Accepted at *Management Science*

Online Appendix for "EnsembleIV: Creating Instrumental Variables from Ensemble Learners for Robust Statistical Inference with ML Generated Variables"

Gordon Burtch
Questrom School of Business, Boston University, Boston, MA, 02215

Edward McFowland III
Harvard Business School, Harvard University, Boston, MA, 02163

Mochen Yang
Minnesota Carlson, University of Minnesota, Minneapolis, MN, 55455

Gediminas Adomavicius
Minnesota Carlson, University of Minnesota, Minneapolis, MN, 55455

Appendix A: Descriptive Evidence of IV Transformation and IV selection Effectiveness

Using both the Bike Sharing and Bank Marketing datasets, we first provide descriptive evidence to demonstrate that the IV transformation step and IV selection step (using LASSO selection for illustration) can indeed create (approximately) valid instruments and select strong ones. In particular, for a given mismeasured covariate $\widehat{X}^{(i)}$, denote $\Psi^{(i)}_{before} = \{\widehat{X}^{(j)}\}_{j \neq i}$ as the set of candidate IVs before any transformation or selection, and denote $\Psi^{(i)}_{after} \subseteq \{\widetilde{Z}^{(j)}\}_{j \neq i}$ as the set of transformed IVs selected by the LASSO step. We then compute $\frac{1}{|\Psi^{(i)}_{\bullet}|} \sum_{z \in \Psi^{(i)}_{\bullet}} |Corr(\widehat{X}^{(i)}, z)|$ to measure the average relevance of selected IVs, and $\frac{1}{|\Psi^{(i)}_{\bullet}|} \sum_{z \in \Psi^{(i)}_{\bullet}} |Corr(\widehat{X}^{(i)} - X, z)|$ to measure the average exclusion of selected IVs. In Figure 1, we plot the distribution of the average relevance and exclusion measures $\forall i \in \{1, \ldots, 100\}$ and across all simulation runs, both before and after the transformation and selection steps.

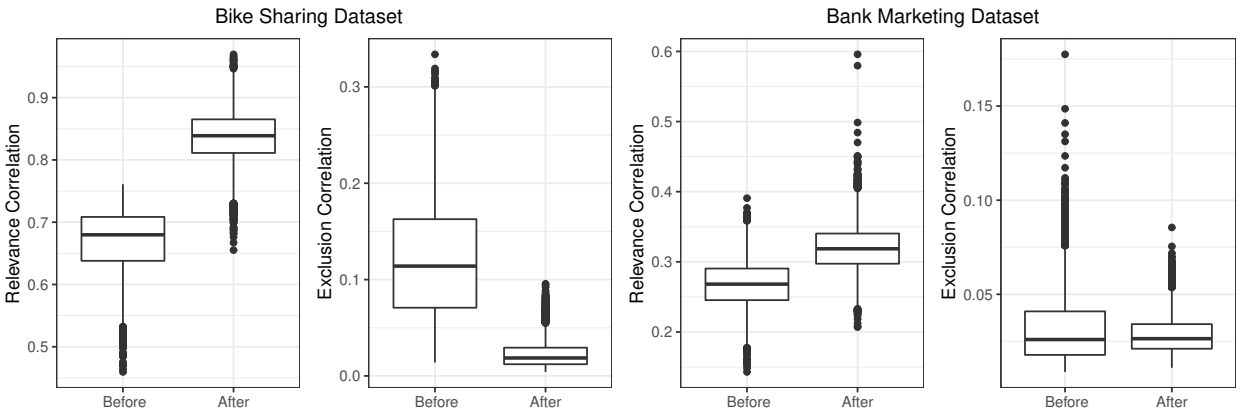

Figure 1 Average Relevance/Exclusion Measures Before and After IV Transformation and Selection

On the Bike Sharing dataset, the average relevance and exclusion measures clearly increase and decrease, respectively, after the IV transformation and selection steps. On the Bank Marketing dataset, the average relevance increase is also clear, whereas the average exclusion measure is already small prior to the IV transformation, though its distribution becomes even narrower after the transformation. Overall, these results demonstrate the good performance of IV transformation and selection steps. That said, one can see that the average value of the exclusion measure is not precisely 0 on either dataset, which indicates that even after transformation, the IVs are almost but not perfectly valid. This further emphasizes the practical importance of selecting strong IVs for estimation.[1]

Appendix B: Validity of Bootstrapped Standard Errors

In this appendix, we provide empirical evidence to support the validity of our bootstrapping procedure to approximate standard errors for EnsembleIV estimates. We do so by showing that the confidence intervals constructed using the bootstrapped standard errors have proper coverage rates. Calculating the coverage rate requires knowledge of the true coefficient value, so we rely on simulation experiments on the Bike Sharing and Bank Marketing datasets. We systematically explore 12 configurations – 2 datasets $\times$ 2 regression specifications (linear and logit) $\times$ 3 IV selection approaches (top-3, PCA, and LASSO). Under each configuration, we follow our proposed bootstrapping procedure, i.e., fixing a single set of labeled / unlabeled partitions and bootstrapping each partition over 100 iterations to obtain 100 sets of EnsembleIV coefficient estimates. The

[1] As Murray (Murray 2006, pg. 128) notes: "Strong and almost valid instruments do tend to bias two-stage least squares only a little. However, weak instruments that are almost valid bias two-stage least squares markedly more than do their strong counterparts."

bootstrapped standard errors are then calculated as the standard deviation over the 100 set of coefficient estimates.

As a first step, we conduct the Kolmogorov-Smirnov (K-S) normality test (Dallal and Wilkinson 1986) to understand the distributional characteristics of the bootstrapped coefficient estimates. We fail to reject the null of normality at 5% significance level across all 12 configurations (minimum p-value $= 0.073$, most p-values range from 0.31 to 0.98). This provides empirical support that the bootstrapped coefficient estimates are approximately normally distributed, and therefore allows us to use the 1.96 critical value to construct 95% confidence intervals and compute coverage rates.

Specifically, let $\{\widehat{\beta_{MLV}}^{(i)}\}_{i=1}^{100}$ denote the coefficient estimates on the ML-generated variable across 100 bootstrapping rounds and let SD denotes their standard deviation (as the standard error estimate). We construct 100 confidence intervals $\left[\widehat{\beta_{MLV}}^{(i)} - 1.96SD, \widehat{\beta_{MLV}}^{(i)} + 1.96SD\right]_{i=1}^{100}$ and calculate the coverage rate as fraction of intervals that contain the true coefficient value 0.5.

Importantly, to obtain a proper coverage rate (i.e., close to 0.95), both the point estimate and standard error need to be close to their true values. On the Bike Sharing data and across different configurations, the point estimates are generally quite close to the true coefficient value (see Table 3 in main manuscript). This enables us to cleanly test the validity of the bootstrapped standard error (i.e., if the coverage rates substantially differ from 0.95 on this dataset, it is due to having invalid standard errors). In contrast, on the Bank Marketing dataset, point estimates still deviate from true coefficient value to some degree (recall that EnsembleIV *reduces* bias but is not guaranteed to *completely eliminates* it). As a result, the Bank Marketing data is not a good testbed for the validity of the bootstrapped standard error.

Across the different simulation configurations of the Bike Sharing data, we find that the empirical coverage rates are all fairly close to 0.95, thereby lending support to the validity of our bootstrapping procedure. The detailed coverage rates are reported in the following Table 1.

