Optimal Categorical Instrumental Variables*

Thomas Wiemann†

April 12, 2023

Abstract

This paper discusses estimation with a categorical instrumental variable in settings with potentially few observations per category. The proposed categorical instrumental variable estimator (CIV) leverages a regularization assumption that implies existence of a latent categorical variable with fixed finite support achieving the same first stage fit as the observed instrument. In asymptotic regimes that allow the number of observations per category to grow at arbitrary small polynomial rate with the sample size, I show that CIV is root-n asymptotically normal and achieves the same asymptotic variance as the oracle IV estimator that presumes knowledge of the optimal instrument.

*I thank Stéphane Bonhomme, Christian Hansen, Thibaut Lamadon, Jonas Lieber, Elena Manresa, Sendhil Mullainathan, Whitney Newey, Guillaume Pouliot, Vitor Possebom, Max Tabord-Meehan, and Alexander Torgovitsky for valuable comments and suggestions, along with participants at the University of Chicago Econometrics advising group. All errors are my own.

†University of Chicago, wiemann@uchicago.edu.

1 Introduction

Poor finite-sample statistical properties of two stage least squares estimators in practice motivate a large number of alternative instrumental variable estimators. Angrist and Krueger (1991), in particular, have generated large interest in the study of both weak and many instruments. Their empirical analysis of returns-to-education has been the leading example to highlight problems of conventional inference when the instruments are only weakly correlated with the endogenous variable (e.g., Bound et al., 1995). One approach to improve statistical precision is to estimate optimal instruments that are aimed at maximizing the strength of the first stage fit (Amemiya, 1974; Chamberlain, 1987; Newey, 1990). However, in the absence of functional form assumptions, optimal instruments need to be nonparametrically estimated which can introduce bias from over-fitting in the first stage.[1] In response to these challenges in optimal instrument estimation, a growing literature considers regularization assumptions on the first stage reduced form that – when leveraged appropriately – allow for second stage estimators with the same asymptotic variance as the oracle estimator that presumes knowledge of the optimal instrument. Increasingly popular in practice is the post-lasso IV estimator of Belloni et al. (2012) that assumes approximate sparsity of the first stage reduced form (see, e.g., Gilchrist and Sands, 2016; Dhar et al., 2022).[2] While effectively a smoothness assumption for the optimal instrument, sparsity can be ill-suited when the instruments are categorical. Simulations with categorical instruments in Angrist and Frandsen (2022) and in this paper highlight that even in settings with many more observations than instruments, post-lasso IV can have worse finite sample behavior than conventional two stage least squares (TSLS).

This paper proposes a new optimal instrumental variable estimator for settings with a large number of categorically encoded instruments. I consider a first-stage regularization assumption designed specifically for categorical instruments: Fixed finite support of the optimal instrument. Levering this assumption, I show that the proposed categorical instrumental

[1] Hansen et al. (2008), for example, suggest that the large number of instruments in Angrist and Krueger (1991) is the primary cause for poor statistical properties of the two stage least squares estimator.

[2] Informally, approximate sparsity presumes that a slowly increasing unknown subset of instruments suffices to approximate the optimal instrument relative to the reduced form estimation error.

variable estimator (CIV) is root-n asymptotically normal when the expected number of observations per category grows at arbitrarily small polynomial rate with the sample size. This asymptotic regime is aimed to approximate the practical settings in which the number of observations per category is small. Further, CIV achieves the same asymptotic variance as the infeasible oracle two stage least squares estimator that presumes knowledge of the optimal instrument, and is semiparametrically efficient when the second stage error is homosekdastic.

The key idea of the categorical instrumental variable estimator is to leverage a latent categorical variable with fewer categories that achieves the same population-level fit in the first stage. Under the assumption that the latent categories are well-separated, this structure allows for the application of exponential inequalities to bound the probability of incorrectly mapping observed categories to the latent categories. In particular, I draw from the literature on estimation of fixed effects in panel data settings under asymptotic regimes where both the number of individuals and the number of time periods grows. The regularization assumption on the optimal instrument that allows for the application of these misclassification bounds in this paper is closely related to the finite support assumption on fixed effects in the panel literature (e.g., Hahn and Moon, 2010; Bonhomme and Manresa, 2015; Bester and Hansen, 2016; Su et al., 2016). Hahn and Moon (2010) show that finite support assumptions substantially decrease the incidental parameter problem associated with increasingly many fixed effects. CIV adapts the K-Means based fixed effects estimator of Bonhomme and Manresa (2015) for estimation of the optimal categorical instrumental variable. To the best of my knowledge, this appears to be the first application of a finite support assumption to the estimation of optimal instruments.

The focus on categorical instrumental variables is motivated by the many examples in empirical economics, including leading examples in the many and weak instruments literature, featuring categorical instruments. Angrist and Krueger (1991), for example, consider interactions between quarter of birth indicators and year and place of birth indicators, resulting in a set of 180 binary instruments. Extensions of their design consider the fully saturated first stage of all interactions between the three sets of indicators resulting in 1530 instrumental variables, each representing a unique combination of quarter, year, and place of birth

(see, e.g., Mikusheva and Sun, 2022; Angrist and Frandsen, 2022). Another key empirical application is the popular "judge IV design" (see, e.g., Kling, 2006; Maestas et al., 2013; Aizer and Doyle Jr, 2015; Bhuller et al., 2020). In these contexts, judge identity is often used as an instrument, yet, the number of cases per judge in the sample can be small.

The advantage of the proposed estimator over alternative optimal IV estimators is that the underlying regularization assumption places restrictions on the data generating process that have straightforward economic interpretations when applied to categorical instruments. In particular, the regularization assumption presumes existence of an unobserved combination of categories that fully captures relevant variation from the instruments. In the Angrist and Krueger (1991) application, for example, the assumption of an optimal instrument with two support points captures the idea that while minimum attendance laws (determining the relation between quarter of birth and years of education) vary across states and cohorts, the only relevant information is whether a student born in a particular quarter and year was constrained or unconstrained by the schooling policy in their state.[3] Similarly, in the judge IV applications, the number of support points corresponds to the number of latent types of judges (e.g., for two support points, one may label them "strict" and "lenient" judges). These interpretations are in strong contrast to the application of approximate sparsity. In the Angrist and Krueger (1991) setting, sparsity presumes that almost all year and state combinations follow the same compulsory schooling law. Similarly, in the judge IV application, sparsity prohibits an equal proportion of, say, lenient and strict judges. The finite support regularization assumption employed in this paper does not restrict the proportion of any particular combination of categories (albeit combinations cannot be asymptotically vanishing).

Related Literature. The paper primarily draws from and contributes to two strands of literature. First, the literature on many instruments that develops estimators robust to asymptotic regimes in which the number of instruments is proportional to the sample size (e.g., Bekker,

[3]The ideal instrument in the setting of Angrist and Krueger (1991) thus is complete information on the compulsory schooling laws in place across all states and cohorts in the data. The proposed CIV estimator aims to achieve the same statistical efficiency as an estimator leveraging such extensive legislative data but without the need to work through legal texts directly. Instead, the first stage efficiently "learns" the compulsory schooling cutoffs.

1994; Angrist and Krueger, 1995; Chao and Swanson, 2005; Hansen et al., 2008; Hausman et al., 2012). Most closely related is Bekker and Van der Ploeg (2005) who provide limiting distributions of two-stage least squares (TSLS), limited information maximum likelihood (LIML), and heteroskedasticity-adjusted estimators under group asymptotics that consider replications of categorical instruments with a constant number of observations per category. In less stringent asymptotic regimes where the number of categories grows at a slower rate than the sample size, their results imply first-order equivalence of the LIML and the oracle IV estimator in the presence of heteroskedasticity when observations are equally distributed across categories and effects are constant.[4] Despite the favorable statistical properties of LIML in settings with categorical instrumental variables and homogeneous effects, its application to causal effects estimation in economics is limited by its strong reliance on constant effects in the linear IV model. Kolesár (2013) shows that under the nonparametric causal model of Imbens and Angrist (1994), the LIML estimand cannot generally be interpreted as a positively (weighted) average of causal effects. In the terminology of Blandhol et al. (2022), LIML does thus not generally admit a weakly causal interpretation. In contrast, the proposed CIV estimator falls in the class of two-step estimators of Kolesár (2013) and therefore admits a weakly causal interpretation in the presences of unobserved heterogeneity.[5]

Second, I draw from the literature on optimal instrumental variable estimators. Optimal instruments are conditional expectations that – in the absence of functional form assumptions – can be nonparametrically estimated (Amemiya, 1974; Chamberlain, 1987; Newey, 1990). Newey (1990) considers approximation of optimal instruments using polynomial sieve regression and characterizes the growth rate of series terms relative to the sample size that allows for root-n consistency. In the setting of categorical variables, the restrictions imply that the number of categories should grow slower than root-n to avoid the many instruments bias. CIV contributes to the literature on optimal IV estimation that leverages regulariza-

[4]Note also that Lemma 6.A of Donald and Newey (2001) implies that LIML using categorical instruments achieves first-order oracle equivalence when the number of categories grows below the sample rate and causal effects are constant.

[5]Albeit not pursued in this paper, CIV can also straightforwardly be combined with nonparametric residualization approaches, including the popular class of double/debiased machine learning estimators of Chernozhukov et al. (2018), which is crucial to obtain convex combinations of causal effects in models with unobserved heterogeneity and controls. See Blandhol et al. (2022).

tion assumptions on the first stage to allow for a larger number of considered instruments. In homoskedastic linear IV models, Donald and Newey (2001) propose instrument selection criteria, Chamberlain and Imbens (2004) consider regularization via a random coefficient assumption, and Okui (2011) suggests first stage estimation via Ridge regression (ℓ_2 regularization). In linear IV models with heteroskedasticity Carrasco (2012) consider ℓ_2 regularization (including Tikhonov regularization) and provide conditions for asymptotic efficiency of the resulting IV estimator in settings when the number of instruments is allowed to grow at faster rate than the sample size. Belloni et al. (2012) apply the lasso and post-lasso (ℓ_1 regularization) to estimate optimal instruments in the setting with very many instruments. The authors provide sufficient conditions for the asymptotic efficiency of the resulting IV estimator, most notably, an approximate sparsity assumption, which presumes that a slowly increasing unknown subset of instruments suffices to approximate the optimal instrument relative to the reduced form estimation error. A common theme in the regularization approaches of these previous approaches is shrinkage of the first stage coefficients to zero. In the setting of categorical instruments, this corresponds to existence of one large latent base category (i.e., the constant) and only a few small deviating latent categories. Settings in which differing latent categories are approximately proportional are not admitted in these shrinkage-to-zero approaches as observed categories cannot be arbitrarily combined. CIV complements these existing estimators by leveraging an alternative regularization assumption that admits approximately proportional latent categories via arbitrary combination of observed categories.

Outline. The remainder of the paper is organized as follows: Section 2 presents the instrument variable framework and develops the proposed estimator. Section 3 states the main theoretical result of the paper. Section 4 provides a simulation exercise to contrast the finite sample performance of CIV with competing estimators, in particular, highlighting the pitfalls of post-lasso IV estimation with categorical instruments. Section 5 revisits the returns to education analysis of Angrist and Krueger (1991). Section 6 concludes.

Notation. It is useful to clarify some notation. For a random variable X, let $\operatorname{supp} X$ denote its support and $|\operatorname{supp} X|$ the cardinality of the support. For a function $f : A \to B$, let $f(A)$

denote its image. For a set $\mathcal{X}$, the indicator function $\mathbb{1}_{\mathcal{X}}(X)$ is equal to one if $X \in \mathcal{X}$ and zero otherwise.

2 The Categorical Instrumental Variable Estimator

This section introduces the econometric framework and defines the categorical instrumental variable estimator. I begin by characterizing the law P_n of the random vector (Y_n, D_n, Z_n, U_n) associated with a single observation. Here, Y_n denotes the outcome, D_n is the endogenous variable of interest, Z_n is the instrumental variable, and U_n are all other determinants of Y_n other than (D_n, Z_n). P_n is is allowed to change with the sample size n. While (Y_n, D_n, Z_n, U_n) and all properties of its distribution are thus implicitly indexed by n, I omit this dependence for notational brevity. It is important to emphasize however that the restrictions on P_n introduced below, including all imposed bounds, are assumed to hold uniformly over n.

Assumption 1 places a restriction on the joint distribution of observables (Y, D, Z) and unobservables U. I focus on the setting with a scalar-valued D. The analysis straightforwardly extends to any fixed number of endogenous variables.

Assumption 1. $\exists \tau_0 \in \mathbb{R} : Y = D\tau_0 + U,\ E[U|Z] = 0.$

The parameter of interest is the IV coefficient τ_0 defined in Assumption 1. In the linear model with homogeneous effects considered here, the coefficient corresponds to the change in the outcome caused by a marginal change in endogenous variable of interest.

Note that mean-independence of U and Z implies the moment condition

$$E\left[(Y - D\tau_0)\left(f(Z) - E[f(Z)]\right)\right] = 0$$

for any measurable function $f : \operatorname{supp} Z \to \mathbb{R}$. If in addition $Cov(D, f(Z)) \neq 0$, a solution to the moment condition is given by

$$\tau_0 = \frac{E[(Y - E[Y])(f(Z) - E[f(Z)])]}{E[(D - E[D])(f(Z) - E[f(Z)])]}. \tag{1}$$

Suppose the econometrician observes a sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ from (Y, D, Z). Equation (1) suggests a sample analog estimator for τ_0 that replaces expectations with sample averages. This sample analog estimator has desirable properties when f is chosen to be the conditional expectation $m_0(Z) \equiv E[D|Z]$ – that is,

$$\hat{\tau}_n^* = \frac{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)\left(m_0(Z_i) - \bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(m_0(Z_i) - \bar{D}_n\right)}, \tag{2}$$

where the law of iterated expectations implies $E[m_0(Z)] = E[D]$. When U is homoskedastic, the asymptotic variance of $\hat{\tau}_n^*$ achieves the semiparametric efficiency bound for estimating τ_0: $Var(U)/Var(m_0(Z))$ (Amemiya, 1974; Chamberlain, 1987; Newey, 1990). The transformed instruments $m_0(Z)$ are thus often termed "optimal instruments." Throughout, I consider settings in which the optimal instruments are strong.

Assumption 2. *$Var(E[D|Z])$ is bounded away from zero.*

Formulating an estimator based on the moment solution (1) has the additional benefit of falling in the class of "two-step" estimators as defined by Kolesár (2013). The author shows that even if the underlying structural model is not additively separable in the structural error U as in Assumption 1, two-step estimators admit interpretation as a convex combination of causal effects under the LATE assumptions of Imbens and Angrist (1994).[6] This starkly contrast the LIML which does not generally permit a weakly causal interpretation in the LATE framework (Kolesár, 2013). Estimation based on (1) and the optimal instrument m_0 thus has both important economic and statistical benefits.

In economic applications, the conditional expectation m_0 is rarely known. The estimator $\hat{\tau}_n^*$ is thus typically infeasible in practice. A growing literature focuses on estimating the optimal instruments such that the asymptotic distribution of the resulting estimator for τ_0 achieves the same asymptotic variance as the infeasible estimator. For example, Newey (1990) considers nearest-neighbor and series regression to approximate m_0. In settings with growing numbers of instruments, Belloni et al. (2012) and Carrasco (2012) consider regularized re-

[6] In addition to stronger exogeneity assumptions, the LATE assumptions include a monotonicity assumption that prohibits simultaneous movements in-and-out of treatment for any increment of the optimal instrument.

gression estimators.

This paper is concerned with estimation of optimal instruments in settings where observed instruments Z are categorical and the number of categories grows with sample size. Assumption 3 regulates the rate at which the number of categories is allowed to grow. In particular, I allow the number of categories to grow such that the expected number of observations per category (i.e., $n \times \Pr(Z = z)$) grows at arbitrarily slow polynomial rate with the sample size.

Assumption 3. $\forall z \in \operatorname{supp} Z, \exists \lambda_z \in (0, 1]$ *such that* $\Pr(Z = z) n^{1-\lambda_z} \to a_z > 0$.

If all $\lambda_z \in (0.5, 1]$ the number of categories grows sufficiently slowly such that the optimal instruments can be estimated by simple least squares of D on the set of indicators $(\mathbb{1}_z(Z))_{z \in \operatorname{supp} Z}$. The interesting cases are thus if for some categories $\lambda_z \in (0, 0.5]$. Since λ_z can be arbitrarily close to 0, this regime can be viewed as approximating settings in which the number of observations per category is small. These settings seem of particular practical importance in economics. For example, a fully saturated first stage in Angrist and Krueger (1991) results in 140 out of 1530 categories to have 20 observations or less.

Assumption 4 is the key regularization assumption that allows for asymptotically negligible estimation error in the optimal instruments for regimes with few numbers of observations per category. It asserts that the optimal instruments m_0 have finite support of cardinality K_0, where K_0 is known.

Assumption 4. $|\operatorname{supp} E[D|Z]| = K_0$, *for known* $K_0 \in \mathbb{N}$.

