Machine Learning Instrument Variables for Causal Inference

Amandeep Singh, Kartik Hosanagar, and Amit Gandhi
The Wharton School, University of Pennsylvania, Philadelphia, PA 19104
{amansin, kartikh}@wharton.upenn.edu
Department of Economics, University of Pennsylvania, Philadelphia, PA 19104
akgandhi@sas.upenn.edu

Instrumental variables (IVs) are a commonly used technique for causal inference from observational data. In practice, the variation induced by IVs can be limited, which yields imprecise or biased estimates of causal effects and renders the approach ineffective for policy decisions. We confront this challenge by formulating the problem of constructing instrumental variables from candidate exogenous data as a machine learning problem. We propose a novel algorithm, called MLIV (machine-learned instrumental variables), which allows learning of instruments and causal inference to be simultaneously performed from sample data. We provide formal asymptotic theory and show root-n consistency and asymptotic efficiency of our estimators hold under very general conditions. Simulations and application to real-world data demonstrate that the algorithm is highly effective and significantly improves the performance of causal inference from observational data.

Key words: Machine Learning, Causal Inference, Econometrics, Empirical Industrial Organization

1. Introduction

One of the most critical challenges in applied empirical research is to draw causal inference from observational data. A central difficulty is endogeneity of variables entering the causal relationship, arising from either omitted variable bias, simultaneity bias, sample selection bias, or measurement errors. Instrumental Variable (IV) methods are among the most frequently used techniques to address endogeneity bias in observational data. Instruments that are correlated with the endogenous variable but are otherwise not associated with the outcome variable can be used to partition the variance of the endogenous variable into endogenous and exogenous components. The method of instrumental variables is based on using the variation in the exogenous component of the endogenous variable induced by the the variation in the instrumental variable to make inference of causal effects.

In recent years, use of the IV method has come under criticism in multiple disciplines (e.g. Bound et al. (1995) and Young (2017) in Economics; Rossi (2014) in Marketing; Yogo (2004), Stock and Yogo (2002), and Hausman et al. (2005) in Finance) because IVs used in practice are often weakly correlated with the endogenous variables, e.g. the instruments used in practice induce limited variation in the endogenous variables leading to impractically large imprecision of estimates of causal effects. In the extreme case when the correlation becomes sufficiently weak, this leads to a formal "weak instruments" problem whereby standard asymptotics break down, and the estimated parameters are no longer even consistent and have a non-standard asymptotic distribution (see Stock and Wright (2000)). The bias becomes worse (Hausman et al. (2005)) when the researcher adds more weak instruments. The problem can be acute for industrial organization models (e.g., Berry et al. (1995), Arellano and Bond (1991), Arellano and Bover (1995), Blundell and Bond (1998), Hendel and Nevo (2006)), which involve estimating parameters from highly nonlinear functions and makes even the detection of weak IV's more challenging.

Constructing strong and valid instruments is, therefore, an important endeavor for causal inference from observational data. In this paper, we approach the problem of constructing strong instrumental variables from exogenous information in causal models as a (supervised) machine learning problem. We first formulate the choice of construction of instruments in a causal model as a decision problem that is amenable to the learning approach. The key empirical problem which arises for the econometrician that is distinct from the machine learner is that, in the typical applied context, the econometrician does not have the luxury of treating the sample in-hand as the training sample for the decision problem. Instead, the sample in-hand is typically the only information from which instruments must be constructed and causal inference must simultaneously be derived. We extend the standard learning framework to develop an algorithm we term "MLIV Algorithm", which allows training of instruments and causal inference to be simultaneously performed from sample data. The MLIVs exploits the variation in the data by treating each observation as a realization of a hypothetical test sample, where the decision rule trained from the remaining data can then be applied. Finally, we provide formal asymptotic theory and demonstrate semiparametric efficiency for machine learned instrument variables.

Existing approaches for addressing weakness in instruments have primarily been focused on constructions that approximate the optimal instrumental variables from the available

exogenous data. Optimal instruments (Amemiya (1974), Chamberlain (1987)) are the ones that provide asymptotically efficient estimators, but it is not feasible to construct optimal instruments from data. Much work (Newey (1990), Newey (1993), Donald et al. (2009), Berry et al. (1995), Blundell and Bond (1998)) has been done to approximate optimal instruments. For instance, in dynamic panel models literature, Blundell and Bond (1998) suggested how difference of consecutive lagged values can provide better inference than directly using the lagged values as instruments. Further, recently, some work (see Belloni et al. (2012)) has applied machine learning approaches to develop approximations to the optimal instruments. One challenge with these approaches is that they require explicit theoretical assumptions on optimal instrument structure, whose relevance will vary by the application.

The MLIV's we propose instead directly use data to learn the strong instruments without requiring any structure on the model and optimal instruments. We therefore provide an alternative and potentially complementary approach to the optimal instrument literature that addresses weak instrument challenges by using the available exogenous information to *algorithmically* optimize the efficiency criterion (Amemiya (1974), Chamberlain (1987), and Newey (1990)). Consequently, posing it as a computational learning problem gets rid of the need to rely on any explicit assumptions regarding the model or optimal instrument structure. Further, it allows access to a broader arsenal of ML and statistical methods. Indeed the contribution of our paper is to demonstrate that the optimal instrument variable problem can be *algorithimcally* solved regardless of the complexity of the model or optimal instrument structure. Further, we are able to provide very general conditions under which our method asymptotically achieves the semi-parametric efficiency bound.

We illustrate the effectiveness of the MLIVs in empirical environments consisting of both linear and nonlinear model parameters. Unlike traditional approaches (e.g., Belloni et al. (2012), Gandhi and Houde (2019), Blundell and Bond (1998)), which have only looked at a few specific models (among the vast set of nonlinear models) individually and tried to come up with analytical approximations for optimal instruments under various assumptions, we propose a method to learn strong instruments without imposing any explicit assumption on the optimality structure or even the model itself. We demonstrate Monte Carlo simulations for (i) a variety of linear causal models and, (ii) a specific nonlinear model Berry et al.

(1995) (henceforth BLP), and showcase, how MLIV's substantially improve the precision of inference in both cases.

For the linear cases, we demonstrate, how MLIVs can still lead to robust estimates even when core assumptions required by many other methods (e.g., "strong sparsity" on the optimal instrument structure as in Belloni et al. (2012)) break down. Since many of these assumptions are hard to come by in applied empirical environments, the MLIVs provide a flexible solution to the weak IV problem.

For the nonlinear case, we consider the BLP model. By allowing consumers to have unobserved preferences towards observed product characteristics, BLP allows for capturing very rich substitution patterns between differentiated products. However, there has been a growing question (see, e.g., Knittel and Metaxoglou (2014)) regarding the reliability of mixed-logit systems to identify consumer heterogeneity. Some work (Reynaert and Verboven (2014), Armstrong (2016)) has argued the potential source for identification challenges are weak instruments. Extant work has relied on mixed-logit model's structure to approximate the analytical form of optimal instruments. We show how the same MLIV formulation can deliver effective results in this setting while providing a more general solution to weak instrument issues for broader a set of models.

We finally end with applying our method to Acemoglu et al. (2001) (henceforth AJR), which faced criticism (Chernozhukov and Hansen (2008)) due to weakness of their instruments across various specifications they considered. We contrast their results against ours and demonstrate how MLIV instruments can help mitigate the curse of weak instruments and deliver stronger identification.

2. Related Literature

As discussed above, the standard existing approach to weak IV concerns is approached through approximation of optimal instruments. Most work on optimal instrument variables in linear models casts the problem as a selection problem among the available exogenous variables (and their transformations e.g., b-splines). Early work on instrument selection goes back to ((Kloek and Mennes 1960) and Amemiya (1966)) where they studied using "selected" principal components of the many available instruments to counter inference issues due to many instruments. Further work by (Kapetanios and Marcellino (2010)) proposed using factor analysis for decomposing the high-dimensional instruments onto a low-dimensional space. Both principal component analysis and factor analysis are not targeted

at approximating the optimal instruments, but rather at coming up with a low dimensional vector that summarizes the high-dimensional instruments, which could potentially yield (Amemiya (1966)) better performance in terms of bias and mean squared error.

Recent work on instrument selection assumes strong sparsity of the optimal instruments structure (i.e. a small set of the available IVs are valid and sufficient for first stage). Work by (Bai and Ng (2009)) demonstrate how boosting can be used for recovering the sparse structure but do not provide any formal proof. (Belloni et al. (2012)) explicitly shows how Lasso can be used for instrument selection among a large set of candidate instruments under the strong sparsity assumption. Further they are also able to prove theoretical consistency and other inference results for their IV estimator. Their proposed approach does not work as well when sparsity is violated, i.e. most instruments are weak, as it selects all of the weak IVs or drops them all. Unlike the extant literature, our learning approach to instrumental variables still exhibits asymptotic guarantees and does not rely on any sparsity assumption on the optimal instrument structure. Further, our approach allows the researcher to apply a broad arsenal of machine learning methods in constructing the instrumental variables.