Table 1 EnsembleIV Bootstrapping Coverage Rates on Bike Sharing Data

	Linear Second Phase Regression			Logistic Second Phase Regression		
	Top-3	PCA	LASSO	Top-3	PCA	LASSO
β_{MLV}	0.95	0.95	0.91	0.96	0.95	0.95

Appendix C: Additional Simulations of Diagnostic Procedure

We simulate the following data generation process:

$$Y = 1 + X + 0.5 \times W + \varepsilon \tag{1}$$

where $X \sim N(0,1)$ represents the ground truth values, $W \sim Uniform[0,1]$ represents an exogenous control variable. Importantly, the idiosyncratic error term, ε, is simulated as the sum of a "peripheral feature" and some exogenous error:

$$\varepsilon = \underbrace{e_1 + e_2 + \mu}_{\text{Peripheral Feature}} + \underbrace{\tau}_{\text{Exogenous Error}} \tag{2}$$

where $e_1, e_2 \sim N(0, \sigma^2)$, $\mu \sim N(0, 0.2^2)$, and $\tau \sim N(0,1)$. We next simulate two mismeasured versions of X (serving as ML-generated variables):

$$\begin{aligned} X_1 &= X + e_1 + e \\ X_2 &= X + e_2 + e \end{aligned} \tag{3}$$

where $e \sim N(0, 0.1^2)$ is a common error component in both variables, which implies that one is only an imperfect instrument for the other (therefore warranting IV transformation). Further, because error components e_1 and e_2 are present in both X_1, X_2 and the peripheral variable, we have introduced correlations between the prediction errors of X_1, X_2 and the peripheral variable. By changing the value of σ, we can alter the strength of the correlations. In the following simulations, we will treat X_1 as the endogenous variable and use X_2 as the (imperfect) candidate instrument to carry out IV transformation and IV regression. We follow the same permutation test discussed above (with 10,000 permutation runs) to test for correlation between the transformed IV and the residual term.

In the first set of simulations, we generate a sample of 5,000 data points and randomly partition it into 3,000, 1,000, and 1,000 observations, for training, testing, and diagnostic, respectively. We vary σ across 20 values along the grid $\{0.02, 0.04, \ldots, 0.4\}$ and, under each choice of σ, repeat the simulations 500 times. We report (i) the correlation between the instrument's prediction error and peripheral variable (calculated based on our diagnostic procedure) both before and after IV transformation (averaged over 500 repetitions); (ii) statistical power – the rate of rejecting the null of zero correlation (i.e., among 500 repetitions, the proportion of simulation runs that return a rejection result); and (iii) the mean and empirical 95% confidence interval of the IV estimate (again, obtained over 500 repetitions). The results are presented in Figure 2.

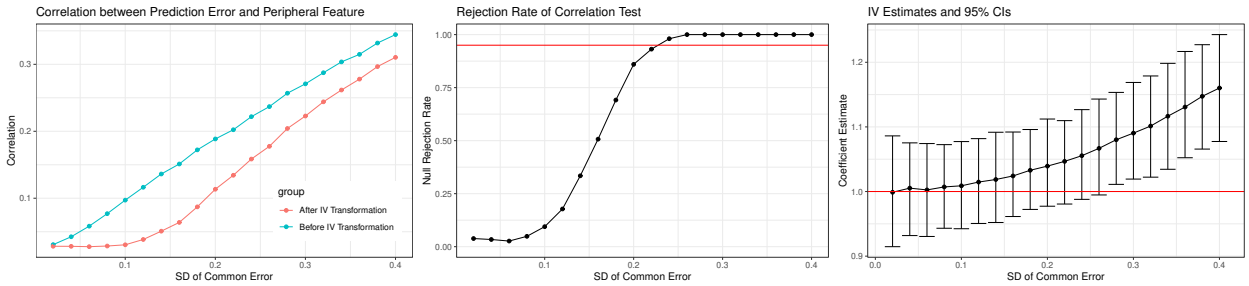

Figure 2 Simulation Results with Varying Strength of Correlation (red line in the second plot indicates 0.95 level; red line in the third plot indicates true coefficient value; results obtained over 500 repetitions)

Across all values of σ, our IV transformation step reduces the correlation between prediction error and peripheral feature (statistically significant with $p < 0.001$). When the correlation before transformation is relatively weak (e.g., $\sigma \leq 0.1$, we see that the post-transformation correlation is very close to 0 (which implies a valid IV). In contrast, when the correlation is sufficiently strong (e.g., $\sigma \geq 0.3$) the subsequent bias due to the peripheral feature challenge noticeably impacts IV estimates. Fortunately, we see that our diagnostic test is able to properly detect the issue by rejecting the null of zero correlation. We also observe that when there is a moderate level of correlation (e.g., $\sigma \in (0.24, 0.28)$), our diagnostic test also correctly rejects the null of zero correlation even though the IV estimates' confidence intervals still cover the true value in the presence of such non-zero correlation. We suspect these large confidence intervals are a byproduct of the considerable variation in the data-generating process and subsequent uncertainty in the coefficient estimates at smaller sample sizes. Therefore, we conduct a second set of simulations where we fix $\sigma = 0.24$ and vary the total sample size across 16 values along the grid $\{5000, 6000 \ldots, 20000\}$. Under each sample size, we keep the relative proportion of training / testing / diagnostic data to $3:1:1$ and repeat the above analyses. The results, presented in Figure 3, show that with larger sample sizes, the estimation bias in IV estimates becomes more evident and the 95% confidence intervals no longer cover the true value. These two sets of simulations confirm that our diagnostic procedure is reliable, possessing sufficient power against the null of zero correlation, even at moderate levels of correlation.

Appendix D: Extension of EnsembleIV to Account for Peripheral Feature Correlation

If the diagnostic procedure indicates a significant correlation between prediction errors and peripheral features, then the original EnsembleIV approach would likely yield inconsistent estimates. To deal with this scenario, we next consider an extension to EnsembleIV that explicitly accounts for the potential correlation

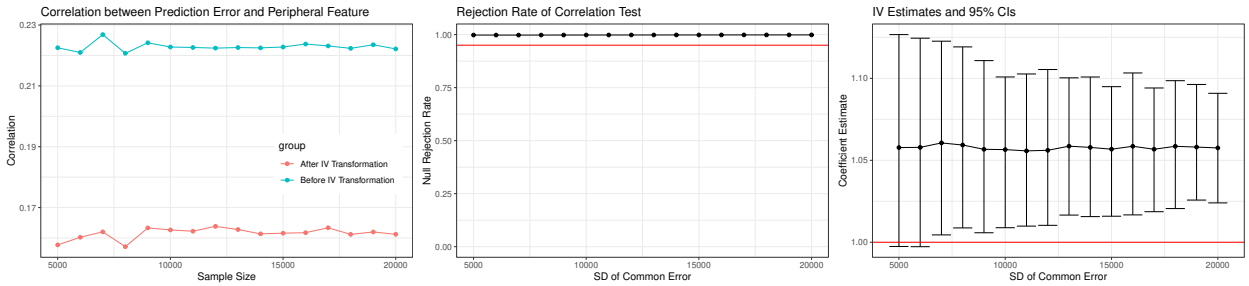

Figure 3 Simulation Results of Varying Sample Size ($\sigma = 0.24$; red line in the second plot indicates 0.95 level; red line in the third plot indicates true coefficient value; results obtained over 500 repetitions)

between prediction errors and peripheral features. Recall the calculation of the λ parameter used in IV transformation:

$$\lambda = \frac{\rho_{Zu}}{\rho_{\widehat{X}u}} = \frac{Cov(Z,u)}{Cov(\widehat{X},u)} \cdot \frac{\sigma_{\widehat{X}}}{\sigma_Z} = \frac{Cov(Z,\varepsilon) - \beta Cov(Z,e)}{Cov(\widehat{X},\varepsilon) - \beta Cov(\widehat{X},e)} \cdot \frac{\sigma_{\widehat{X}}}{\sigma_Z} \tag{4}$$

In the presence of peripheral feature correlation, $Cov(Z,\varepsilon)$ and $Cov(\widehat{X},\varepsilon)$ are not zero, but they can still be estimated using D_{test}. Similar to the diagnostic procedure, we estimate the unbiased first-phase regression on D_{train}, obtain the residuals on D_{test} as a proxy for ε, then compute the empirical correlations between said residuals and individual learner predictions on D_{test}. The unknown β coefficient can be approximated by the sample estimate on the full labeled data, jointly on D_{train} and D_{test} (i.e., based on the unbiased regression). Together, we calculate a *modified* λ value that accounts for the peripheral feature correlations, which can then be used to transform the corresponding IV on $D_{unlabel}$. Please note that cross-fitting can also be applied here, in the same way that it is applied to the original EnsembleIV approach, enabling the full labeled data to eventually be used in estimating λ.