Assumption 4 implies existence of a latent categorical instrument that captures all relevant information about the endogenous variable. That is, there exists a partition $(\mathcal{Z}_g^0)_{g=1}^{K_0}$ of $\operatorname{supp} Z$ such that for all values of the latent category $g \in \{1, \ldots, K_0\}$, the optimal instrument takes the same value: $m_0(z') = m_0(z), \forall z', z \in \mathcal{Z}_g^0$. When this partition is known, estimation of optimal instruments simplifies to a simple least squares estimator of D on the indicators $(\mathbb{1}_{\mathcal{Z}_g^0}(Z))_{g=1}^{K_0}$. In practice, the partition is unknown and needs to be estimated. However, the finite support assumption essentially allows for estimation of the partition at faster than root-n rate such that it does not affect the asymptotic distribution.

The application of a Assumption 4 along with the rate condition in Assumption 3 to optimal

instruments appears novel, however, finite support assumptions have grown increasingly popular in longitudinal data settings where the categories are individual identifiers. In these cases, Assumption 4 corresponds to the group-fixed effects assumption and Assumption 3 regulates the rates at which the cross-section and the time dimension grow. Hahn and Moon (2010) show that under a finite support assumption, the incidental parameter problem due to a growing number of fixed effects is substantially reduced. Bester and Hansen (2016) consider grouped fixed effects when the grouping is known. Bonhomme and Manresa (2015) and Su et al. (2016), among others, consider settings with unknown groups and parameters.

Assumptions 1-4 motivate the categorical instrumental variable estimator (CIV) for τ_0:

$$\hat{\tau}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)\left(\hat{m}_n(Z_i) - \bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_n(Z_i) - \bar{D}_n\right)^2}, \tag{3}$$

where $\hat{m}_n(Z_i)$ is an estimator for $m_0(Z_i)$ defined by

$$\hat{m}_n = \underset{\substack{m\colon \operatorname{supp} Z \to \mathcal{M} \\ |m(\operatorname{supp} Z)| = K_0}}{\arg\min} \sum_{i=1}^{n} \left(D_i - m(Z_i)\right)^2. \tag{4}$$

In particular, the estimator simultaneously estimates the partition of $\operatorname{supp} Z$ and the associated values of the optimal instruments.

It is important to enforce the finite support assumption of the optimal instrument in estimation by restricting the support of the image of m in (4). For this purpose, I apply a K_0-Means based estimator to estimate $\hat{m}_n$. This follows the approach of Bonhomme and Manresa (2015) who apply the procedure to estimate individual and time fixed effects in longitudinal data setting. Appendix B provides implementation details along with pseudo-code for the considered algorithms. In the asymptotic analysis, I follow the conventional approach and abstract from optimization error.

3 Asymptotic Analysis

I now state additional assumptions leveraged in the asymptotic analysis. Assumption 5 places moment restrictions on the unobservable U and the first stage conditional expectation function residual $V \equiv D - m_0(Z)$. Assumption 6 regulates the tails of the first stage residual to be decaying at some polynomial rate. Analogously to Bonhomme and Manresa (2015), Assumption 6 along with a dependence restriction (see Assumption 10 below) allows for application of exponential inequalities that are key to bound the probability of misclassifying values of the instrument Z in estimation of the optimal instruments.

Assumption 5. $\exists L < \infty$ *such that* $E[U^4] \leq L$ *and* $E[V^4] \leq L$.

Assumption 6. $\exists b_1, b_2 : \Pr(|V| > v) \leq \exp\left\{1 - \left(\frac{v}{b_1}\right)^{b_2}\right\}, \forall v > 0.$

Assumption 7 and 8 ensure that the optimal instrument themselves are non-vanishing and are not asymptotically equivalent, respectively.

Assumption 7. $\Pr(E[D|Z] = d_z)$ *is bounded away from zero for all* $d_z \in \operatorname{supp} E[D|Z]$.

Assumption 8. $\exists c > 0$ *such that,* $(d_z - \tilde{d}_z)^2 \geq c, \forall d_z \neq \tilde{d}_z \in \operatorname{supp} E[D|Z]$.

Finally, I assume that the set of potential values of the optimal instrument $\mathcal{M}$ in (4) is compact, and that the econometrician observes independent samples from (Y, D, Z).

Assumption 9. $\operatorname{supp} E[D|Z] \subset \mathcal{M}$ *and* $\mathcal{M} \subset \mathbb{R}$ *is compact.*

Assumption 10. *The data is an i.i.d. sample* $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ *from* (Y, D, Z).

Theorem 1 states the main theoretical result of the paper. In particular, it shows that assumptions 1-10 are sufficient conditions for the CIV estimator $\hat{\tau}_n$ to achieve the same asymptotic distribution as the infeasible estimator $\hat{\tau}_n^*$ that presumes knowledge of the optimal instruments. Further, in homoskedastic settings, the estimator achieves the semiparametric efficiency bound.[7]

[7]Note that in the heteroskedastic setting, weighting observations proportional to their variance can improve the asymptotic variance. Since this approach follows standard generalized least squares arguments, I omit further discussion here.

Theorem 1. *Let assumptions 1-10 hold. Then, as* $n \to \infty$*,*

$$\sqrt{n}\left(\hat{\tau}_n - \tau_0\right)/\sigma \xrightarrow{d} N\left(0,1\right),$$

where $\sigma = \sqrt{Var(m_0(Z)U)}/Var(m_0(Z))$*. If in addition,* U *is homoskedastic, then* $\hat{\tau}_n$ *is semiparametrically efficient for estimating* τ_0*.*

Proof. See Appendix A. □

Remark 1. *The proof of Theorem 1 proceeds in three steps: First, I characterize the convergence rate of* $\hat{m}_n$*. Second, I prove that* $\hat{\tau}_n$ *converges to the infeasible IV estimator* $\tilde{\tau}_n$ *that presumes knowledge of the* $(\mathcal{Z}_g^0)_{g=1}^{K_0}$ *at arbitrary polynomial rate. Finally, I characterize the distribution of* $\tilde{\tau}_n$*. The first step heavily leverages the arguments of Bonhomme and Manresa (2015) with adjustments to accommodate random numbers of observations per category. See Appendix A for details.*

The result of Theorem 1 continues to hold when σ is replaced by a consistent estimator such as

$$\hat{\sigma}_n \equiv \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{m}_n(Z_i)^2\left(Y_i - D_i\hat{\tau}_n\right)^2} \Bigg/ \left(\frac{1}{n}\sum_{i=1}^{n}\hat{m}_n(Z_i)^2\right).$$

4 Monte Carlo Simulation

This section discusses a Monte Carlo simulation exercise to illustrate finite sample behavior of the proposed CIV estimator and highlight key challenges of post-lasso IV estimators for estimation with categorical instrumental variables.

The data generating process considered here adapts the design of Bekker and Van der Ploeg (2005). The second stage is given by

$$Y_i = D_i\tau_0 + U_i,$$

where D_i is a scalar-valued endogenous variable. The parameter of interest τ_0 is varied across

simulations. The first stage is given by

$$D_i = m_0(Z_i) + V_i,$$

where Z_i is a categorical instrumental variable taking values in $\mathcal{Z} = \{1, \ldots, 50\}$ and V_i is the first stage error satisfying $E[V_i|Z_i] = 0$. For each observation i, the first and second stage error draws from a bivariate normal where their covariance depends on the value of the observed instrument:

$$Cov(U_i, V_i|Z_i = z) = \begin{bmatrix} \sigma_U^2(z) & \frac{1}{2}\sigma_U(z)\sigma_V(z) \\ \frac{1}{2}\sigma_U(z)\sigma_V(z) & \sigma_V^2(z) \end{bmatrix}$$

where for each $z \in \mathcal{Z}$, the parameters $\sigma_U(z)$ and $\sigma_V(z)$ are independent draws from a uniform $U(\frac{1}{2}, \frac{3}{2})$.

To allow for easy assessment of estimator performance across settings with different numbers of observations per observed category, I consider a balanced design where each possible value of the instrument is associated with the same number of observations n_z in the sample. The optimal instrument m_0 is constructed by first partitioning $\mathcal{Z}$ into K_0 approximately equal subsets and then assigning evenly-spaced values in the interval $[-\frac{p}{2}, \frac{p}{2}]$ to each subset.[8]

Figure 1 plots power curves for the hypothesis test $H_0 : \tau_0 = 0$ at significance level $\alpha = 0.05$ associated with five estimators for a setting where the optimal instrument has two support points: $-\frac{1}{2}$ and $\frac{1}{2}$.[9] The considered estimators are CIV, the infeasible oracle IV estimator $\hat{\tau}_n^*$ that presumes knowledge of the optimal instrument m_0, the post-lasso IV estimator (rlasso-IV) proposed by Belloni et al. (2012), and the feasible TSLS estimator that uses the observed instruments. Standard errors used for testing are heteroskedasticity robust.

Panels (a)-(d) of Figure 1 correspond to samples with 30, 50, 100, and 150 observations per category, respectively. Several results are worth highlighting. First, for all considered

[8]For example, for $K_0 = 2$ and $p = 1$, $m_0(z) = -0.5$ for $z \in \{1, \ldots, 25\}$ and $m_0(z) = 0.5$ for $z \in \{26, \ldots, 50\}$.

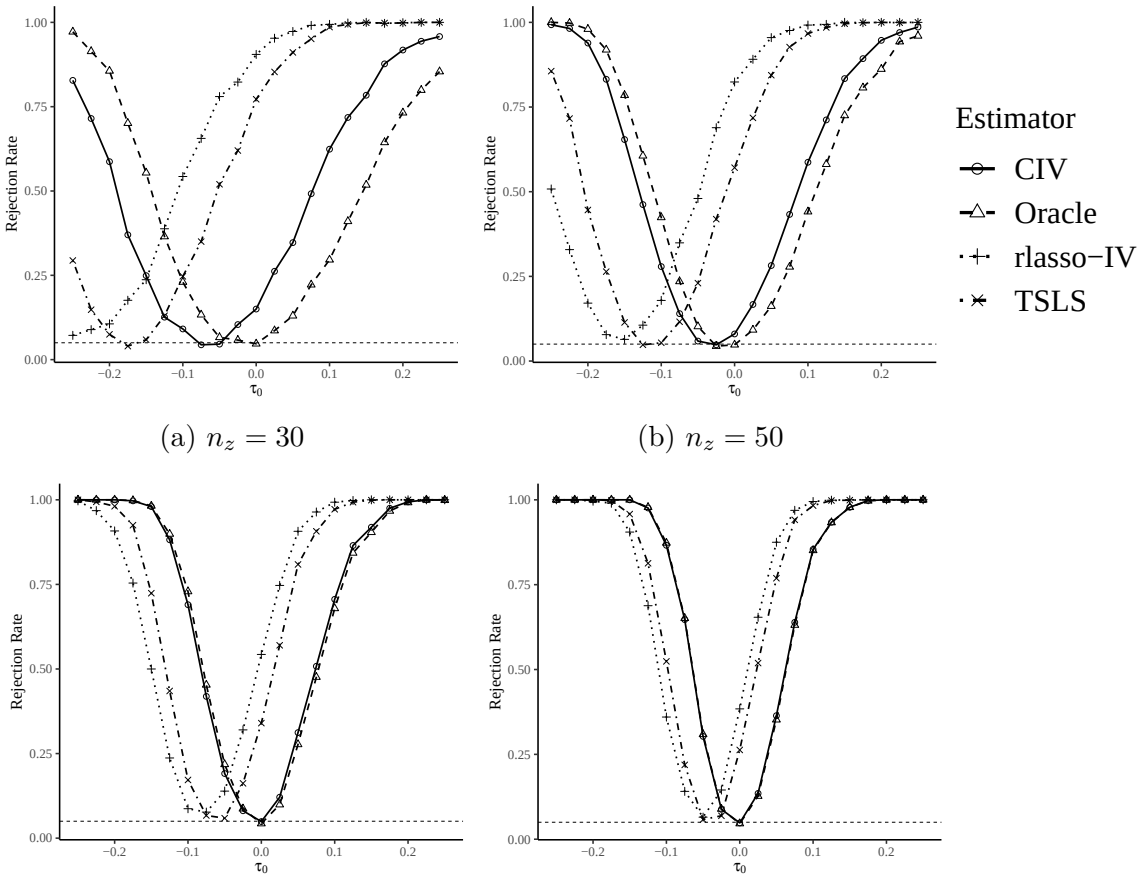
[9]Figure 5 in the Appendix revisits the simulation exercise with four support point evenly-spaced in the interval $[-1, 1]$. The qualitative results are unaffected.

category sizes, the post-lasso IV estimator of Belloni et al. (2012) has severely distorted size. With as many as 150 observations per category, the false-rejection rate (i.e., at $\tau_0 = 0$) associated with post-lasso IV is 35% rather than the desired 5%. Perhaps surprisingly, TSLS obtains better (albeit still substantially distorted) false rejection rates: At 150 observations per category, the false-rejection rate of TSLS is approximately 25%. This illustrates that ill-suited regularization in the first stage can result in worse estimator behavior in practice than no regularization. Second, in strong contrast to the post-lasso IV and TSLS, the proposed CIV estimator is close in size to the oracle estimator with 50 observations per category (in comparison, post-lasso IV and TSLS have size of approximately 0.75 and 0.55, respectively). At 100 observations per category, CIV is near indistinguishable from the oracle estimator. The estimator thus succeeds in leveraging the underlying low-dimensional structure of the DGP.

Figure 2 plots power curves for the same DGP for alternative shrinkage-based estimators. In addition to the CIV estimator included as a reference point, the figure plots IV estimators using lasso and ridge estimators in the first stage with 10-fold cross-validated penalty parameters (cvlasso-IV and cvridge-IV, respectively), as well as an IV estimator using a random forest in the first stage (randomForest-IV). As before, the CIV outperforms the competing shrinkage estimators for all considered sample sizes. The results thus further reflect the poor performance of popular machine learning methods in settings with categorical variables.

These results in Figure 1 and Figure 2 highlight the advantage of CIV to effectively regularize coefficients associated with categorical variables. In particular, lasso attempts to set first stage coefficients associated with the 50 available instruments to zero. Since the optimal instrument has two support points, each associated with half the categories, the post-lasso estimator should estimate at least 25 separate category means in the first stage. Post-lasso IV applied to categorical instruments is thus not able to sufficiently regularize first stage fit. The strength of CIV lies in its ability to automatically combine categories with similar first stage fit such that their coefficients can be jointly estimated. When the categories are combined correctly, this substantially reduces first stage estimation noise. For example, rather than using n_z observations to estimate the first stage coefficient corresponding to a

Figure 1: Power Curves ($K_0 = 2$, $p = 1$)



(a) $n_z = 30$ (b) $n_z = 50$

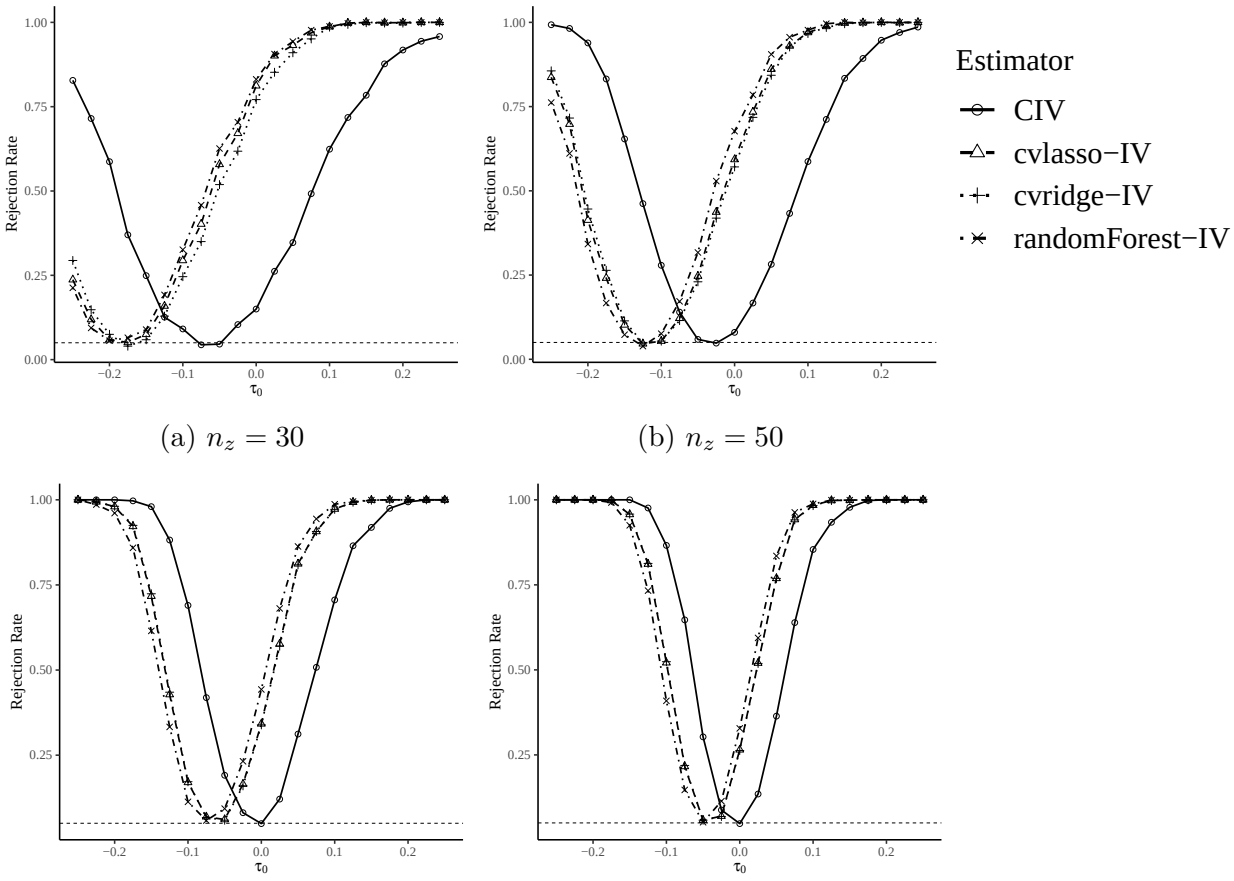
(c) $n_z = 100$ (d) $n_z = 150$

Notes. Simulation results are based on 1,000 simulations with $K_0 = 2$ and $p = 1$. The power curves plot the rejection rate of testing $H_0 : \tau_0 = 0$ at significance level $\alpha = 0.05$ as a function of of the true coefficient τ_0. CIV denotes the categorical IV estimator with known K_0. Oracle denotes the infeasible two stage least squares (TSLS) estimator that presumes knowledge of the optimal instruments. rlasso-IV denotes the post-lasso IV estimator as proposed in Belloni et al. (2012). TSLS denotes the feasible TSLS estimator that uses the observed categorical instruments. Standard errors used for testing are heteroskedasticity robust.

particular category as in TSLS or post-IV lasso, CIV uses $25 \times n_z$ observations after correctly combining observed instrument values. It is this strategy that allows for the near-oracle performance of CIV seen in the simulation.