Although work on estimating optimal instrument variables in nonlinear models is somewhat limited, the issues with weak IVs are well documented. Empirical Industrial Organization models are often characterized by a set of nonlinear moment conditions and a large number of (potentially weak) instruments. For instance, dynamic panel models in the production function literature (Melitz and Polanec (2015), Wang and Yu (2012), Ackerberg et al. (2006)) derive identification from lagged values of the time series, which are often seen as weak instruments. Blundell and Bond (1998) discuss how traditional estimators (Arellano and Bover (1995)) suffer from weak instrument issues and propose transformations of lagged values that deliver more efficient results. Similarly, Armstrong (2016) argues that many of the empirical challenges associated with identifying substitution patterns in random-coefficient (or mixed) logit models such as Berry et al. (1995) are due to weak instruments. Gandhi and Houde (2019) discuss methods of analytically approximating (using explicit assumption of random-coefficients model) the optimal instruments for BLP. Our approach to strong IV's provides a more general and more data-driven solution to such identification challenges without relying on the explicit details (e.g., AR(1), logit) of the model.

Our paper is also related to sample-splitting in the econometric literature. Angrist et al. (1993) and Angrist et al. (1999) introduced the idea of sample splitting to estimate first stage estimates, and argued for their superior small sample properties. Sample splitting plays a key role in defining the properties of our estimators by introducing orthogonality conditions in our asymptotic proofs. We also relate to Ayyagari (2010), Newey and Robins (2018) and Chernozhukov et al. (2016), which make use of sample splitting to identify nuisance parameters in partially linear models.

Finally, at a high level, our work is also broadly related to emerging work on machine learning and causal inference. Belloni et al. (2014) propose a double selection procedure, where they first use Lasso to select among high dimensional covariates to improve the performance of treatment effect estimator. Wager and Athey (2018) discuss how random forest can be used to estimate heterogeneous treatment effects. Many recent methods like "Deep IV" (see Hartford et al. (2017)), "Dual IV" (see Muandet et al. (2019)), "Deep GMM" (see Bennett et al. (2019)), "Adversarial GMM" (see Lewis and Syrgkanis (2018)), and "Kernal IV" (see Singh et al. (2019)) have come up that involve using machine learning for causal identification. However, they study a fundamentally different problem of non-parametric IV regression i.e. estimate f such that $y = f(x) + \epsilon$. Relatedly, other methods like (Fan and Liao (2014) and Gautier and Rose (2011)) describe how machine learning methods can be used for inference with high-dimensional endogeneous regressors.

3. Learning Optimal Instrumental Variables

Consider the causal model framework characterized by P moment conditions:

$$m(\theta_0) = \mathbb{E}\left[f(x_i, z_i; \theta)\right]_{\theta=\theta_0} = 0$$

where θ is $K \times 1$ vector of parameters; $f(\cdot)$ is $P \times 1$ vector of (nonlinear) functions; x_i refer to model variables (data), and z_i refers to the instrument variables. We consider the specific subset set of estimators referred to as instrumental variable estimators such that the moment conditions have the following form:

$$f(x_i, z_i; \theta) = \underbrace{\xi(x_i; \theta)}_{1\times 1} \cdot \underbrace{z_i}_{P\times 1} \tag{1}$$

and satisfy strong exogeneity,

$$\mathbb{E}(\xi(x_i; \theta)|Z) = 0 \tag{2}$$

For instance, for a linear regression ($y = \theta x + \epsilon$) moments are given by $\mathbb{E}\left[(y - \theta x) \cdot z\right]$. The generalized method of moments estimator (Hansen (1982)) $\hat{\theta}_{GMM}$ minimizes the following objective function.

$$Q_N(\theta) = [\frac{1}{N}\sum_{i=1}^{i=N} \mathbf{f(x_i, z_i, \theta)}]' \mathbf{W_N} [\frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{i=N}} \mathbf{f(x_i, z_i, \theta)}],$$

where the $P \times P$ weighting matrix W_N is a symmetric positive definite, possible stochastic with finite probability limit. N is used to indicate the estimator's dependence on sample. Under some sufficient conditions, the GMM estimator is consistent and aymptotically normal.

3.1. Learning Criterion For Optimal Instrumental Variables

Strong exogeneity implies that any choice of the transformation $H(\cdot)$ of the instruments z also corresponds to a set of instruments for the GMM problem i.e.,

$$\mathbb{E}\left[\xi(x_i; \theta)|Z\right] = 0 \implies \mathbb{E}\left[\xi(x_i; \theta) \cdot H(z_i)\right] = 0$$

The problem faced by an econometrician is to choose an "*efficient*" rule $H(\cdot)$. The most *efficient* set of instruments $H(z_i)$, referred to as optimal instruments (Chamberlain (1987), Amemiya (1974), Hansen (1985)), are such that:

$$H = \arg\min_h V(\theta_0; h(z)) \tag{3}$$

where V refers to the asymptotic variance of the parameters θ_0 minimized over all possible functions h. Keeping computational concerns aside, estimating optimal H is infeasible as it depends on asymptotic variance V (which is unknown) and true model parameters θ_0 (which can be estimated once the instruments are estimated).

To get around these issues, extant work (Berry et al. (1995), Blundell and Bond (1998), Gandhi and Houde (2019)) has often used approximations of optimal instruments based on the structure of the specific model. For instance, Berry et al. (1995) argue that the observed characteristics of the product, the sum of characteristics of the other products produced by the same firm, and the sum of characteristics of the rival products are good approximations of optimal instruments. Similarly, in context of dynamic panel models, Blundell and Bond (1998) propose using lagged differences over lagged values as better approximations. And in linear models such as 2SLS, it is not uncommon for researchers

to try quadratic and cubic forms of the available instruments in search for an efficient estimator. However, it is in general hard to know which transformations are more efficient without making additional assumptions about the optimal instrument structure.

In contrast, we take a completely data-driven and algorithmic approach to estimate efficient instruments from data, thereby avoiding any reliance on model or optimal instrument structure assumptions. We denote $H(Z;\eta)$ as a class of instrumental variable functions parameterized by η. Such functional approximation models could encompass any of a wide variety of ML approaches ranging from polynomials and splines to sparse regressions, and neural nets.

We propose to directly minimize the variance of the estimator (trace of $\hat{V}$) given by:

$$\begin{aligned} &\min_{\eta} \left[\mathcal{L}(\eta) = \text{tr}(\hat{V}_N(\theta;\eta))\right] \\ \text{s.t. } \theta \quad &\in \quad \arg\min_{\theta'} \hat{Q}_N(\theta';\eta) \end{aligned} \tag{4}$$

where $\hat{Q}_N$ refers to the sample GMM criterion given by:

$$\hat{Q}_N(\theta;\eta) = \left[\frac{1}{N}\sum_{i=1}^{i=N} f(x_i, H(z_i;\eta);\theta)\right]' \mathbf{W_N} \left[\frac{1}{N}\sum_{i=1}^{i=N} f(x_i, H(z_i;\eta);\theta)\right]$$

and $\hat{V}$ is variance-covariance matrix of model parameters θ.

$$\hat{V}_N(\theta;\eta) = (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}\hat{S}\hat{W}\hat{G}(\hat{G}'\hat{W}\hat{G})^{-1}$$

$$\hat{G}_N = \frac{1}{N}\sum_{i=1}^{i=N} \frac{\partial}{\partial\theta}\mathbf{f}(\mathbf{x_i}, \mathbf{H}(\mathbf{z_i};\eta);\theta)$$

$$\hat{S}_N = \frac{1}{N}\sum_{i=1}^{i=N} \mathbf{f}(\mathbf{x_i}, \mathbf{H}(\mathbf{z_i};\eta);\theta)\mathbf{f}(\mathbf{x_i}, \mathbf{H}(\mathbf{z_i};\eta);\theta)'$$

This sets up a bi-level optimization problem involving the estimation of the model parameters using the GMM criterion and the efficient instruments using the variance minimization criterion. Each optimization problem depends on the other. Recent work in computer science (see Gould et al. (2016), Samuel and Tappen (2009)) has demonstrated how gradient descent methods can be used to solve similar bi-level optimization problems. For instance, generative adversarial networks (Li et al. (2018), Goodfellow et al. (2014)) are trained solving a bi-level program. Zoph and Le (2016) demonstrates how search over

optimal neural network architecture can be posed as a bi-level program and be tractably solved. We follow a similar procedure and apply gradient descent as outlined next.

Algorithm 1 Gradient Descent for Learning ML Instruments

1. Select a machine learning method H.
2. Initialize $\eta^{(0)}$ to get an initial estimate of instruments $H(z;\eta^{(0)})$.
3. Estimate $\hat{\theta}^{(0)}$ minimizing GMM criterion using the generated instruments.
4. Update η such that:

$$\eta_i = \eta_{i-1} - \nu \left[\frac{\partial \mathcal{L}}{\partial \eta} + \frac{\partial \mathcal{L}}{\partial \theta} \frac{\partial \theta}{\partial \eta} \right]_{\theta=\theta^{(i-1)}, \eta=\eta^{(i-1)}}$$

5. Update the new set of instruments $H(z;\eta_i)$ and solve the GMM criterion to compute θ_i using the set of new instruments.
6. Iterate until stopping criterion is reached.

