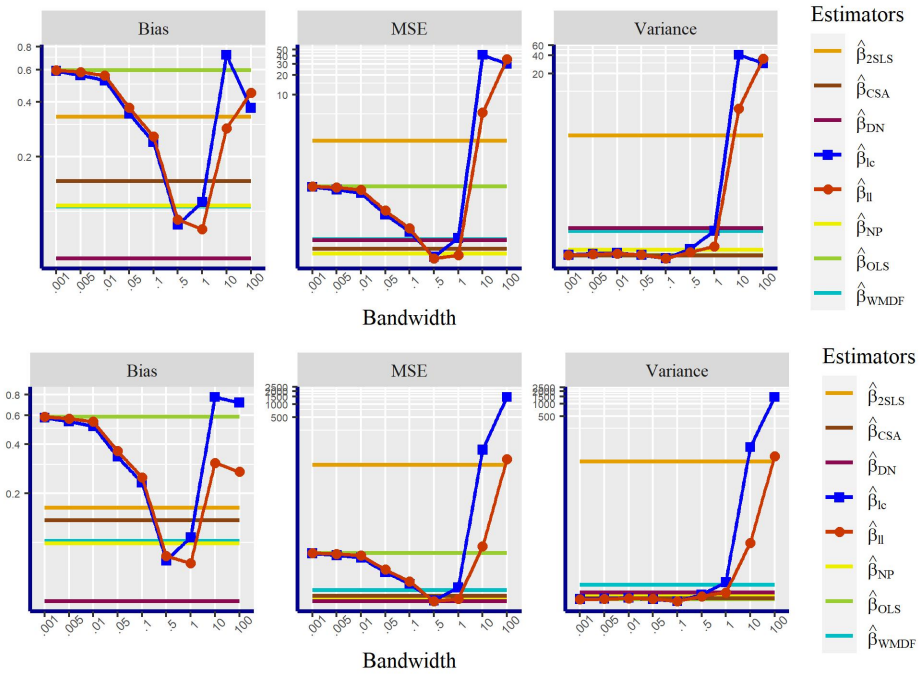
In the third design, $\hat{\beta}_{ll}$ and $\hat{\beta}_{lc}$ behave the same way, but $\hat{\beta}_{lc}$'s bias and variance are very high for larger bandwidths. Note that here, $\hat{\beta}_{ll}$ with $h = 100$ does not coincide with $\hat{\beta}_{2SLS}$, although it is close. For such a design, a higher bandwidth is needed to observe the equivalence between the two estimators. The 2SLS estimator performs particularly poorly because of the relationship between x_i and z_i which features a decreasing linear relationship as n increases. There is a global minimum within the bias that $\hat{\beta}_{NP}$ is able to approximate well in all the cases. Combined with a lower variance, $\hat{\beta}_{NP}$ outperforms $\hat{\beta}_{2SLS}$ in all dimensions, and for all levels of endogeneity.

Figure 5: Design 3, low endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.

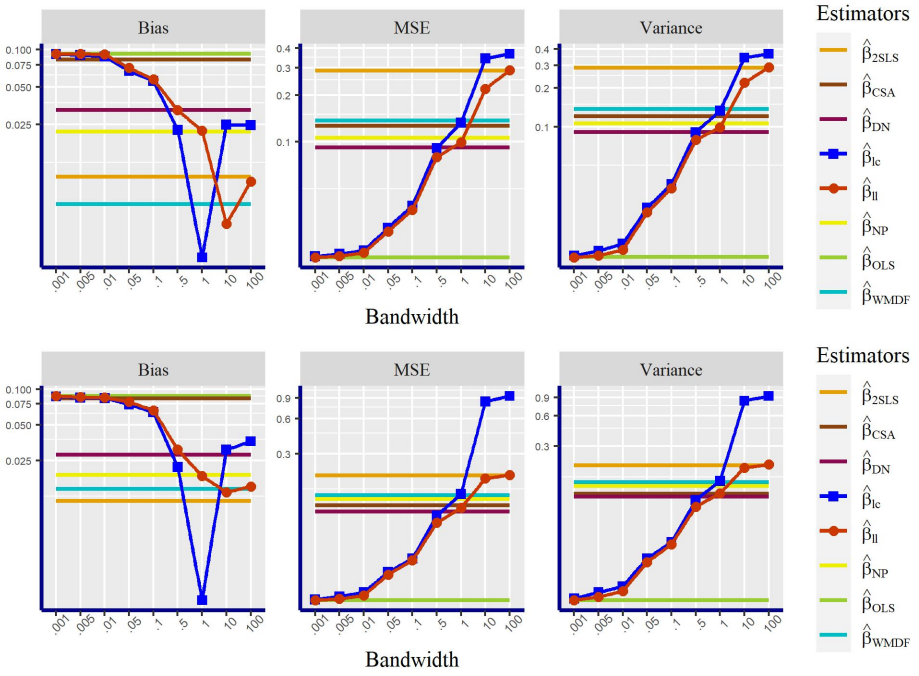
Figure 6: Design 3, high endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.

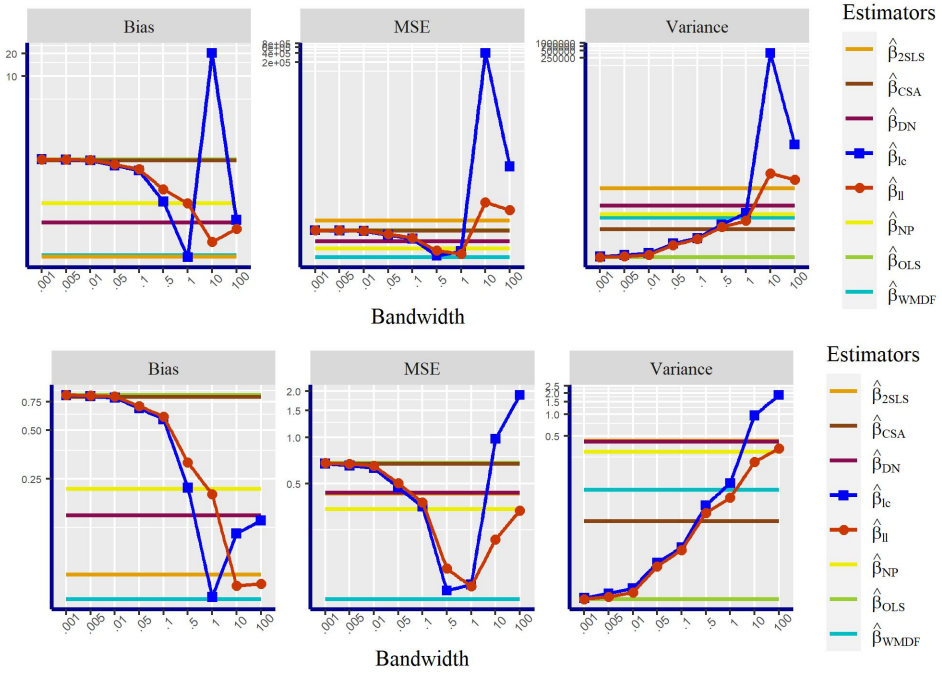
Design 4 features a lower bias for $\hat{\beta}_{NP}$ than $\hat{\beta}_{2SLS}$ when endogeneity is low, but not for moderate and high endogeneity levels. Thanks to a lower variance, $\hat{\beta}_{NP}$ either outperforms $\hat{\beta}_{2SLS}$ and $\hat{\beta}_{WMDF}$ or performs as well in terms of MSE.

Figure 7: Design 4, low endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \phi(z_i) + v_i$.

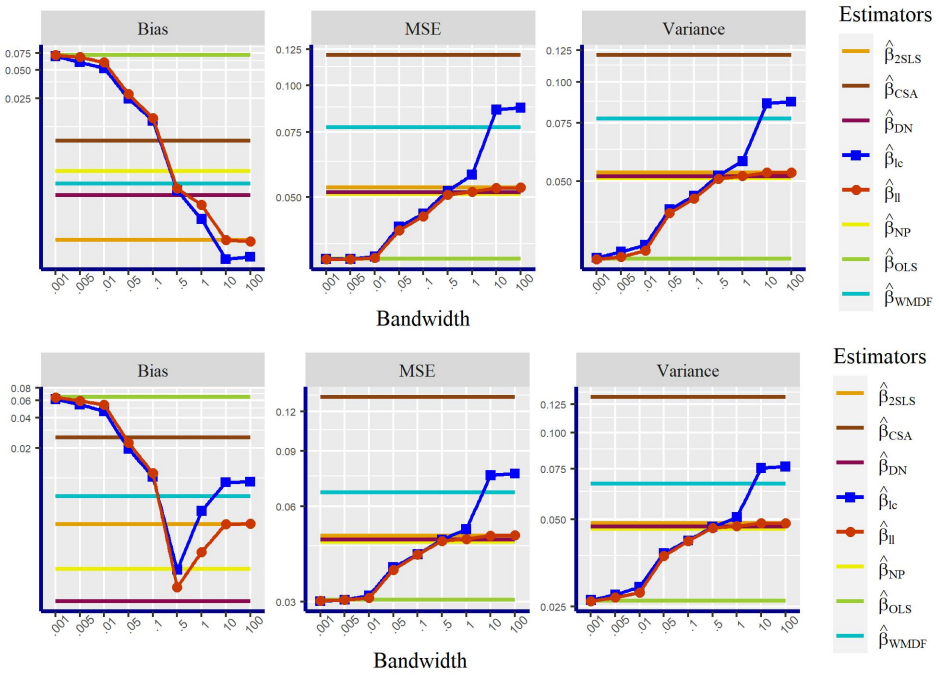
Figure 8: Design 4, high endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \phi(z_i) + v_i$.

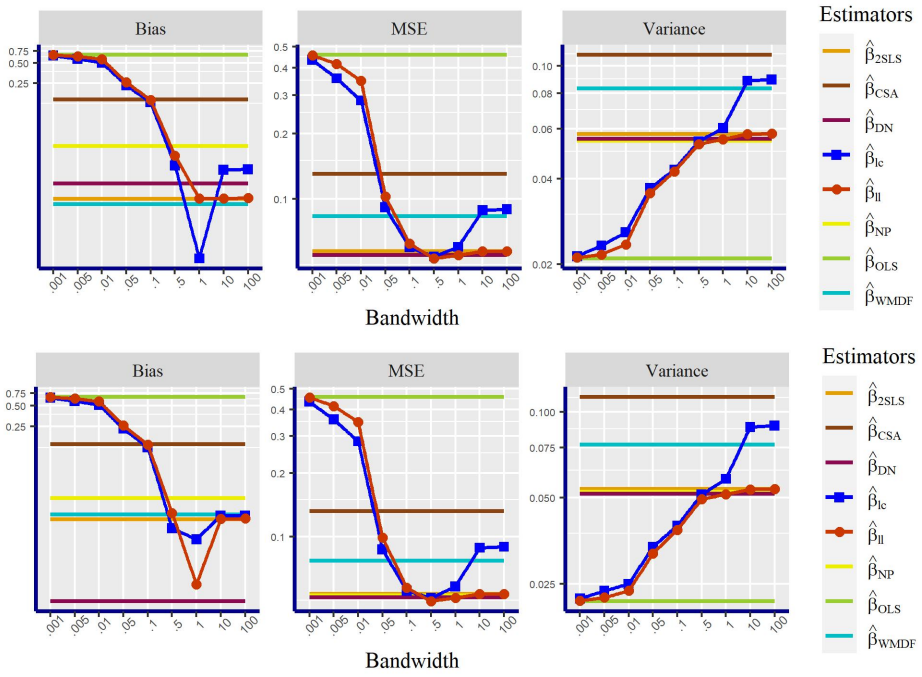
In design 5, $\hat{\beta}_{NP}$ and $\hat{\beta}_{2SLS}$ perform equivalently in terms of MSE. For high samples, $\hat{\beta}_{NP}$ approximates well the optimal bias-variance tradeoff as it aligns with the lowest MSEs of $\hat{\beta}_{ll}$.

Figure 9: Design 5, low endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \mathbb{1}\{z_i + v_i > 0\}$, $z_i \sim \mathcal{N}(-0.5, 1)$.

Figure 10: Design 5, high endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \mathbb{1}\{z_i + v_i > 0\}$, $z_i \sim \mathcal{N}(-0.5, 1)$.

The cross validation criterion used in the choice of the optimal bandwidth for $\hat{\beta}_{NP}$ is not suitable to select the bandwidth leading to the smallest bias. However it allows $\hat{\beta}_{NP}$ to outperform its competitors in almost all the designs shown above. Nevertheless, this criterion is not tailored towards minimizing some error function related to the estimator $\hat{\beta}_{NP}$. Rather, it focuses on the mean squared prediction error of the first stage, for which one working with such models often has little interest. Hence, there is room for further improvement by designing criteria targeting the mean-squared error of the k-IV estimator. Moreover, using the local linear estimator over the local constant one is preferable due to the sometimes unstable behavior of $\hat{\beta}_{lc}$, in particular for small samples and linear first stage relationships.

4.2 Bandwidth selection

The estimator is consistent as long as the bandwidth goes to zero as the sample size goes to infinity, but it does not propose a way to select the bandwidth in finite sample. The selection method used in the previous section is natural and straightforward, but it optimizes a first stage related criterion, while the end goal is the estimation of the second stage parameters. In this section, I consider three different methods on top of the one previously used. Two of these methods are based on an information criterion (and assume homoskedasticity), whereas the last one is a leave-one-out cross validation criterion using $\hat{\beta}$ rather than $\hat{y}$. Moreover, given the simulations results displayed above, I only consider the local linear kernel estimator in the first stage and not the local constant one. Kernel estimators are linear, as $\hat{X}$ can be written as $\hat{X} = LX$ where L is the matrix of weights built using kernels following expressions (5) and (6).

Following Hurvich *et al.* (1998), I look at an alternative to leave-one-out cross validation in the first stage, by minimizing the following criterion function:

$$AIC_1(h) \equiv \ln \hat{\sigma}_v^2 + \frac{1 + tr(L)/n}{1 - (tr(L+2))/n}$$

where $\hat{\sigma}_v^2$ is an estimate of the variance of the first stage error term v_i obtained as follows: $\hat{\sigma}_v^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{g}(x_i)\right)^2 = X'(I_n - L)'(I_n - L)X/n$. I call the k-iv estimator using the bandwidth that minimizes that function $\hat{\beta}_{AIC1}$.

The second information criterion is related to the second stage equation, rather than the first:

$$AIC_2(h) \equiv \ln \hat{\sigma}_u^2 + \frac{1 + tr(L_X)/n}{1 - (tr(L_X + 2))/n}$$

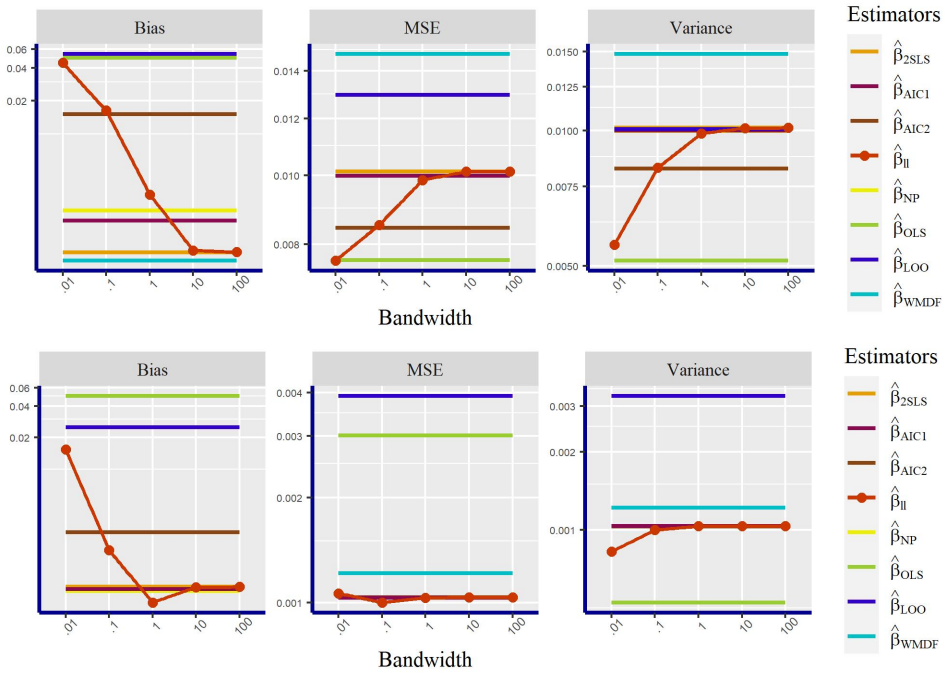
where $\hat{\sigma}_u^2$ is an estimate of the variance of the second stage error term u_i obtained as follows: $\hat{\sigma}_u^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(y_i - x_i\hat{\beta}\right)^2 = Y'(I_n - L_X)'(I_n - L_X)Y/n$ and $L_X \equiv X\left(\hat{X}'X\right)^{-1}\hat{X}'$ I call the k-iv estimator using the bandwidth that minimizes that function $\hat{\beta}_{AIC2}$.