We apply this approach to the aforementioned simulation. We specifically focus on $\sigma \in \{0.1, 0.12, \ldots, 0.4\}$ because they represent situations where our original (unmodified) IV transformation step fails to create valid IVs (as shown in the first plot of Figure 2). Under each value of σ, we simulate a sample of 14,000 data points and randomly partition it into 3,000 as D_{train}, 1,000 as D_{test}, and 10,000 as $D_{unlabel}$. As before, we treat X_1 as the endogenous variable and X_2 as the candidate IV. We use D_{train} and D_{test} to estimate the modified λ, then carry out IV transformation and estimation on $D_{unlabel}$. In Figure 4, we present the mean and empirical 95% confidence interval of the IV estimate (implementing 500 repetitions).

Compared with the third plot in Figure 2, it is clear that the extended EnsembleIV approach can recover empirically unbiased coefficient estimates in the presence of the peripheral feature challenge, indicating that

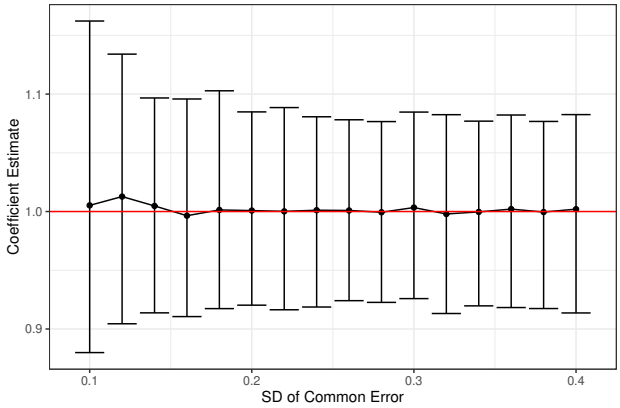

Figure 4 Estimation Results of Extended EnsembleIV (red line indicates true coefficient value; results obtained over 500 repetitions)

the modified IV transformation step is able to remove the correlations between prediction error and peripheral feature.

Appendix E: Statistical Properties

We will now present the asymptotic properties of our EnsembleIV estimator. We begin by establishing the scenario and available data. There are two sets of data: $D_{label} = \{(Y_i, X_i, V_i)\}_{i=1\ldots n_l}$ and $D_{unlabel} = \{(Y_i, \cdot, V_i)\}_{j=1\ldots n_u}$, where V_i represents a vector of features (e.g., text); $X_i = f(V_i)$ is a property of the features (e.g., textual sentiment), which is missing for all elements in $D_{unlabel}$; Y_i is some outcome of interest, and each tuple of data is drawn i.i.d. from some population of interest. Further, D_{label} is partitioned into two subsets: $D_{test} = \{(X_i, V_i)\}_{i=1\ldots n_t}$ and $D_{train} = \{(X_i, V_i)\}_{i=n_{t+1}\ldots n_l}$, where $|D_{test}| = n_t$, $|D_{train}| = n_{tr} = n_l - n_t$, and $\psi = \frac{n_t}{n_l}$ represents the (fixed) proportion of the labeled data allocated to D_{test}. Please note that, in the second-phase regression, there may be a set of exogenous control covariates, $\boldsymbol{W}$. We omit them for brevity in the following theoretical discussions, because they are exogenous with respect to the ML-generated covariate.

We first impose a set of common regularity conditions on f, the function that describes ML predictions. These conditions ensure f is well-behaved for the eventual purpose of estimation, which puts mild limitations on permissible function classes. Specifically, we impose the following regularity conditions:

- $|f(V)| \leq C_1$ for some constant C_1.
- $E[f(V)^2] \in (0, \infty)$.

Now we turn our attention to the estimation of f using D_{train}. Specifically, there is $\widehat{\mathcal{F}} = \{\hat{f}_j\}_{j=1\ldots M}$ a set of $M > 1$ unique models, where each element $\hat{f}_j(V)$ is capable of providing a distinct prediction of X given

V, represented by

$$\begin{aligned}\widehat{X}^{(j)} &= \hat{f}_j(V) \\ &= f(V) + \Big(\hat{f}_j(V) - f(V)\Big) \\ &= f(V) + e_j(V).\end{aligned}$$

While EnsembleIV proposes to construct $\widehat{\mathcal{F}}$ via an ensemble machine learning method, the use of ensemble learners is not a strict requirement for the application of the theory derived in this section. Instead, we consider the following mild regularity conditions on the prediction errors $e_j(V)$, i.e., the prediction error associated with the j-th model:

- $E[e_j(V) \mid V] = 0$.
- $E\big[e_j^2(V)\big] = \frac{C_2^{(j)}}{n_{tr}^{\gamma}}$ for some $\gamma > 0$ and constants $\{C_2^{(j)}\}_{j=1}^{M}$ with $C_2^{(j)} \in (0, \infty)$.
- $|E[e_j(V)e_k(V)]| = \frac{C_3^{(j,k)}}{n_{tr}^{\nu}}$ for some $\nu \geq \gamma$ and constants $\{C_3^{(j,k)}\}_{j,k=1}^{M}$ with $C_3^{(j,k)} \in (0, \infty)$.
- $E\big[e_j^4(V)\big] \leq \frac{C_4^{(j)}}{n_{tr}^{\eta}}$ for some $\eta \leq 2\min\{\gamma, \nu\}$ and constants $\{C_4^{(j)}\}_{j=1}^{M}$ with $C_4^{(j)} \in (0, \infty)$.

Next, we consider the main structural equation of interest. For simplicity, and without loss of generality, let us define $\widehat{X} \coloneqq \widehat{X}^{(1)}$ and $Z \coloneqq \widehat{X}^{(2)}$ to be two distinct estimations of f, with corresponding errors $e_{\widehat{X}} \coloneqq e_1(V)$ and $e_Z \coloneqq e_2(V)$. For clarity, we restate a simplified version of the structural equation from Eq.1-Eq.2 in the main text:

$$\begin{aligned}Y &= X\beta + \varepsilon \\ &= \widehat{X}\beta + (\varepsilon - e_{\widehat{X}}\beta) \\ &= \widehat{X}\beta + u.\end{aligned} \tag{5}$$

To reiterate, we have removed the presence of exogenous control variables $\boldsymbol{W}$ for mathematical convenience. As discussed in Section 2.1 of the main manuscript (adapting to the notations and setup in this part), we make the following standard and mild assumptions about the idiosyncratic error term ε:

Assumption I: $E[\varepsilon \mid X] = 0$.

Assumption II: Given $\widehat{X}^{(i)} = X + e_i(V)$, $E[\varepsilon e_i(V) \mid X] = 0,\ \forall i \in \{1, \ldots, M\}$.

Assumption I ensures that, in the absence of measurement error, we are working with a correctly specified regression equation free from other sources of endogeneity, while Assumption II ensures that prediction errors associated with each individual learner do not have a direct impact on outcome Y beyond the covariates already included in the regression. Furthermore, to more clearly characterize the various sources of error in

the EnsembleIV estimator in finite samples, and to facilitate the demonstration of its asymptotic behavior, we also make the following homoscedasticity assumption:

Assumption III: $E[\varepsilon^2 \mid X] = \sigma_\varepsilon^2 < \infty$.