A potential stopping criterion for our algorithm is convergence of the parameters. In machine learning, *early stopping* is employed to avoid overfitting. So gradient descent iterations are not carried out till convergence but rather stopped once the loss measure computed on hold-out data starts increasing. Further, depending on the complexity of GMM criterion, GD iterations can sometimes be carried out lazily i.e., instead of re-solving the GMM optimization criterion, θ can be treated constant for a certain number of iterations and η can be updated simply as:

$$\eta_i = \eta_{i-1} - \nu \left[\frac{\partial \mathcal{L}(\theta^{(j)})}{\partial \eta} \right]_{\eta=\eta^{(i-1)}}$$

3.1.1. Gauss Newton Regression: To leverage the full power of MLIV approach we cast the GMM optimization problem as Gauss-Newton regression which allows to conduct the GMM step iteratively and lazily. Gauss newton approximation allows us to lazily solve

[0] Using implicit function theorem we can show that

$$\frac{\partial \theta}{\partial \eta} = -\left[\hat{Q}_{\theta\theta}(\eta, \theta(\eta)) \right]^{-1} \left[\hat{Q}_{\eta\theta}(\eta, \theta(\eta)) \right]$$

where $\hat{Q}_{\theta\theta} \doteq \nabla^2_{\theta\theta} \hat{Q}(\theta;\eta) \in \mathbb{R}^{K \times K}$ and $\hat{Q}_{\eta\theta} \doteq \frac{\partial}{\partial \eta} \nabla_\theta \hat{Q}(\theta;\eta)$.

the GMM optimization problem i.e., avoid the need for convergence on θ to update η. The sample moments can be approximated by using taylor expansion as follows:

$$\begin{aligned} f_N(\theta) &= f_N(\hat{\theta}) + G_N(\hat{\theta})(\theta - \hat{\theta}) + Residual(\theta) \\ &\approx \Big[f_N(\hat{\theta}) - G_N(\hat{\theta})\hat{\theta}\Big] - \Big[-G_N(\hat{\theta})\Big]\theta \\ &= v - F\theta \end{aligned} \tag{5}$$

where,

$$\begin{aligned} v &= f_N(\hat{\theta}) + F\hat{\theta} \\ F &= -G_N(\hat{\theta}) \end{aligned} \tag{6}$$

Hence, the GMM objective function can be written using the linear moment approximation as follows:

$$\arg\min_{\theta} Q(\theta, \hat{W}) = \Big[v - F\hat{\theta}\Big]' \hat{W} \Big[v - F\hat{\theta}\Big] \tag{7}$$

and the linear GMM estimator has closed form solution:

$$\hat{\theta}^i = (F'\hat{W}F)^{-1}F'\hat{W}v \tag{8}$$

Thus the iterative step for the GMM optimization is given as follows:

$$\hat{\theta}^i = \hat{\theta}^{i-1} + (F'\hat{W}F)^{-1}F'\hat{W}f_N(\hat{\theta}^{i-1}) \tag{9}$$

Hence, the GMM problem can be solved iteratively and made part of MLIV gradient descent algorithm. We detail the complete algorithm in Algorithm 2.

Algorithm 2 Gradient Descent for Learning ML Instruments using GNR

1. Select a machine learning method H.
2. Initialize $\eta^{(0)}$ to get an initial estimate of instruments $H(z;\eta^{(0)})$.
3. Estimate $\hat{\theta}^{(0)}$ by minimizing GMM criterion to convergence using the generated instruments.
4. **M - IV GD Steps**: Conduct M lazy update steps for η such that:

$$\eta_j = \eta_{j-1} - \nu \left[\frac{\partial \mathcal{L}(\theta^{(i-1)})}{\partial \eta} \right]_{\eta=\eta^{(j-1)}}$$

5. Update the new set of instruments $H(z;\eta_i)$
6. **N - GMM GD Steps**: Estimate $\hat{\theta}^{(i)}$ by conducting N number of GNR steps with the new set of generated instruments:

$$\hat{\theta}^j = \hat{\theta}^{j-1} + (F'\hat{W}F)^{-1}F'\hat{W}f_N(\hat{\theta}^{j-1})$$

7. Iterate until stopping criterion is reached.

Cross-validation: Our algorithm implies that any machine learning method that relies on gradient descent can be used to construct the instruments. This includes a large family of methods including Lasso, Ridge, Neural Nets, gradient boosting, etc. We believe there is no universal ML method across empirical contexts. Further, it is not necessary for an econometricians to make that choice a priori. A common approach for ML method selection in predictive modeling is cross-validation (Webb (2003)) . Webb (2003) discusses how cross-validation entails partitioning data set of size n samples into two parts. The model parameters are estimated using one set (by minimising some optimisation criterion) and the goodness-of-fit criterion evaluated on the second set. In k-cross-validation the second set consists of $\lfloor \frac{n}{k} \rfloor$ samples. The idea behind k-cross-validation is to create a number of partitions (validation datasets) from the training dataset and fitting the model to the training dataset (sans the validation data). The model is then evaluated against each validation dataset and the results are averaged to obtain the cross-validation error. The econometrician can compare the performance of various ML methods and choose the one which gives the lowest cross-validated error in the validation data. Most machine learning algorithms also involve varying degrees of regularization, usually through hyperparameters

Table 1 Examples of Criterion for ML Algorithms

	Objective Criterion	Hyperparameters	$H(z;\eta)$
Lasso	$\mathcal{L}(\eta)+\lambda\|\|\eta\|\|_1$	λ	$\mathbf{z}\eta$
Ridge	$\mathcal{L}(\eta)+\lambda\|\|\eta\|\|^2$	λ	$\mathbf{z}\eta$
Elastic Nets	$\mathcal{L}(\eta)+\lambda_1\|\|\eta\|\|_1+\lambda_2\|\|\eta\|\|^2$	λ_1 and λ_2	$\mathbf{z}\eta$
Kernel Regularized Least Squares	$\mathcal{L}(\eta)+\lambda\|\|H\|\|_K^2$	λ	$\sum_{i=1}^{i=N}\eta_i k(z_i,z)$
Neural Nets	$\mathcal{L}(\eta)$	Network Architecture[1]	$\text{nnet}(\mathbf{z};\eta)$

to avoid overfitting. We recommend using cross-validation to tune the hyperparameters (including number of iterations) by splitting the data into training and validation sets. Table (1) discusses a few machine learning algorithms and the corresponding regularized objective function. The regularized objective function contains hyperparameters, which can be tuned using cross-validation on the validation set. If econometrician is working with a large dataset or the problem is computationally expensive, she can also consider just hold-out validation (Webb (2003)) that entails partitioning into two independent samples i.e training and validation.

However, learning H by minimizing (4) directly might lead to inconsistent estimates, because if η is very high dimensional, then $\mathbb{E}\left[\xi(x_i;\theta_0)\cdot H(z_i;\eta(\mathbf{D}))\right]\neq 0$. The problem of inconsistency arises because we use the same data D, to learn optimal instruments and then use the learned instruments to drive inference on same D. Next we discuss, how by embedding our learning problem within sample-splitting can address these concerns and preserve standard asymptotic theory in instrumental variables.

3.2. The MLIV Estimator Recipe

We introduce a general class of estimators termed as MLIV estimators. Every MLIV instrument is characterized by a choice of "learning" algorithm. The general procedure to estimate an MLIV instrument proceeds as follows:

1. **Outer Loop**: Split the data into a K-fold partition, such that each partition D_k has size $\left\lfloor\frac{n}{k}\right\rfloor$. For each partition k, define D_k^c to be the excluded data.

[1] Other hyperparameters include number of iterations, step size (ν) for GD, and M and N for GNR-MLIV.

2. **Inner Loop**: For each partition k, learn the optimal instrument function $H_k(\cdot;\eta_k)$, as described in previous section using only the excluded data D_k^c. The hyperparameters pertinent to H should be tuned through cross-validation (i.e. inner loop cross-validation) using only excluded data D_k^c.
3. For each partition k, generate new instruments such that $\hat{z}_k = H(z_k;\eta_k)$ i.e., transformation learned on D_k^c is applied on D_k.
4. Construct moment conditions using the set of generated instruments. Formally parameters of interest are estimated as follows.

$$\hat{\theta}_{ML} \in \arg\min[\frac{1}{N}\sum_{i=1}^{i=N}\mathbf{f}(\mathbf{x_i},\mathbf{h}(\mathbf{z_i};\eta_{\mathbf{k(i)}});\theta)]'\mathbf{W_N}[\frac{\mathbf{1}}{\mathbf{N}}\sum_{\mathbf{i=1}}^{\mathbf{i=N}}\mathbf{f}((\mathbf{x_i},\mathbf{h}(\mathbf{z_i};\eta_{\mathbf{k(i)}});\theta)]$$

Cross-learned instruments by construction preserve asymptotic theory as we discuss next.

3.3. Asymptotic Distribution

In this section, we provide formal asymptotic theory for MLIV estimators. We build on recent work by (Chernozhukov et al. (2016)) and provide formal conditions for the asymptotic efficiency of MLIV estimators. Our proofs only require very general mean-square consistency conditions. Many off the shelf machine learning algorithms have been shown to be mean square consistent (for instance, see Scornet et al. (2015) for random forests, see Steinwart (2005) for support vector machines for similar results). Moreover, the applied researcher can still use ML toolkits for which mean square consistency results are not available. The trade-off is that she might not achieve the semi-parametric efficiency bound (i.e., higher standard errors), but nevertheless, still, achieve very effective performance for a broad range of empirical settings with minimal assumptions.