To further highlight the importance of correctly combining the observed categories to take advantage of efficiency gains, Figure 3 revisits the simulation exercise but further separates the two support point of the optimal instrument. Using $m_0(z) = -1$ for $z \in \{1, \ldots, 25\}$ and $m_0(z) = 1$ for $z \in \{26, \ldots, 50\}$, the setting thus corresponds to stronger first stage. Again,

Figure 2: Power Curves ($K_0 = 2$, $p = 1$)



(a) $n_z = 30$ (b) $n_z = 50$

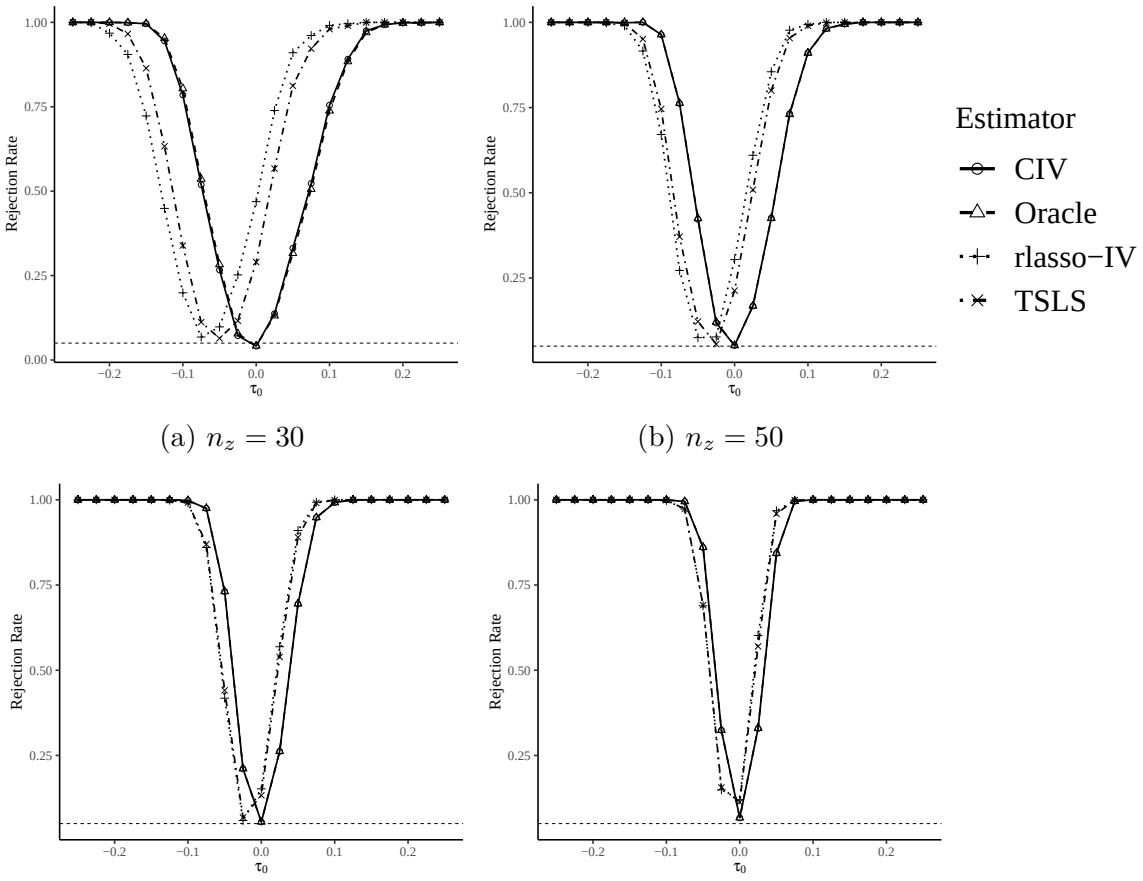
(c) $n_z = 100$ (d) $n_z = 150$

Notes. Simulation results are based on 1,000 simulations with $K_0 = 2$ and $p = 1$. The power curves plot the rejection rate of testing $H_0 : \tau_0 = 0$ at significance level $\alpha = 0.05$ as a function of of the true coefficient τ_0. CIV denotes the categorical IV estimator with known K_0. cvlasso-IV and cvridge-IV denotes IV estimators that compute first stage using lasso or ridge regression, respectively, where the shrinkage parameter is determined using 10-fold cross-validation. randomForest-IV denotes an IV estimator that compute the first stage using a random forest. Standard errors used for testing are heteroskedasticity robust.

panels (a)-(d) plot the power curves based on samples with 30, 50, 100, and 150 observations per category, respectively. In comparison to earlier results, CIV attains near oracle size and power even with 30 observations per category. Despite the much stronger first stage, both post-lasso IV and TSLS, however, continue to suffer greatly and do not attain nominal size levels even with 150 observations per category.[10]

[10]Appendix C provides additional results for the alternative shrinkage-based estimators.

Figure 3: Power Curves ($K_0 = 2$, $p = 2$)



(a) $n_z = 30$ (b) $n_z = 50$

(c) $n_z = 100$ (d) $n_z = 150$

Notes. Simulation results are based on 1,000 simulations with $K_0 = 2$ and $p = 2$. The power curves plot the rejection rate of testing $H_0 : \tau_0 = 0$ at significance level $\alpha = 0.05$ as a function of of the true coefficient τ_0. Estimators are the same as those in Figure 1. Standard errors used for testing are heteroskedasticity robust.

5 Application to Returns to Schooling

This section revisits the returns to schooling analysis of Angrist and Krueger (1991) in a sample of 329,509 American men born between 1930 and 1939. The authors use quarter of birth (QOB) indicators as instruments for the highest grade completed. This approach is motivated by two arguments. First, quarter of birth is plausibly exogenous with other determinants of wages. Second, children born in later quarters attain the minimum dropout age after having completed more schooling. While the QOB instrument thus appears a valid approach to instrument for years of schooling, it averages over potential heterogeneity in educational policy. A large number of categorical instruments arises when interactions

between QOB and indicators for year of birth (YOB) and place of birth (POB) are formed. These interactions capture the fact that mandatory schooling laws differ across cohorts and states. In particular, the interactions account for the possibility that two students born in the same QOB may differ on whether they can dropout depending on the particular policy in place in their state at the corresponding year.

A fully interacted specification results in 1530 excluded instruments and 510 control variables. The setting of Angrist and Krueger (1991) is thus an interesting application for assessing the performance of the proposed estimator. This paper joins the large literature on many or weak instrument estimators using the Angrist and Krueger (1991) application as an empirical illustration (among many others, see, e.g., Bound et al., 1995; Angrist and Krueger, 1995; Angrist et al., 1999; Donald and Newey, 2001; Hansen et al., 2008; Angrist and Frandsen, 2022; Mikusheva and Sun, 2022). With the exception of Angrist and Frandsen (2022) and Mikusheva and Sun (2022), most empirical analyzes of the Angrist and Krueger (1991) data consider disjoint interactions of QOB with YOB and POB, respectively, leading to 180 (rather than 1530) excluded instruments. As highlighted in Blandhol et al. (2022), causal interpretations of such specifications likely violate the monotonicity correctness of the first stage. Monotonicity correctness is guaranteed mechanically by the fully saturated interaction specification considered in this paper.

The CIV estimator considered here assumes $K_0 = 2$. The reason for restricting the optimal instrument to two support points is that a student's dropout decision either is or is not constrained by the mandatory attendance law in place in their state and corresponding year. The true underlying instrument is thus whether a particular student is induced or not induced to attend school for an additional year. While it would be ideal in this setting to know of the specific educational policies across all states and years in the cohort, the approach of CIV is to learn the relevant cutoffs from data.

Column (10) in Table 1 provides estimates and heteroskedasticity robust standard errors of six estimators at the full sample of Angrist and Krueger (1991).[11] The considered estimators

[11]LIML standard errors are not heteroskedasticity robust. Existing implementations of the CSE proposed Hansen et al. (2008) are computationally infeasible for the considered dataset.

are linear regression (OLS), TSLS, CIV with $K_0 = 2$ and $K_0 = 3$, respectively, two post-lasso IV estimators as proposed by Belloni et al. (2012), and LIML. The difference between the post-lasso IV estimators lies in the construction of the interactions. rlasso-IV-1 first includes main effects, then second order effects via interactions of QOB with YOB and POB, respectively, and finally third order effects by the remaining interactions between QOB and YOB and POB. In contrast, rlasso-IV-2 constructs interactions between the three sets of indicators to generate 1530 non-overlapping cells of the data.

The results at the full sample size in column (10) show similarity between OLS and TSLS estimates, with the CIV estimators and rlasso-IV-1 being between OLS and the LIML estimate. rlasso-IV-2, however, does not return valid estimates because no instruments are selected in the first stage. Note that without variable selection, both instrument construction approaches (rlasso-IV-1 and rlasso-IV-2) are identical as they imply the same first stage fitted values. Indeed, the TSLS coefficients are numerically identical regardless of the considered approach. With variable selection, however, the construction of indicators is crucial as it determines the composition of the baseline category. As a consequence, while sparsity may be an appropriate assumption in one indicator specification, it may not be appropriate in another.[12] Much akin the approach of Donald and Newey (2001) who consider an ordered list of instruments in increasing importance from which to choose from, applications of lasso-based estimators for categorical variables require careful consideration of the formed categories from the researcher. Importantly, this caveat does not apply to the regularization approach of the proposed CIV estimator, which combines indicators rather than setting individual coefficients to zero.[13]

Although substantial similarities between different candidate IV estimators and OLS in the Angrist and Krueger (1991) may cause unease, the true returns to education coefficient in the Angrist and Krueger (1991) application is unknown, making comparisons between

[12]Importantly, the fact that no instruments are selected in rlasso-IV-2 does not imply that none of them are important – rather, in this application, it is a consequence of the fact that almost *all* of the associated coefficients are different from zero leading to a violation of the underlying approximate sparsity assumption.

[13]Note that the post-lasso estimator allows for multiple sets of indicators. However, to achieve the same first stage fitted values as CIV, researchers would be required to include the power set, which exceeds the accommodated variable growth rates of Belloni et al. (2012). In the Angrist and Krueger (1991) application the power set corresponds to 2^{1530} indicators.

estimates ultimately inconclusive. Note further that in the nonparametric causal model of Imbens and Angrist (1994), TSLS, CIV, and the rlasso-based IV estimators target the same convex combination of causal effects. This is because each is a two-step estimator in the sense of Kolesár (2013) where the scalar-valued instrument estimated in the first step is the conditional expectation of schooling given quarter of birth, year of birth, and place of birth. In contrast, Kolesár (2013) implies that the LIML estimator generally targets a different and potentially non-convex combination of causal effects. LIML does thus not have the same robustness properties in the presence of unobserved heterogeneity as the alternative IV estimators, which further complicates the comparisons of point estimates.

In an attempt to allow for some insight into relative estimator properties in this application, I consider a subsampling exercise that compares differences between two approximations of the sampling distributions of the estimators: For $p \in \{10, 20, \ldots, 90\}$, randomly subsample $p\%$ of the Angrist and Krueger (1991) data (without replacement) and calculate the seven candidate estimators. Repeat this 1000 times and compare the mean coefficient estimates, the mean standard error estimates, as well as the subsampling standard deviation of the 1000 estimates. Large differences in the mean standard error estimates and the subsampling standard deviation suggest poor approximation of the sampling distribution, in particular, that the corresponding standard error does not fully capture the underlying sampling uncertainty.

Columns (1)-(9) of Table 1 show the results of the subsampling exercise for increasing sample sizes. As expected, the OLS and TSLS estimates do not vary substantially across sample sizes. For small sample sizes, the mean TSLS point estimate is almost indistinguishable from the OLS estimate caused by a first stage with near perfect fit. At 90% of the full sample size, the mean TSLS point estimate is 0.071 which is a difference of only 0.004 to the mean OLS point estimate of 0.067.

In contrast, CIV with $K_0 = 2$ achieves the same difference to the mean OLS point estimate as TSLS at just 20% of the sample size. At 90% of the full sample size, the mean CIV ($K_0 = 2$) point estimate is 0.078. Further for all sample sizes, the mean standard error and the subsampling standard deviation of CIV ($K_0 = 2$) are close, suggesting the asymptotic approximation characterizes the sampling uncertainty well. The CIV point estimates are not

substantially affected when the optimal instrument is allowed an additional support point with $K_0 = 3$. However, the increase in support points increases the variance of the estimates by about an order of magnitude across sample sizes.

The mean rlasso-IV-1 estimates are stable for sample sizes of 30-100% of the full sample. This is due to the fact that – whenever lasso in the first stage does not shrink all coefficients to zero – only the 4th quarter of birth indicator is selected. For example, at 50% of the full sample (Column (5)), rlasso-IV-1 selects instruments in 812 out of 1,000 subsampling repetitions of which 778 select only the 4th quarter of birth indicator. At 90% of the full sample (Column (9)), 996 out of 1,000 select the 4th quarter of birth indicator. In contrast, the alternative specification of indicators implemented for rlasso-IV-2 does not select any instruments in the first stage in any of the 1,000 repetitions for sample sizes larger than 40%. These results thus further highlight the poor applicability of (approximate) sparsity based shrinkage in settings with categorical variables.

Finally, the point estimates of the LIML estimator are highly volatile for smaller sample sizes. This is reflected both in differences of the mean point estimates across columns (1)-(6) as well as the large subsampling standard deviation of the estimates which is at times several magnitudes larger than the corresponding standard error. While the differences in point estimates between LIML and the other candidate estimators are not easily interpretable due to differences in the estimands under unobserved heterogeneity in causal effects, the subsampling results suggest that the LIML estimator is poorly characterized by the conventional asymptotic approximation for smaller versions of the Angrist and Krueger (1991) data.[14]

In summary, the subsampling exercise shows that by effectively regularizing the first stage, CIV with two support points can produce second stage estimates that substantially differ from TSLS estimates in the Angrist and Krueger (1991) data. These estimates appear to be well-characterized by the conventional asymptotic approximation even at smaller subsets of the sample.

[14]Because consistency of the subsampling distribution relies on asymptotic regimes that let the share of the subsampled data tend to zero, comparisons between standard errors and the subsampling standard deviation in Table 1 may be most informative for small sample shares (e.g., columns (1)-(6)).