That is, in the absence of measurement error, the variance of ε conditioned on X is homogeneous and finite. This assumption is notably standard practice in the econometrics literature, resulting in simpler and more intuitive empirical and theoretical modeling. Regression estimation with infinite variance is extremely challenging as most standard estimators rely on various smoothness properties of the underlying data-generating process, which are violated in the presence of infinite variance. Homogeneous variance (i.e., homoscedasticity) is also commonly assumed as it enables more reliable (unbiased and consistent) estimation of the regression coefficient standard errors and, thus, inference on the coefficients themselves.[2]

Our focal context occurs when the estimation of β in (5) is carried out on $D_{unlabel}$ with $\widehat{X}$ in place of the unobserved X. The implicit objective is to leverage the larger sample size of $D_{unlabel}$ for estimation, given $n_u >> n_l$. However, standard estimation of β by Ordinary Least Squares (OLS) using $D_{unlabel}$ in (5) is generally biased and inconsistent because machine learning models generally produce imperfect predictions of X, and the resulting measurement error yields unreliable inference on β.

To move forward, despite the imperfect predictions of machine learning models, EnsembleIV introduces the quantity λ, which is used to build an instrumental variable (IV) for $\widehat{X}$. We restate λ with slight notational adjustments from Equation 3 in the main text:

$$\begin{aligned}\lambda &= \frac{\rho_{Zu}}{\rho_{\widehat{X}u}} \\ &= \frac{\mathrm{Cov}(Z,u)}{\mathrm{Cov}(\widehat{X},u)} \cdot \frac{\sigma_{\widehat{X}}}{\sigma_Z} \\ &= \frac{\mathrm{Cov}(Z,e_{\widehat{X}})}{\mathrm{Cov}(\widehat{X},e_{\widehat{X}})} \cdot \frac{\sigma_{\widehat{X}}}{\sigma_Z},\end{aligned} \tag{6}$$

where the final equality follows from the application of our above assumptions. Moreover, we can also consider a plug-in estimator of λ:

$$\hat{\lambda} = \frac{\widehat{\mathrm{Cov}}(Z,e_{\widehat{X}})}{\widehat{\mathrm{Cov}}(\widehat{X},e_{\widehat{X}})} \cdot \frac{\hat{\sigma}_{\widehat{X}}}{\hat{\sigma}_Z}, \tag{7}$$

substituting each theoretical $\mathrm{Cov}(\cdot,\cdot)$ and σ with an empirical analog $\widehat{\mathrm{Cov}}(\cdot,\cdot)$ and $\hat{\sigma}$ respectively, computed from an available sample of data, where all the necessary variables are observed. Having defined our setting,

[2] Note that this homoscedasticity assumption is not strictly necessary for EnsembleIV estimation and inference because we bootstrap our standard errors.

assumptions, and relevant quantities, we will now begin establishing the theoretical results that enable EnsembleIV, the first of which will focus on the properties of λ and $\hat{\lambda}$. Note that all theoretical results implicitly rely on Assumptions I-III as well as the regularity conditions on f and $e_j(V)$ discussed above.

LEMMA 1 (**Nevo and Rosen (2012)**). *Let* $\widetilde{Z} = \sigma_{\widehat{X}} Z - \lambda \sigma_Z \widehat{X}$, *then* $Cov(\widetilde{Z}, u) = 0$.

Proof. This result follows directly from the definition of λ:

$$\begin{aligned}\text{Cov}(\widetilde{Z}, u) &= \sigma_{\widehat{X}} \text{Cov}(Z, u) - \frac{\text{Cov}(Z, u)}{\text{Cov}(\widehat{X}, u)} \cdot \frac{\sigma_{\widehat{X}}}{\sigma_Z} \sigma_Z \text{Cov}(\widehat{X}, u) \\ &= 0. \quad \square\end{aligned}$$

COROLLARY 1. *Given a sample of data elements* $D_n = \left(\widehat{\boldsymbol{X}}, \boldsymbol{Z}, \boldsymbol{u}\right) = \left\{\left(\widehat{X}_i, Z_i, u_i\right)\right\}_{i=1\ldots n}$, $\widetilde{Z}_i = \hat{\sigma}_{\widehat{X}} Z_i - \hat{\lambda}_n \hat{\sigma}_Z \widehat{X}_i$, $\widetilde{\boldsymbol{Z}} = \{\widetilde{Z}_i\}_{i=1\ldots n}$, *let*

$$\begin{aligned}\hat{\lambda}_n &= \frac{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})(u_i - \bar{u})}{\frac{1}{n-1}\sum_{i=1}^n (\widehat{X}_i - \bar{\widehat{X}})(u_i - \bar{u})} \cdot \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (\widehat{X}_i - \bar{\widehat{X}})^2}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})^2}} \\ &= \frac{\widehat{Cov}(\boldsymbol{Z}, \boldsymbol{u})}{\widehat{Cov}(\widehat{\boldsymbol{X}}, \boldsymbol{u})} \cdot \frac{\hat{\sigma}_{\widehat{X}}}{\hat{\sigma}_Z},\end{aligned}$$

then

$$\widehat{Cov}(\widetilde{\boldsymbol{Z}}, \boldsymbol{u}) = 0.$$

From Lemma 1, we have that the population level transformation of the potentially endogenous instrument Z provides a new instrument, $\widetilde{Z}$, that theoretically meets the exclusion condition. Corollary 1 extends this result to a fixed dataset. More specifically, it establishes that transforming the sample vector $\boldsymbol{Z}$ by the $\hat{\lambda}_n$ computed on the dataset, provides a new sample vector $\widetilde{\boldsymbol{Z}}$ that meets the exclusion condition for the given dataset. While it is possible to compute $\hat{\lambda}_n$ on D_{test} (i.e., $\hat{\lambda}_{n_t}$), it is generally not possible to compute $\hat{\lambda}_{n_u}$ on $D_{unlabel}$. And, given $n_t >> n_u$, our goal is to leverage $D_{unlabel}$ for inference. Therefore, we next analyze the relationship between $\hat{\lambda}_{n_t}$ and $\hat{\lambda}_{n_u}$.

LEMMA 2. *Let* $\hat{\lambda}_{n_t}$ *be the (computable) plug-in estimator for* λ *based on* D_{test}, *and* $\hat{\lambda}_{n_u}$ *be the (uncomputable) plug-in estimator for* λ *based on* $D_{unlabel}$, *then* $\hat{\lambda}_{n_t} \xrightarrow{a.s.} \hat{\lambda}_{n_u}$ *as* $n_u, n_t \to \infty$.

Proof. We start by defining useful quantities:

$$\begin{aligned}\theta_{n_{tr}} &= \left(\text{Cov}(\boldsymbol{Z}, \boldsymbol{e}_{\widehat{\boldsymbol{X}}}), \text{Cov}(\widehat{\boldsymbol{X}}, \boldsymbol{e}_{\widehat{\boldsymbol{X}}}), \sigma_{\widehat{X}}, \sigma_Z\right), \\ \widehat{\theta}_{n_{tr},n} &= \left(\widehat{\text{Cov}}_n(\boldsymbol{Z}, \boldsymbol{e}_{\widehat{\boldsymbol{X}}}), \widehat{\text{Cov}}_n(\widehat{\boldsymbol{X}}, \boldsymbol{e}_{\widehat{\boldsymbol{X}}}), \hat{\sigma}_{\widehat{X},n}, \hat{\sigma}_{Z,n}\right), \\ g\left((a,b,c,d)\right) &= \frac{a}{b} \cdot \frac{c}{d}.\end{aligned}$$

Note that g is continuous everywhere except where b or d are zero. Further, we define $\lambda_{n_{tr}}$, making explicit its reliance on the training data sample size n_{tr}:

$$\begin{aligned}
\lambda_{n_{tr}} &= g(\theta_{n_{tr}}) \\
&= \frac{\mathrm{Cov}(Z, e_{\widehat{X}})}{\mathrm{Cov}(\widehat{X}, e_{\widehat{X}})} \cdot \frac{\sigma_{\widehat{X}}}{\sigma_Z} \\
&= \frac{\mathrm{Cov}(Z, e_{\widehat{X}})}{\mathrm{Var}(e_{\widehat{X}})} \cdot \frac{\sqrt{\mathrm{Var}(f(V) + e_{\widehat{X}})}}{\sqrt{\mathrm{Var}(f(V) + e_Z)}} \\
&= \frac{\Theta({n_{tr}}^{-\nu})}{\Theta({n_{tr}}^{-\gamma})} \cdot \frac{\sqrt{\mathrm{Var}(f(V) + \Theta({n_{tr}}^{-\gamma}))}}{\sqrt{\mathrm{Var}(f(V) + \Theta({n_{tr}}^{-\gamma}))}} \\
&= \Theta\big({n_{tr}}^{-(\nu-\gamma)}\big) \cdot \Theta(1) \\
&= \Theta({n_{tr}}^{-\alpha}) \quad \text{where } \alpha = \nu - \gamma \geq 0.
\end{aligned} \tag{8}$$