ASSUMPTION 1. *There exists some* η^* *such that for each* $k = 1, 2 \ldots K$ $\int \Big[h(z,\eta^*) - h(z,\hat{\eta}_k)\Big]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0$.

The literature assumes much stronger variants of this assumption to show consistency. For instance, traditional two-step estimators (also known as M-step estimators) that involve estimation of some first step (in our case $\hat{\eta}$) assume that $\hat{\eta} \xrightarrow{p} \eta^*$ for some η^* (see Wooldridge (2010)) to show consistency. However, for MLIV estimator we require a much weaker assumption i.e., existence of a mean square limit η^*. Note that $\hat{\eta} \xrightarrow{p} \eta^*$ directly implies existence of mean square limit under mild regularity conditions on h, however it is not necessarily true the other way round.

LEMMA 1. *If assumption 1 is satisfied and for every* $k = 1, 2 \ldots, K$ $\mathbb{E}[b(w, \theta, \hat{\eta}_k)] < \infty$, *where* $b(w_i, \theta, \hat{\eta}_k) = \xi(x_i, \theta)^4 (h(z_i, \eta^*) - h(z_i, \hat{\eta}_k))^2$ *and* $\theta \in \Theta$ *such that* Θ *is compact; Then,*

$$\hat{f}(\theta) \xrightarrow{p} \mathbb{E}[f(w_i, \theta, \eta^*)]$$

where,

$$\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} f(w_i, \theta, \hat{\eta}_k)$$

We assume that the fourth moments are bounded. This is a very standard assumption used in the literature. With this lemma in place, consistency of MLIVs can be established as discussed next.

THEOREM 1. *Let* $\{x_i, z_i\} \equiv w_i$ *be i.i.d random variables distributed by* $\mathbb{P}$. *If*

1. *Identification:* $f^{\eta^*}(\theta) = \mathbb{E}\left[f(w_i, \theta, \eta^*))\right] = 0$ *if and only if* $\theta = \theta_0$
2. $f^{\eta^*}(\theta)$ *is continuous at each* θ *with probability one*
3. $E\left[\sup_{\theta \in \Theta} ||f(w_i, \theta, \eta^*))||\right]$ *is finite;* $\theta \in \Theta$ *such that* Θ *is compact; and* $W_N \xrightarrow{p} W$ *positive definite;*
4. *Assumption (1) is satisfied and* $\mathbb{E}[b(w_i, \theta, \hat{\eta}_k] < \infty$ *for each* $k = 1, 2 \ldots, K$ *for* $\theta \in \Theta$;

then $\hat{\theta}_N \xrightarrow{p} \theta_0$.

Once the results of lemma 1 are established, it immediately follows $\sup_{\theta \in \Theta} ||\hat{f}(\theta) - \mathbb{E}[f(w_i, \eta^*, \theta)]|| \xrightarrow{p} 0$ (see lemma 2.4 of Newey and McFadden (1994)). Next, consistency can shown in a straightforward way (for technical details see theorem 2.1 of Newey and McFadden (1994)).

Next, we demonstrate that MLIV estimator asymptotically achieve the efficiency bound i.e.,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathbb{N}(0, V_0)$$

ASSUMPTION 2. *For each* $k = 1, 2 \ldots K$ $\int \left[h(z_i, \eta_0) - h(z_i, \hat{\eta}_k)\right]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0$.

This assumption implies that the instruments are estimated in a mean square consistent manner. This assumption along with MLIV procedure allows us to show asymptotic efficiency of our estimator. Many off the shelf machine learning algorithms have been shown to be mean square consistent. For instance, see Scornet et al. (2015) for random forests, see Steinwart (2005) for support vector machines for similar results.

LEMMA 2. *If the instruments are estimated in a mean square consistent manner for each* $k = 1, 2 \dots K$ *i.e* $\int \Big[h(z_i, \eta_0) - h(z_i, \hat{\eta}_k)\Big]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0$ *then for each* $k = 1, 2 \dots K$*;*

$$\int \Big[f(w, \theta_0, \eta_0) - f(w, \theta_0, \hat{\eta}_k)\Big]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0 \tag{10}$$

See Appendix for proof.

LEMMA 3. *If assumption 2 is satisfied then estimation of* $\hat{\eta}$ *does not have any asymptotic effect on the moment conditions for a K fold MLIV estimator i.e*

$$\sqrt{n}\hat{f}(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} f(x_i, z_i, \theta_0, \eta_0) + o_p(1) \tag{11}$$

where $\hat{f}$ *is given as*

$$\hat{f}(\theta_0) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} f(x_i, z_i, \theta_0, \hat{\eta}_k) \tag{12}$$

η_k *is the estimator for the* k^{th} *fold estimated using data excluding the* k^{th} *fold (i.e.,* D_k^c*).*

See Appendix for proof. The MLIV procedure allows for orthogonality restrictions that play an important role in proving asymptotic efficiency.

Next, another result required for asymptotic normality result is the convergence of the jacobian term i.e $\hat{G}(\hat{\theta}) = \frac{\partial \hat{f}(\hat{\theta})}{\partial \theta}$ to $G = \mathbb{E}[\frac{\partial f(w_i, \theta, \eta_0)}{\partial \theta} | \theta = \theta_0]$. Next we state the formal assumptions required for convergence of jacobian.

ASSUMPTION 3. *G exists and there is a neighbourhood* $\mathcal{N}$ *of* θ_0 *such that*

1. *For each* $k = 1, 2, \dots, K$*, and* $i \in \mathbb{N}^{[1,P]}$ *and* $j \in \mathbb{N}^{[1,D]}$ $\int |\frac{\partial f(w, \theta_0, \eta_0)_i}{\partial \theta_j} - \frac{f(w, \theta_0, \hat{\eta}_k)_i}{\partial \theta_j}| d\mathbb{P}_0(w) \xrightarrow{p} 0$
2. *for all* $||\eta - \eta_0||$ *small enough* $f(w_i, \theta_0, \eta)$ *is differentiable in* θ *on* $\mathcal{N}$ *with probability approaching 1.*
3. *there exists* $\xi > 0$ *and some* $u(w_i)$*, such* $\mathbb{E}[u(w_i)] = O_p(1)$ *(or* $\mathbb{E}[u(w_i)] < \infty$*), and for* $\theta \in \mathcal{N}$ *and* $||\eta - \eta_0||$ *small enough, such that*

$$\Big|\Big|\frac{\partial f(w, \theta, \eta)_i}{\partial \theta_j} - \frac{f(w, \theta_0, \eta)_i}{\partial \theta_j}\Big|\Big| \leq u(w_i)||\theta - \theta_0||^{\xi}$$

Assumption 3.2 and 3.3 are standard assumptions used in the literature. Assumption 3.1 directly follows if h is estimated in a mean square consistent manner (see lemma 4).

LEMMA 4. *If the instruments are estimated in a mean square consistent manner for each* $k = 1, 2 \dots K$ *i.e* $\int \left[h(z_i, \eta_0) - h(z_i, \hat{\eta}_k)\right]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0$ *and* $\mathbb{E}\left[\left(\frac{\partial \xi_i(\theta_0)}{\partial \theta_j}\right)^2\right] < \infty$ *then for each* $k = 1, 2 \dots K$*;*

$$\int \left| \frac{\partial f(w, \theta_0, \eta_0)_i}{\partial \theta_j} - \frac{f(w, \theta_0, \hat{\eta}_k)_i}{\partial \theta_j} \right| d\mathbb{P}_0(w) \xrightarrow{p} 0 \tag{13}$$

See Appendix for proof.

LEMMA 5. *If assumption 3 is satisfied and for any* $\bar{\theta} \xrightarrow{p} \theta$*,* $\hat{f}$ *is differentiable at* $\bar{\theta}$ *with probability approaching one, then* $\hat{G}(\bar{\theta}) \xrightarrow{p} G$*. where* $\hat{G}(\bar{\theta}) = \frac{\partial \hat{f}(\bar{\theta})}{\partial \theta}$ *and* $G = \mathbb{E}\left[\frac{\partial f(w_i, \theta, \eta_0)}{\partial \theta} | \theta = \theta_0\right]$.

See Appendix for proof. Once this lemma is established, this along with other lemmas allows us to show asymptotic normality.

THEOREM 2. *Let* $\{x_i, z_i\} \equiv w_i$ *be i.i.d random variables distributed according to some* $\mathbb{P}$*. If*

1. $\hat{W} \xrightarrow{p} W$ *and* W *is positive semi-definite;* $G'WG$ *is non-singular for* $G := \mathbb{E}\left[\nabla_\theta f(w_i, \theta_0, \eta_0)\right]$ *and* $S := \mathbb{E}\left[f(w_i, \theta_0, \eta_0) f(w_i, \theta_0, \eta_0)'\right]$
2. *For each* $k = 1, 2 \dots, K$ $f(w_i, \theta, \hat{\eta}_k)$ *is continuously differentiable in some neighbourhood* $\mathcal{N}$ *of* θ_0*.*
3. *For each* $k = 1, 2 \dots, K$ $\mathbb{E}\left[\sup_{\theta \in \Theta} ||\nabla_\theta f(w_i, \theta, \hat{\eta}_k)||\right] < \infty$
4. *Assumption 1-3 is satisfied.*

If the GMM estimator is consistent, then

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, ((G'WG)^{-1} G'WSWG(G'WG)^{-1}))$$

Once lemma 1-5 are established, we can invoke Newey and McFadden (1994)'s theorem 3.4 and asymptotic normality follows in a straightforward way. In the next theorem we show that asymptotic variance V_0 is consistently estimated by $\hat{V}$.