Table 1: Estimating Returns to Schooling

		32,950 (1)	65,901 (2)	98,852 (3)	131,803 (4)	167,754 (5)	197,705 (6)	230,656 (7)	263,607 (8)	296,558 (9)	329,509 (10)
OLS	Mean $\hat{\tau}_n$	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067	0.067
	Mean $se(\hat{\tau}_n)$	0.001	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000
	Std. Dev. $\hat{\tau}_n$	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000	-
TSLS	Mean $\hat{\tau}_n$	0.067	0.067	0.068	0.069	0.069	0.069	0.070	0.070	0.071	0.071
	Mean $se(\hat{\tau}_n)$	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
	Std. Dev. $\hat{\tau}_n$	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.003	0.002	-
CIV ($K_0 = 2$)	Mean $\hat{\tau}_n$	0.070	0.071	0.072	0.073	0.074	0.074	0.075	0.076	0.078	0.078
	Mean $se(\hat{\tau}_n)$	0.010	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.008
	Std. Dev. $\hat{\tau}_n$	0.008	0.009	0.008	0.007	0.006	0.006	0.005	0.005	0.004	-
CIV ($K_0 = 3$)	Mean $\hat{\tau}_n$	0.069	0.070	0.069	0.072	0.074	0.077	0.075	0.077	0.074	0.074
	Mean $se(\hat{\tau}_n)$	0.035	0.019	0.368	0.040	0.018	0.073	0.021	0.016	0.060	0.060
	Std. Dev. $\hat{\tau}_n$	0.037	0.021	0.137	0.062	0.024	0.106	0.030	0.019	0.087	-
rlasso-IV-1	Mean $\hat{\tau}_n$	0.128	0.105	0.085	0.086	0.086	0.086	0.088	0.086	0.086	0.086
	Mean $se(\hat{\tau}_n)$	0.019	0.033	0.037	0.036	0.035	0.033	0.031	0.029	0.027	0.025
	Std. Dev. $\hat{\tau}_n$	0.037	0.044	0.032	0.027	0.025	0.020	0.017	0.013	0.009	-
rlasso-IV-2	Mean $\hat{\tau}_n$	0.098	0.166	0.046	-	-	-	-	-	-	-
	Mean $se(\hat{\tau}_n)$	0.043	0.040	0.035	-	-	-	-	-	-	-
	Std. Dev. $\hat{\tau}_n$	0.077	-	-	-	-	-	-	-	-	-
LIML	Mean $\hat{\tau}_n$	0.127	0.053	0.128	0.108	0.080	0.096	0.097	0.099	0.102	0.102
	Mean $se(\hat{\tau}_n)$	0.067	0.055	0.033	0.026	0.024	0.020	0.019	0.017	0.016	0.014
	Std. Dev. $\hat{\tau}_n$	1.886	4.120	0.676	0.459	0.710	0.168	0.067	0.034	0.020	-

Notes. CIV ($K_0 = 2$) denotes the CIV estimator where the optimal instrument is restricted to two support points. rlasso-IV-1 and rlasso-IV-2 denote the post-lasso IV estimators of Belloni et al. (2012) under two indicator constructions, using main effects first and separating the categories into disjoint cells, respectively. Columns (1)-(9) correspond to estimation results of 1000 repeated randomly chosen subsamples of the Angrist and Krueger (1991) data of size n. Mean $\hat{\tau}_n$, Mean se $(\hat{\tau}_n)$, and Std. Dev. $\hat{\tau}_n$ correspond to the mean point estimate, mean standard error estimate, and the standard deviation of the point estimates across the 1000 subsamples. Column (10) applies the estimators to the full data and provides point and standard error estimates. With exception of the LIML, standard errors are heteroskedasticity robust. Computation of the CSE standard errors for the LIML required more than 100-times the runtime of the other standard errors, which prohibited their estimation in this exercise.

6 Conclusion

This paper considers estimation with categorical instrumental variables when the number of observations per category is small. The proposed categorical instrumental variable estimator is motivated by a first-stage regularization assumption that restricts the unknown optimal instrument to have fixed finite support. In asymptotic regimes that allow the number of observations to grow at arbitrarily slow polynomial rate with the sample, I show that the proposed estimator is root-n asymptotically normal and attains the same asymptotic variance as the infeasible oracle instrument variable estimator that presumes knowledge of the optimal instruments. Further, the proposed estimator is semiparametrically efficient when the second stage is homoskedastic. A simulation exercise illustrates the finite sample performance of the proposed CIV estimator and highlights pitfalls associated with lasso-based IV estimators in the setting of categorical instruments. Further, the application to Angrist and Krueger (1991) shows that CIV is more stable than LIML while being robust to the many instrument bias that TSLS suffers from.

The key advantage of the proposed CIV estimator is the appeal of the finite support assumption over alternative first stage regularization assumptions such as (approximate) sparsity. As showcased in the simulation and application, sparsity can have unintended implications in settings with categorical instruments with substantial practical consequences. In particular, the assumption leveraged in this paper does not restriction the proportion of observations across latent categories. CIV thus appears a suitable and easily applicable alternative to existing estimators in important empirical settings similar to Angrist and Krueger (1991) or judge IV designs.

References

Aizer, A. and Doyle Jr, J. J. (2015). Juvenile incarceration, human capital, and future crime: Evidence from randomly assigned judges. *Quarterly Journal of Economics*, 130(2):759–803.

Amemiya, T. (1974). Multivariate regression and simultaneous equation models when the dependent variables are truncated normal. *Econometrica*, pages 999–1012.

Angrist, J. D. and Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, 40(S1):S97–S140.

Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.

Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.

Arthur, D. and Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, page 1027–1035. Society for Industrial and Applied Mathematics Philadelphia, PA.

Bekker, P. A. (1994). Alternative approximations to the distributions of instrumental variable estimators. *Econometrica: Journal of the Econometric Society*, pages 657–681.

Bekker, P. A. and Van der Ploeg, J. (2005). Instrumental variable estimation based on grouped data. *Statistica Neerlandica*, 59(3):239–267.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Bester, C. A. and Hansen, C. B. (2016). Grouped effects estimators in fixed effects models. *Journal of Econometrics*, 190(1):197–208.

Bhuller, M., Dahl, G. B., Løken, K. V., and Mogstad, M. (2020). Incarceration, recidivism, and employment. *Journal of Political Economy*, 128(4):1269–1324.

Blandhol, C., Bonney, J., Mogstad, M., and Torgovitsky, A. (2022). When is TSLS actually LATE? *BFI Working Paper*, (2022-16).

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Carrasco, M. (2012). A regularization approach to the many instruments problem. *Journal of Econometrics*, 170(2):383–398.

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

Chamberlain, G. and Imbens, G. (2004). Random effects estimators with many instrumental variables. *Econometrica*, 72(1):295–306.

Chao, J. C. and Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1).

Dhar, D., Jain, T., and Jayachandran, S. (2022). Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in india. *American Economic Review*, 112(3):899–927.

Donald, S. G. and Newey, W. K. (2001). Choosing the number of instruments. *Econometrica*, 69(5):1161–1191.

Gilchrist, D. S. and Sands, E. G. (2016). Something to talk about: Social spillovers in movie consumption. *Journal of Political Economy*, 124(5):1339–1382.

Hahn, J. and Moon, H. R. (2010). Panel data models with finite number of multiple equilibria. *Econometric Theory*, 26(3):863–881.

Hansen, C., Hausman, J., and Newey, W. (2008). Estimation with many instrumental variables. *Journal of Business & Economic Statistics*, 26(4):398–422.

Hansen, P., Mladenović, N., and Pérez, J. A. M. (2010). Variable neighbourhood search: methods and applications. *Annals of Operations Research*, 175(1):367–407.

Hausman, J. A., Newey, W. K., Woutersen, T., Chao, J. C., and Swanson, N. R. (2012). Instrumental variable estimation with heteroskedasticity and many instruments. *Quantitative Economics*, 3(2):211–255.

Imbens, G. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, pages 467–475.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *American Economic Review*, 96(3):863–876.

Kolesár, M. (2013). Estimation in an instrumental variables model with treatment effect heterogeneity. Working Paper.

Maestas, N., Mullen, K. J., and Strand, A. (2013). Does disability insurance receipt discourage work? using examiner assignment to estimate causal effects of ssdi receipt. *American economic review*, 103(5):1797–1829.

Mikusheva, A. and Sun, L. (2022). Inference with many weak instruments. *Review of Economic Studies*, 89(5):2663–2686.

Newey, W. K. (1990). Efficient instrumental variables estimation of nonlinear models. *Econometrica*, pages 809–837.

Okui, R. (2011). Instrumental variable estimation in the presence of many moment conditions. *Journal of Econometrics*, 165(1):70–86.

Su, L., Shi, Z., and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

A Proof of Theorem 1

The proof of Theorem 1 proceeds in three steps: First, I begin the proof with a set of lemmas to characterize the asymptotic properties of the estimator $\hat{m}_n$. The proof of lemmas 1-3 heavily leverages the arguments of Bonhomme and Manresa (2015) with adaptations to accommodate the unbalanced categorical variable Z. Most notably, I prove and apply Lemma 4 to account for a random number of observations per category. The proofs of these lemmas are included here for completeness.

Second, I proof that $\hat{\tau}_n$ tends to the infeasible two stage least squares estimator $\tilde{\tau}_n$ at rate $n^{-\delta}$ for any $\delta > 0$, where $\tilde{\tau}_n$ is defined by

$$\tilde{\tau}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \bar{Y}_n\right)\left(\tilde{m}_n(Z_i) - \bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(\tilde{m}_n(Z_i) - \bar{D}_n\right)^2}, \tag{5}$$

and where $\tilde{m}_n$ is the infeasible least squares estimator defined by

$$\tilde{m}_n(z) \equiv \sum_{g=1}^{K_0} \mathbb{1}\{z \in \mathcal{Z}_g^0\}\tilde{\alpha}_g, \quad \tilde{\alpha} \equiv \underset{\alpha \in \mathcal{M}^{K_0}}{\arg\min} \sum_{i=1}^{n}\left(D_i - \sum_{g=1}^{K_0} \mathbb{1}\{Z_i \in \mathcal{Z}_g^0\}\alpha_g\right)^2. \tag{6}$$

This estimator is infeasible because it presumes knowledge of the unknown partition $(\mathcal{Z}_g^0)_{g=1}^{K_0}$ of $\operatorname{supp} Z$ for which the conditional expectation function m_0 takes the same value.

Finally, I characterize the asymptotic distribution of $\tilde{\tau}_n$.

Notation. The cardinality of the support of the instrument is $|\operatorname{supp} Z| \equiv K_Z$. A η-neighborhood around α_0 is denoted by $\mathcal{N}_{\alpha_0}(\eta) \equiv \{\alpha \in \mathcal{M}^{K_0} : \|\alpha - \alpha_0\| < \eta\}$.

A.1 Asymptotic properties of $\hat{m}_n$

Lemma 1. *Let the assumptions of Theorem 1 hold. Then*

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_n(Z_i) - m_0(Z_i)\right)^2 = o_p(1).$$

Proof. Define

$$\hat{Q}(m) = \frac{1}{n}\sum_{i=1}^{n}(D_i - m(Z_i))^2 = \frac{1}{n}\sum_{i=1}^{n}(m_0(Z_i) + V_i - m(Z_i))^2$$

and

$$\tilde{Q}(m) = \frac{1}{n}\sum_{i=1}^{n}(m_0(Z_i) - m(Z_i))^2 + \frac{1}{n}\sum_{i=1}^{n}V_i^2.$$

Now

$$\hat{Q}(m) - \tilde{Q}(m) = \frac{2}{n}\sum_{i=1}^{n}V_i(m_0(Z_i) - m(Z_i)) = \frac{2}{n}\sum_{i=1}^{n}V_i m_0(Z_i) - \frac{2}{n}\sum_{i=1}^{n}V_i m(Z_i).$$

Hence, for any $m : \operatorname{supp} Z \to \mathcal{M}, |m(\operatorname{supp} Z)| = K_0$,

$$\frac{1}{n}\sum_{i=1}^{n}V_i m(Z_i) = \frac{1}{n}\sum_{i=1}^{n}\sum_{z\in\operatorname{supp} Z}\mathbb{1}_z(Z_i)V_i m(k) = \frac{1}{K_Z}\sum_{z\in\operatorname{supp} Z}m(k)\left(\frac{K_Z}{n}\sum_{i=1}^{n}\mathbb{1}_z(Z_i)V_i\right),$$

so by Cauchy-Schwarz

$$\left(\frac{1}{K_Z}\sum_{z\in\operatorname{supp} Z}m(k)\left(\frac{K_Z}{n}\sum_{i=1}^{n}\mathbb{1}_z(Z_i)V_i\right)\right)^2 \le \left(\frac{1}{K_Z}\sum_{z\in\operatorname{supp} Z}m(k)^2\right)\left(\frac{1}{K_Z}\sum_{z\in\operatorname{supp} Z}\left(\frac{K_Z}{n}\sum_{i=1}^{n}\mathbb{1}_z(Z_i)V_i\right)^2\right).$$

The first term is $O_p(1)$ by Assumption 9. For the second term, we have

$$\frac{1}{K_Z}\sum_{z\in\operatorname{supp} Z}\left(\frac{K_Z}{n}\sum_{i=1}^{n}\mathbb{1}_z(Z_i)V_i\right)^2 = \frac{K_Z}{n^2}\sum_{z\in\operatorname{supp} Z}\sum_{i=1}^{n}\sum_{j=1}^{n}V_iV_j\mathbb{1}_z(Z_i)\mathbb{1}_z(Z_j).$$

Taking expectations results in

$$E\left[\frac{K_Z}{n^2}\sum_{z\in\operatorname{supp} Z}\sum_{i=1}^{n}\sum_{j=1}^{n}V_iV_j\mathbb{1}_z(Z_i)\mathbb{1}_z(Z_j)\right] \overset{[1]}{=} \frac{K_Z}{n^2}\sum_{j=1}^{n}E\left[V_i^2\right] \overset{[2]}{\le} \frac{K_Z}{n}L \overset{[3]}{=} o(1),$$

where [1] follows from Assumption 10, [2] follows from Assumption 5, and [3] is a consequence of Assumption 3 and the fact that $\sum_{z\in\operatorname{supp} Z}\Pr(Z = z) = 1$ which implies $K_Z = o(n)$. It

then follows that

$$\sup_{\substack{m:\,\mathrm{supp}\,Z\to\mathcal{M}\\ |m(\mathrm{supp}\,Z)|=K_0}} \left|\hat{Q}(m)-\tilde{Q}(m)\right| = o_p(1). \tag{7}$$

To complete the proof of Lemma 1, consider

$$\tilde{Q}(\hat{m}) \overset{[1]}{=} \hat{Q}(\hat{m}) + o_p(1) \overset{[2]}{\leq} \hat{Q}(m_0) + o_p(1) \overset{[3]}{=} \tilde{Q}(m_0) + o_p(1)$$

where [1] and [3] follows from (7), and [2] follows from definition of $\hat{m}$. Then

$$\tilde{Q}(\hat{m}) - \tilde{Q}(m_0) = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{m}_n(Z_i) - m_0(Z_i)\right)^2 = o_p(1).$$

□

Note that, since $\mathrm{supp}\,Z$ is finite and m in (4) takes K_0 unique values, each such function $m : \mathrm{supp}\,Z \to \mathcal{M}, |m(\mathrm{supp}\,Z)| = K_0$ is fully characterized by a partition γ of $\mathrm{supp}\,Z$ in K_0 sets and a corresponding vector of coefficients $\alpha = (\alpha_1, \ldots, \alpha_{K_0})$ – i.e.,

$$m(z) = \alpha_{g_z}, \tag{8}$$

where g_z denotes the set associated with the value $z \in \mathrm{supp}\,Z$. Throughout, let α^0 and g^0 denote the coefficients and partition that characterize the true conditional expectation function m_0.

By (8), it is possible to re-cast the estimator $\hat{m}_n$ in (4) as

$$(\hat{\alpha}_n, \hat{g}_n) \equiv \underset{\substack{\alpha\in\mathcal{M}^{K_0}\\ \gamma\in\Gamma_{K_0}}}{\arg\min} \sum_{i=1}^{n} \left(D_i - \alpha_{g_{Z_i}}\right)^2, \tag{9}$$

where Γ_{K_0} denotes the set of all possible partitions of $\mathrm{supp}\,Z$ into K_0 sets. This representation follows the group fixed effects estimator of Bonhomme and Manresa (2015). I adopt it here because it allows for separate analysis of estimation properties of the partition and the

coefficients.

Since m in (8) is invariant to relabeling of the coefficients α and partitioning sets g, it is useful to consider the Hausdorff distance d_H in $\mathbb{R}^{K_0}$ to characterize the asymptotic properties of the estimators $\hat{\alpha}$. In particular, define

$$d_H(a,b)^2 = \max\left\{ \max_{g\in\{1,\dots,K_0\}} \left(\min_{\tilde{g}\in\{1,\dots,K_0\}} (\alpha_{\tilde{g}} - b_g)^2 \right), \max_{\tilde{g}\in\{1,\dots,K_0\}} \left(\min_{g\in\{1,\dots,K_0\}} (\alpha_{\tilde{g}} - b_g)^2 \right) \right\}.$$

Lemma 2. *Let the assumptions of Theorem 1 hold. Then, as $n \to \infty$,*

$$d_H(\hat{\alpha}, \alpha^0) \xrightarrow{p} 0.$$

Proof. The proof proceeds in two steps. I first show that for all $g \in \{1,\dots,K_0\}$,

$$\min_{\tilde{g}\in\{1,\dots,K_0\}} \left(\hat{\alpha}_{\tilde{g}} - \alpha_g^0\right)^2 = o_p(1). \tag{10}$$

Let $g \in \{1,\dots,K_0\}$. We have

$$\frac{1}{n}\sum_{i=1}^n \left(\min_{\tilde{g}\in\{1,\dots,K_0\}} \mathbb{1}\{g_{Z_i}^0 = g\}(\hat{\alpha}_{\tilde{g}} - \alpha_g^0)^2 \right) = \left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}\{g_{Z_i}^0 = g\} \right) \left(\min_{\tilde{g}\in\{1,\dots,K_0\}} (\hat{\alpha}_{\tilde{g}} - \alpha_g^0)^2 \right).$$

Since by Assumption 7, the first term is non-vanishing, it suffices to show that the left hand side is $o_p(1)$ for all $g \in \{1,\dots,K_0\}$. In particular,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \left(\min_{\tilde{g}\in\{1,\dots,K_0\}} \mathbb{1}\{g_{Z_i}^0 = g\}(\hat{\alpha}_{\tilde{g}} - \alpha_g^0)^2 \right) \leq & \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}\{g_{Z_i}^0 = g\}(\hat{\alpha}_{\hat{g}_{Z_i}} - \alpha_g^0)^2 \right) \\
\leq & \frac{1}{n}\sum_{i=1}^n (\hat{\alpha}_{\hat{g}_{Z_i}} - \alpha_{g_{Z_i}^0}^0)^2 \\
= & \frac{1}{n}\sum_{i=1}^n (\hat{m}(Z_i) - m_0(Z_i))^2 \\
= & o_p(1),
\end{aligned}$$

where the final equality follows from Lemma 1.