The leave-one-out cross validation second stage based criterion uses $\hat{\beta}_{-i}$, the k-iv estimator using all but observation i:

$$M\hat{S}E_{LOO}(h) \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\beta}_{-i} - \hat{\beta}\right)^2.$$

I call the k-iv estimator using the bandwidth that minimizes that function $\hat{\beta}_{LOO}$. [7] To sum up, two estimators use a bandwidth that optimizes a first stage criterion ($\hat{\beta}_{AIC1}$ and $\hat{\beta}_{NP}$) and two estimators use a bandwidth that optimizes a second stage criterion ($\hat{\beta}_{AIC2}$ and $\hat{\beta}_{LOO}$). I compare these estimators to the OLS, 2SLS and WMDF ones (the estimator computed over different bandwidths values is also shown).

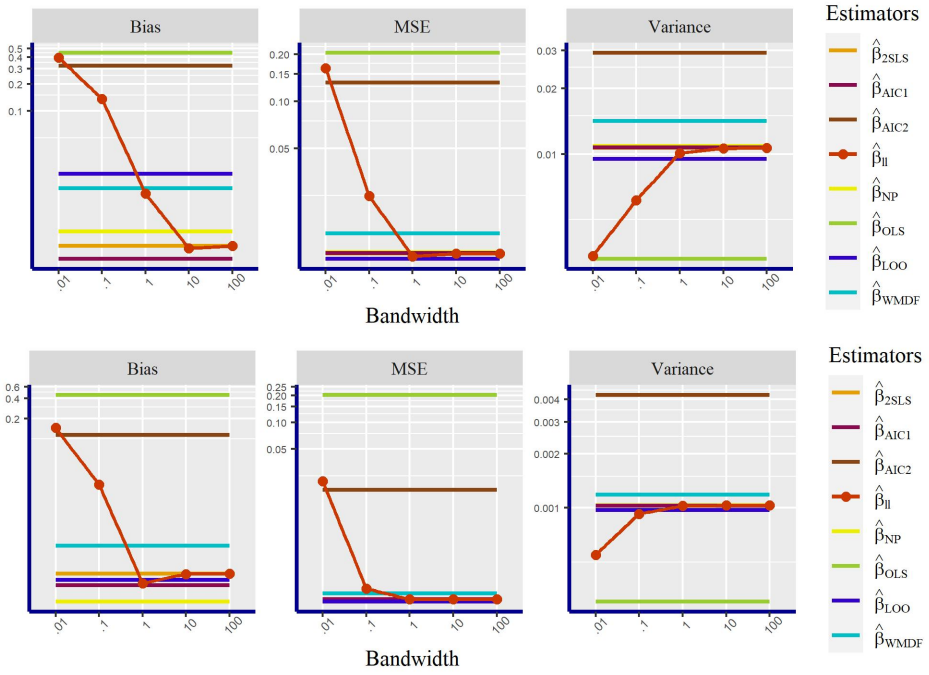
Figure 11: Design 1, low endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = z_i + v_i$.

[7] The following relationship highlighted by Kline *et al.* (2020): $y_i - x_i\hat{\beta}_{-i} = \frac{y_i - x_i\hat{\beta}}{1 - P_{ii}}$ where $P_{ii} \equiv x_i'\left[\sum_{i=1}^{n} x_i x_i'\right]^{-1} x_i$ provides a computationally friendly way to implement this criterion function.

Figure 12: Design 1, high endogeneity, $n = 100$ (top), $n = 1000$ (bottom)

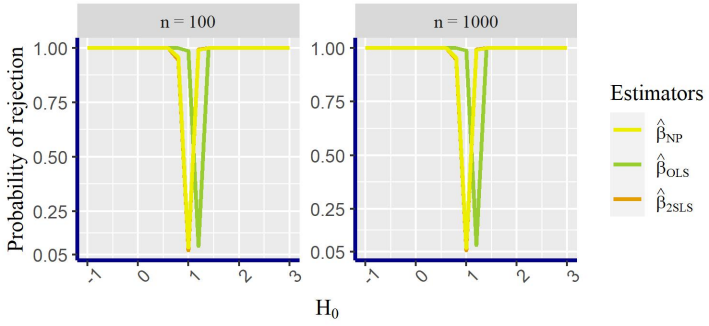


Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1,000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = z_i + v_i$.

4.3 Inference

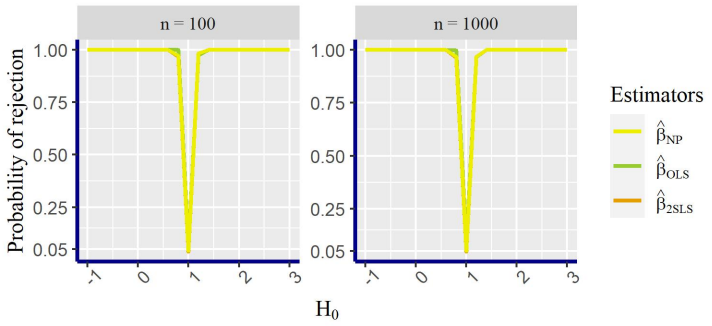
This section focuses on the power properties of the k-IV estimator compared to its competitors. The true parameter in the simulation being $\beta_0 = 1$, The power curves below apply on a grid ranging from -1 to 3. $\hat{\beta}_{lc}$ and $\hat{\beta}_{ll}$ are omitted as their curve differs according to the bandwidth but a bandwidth is not selected for these estimators. Inference using $\hat{\beta}_{NP}$ makes use of the estimate of the asymptotic variance shown in 3. The power curve for $\hat{\beta}_{NP}$ is centered around the true value in all the designs, whereas each other estimator displays either a low level or a shift away from β_0.

Figure 13: Design 1, high endogeneity



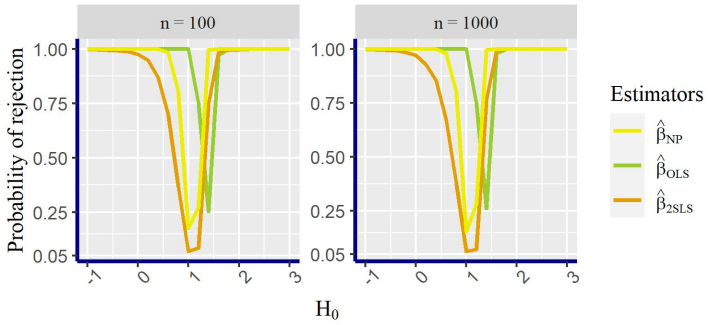
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = 2z_i + v_i$.

Figure 14: Design 1, low endogeneity



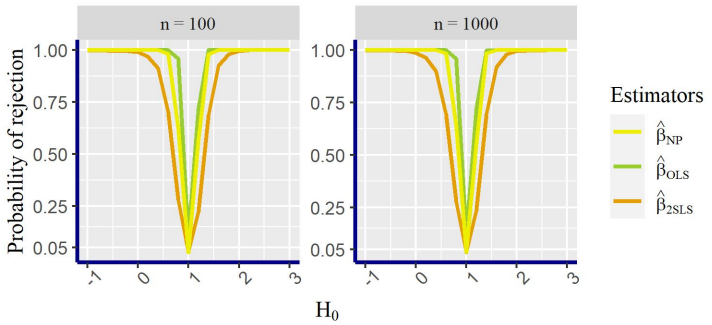
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = 2z_i + v_i$.

Figure 15: Design 2, high endogeneity



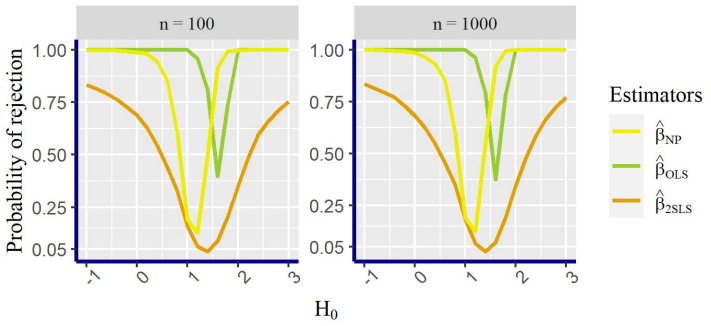
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \ln(|z_i|) + v_i$.

Figure 16: Design 2, low endogeneity



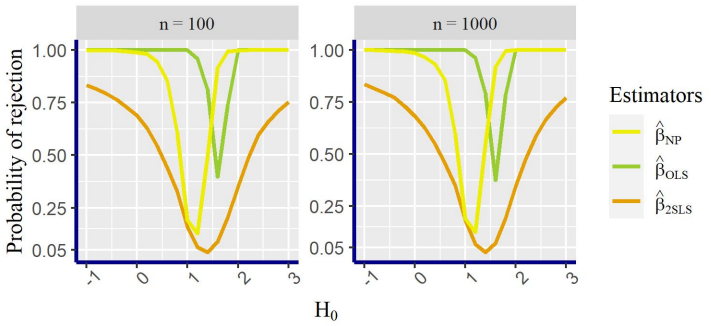
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \ln(|z_i|) + v_i$.

Figure 17: Design 3, high endogeneity



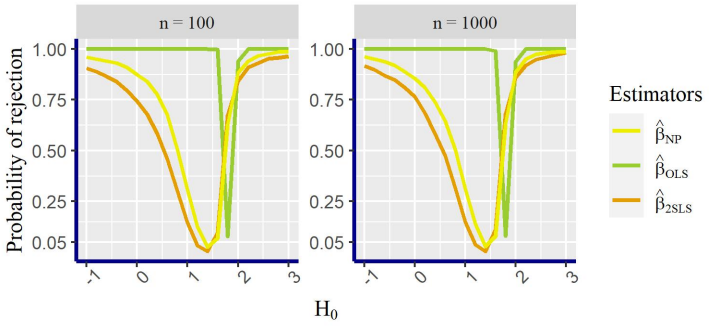
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right)„ $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.

Figure 18: Design 3, low endogeneity



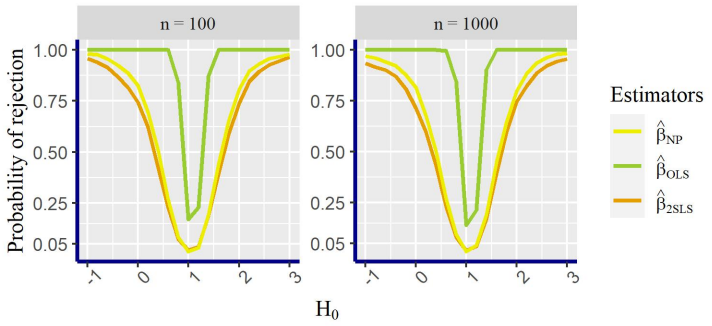
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.

Figure 19: Design 4, high endogeneity



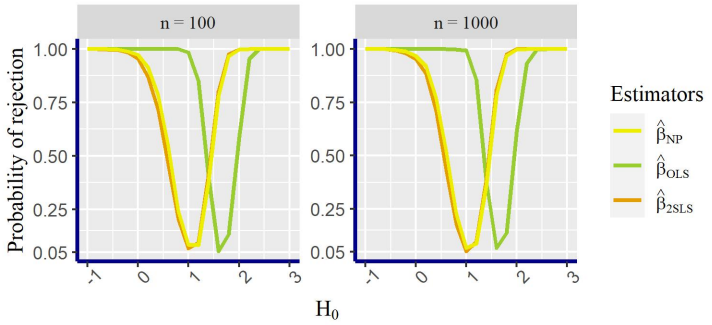
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \phi\left(z_i\right) + v_i$.

Figure 20: Design 4, low endogeneity



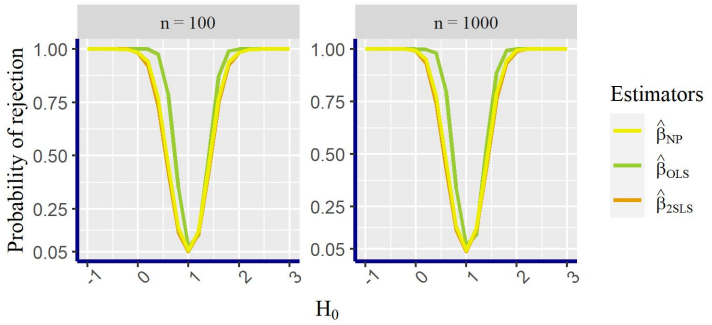
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \phi(z_i) + v_i$.

Figure 21: Design 5, high endogeneity



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \mathbb{1}\{z_i + v_i > 0\}$, $z_i \sim \mathcal{N}(-0.5, 1)$.

Figure 22: Design 5, low endogeneity



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (left), $n = 1,000$ (right), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \mathbb{1}\{z_i + v_i > 0\}$, $z_i \sim \mathcal{N}(-0.5, 1)$.

5 Empirical application: Chalfin (2015)

I apply the k-IV estimator on an empirical data set used in Chalfin (2015). The paper explores the effect of Mexican immigration on crime in the USA and attempts to measure its impact on various types of crimes.