THEOREM 3. *If assumption 3 and 2 are satisfied and* $G'WG$ *is nonsingular then* $\hat{S} \xrightarrow{p} S$ *and* $\hat{V} \xrightarrow{p} V_0$

See Appendix for proof.

In this theorem we show that MLIV estimators achieve the semi-parametric efficiency bound and the sample analog of variance estimator consistently estimates the asymptotic variance V_0.

Finally, let us consider the case when the optimization routine is not mean square consistent or the routine can not guarantee the global optima. In such a scenario as long as assumption 1 is not violated i.e the estimation procedure has "some" mean square limit, then all asymptotic properties (except efficiency) still hold and can be shown using the same arguments. The only trade-off is that we might get higher standard errors than if we knew the true optimal instrument function H. However, from an applied econometrician's perspective using MLIV's will nonetheless provide significant gains over just using raw exogenous data.

Next, we demonstrate through simulations in Section 5 that MLIVs exhibit desirable small sample properties and could help alleviate the problem of weak instruments for a wide variety of applied problems.

4. Linear IV and the Learning Heuristic

Next, we consider the special case of linear causal models and discuss an assumption to simplify the estimation of MLIV instruments. Consider the following linear causal model framework

$$y = \theta^{(0)} + (x^{(1)})'\theta^{(1)} + (x^{(2)})'\theta^{(2)} + \epsilon \tag{14}$$

The observed covariates $x_i \equiv (x_i^{(1)}, x_i^{(2)})$ possesses both exogeneous $(x_i^{(1)})$ and endogeneous $(x_i^{(2)})$ variables. The econometric error ϵ satisfies:

$$\mathbb{E}(\epsilon \mid x^{(1)}, z) = 0 \implies \mathbb{E}(H(x^{(1)}, z) \cdot \epsilon) = 0 \tag{15}$$

where z_i refers to the exogenous variables from which instruments will be constructed. In the case of conditional homoscedasticity[2] considered by Chamberlain (1987), the optimal such rule that minimize the variance of $\hat{\theta}_{GMM}$ are given by

$$H_i\left(x_i^{(1)}, z_i\right) = \mathbb{E}[x_i^{(2)} | x_i^{(1)}, z_i]$$

Thus, the MLIV optimization routine, simplifies to that of a supervised machine learning problem, with the choice of the rule H_i as: predict $x^{(2)}$ from $\left\{x^{(1)}, z\right\}$. This makes the

[2] Note that Belloni et al. (2012) demonstrate results for their Lasso based estimator under similar assumptions.

estimation of the linear MLIV simpler and allows the use of most off-the-shelf ML toolkits. Further, in the linear case with homoscedasticity the asymptotic efficiency of MLIVs is directly implied from mean square consistency of the ML method.

Choosing ML Methods: As before, cross-validation can also be used to select ML methods, along with hyperparameter tuning. For every fold k in our algorithm, we can further partition the excluded data D_k^c into training and validation sets. The econometrician can compare the performance of various ML methods and choose the one which gives the lowest cross-validated error in the validation data.

5. Simulation Studies

In this section, we conduct multiple simulation studies, where we demonstrate how MLIVs can help mitigate the curse of weak instruments. We consider two cases: (ii) linear causal model with many weak instruments, none of which are strong, and (ii) random coefficients logit model of demand.

5.1. Many Instruments with Strong Sparsity

We first consider the case in which the first stage is sparse, that is, there are many available instruments but there exist only a small set of strong instruments whose identity is unknown (Belloni et al. (2012)). Availability of many candidate instruments is a commonplace in many empirical problems. For instance, a very common problem is to estimate the effect of expenditure on advertisements, on product demand. However, in many cases researchers, have access only to aggregate level demand data, that creates potential endogeneity concerns with respect to ad spends. To correct for this, researchers (see Dinner et al. (2013)) have used online advertising cost in similar but different markets as instruments. Such instruments also commonly know as Hausman-type instruments (see Hausman et al. (1994) and Hausman (1996)) are very common in literature and have been used in many other problems as well. However, such instruments always come under the scanner due to their high-dimensional nature and potential weakness. In this simulation, we consider a similar cases, where researcher has access to many instruments, but out of all candidate instruments, only a small subset are strong enough for causal estimation. This is also where the scope of our method overlaps with existing literature on instrument selection (Belloni et al. (2012), Bai and Ng (2009), Kapetanios and Marcellino (2010) and Amemiya (1966)). Compared to prior work which largely looks at the many-instruments problem as

a selection problem, our method, on the other hand, attempts to directly learn the optimal function H. We demonstrate that this alternative framing of the problem delivers many benefits to the econometrician.

To contrast our results with the Post Lasso method proposed by Belloni et al. (2012), we consider a simulation design similar to theirs (which further let us use their code). Specifically, the data $D \equiv \{(y_i, x_i \equiv (x_i^{(1)}, x_i^{(2)}), z_i)_i \mid i = 1, 2, ..n\}$ is generated using the following DGP.

$$y = \theta d + e \tag{16}$$

$$d = H(z)\Pi + \nu \tag{17}$$

$$\begin{pmatrix} e \\ \nu \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e & \sigma_{e\nu} \\ \sigma_{\nu e} & \sigma_\nu \end{bmatrix}\right) \tag{18}$$

$$z \sim N_J(0, 1) \tag{19}$$

where $\theta = 0.75$ is the parameter of interest. We also let $\sigma_\nu = 1$, $\sigma_e = 1$ and $J = 500$ for all our simulations.

For other parameters we consider various settings. We run simulations for sample size n of $1,500$ and $1,000$; we consider two different values for $\sigma_{e\nu}$: 0.3, and 0.5. For $H(\cdot)$, we assume, $H(z) = [\underbrace{1, 1, .., 1, 1}_{s}, \underbrace{0, 0, .., 0}_{J-s}]z$, with $s = 25$ and 50. The strong sparsity assumption is met because a small subset of the 500 candidate instruments are valid. We use $\Pi = \frac{5}{s}$ which ensures that the instruments have the same impact on d independent of the value of s.

For each setting of simulation parameters, Table (2) reports results for 2SLS with Standard, MLIVs and Belloni Lasso (Belloni et al. (2012)). To generate MLIVs, we use 3-cross learner estimator (i.e. cross-learner with $k = 3$ folds) with Lasso as the ML method to estimate the function $H(\cdot)$. By choosing Lasso as our ML method in stage 1 of 2SLS, we can more directly compare our method with the post Lasso method for instrument selection (Belloni et al. (2012)). For all methods we report mean bias, mean standard error and root mean square error for each estimated parameter.

As expected, we find improvements in both bias and efficiency with increase in sample size for all three estimators. However, across the board, MLIVs result in far lower bias compared to both standard instruments and Post Lasso. Further, as suggested by (20) we

see a significant increase in bias for post lasso for $s = 50$ compared to $s = 25$. In contrast, MLIVs are not as impacted by an increase in the number of relevant instruments.

It is interesting to note that MLIV is associated with lower bias than (Belloni et al. (2012)) even though both techniques use Lasso as the underlying ML method. This could be, since we directly use $\hat{H}$, for parameter estimation instead of multiple selected instruments. This can be readily seen from the expression of 2SLS small-sample bias derived by Hahn and Hausman (2005):

$$E[\theta_{2SLS}] - \theta \approx \frac{J\sigma_{e\nu}}{nR^2 var(x^{(2)})} \tag{20}$$

where R^2, denote the first stage R^2 and J denotes the number of instruments. The expression dictates that as long as additional instruments do not add much to first stage R^2, it is strictly better to use fewer instruments. So given we can get a good approximation of the optimal instruments (similar R^2 as Belloni et al. (2012) but we use only one instrument) we should end up with lower bias compared to selection methods. Further, within the literature on instrument selection, formal asymptotics have only been shown for Lasso (Belloni et al. (2012)), but as shown our method results in asymptotically efficient estimates for a wide range of ML and statistical methods. This allows us to use a wide range of ML methods, which in turn can improve first stage R^2 and in turn further reduce the bias.