Next, I show that for all $\tilde{g} \in \{1, \ldots, K_0\}$,

$$\min_{g \in \{1,\ldots,K_0\}} \left(\hat{\alpha}_{\tilde{g}} - \alpha_g^0\right)^2 \xrightarrow{p} 0. \tag{11}$$

Define

$$\sigma(g) \equiv \underset{\tilde{g} \in \{1,\ldots,K_0\}}{\arg\min} \left(\hat{\alpha}_{\tilde{g}} - \alpha_g^0\right)^2.$$

By the triangle inequality, it holds that

$$\left|\hat{\alpha}_{\sigma(g)} - \hat{\alpha}_{\sigma(\tilde{g})}\right| \geq \left|\alpha_g^0 - \alpha_{\tilde{g}}^0\right| - \left|\hat{\alpha}_{\sigma(g)} - \alpha_g^0\right| - \left|\hat{\alpha}_{\sigma(\tilde{g})} - \alpha_{\tilde{g}}^0\right|$$

where $\left|\hat{\alpha}_{\sigma(g)} - \alpha_g^0\right|$ and $\left|\hat{\alpha}_{\sigma(\tilde{g})} - \alpha_{\tilde{g}}^0\right|$ are $o_p(1)$ by the first result (10), and $\left|\alpha_g^0 - \alpha_{\tilde{g}}^0\right| > 0$ by Assumption 8. Thus $\sigma(g) \neq \sigma(\tilde{g})$ with probability approaching one, implying that the inverse σ^{-1} is well-defined.

Now, with probability approaching one, we have

$$\begin{aligned}
\min_{g \in \{1,\ldots,K_0\}} \left(\hat{\alpha}_{\tilde{g}} - \alpha_g^0\right)^2 &\leq \left(\hat{\alpha}_{\tilde{g}} - \alpha_{\sigma^{-1}(\tilde{g})}^0\right)^2 \\
&\overset{[1]}{=} \min_{h \in \{1,\ldots,K_0\}} \left(\hat{\alpha}_h - \alpha_{\sigma^{-1}(\tilde{g})}^0\right)^2 \\
&\overset{[2]}{=} o_p(1),
\end{aligned}$$

where [1] follows from $\tilde{g} = \sigma(\sigma^{-1}(\tilde{g}))$, and [2] follows from the first result (10).

Combining (10) and (11) completes the proof. □

The proof of Lemma 2 shows that there exists a permutation $\sigma : \{1, \ldots, K_0\} \to \{1, \ldots, K_0\}$ such that $\left(\hat{\alpha}_{\sigma(g)} - \alpha_g^0\right)^2 = o_p(1)$. It is thus possible take $\sigma(g) = g$ by simply relabeling the elements of $\hat{\alpha}$. The remainder of the proof adopts this convention.

Lemma 3. *For $\eta > 0$ small enough, as $n \to \infty$, we have for all $\delta > 0$*

$$\sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} = o_p(n^{-\delta}).$$

Proof. From the definition of $\hat{g}$ in (9), we have for all $g \in \{1, \ldots, K_0\}$,

$$\begin{aligned}
\mathbb{1}\{\hat{g}_z(\alpha) = g\} &= \mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_z(Z_i)(Y_i - \alpha_g)^2 \leq \min_{\tilde{g} \in \{1,\ldots,K_0\}} \sum_{i=1}^{n} \mathbb{1}_z(Z_i)(Y_i - \alpha_{\tilde{g}})^2\right\} \\
&\leq \mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_z(Z_i)(Y_i - \alpha_g)^2 \leq \sum_{i=1}^{n} \mathbb{1}_z(Z_i)(Y_i - \alpha_{g^0_z})^2\right\}.
\end{aligned}$$

As a consequence,

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} = \sum_{g=1}^{K_0} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{g^0_{Z_i} \neq g\} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) = g\} \leq \frac{1}{n} \sum_{i=1}^{n} \sum_{g=1}^{K_0} M_{Z_i g}(\alpha), \tag{12}$$

where for all $z \in \operatorname{supp} Z$

$$\begin{aligned}
M_{zg}(\alpha) &\equiv \mathbb{1}\{g^0_z \neq g\} \mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_z(Z_i)(D_i - \alpha_g)^2 \leq \sum_{i=1}^{n} \mathbb{1}_z(Z_i)(D_i - \alpha_{g^0_z})^2\right\} \\
&= \mathbb{1}\{g^0_z \neq g\} \mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_z(Z_i)\left(2D_i(\alpha_{g^0_z} - \alpha_g) - (\alpha_{g^0_z} - \alpha_g)(\alpha_{g^0_z} + \alpha_g)\right) \leq 0\right\} \\
&= \mathbb{1}\{g^0_z \neq g\} \mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_z(Z_i)(\alpha_{g^0_z} - \alpha_g)\left(V_i + \alpha^0_{g^0_z} - \frac{(\alpha_{g^0_z} + \alpha_g)}{2})\right) \leq 0\right\},
\end{aligned}$$

where I have substituted for $D_i = \alpha^0_{g^0_{Z_i}} + V_i$ and rearranged terms for simplification.

Now, whenever $\alpha \in \mathcal{N}_{\alpha_0}(\eta)$, it is possible to bound $M_{zg}(\alpha)$ by a quantity that does not depend on α. In particular, note that

$$M_{zg}(\alpha) \leq \max_{\tilde{g} \neq g} \mathbb{1}\left\{\sum_{i=1}^{n} \mathbb{1}_z(Z_i)(\alpha_{\tilde{g}} - \alpha_g)\left(V_i + \alpha^0_{\tilde{g}} - \frac{\alpha_{\tilde{g}} + \alpha_g}{2}\right) \leq 0\right\}.$$

Further, by simple application of the triangle inequality

$$
\begin{aligned}
&\left|\sum_{i=1}^{n}\mathbb{1}_z(Z_i)(\alpha_{\tilde{g}}-\alpha_g)\left(V_i+\alpha^0_{\tilde{g}}-\frac{\alpha_{\tilde{g}}+\alpha_g}{2}\right)-\sum_{i=1}^{n}\mathbb{1}_z(Z_i)(\alpha^0_{\tilde{g}}-\alpha^0_g)\left(V_i+\alpha^0_{\tilde{g}}-\frac{\alpha^0_{\tilde{g}}+\alpha^0_g}{2}\right)\right| \\
\leq&\left|\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\left[(\alpha_{\tilde{g}}-\alpha_g)-(\alpha^0_{\tilde{g}}-\alpha^0_g)\right]V_i\right| \\
&+\left|\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\left[(\alpha_{\tilde{g}}-\alpha_g)\left(\alpha^0_{\tilde{g}}-\frac{\alpha_{\tilde{g}}+\alpha_g}{2}\right)-(\alpha^0_{\tilde{g}}-\alpha^0_g)\left(\alpha^0_{\tilde{g}}-\frac{\alpha^0_{\tilde{g}}+\alpha^0_g}{2}\right)\right]\right|.
\end{aligned}
\tag{13}
$$

For the first term in (13), it holds that

$$
\begin{aligned}
&\left|\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\left[(\alpha_{\tilde{g}}-\alpha_g)-(\alpha^0_{\tilde{g}}-\alpha^0_g)\right]V_i\right| \\
\overset{[1]}{\leq}&\left(\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\right)\left[(\alpha_{\tilde{g}}-\alpha_g)-(\alpha^0_{\tilde{g}}-\alpha^0_g)\right]\left(\frac{1}{\sum_{i=1}^{n}\mathbb{1}_z(Z_i)}\sum_{i=1}^{n}\mathbb{1}_z(Z_i)V_i^2\right)^{\frac{1}{2}} \\
\overset{[2]}{\leq}&2\left(\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\right)\sqrt{\eta}\left(\frac{1}{\sum_{i=1}^{n}\mathbb{1}_z(Z_i)}\sum_{i=1}^{n}\mathbb{1}_z(Z_i)V_i^2\right)^{\frac{1}{2}},
\end{aligned}
$$

where [1] follows from Jensen's inequality, and [2] follows from $\alpha\in\mathcal{N}_{\alpha^0}(\eta)$. Similarly, for the second term in (13), it holds that

$$
\begin{aligned}
&\left|\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\left[(\alpha_{\tilde{g}}-\alpha_g)\left(\alpha^0_{\tilde{g}}-\frac{\alpha_{\tilde{g}}+\alpha_g}{2}\right)-(\alpha^0_{\tilde{g}}-\alpha^0_g)\left(\alpha^0_{\tilde{g}}-\frac{\alpha^0_{\tilde{g}}+\alpha^0_g}{2}\right)\right]\right| \\
=&\frac{1}{2}\left(\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\right)\left|\left[(\alpha_{\tilde{g}}-\alpha_g)-\left(\alpha^0_{\tilde{g}}-\alpha^0_g\right)\right]\left(\alpha^0_{\tilde{g}}-\alpha^0_g\right)+(\alpha_{\tilde{g}}-\alpha_g)\left[(\alpha^0_{\tilde{g}}+\alpha^0_g)-(\alpha_{\tilde{g}}+\alpha_g)\right]\right| \\
\leq&\frac{1}{2}\left(\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\right)\left(\left|\left[(\alpha_{\tilde{g}}-\alpha_g)-\left(\alpha^0_{\tilde{g}}-\alpha^0_g\right)\right]\left(\alpha^0_{\tilde{g}}-\alpha^0_g\right)+\left(\alpha^0_{\tilde{g}}-\alpha^0_g\right)\left[(\alpha^0_{\tilde{g}}+\alpha^0_g)-(\alpha_{\tilde{g}}+\alpha_g)\right]\right|\right. \\
&\left.+\left|(\alpha_{\tilde{g}}-\alpha_g)\left[(\alpha^0_{\tilde{g}}+\alpha^0_g)-(\alpha_{\tilde{g}}+\alpha_g)\right]-\left(\alpha^0_{\tilde{g}}-\alpha^0_g\right)\left[(\alpha^0_{\tilde{g}}+\alpha^0_g)-(\alpha_{\tilde{g}}+\alpha_g)\right]\right|\right) \\
\overset{[1]}{\leq}&\left(\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\right)\left(|\alpha^0_{\tilde{g}}-\alpha^0_g|\sqrt{\eta}+2\eta\right) \\
\overset{[2]}{\leq}&\left(\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\right)C\sqrt{\eta},
\end{aligned}
$$

with $C\equiv|\alpha^0_{\tilde{g}}-\alpha^0_g|+2$ a constant independent of η,n and α, and where [1] follows from $\alpha\in\mathcal{N}_{\alpha^0}(\eta)$, and [2] follows from $\sqrt{\eta}\geq\eta$ for $\eta\in[0,1]$.

Therefore,

$$
\begin{aligned}
M_{zg}(\alpha) &\leq \max_{\tilde{g}\neq g} \mathbb{1}\Bigg\{ \sum_{i=1}^{n} \mathbb{1}_z(Z_i)(\alpha^0_{\tilde{g}} - \alpha^0_g)\left(V_i + \alpha^0_{\tilde{g}} - \frac{\alpha^0_{\tilde{g}} + \alpha^0_g}{2}\right) \\
&\qquad \leq 2\left(\sum_{i=1}^{n} \mathbb{1}_z(Z_i)\right)\sqrt{\eta}\left(\frac{1}{\sum_{i=1}^{n}\mathbb{1}_z(Z_i)}\sum_{i=1}^{n}\mathbb{1}_z(Z_i)V_i^2\right)^{\frac{1}{2}} + \left(\sum_{i=1}^{n}\mathbb{1}_z(Z_i)\right)C\sqrt{\eta}\Bigg\} \\
&= \max_{\tilde{g}\neq g} \mathbb{1}\Bigg\{ \frac{1}{\sum_{i=1}^{n}\mathbb{1}_z(Z_i)}\sum_{i=1}^{n}\mathbb{1}_z(Z_i)V_i(\alpha^0_{\tilde{g}} - \alpha^0_g) \\
&\qquad \leq 2\sqrt{\eta}\left(\frac{1}{\sum_{i=1}^{n}\mathbb{1}_z(Z_i)}\sum_{i=1}^{n}\mathbb{1}_z(Z_i)V_i^2\right)^{\frac{1}{2}} + C\sqrt{\eta} - \frac{1}{2}\left(\alpha^0_{\tilde{g}} - \alpha^0_g\right)^2\Bigg\},
\end{aligned}
\tag{14}
$$

where the right hand side does not depend on α. Hence, $\sup_{\alpha\in\mathcal{N}_{\alpha^0}(\eta)} M_{zg}(\alpha) \leq \tilde{M}_{zg}$, for $\eta < 1$, where $\tilde{M}_{zg}$ denotes the final term in (14). Combining with (12) then implies

$$
\sup_{\alpha\in\mathcal{N}_{\alpha^0}(\eta)} \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} \leq \frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{K_0}\tilde{M}_{Z_i g},
$$

and therefore for any $\epsilon > 0$ and $\delta > 0$

$$
\begin{aligned}
&\Pr\left(\sup_{\alpha\in\mathcal{N}_{\alpha^0}(\eta)} \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} > \epsilon n^{-\delta}\right) \\
\leq &\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{K_0}\tilde{M}_{Z_i g} > \epsilon n^{-\delta}\right) \\
\overset{[1]}{\leq} &\frac{E\left[\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{K_0}\tilde{M}_{Z_i g}\right]}{\epsilon n^{-\delta}} \\
= &\frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{g=1}^{K_0}\Pr(\tilde{M}_{Z_i g} = 1)}{\epsilon n^{-\delta}},
\end{aligned}
$$

where [1] follows from Markov's inequality. It thus suffices to show that $\forall z \in \operatorname{supp} Z$, $g \in \{1, \ldots, K_0\}$, and $\delta > 0$,

$$
\Pr(\tilde{M}_{zg} = 1) = o(n^{-\delta}). \tag{15}
$$

For this purpose, consider

$$\begin{aligned}\Pr\left(\tilde{M}_{zg}\right) &\overset{[1]}{\leq} \sum_{\tilde{g}\neq g} \Pr\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_z(Z_i)} \sum_{i=1}^n \mathbb{1}_z(Z_i) V_i(\alpha_{\tilde{g}}^0 - \alpha_g^0)\right. \\ &\qquad \left. \leq 2\sqrt{\eta}\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_z(Z_i)} \sum_{i=1}^n \mathbb{1}_z(Z_i) V_i^2\right)^{\frac{1}{2}} + C\sqrt{\eta} - \frac{1}{2}\left(\alpha_{\tilde{g}}^0 - \alpha_g^0\right)^2\right) \\ &\overset{[2]}{\leq} \sum_{\tilde{g}\neq g}\left[\Pr\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_z(Z_i)} \sum_{i=1}^n \mathbb{1}_z(Z_i) V_i(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq 2\sqrt{\eta}\sqrt{L} + C\sqrt{\eta} - \frac{1}{2}\left(\alpha_{\tilde{g}}^0 - \alpha_g^0\right)^2\right)\right. \\ &\qquad \left. + \Pr\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_z(Z_i)} \sum_{i=1}^n \mathbb{1}_z(Z_i) V_i^2 \geq L\right)\right],\end{aligned} \tag{16}$$

where [1] follows from the union bound, and [2] follows from the triangle inequality. I now consider each term separately.