We note that $\alpha = \nu - \gamma \geq 0$ is a natural restriction because, as it can be interpreted as comparing the rate that cross-learner error correlations decay (ν) to the rate of marginal prediction-error variance decay (γ), and by the (conditional) Cauchy–Schwarz inequality,

$$\begin{aligned}
|\mathbb{E}[e_j(V) e_k(V)]| &\leq \sqrt{\mathbb{E}\big[e_j^2(V)\big] \cdot \mathbb{E}[e_k^2(V)]} \\
&\leq \frac{\sqrt{C_2^{(j)}\, C_2^{(k)}}}{n_{tr}^{\gamma}} \\
&= \Theta\big(n_{tr}^{-\gamma}\big)\,.
\end{aligned} \tag{9}$$

Therefore, (9) shows that one may take $\nu \geq \gamma$ given that the cross-learning error decay rate is not slower than the prediction-error variance decay. It may be the same order (then $\nu = \gamma$ and $\alpha = 0$), or it may be a smaller order (then $\nu > \gamma$ and $\alpha > 0$), or it may be exactly zero. As a result, $\lambda_{n_{tr}}$ converges to some constant (specifically zero if $\alpha > 0$). We also define $\hat{\lambda}_{n_{tr},n}$, making explicit its similar reliance on n_{tr} and additional reliance on the size (n) of a separate data sample used for plug-in estimation:

$$\begin{aligned}
\hat{\lambda}_{n_{tr},n} &= g(\widehat{\theta}_{n_{tr},n}) \\
&= \frac{\widehat{\mathrm{Cov}}_n(Z, e_{\widehat{X}})}{\widehat{\mathrm{Cov}}_n(\widehat{X}, e_{\widehat{X}})} \cdot \frac{\hat{\sigma}_{\widehat{X},n}}{\hat{\sigma}_{Z,n}} \\
&\xrightarrow[n\to\infty]{\text{a.s.}} \frac{\mathrm{Cov}(Z, e_{\widehat{X}})}{\mathrm{Cov}(\widehat{X}, e_{\widehat{X}})} \cdot \frac{\sigma_{\widehat{X}}}{\sigma_Z} \\
&= g(\theta_{n_{tr}}) \\
&= \lambda_{n_{tr}}.
\end{aligned} \tag{10}$$

Note that (10) follows from $\widehat{\theta}_{n_{tr},n} \xrightarrow[n\to\infty]{\text{a.s.}} \theta_{n_{tr}}$ and therefore $g(\widehat{\theta}_{n_{tr},n}) \xrightarrow[n\to\infty]{\text{a.s.}} g(\theta_{n_{tr}})$, by the Strong law of large numbers and the Continuous mapping theorem. Importantly, (8) ensures that (10) holds for any value of n_{tr}, including $n_{tr} \to \infty$ where $\mathrm{Cov}(\widehat{X}, e_{\widehat{X}}) \to 0$.

Finally, we define $\hat{\lambda}_{n_t} \equiv \hat{\lambda}_{n_{tr},n_t}$ and $\hat{\lambda}_{n_u} \equiv \hat{\lambda}_{n_{tr},n_u}$, and have

$$|\hat{\lambda}_{n_t} - \hat{\lambda}_{n_u}| \leq |\hat{\lambda}_{n_t} - \lambda_{n_{tr}}| + |\hat{\lambda}_{n_u} - \lambda_{n_{tr}}|$$

$$\xrightarrow[n_t,n_u\to\infty]{\text{a.s.}} 0,$$

by the Triangle inequality and the application of (10) for both $\hat{\lambda}_{n_{tr},n_t}$ and $\hat{\lambda}_{n_{tr},n_u}$. □

Lemma 2, demonstrates that $\hat{\lambda}_{n_t}$ is actually converging almost-surely to $\hat{\lambda}_{n_u}$, allowing us to use $\hat{\lambda}_{n_t}$ in its place, asymptotically. We can therefore obtain the asymptotic behavior of the EnsembleIV estimator on unlabeled data $D_{unlabel}$, by first analyzing its behavior without the transformation of Z by $\hat{\lambda}_{n_t}$

THEOREM 1. *Let $\hat{\beta}_{IV}$ be an instrumental variable estimator of β in (5) estimated on unlabeled data $D_{unlabel}$, where the instrument Z, is a variable with prediction error that asymptotically meets the exclusion condition. Under Assumptions I-III as well as the regularity conditions on f and $e_j(V)$, as $n_u, n_{tr} \to \infty$ with $n_u = o(n_{tr}^{2\nu})$*

$$\sqrt{n_u}(\hat{\beta}_{IV} - \beta) \rightsquigarrow \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{E[f(V)^2]}\right).$$

Proof. The IV estimator is defined as:

$$\begin{aligned}\hat{\beta}_{IV} &= \left(\frac{1}{n_u}\sum_{i=1}^{n_u} Z_i \widehat{X}_i\right)^{-1}\left(\frac{1}{n_u}\sum_{i=1}^{n_u} Z_i Y_i\right) \\ &= \left(\frac{1}{n_u}\sum_{i=1}^{n_u} Z_i \widehat{X}_i\right)^{-1}\left(\frac{1}{n_u}\sum_{i=1}^{n_u} Z_i \left(X_i\beta + \varepsilon_i\right)\right). \end{aligned} \tag{11}$$

As is common in instrumental variable analysis, we are specifically interested in the behavior of $\hat{\beta}_{IV}$ given a set of covariates/instruments. In our setting, this means conditioning on the fitted learners $\widehat{\mathcal{F}}$ – or equivalently, on D_{train} and any algorithmic randomness used to fit the learners – such that each $\hat{f}_j$ is treated as observed and fixed when estimating $\hat{\beta}_{IV}$ in $D_{unlabeled}$. Conditional on $\widehat{\mathcal{F}}$, the unlabeled observations $\{(V_i, \varepsilon_i)\}_{i=1}^{n_u}$ remain random and i.i.d. For notational simplicity, we will leave the conditioning of $\hat{\beta}_{IV}$ implicit, unless if proves particularly fruitful to be explicit. We can now proceed by recognizing that the asymptotic bias and variance of $\hat{\beta}_{IV}$ will be dictated by the limit of the right summand in (11) and therefore consider it's components in service of analyzing our final object of interest

$$Z_{IV} \equiv \sqrt{n_u}(\hat{\beta}_{IV} - \beta).$$

We begin by considering the denominator of $\hat{\beta}_{IV}$,

$$\begin{aligned}\frac{1}{n_u}\sum_{i=1}^{n_u} Z_i\widehat{X}_i &= \frac{1}{n_u}\sum_i \left(f(V_i)+e_2(V_i)\right)\left(f(V_i)+e_1(V_i)\right)\\ &= \frac{1}{n_u}\sum_i^{n_u} f(V_i)^2 + \left(e_2(V_i)+e_1(V_i)\right)f(V_i) + \left(e_2(V_i)e_1(V_i)\right)\\ &= \bar{f^2} + \frac{1}{n_u}\sum_i^{n_u} A_i\end{aligned}$$

defining $\bar{f^2} \equiv \frac{1}{n_u}\sum_i f(V_i)^2$ and $A_i \equiv \left(e_2(V_i)+e_1(V_i)\right)f(V_i) + e_2(V_i)e_1(V_i)$. We analyze the first central moment of $\frac{1}{n_u}\sum_i^{n_u} A_i$,

$$\begin{aligned}E\left[\frac{1}{n_u}\sum_{i=1}^{n_u} A_i\right] &= E\left[\frac{1}{n_u}\sum_i^{n_u}\left(e_2(V_i)+e_1(V_i)\right)f(V_i)\right] + E\left[\frac{1}{n_u}\sum_i^{n_u}\left(e_2(V_i)e_1(V_i)\right)\right]\\ &= 0 + O\left(\frac{1}{n_{tr}^{\nu}}\right)\\ &= O\left(\frac{1}{n_{tr}^{\nu}}\right).\end{aligned} \tag{12}$$