5.2. Many Weak Instruments

We next consider the case of a linear causal model with many instruments, all of which are weak (i.e., Belloni et al. (2012)'s sparsity assumption breaks down). As discussed before, a lot of empirical problems are characterized by availability of many candidate instruments, and in many of those problems it might be the case that all candidate instruments are important. To our best knowledge, current literature offers limited guidance on how to estimate strong instruments in such scenarios. The issue with many weak instruments is that, when the sparsity assumption breaks down, variable selection methods like lasso or boosting tend not to select any variable or select all variables, which leads to poor asymptotics. Since MLIVs allow for using any machine learning algorithm without compromising on asymptotic theory, it provides for a potential solution to the many-weak instrument problem. The underlying assumption that allows for identification in the case of many weak instruments is that even though each instrument individually is weak, there may

Table 2 Simulation (b)

	True		Standard Instruments		ML Instruments			Belloni Lasso		
		Bias	St. Err.	RMSE	Bias	St. Err.	RMSE	Bias	St. Err.	RMSE
	$n=1,000$	$s=50$	$\sigma_{e\nu}=0.5$							
θ	0.75	0.253	0.029	0.254	0.000	0.074	0.068	0.087	0.096	0.301
	$n=1,000$	$s=25$	$\sigma_{e\nu}=0.5$							
θ	0.75	0.170	0.024	0.172	-0.003	0.034	0.031	0.022	0.032	0.031
	$n=1,500$	$s=50$	$\sigma_{e\nu}=0.5$							
θ	0.75	0.199	0.026	0.200	0.001	0.047	0.053	0.042	0.068	0.269
	$n=1,500$	$s=25$	$\sigma_{e\nu}=0.5$							
θ	0.75	0.122	0.021	0.123	-0.001	0.027	0.025	-0.002	0.025	0.027
	$n=1,000$	$s=50$	$\sigma_{e\nu}=0.3$							
θ	0.75	0.155	0.030	0.158	0.003	0.074	0.064	0.027	0.104	0.277
	$n=1,000$	$s=25$	$\sigma_{e\nu}=0.3$							
θ	0.75	0.106	0.025	0.108	-0.004	0.034	0.033	0.015	0.031	0.030
	$n=1,500$	$s=50$	$\sigma_{e\nu}=0.3$							
θ	0.75	0.115	0.027	0.117	0.000	0.047	0.042	-0.004	0.070	0.250
	$n=1,500$	$s=25$	$\sigma_{e\nu}=0.3$							
θ	0.75	0.071	0.022	0.073	-0.002	0.027	0.025	-0.005	0.025	0.027

exist a function H of all weak instruments which is strong. To demonstrate this, consider the following simulation design.

The data $D \equiv \{(y_i, x_i, z_i)_i \mid i = 1, 2, ..n\}$ is generated using the following data generating process (DGP).

$$y = \theta^{(0)} + x\theta^{(1)} + e \tag{21}$$

$$x = \Pi^{(0)} + H(z)'\Pi^{(1)} + \nu \tag{22}$$

$$\begin{pmatrix} e \\ \nu \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e & \sigma_{e\nu} \\ \sigma_{\nu e} & \sigma_\nu \end{bmatrix} \right) \tag{23}$$

$$z \sim N_J(0, 1) \tag{24}$$

where $(\theta^{(0)}, \theta^{(1)}) = (-0.90, 0.75)$ are the parameters of interest. We also let $\sigma_\nu = 1$, $\sigma_e = 1$ and $J = 500$ for all our simulations. The correlation structure in (23), leads to $\mathbb{E}(x \cdot e) \neq 0$ implying an endogeneity issue. For other parameters we consider various settings. We run

simulations for sample size n of $1,000$ and $1,500$; we consider two different values for $\sigma_{e\nu}$: 0.3, and 0.5. For $H(\cdot)$, we assume, $H(z) = [\underbrace{1, 1, .., 1, 1}_{J}]z$, with $\Pi^{(1)} = 0.05$ and $\Pi^{(1)} = 0.03$ to simulate different strengths of instruments.

For each setting of simulation parameters, Table (3) reports results for 2SLS with standard instruments (i.e. using all instruments)[3], MLIVs with the linear heuristic, and MLIVs using Gradient Descent (the more general GMM approach). For all three methods we report mean bias, mean standard error and root mean square error for each estimated parameter. To generate MLIVs from 500 weakly predictive variables using the linear heuristic, we use an estimator with 3-fold sample splitting on outer loop. For each fold k, we use excluded data D_k^c to train 6 machine learning algorithms: lasso regression, ridge regression, elastic-net, KRLS, XGBoost and neural nets. We tune the hyperparameters of each machine learning algorithm using 4-fold cross-validation (inner loop cross-validation) in the data D_k^c (i.e. D_k^c is further partitioned into a training set and a validation set as described earlier). We then choose the machine learning algorithm with the lowest cross-validation error and use that for the instrument transformation for data D_k. In our simulations, ridge regression outperformed all the other algorithms. For MLIVs using the more general gradient descent approach, we use ridge regression as the ml and employ a 3-fold cross-learner. Within each fold, we use a 80-20 (train and validation) split to tune the number of iterations for gradient descent.

As the number of instruments is very high (J=500 with N being under 1500 in all simulations), the first stage with standard instruments fits closely with the true values of the endogenous variables. In this case, the IV estimator resembles the OLS estimator (efficient but biased). Hence, the bias is very high with standard instruments. In contrast, MLIV (both linear heuristic and GD MLIV) is able to achieve nearly the same standard errors but considerably lower bias across all parameter settings. Next, we discuss results for two addition settings for linear models, namely a setting with many instruments that satisfy strong sparsity and another in which there are only a few weak IVs. MLIV continues to perform well in these settings as well.

[3] Post lasso (Belloni et al. (2012)) delivers similar results as standard instruments (as lasso did not select any instrument across our simulations). Even for MLIV with Lasso we find similar results as OLS as Lasso did not select any instrument across all our simulations.

Table 3 Simulation (a)

	True	Standard Instruments			Linear MLIV			GD MLIV		
		Bias	St. Err.	RMSE	Bias	St. Err.	RMSE	Bias	St. Err.	RMSE
			$n=1,000$	$\sigma_{e\nu}=0.5$	$\Pi^{(2)}=0.05$					
$\theta^{(0)}$	-0.90	-0.042	0.030	0.052	-0.013	0.033	0.042	-0.008	0.036	0.023
$\theta^{(1)}$	0.75	0.135	0.022	0.136	0.027	0.040	0.047	0.019	0.064	0.079
			$n=1,500$	$\sigma_{e\nu}=0.5$	$\Pi^{(2)}=0.05$					
$\theta^{(0)}$	-0.90	-0.029	0.025	0.039	0.004	0.027	0.033	0.002	0.030	0.023
$\theta^{(1)}$	0.75	0.114	0.019	0.114	-0.009	0.030	0.032	-0.010	0.053	0.042
			$n=1,000$	$\sigma_{e\nu}=0.3$	$\Pi^{(2)}=0.05$					
$\theta^{(0)}$	-0.90	-0.025	0.031	0.041	0.001	0.035	0.033	0.004	0.037	0.026
$\theta^{(1)}$	0.75	0.076	0.023	0.078	-0.014	0.054	0.045	-0.009	0.061	0.049
			$n=1,500$	$\sigma_{e\nu}=0.3$	$\Pi^{(2)}=0.05$					
$\theta^{(0)}$	-0.90	-0.017	0.026	0.033	0.002	0.027	0.030	0.007	0.030	0.022
$\theta^{(1)}$	0.75	0.071	0.020	0.071	0.007	0.032	0.027	-0.001	0.051	0.039
			$n=1,000$	$\sigma_{e\nu}=0.5$	$\Pi^{(2)}=0.03$					
$\theta^{(0)}$	-0.90	-0.117	0.030	0.121	-0.032	0.055	0.032	0.034	0.050	0.051
$\theta^{(1)}$	0.75	0.255	0.029	0.256	-0.081	0.142	0.081	-0.027	0.123	0.103
			$n=1,500$	$\sigma_{e\nu}=0.5$	$\Pi^{(2)}=0.03$					
$\theta^{(0)}$	-0.90	-0.062	0.025	0.067	-0.063	0.035	0.063	-0.022	0.035	0.040
$\theta^{(1)}$	0.75	0.218	0.026	0.219	-0.011	0.075	0.011	0.060	0.079	0.076
			$n=1,000$	$\sigma_{e\nu}=0.3$	$\Pi^{(2)}=0.03$					
$\theta^{(0)}$	-0.90	-0.046	0.032	0.056	-0.037	0.055	0.037	-0.005	0.046	0.061
$\theta^{(1)}$	0.75	0.145	0.030	0.147	-0.082	0.136	0.082	0.007	0.107	0.149
			$n=1,500$	$\sigma_{e\nu}=0.3$	$\Pi^{(2)}=0.03$					
$\theta^{(0)}$	-0.90	-0.038	0.026	0.047	-0.072	0.035	0.072	-0.004	0.036	0.032
$\theta^{(1)}$	0.75	0.135	0.027	0.135	0.004	0.075	0.004	0.009	0.084	0.060

5.3. Few Exogeneous Variables or Many Weak Technical Instruments

We finally discuss the case in which we are faced with a small number of exogenous variables, all of which are weak. To deal with such a setup, extant literature (e.g., Newey and Powell (2003)) in non-parametric methods recommends constructing multiple basis/polynomial functions from the available exogenous data. This translates the few instrument problem to that of many available instruments or many technical instruments (see Chernozhukov and Hansen (2013)). However, extant literature provides limited guidance on how to construct those functional approximations, and thus, the choice relies heavily on econometrician's discretion. Moreover, recent literature Deaner (2019) has shown

that misspecification errors by the econometrician in non-parameteric settings could lead to huge bias in estimated parameters. To demonstrate this, we consider the following simulation design. The data $D \equiv \{(y_i, x_i, z_i)_i \mid i = 1, 2, \ldots n\}$ is generated using the following data generating process (DGP).

$$y = \theta^{(0)} + x\theta^{(1)} + e \tag{25}$$

$$x = \Pi^{(0)} + H(z)'\Pi^{(1)} + \nu \tag{26}$$

$$\begin{pmatrix} e \\ \nu \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_e & \sigma_{e\nu} \\ \sigma_{\nu e} & \sigma_\nu \end{bmatrix} \right) \tag{27}$$

$$z \sim N(0, 1) \tag{28}$$

where $(\theta^{(0)}, \theta^{(1)}) = (-0.90, 0.75)$ are the parameters of interest. We also let $\Pi^{(0)} = 0.3$, $\Pi^{(1)} = 1$, $\sigma_\nu = 1$ and $\sigma_e = 1$ for all our simulations. The correlation structure in (27), leads to $\mathbb{E}(x \cdot e) \neq 0$, rendering x endogenous.