Focusing on the first term, fix $\eta \geq 0$ such that $\eta \leq \min\{1, \tilde{\eta}\}$ where

$$\tilde{\eta} < \left(\frac{c}{2\sqrt{L} + C}\right)^2$$

with c defined by Assumption 8. Denote

$$\tilde{c}_{\tilde{g},g} \equiv 2\sqrt{\eta}\sqrt{L} + C\sqrt{\eta} - \frac{1}{2}\left(\alpha_{\tilde{g}}^0 - \alpha_g^0\right)^2.$$

Note that the choice of η above implies that $\tilde{c}_{\tilde{g},g} < 0$ for all combinations $\tilde{g} \neq g$. Further, fix

$\tilde{\lambda} > 0$ such that $\tilde{\lambda} < \lambda_z$ as defined by Assumption 3 and consider

$$
\begin{aligned}
&\Pr\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_z(Z_i)} \sum_{i=1}^n \mathbb{1}_z(Z_i) V_i(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq \tilde{c}_{\tilde{g},g}\right) \\
=&\Pr\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_z(Z_i)} \sum_{i=1}^n \mathbb{1}_z(Z_i) V_i(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq \tilde{c}_{\tilde{g},g} \middle| \sum_{i=1}^n \mathbb{1}_z(Z_i) > n^{\tilde{\lambda}}\right) \Pr\left(\sum_{i=1}^n \mathbb{1}_z(Z_i) > n^{\tilde{\lambda}}\right) \\
&+ \Pr\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_z(Z_i)} \sum_{i=1}^n \mathbb{1}_z(Z_i) V_i(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq \tilde{c}_{\tilde{g},g} \middle| \sum_{i=1}^n \mathbb{1}_z(Z_i) \leq n^{\tilde{\lambda}}\right) \Pr\left(\sum_{i=1}^n \mathbb{1}_z(Z_i) \leq n^{\tilde{\lambda}}\right) \\
\leq&\Pr\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_z(Z_i)} \sum_{i=1}^n \mathbb{1}_z(Z_i) V_i(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq \tilde{c}_{\tilde{g},g} \middle| \sum_{i=1}^n \mathbb{1}_z(Z_i) > n^{\tilde{\lambda}}\right) \\
&+ \Pr\left(\sum_{i=1}^n \mathbb{1}_z(Z_i) \leq n^{\tilde{\lambda}}\right),
\end{aligned}
\tag{17}
$$

where the inequality follows from probabilities being bounded by 1. For the first term, it holds for any $\delta > 0$ that

$$
\begin{aligned}
&\Pr\left(\frac{1}{\sum_{i=1}^n \mathbb{1}_z(Z_i)} \sum_{i=1}^n \mathbb{1}_z(Z_i) V_i(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq \tilde{c}_{\tilde{g},g} \middle| \sum_{i=1}^n \mathbb{1}_z(Z_i) > n^{\tilde{\lambda}}\right) \\
\overset{[1]}{=}&\Pr\left(\frac{1}{N_z} \sum_{i=1}^{N_z} V_{iz}(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq \tilde{c}_{\tilde{g},g} \middle| N_z > n^{\tilde{\lambda}}\right) \\
\overset{[2]}{\leq}&\Pr\left(\left|\frac{1}{N_z} \sum_{i=1}^{N_z} V_{iz}\right| \geq \frac{|\tilde{c}_{\tilde{g},g}|}{|\alpha_{\tilde{g}}^0 - \alpha_g^0|} \middle| N_z > n^{\tilde{\lambda}}\right) \\
\overset{[3]}{\leq}&\Pr\left(\left|\frac{1}{\lfloor n^{\tilde{\lambda}} \rfloor} \sum_{i=1}^{\lfloor n^{\tilde{\lambda}} \rfloor} V_{iz}\right| \geq \frac{|\tilde{c}_{\tilde{g},g}|}{|\alpha_{\tilde{g}}^0 - \alpha_g^0|}\right) \\
\overset{[4]}{=}&o(n^{-\delta}),
\end{aligned}
\tag{18}
$$

where [1] takes $V_{iz} \equiv (V_i | Z_i = z)$ and $N_z \equiv \sum_{i=1}^n \mathbb{1}_z(Z_i)$, [2] follows from $\tilde{c}_{\tilde{g},g} < 0$, and [3] follows from $\Pr(|\frac{1}{n}\sum_{i=1}^n V_{iz}| \geq b) \leq \Pr(|\frac{1}{\tilde{n}}\sum_{i=1}^{\tilde{n}} V_{iz}| \geq b), \forall \tilde{n} \leq n, b > 0$, and $N_z \perp\!\!\!\perp (\sum_{i=1}^{\lfloor n^{\tilde{\lambda}} \rfloor} V_{iz})$. Finally, [4] follows by application of Lemma B.5 in Bonhomme and Manresa (2015) where I take $T \equiv \lfloor n^{\tilde{\lambda}} \rfloor$, $z_t = V_{iz}$, and $z = \frac{|\tilde{c}_{\tilde{g},g}|}{|\alpha_{\tilde{g}}^0 - \alpha_g^0|}$.[15] The lemma applies by assumptions 6 and 10.

[15]Note that Lemma B.5 implies the rate $o(\lfloor n^{(\tilde{\lambda})} \rfloor^{-\delta})$, but since this holds for any $\delta > 0$ and $\tilde{\lambda} > 0$ is a fixed constant, the stated result follows.

To bound the second term in (17), I prove a simple concentration inequality in Lemma 4, whose application implies for any $\delta > 0$,

$$\Pr\left(\sum_{i=1}^{n} \mathbb{1}_z(Z_i) \leq n^{a_z \tilde{\lambda}}\right) = o(n^{-\delta}), \quad \forall z \in \operatorname{supp} Z. \tag{19}$$

Lemma 4. *Let X_n be a Binomial random variable with n trials and success probability $p_n = an^{\lambda-1}$ for fixed $a > 0$ and $\lambda \in (0,1)$. Then, $\forall \delta > 0$ and $\tilde{\lambda} > 0 : \lambda > \tilde{\lambda}$,*

$$\Pr\left(X_n \leq an^{\tilde{\lambda}}\right) = o(n^{-\delta}).$$

Proof. By Chernoff's inequality,

$$\begin{aligned}
&\Pr\left(X_n \leq an^{\tilde{\lambda}}\right) \\
&\leq \exp\left\{-n\left[an^{\tilde{\lambda}-1}\log\left(\frac{an^{\tilde{\lambda}-1}}{an^{\lambda-1}}\right) + \left(1 - an^{\tilde{\lambda}-1}\right)\log\left(\frac{1-an^{\tilde{\lambda}-1}}{1-an^{\lambda-1}}\right)\right]\right\} \\
&= \exp\left\{-n\left[-an^{\tilde{\lambda}-1}(\lambda-\tilde{\lambda})\log(n) + \left(1-an^{\tilde{\lambda}-1}\right)\left(\log\left(1-an^{\tilde{\lambda}-1}\right) - \log\left(1-an^{\lambda-1}\right)\right)\right]\right\}.
\end{aligned}$$

It is possible to bound the terms involving the logarithm via the following simple inequalities:

$$\begin{aligned}
\log(n) &\leq \gamma(n^{1/\gamma} - 1), \quad \forall n, \gamma > 0, \\
\frac{-x}{1-x} \leq \log(1-x) &\leq -x, \quad \forall x \in [0,1).
\end{aligned}$$

Therefore, fixing $\gamma > 0 : \lambda > \tilde{\lambda} + \frac{1}{\gamma}$,

$$\begin{aligned}
&\Pr\left(X_n \leq an^{\tilde{\lambda}}\right) \\
&\leq \exp\left\{-n\left[-an^{\tilde{\lambda}-1}(\lambda-\tilde{\lambda})\gamma\left(n^{1/\gamma}-1\right) + \left(1-an^{\tilde{\lambda}-1}\right)\left(an^{\lambda-1} - \frac{an^{\tilde{\lambda}-1}}{1-an^{\tilde{\lambda}-1}}\right)\right]\right\} \\
&= \exp\left\{-\left[an^{\lambda} + an^{\tilde{\lambda}}(\lambda-\tilde{\lambda})\gamma - an^{\tilde{\lambda}+1/\gamma}(\lambda+\tilde{\lambda})\gamma - a^2 n^{\lambda-\tilde{\lambda}-1} - an^{\tilde{\lambda}}\right]\right\} \\
&= \exp\left\{-n^{\lambda}\left[a + an^{\tilde{\lambda}-\lambda}(\lambda-\tilde{\lambda})\gamma - an^{\tilde{\lambda}+1/\gamma-\lambda}(\lambda-\tilde{\lambda})\gamma - a^2 n^{\tilde{\lambda}-1} - an^{\tilde{\lambda}-\lambda}\right]\right\},
\end{aligned}$$

where the term in brackets tends to $a > 0$ as $n \to \infty$. Finally, note that for any $\delta, \lambda > 0$,

$$\exp\left\{-n^{\lambda}\right\} = o(n^{-\delta}),$$

which completes the proof.

□

Combining (17)-(19) then implies for any $\delta > 0$,

$$\Pr\left(\frac{1}{\sum_{i=1}^{n} \mathbb{1}_z(Z_i)} \sum_{i=1}^{n} \mathbb{1}_z(Z_i) V_i(\alpha_{\tilde{g}}^0 - \alpha_g^0) \leq \tilde{c}_{\tilde{g},g}\right) = o(n^{-\delta}). \tag{20}$$

Focusing now on the second term in (16) and following similar arguments as before, we have for any $\delta > 0$

$$\begin{aligned}
&\Pr\left(\frac{1}{\sum_{i=1}^{n} \mathbb{1}_z(Z_i)} \sum_{i=1}^{n} \mathbb{1}_z(Z_i) V_i^2 \geq L\right) \\
\leq &\Pr\left(\frac{1}{\lfloor n^{\lambda-\tilde{\lambda}} \rfloor} \sum_{i=1}^{\lfloor n^{\lambda-\tilde{\lambda}} \rfloor} V_{iz}^2 \geq L\right) + \Pr\left(\sum_{i=1}^{n} \mathbb{1}_z(Z_i) \leq n^{\lambda-\tilde{\lambda}}\right) \\
\leq &\Pr\left(\left|\frac{1}{\lfloor n^{\lambda-\tilde{\lambda}} \rfloor} \sum_{i=1}^{\lfloor n^{\lambda-\tilde{\lambda}} \rfloor} V_{iz}^2 - E[V_{iz}^2]\right| \geq L - E[V_{iz}^2]\right) + o(n^{-\delta}).
\end{aligned}$$

Finally, we can again apply Lemma B.5 in Bonhomme and Manresa (2015) where I take $T \equiv \lfloor n^{\tilde{\lambda}} \rfloor$, $z_t = V_{iz}^2 - E[V_{iz}^2]$, and $z = L - E[V_{iz}^2]$. Note that $L - E[V_{iz}^2] > 0$ is implied by Assumption 5. As a consequence, for any $\delta > 0$

$$\Pr\left(\frac{1}{\sum_{i=1}^{n} \mathbb{1}_z(Z_i)} \sum_{i=1}^{n} \mathbb{1}_z(Z_i) V_i^2 \geq L\right) = o(n^{-\delta}). \tag{21}$$

Combining with (16) and (20) shows (15) and thus completes the proof.

□

A.2 Convergence of $\hat{\tau}_n$ to $\tilde{\tau}_n$

Lemma 5. *Let the assumptions of Theorem 1 hold. Then, for any $\delta > 0$,*

$$\hat{\tau}_n = \tilde{\tau}_n + o_p(n^{-\delta}),$$

where $\tilde{\tau}_n$ is the infeasible two stage least squares estimator defined in (5)*.*

Proof. We have

$$\begin{aligned}
&\hat{\tau}_n - \tilde{\tau}_n \\
=&\frac{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)\left(\hat{m}_n(Z_i) - \bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(\hat{m}_n(Z_i) - \bar{D}_n\right)} - \frac{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)\left(\tilde{m}_n(Z_i) - \bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(\tilde{m}_n(Z_i) - \bar{D}_n\right)} \\
=&\frac{\frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)\left(\hat{m}_n(Z_i) - \tilde{m}_n(Z_i)\right)}{\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(\hat{m}_n(Z_i) - \bar{D}_n\right)} + \left(\frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)\left(\tilde{m}_n(Z_i) - \bar{D}_n\right)\right) \\
&\quad \times \left[\left(\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(\hat{m}_n(Z_i) - \bar{D}_n\right)\right)^{-1} - \left(\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(\tilde{m}_n(Z_i) - \bar{D}_n\right)\right)^{-1}\right].
\end{aligned}$$

Note first that for the first term, it holds by Cauchy-Schwarz that

$$\frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)\left(\hat{m}_n(Z_i) - \tilde{m}_n(Z_i)\right) \leq \left(\frac{1}{n}\sum_{i=1}^n \left(Y_i - \bar{Y}_n\right)^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \left(\hat{m}_n(Z_i) - \tilde{m}_n(Z_i)\right)^2\right)^{1/2}.$$

Further,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(\hat{m}_n(Z_i) - \bar{D}_n\right) \\
=&\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(m_0(Z_i) - \bar{D}_n + \hat{m}_n(Z_i) - m_0(Z_i)\right) \\
\overset{[1]}{\leq}&\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)\left(m_0(Z_i) - \bar{D}_n\right) + \left(\frac{1}{n}\sum_{i=1}^n \left(D_i - \bar{D}_n\right)^2\right)^{1/2} \left(\frac{1}{n}\sum_{i=1}^n \left(\hat{m}_n(Z_i) - m_0(Z_i)\right)^2\right)^{1/2} \\
\overset{[2]}{=}&Var(E[D|Z]) + o_p(1)
\end{aligned} \tag{22}$$

where [1] follows from Cauchy-Schwarz, and [2] follows from Lemma 1, Assumption 5 which

implies $\frac{1}{n}\sum_{i=1}^{n}\left(D_i-\bar{D}_n\right)^2 = O_p(1)$, and the fact that by Assumption 10 and the weak law of large numbers $\frac{1}{n}\sum_{i=1}^{n}\left(m_0(Z_i)-\bar{D}_n\right)^2 = Var(E[D|Z]) + o_p(1)$. Because by Assumption 2 $Var(E[D|Z])$ is bounded away from zero, (22) implies by the continuous mapping theorem

$$\left[\frac{1}{n}\sum_{i=1}^{n}\left(D_i-\bar{D}_n\right)\left(\hat{m}_n(Z_i)-\bar{D}_n\right)\right]^{-1} = O_p(1).$$

Finally, using similar arguments as above, it holds by assumptions 5 and 9 and consistency of the infeasible least squares estimator $\tilde{m}_n$ under the conditions of Lemma 1 that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_n\right)^2 &= O_p(1),\\ \frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_n\right)\left(\tilde{m}_n(Z_i)-\bar{D}_n\right) &= O_p(1).\end{aligned}$$

Therefore, it suffices to show that for any $\delta > 0$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_n(Z_i)-\tilde{m}_n(Z_i)\right)^2 = o_p(n^{-\delta}) \tag{23}$$

and

$$\left(\hat{\sigma}_n^2\right)^{-1} - \left(\tilde{\sigma}_n^2\right)^{-1} = o_p(n^{-\delta}), \tag{24}$$

where

$$\hat{\sigma}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(D_i-\bar{D}_n\right)\left(\hat{m}_n(Z_i)-\bar{D}_n\right), \quad \text{and} \quad \tilde{\sigma}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\left(D_i-\bar{D}_n\right)\left(\tilde{m}_n(Z_i)-\bar{D}_n\right).$$

For this purpose, define

$$\bar{Q}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(D_i-\alpha_{g^0_{Z_i}}\right)^2, \qquad \text{and} \qquad \hat{Q}(\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(D_i-\alpha_{\hat{g}_{Z_i}(\alpha)}\right)^2. \tag{25}$$

Note that for η satisfying the condition of Lemma 3, it holds for any $\delta > 0$ that

$$
\begin{aligned}
&\sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \left| \bar{Q}(\alpha) - \hat{Q}(\alpha) \right| \\
&= \sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \Bigg| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} \left(D_i - \alpha_{\hat{g}_{Z_i}(\alpha)} \right)^2 \\
&\qquad\qquad + \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) = g^0_{Z_i}\} \left(D_i - \alpha_{\hat{g}_{Z_i}(\alpha)} \right)^2 - \frac{1}{n} \sum_{i=1}^{n} \left(D_i - \alpha_{g^0_{Z_i}} \right)^2 \Bigg| \\
&= \sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} \left(D_i - \alpha_{\hat{g}_{Z_i}(\alpha)} \right)^2 - \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} \left(D_i - \alpha_{g^0_{Z_i}} \right)^2 \right| \\
&= \sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} \left[\left(D_i - \alpha_{\hat{g}_{Z_i}(\alpha)} \right)^2 - \left(D_i - \alpha_{g^0_{Z_i}} \right)^2 \right] \right| \\
&\leq \sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} \left| \left(D_i - \alpha_{\hat{g}_{Z_i}(\alpha)} \right)^2 - \left(D_i - \alpha_{g^0_{Z_i}} \right)^2 \right| \\
&\overset{[1]}{\leq} \sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} \left| \left(D_i - \alpha_{\hat{g}_{Z_i}(\alpha)} \right)^2 - \left(D_i - \alpha_{g^0_{Z_i}} \right)^2 \right|^2 \right)^{1/2} \\
&\leq \left(\sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} \right)^{1/2} \left(\sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \frac{1}{n} \sum_{i=1}^{n} \left| \left(D_i - \alpha_{\hat{g}_{Z_i}(\alpha)} \right)^2 - \left(D_i - \alpha_{g^0_{Z_i}} \right)^2 \right|^2 \right)^{1/2} \\
&\overset{[2]}{=} o_p(n^{-\delta})
\end{aligned}
$$

where [1] follows from Cauchy-Schwarz, and [2] follows from Lemma 3 and the fact that the second term is $O_p(1)$ under Assumption 5.