We also analyze the second central moment of $\frac{1}{n_u}\sum_i^{n_u} A_i$, noting that conditional on $\widehat{\mathcal{F}}$ the sequence $\{A_i\}_{i=1}^{n_u}$ is i.i.d.:

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{n_u}\sum_{i=1}^{n_u} A_i\right] &= \frac{1}{n_u^2}\sum_{i=1}^{n_u}\mathrm{Var}\left[\left(e_2(V_i)+e_1(V_i)\right)f(V_i)+e_2(V_i)e_1(V_i)\right]\\ &= \frac{1}{n_u^2}\sum_{i=1}^{n_u}\Big[\mathrm{Var}\left[\left(e_2(V_i)+e_1(V_i)\right)f(V_i)\right] + \mathrm{Var}\left[e_2(V_i)e_1(V_i)\right]\\ &\quad + 2\cdot\mathrm{Cov}\left[\left(e_2(V_i)+e_1(V_i)\right)f(V_i), e_2(V_i)e_1(V_i)\right]\Big]\\ &\le \frac{1}{n_u^2}\sum_{i=1}^{n_u}\Big[\mathrm{Var}\left[\left(e_2(V_i)+e_1(V_i)\right)f(V_i)\right] + \sqrt{E[e_2^4(V_i)]\cdot E[e_1^4(V_i)]}\\ &\quad + 2\cdot\sqrt{\mathrm{Var}[\left(e_2(V_i)+e_1(V_i)\right)f(V_i)]\cdot\mathrm{Var}[e_2(V_i)e_1(V_i)]}\Big]\\ &= \frac{1}{n_u^2}\sum_{i=1}^{n_u}\left[O\left(\frac{1}{n_{tr}^{\gamma}}\right) + O\left(\frac{1}{n_{tr}^{\eta}}\right) + O\left(\frac{1}{n_{tr}^{(\gamma+\eta)/2}}\right)\right]\\ &= O\left(\frac{1}{n_u}\frac{1}{n_{tr}^{\gamma}}\right) + O\left(\frac{1}{n_u}\frac{1}{n_{tr}^{\eta}}\right).\end{aligned} \tag{13}$$

We note that the inequality in (13) follows from the fact that $\mathrm{Var}\left[e_2e_1\right] = E\left[e_2^2e_1^2\right] - \left(E\left[e_2e_1\right]\right)^2 \le \sqrt{E[e_2^4]\,E[e_1^4]}$ and that $\left|\mathrm{Cov}\left[\left(e_2+e_1\right)f, e_2e_1\right]\right| \le \sqrt{\mathrm{Var}\left[\left(e_2+e_1\right)f\right]\mathrm{Var}\left[e_2e_1\right]}$ and, both by the Cauchy–Schwarz inequality.

Therefore, we can finally express the denominator component as

$$\begin{aligned}\frac{1}{n_u}\sum_{i=1}^{n_u} Z_i\widehat{X}_i &= \bar{f}^2 + \frac{1}{n_u}\sum_{i=1}^{n_u} A_i \\ &= \bar{f}^2 + E\left[\frac{1}{n_u}\sum_{i=1}^{n_u} A_i\right] + \left(\frac{1}{n_u}\sum_{i=1}^{n_u} A_i - E[A_i]\right) \\ &= \bar{f}^2 + O\left(\frac{1}{n_{tr}^{\nu}}\right) + \left(O_p\left(\frac{1}{\sqrt{n_u}}\frac{1}{n_{tr}^{\gamma/2}}\right) + O_p\left(\frac{1}{\sqrt{n_u}}\frac{1}{n_{tr}^{\eta/2}}\right)\right) \\ &= \bar{f}^2 + O\left(\frac{1}{n_{tr}^{\nu}}\right) + R^{(D)}_{n_u,n_{tr}}\end{aligned} \tag{14}$$

where the second equality follows from adding zero, the third equality follows from combining (12)-(13) with Chebyshev's inequality, and the final equality follows from defining $R^{(D)}_{n_u,n_{tr}} \equiv O_p\left(\frac{1}{\sqrt{n_u}}\frac{1}{n_{tr}^{\gamma/2}}\right) + O_p\left(\frac{1}{\sqrt{n_u}}\frac{1}{n_{tr}^{\eta/2}}\right)$ as the O_p remainder terms from the denominator.

Next, we consider the numerator of $\hat{\beta}_{IV}$,

$$\begin{aligned}\frac{1}{n_u}\sum_{i=1}^{n_u} Z_iY_i &= \frac{1}{n_u}\sum_{i=1}^{n_u}(Z_iX_i\beta + Z_i\varepsilon_i) \\ &= \frac{1}{n_u}\sum_{i=1}^{n_u}((f(V_i)+e_2(V_i))f(V_i)\beta + (f(V_i)+e_2(V_i))\varepsilon_i) \\ &= \frac{1}{n_u}\sum_{i=1}^{n_u} f(V_i)^2\beta + f(V_i)e_2(V_i)\beta + f(V_i)\varepsilon_i + e_2(V_i)\varepsilon_i \\ &= \beta\bar{f}^2 + \frac{1}{n_u}\sum_{i}^{n_u} f(V_i)\varepsilon_i + B_i\end{aligned}$$

defining $B_i \equiv f(V_i)e_2(V_i)\beta + e_2(V_i)\varepsilon_i$. We analyze the first central moment of $\frac{1}{n_u}\sum_i^{n_u} B_i$,

$$\begin{aligned}E\left[\frac{1}{n_u}\sum_{i=1}^{n_u} B_i\right] &= \beta E\left[\frac{1}{n_u}\sum_{i} f(V_i)e_2(V_i)\right] + E\left[\frac{1}{n_u}\sum_{i} e_2(V_i)\varepsilon_i\right] \\ &= 0 + 0.\end{aligned} \tag{15}$$

We also analyze the second central moment of the $\frac{1}{n_u}\sum_i^{n_u} B_i$, recognizing that conditional on $\widehat{\mathcal{F}}$ the sequence $\{B_i\}_{i=1}^{n_u}$ is i.i.d.

$$\begin{aligned}\mathrm{Var}\left[\frac{1}{n_u}\sum_{i=1}^{n_u} B_i\right] &= \mathrm{Var}\left[\frac{1}{n_u}\sum_{i=1}^{n_u} f(V_i)e_2(V_i)\beta + e_2(V_i)\varepsilon_i\right] \\ &= \frac{1}{n_u^2}\sum_{i=1}^{n_u} \mathrm{Var}\left[f(V_i)e_2(V_i)\beta\right] + \mathrm{Var}\left[e_2(V_i)\varepsilon_i\right] + 2\,\mathrm{Cov}[\beta f(V_i)e_2(V_i),\, e_2(V_i)\varepsilon_i] \\ &= \frac{1}{n_u^2}\sum_{i=1}^{n_u}\left[O\left(\frac{1}{n_{tr}^{\gamma}}\right) + O\left(\frac{1}{n_{tr}^{\gamma}}\right) + 0\right] \\ &= O\left(\frac{1}{n_u}\frac{1}{n_{tr}^{\gamma}}\right).\end{aligned} \tag{16}$$