While answers to that question vary in the literature, the author emphasizes the endogeneity of a commonly used instrument for Mexican immigration flows, the network instrument, due to the inclusion of the probability of migration conditional on being eligible in its definition. That probability can depend on unobserved characteristics related to labor market characteristics in the US for instance, and is thus endogenous. The proposed instrument omits that probability, becoming an instrument based not on the predicted change in the number of migrants living in a given city in a given year, but on the predicted change in the number of migrants if the whole pool of eligible migrants migrates. The author uses metropolitan statistical area (MSA)-level data from three US census, and regresses the difference in crime between two censuses for a particular MSA on the difference in the Mexican immigrant share, using the proposed instrument. I refer the reader to Chalfin (2015, p.222) for more information. The paper has seven specifications (corresponding to each type of crime), and considers the following model:

$$\Delta y_{i,t} = \beta_0 + \beta_1 \Delta x_{i,t} + \beta_2' w_{i,t} + \alpha_j \times \delta_t + u_{i,t}, \tag{13}$$

where i refers to a MSA and t refers to a year. There are $n = 92$ MSAs in total, and three years: 1980, 1990 and 2000. So $\Delta y_{i,t}$ is the change in log crimes per capita in MSA i between year t and year $t-1$, and the regressor of interest is $\Delta x_{i,t}$ which represents the change in the foreign-born Mexican population share between year t and year $t-1$. Control variables are denoted by the vector $w_{i,t}$, and include demographic changes in each MSA. The variable $\alpha_j \times \delta_t$ is the interaction between the region in which MSA i is located and time fixed effects. There are six regions of MSAs and since the variables are differenced, two time periods are considered so $T = 2$. The change in the foreign-born Mexican population share being endogenous, Chalfin (2015) uses the 2SLS estimator by first estimating

$$\Delta x_{i,t} = \pi_0 + \pi_1 z_{i,t} + \pi_2' w_{i,t} + \alpha_j \times \delta_t + v_{i,t}.$$

One then gets $\widehat{\Delta x_{i,t}}$, and uses these fitted values in the second stage regression (13). In the paper, both regressions are weighted by the 1980 MSA population that I denote ω_i, i.e. all the variables indexed by i are multiplied by the same weight ω_i, including the intercept and the fixed effects dummy variables. In this context, the first stage equation I estimate is

$$\Delta x_{i,t}^{\omega} = g_0\left(z_{i,t}^{\omega}\right) + \alpha_j \times \delta_t + v_{i,t},$$

where $\Delta x_{i,t}^{\omega} \equiv \Delta x_{i,t}\omega_i$ and $z_{i,t}^{\omega} \equiv z_{i,t}\omega_i$ correspond to the endogenous variable and the instrument after re-weighting by ω_i. I assume that the interacted fixed effect $\alpha_j \times \delta_t$ is uncorrelated with $z_{i,t}^{\omega}$. While this random effect assumption differs from Chalfin (2015)'s fixed effect assumption, it allows me to sink the fixed effect with the error term $v_{i,t}$ so that the convergence properties of the k-IV estimator are unaltered (see Li & Racine, 2007, Theorem 19.1, p.577). Aware of the limitations implied by this assumption, I leave the case of fixed effects in the first stage for future research. Thus, in the first stage, I do not distinguish between the MSA dimension and the time dimension, and the local linear estimator is used to estimate $g_0\left(z_{i,t}^{\omega}\right)$. The bandwidth is chosen via the same cross validation criterion as the one used for $\hat{\beta}_{NP}$ in section 4. Let $d_{j,t}^{i}$ be a dummy variable equal to one if MSA i belongs to region j and if it is observed at time t, zero else, i.e.:

$$d_{j,t}^{i} \equiv \begin{cases} 1 & \text{if MSA } i \text{ belongs to region } j \text{ and is observed at time } t, \\ 0 & \text{else.} \end{cases}.$$

Let X be the matrix that stacks the vectors $\left(1 \; x_{i,t} \; w_{i,t}' \; d_{1,1990}^{i}, \; d_{1,2000}^{i}, \; d_{2,1990}^{i}, ..., d_{6,2000}^{i}\right)$ vertically. Let $\hat{X}$ be the same matrix as X but $\Delta x_{i,t}$ is replaced by $\frac{\hat{g}\left(z_{i,t}^{\omega}\right)}{\omega_i}$ and Y the $n \times T$ vector that contains the $y_{i,t}$. I denote $W^{1/2}$ the $n \times T$ diagonal matrix containing the square root of the population weights $\omega_i^{1/2}$. Let

$$\hat{X}_{\omega} \equiv W^{1/2}\hat{X},$$
$$X_{\omega} \equiv W^{1/2}X,$$
$$Y_{\omega} \equiv W^{1/2}Y.$$

The k-IV estimator[8] is defined as:

$$\begin{aligned}\hat{\beta} &\equiv \left(\hat{X}_\omega' X_\omega\right)^{-1} \hat{X}_\omega' Y_\omega \\ &= \left(\hat{X}' W^{1/2} W^{1/2} X\right) \hat{X}' W^{1/2} W^{1/2} Y \\ &= \left(\hat{X}' W X\right) \hat{X}' W Y.\end{aligned}$$

I use the following estimator for the asymptotic variance to conduct inference:

$$\hat{\Sigma}_\omega \equiv \left(\left(\frac{1}{T}\sum_{t=1}^{T} \frac{1}{n}\sum_{i=1}^{n} \hat{x}_{i,t}\omega_i \hat{x}_{i,t}' \right)^{-1} \frac{1}{T}\sum_{t=1}^{T} \frac{1}{n}\sum_{i=1}^{n} \hat{x}_{i,t}\omega_i \hat{u}_{i,t}^2 \hat{x}_{i,t}' \left(\frac{1}{T}\sum_{t=1}^{T} \frac{1}{n}\sum_{i=1}^{n} \hat{x}_{i,t}\omega_i \hat{x}_{i,t}' \right)^{-1} \right),$$

where $\hat{u}_i$, $i = 1, ..., n$ are the residuals of the weighted regression from (13). Table 1 reports the OLS, 2SLS and k-IV estimates for each type of crime, along with their standard errors in parenthesis. The standard errors are clustered at the MSA level for the OLS and 2SLS estimates. Having not developed a theory of clustered standard errors for the k-IV estimator in this paper, I use $\hat{\Sigma}_\omega$ as a heteroskedasticity robust covariance matrix estimator. In spite of the presence of the weights, it can be shown that $\hat{\Sigma}_\omega$ is heteroskedasticity robust after redefining notation and using the statement in Theorem 3.

Table 1: OLS, 2SLS and k-IV estimates of the effect of Mexican immigration on Crimes Reported to Police

Violent crimes				Property crimes		
Murder	Rape	Robbery	Agg. assault	Burglary	Larceny	Vehicle theft
Panel A: OLS estimates						
0.075^{***}	0.002	0.095^{***}	0.037	0.025	0.000	0.042
(0.028)	(0.029)	(0.025)	(0.029)	(0.019)	(0.023)	(0.028)
Panel B: 2SLS estimates						
-0.022	-0.131^{***}	-0.049	0.197^{***}	-0.092	-0.105^{**}	-0.149^{*}
(0.102)	(0.062)	(0.094)	(0.075)	(0.057)	(0.047)	(0.081)
Panel C: k-IV estimates						
0.084^{***}	0.013	0.069^{**}	0.055^{*}	0.006	-0.054^{**}	-0.007
(0.029)	(0.027)	(0.033)	(0.039)	(0.028)	(0.025)	(0.036)

Note: OLS, 2SLS and k-IV estimates. All the regressions include region and year fixed effects. All models are weighted by 1980 MSA population and standard errors for the OLS and 2SLS estimates are clustered by region. One star indicates significance at the 10% level, two stars at the 5% level, and three stars at the 1% level.

The k-IV estimates feature lower standard errors than the 2SLS estimates, and are always between the OLS and 2SLS estimates (except for the murder case). They corroborate the significance of Mexican immigration on murder and robbery cases, and support the findings of the 2SLS estimator on larceny. However, no significance is found for vehicle theft nor rape, confirming the OLS results against the 2SLS ones.

6 Conclusion

Linear models featuring endogeneity are mostly estimated through two stage procedures in the applied literature. The first stage is a prediction problem in and of itself, that can have an influence on the second stage, the estimation problem. By allowing more flexibility in the first stage through the use of kernel estimators, one can obtain less biased and more efficient estimates than the traditional estimation tools.

[8] Note that since the intercept is also being weighted for the 2SLS estimates, $\hat{\beta}_{NP}$ and $\hat{\beta}_{2SLS}$ do not coincide for a high value of the bandwidth. Moreover, the first stage regression for $\hat{\beta}_{2SLS}$ contains the additional control variables present in the main regressions.

In particular, using a local linear estimator in the first stage not only includes the 2SLS estimator as a special case, but also provides ways to improve the estimates through the selection of the bandwidth. That additional flexibility is particularly useful when the relationship between the endogenous variable and the instruments is highly nonlinear, and when the degree of endogeneity is low. The nature of that relationship often being overlooked, a flexible estimator can potentially avoid the risk of using a misspecified linear model. In such cases, the MSE can be improved upon through a lower variance compared to the 2SLS estimator, as emphasized in the simulation study of section 4. The k-IV estimator also features reliable inference in all of the designs shown above, making it a candidate of choice over the 2SLS estimator. The estimator applied to Chalfin (2015)'s data set corroborates the OLS estimates in the cases of murder, rape, robbery, and vehicle theft whereas it supports the conclusions drawn by the 2SLS estimator for larceny and aggravated assault to a lesser extent.

This estimator is however likely to be subject to some curse of dimensionality as the number of instruments increases, making the LASSO and post LASSO estimators of Belloni *et al.* (2012) better suited alternatives in that case.

The choice of the bandwidth can be made based on the first stage problem using classical methods such as cross validation, and the resulting estimator is competitive ($\hat{\beta}_{NP}$ in the simulations), but since the second stage is of higher relevance to the researcher, the estimator can further be improved by choosing a criterion function related to the second stage estimates. A second stage cross validation criterion might not necessarily be relevant either, since one working with models such as the one described in Assumption 1 might rather be interested in the properties of the estimator itself, not its implied predictions.

References

Antoine, Bertille, & Lavergne, Pascal. 2014. Conditional moment models under semi-strong identification. *Journal of econometrics*, **182**(1), 59–69.

Belloni, Alexandre, Chen, Daniel, Chernozhukov, Victor, & Hansen, Christian. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, **80**(6), 2369–2429.

Bierens, Herman J. 1987. Kernel estimators of regression functions. *Pages 99–144 of: Advances in econometrics: Fifth world congress*, vol. 1.

Cattaneo, Matias D, & Jansson, Michael. 2018. Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency. *Econometrica*, **86**(3), 955–995.

Chalfin, Aaron. 2015. The long-run effect of mexican immigration on crime in us cities: Evidence from variation in mexican fertility rates. *American economic review*, **105**(5), 220–25.

Chamberlain, Gary. 1987. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of econometrics*, **34**(3), 305–334.

Donald, Stephen G., & Newey, Whitney K. 2001. Choosing the number of instruments. *Econometrica*, **69**(5), 1161–1191.

Greene, William H. 2003. *Econometric analysis*. Pearson Education India.

Hansen, Bruce E. 2008. Uniform convergence rates for kernel estimation with dependent data. *Econometric theory*, **24**(3), 726–748.

Hurvich, Clifford M, Simonoff, Jeffrey S, & Tsai, Chih-Ling. 1998. Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the royal statistical society: Series b (statistical methodology)*, **60**(2), 271–293.

Jorgenson, Dale W, & Laffont, Jean-Jacques. 1974. Efficient estimation of nonlinear simultaneous equations with additive disturbances. *Pages 615–640 of: Annals of economic and social measurement, volume 3, number 4*. NBER.

Kline, Patrick, Saggio, Raffaele, & Sølvsten, Mikkel. 2020. Leave-out estimation of variance components. *Econometrica*, **88**(5), 1859–1898.

Lee, Seojeong, & Shin, Youngki. 2018. Optimal estimation with complete subsets of instruments. *Mcmaster university, department of economics, working paper series*, **15**.

Levinsohn, James, & Petrin, Amil. 2003. Estimating production functions using inputs to control for unobservables. *The review of economic studies*, **70**(2), 317–341.

Li, Qi, & Racine, Jeffrey Scott. 2007. *Nonparametric econometrics: theory and practice*. Princeton University Press.

Linton, Oliver. 2002. Edgeworth approximations for semiparametric instrumental variable estimators and test statistics. *Journal of econometrics*, **106**(2), 325–368.

Masry, Elias. 1996. Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of time series analysis*, **17**(6), 571–599.

Newey, Whitney K. 1990. Efficient instrumental variables estimation of nonlinear models. *Econometrica: Journal of the econometric society*, 809–837.

Newey, Whitney K. 1994. The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the econometric society*, 1349–1382.

Newey, Whitney K, & McFadden, Daniel. 1994. Large sample estimation and hypothesis testing. *Handbook of econometrics*, **4**, 2111–2245.

Olley, G. Steven, & Pakes, Ariel. 1996. The dynamics of productivity in the telecommunications equipment industry. *Econometrica*, **64**(6), 1263–1297.

Pakes, Ariel, & Olley, Steven. 1995. A limit theorem for a smooth class of semiparametric estimators. *Journal of econometrics*, **65**(1), 295–332.

Pakes, Ariel, & Pollard, David. 1989. Simulation and the asymptotics of optimization estimators. *Econometrica: Journal of the econometric society*, 1027–1057.

Theil, Henri. 1958. *Economic forecasts and policy*. North Holland.

Wang, Xiaohu, & Yu, Jun. 2016. Double asymptotics for explosive continuous time models. *Journal of econometrics*, **193**(1), 35–53.

Young, Alwyn. 2017. *Consistency without inference: instrumental variables in practical application*. Tech. rept. Working Paper.

7 Appendix

The appendix includes proofs of the results of the paper, as well as supplementary material.

7.1 Proof of Theorem 1

Proof. Let

$$x_{0,i} \equiv \begin{pmatrix} 1 & g_0(z_i) & x_{2,i}' \end{pmatrix}',$$

$$X_0 \equiv \begin{pmatrix} 1 & g_0(z_1) & x_{2,1}' \\ 1 & g_0(z_2) & x_{2,2}' \\ \vdots & \vdots & \vdots \\ 1 & g_0(z_n) & x_{2,n}' \end{pmatrix},$$

and write the estimator as

$$\hat{\beta}-\beta_0=\left(\hat{X}'X\right)^{-1}\hat{X}'U=\left[\frac{\left(\hat{X}-X_0\right)'X+X_0'X}{n}\right]^{-1}\left[\frac{X_0'U+\left(\hat{X}-X_0\right)'U}{n}\right]. \tag{14}$$

The only difference between $\hat{X}$ and X_0 is in the second column, where the conditional expectation of $x_{1,i}$ given z_i, $g_0(z_i)$, is replaced by its kernel estimator $\hat{g}(z_i)$. The proof consists in showing that $\frac{\left(\hat{X}-X_0\right)'X}{n}$ and $\frac{\left(\hat{X}-X_0\right)'U}{n}$ are both $o_{\mathbb{P}}(1)$, and then use a weak law of large numbers on $\frac{X_0'X}{n}$ and $\frac{X_0'U}{n}$. For this, I use results on uniform convergence of the kernel estimator to show that when $\hat{g}(z_i)-g_0(z_i)$ is multiplied by u_i or x_i and summed over, the result is $o_{\mathbb{P}}(1)$. First, write

$$\begin{aligned}\frac{\left(\hat{X}-X_0\right)'U}{n}&=\frac{1}{n}\begin{pmatrix}0 & 0 & \cdots & 0\\ \hat{g}(z_1)-g_0(z_1) & \hat{g}(z_2)-g_0(z_2) & \cdots & \hat{g}(z_n)-g_0(z_n)\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{pmatrix}\begin{pmatrix}u_1\\ u_2\\ \vdots\\ u_n\end{pmatrix}\\ &=\begin{pmatrix}0\\ \frac{1}{n}\sum_{i=1}^{n}u_i\left(\hat{g}(z_i)-g_0(z_i)\right)\\ 0\\ \vdots\\ 0\end{pmatrix}.\end{aligned} \tag{15}$$

Under Assumption 3 and the assumption that $h\to 0$ and $\ln(n)/nh\to 0$, I have[9]

$$\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}(z_j)-g_0(z_j)\right|=o_{\mathbb{P}}(1),$$

so, for some $\varepsilon>0$:

$$\mathbb{P}\left(\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}(z_j)-g_0(z_j)\right|<\varepsilon\right)\to 1. \tag{16}$$

Hence, $\forall i$:

$$\mathbb{P}\left(|u_i|\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}(z_j)-g_0(z_j)\right|<|u_i|\,\varepsilon\right)\to 1,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}|u_i|\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}(z_j)-g_0(z_j)\right|<\frac{1}{n}\sum_{i=1}^{n}|u_i|\,\varepsilon\right)\to 1.$$

Since

$$\frac{1}{n}\sum_{i=1}^{n}|u_i|\left|\hat{g}(z_j)-g_0(z_j)\right|\leq\frac{1}{n}\sum_{i=1}^{n}|u_i|\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}(z_j)-g_0(z_j)\right|\quad a.s.,$$

then I also have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}|u_i|\left|\hat{g}(z_i)-g_0(z_i)\right|<\frac{1}{n}\sum_{i=1}^{n}|u_i|\,\varepsilon\right)\to 1,$$