Now, for both $\hat{\alpha}$ and the infeasible least squares coefficients $\tilde{\alpha}$, it holds for any $\delta > 0$ that

$$
\bar{Q}(\hat{\alpha}) - \hat{Q}(\hat{\alpha}) = o_p(n^{-\delta}), \qquad \bar{Q}(\tilde{\alpha}) - \hat{Q}(\tilde{\alpha}) = o_p(n^{-\delta}). \tag{26}
$$

To see this, fix $\varepsilon > 0$ and consider

$$\begin{aligned}
&\Pr\left(\left|\bar{Q}(\hat{\alpha}) - \hat{Q}(\hat{\alpha})\right| > \varepsilon n^{-\delta}\right) \\
\overset{[1]}{=}&\Pr\left(\left|\bar{Q}(\hat{\alpha}) - \hat{Q}(\hat{\alpha})\right| > \varepsilon n^{-\delta} \middle| \hat{\alpha} \in \mathcal{N}_{\alpha^0}(\eta)\right) \Pr\left(\hat{\alpha} \in \mathcal{N}_{\alpha^0}(\eta)\right) \\
&\quad + \Pr\left(\left|\bar{Q}(\hat{\alpha}) - \hat{Q}(\hat{\alpha})\right| > \varepsilon n^{-\delta} \middle| \hat{\alpha} \notin \mathcal{N}_{\alpha^0}(\eta)\right) \Pr\left(\hat{\alpha} \notin \mathcal{N}_{\alpha^0}(\eta)\right) \\
\overset{[2]}{\leq}&\Pr\left(\left|\bar{Q}(\hat{\alpha}) - \hat{Q}(\hat{\alpha})\right| > \varepsilon n^{-\delta} \middle| \hat{\alpha} \in \mathcal{N}_{\alpha^0}(\eta)\right) + \Pr\left(\hat{\alpha} \notin \mathcal{N}_{\alpha^0}(\eta)\right) \\
\overset{[3]}{\leq}&\Pr\left(\sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \left|\bar{Q}(\alpha) - \hat{Q}(\alpha)\right| > \varepsilon n^{-\delta}\right) + o(1) \\
\overset{[4]}{=}&o(1),
\end{aligned}$$

where [1] follows from the law of total probability, [2] follows from probabilities being bounded by one, [3] follows from consistency of $\hat{\alpha}$ by Lemma 2, and [4] follows from (26). The arguments for the infeasible least squares coefficients are analogous.

As a consequence,

$$0 \leq \bar{Q}(\hat{\alpha}) - \bar{Q}(\tilde{\alpha}) = \hat{Q}(\hat{\alpha}) - \hat{Q}(\tilde{\alpha}) + o_p(n^{-\delta}) \leq o_p(n^{-\delta}), \tag{27}$$

where the inequalities follow from the definition of $\hat{\alpha}$ and $\tilde{\alpha}$ (minimizing, $\hat{Q}$ and $\bar{Q}$, respectively), and the equality follows from (26).

Note further that

$$\begin{aligned}
&\bar{Q}(\hat{\alpha}) - \bar{Q}(\tilde{\alpha}) \\
=&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\alpha}_{g^0_{Z_i}} - \tilde{\alpha}_{g^0_{Z_i}}\right)^2 + \frac{2}{n}\sum_{i=1}^{n}\left(D_i - \tilde{\alpha}_{g^0_{Z_i}}\right)\left(\hat{\alpha}_{g^0_{Z_i}} - \tilde{\alpha}_{g^0_{Z_i}}\right) \\
=&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\alpha}_{g^0_{Z_i}} - \tilde{\alpha}_{g^0_{Z_i}}\right)^2,
\end{aligned} \tag{28}$$

where the equality follows from

$$\begin{aligned}
&\sum_{i=1}^{n}\left(D_i-\tilde{\alpha}_{g^0_{Z_i}}\right)\left(\hat{\alpha}_{g^0_{Z_i}}-\tilde{\alpha}_{g^0_{Z_i}}\right)\\
=&\sum_{i=1}^{n}\sum_{g=1}^{K_0}\mathbb{1}\{g^0(Z_i)=g\}\left(D_i-\tilde{\alpha}_{g^0_{Z_i}}\right)(\hat{\alpha}_g-\tilde{\alpha}_g)\\
=&\sum_{g=1}^{K_0}(\hat{\alpha}_g-\tilde{\alpha}_g)\sum_{i=1}^{n}\mathbb{1}\{g^0(Z_i)=g\}(D_i-\tilde{\alpha}_g)\\
=&\sum_{g=1}^{K_0}(\hat{\alpha}_g-\tilde{\alpha}_g)\times 0
\end{aligned}$$

because the infeasible least squares coefficients $(\tilde{\alpha}_g)_{g=1}^{K_0}$ correspond to the sample average of observations with $\mathbb{1}\{g^0(Z_i)=g\}$ for all $g\in\{1,\dots,K_0\}$. Combining (27) and (28) then implies

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\alpha}_{g^0_{Z_i}}-\tilde{\alpha}_{g^0_{Z_i}}\right)^2=o_p(n^{-\delta}). \tag{29}$$

Now, for any $\delta>0$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\hat{m}_n(Z_i)-\tilde{m}_n(Z_i))^2\\
\overset{[1]}{=}&\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\alpha}_{\hat{g}_{Z_i}(\hat{\alpha})}-\tilde{\alpha}_{g^0_{Z_i}}\right)^2\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{g}_{Z_i}(\hat{\alpha})\neq g^0_{Z_i}\}\left(\hat{\alpha}_{\hat{g}_{Z_i}(\hat{\alpha})}-\tilde{\alpha}_{g^0_{Z_i}}\right)^2+\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{g}_{Z_i}(\hat{\alpha})=g^0_{Z_i}\}\left(\hat{\alpha}_{g^0_{Z_i}}-\tilde{\alpha}_{g^0_{Z_i}}\right)^2\\
\overset{[2]}{\leq}&\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{g}_{Z_i}(\hat{\alpha})\neq g^0_{Z_i}\}\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\alpha}_{\hat{g}_{Z_i}(\hat{\alpha})}-\tilde{\alpha}_{g^0_{Z_i}}\right)^4\right)^{1/2}+o_p(n^{-\delta})
\end{aligned} \tag{30}$$

where [1] follows from the alternative representation of $\hat{m}_n$ and $\tilde{m}_n$, [2] follows from Cauchy-Schwarz and (29). By Assumption 9, $\frac{1}{n}\sum_{i=1}^{n}(\hat{\alpha}_{\hat{g}_{Z_i}(\hat{\alpha})}-\tilde{\alpha}_{g^0_{Z_i}})^4=O_p(1)$. Finally, taking η

satisfying the condition of Lemma 3 and fixing $\varepsilon > 0$, it holds that

$$\begin{aligned}
&\Pr\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{g}_{Z_i}(\hat{\alpha}) \neq g^0_{Z_i}\} > \varepsilon n^{-\delta}\right) \\
&\overset{[1]}{\leq} \Pr\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{g}_{Z_i}(\hat{\alpha}) \neq g^0_{Z_i}\} > \varepsilon n^{-\delta} | \hat{\alpha} \in \mathcal{N}_{\alpha^0}(\eta)\right) + \Pr\left(\hat{\alpha} \notin \mathcal{N}_{\alpha^0}(\eta)\right) \\
&\overset{[2]}{\leq} \Pr\left(\sup_{\alpha \in \mathcal{N}_{\alpha^0}(\eta)} \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\hat{g}_{Z_i}(\alpha) \neq g^0_{Z_i}\} > \varepsilon n^{-\delta}\right) + o_p(1) \\
&\overset{[3]}{\leq} o_p(1),
\end{aligned}$$

where [1] follows from the law of total probability and bounding probabilities by one, [2] follows from Lemma 2, and [3] follows from Lemma 3. Combining with (30) then implies (23).

I conclude the proof by showing (24). For this purpose, fix $\varepsilon > 0$ and take $Var(E[D|Z]) > \tilde{\eta} > 0$. Consider

$$\begin{aligned}
&\Pr\left(\left|\left(\hat{\sigma}_n^2\right)^{-1} - \left(\tilde{\sigma}_n^2\right)^{-1}\right| > ^{-\delta}\right) \\
&\overset{[1]}{\leq} \Pr\left(\left|\left(\hat{\sigma}_n^2\right)^{-1} - \left(\tilde{\sigma}_n^2\right)^{-1}\right| > ^{-\delta} \left|\hat{\sigma}_n^2, \tilde{\sigma}_n^2 \in \mathcal{N}_{Var(E[D|Z])}(\tilde{\eta})\right.\right) + \Pr\left(\hat{\sigma}_n^2, \tilde{\sigma}_n^2 \in \mathcal{N}_{Var(E[D|Z])}(\tilde{\eta})\right) \\
&\overset{[2]}{\leq} \Pr\left(\left|\hat{\sigma}_n^2 - \tilde{\sigma}_n^2\right| > \frac{\varepsilon}{\tilde{L}} n^{-\delta} \middle| \hat{\sigma}_n^2, \tilde{\sigma}_n^2 \in \mathcal{N}_{Var(E[D|Z])}(\tilde{\eta})\right) + o(1),
\end{aligned}$$

where [1] follows again from the law of total probability and bounding probabilities by one, and [2] follows from Lipschitz continuity with Lipschitz constant $0 < \tilde{L} < \infty$ of $1/x$ when x is bounded away from zero as well as consistency of $\hat{\sigma}_n^2$ by (22) and consistency of $\tilde{\sigma}_n^2$ by the same arguments. Finally,

$$\begin{aligned}
\left|\hat{\sigma}_n^2 - \tilde{\sigma}_n^2\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}\left(D_i - \bar{D}_n\right)\left(\hat{m}_n(Z_i) - \tilde{m}_n(Z_i)\right)\right| \\
&\overset{[1]}{\leq} \left(\frac{1}{n}\sum_{i=1}^{n}\left(D_i - \bar{D}_n\right)^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}_n(Z_i) - \tilde{m}_n(Z_i)\right)^2\right)^{1/2} \\
&\overset{[2]}{=} o_p(n^{-\delta}),
\end{aligned}$$

where [1] follows from Cauchy-Schwarz, and [2] follows from (23) and $\frac{1}{n}\sum_{i=1}^{n}\left(D_i-\bar{D}_n\right)^2 = O_p(1)$ by Assumption 5. This concludes the proof. □

A.3 Asymptotic distribution of $\hat{\tau}_n$

To finish the proof of Theorem 1, note that by application of Lemma 5, it holds for any $\delta > 0$ that

$$\sqrt{n}\left(\hat{\tau}_n - \tau_0\right) = \sqrt{n}\left(\tilde{\tau}_n - \tau_0\right) + o_p(n^{-\delta}). \tag{31}$$

It thus suffices to study the asymptotic distribution of the infeasible two stage least squares coefficient $\tilde{\tau}_n$. In particular,

$$\begin{aligned}\sqrt{n}\left(\tilde{\tau}_n - \tau_0\right) =& \sqrt{n}\left[\frac{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\bar{Y}_n\right)\left(\tilde{m}_n(Z_i)-\bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(D_i-\bar{D}_n\right)\left(\tilde{m}_n(Z_i)-\bar{D}_n\right)} - \tau_0\right] \\ =& \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(U_i-\bar{U}_n\right)\left(\tilde{m}_n(Z_i)-\bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(D_i-\bar{D}_n\right)\left(\tilde{m}_n(Z_i)-\bar{D}_n\right)}.\end{aligned}$$

Focusing on the numerator,

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(U_i-\bar{U}_n\right)\left(\tilde{m}_n(Z_i)-\bar{D}_n\right) \\ &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(U_i-\bar{U}_n\right)\left(m_0(Z_i)-\bar{D}_n\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(U_i-\bar{U}_n\right)\left(\tilde{m}_n(Z_i)-m_0(Z_i)\right) \\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(U_i-\bar{U}_n\right)\left(m_0(Z_i)-\bar{D}_n\right) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{g=1}^{K_0}\mathbb{1}\{g_{Z_i}^0 = g\}\left(U_i-\bar{U}_n\right)\left(\tilde{\alpha}_g-\alpha_g^0\right) \\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(U_i-\bar{U}_n\right)\left(m_0(Z_i)-\bar{D}_n\right) + \sum_{g=1}^{K_0}\left(\tilde{\alpha}_g-\alpha_g^0\right)\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbb{1}\{g_{Z_i}^0 = g\}\left(U_i-\bar{U}_n\right)\right] \\ =&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(U_i-\bar{U}_n\right)\left(m_0(Z_i)-\bar{D}_n\right) + o_p(1)\end{aligned}$$

By standard application of central limit theorem and Slutsky's lemma under assumptions 2,

5, and 10, it thus holds that

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(U_i-\bar{U}_n\right)\left(\tilde{m}_n(Z_i)-\bar{D}_n\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(D_i-\bar{D}_n\right)\left(\tilde{m}_n(Z_i)-\bar{D}_n\right)} \xrightarrow{d} N\left(0,\sigma^2\right), \quad \sigma^2=\frac{Var(m_0(Z)U)}{Var(m_0(Z))^2}.$$

Further, when U is homoskedastic, $\sigma^2 = Var(U)/Var(m_0(Z))$ which is the semiparametric efficiency bound (Chamberlain, 1987). In combination with (31), this completes the proof of Theorem 1.

B Implementation Details

This appendix provides pseudo-code for the optimal instrumental variable estimator via K-Means. In particular, as in Bonhomme and Manresa (2015), I consider LLoyd's algorithm, which proceeds in two steps. First, given a vector of coefficients α, we can estimate an associated combination of categories as

$$\hat{g}_z = \underset{g \in \{1,\dots,K_0\}}{\arg\min} \sum_{i:\, Z_i = z} \left(D_i - \alpha_g\right)^2. \tag{32}$$

Given a partition defined by $(g_z)_{z \in \operatorname{supp} Z}$, it is possible to estimate a vector of coefficients as

$$\hat{\alpha} = \underset{\alpha \in \mathcal{M}^{K_0}}{\arg\min} \sum_{i=1}^{n} \left(D_i - \alpha_{g_{Z_i}}\right)^2. \tag{33}$$

This suggests a simple iterative algorithm for computing the optimal instrument estimator where equation (4) and (5) are repeatedly computed in turn. Algorithm 1 presents this first computational approach. Note that while computationally attractive, the procedure does not guarantee convergence to a local optima.