Therefore, we can finally express the numerator component as

$$\begin{aligned}\frac{1}{n_u}\sum_{i=1}^{n_u} Z_i Y_i &= \frac{1}{n_u}\sum_{i}^{n_u} f(V_i)\varepsilon_i + \beta\bar{f}^2 + \frac{1}{n_u}\sum_{i}^{n_u} B_i \\ &= \frac{1}{n_u}\sum_{i}^{n_u} f(V_i)\varepsilon_i + \beta\bar{f}^2 + E\left[\frac{1}{n_u}\sum_{i=1}^{n_u} B_i\right] + \left(\frac{1}{n_u}\sum_{i=1}^{n_u} B_i - E[B_i]\right) \\ &= \frac{1}{n_u}\sum_{i}^{n_u} f(V_i)\varepsilon_i + \beta\bar{f}^2 + O_p\left(\frac{1}{\sqrt{n_u}}\frac{1}{n_{tr}^{\gamma/2}}\right) \\ &= \frac{1}{n_u}\sum_{i}^{n_u} f(V_i)\varepsilon_i + \beta\bar{f}^2 + R^{(N)}_{n_u,n_{tr}}\end{aligned} \tag{17}$$

where the second equality follows from adding zero, the third equality follows from combining (15)-(16) with Chebyshev's inequality, and the final equality follows from defining $R^{(N)}_{n_u,n_{tr}} \equiv O_p\left(\frac{1}{\sqrt{n_u}}\frac{1}{n_{tr}^{\gamma/2}}\right)$ as the remainder terms from the numerator.

We can therefore combine (14) and (17) to finalize our analysis of $\hat{\beta}_{IV}$ from (11) with

$$\hat{\beta}_{IV} = \frac{\frac{1}{n_u}\sum_{i}^{n_u} f(V_i)\varepsilon_i + \beta\bar{f}^2 + R^{(N)}_{n_u,n_{tr}}}{\bar{f}^2 + O\left(\frac{1}{n_{tr}^{\nu}}\right) + R^{(D)}_{n_u,n_{tr}}}$$

Moreover, we now can analyze our main object of interest

$$\begin{aligned}Z_{IV} &= \sqrt{n_u}(\hat{\beta}_{IV} - \beta) \\ &= \sqrt{n_u}\left(\frac{\frac{1}{n_u}\sum_{i=1}^{n_u} f(V_i)\varepsilon_i + \beta\bar{f}^2 + R^{(N)}_{n_u,n_{tr}}}{\bar{f}^2 + O\left(\frac{1}{n_{tr}^{\nu}}\right) + R^{(D)}_{n_u,n_{tr}}} - \beta\right) \\ &= \sqrt{n_u}\left(\frac{\frac{1}{n_u}\sum_{i=1}^{n_u} f(V_i)\varepsilon_i + R^{(N)}_{n_u,n_{tr}} - \beta\left(O\left(\frac{1}{n_{tr}^{\nu}}\right) + R^{(D)}_{n_u,n_{tr}}\right)}{\bar{f}^2 + O\left(\frac{1}{n_{tr}^{\nu}}\right) + R^{(D)}_{n_u,n_{tr}}}\right) \\ &= \sqrt{n_u}\left(\frac{\frac{1}{n_u}\sum_{i=1}^{n_u} f(V_i)\varepsilon_i - \beta\left(O\left(\frac{1}{n_{tr}^{\nu}}\right) + O_p\left(\frac{1}{\sqrt{n_u}}\frac{1}{n_{tr}^{\gamma/2}}\right) + O_p\left(\frac{1}{\sqrt{n_u}}\frac{1}{n_{tr}^{\eta/2}}\right)\right)}{\bar{f}^2 + O\left(\frac{1}{n_{tr}^{\nu}}\right) + R^{(D)}_{n_u,n_{tr}}}\right) \\ &= \left(\frac{\frac{1}{\sqrt{n_u}}\sum_{i=1}^{n_u} f(V_i)\varepsilon_i - \beta\left(O\left(\frac{\sqrt{n_u}}{n_{tr}^{\nu}}\right) + O_p\left(\frac{1}{n_{tr}^{\gamma/2}}\right) + O_p\left(\frac{1}{n_{tr}^{\eta/2}}\right)\right)}{\bar{f}^2 + O\left(\frac{1}{n_{tr}^{\nu}}\right) + o_p(1)}\right) \\ &\underset{n_{tr}\to\infty}{\longrightarrow} \frac{\frac{1}{\sqrt{n_u}}\sum_{i=1}^{n_u} f(V_i)\varepsilon_i + o_p(1)}{\bar{f}^2 + o_p(1)}.\end{aligned}$$

Finally, by the CLT,

$$\frac{1}{\sqrt{n_u}}\sum_{i=1}^{n_u} f(V_i)\varepsilon_i \rightsquigarrow \mathcal{N}\left(0,\, \sigma_\varepsilon^2 E\left[f(V)^2\right]\right),$$

and since $\bar{f}^2 \xrightarrow{p} E[f(V)^2]$ by the LLN, from Slutsky's theorem we have

$$\frac{\frac{1}{\sqrt{n_u}}\sum_{i=1}^{n_u} f(V_i)\varepsilon_i}{E[f(V)^2] + o_p(1)} \rightsquigarrow \mathcal{N}\left(0,\, \frac{\sigma_\varepsilon^2}{E[f(V)^2]}\right),$$

yielding our desired result:

$$Z_{IV} \rightsquigarrow \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{E[f(V)^2]}\right).$$

□

COROLLARY 2. *Let* $\widetilde{Z}_i = \hat{\sigma}_{\widehat{X}} Z_i - \hat{\lambda}_{n_t} \hat{\sigma}_Z \widehat{X}_i$ *be the transformation of each* Z_i *in* $D_{unlabel}$ *and let* $\hat{\beta}_{EIV}$ *be an estimator with instrument* $\widetilde{Z}$ *defined by the conditions in Theorem 1, then by Corollary 1 and Lemma 2, as additionally* $n_t \to \infty$

$$\sqrt{n_u}(\hat{\beta}_{EIV} - \beta) \rightsquigarrow \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{E[f(V)^2]}\right).$$

The specific behavior of $\hat{\beta}_{EIV}$ *is dictated by* γ, ν, η, *and the convergence rate of* $\hat{\lambda}_{n_t}$ *to* $\lambda_{n_{tr}}$.

Remark: The challenge of conducting inference after the use of flexible machine learning algorithms for model selection (i.e, post-selection inference) has been considered extensively (Berk et al. 2013, Belloni et al. 2014, Chernozhukov et al. 2015, Belloni et al. 2016, 2017). One critical challenge this literature attempts to address occurs when the outcome model in which inference is being carried out is selected in a data-driven manner (Berk et al. 2013). A second challenge occurs when the variable of inferential interest (e.g., treatment) in the outcome model is not observed exogenously, and therefore the treatment and the outcome are jointly determined by a subset of variables. Data-dependent selection for these control variables can create a type of "regularization bias" by incorrectly estimating the functional form with which the control variables enter the treatment or the outcome models, including failing to include relevant controls (Chernozhukov et al. 2015). Our setting described in Equation 1 (main manuscript) assumes an outcome model where $E[\varepsilon \mid X, \boldsymbol{W}] = 0$ and the variables used to build X (e.g., textual features that determine sentiment) have no direct impact on the outcome conditional on X. Therefore, data-driven selection of the treatment model (i.e., the model of X) will not lead to estimation bias of β in the outcome model. Essentially, in Equation 1 (main manuscript) X is exogenous by assumption, hence the only potential source of bias in Equation 2 (main manuscript) arises from using $\widehat{X}$ in place of X due to prediction errors. There may exist contexts where these assumptions do not hold, and the extensive and focused statistical innovations required to relax these assumptions would be of future interest.

Appendix F: Using Subset of Individual Learners as Endogenous Covariates and IVs

On the Bike Sharing dataset, we first built a random forest of 100 trees and then constructed predictions based on subsets of 50 trees. Among all $\binom{100}{50}$ different ways to select 50 trees out of 100, we randomly

sampled 100, to maintain the same number of covariates as in EnsembleIV with individual trees, enabling an apples-to-apples comparison. The estimation results are reported in Table 2, below. We can see that, compared to using individual trees as covariates, using subsets of 50 trees results in slightly more biased point estimates, inflated standard errors, and overall worse estimation MSE scores.