[9] See Li & Racine (2007), Theorem 2.6 in chapter 2. The statement of the theorem for a q dimensional vector of regressors is $\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}(z_j)-g_0(z_j)\right|=O\left(\frac{(\ln n)^{1/2}}{(nh_1\ldots h_q)^{1/2}}+\sum_{s=1}^{q}h_s^2\right)\ a.s.$

and as

$$\frac{1}{n}\sum_{i=1}^{n} u_i\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right) \leq \frac{1}{n}\sum_{i=1}^{n}\left|u_i\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right| \quad a.s., \tag{17}$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} u_i\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right) < \frac{1}{n}\sum_{i=1}^{n}\left|u_i\right|\varepsilon\right) \geq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|u_i\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right| < \frac{1}{n}\sum_{i=1}^{n}\left|u_i\right|\varepsilon\right) \to 1.$$

The fact that $\mathbb{E}\left[\left|u_i^2\right|\right] < \infty$ that implies $\mathbb{E}\left[\left|u_i\right|\right] < \infty$ ensures that the right hand side of the inequality inside the probability is finite as n goes to infinity, and since $\varepsilon > 0$ is arbitrary:

$$\frac{1}{n}\sum_{i=1}^{n} u_i\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right) = o_{\mathbb{P}}\left(1\right), \tag{18}$$

and

$$\frac{\left(\hat{X}-X_0\right)' U}{n} = o_{\mathbb{P}}\left(1\right).$$

Consider now the term $\frac{\left(\hat{X}-X_0\right)' X}{n}$ in (14). It corresponds to

$$\begin{aligned}\frac{\left(\hat{X}-X_0\right)' X}{n} &= \frac{1}{n}\begin{pmatrix} 0 & 0 & \cdots & 0 \\ \hat{g}\left(z_1\right)-g_0\left(z_1\right) & \hat{g}\left(z_2\right)-g_0\left(z_2\right) & \cdots & \hat{g}\left(z_n\right)-g_0\left(z_n\right) \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}\begin{pmatrix} 1 & x_{1,1} & x'_{2,1} \\ 1 & x_{1,2} & x'_{2,2} \\ 1 & x_{1,3} & x'_{2,3} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x'_{2,n} \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ \frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right) & \frac{1}{n}\sum_{i=1}^{n} x_{1,i}\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right) & \frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right) x'_{2,i} \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \end{pmatrix}.\end{aligned}$$

Assumption 4 ensures that each component of x_i have a finite expectation in absolute value, so I can use the same argument as for the term in (15) to show that

$$\frac{\left(\hat{X}-X_0\right)' X}{n} = o_{\mathbb{P}}\left(1\right), \tag{19}$$

by replacing u_i by each component of x_i from expression (16) to (18). Thus:

$$\hat{\beta}-\beta_0 = \left[\frac{X_0' X}{n} + o_{\mathbb{P}}\left(1\right)\right]^{-1}\left[\frac{X_0' U}{n} + o_{\mathbb{P}}\left(1\right)\right].$$

Because the random vector $x_{2,i}$ is exogenous, I have

$$\mathbb{E}\left[x_{0,i} x_i'\right] = \mathbb{E}\begin{bmatrix} 1 & x_{1,i} & x'_{2,i} \\ g_0\left(z_i\right) & g_0\left(z_i\right) x_{1,i} & g_0\left(z_i\right) x'_{2,i} \\ x_{2,i} & x_{2,i} x_{1,i} & x_{2,i} x'_{2,i} \end{bmatrix},$$

where $\mathbb{E}\left[g_0\left(z_i\right) x_{1,i}\right] = \mathbb{E}\left[g_0\left(z_i\right)\left(g_0\left(z_i\right)+v_i\right)\right] = \mathbb{E}\left[g_0^2\left(z_i\right)\right]$ and $\mathbb{E}\left[x_{2,i} x_{1,i}\right] = \mathbb{E}\left[x_{2,i}\left(g_0\left(z_i\right)+v_i\right)\right] = \mathbb{E}\left[g_0\left(z_i\right) x_{2,i}\right]$ are implied by Assumption 1. Hence:

$$\mathbb{E}\left[x_{0,i} x_i'\right] = \mathbb{E}\left[x_{0,i} x_{0,i}'\right], \tag{20}$$

which is nonsingular by Assumption 4. Coupled with Assumption 2, I invoke Khinchin's weak law of large numbers (see Li & Racine, 2007, Lemma A.2, p.668) to get

$$\frac{X_0'U}{n} = \frac{1}{n}\sum_{i=1}^{n} x_{0,i}u_i \xrightarrow{\mathbb{P}} \mathbb{E}\left[x_{0,i}u_i\right] = 0,$$

and

$$\frac{X_0'X}{n} = \frac{1}{n}\sum_{i=1}^{n} x_{0,i}x_i' \xrightarrow{\mathbb{P}} \mathbb{E}\left[x_{0,i}x_i'\right] = \mathbb{E}\left[x_{0,i}x_{0,i}'\right]$$

Finally:

$$\hat{\beta} - \beta_0 = o_{\mathbb{P}}(1).$$

□

7.2 Proof of Lemma 1

Proof. Consider the local linear estimator of $g_0(z)$ for some value z that is the solution to the following problem:

$$\min_{\{a,b\}} \left\{ \sum_{i=1}^{n} \left(x_i - a - (z_i - z)\, b\right)^2 K\left(\frac{z_i - z}{h}\right) \right\}.$$

Taking the limit of this problem as $h \to \infty$ yields

$$\min_{\{a,b\}} \left\{ \sum_{i=1}^{n} \left(x_i - a - (z_i - z)\, b\right)^2 K(0) \right\},$$

which amounts to a linear regression of x_i on $z_i - z$ and a constant (as $K(0)$ will not enter the first order conditions). It is equivalent to fitting the model:

$$x_i = a + (z_i - z)\, b + v_i,$$

so that $g(z) = \mathbb{E}\left[x_i | z_i = z\right] = a$.[10] The solutions to that problem are:

$$\hat{b}(z) = \frac{cov(x_i, z_i - z)}{var(z_i - z)} = \frac{cov(x_i, z_i)}{var(z_i)},$$
$$\hat{a}(z) = \bar{x} - (\bar{z} - z)\,\hat{b}(z),$$

and

$$\begin{aligned}\hat{g}(z) &= \hat{\mathbb{E}}\left[x_i | z_i = z\right] \\ &= \hat{a}(z) + \hat{b}(z)(z - z) \\ &= \bar{x} - (\bar{z} - z)\,\hat{b}(z).\end{aligned}$$

Note that $\hat{b}(z)$ is equal to the least squares estimate of the slope of the regression of x_i on z_i (and an intercept) that we denote $\hat{\alpha}_1$. The predicted value of $g_0(z)$ in the case of an OLS regression is

$$\begin{aligned}\hat{\alpha}_0 + \hat{\alpha_1} z &= \bar{x} - \hat{\alpha}_1 \bar{z} + \hat{\alpha_1} z \\ &= \bar{x} - (\bar{z} - z)\,\hat{\alpha_1} \\ &= \bar{x} - (\bar{z} - z)\,\hat{b}(z).\end{aligned}$$

[10] This notation is borrowed from Bruce Hansen's lecture notes on nonparametric econometrics.

Thus, a local linear estimator becomes "global" as the length of the bandwidth increases to infinity. For the purpose of the model considered in the paper, the first stage computes predictions not at an arbitrary value z but at the observed values z_i, $i = 1, ..., n$. So the predicted values of $g_0(z_i)$ using the local linear estimator with a bandwidth equal to infinity are the same as the predicted values of $g_0(z_i)$ using an OLS estimation of the regression of x_i on z_i and a constant, i.e. $\hat{g} \equiv \begin{pmatrix} \hat{g}(z_1) \\ \hat{g}(z_2) \\ \vdots \\ \hat{g}(z_n) \end{pmatrix} = P_Z X$. The estimator of β_0 is then

$$\hat{\beta} \overset{h \to \infty}{\to} (X' P_Z X)^{-1} X' P_Z Y = \hat{\beta}_{2SLS}.$$

Note that, due to the idempotence of P_Z, one has $\hat{\beta}_{2SLS} = \left(\hat{X}'X\right)^{-1} \hat{X}'Y = \left(\hat{X}'\hat{X}\right)^{-1} \hat{X}'Y$ so the 2SLS estimator can also be computed by regressing y_i on $\hat{x}_i$. It is not the case for an arbitrary h however as the hat matrix L is neither symmetric nor idempotent.[11] □

7.3 Proof of Theorem 2

Proof. Write

$$\begin{aligned} \sqrt{n}\left(\hat{\beta} - \beta_0\right) &= \left(\hat{X}'X\right)^{-1} \sqrt{n} \hat{X}'U \\ &= \left[\frac{\left(\hat{X} - X_0\right)' X + X_0' X}{n}\right]^{-1} \sqrt{n} \left[\frac{X_0' U + \left(\hat{X} - X_0\right)' U}{n}\right]. \end{aligned}$$

The proof is similar to the consistency proof, except that the term $\sqrt{n}\frac{\left(\hat{X} - X_0\right)' U}{n}$ has to be handled using a Central Limit Theorem. That term can be written

$$\sqrt{n}\frac{\left(\hat{X} - X_0\right)' U}{n} = \begin{pmatrix} 0 \\ \sqrt{n}\frac{1}{n}\sum_{i=1}^n u_i \left(\hat{g}(z_i) - g_0(z_i)\right) \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

From (17), I know that, for some $\varepsilon > 0$:

$$\begin{aligned} \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n u_i \left(\hat{g}(z_i) - g_0(z_i)\right) < \frac{1}{n}\sum_{i=1}^n |u_i| \varepsilon\right) \to 1, \\ \mathbb{P}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^n u_i \left(\hat{g}(z_i) - g_0(z_i)\right) < \sqrt{n}\frac{1}{n}\sum_{i=1}^n |u_i| \varepsilon\right) \to 1. \end{aligned}$$

[11] Note that, due to the symmetry and idempotence of P_Z, one has $\hat{\beta}_{2SLS} = \left(\hat{X}'X\right)^{-1} \hat{X}'Y = (X' P_Z X)^{-1} \hat{X}' P_Z Y = (X' P_Z P_Z X)^{-1} \hat{X}' P_Z Y = \left(\hat{X}'\hat{X}\right)^{-1} \hat{X}'Y$ so the 2SLS estimator can also be computed by regressing y_i on $\hat{x}_i$. It is not the case for an arbitrary h however as the hat matrix L is neither symmetric nor idempotent. The estimator in this case would be $\frac{\frac{1}{n}\sum_{i=1}^n \hat{g}(z_i) y_i}{\frac{1}{n}\sum_{i=1}^n \hat{g}(z_i)\hat{g}(z_i)} \neq \frac{\frac{1}{n}\sum_{i=1}^n \hat{g}(z_i) y_i}{\frac{1}{n}\sum_{i=1}^n \hat{g}(z_i) x_i}$.

By Lindeberg-Lévy's central limit Theorem (see Greene, 2003, Chapter 4), since $\mathbb{E}[|u_i|] \equiv \zeta$ and $\mathbb{E}\left[|u_i|^2\right]$, I have:

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}|u_i| \xrightarrow{d} \mathcal{N}\left(\zeta, \xi^2\right),$$

where $\xi^2 \equiv \mathbb{V}[|u_i|] < \infty$. This way, I obtain that $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}|u_i| = O_{\mathbb{P}}(1)$ and $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}u_i\left(\hat{g}(z_i) - g_0(z_i)\right) = o_{\mathbb{P}}(1)$ from ε being arbitrarily small.

Now, the term $\frac{(\hat{X}-X_0)'X}{n}$ can be handled the same way as for Theorem 1, so that I have $\frac{(\hat{X}-X_0)'X}{n} = o_{\mathbb{P}}(1)$ following the argument leading to (19). Using Khinchin's law of large numbers (see Li & Racine, 2007, Lemma A.2, p.668),

$$\frac{X_0'X}{n} \xrightarrow{\mathbb{P}} \mathbb{E}[x_{0,i}x_i'] = \mathbb{E}\left[x_{0,i}x_{0,i}'\right],$$

from (20). Assumptions 2 and 4 allow Lindeberg-Lévy's central limit Theorem to apply to $\sqrt{n}\frac{X_0'U}{n}$ so that

$$\sqrt{n}\frac{X_0'U}{n} = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}x_{0,i}u_i \xrightarrow{d} \mathcal{N}(0, \Omega_0),$$

where Ω_0 is defined in (8). Finally, combining the terms and applying Slutsky's Theorem:

$$\left(\frac{X_0'X}{n}\right)^{-1}\sqrt{n}\frac{X_0'U}{n} \xrightarrow{d} \mathbb{E}\left[x_{0,i}x_{0,i}'\right]\mathcal{N}(0, \Omega_0),$$

so

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1}\Omega_0\left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1}\right),$$

which corresponds to the definition of Σ. In the homoskedastic case, $\sigma^2(z_i, x_{2,i}) = \sigma^2$ so Σ simplifies to:

$$\begin{aligned}\Sigma &= \left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1}\mathbb{E}\left[x_{0,i}\sigma^2 x_{0,i}'\right]\left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1} \\ &= \sigma^2\left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1}\mathbb{E}\left[x_{0,i}\sigma^2 x_{0,i}'\right]\left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1} \\ &= \sigma^2\left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1},\end{aligned}$$

so that

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1}\right).$$

□

7.4 Proof of Theorem 3

Proof. Recall the estimator of the asymptotic variance in the heteroskedastic case, along with the dimensions of each block:

$$\hat{\Sigma} \equiv \underset{[(d+2)\times(d+2)]}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{x}_i'\right)^{-1}}\underset{[(d+2)\times(d+2)]}{\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{u}_i^2\hat{x}_i'}\underset{[(d+2)\times(d+2)]}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{x}_i'\right)^{-1}}, \tag{21}$$

I recall the notation

$$\begin{aligned}\hat{x}_i &= \left(1,\ \hat{g}(z_i)\ \ x_{2,i}'\right)', \\ x_{0,i} &= \left(1,\ g_0(z_i)\ \ x_{2,i}'\right)'.\end{aligned}$$

Assume the conditions need for Theorem 3 are satisfied. I decompose the first term

$$\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{x}_i' = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{x}_i - x_{0,i}\right)\hat{x}_i' + x_{0,i}\left(\hat{x}_i - x_{0,i}\right)' + x_{0,i}x_{0,i}'. \tag{22}$$

The proof consists in showing that the first two terms are $o_{\mathbb{P}}(1)$, and using Khinchin's law of large numbers on the last term. Let 0^d be a d-dimensional vector of zeros and $0^{d\times d}$ a $d \times d$ matrix of zeros. I have:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{x}_i - x_{0,i}\right)\hat{x}_i' &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} 0 \\ \left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right) \\ 0^d \end{pmatrix}\begin{pmatrix} 0 & \hat{g}\left(z_i\right) & x_{2,i}' \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)\hat{g}\left(z_i\right) & \frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)x_{2,i}' \\ 0^d & 0^d & 0^{d\times d} \end{pmatrix}, \end{aligned} \tag{23}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}x_{0,i}\left(\hat{x}_i - x_{0,i}\right)' = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{n}\sum_{i=1}^{n}g_0\left(z_i\right)\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right) & \frac{1}{n}\sum_{i=1}^{n}x_{2,i}'\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right) \\ 0^d & 0^d & 0^{d\times d} \end{pmatrix}. \tag{24}$$

I now tackle expression (23). I consider the diagonal element

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)\hat{g}\left(z_i\right).$$

For some $\varepsilon > 0$:

$$\mathbb{P}\left(\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \varepsilon\right) \to 1.$$