Algorithm 1 LLoyd's Algorithm for Optimal Instrument Estimation

Required input: $\mathbf{D} \in \mathbb{R}^n$ endogenous variable vector, $\mathbf{Z} \in (\operatorname{supp} Z)^n$ feature vector, K_0 cardinality of the support of the optimal instrument, $\alpha^{(0)} \in \mathcal{M}^K$ initial set of conditional means;

procedure LLoyd
  $\hat{\alpha} \leftarrow \alpha^{(0)}$
  repeat
    for $z \in \operatorname{supp} Z$ **do**
      $\hat{g}_z \leftarrow \arg\min_{k \in \{1,\dots,K_0\}} \sum_{i:\, Z_i = z} \left(D_i - \hat{\alpha}_g\right)^2$ ▷ Infer partition
    $\hat{\alpha} \leftarrow \arg\min_{\alpha \in \mathcal{M}^{K_0}} \sum_{i=1}^{n} \left(D_i - \alpha_{\hat{g}_{Z_i}}\right)^2$ ▷ Compute conditional means
  until convergence

Return: $(\hat{g}_z)_{z \in \operatorname{supp} Z}$ estimated partition, $\hat{\alpha} \in \mathcal{M}^{K_0}$ estimated coefficients

As an alternative to LLoyd's algorithm, it is also possible to adapt the variable neighborhood search algorithm of Hansen et al. (2010). The pseudocode for the adapted variable search neighborhood search algorithm for optimal instrument estimation is presented in Algorithm

2. It combines local search with random perturbations to the partition of $\operatorname{supp} Z$.

Algorithm 2 Adapted Variable Neighborhood Search (VNS)

Required input: $\mathbf{D} \in \mathbb{R}^n$ endogenous variable vector, $\mathbf{Z} \in (\operatorname{supp} Z)^n$ feature vector, K_0 cardinality of the support of the optimal instrument, $\alpha^{(0)} \in \mathcal{M}^K$ initial set of conditional means; $\overline{N}$ maximum number of neighborhood reassignments;

procedure VNS
  $\alpha \leftarrow \alpha^{(0)}$
  for $z \in \operatorname{supp} Z$ **do**
    $\hat{g}_z \leftarrow \arg\min_{g \in \{1,\dots,K_0\}} \sum_{i:\, Z_i = v} (D_i - \hat{\alpha}_g)^2$ ▷ Infer initial partition
  repeat
    for $\overline{n} \in \{1, \dots, \overline{N}\}$ **do**
      $(\tilde{g})_{g=1}^{K_0} \leftarrow$ Perturb$(\hat{g}, \overline{n})$ ▷ Randomly reassign $\overline{n}$ values of z
      $\tilde{\alpha}^{(0)} \leftarrow \arg\min_{\alpha \in \mathcal{M}^{K_0}} \sum_{i=1}^n \left(D_i - \alpha_{\tilde{g}_{Z_i}}\right)^2$ ▷ Compute conditional means
      $\tilde{g}, \tilde{\alpha} \leftarrow$ Lloyd$(\mathbf{D}, \mathbf{Z}, K_0, \tilde{\alpha}^{(0)})$ ▷ Compute Lloyd
      repeat ▷ Search for local optimum
        for $z \in \operatorname{supp} Z$ **do**
          for $g \in \{1, \dots, K_0\} \setminus \tilde{g}_z$ **do**
            $(\tilde{g}')_{g=1}^{K_0} \leftarrow (\tilde{g})_{g=1}^{K_0}$
            $\tilde{g}'_z \leftarrow g$ ▷ Reassign z to cluster g
            $\tilde{\alpha}' \leftarrow \arg\min_{\alpha \in \mathcal{M}^{K_0}} \sum_{i=1}^n \left(D_i - \alpha_{\tilde{g}'_{Z_i}}\right)^2$
            if $\sum_{i=1}^n \left(D_i - \tilde{\alpha}_{\tilde{g}_{Z_i}}\right)^2 > \sum_{i=1}^n \left(D_i - \tilde{\alpha}'_{\tilde{g}'_{Z_i}}\right)^2$ **then**
              $(\tilde{g})_{g=1}^{K_0}, \tilde{\alpha} \leftarrow (\tilde{g}')_{g=1}^{K_0}, \tilde{\alpha}'$ ▷ Update conditional on improvement
      until convergence
      if $\sum_{i=1}^n \left(D_i - \hat{\alpha}_{\hat{g}_{Z_i}}\right)^2 > \sum_{i=1}^n \left(D_i - \tilde{\alpha}_{\tilde{g}_{Z_i}}\right)^2$ **then**
        $(\hat{g})_{g=1}^{K_0}, \hat{\alpha} \leftarrow (\tilde{g})_{g=1}^{K_0}, \tilde{\alpha}$ ▷ Update conditional on improvement
        Break ▷ Break out of for-loop to reset $\overline{n} = 1$
  until convergence

Return: $(\hat{g}_z)_{z \in \operatorname{supp} Z}$ estimated partition, $\hat{\alpha} \in \mathcal{M}^{K_0}$ estimated coefficients

Both procedures require an initial guess of the values of the optimal instrument $\alpha^{(0)}$ on which convergence to a global minimum depends. Bonhomme and Manresa (2015) address this by considering a large number of random initializations and selecting the in-sample loss-minimizing estimator. For further improvement, it is also possible to consider an adaptation of the KMeans++ initialization (Arthur and Vassilvitskii, 2006). Algorithm 3 presents the corresponding pseudo code.

Algorithm 3 ADAPTED KMEANS++ INITIALIZATION

Required input: $\mathbf{D} \in \mathbb{R}^n$ endogenous variable vector, $\mathbf{Z} \in (\operatorname{supp} Z)^n$ feature vector, K_0 cardinality of the support of the optimal instrument;

procedure ADAPTED KMEANS++ INITIALIZATION

$\mathcal{C} \leftarrow \emptyset$ ▷ Initialize centers

$\alpha^{(0)} \leftarrow \emptyset$ ▷ Initialize conditional means

$R \leftarrow \frac{1}{K_z} \iota_{K_z}$ ▷ Initialize with equal distances to centers

for $g \in \{1, \ldots, K_0\}$ **do**

$z \sim \mathcal{P}\left(\operatorname{supp} Z \setminus \mathcal{C}; \left\{\frac{r_z^2}{\sum r_{\tilde{z}}^2}\right\}_{z=1}^{K_z-(K_0-1)}\right)$ ▷ Draw new center z w.p. $\frac{r_z^2}{\sum r_{\tilde{z}}^2}$

$\mathcal{C} \leftarrow \mathcal{C} \cup \{z\}$

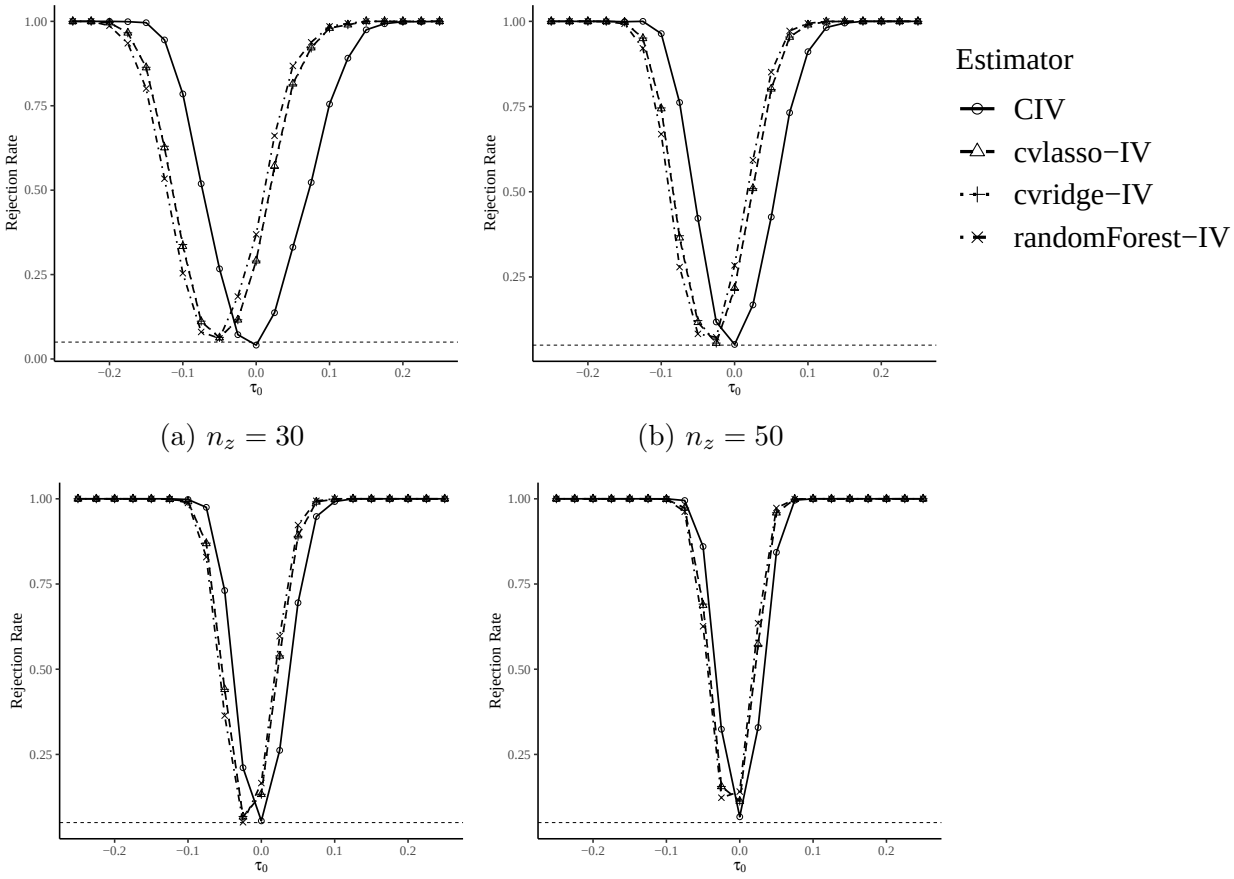
$\alpha^{(0)} \leftarrow \left[\alpha^{(0)}, \frac{\sum_{i:\, Z_i = z} D_i}{|\{i : Z_i = z\}|}\right]$ ▷ Add new conditional mean

Return: $\alpha^{(0)} \in \mathcal{M}^{K_0}$ starting values for coefficient vector of the optimal instruments

C Additional Simulation Results

This appendix provides simulation results complementary to the results in Section 4. In particular, Figure 4 provides results for the three additional machine learning-based IV estimators for the same DGP as Figure 3 in the main text. Further, Figures 5 and 6 present results where the optimal instrument has four support points (i.e., $K_0 = 4$). The qualitative results are unaffected.

Figure 4: Power Curves ($K_0 = 2$, $p = 2$)

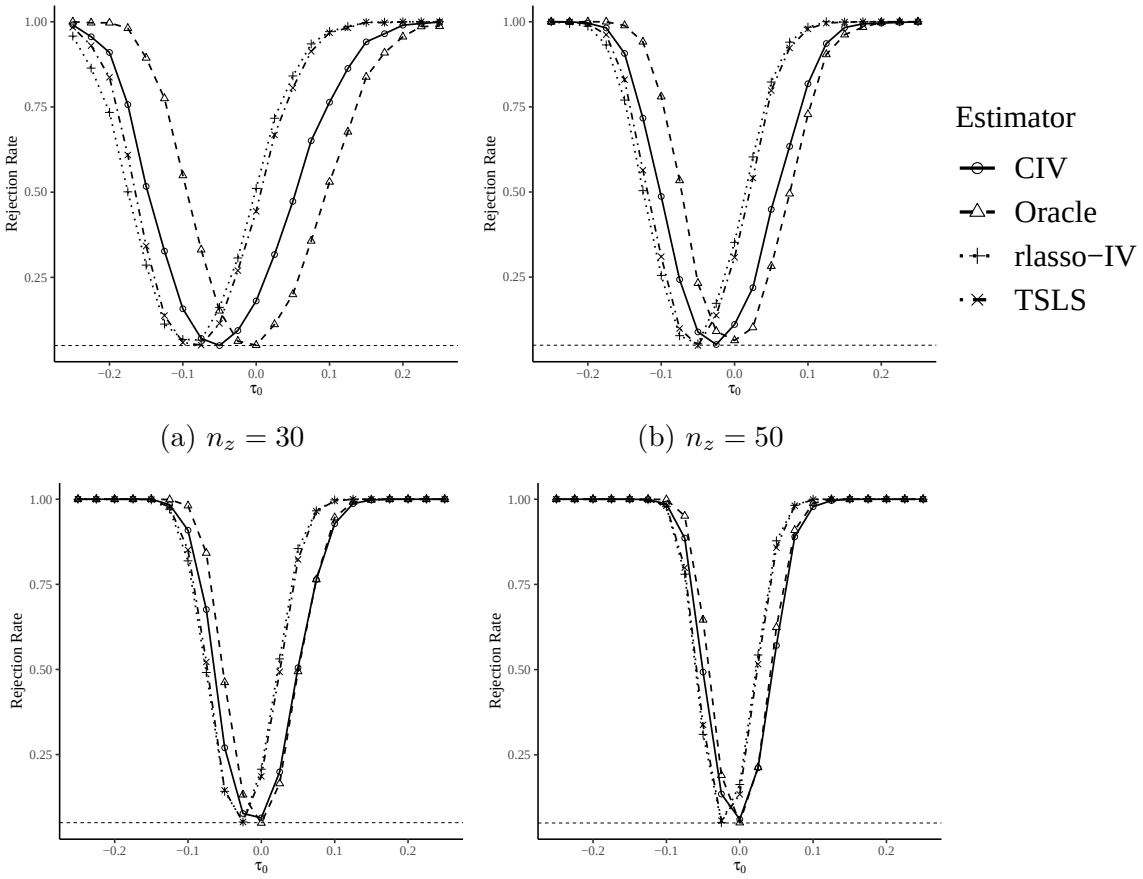


(a) $n_z = 30$ (b) $n_z = 50$

(c) $n_z = 100$ (d) $n_z = 150$

Notes. Simulation results are based on 1,000 simulations with $K_0 = 2$ and $p = 1$. The power curves plot the rejection rate of testing $H_0 : \tau_0 = 0$ at significance level $\alpha = 0.05$ as a function of of the true coefficient τ_0. CIV denotes the categorical IV estimator with known K_0. cvlasso-IV and cvridge-IV denotes IV estimators that compute first stage using lasso or ridge regression, respectively, where the shrinkage parameter is determined using 10-fold cross-validation. randomForest-IV denotes an IV estimator that compute the first stage using a random forest. Standard errors used for testing are heteroskedasticity robust.

Figure 5: Power Curves ($K_0 = 4$, $p = 2$)

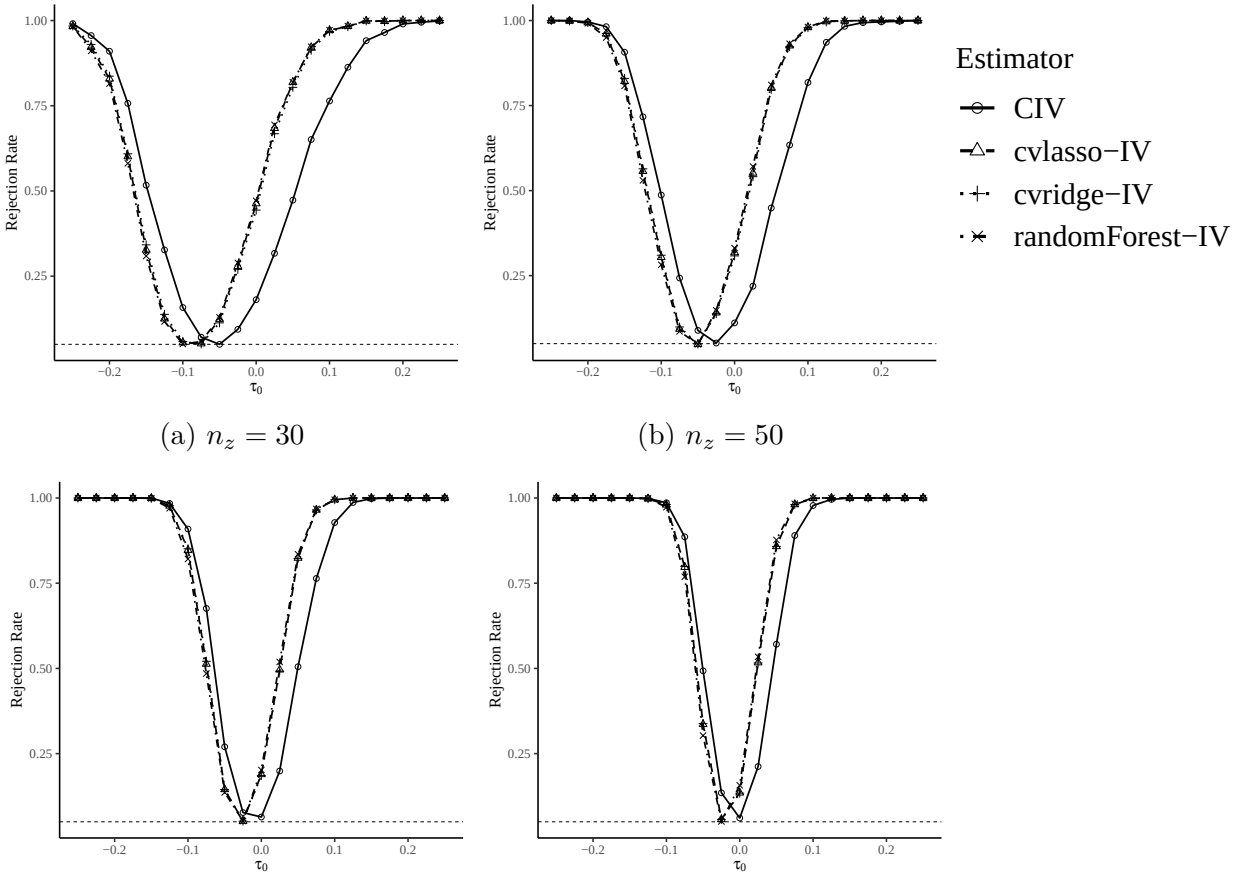


(a) $n_z = 30$ (b) $n_z = 50$

(c) $n_z = 100$ (d) $n_z = 150$

Notes. Simulation results are based on 1,000 simulations with $K_0 = 4$ and $p = 2$. The power curves plot the rejection rate of testing $H_0 : \tau_0 = 0$ at significance level $\alpha = 0.05$ as a function of of the true coefficient τ_0. CIV denotes the categorical IV estimator with known K_0. Oracle denotes the infeasible two stage least squares (TSLS) estimator that presumes knowledge of the optimal instruments. rlasso-IV denotes the post-lasso IV estimator as proposed in Belloni et al. (2012). TSLS denotes the feasible TSLS estimator that uses the observed categorical instruments. Standard errors used for testing are heteroskedasticity robust.

Figure 6: Power Curves ($K_0 = 4$, $p = 2$)



(a) $n_z = 30$ (b) $n_z = 50$

(c) $n_z = 100$ (d) $n_z = 150$

Notes. Simulation results are based on 1,000 simulations with $K_0 = 4$ and $p = 2$. The power curves plot the rejection rate of testing $H_0 : \tau_0 = 0$ at significance level $\alpha = 0.05$ as a function of of the true coefficient τ_0. Estimators are the same as those in Figure 4. Standard errors used for testing are heteroskedasticity robust.