For other parameters we consider various settings. We run simulations for sample size n of $1,500$ and $1,000$; we consider two different values for $\sigma_{e\nu}$: 0.3, and 0.5. We also carry out simulations for two different specifications of $H(z) = \cos(100 \cdot z)$ and $H(z) = \sin(100 \cdot z)$.

We use polynomial functions of exogenous data z i.e. $\{z, z^2, z^3, \ldots, z^{50}\}$[4] as instruments to approximate cos and sin functions. For each setting of simulation parameters, Table (4) reports results for 2SLS with Standard Instruments vs MLIVs and Fuller. To generate MLIVs, we use a 3-cross sample splitting in outer loop. For every fold k, we use excluded data D_k^c to train a super learner i.e an ensemble of 7 machine learning algorithms: lasso regression, ridge regression, elastic-net, random forest, XGBoost, random forest and neural nets. We tune the hyperparameters and relative weights of each machine learning algorithm using 4 cross-validation in the data D_k^c. For all methods, we report mean bias, mean standard error and root mean square error for each estimated parameter across 100 runs of the simulation.

In all cases, we find that the bias of the estimates is lower with our proposed method relative to standard 2SLS, Fuller, and Belloni's approach. This is in line with how even small misspecification errors can create huge bias (Deaner (2019)). Further, even with

[4] While using standard instruments, we get collinearity issues in our simulations, and thus resort to using only upto 5^{th} power for the simulations involving the standard 2SLS estimator.

Table 4 Simulation (c)

	True	Standard Instruments			ML Instruments			Fuller Instruments			Belloni Lasso		
		Bias	St. Err.	RMSE	Bias	St. Err.	RMSE	Bias	St. Err.	RMSE	Bias	St. Err.	RMSE
$n = 1,000$ $\sigma_{e\nu} = 0.5$ $H(z) = \cos(100 \cdot z)$													
$\theta^{(1)}$	0.75	-0.975	1.464	3.816	-0.010	0.121	0.120	-0.766	9.390	12.926	-0.174	4.136	7.829
$n = 1,500$ $\sigma_{e\nu} = 0.5$ $H(z) = \cos(100 \cdot z)$													
$\theta^{(1)}$	0.75	-1.210	1.756	4.805	-0.002	0.064	0.102	-2.983	10.807	16.042	-3.407	12.393	17.590
$n = 1,000$ $\sigma_{e\nu} = 0.5$ $H(z) = \sin(100 \cdot z)$													
$\theta^{(1)}$	0.75	0.866	1.448	3.793	0.011	0.099	0.118	-1.307	9.163	12.827	-0.411	0.775	3.372
$n = 1,500$ $\sigma_{e\nu} = 0.5$ $H(z) = \sin(100 \cdot z)$													
$\theta^{(1)}$	0.75	-1.384	1.759	4.802	-0.012	0.069	0.071	-2.591	11.221	16.201	0.069	3.478	8.748
$n = 1,000$ $\sigma_{e\nu} = 0.3$ $H(z) = \cos(100 \cdot z)$													
$\theta^{(1)}$	0.75	-0.969	1.383	3.590	0.010	0.123	0.145	0.612	8.803	12.069	-0.543	4.323	7.527
$n = 1,500$ $\sigma_{e\nu} = 0.3$ $H(z) = \cos(100 \cdot z)$													
$\theta^{(1)}$	0.75	-1.410	1.649	4.484	-0.021	0.062	0.098	-2.893	10.074	15.035	-0.475	30.738	25.418
$n = 1,000$ $\sigma_{e\nu} = 0.3$ $H(z) = \sin(100 \cdot z)$													
$\theta^{(1)}$	0.75	-0.832	1.375	3.581	0.12	0.106	0.152	-1.242	8.492	11.942	-0.768	0.261	1.629
$n = 1,500$ $\sigma_{e\nu} = 0.3$ $H(z) = \sin(100 \cdot z)$													
$\theta^{(1)}$	0.75	-1.312	1.658	4.509	-0.022	0.092	0.119	-2.461	10.439	15.117	0.022	3.193	8.215

increase in sample size the performance of 2SLS, Fuller and Belloni-Lasso does not improve. In contrast, for MLIVs, we find non-trivial improvements in both bias and efficiency as the sample size increases. These gains can arguably be attributed to more precise "learning" of the optimal instruments with more data. We now turn to evaluate the performance of MLIVs with a nonlinear model.

5.4. Random-Coefficients Logit Model of Demand

Now we evaluate the use of MLIVs in a random coefficients logit model, also known as BLP (Berry et al. (1995)). We present some introductory information on BLP, followed by our simulation design and results. For more details we encourage readers to refer to Berry et al. (1995) and Nevo (2001).

There are T markets with J_t differentiated products. Each market has I_t individuals who choose between the available J_t products or the outside good. The utility a consumer i

gets by purchasing a product j is given by :

$$u_{ijt} = x'_{jt}\theta_i + \xi_{jt} + \epsilon_{ijt} \tag{29}$$

where $x_{jt} \in \mathbb{R}^k$ refers to the vector of observable (to the econometrician) product characteristics; ξ_{jt} refer to product characteristics observed by consumers and producers but unobservable to the econometrician. $\theta_i \in \mathbb{R}^k$ refers to consumer i's preference for the k observable product characteristics. The random coefficient for characteristic k for individual i can be represented as $\theta_i^k = \theta^k + \sigma^k \nu_i^k$, where ν_i^k is random variable with mean 0 and unit variance, so that θ^k represents the mean preference of consumers towards characteristic k and σ^k denotes its standard deviation. Similar to Berry and Haile (2014), we divide the product characteristics space into two subsets: $\vec{\mathbf{x}}_{jt} \equiv \{\vec{\mathbf{x}}_{jt}^{(1)}, \vec{\mathbf{x}}_{jt}^{(2)}\}$, such that $\vec{\mathbf{x}}_{jt}^{(1)} \in \mathbb{R}^{k^{(1)}}$ refer to the product characteristics for which consumers have homogeneous preferences and $\vec{\mathbf{x}}_{jt}^{(2)} \in \mathbb{R}^{k^{(2)}}$ for which consumers have heterogeneous preferences.

We can express consumer i's indirect utility as follows:

$$u_{ijt} = \underbrace{x_{jt}\theta + \xi_{jt}}_{\delta_{jt}} + \underbrace{\sum_k x_{jt}^k \sigma^k \nu_i^k}_{\mu_{ijt}} + \epsilon_{ijt} \tag{30}$$

Assuming ϵ_{ijt} follows a Type I extreme value distribution (T1EV), the market share function is as follows.

$$\hat{s}_{jt}(x_j; \delta, \Sigma) = \int \frac{\exp(\delta_{jt} + \mu_{ijt}(\nu))}{1 + \sum_{j=1}^{j=J_t} \exp(\delta_{jt} + \mu_{ijt}(\nu))} dP_\nu(\nu) \tag{31}$$

For empirical applications the above integral is approximated by taking monte-carlo draws from P_ν for sufficient number of individuals, such that.

$$\hat{s}_{jt}(x_j; \delta, \Sigma) = \frac{1}{ns} \sum_{i=1}^{ns} \frac{\exp(\delta_{jt} + \mu_{ijt}(\nu_i))}{1 + \sum_{j=1}^{j=J_t} \exp(\delta_{jt} + \mu_{ijt}(\nu_i))}$$

5.4.1. Identification and Instruments Identification of demand models for differentiated goods relies on moment restrictions on unobserved product characteristics introduced in Berry et al. (1995). The authors assume that the unobserved quality ξ_{jt} is independent of the market structure $Z_t \equiv \{x_{1t}, x_{2t}, .., x_{Jt}\}$.

$$\mathbb{E}[\xi_{jt} | Z_{jt}] = 0$$

There are many economic motivations behind these restrictions (e.g., product characteristics are costly to adjust in response to demand shocks in ξ_{jt}). However, our purpose here is not to justify the assumption, but rather to test its empirical robustness under weak instruments. Since, x_{jt} is a high dimensional vector, literature has proposed certain transformations of the observed product characteristics that better approximate the optimal instruments. We will compare the performance of our instruments with a popular transformation of market structure known as BLP Style instruments (Berry et al. (1995), Nevo (2001)) given by observed characteristics of the product, and sum of the characteristics of the rival products.[5]

For our simulation design, we conduct simulations similar to Gandhi and Houde (2019), with minor differences. We consider 10 products (offered by 10 different firms) in 50 markets. The data $D \equiv \left\{ \{s_{jt}, \vec{\mathbf{x}}_{jt}\}_{j=0}^{j=J_t} \right\}_{t=1}^{t=T}$ is generated such that

$$\epsilon_{ijt} \sim T1EV(0,1) \tag{32}$$

$$\xi_{jt} \sim N(0,1) \tag{33}$$

$$x \sim N_k(0, I_k) \tag{34}$$

We consider the case of iid random coefficients.

$$\vec{\nu} \sim N(0, I_k) \tag{35}$$

We carry out our simulations for two different settings in which we vary the number of observable characteristics and the number of characteristics for which customers have heterogeneous preferences. In the first setting, we assume two observable characteristics ($k=2$) with heterogeneity allowed for only one covariate. In the second setting, we assume four observable characteristics ($k=4$) and consumers have heterogeneous preferences for all of them.. The true values of these parameters are $\left\{\theta^{(1)}, \theta^{(2)}, \sigma^{(2)} \equiv (1.0, -0.5, 0.5)\right\}$ and $\left\{\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}, \sigma^{(1)}, \sigma^{(2)}, \sigma^{(3)}, \sigma^{(4)} \equiv (1.0, -0.5, 0.5, 1.0, 2.0, 2.0, 2.0, 2.0)\right\}$ for our simulation studies 1 and 2 in Table (6) respectively.