Hence, $\forall i$:

$$\mathbb{P}\left(\left|\hat{g}\left(z_i\right)\right|\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \left|\hat{g}\left(z_i\right)\right|\varepsilon\right) \to 1,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\varepsilon\right) \to 1.$$

Now, by the triangle inequality:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right| &= \frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right) - g_0\left(z_i\right) + g_0\left(z_i\right)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right| + \frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| + \frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right| \\ &= \sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| + \frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right| \\ &= o_{\mathbb{P}}(1) + \frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|. \end{aligned}$$

Using the assumption that $\mathbb{E}\left[\left|g_0\left(z_i\right)\right|\right]<\infty$, I then have

$$\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right| \leq o_{\mathbb{P}}\left(1\right)+\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|<\infty.$$

Hence:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right| \sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right)-g_0\left(z_j\right)\right|<\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\varepsilon\right)\rightarrow 1.$$

Since

$$\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right| \sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right)-g_0\left(z_j\right)\right| \quad a.s.,$$

then I also have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right|<\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\varepsilon\right)\rightarrow 1,$$

and as

$$\frac{1}{n}\sum_{i=1}^{n}\hat{g}\left(z_i\right)\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right) \leq \frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right| \quad a.s., \tag{25}$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{g}\left(z_i\right)\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right)<\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\varepsilon\right) \geq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right|<\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)\right|\varepsilon\right)\rightarrow 1.$$

Finally:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right)\hat{g}\left(z_i\right)=o_{\mathbb{P}}\left(1\right).$$

The term $\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right)x_{2,i}'$ can be handled analogously, using the assumption $\mathbb{E}\left[\left|x_{2,i}\right|\right]<\infty$. I finally get

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{x_i}-x_{0,i}\right)\hat{x}_i'=o_{\mathbb{P}}\left(1\right).$$

I now turn to expression (24). The term $\frac{1}{n}\sum_{i=1}^{n}x_{2,i}'\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right)$ has already been covered above, so I consider

$$\frac{1}{n}\sum_{i=1}^{n}g_0\left(z_i\right)\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right).$$

For some $\varepsilon>0$:

$$\mathbb{P}\left(\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right)-g_0\left(z_j\right)\right|<\varepsilon\right)\rightarrow 1.$$

Hence, $\forall i$:

$$\mathbb{P}\left(\left|g_0\left(z_i\right)\right| \sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right)-g_0\left(z_j\right)\right|<\left|g_0\left(z_i\right)\right|\varepsilon\right)\rightarrow 1,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right| \sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right)-g_0\left(z_j\right)\right|<\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|\varepsilon\right)\rightarrow 1.$$

Since

$$\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right| \leq \frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right| \sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right)-g_0\left(z_j\right)\right| \quad a.s.,$$

then I also have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right|<\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|\varepsilon\right)\to 1,$$

and as

$$\frac{1}{n}\sum_{i=1}^{n}g_0\left(z_i\right)\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right)\leq\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right|\quad a.s.,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}g_0\left(z_i\right)\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right)<\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|\varepsilon\right)\geq\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|\left|\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right|<\frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|\varepsilon\right)\to 1.$$

Finally:

$$\frac{1}{n}\sum_{i=1}^{n}g_0\left(z_i\right)\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right)=o_{\mathbb{P}}\left(1\right).$$

To sum up, the following holds:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{x}_i-x_{0,i}\right)\hat{x}_i'&=o_{\mathbb{P}}\left(1\right),\\ \frac{1}{n}\sum_{i=1}^{n}x_{0,i}\left(\hat{x}_i-x_{0,i}\right)'&=o_{\mathbb{P}}\left(1\right),\\ \frac{1}{n}\sum_{i=1}^{n}x_{2,i}'\left(\hat{g}\left(z_i\right)-g_0\left(z_i\right)\right)&=o_{\mathbb{P}}\left(1\right),\end{aligned}$$

so that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{x}_i'&=o_{\mathbb{P}}\left(1\right)+o_{\mathbb{P}}\left(1\right)+\frac{1}{n}\sum_{i=1}^{n}x_{0,i}x_{0,i}'\\ &\xrightarrow{\mathbb{P}}\mathbb{E}\left[x_{0,i}x_{0,i}'\right],\end{aligned}\tag{26}$$

by Khinchin's law of large numbers.

I now handle the middle term of (21):

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{u}_i^2\hat{x}_i'&=\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\left(\left(x_i'\left(\beta_0-\hat{\beta}\right)\right)^2+u_i^2+2u_ix_i'\left(\beta_0-\hat{\beta}\right)\right)\hat{x}_i'\\ &=\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\left(o_{\mathbb{P}}\left(1\right)+u_i^2+o_{\mathbb{P}}\left(1\right)\right)\hat{x}_i'\\ &=\frac{1}{n}\sum_{i=1}^{n}\hat{x}_iu_i^2\hat{x}_i'+o_{\mathbb{P}}\left(1\right)\\ &=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{x}_i-x_{0,i}\right)u_i^2\hat{x}_i'+x_{0,i}u_i^2\hat{x}_i'\\ &=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{x}_i-x_{0,i}\right)u_i^2\hat{x}_i'+x_{0,i}u_i^2\left(\hat{x}_i-x_{0,i}\right)'+x_{0,i}u_i^2x_{0,i}'\end{aligned}\tag{27}$$

The proof consists in showing that the first two terms are $o_{\mathbb{P}}(1)$ as was done for (22) and using the law of large numbers on the last term. I have:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\hat{x}_i - x_{0,i}\right)u_i^2\hat{x}_i' &= \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} 0 \\ \left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)u_i^2 \\ 0^d \end{pmatrix}\begin{pmatrix}1 & \hat{g}\left(z_i\right) & x_{2,i}'\end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)u_i^2\hat{g}\left(z_i\right) & \frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)u_i^2 x_{2,i}' \\ 0^d & 0^d & 0^{d\times d}\end{pmatrix}, \quad (28)\end{aligned}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}x_{0,i}u_i^2\left(\hat{x}_i - x_{0,i}\right)' = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{n}\sum_{i=1}^{n}g_0\left(z_i\right)u_i^2\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right) & \frac{1}{n}\sum_{i=1}^{n}x_{2,i}'u_i^2\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right) \\ 0^d & 0^d & 0^{d\times d}\end{pmatrix}. \quad (29)$$

Let $|x_{2,i}|$ denote the vector of absolute values of each component of $x_{2,i}$.

I now tackle the first term, expression (28). I consider the diagonal element

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)u_i^2\hat{g}\left(z_i\right).$$

For some $\varepsilon > 0$:

$$\mathbb{P}\left(\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \varepsilon\right) \to 1.$$

Hence, $\forall i$:

$$\mathbb{P}\left(\left|\hat{g}\left(z_i\right)u_i^2\right|\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \left|\hat{g}\left(z_i\right)u_i^2\right|\varepsilon\right) \to 1,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)u_i^2\right|\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)u_i^2\right|\varepsilon\right) \to 1.$$

Now,

$$\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)u_i^2\right| = \frac{1}{n}\sum_{i=1}^{n}\left|\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)u_i^2 + g_0\left(z_i\right)u_i^2\right|.$$

By the triangle inequality:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left|\hat{g}\left(z_i\right)u_i^2\right| &= \frac{1}{n}\sum_{i=1}^{n}\left|\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)u_i^2 + g_0\left(z_i\right)u_i^2\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\left|\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)u_i^2\right| + \frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)u_i^2\right| \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right|u_i^2 + \frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)u_i^2\right| \\ &= \sup_{\{z_j\in\mathcal{Z}\}}\left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right|\frac{1}{n}\sum_{i=1}^{n}u_i^2 + \frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)u_i^2\right| \\ &= o_{\mathbb{P}}(1)\,O_{\mathbb{P}}(1) + \frac{1}{n}\sum_{i=1}^{n}\left|g_0\left(z_i\right)\right|u_i^2.\end{aligned}$$

Using the assumption that $\mathbb{E}\left[\left|g_0\left(z_i\right)u_i^2\right|\right] < \infty$, I then have, since $\frac{1}{n}\sum_{i=1}^n \left|g_0\left(z_i\right)\right| u_i^2 \leq o_{\mathbb{P}}\left(1\right) O_{\mathbb{P}}\left(1\right) + \frac{1}{n}\sum_{i=1}^n \left|g_0\left(z_i\right)\right| u_i^2$:

$$\frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| < \infty.$$

Hence:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \sup_{\{z_j \in \mathcal{Z}\}} \left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \varepsilon\right) \to 1.$$

Since

$$\frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \left|\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right| \leq \frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \sup_{\{z_j \in \mathcal{Z}\}} \left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| \quad a.s.,$$

then I also have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \left|\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right| < \frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \varepsilon\right) \to 1,$$

and as

$$\frac{1}{n}\sum_{i=1}^n \hat{g}\left(z_i\right)u_i^2\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right) \leq \frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \left|\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right| \quad a.s.,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \hat{g}\left(z_i\right)u_i^2\left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right) < \frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \varepsilon\right) \geq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \left|\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right| < \frac{1}{n}\sum_{i=1}^n \left|\hat{g}\left(z_i\right)u_i^2\right| \varepsilon\right) \to 1.$$

Finally:

$$\frac{1}{n}\sum_{i=1}^n \left(\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right)u_i^2\hat{g}\left(z_i\right) = o_{\mathbb{P}}\left(1\right).$$

I now look at the second term, expression (29).

Uniform convergence of $\hat{g}\left(z_i\right)$ means that

$$\sup_{\{z_j \in \mathcal{Z}\}} \left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| = o_{\mathbb{P}}\left(1\right),$$

so, for some $\varepsilon > 0$:

$$\mathbb{P}\left(\sup_{\{z_j \in \mathcal{Z}\}} \left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \varepsilon\right) \to 1. \tag{30}$$

Hence, $\forall i$:

$$\mathbb{P}\left(\left|g_0\left(z_i\right)u_i^2\right| \sup_{\{z_j \in \mathcal{Z}\}} \left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \left|g_0\left(z_i\right)u_i^2\right| \varepsilon\right) \to 1,$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \left|g_0\left(z_i\right)u_i^2\right| \sup_{\{z_j \in \mathcal{Z}\}} \left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| < \frac{1}{n}\sum_{i=1}^n \left|g_0\left(z_i\right)u_i^2\right| \varepsilon\right) \to 1.$$

Since

$$\frac{1}{n}\sum_{i=1}^n \left|g_0\left(z_i\right)u_i^2\right| \left|\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right| \leq \frac{1}{n}\sum_{i=1}^n \left|g_0\left(z_i\right)u_i^2\right| \sup_{\{z_j \in \mathcal{Z}\}} \left|\hat{g}\left(z_j\right) - g_0\left(z_j\right)\right| \quad a.s.,$$

then I also have

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \left|g_0\left(z_i\right)u_i^2\right| \left|\hat{g}\left(z_i\right) - g_0\left(z_i\right)\right| < \frac{1}{n}\sum_{i=1}^n \left|g_0\left(z_i\right)u_i^2\right| \varepsilon\right) \to 1,$$

and as

$$\frac{1}{n}\sum_{i=1}^{n} g_0(z_i)\, u_i^2\, (\hat{g}(z_i) - g_0(z_i)) \leq \frac{1}{n}\sum_{i=1}^{n} \left|g_0(z_i)\, u_i^2\right| |\hat{g}(z_i) - g_0(z_i)| \quad a.s., \tag{31}$$

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} g_0(z_i)\, u_i^2\, (\hat{g}(z_i) - g_0(z_i)) < \frac{1}{n}\sum_{i=1}^{n} \left|g_0(z_i)\, u_i^2\right| \varepsilon\right) \geq \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} \left|g_0(z_i)\, u_i^2\right| |\hat{g}(z_i) - g_0(z_i)| < \frac{1}{n}\sum_{i=1}^{n} \left|g_0(z_i)\, u_i^2\right| \varepsilon\right) \to 1, \tag{32}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} g_0(z_i)\, u_i^2\, (\hat{g}(z_i) - g_0(z_i)) = o_{\mathbb{P}}(1).$$

The term $\frac{1}{n}\sum_{i=1}^{n} x'_{2,i} u_i^2\, (\hat{g}(z_i) - g_0(z_i))$ can be handled analogously by using the assumption that $\mathbb{E}\left[\left|x_{2,i} u_i^2\right|\right] < \infty$ to get

$$\frac{1}{n}\sum_{i=1}^{n} x'_{2,i} u_i^2\, (\hat{g}(z_i) - g_0(z_i)) = o_{\mathbb{P}}(1).$$

The term $(\hat{g}(z_i) - g_0(z_i))\, u_i^2 x'_{2,i}$ is the same as $x'_{2,i} u_i^2\, (\hat{g}(z_i) - g_0(z_i))$ and can be handled analogously by assuming $\mathbb{E}\left[|x_{2,i}|\, u_i^2\right] < \infty$. Finally:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} (\hat{x}_i - x_{0,i})\, u_i^2 \hat{x}'_i &= o_{\mathbb{P}}(1), \\
\frac{1}{n}\sum_{i=1}^{n} x_{0,i} u_i^2\, (\hat{x}_i - x_{0,i})' &= o_{\mathbb{P}}(1), \\
\frac{1}{n}\sum_{i=1}^{n} x'_{2,i} u_i^2\, (\hat{g}(z_i) - g_0(z_i)) &= o_{\mathbb{P}}(1), \\
\frac{1}{n}\sum_{i=1}^{n} (\hat{g}(z_i) - g_0(z_i))\, u_i^2 x'_{2,i} &= o_{\mathbb{P}}(1),
\end{aligned}$$

and what is left of (27) is

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \hat{u}_i^2 \hat{x}'_i &= \frac{1}{n}\sum_{i=1}^{n} x_{0,i} u_i^2 x'_{0,i} + o_{\mathbb{P}}(1) \\
&\xrightarrow{\mathbb{P}} \mathbb{E}\left[x_{0,i} u_i^2 x'_{0,i}\right],
\end{aligned} \tag{33}$$

by Khinchin's law of large numbers and the fact that $\mathbb{E}\left[x_{0,i} u_i^2 x'_{0,i}\right]$ is finite. Finally, assembling (26) and (33):

$$\hat{\Sigma} = \left(\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \hat{x}'_i\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \hat{u}_i^2 \hat{x}'_i \left(\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \hat{x}'_i\right)^{-1} \xrightarrow{\mathbb{P}} \left(\mathbb{E}\left[x_{0,i} x'_{0,i}\right]\right)^{-1} \mathbb{E}\left[x_{0,i} u_i^2 x'_{0,i}\right] \left(\mathbb{E}\left[x_{0,i} x'_{0,i}\right]\right)^{-1} = \Sigma.$$

Consider now the estimator of the asymptotic variance in the homoskedastic case

$$\hat{\sigma}^2 \left(\frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \hat{x}'_i\right)^{-1},$$

where $\hat{\sigma}^2 \equiv \frac{1}{n-(d+2)}\sum_{i=1}^{n} \hat{u}_i^2$. The second term converges in probability to $\mathbb{E}\left[x_{0,i} x'_{0,i}\right]$ following (26). I now turn to the estimator of the variance of the error term:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-(d+2)}\sum_{i=1}^{n}\left(y_i - x_i'\hat{\beta}\right)^2 \\ &= \frac{1}{n-(d+2)}\sum_{i=1}^{n}\left(\left(x_i'\left(\beta_0-\hat{\beta}\right)\right)^2 + u_i^2 + 2u_i x_i'\left(\beta_0-\hat{\beta}\right)\right) \\ &= \frac{1}{n-(d+2)}\sum_{i=1}^{n} u_i^2 + o_{\mathbb{P}}(1) \\ &\xrightarrow{\mathbb{P}} \mathbb{E}\left[u_i^2\right] = \mathbb{E}\left[\mathbb{E}\left[u_i^2|z_i, x_{2,i}'\right]\right] = \sigma^2,\end{aligned}$$

where the $o_{\mathbb{P}}(1)$ come from the fact that $\hat{\beta} \xrightarrow{\mathbb{P}} \beta_0$ from Assumptions 2, and 3, and the conditions $h \to 0$ and $\ln(n)/nh \to 0$. Combining the terms:

$$\hat{\sigma}^2\left(\frac{1}{n}\sum_{i=1}^{n}\hat{x}_i\hat{x}_i'\right)^{-1} \xrightarrow{\mathbb{P}} \sigma^2\left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1}.$$

□

7.5 Extension to multiple instruments

In this section, I consider the same estimator as in (4) but I allow the function g_0 to depend on multiple instruments.