Table 2 EnsembleIV Estimates with Predictions from Subsets of Trees

		Individual Trees			Subsets of 50 Trees		
	True	Top-3	PCA	LASSO	Top-3	PCA	LASSO
β_0	1.0	1.026	1.015	1.058	0.956	0.941	0.903
		(0.063)	(0.063)	(0.062)	(0.069)	(0.073)	(0.073)
β_{MLV}	0.5	0.494	0.496	0.487	0.510	0.513	0.521
		(0.013)	(0.013)	(0.013)	(0.015)	(0.015)	(0.016)
β_{W_1}	2.0	2.000	2.000	2.000	2.000	2.000	2.000
		(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
β_{W_2}	1.0	1.000	1.000	1.000	1.000	1.000	1.000
		(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Estimation MSE		0.005	0.004	0.008	0.007	0.009	0.015

Appendix G: Variance-Weighted EnsembleIV Estimates

The following Tables 3-4 report estimation results when EnsembleIV's individual estimators are weighted by their respective variances before aggregation.

Table 3 Variance-Weighted Evaluation Results on Bike Sharing Data

		Linear Second Phase Regression					Logistic Second Phase Regression				
	True	Biased	Unbiased	Ens.IV (Top-3)	Ens.IV (PCA)	Ens.IV (LASSO)	Biased	Unbiased	Ens.IV (Top-3)	Ens.IV (PCA)	Ens.IV (LASSO)
β_0	1.0	0.756	0.999	1.027	1.017	1.059	0.711	0.954	0.982	0.975	1.015
		(0.070)	(0.111)	(0.063)	(0.063)	(0.062)	(0.201)	(0.405)	(0.180)	(0.179)	(0.176)
β_{MLV}	0.5	0.553	0.500	0.494	0.496	0.487	0.552	0.531	0.492	0.495	0.485
		(0.014)	(0.023)	(0.013)	(0.013)	(0.013)	(0.045)	(0.095)	(0.039)	(0.039)	(0.038)
β_{W_1}	2.0	2.000	2.000	2.000	2.000	2.000	1.976	2.057	1.974	1.979	1.976
		(0.003)	(0.007)	(0.003)	(0.003)	(0.003)	(0.054)	(0.142)	(0.054)	(0.055)	(0.055)
β_{W_2}	1.0	1.000	1.000	1.000	1.000	1.000	0.987	1.029	0.987	0.989	0.988
		(0.002)	(0.004)	(0.002)	(0.002)	(0.002)	(0.027)	(0.072)	(0.027)	(0.027)	(0.027)
Estimation MSE		0.067	0.013	0.005	0.004	0.008	0.133	0.206	0.039	0.039	0.037

Appendix H: EnsembleIV Estimates based on LIML / Fuller Methods

The following Tables 5-6 report EnsembleIV results when IV estimations are done via the Limited Information Maximum Likelihood (LIML) method and the Fuller method, respectively.

Table 4 Variance-Weighted Evaluation Results on Bank Marketing Data

		Linear Second Phase Regression					Logistic Second Phase Regression				
	True	Biased	Unbiased	Ens.IV (Top-3)	Ens.IV (PCA)	Ens.IV (LASSO)	Biased	Unbiased	Ens.IV (Top-3)	Ens.IV (PCA)	Ens.IV (LASSO)
β_0	1.0	1.044	1.002	1.008	1.006	1.017	1.041	0.980	1.004	1.002	1.012
		(0.011)	(0.028)	(0.013)	(0.013)	(0.012)	(0.036)	(0.102)	(0.041)	(0.041)	(0.040)
β_{MLV}	0.5	0.280	0.503	0.425	0.437	0.352	0.292	0.505	0.441	0.454	0.366
		(0.043)	(0.096)	(0.049)	(0.050)	(0.039)	(0.140)	(0.291)	(0.165)	(0.167)	(0.136)
β_{W_1}	2.0	2.000	2.000	2.000	2.000	2.000	1.999	1.998	2.001	2.001	2.000
		(0.002)	(0.005)	(0.002)	(0.002)	(0.002)	(0.029)	(0.094)	(0.029)	(0.029)	(0.029)
β_{W_2}	1.0	1.000	0.999	1.000	1.000	1.000	1.000	0.998	1.000	1.001	1.000
		(0.001)	(0.003)	(0.001)	(0.001)	(0.001)	(0.015)	(0.047)	(0.015)	(0.015)	(0.015)
Estimation MSE		0.052	0.010	0.008	0.007	0.024	0.067	0.106	0.034	0.033	0.039

Table 5 Results on Bike Sharing Data based on LIML/Fuller Methods (Linear Second-Phase)

	True	Biased	Unbiased	Ens.IV (Top-3)	Ens.IV (PCA)	Ens.IV (LASSO)	Ens.IV (Top-3)	Ens.IV (PCA)	Ens.IV (LASSO)
				LIML Method			Fuller Method		
β_0	1.0	0.756	0.999	0.954	0.938	0.902	0.952	0.936	0.899
		(0.070)	(0.111)	(0.074)	(0.076)	(0.075)	(0.076)	(0.077)	(0.077)
β_{MLV}	0.5	0.553	0.500	0.510	0.513	0.521	0.510	0.514	0.522
		(0.014)	(0.023)	(0.015)	(0.016)	(0.016)	(0.015)	(0.016)	(0.016)
β_{W_1}	2.0	2.000	2.000	2.001	2.001	2.001	2.000	2.000	2.000
		(0.003)	(0.007)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
β_{W_2}	1.0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		(0.002)	(0.004)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Estimation MSE		0.071	0.015	0.008	0.010	0.016	0.008	0.011	0.017

Table 6 Results on Bank Marketing Data based on LIML/Fuller Methods (Linear Second-Phase)

	True	Biased	Unbiased	Ens.IV (Top-3)	Ens.IV (PCA)	Ens.IV (LASSO)	Ens.IV (Top-3)	Ens.IV (PCA)	Ens.IV (LASSO)
				LIML Method			Fuller Method		
β_0	1.0	1.044	1.002	1.009	1.007	1.017	1.009	1.007	1.018
		(0.011)	(0.028)	(0.012)	(0.012)	(0.012)	(0.010)	(0.010)	(0.010)
β_{MLV}	0.5	0.280	0.503	0.422	0.437	0.349	0.421	0.437	0.350
		(0.043)	(0.096)	(0.046)	(0.046)	(0.038)	(0.040)	(0.041)	(0.032)
β_{W_1}	2.0	2.000	2.000	2.000	2.000	2.000	2.000	2.000	2.000
		(0.002)	(0.005)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
β_{W_2}	1.0	1.000	0.999	1.000	1.000	1.000	1.000	1.000	1.000
		(0.001)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Estimation MSE		0.051	0.011	0.008	0.006	0.025	0.008	0.006	0.024

References

Belloni A, Chernozhukov V, Fernández-Val I, Hansen C (2017) Program evaluation and causal inference with high-dimensional data. *Econometrica* 85(1):233–298.

Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81(2):608–650, URL https://doi.org/10.1093/restud/rdt044.

Belloni A, Chernozhukov V, Wei Y (2016) Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics* 34(4):606–619.

Berk R, Brown L, Buja A, Zhang K, Zhao L (2013) Valid post-selection inference. *The Annals of Statistics* 41(2):802–837.

Chernozhukov V, Hansen C, Spindler M (2015) Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review* 105(5):486–90, URL http://dx.doi.org/10.1257/aer.p20151022.

Dallal GE, Wilkinson L (1986) An analytic approximation to the distribution of lilliefors's test statistic for normality. *The American Statistician* 40(4):294–296.

Murray MP (2006) Avoiding invalid instruments and coping with weak instruments. *Journal of economic Perspectives* 20(4):111–132.

Nevo A, Rosen AM (2012) Identification with imperfect instruments. *Review of Economics and Statistics* 94(3):659–671.