Assumption 5. *The random variables satisfy the following conditions*

$$y_i = a + x_{1,i}b + x_{2,i}'c + u_i, \tag{34}$$
$$x_{1,i} = g_0(z_i) + v_i, \tag{35}$$
$$\mathbb{E}\left[u_i|z_i, x_{2,i}\right] = \mathbb{E}\left[v_i|z_i, x_{2,i}\right] = 0. \tag{36}$$

where $x_{1,i}$ is of dimension 1, $x_{2,i}$ is of dimension d, $z_i \equiv (z_{i,1}, z_{i,2}, ..., z_{i,q})'$ is of dimension q and y_i is univariate. The joint pdf of z_i is denoted $f_Z(z) = f_Z(z_1, z_2, ..., z_q)$.

The sample considered is still i.i.d., as stated in the following assumption.

Assumption 6. *An i.i.d. random sample of size n is available: $\{(y_i, x_i, z_i), i = 1, ..., n\}$.*

The same k-IV estimator as (4) is considered, but $\hat{g}(z_i) = \hat{g}(z_{1,i}, z_{2,i}, ..., z_{q,i})$ is obtained via the product of univariate kernel functions with a different bandwidth for each $z_{j,i}, j = 1, ..., q$. Let the product kernel be denoted by $\mathcal{K}\left(\frac{z_i - z}{h}\right) \equiv K\left(\frac{z_{1,i}-z_1}{h_q}\right) \times ... \times K\left(\frac{z_{q,i}-z_q}{h_q}\right)$. The estimator is defined as in (4), but the kernel functions $K\left(\frac{z_i-z}{h}\right)$ appearing in expressions (5) and (6) are replaced by $\mathcal{K}\left(\frac{z_i-z}{h}\right)$. Thus, $\hat{X} = LX$ as before but the weights in L contain the product kernel instead of the original kernel. The notation becomes

$$\hat{X} \equiv \begin{pmatrix} 1 & \hat{g}(z_{1,1}, z_{2,1}, ..., z_{q,1}) & x_{2,1}' \\ 1 & \hat{g}(z_{1,2}, z_{2,2}, ..., z_{q,2}) & x_{2,2}' \\ \vdots & \vdots & \vdots \\ 1 & \hat{g}(z_{1,n}, z_{2,n}, ..., z_{q,n}) & x_{2,n}' \end{pmatrix},$$

$$\hat{\beta} = \left(\hat{X}'X\right)^{-1}\hat{X}'Y. \tag{37}$$

The next assumption is a set of conditions needed to ensure convergence of the kernel estimator is uniform. The conditions are the same as for the single instrument case, except that here the product kernel is the product of the K.

Assumption 7. *Assume,* $\forall i$:

(i)$f_Z(z_i)$ *is differentiable almost surely over* S, $g_0(z_i)$ *is twice differentiable almost surely over* S, $|m(z_i) - m(z)| \leq C|z_i - z|$ $\forall i$ *for some* $C > 0$, *where* $m = g_0''$ *and* f_Z'.

(ii) $\mathbb{E}\left[v_i^2|z\right] \equiv \sigma_v^2(z)$ *is a continuous function, and* $\inf_{\{z \in S\}} f_Z(z) \geq \delta > 0$

(iii) The kernel K *is symmetric, bounded and has compact support (i.e., for scalars* u, u', $\exists c > 0$ *such that* $K(u) = 0$ *for* $|u| \geq c$*). Assume* $\left||u|^l K(u) - |u'|^l K(u')\right| \leq C_2|u - u'|$ $\forall 0 \leq l \leq 3$.

Assumption 8. $\mathbb{E}\left[x_{0,i}x_{0,i}'\right]$ *is nonsingular and* Ω_0 *is a finite positive definite matrix.*

Theorem 4. *(consistency result with multiple instruments) If Assumption 5, 6, 7 and 8 are satisfied, and if: (bandwidths conditions for the kernel estimator) the bandwidths* $\max_{1 \leq s \leq q}\{h_s\} \to 0$, $h_1 h_2 ... h_q \overset{n \to \infty}{\to} 0$,*and* $\ln(n)/(nh_1h_2...h_q) \overset{n \to \infty}{\to} 0$ *then, for* $\hat{\beta}$ *defined as in (37):*

$$\hat{\beta} - \beta_0 = o_{\mathbb{P}}(1).$$

The theorem extends to the case where multiple instruments are considered, and thus requires each bandwidth to go to zero as well as the ratio of the logarithm of the sample size to the product of the sample size and the bandwidths. The proof follows the same steps as the one for Theorem 1.

Proof. The proof is similar to the one for Theorem 1, but the conditions on the bandwidths are extended to the case where there are multiple instruments. Let $z_i \equiv (z_{1,i}, z_{2,i}, ..., z_{q,i})$ and write the estimator from (37) as

$$\hat{\beta} - \beta_0 = \left(\hat{X}'X\right)^{-1}\hat{X}'U = \left[\frac{\left(\hat{X} - X_0\right)'X + X_0'X}{n}\right]^{-1}\left[\frac{X_0'U + \left(\hat{X} - X_0\right)'U}{n}\right].$$

Under Assumption 7 and the assumption that $\max_{1 \leq s \leq q}\{h_s\} \to 0, h_1 h_2 ... h_q \overset{n \to \infty}{\to} 0$, and $\ln(n)/(nh_1h_2...h_q) \overset{n \to \infty}{\to} 0$, I have[12]

$$\sup_{\{z_j \in \mathcal{Z}\}} |\hat{g}(z_j) - g_0(z_j)| = o(1) \quad a.s.$$

The rest of the proof can then start from expression (16) all the way to the end result. □

Theorem 5. *Under Assumptions 5, 6, 7, 8, if* $h \overset{n \to \infty}{\to} 0$ *and* $\ln(n)/nh \overset{n \to \infty}{\to} 0$ *then, with* $\hat{\beta}$ *defined in (37):*

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \overset{d}{\to} \mathcal{N}(0, \Sigma).$$

where $\Sigma \equiv \left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1}\Omega_0\left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1}$. *If, in addition,* $\mathbb{V}\left[u_i|z_i, x_{2,i}\right] = \sigma^2 < \infty$ $\forall i = 1, ..., n$, *then:*

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \overset{d}{\to} \mathcal{N}\left(0, \sigma^2\left(\mathbb{E}\left[x_{0,i}x_{0,i}'\right]\right)^{-1}\right).$$

Proof. As for the consistency proof, the argument used for the asymptotic normality follows the lines of the proof of Theorem 2, so I refer the reader to that proof. □

[12] See Li & Racine (2007), Theorem 2.6 in chapter 2.

7.6 Additional simulation material

7.6.1 Methods description

This section documents the estimators used in section 4. Let $X \equiv \begin{pmatrix} x_1 & w_{1,1} & w_{1,2} \\ x_2 & w_{2,1} & w_{2,2} \\ \vdots & \vdots & \vdots \\ x_n & w_{n,1} & w_{n,2} \end{pmatrix}$, $y \equiv \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$, $Z \equiv \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$. Define the projection matrix $P_Z \equiv Z\left(Z'Z\right)^{-1}Z'$ and $M_Z \equiv I_n - P_Z$. Then the OLS and 2SLS estimators are defined as:

$$\hat{\beta}_{OLS} \equiv \left(X'X\right)^{-1}X'y$$

$$\hat{\beta}_{2SLS} \equiv \left(X'P_ZX\right)^{-1}X'P_Zy$$

The estimator of Antoine & Lavergne (2014) is based on the first conditional moment restriction from Assumption 1 and defined as:

$$\hat{\beta}_{WMDF} \equiv \left(X'\left(\tilde{K} - \tilde{\lambda}I_n\right)X\right)^{-1}X'\left(\tilde{K} - \tilde{\lambda}I_n\right)Y,$$

where $\tilde{K}$ is a $n \times n$ matrix with diagonal elements of zero and off-diagonal elements $K_{ij} \equiv K\left(z_i - z_j\right)$ for some density $K()$. As recommended by the authors, I use the normal density. The parameter $\tilde{\lambda}$ is defined as:

$$\tilde{\lambda} \equiv \left[\lambda_{WMD} - \left(1 - \lambda_{WMD}\right)/n\right] / \left[1 - \left(1 - \lambda_{WMD}\right)/n\right],$$

where λ_{WMD} is the smallest eigenvalue of $\left(X^{*\prime}X^*\right)^{-1}\left(X^{*\prime}\tilde{K}X^*\right)$ with $X^* \equiv \begin{pmatrix} Y & X \end{pmatrix}$.

The estimators of Donald & Newey (2001) and Lee & Shin (2018) are the 2SLS estimators using the combination of instruments that minimize some feasible approximate MSE criterion functions.

In the case of Lee & Shin (2018):

$$\hat{k} \equiv \underset{k \in K}{\operatorname{argmin}}\{S_\lambda(k)\} = \underset{k \in K}{\operatorname{argmin}}\left\{\tilde{\sigma}_{\lambda\varepsilon}\frac{k^2}{n} + \tilde{\sigma}_\varepsilon^2\left[\tilde{\lambda}'\tilde{H}^{-1}\tilde{e}_f^k\tilde{H}^{-1}\tilde{\lambda} - \tilde{\lambda}'\tilde{H}^{-1}\tilde{\xi}_f^k\tilde{H}^{-1}\tilde{\xi}_f^k\tilde{H}^{-1}\tilde{\lambda}\right]\right\}$$

and

$$\hat{\beta}_{CSA} \equiv \frac{X'P^{\hat{k}}y}{X'P^{\hat{k}}X}$$

where

$$P^{\hat{k}} \equiv \frac{1}{M}\sum_{m=1}^{M}P_m^{\hat{k}} = \frac{1}{M}\sum_{m=1}^{M}Z_m\left(Z_m'Z_m\right)^{-1}Z_m'$$

is the average of projection matrices among all the projection matrices made with $\hat{k}$ instruments. So Z_m is the matrix that stacks the m^{th} combination of $\hat{k}$ instruments.There are $M = C_K^k = \frac{K!}{k!(K-k)!}$ such matrices. The reader is referred to Lee & Shin (2018) section 3.2 for the definition of the variables with a tilde in $S_\lambda(k)$, which are estimates. In this simulation study, a single endogenous variable is considered, so the minimization problem does not depend on λ. It is equivalent to set $\tilde{\lambda} = 1$.

In the case of Donald & Newey (2001), $\hat{k}$ is

$$\hat{k} \equiv \underset{k \in K}{\operatorname{argmin}} \{\tilde{\lambda}' S_{\lambda}(k) \tilde{\lambda}\} = \underset{k \in K}{\operatorname{argmin}} \left\{ \tilde{\lambda}' \left(\tilde{H}^{-1} \left[\tilde{\sigma}_{uv}^2 \frac{k^2}{n} + \tilde{\sigma}_{\varepsilon}^2 \frac{\tilde{f}'(I_n - P_k)\tilde{f}}{n} \right] \tilde{H}^{-1} \right) \tilde{\lambda} \right\}$$

and

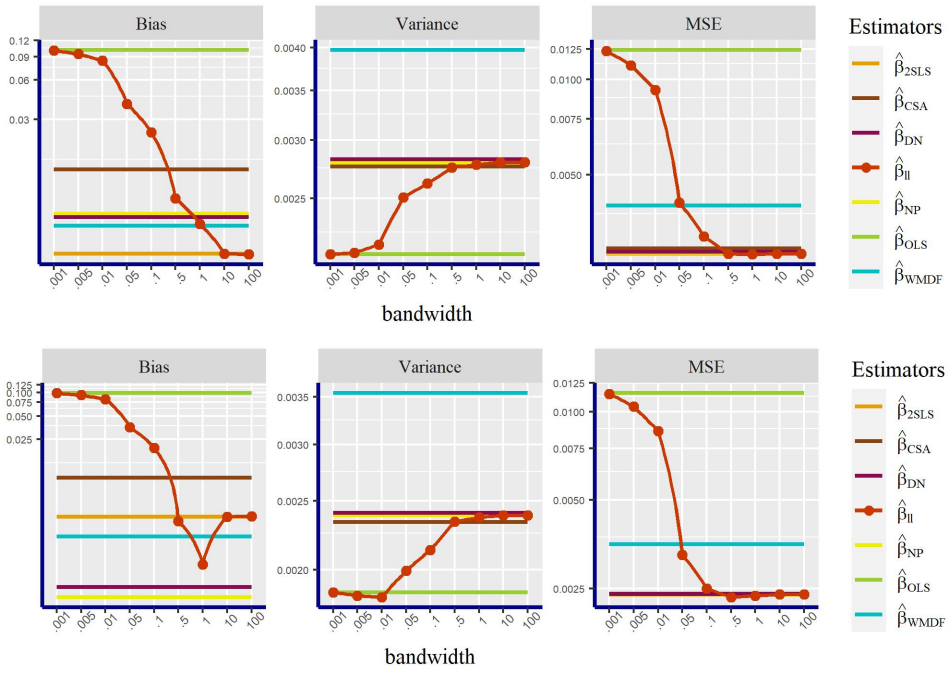
$$\hat{\beta}_{DN} \equiv \frac{X' P_{\hat{k}} y}{X' P_{\hat{k}} X}$$

where P_k is the projection matrix using the combination k instruments. Again, for this simulation study, $\tilde{\lambda}$ can be set to 1. All the quantites with a tilde are estimated quantities, and the reader is referred to Donald & Newey (2001) for more details. Both Lee & Shin (2018) and Donald & Newey (2001)'s estimations are done using the code provided by Younki Shin's Github.

7.6.2 Additional graphs

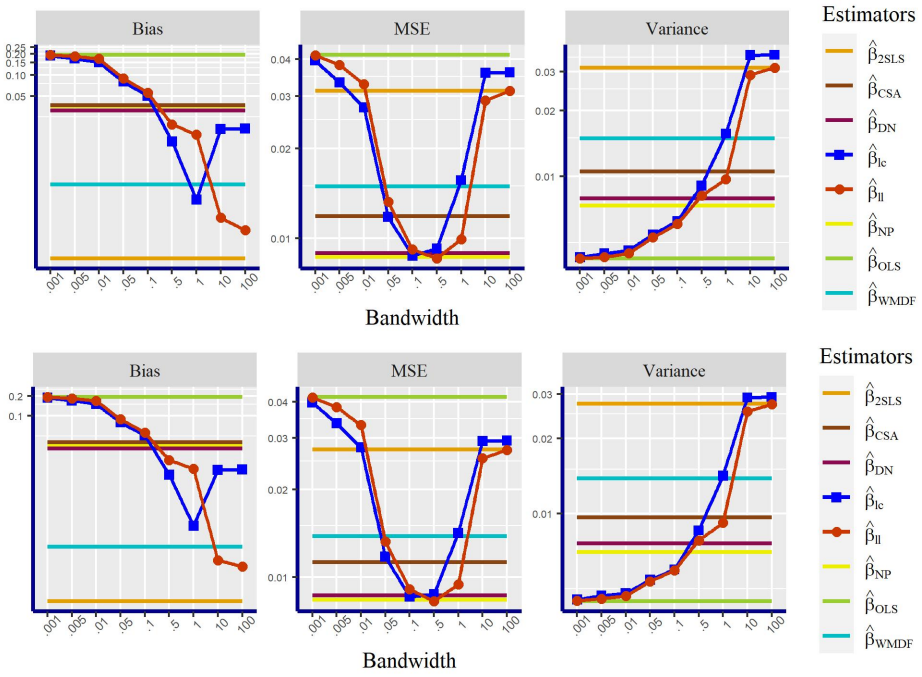
The following graphs complement the ones shown in section 4.

Figure 23: Design 1, moderate endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



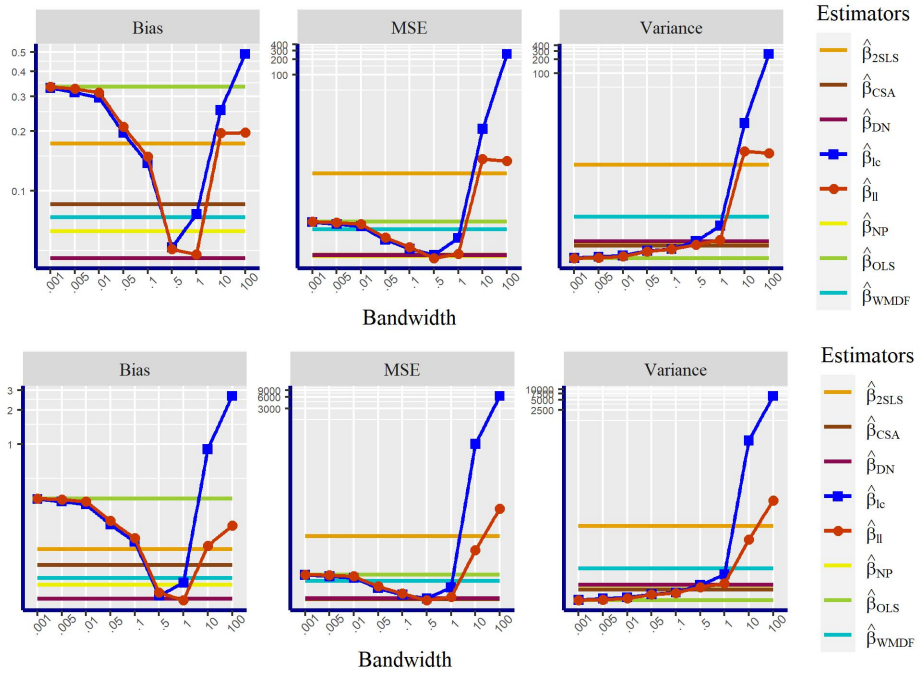
Note: The graphs report results of 1,000 rounds of simulation. The estimator using the local constant kernel estimator is omitted for scaling purposes. $n = 100$ (top), $n = 1000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.5$ and $x_i = 2z_i + v_i$.

Figure 24: Design 2, moderate endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



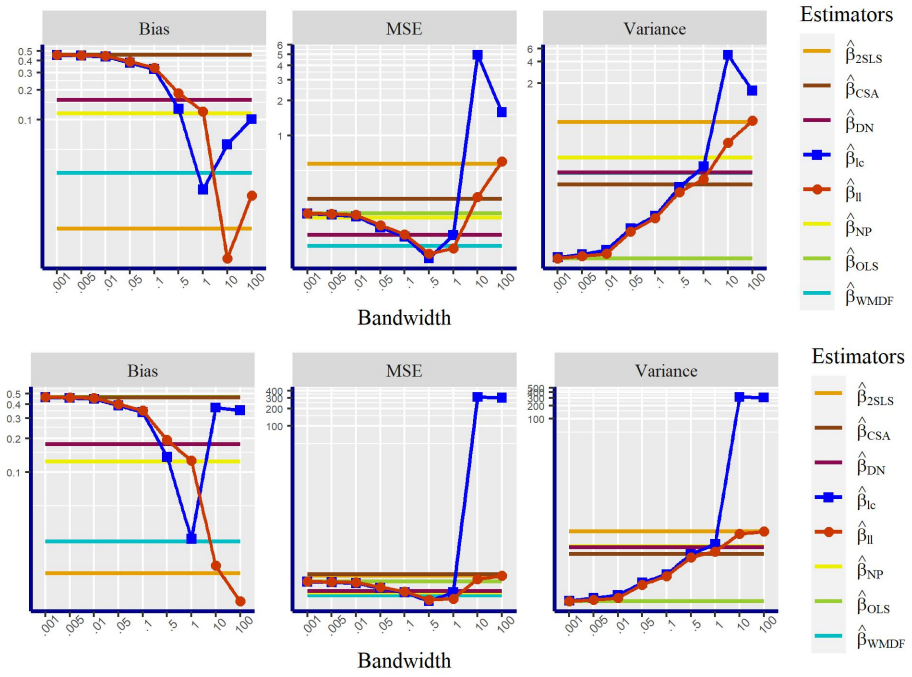
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.5$ and $x_i = \ln(|z_i|) + v_i$.

Figure 25: Design 3, moderate endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.5$ and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.

Figure 26: Design 4, moderate endogeneity, $n = 100$ (top), $n = 1000$ (bottom)



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$ (top), $n = 1000$ (bottom), $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.5$ and $x_i = \phi(z_i) + v_i$.

7.6.3 Additional tables

The following tables show the relative bias, variance and MSE of the different estimators compared to the 2SLS estimator.

Table 2: Relative bias, Design 1

n	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.1	183.909	1	6.679	4.568
100	0.5	282.499	1	9.061	3.288
100	0.9	227.006	1	5.577	4.091
1000	0.1	68.438	1	1.238	0.813
1000	0.5	610.058	1	1.317	1.739
1000	0.9	15663.783	1	57.972	32.531

Note: The graphs report results of 1,000 rounds of simulation. The estimator using the local constant kernel estimator is omitted for scaling purposes and $x_i = 2z_i + v_i$.

Table 3: Relative variance, Design 1

n	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.1	0.796	1	0.993	1.284
100	0.5	0.764	1	0.997	1.273
100	0.9	0.671	1	1.001	1.280
1000	0.1	0.801	1	0.999	1.203
1000	0.5	0.759	1	1.000	1.182
1000	0.9	0.675	1	1.000	1.220

Note: The graphs report results of 1,000 rounds of simulation. The estimator using the local constant kernel estimator is omitted for scaling purposes and $x_i = 2z_i + v_i$.

Table 4: Relative MSE, Design 1

n	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.1	0.953	1	0.993	1.284
100	0.5	4.689	1	1.001	1.273
100	0.9	13.452	1	1.008	1.284
1000	0.1	2.442	1	0.999	1.202
1000	0.5	40.021	1	1.000	1.182
1000	0.9	135.582	1	1.002	1.220

Note: The graphs report results of 1,000 rounds of simulation. The estimator using the local constant kernel estimator is omitted for scaling purposes and $x_i = 2z_i + v_i$.

Table 5: Relative bias, Design 2

n	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.1	10.326	1	1.212	1.375
100	0.5	63.486	1	11.251	1.626
100	0.9	89.955	1	16.308	2.982
1000	0.1	28.318	1	1.803	0.556
1000	0.5	702.885	1	41.375	4.440
1000	0.9	658.758	1	40.917	1.090

Note: The graphs report results of 1,000 rounds of simulation and $x_i = \ln(|z_i|) + v_i$.

Table 6: Relative variance, Design 2

n	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.1	0.154	1	0.247	0.482
100	0.5	0.151	1	0.237	0.467
100	0.9	0.163	1	0.247	0.492
1000	0.1	0.156	1	0.275	0.489
1000	0.5	0.147	1	0.255	0.479
1000	0.9	0.178	1	0.297	0.474

Note: The graphs report results of 1,000 rounds of simulation and $x_i = \ln(|z_i|) + v_i$.

Table 7: Relative MSE, Design 2

n	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.1	0.201	1	0.247	0.482
100	0.5	1.456	1	0.277	0.468
100	0.9	4.402	1	0.386	0.497
1000	0.1	0.748	1	0.277	0.489
1000	0.5	14.192	1	0.303	0.479
1000	0.9	46.059	1	0.474	0.474

Note: The graphs report results of 1,000 rounds of simulation and $x_i = \ln(|z_i|) + v_i$.

Table 8: Relative bias, Design 3

n	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.1	1.495	1	0.284	0.330
100	0.5	1.553	1	0.271	0.127
100	0.9	1.840	1	0.336	0.150
1000	0.1	1.239	1	0.074	0.016
1000	0.5	1.277	1	0.070	0.038
1000	0.9	1.410	1	0.074	0.034

Note: The graphs report results of 1,000 rounds of simulation and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.

Table 9: Relative variance, Design 3

n	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.1	0.004	1	0.013	0.262
100	0.5	0.004	1	0.008	0.135
100	0.9	0.014	1	0.016	0.450
1000	0.1	0.000	1	0.003	0.014
1000	0.5	0.000	1	0.001	0.004
1000	0.9	0.001	1	0.003	0.011

Note: The graphs report results of 1,000 rounds of simulation and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.

Table 10: Relative MSE, Design 3

n	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.1	0.006	1	0.013	0.262
100	0.5	0.043	1	0.009	0.133
100	0.9	0.270	1	0.024	0.418
1000	0.1	0.004	1	0.003	0.014
1000	0.5	0.022	1	0.001	0.004
1000	0.9	0.241	1	0.003	0.010

Note: The graphs report results of 1,000 rounds of simulation and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.

Table 11: Relative bias, Design 4

n	σ_u^2	σ_v^2	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.25	0.25	0.1	58.388	1	5.733	2.965
100	0.5	1	0.5	171.753	1	38.873	27.286
100	1	0.5	0.5	163.285	1	26.530	12.044
100	1	1	0.1	10.405	1	0.761	4.355
100	1	1	0.5	157.673	1	35.119	24.251
100	1	1	0.9	39.100	1	7.223	7.873
1000	0.25	0.25	0.1	238.702	1	6.114	0.205
1000	0.5	1	0.5	700.878	1	38.139	8.953
1000	1	0.5	0.5	302.182	1	11.511	0.980
1000	1	1	0.1	17.253	1	0.307	1.124
1000	1	1	0.5	186.806	1	10.608	2.329
1000	1	1	0.9	127.449	1	8.014	0.705

Note: The graphs report results of 1,000 rounds of simulation and $x_i = \phi(z_i) + v_i$.

Table 12: Relative variance, Design 4

n	σ_u^2	σ_v^2	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.25	0.25	0.1	0.188	1	0.819	0.926
100	0.5	1	0.5	0.014	1	0.622	0.779
100	1	0.5	0.5	0.053	1	0.675	1.084
100	1	1	0.1	0.053	1	0.556	1.137
100	1	1	0.5	0.032	1	0.680	1.008
100	1	1	0.9	0.004	1	0.469	0.628
1000	0.25	0.25	0.1	0.187	1	0.830	0.859
1000	0.5	1	0.5	0.038	1	0.827	0.899
1000	1	0.5	0.5	0.079	1	0.826	0.848
1000	1	1	0.1	0.066	1	0.812	0.863
1000	1	1	0.5	0.055	1	0.822	0.878
1000	1	1	0.9	0.016	1	0.825	0.863

Note: The graphs report results of 1,000 rounds of simulation and $x_i = \phi(z_i) + v_i$.

Table 13: Relative MSE, Design 4

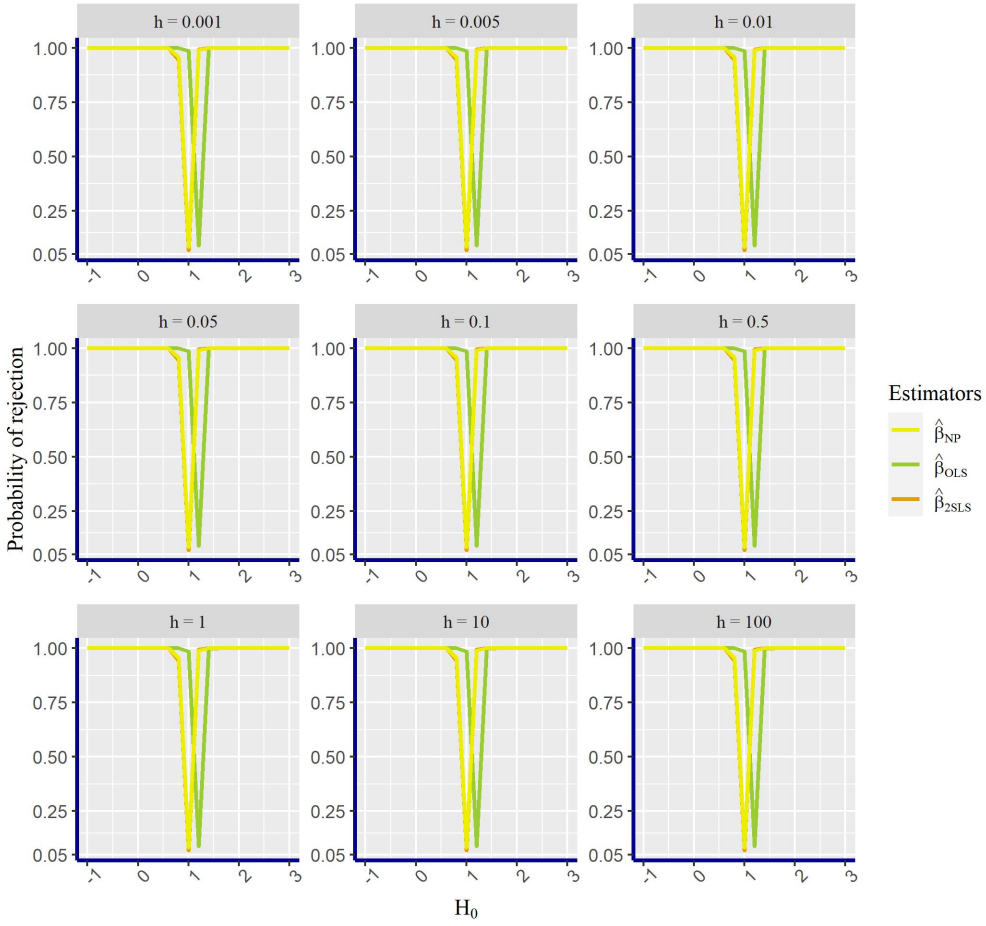
n	σ_u^2	σ_v^2	σ_{uv}	$\hat{\beta}_{OLS}$	$\hat{\beta}_{2SLS}$	$\hat{\beta}_{NP}$	$\hat{\beta}_{WMDF}$
100	0.25	0.25	0.1	2.795	1	0.844	0.932
100	0.5	1	0.5	1.167	1	0.681	0.808
100	1	0.5	0.5	3.659	1	0.770	1.104
100	1	1	0.1	0.092	1	0.556	1.143
100	1	1	0.5	0.951	1	0.726	1.029
100	1	1	0.9	1.094	1	0.506	0.672
1000	0.25	0.25	0.1	25.304	1	0.846	0.858
1000	0.5	1	0.5	31.470	1	0.920	0.904
1000	1	0.5	0.5	59.608	1	0.911	0.849
1000	1	1	0.1	0.657	1	0.811	0.863
1000	1	1	0.5	15.897	1	0.873	0.880
1000	1	1	0.9	53.686	1	1.035	0.862

Note: The graphs report results of 1,000 rounds of simulation and $x_i = \phi(z_i) + v_i$.

7.6.4 Additional power curves

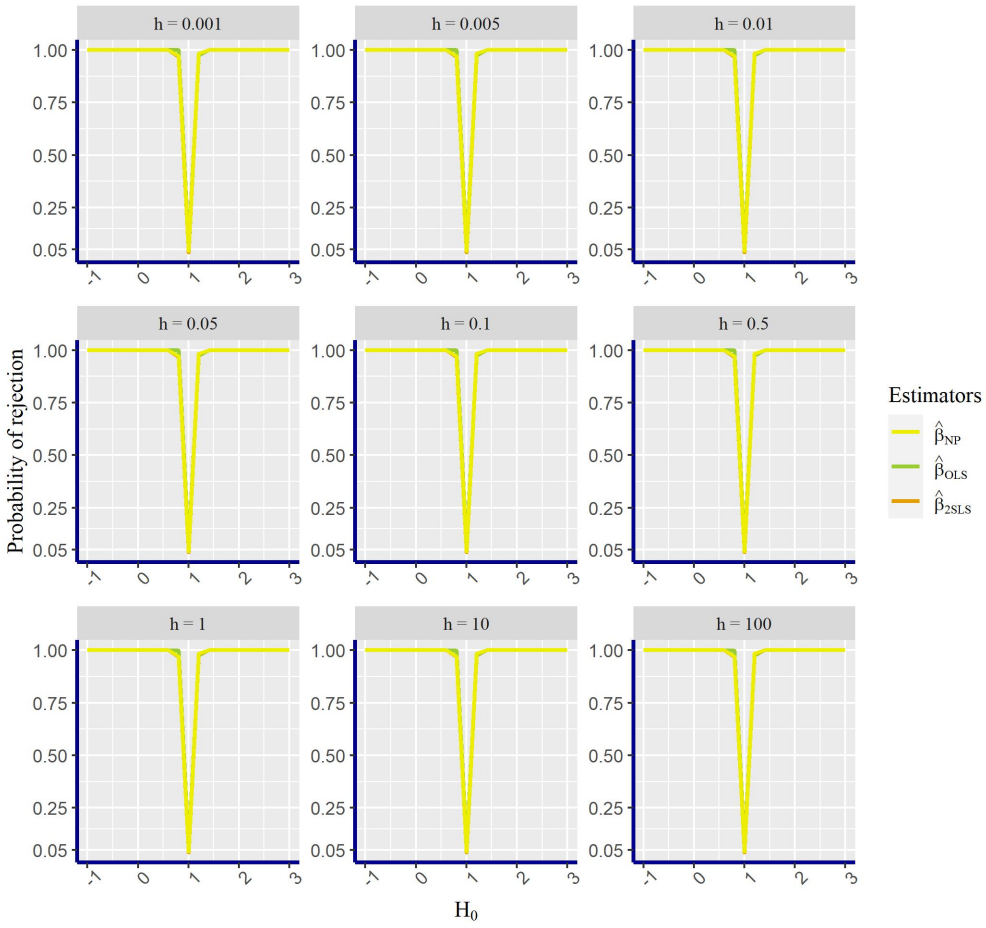
The following graphs display power curves as the bandiwdth changes. Hence, only the curves for $\hat{\beta}_{ll}$ adn $\hat{\beta}_{lc}$ change.

Figure 27: Design 1, high endogeneity, $n = 100$



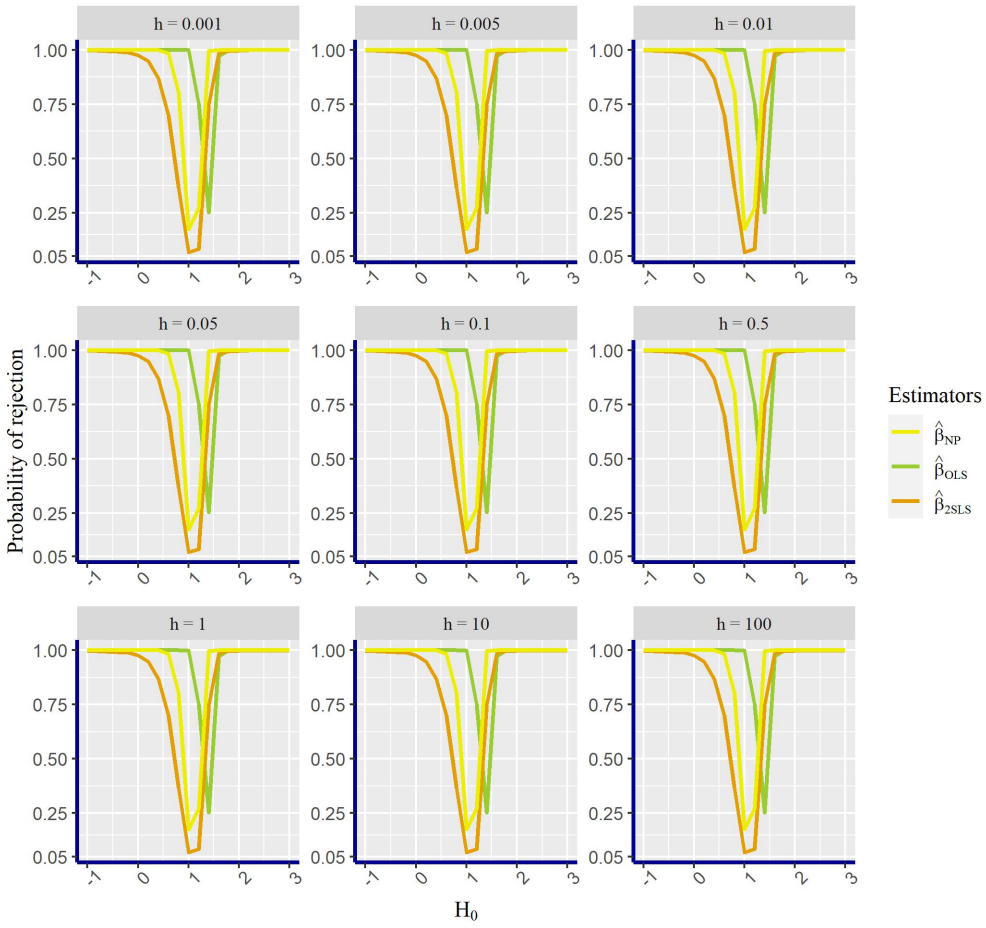
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$, $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = 2z_i + v_i$.

Figure 28: Design 1, low endogeneity, $n = 100$



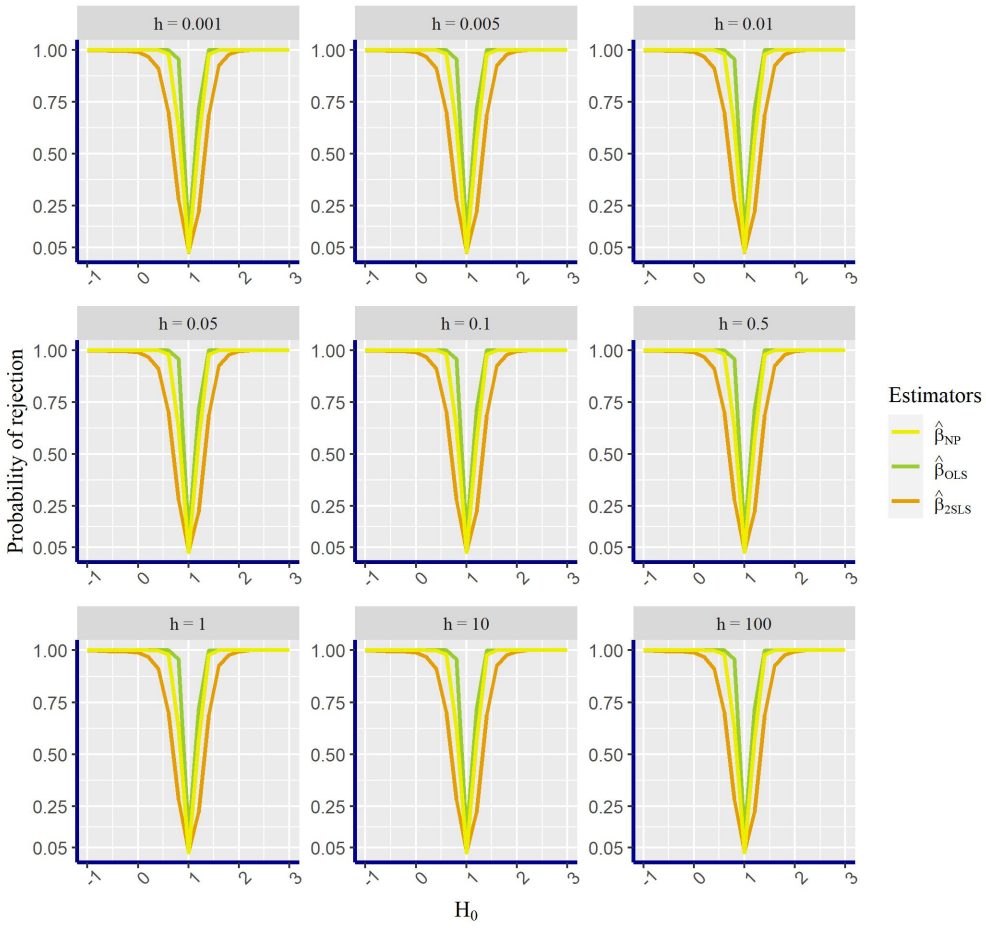
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$, $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = 2z_i + v_i$.

Figure 29: Design 2, high endogeneity, $n = 100$



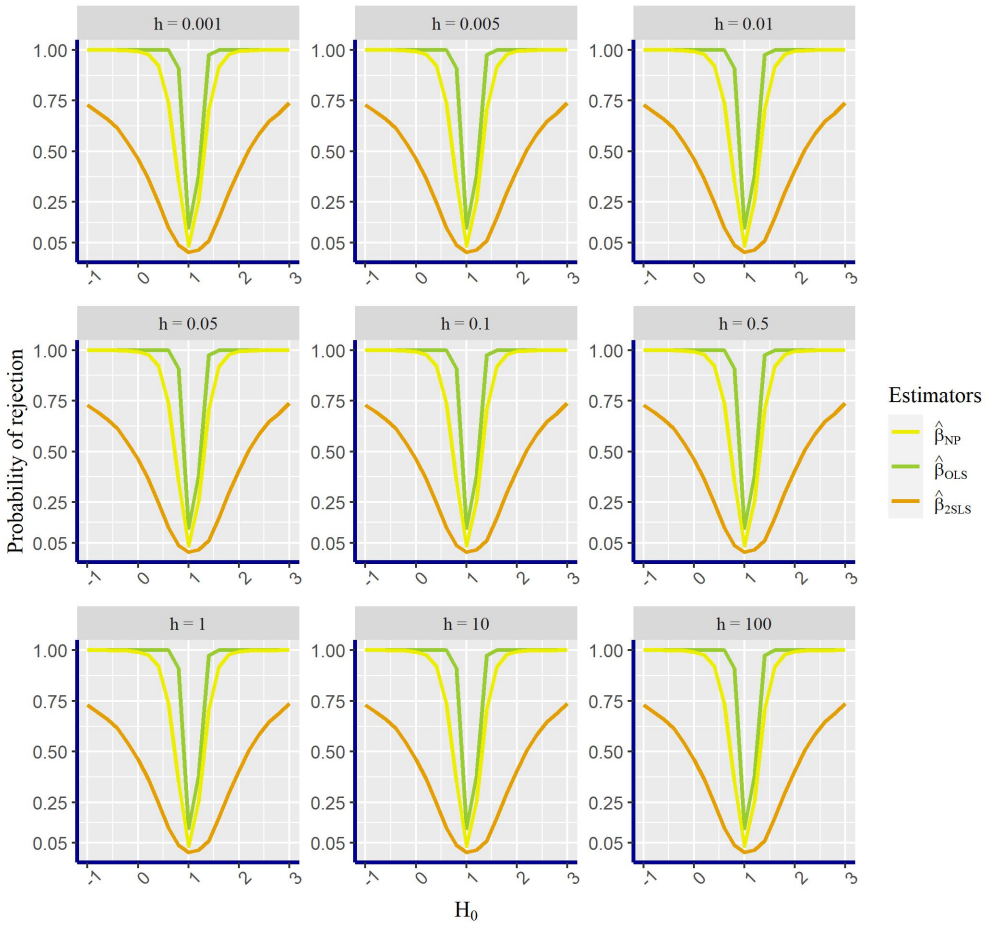
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$, $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \ln(|z_i|) + v_i$.

Figure 30: Design 2, low endogeneity, $n = 100$



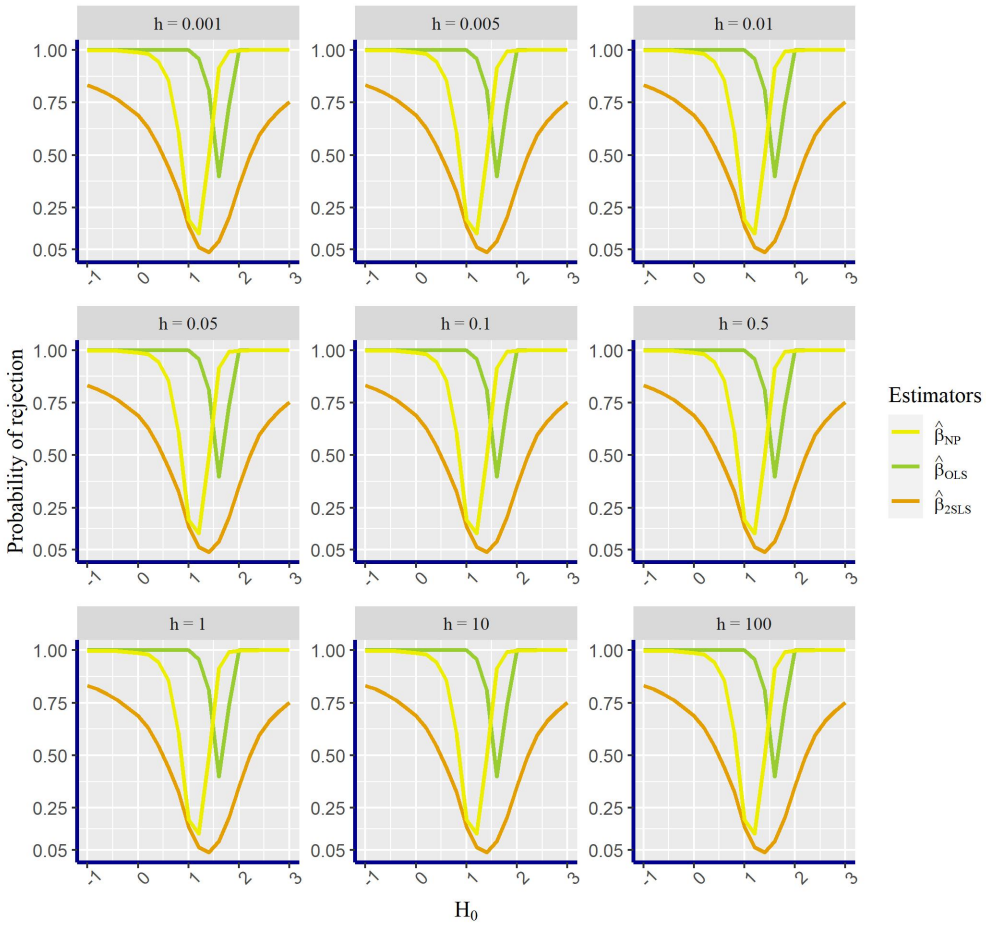
Note: The graphs report results of 1,000 rounds of simulation. $n = 100$, $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \ln(|z_i|) + v_i$.

Figure 31: Design 3, low endogeneity, $n = 100$



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$, $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.1$ and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.

Figure 32: Design 3, high endogeneity, $n = 100$



Note: The graphs report results of 1,000 rounds of simulation. $n = 100$, $\sigma_u^2 = 1$, $\sigma_v^2 = 1$, $\sigma_{uv} = 0.9$ and $x_i = \frac{1}{n^{\frac{1}{4}}}\left(3z_i - z_i^3\right) + \frac{1}{\sqrt{n}}\left(z_i^2 - 1\right) + v_i$.