We use a neural network to parameterize the optimal instrument function $H(\cdot)$. One benefit of using neural networks is that it gives access to readily available tools that allow

[5] For simplicity, our simulation design assumes one product per firm so we do not consider the third BLP-style instrument, namely the sum of characteristics of other products by the same firm.

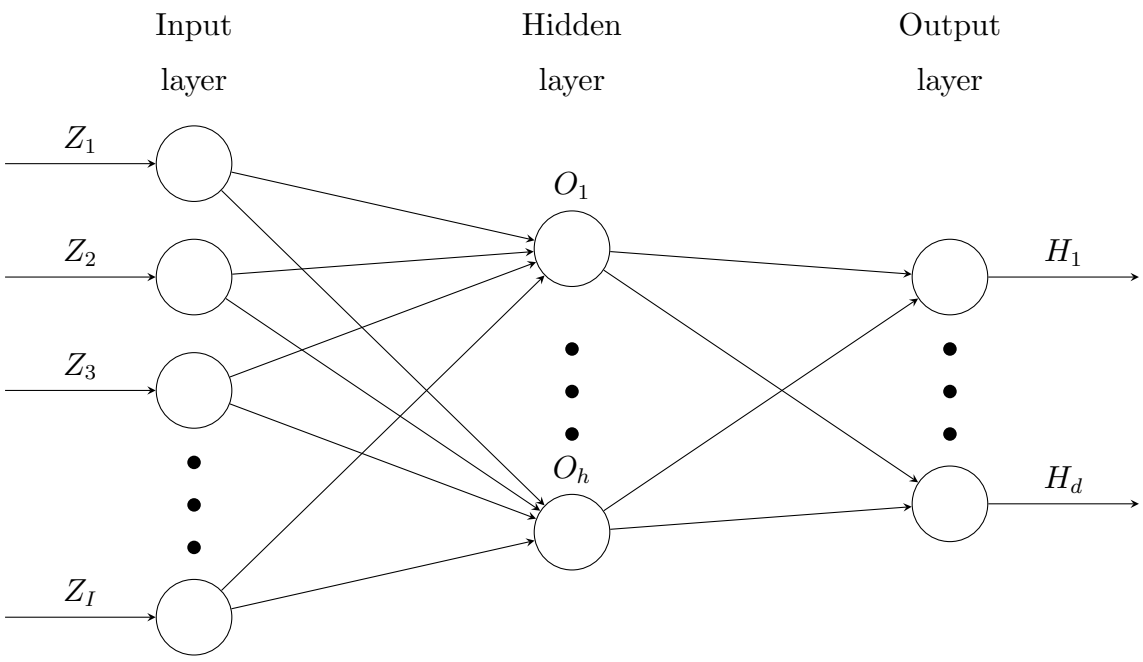

Figure 5 MLIVs through Neural Nets

for computing analytical derivatives, making it computationally easier to carry out GD. We use a neural network of dimension $[(k^{(1)}+k^{(2)})\cdot J]\times[O]\times[k^{(2)}]$. The input layer takes in all candidate instruments. With J products and $(k^{(1)}+k^{(2)})$ observed product characteristics which are assumed to be exogenous, that leaves us with $[(k^{(1)}+k^{(2)})\cdot J]$ nodes in the input layer. O denotes the number of units in the hidden layer. For the simulation below, we set this equal to 200 but this parameter – as well as the number of layers in the neural network – can be treated as hyperparameters that need to be estimated along with the weights in the network (thus, the estimates below are conservative estimates of the improvement delivered by MLIVs). The output layer denotes the number of parameters to be estimated. The optimal instruments for the linear parameters (θ) are $\vec{x_{jt}}$, as dictated by theory (Newey (1990)) and hence we only require instruments for the non-linear parameters (σ) $(dim = k^{(2)})$. We use gradient descent as previously discussed in each k fold to learn the parameters η_k of the neural network. It is worth noting that BLP assumes that the instruments in the input layer have to be combined in a specific way (sum of rival product characteristics), whereas we estimate the instrument function $H(\cdot)$ more effectively using a Neural Net.

We carry out 2 fold cross-learner such that each fold has 25 markets each. We use 8 markets as validation set to determine the number of GD iterations and stop once the

Table 6 Simulation (b)

	True	BLP Instruments		ML Instruments	
		Bias	St. Err.	Bias	St. Err.
$\theta^{(1)}$	1.00	0.055	0.01	0.04	0.004
$\theta^{(2)}$	-0.50	0.195	0.035	0.02	0.005
$\sigma^{(2)}$	0.50	0.560	0.740	0.18	0.050
$\theta^{(1)}$	1.00	-0.371	0.090	-0.008	0.011
$\theta^{(2)}$	-0.50	-0.029	0.027	-0.053	0.006
$\theta^{(3)}$	0.50	0.041	0.035	0.044	0.004
$\theta^{(4)}$	1.00	-0.180	0.107	-0.026	0.006
$\sigma^{(1)}$	2.00	-1.280	1.892	-0.129	0.109
$\sigma^{(2)}$	2.00	-0.015	0.855	0.004	0.167
$\sigma^{(3)}$	2.00	0.745	1.495	0.282	0.076
$\sigma^{(4)}$	2.00	-0.785	1.632	-0.152	0.098

variance of the estimator (validation error) starts increasing in the validation set. Table (6) reports average bias and average standard error for estimated parameters. We report results for BLP instruments and MLIVs. We find that consistently across both specifications MLIVs lead to much lower standard errors and bias on the heterogeneity parameters, compared to the BLP instruments, which are unable to identify heterogeneity parameters (σ). Further, we also observe significant reduction in bias for linear parameters with the use of MLIVs.

6. The Colonial Origins of Comparative Development

We finally consider an application of our method to a real-world dataset. We consider Acemoglu, Johnson and Robinson(2001) - henceforth AJR(2001), where the authors study whether differences in economic development across countries can be explained by differences in the quality of their institutions (rule of law, property rights, etc). However, it is possible that economic development leads to good institutions or that other unobservables drive both institutional structure and economic development, leading to an endogeneity issue in estimating the impact of institutions on economic performance. To address this endogeneity issue, authors use the mortality rates expected by the first European settlers in the colonies as an instrument for the quality of institutions. The argument is that the settler mortality rates affected their colonization strategies which, in turn, affected their

path of institutional development that culminated in today's institutions. The AJR paper serves as an excellent example to test our cross learner algorithm because it has come under criticism recently due to the weakness of instruments across the various specifications (Chernozhukov and Hansen (2008)). Another reason to consider applying our approach to the AJR data is that their sample size is relatively small ($n = 64$), allowing us to investigate whether the cross learner algorithm is beneficial with small datasets.

Authors report the 2SLS estimates from the following reduced form regression.

$$\log(y_i) = \mu + \alpha R_i + X_i'\gamma + \epsilon_i \tag{36}$$

where y_i is the income per capita in country i, R_i is the protection against expropriation (measure of institution quality) and X_i is a vector of other covariates. To address the endogeneity issue the authors use the following first stage.

$$R_i = \zeta + \theta Z_i + X_i'\delta + \nu_i \tag{37}$$

where Z_i is logarithm of the settler mortality rate in $1,000$ mean strength.

We replicate the model specifications, from Table (7) of AJR(2001), which came under criticism due to weak instruments. To contrast their results with ours, we learn the MLIVs $H(\cdot)$ by estimating a slightly different first stage.

$$R_i = \zeta + \theta \hat{H}_i(Z_i) + X_i'\delta + \nu_i \tag{38}$$

For learning the MLIVs we employ a 3-fold sample splitting procedure. We employ multiple machine learning methods and select the winning ML estimator (random forest) based on cross-validation error.

We find that across all specifications, MLIVs seem much stronger than AJR instruments, as measured by the F value (see Stock and Yogo (2002)). A higher F value, generally indicates stronger instruments. Higher power of MLIVs leads to lower standard errors across the board. AJR instruments are not able to identify the constant term, which could potentially be due to weak instruments. On the other hand, MLIVs are able to identify the constant term across all specifications. Further, the estimates of the impact of institution quality on per capita income generated using MLIVs are lower than corresponding AJR estimates.

Table 7 AJR(2001) Table 4

	AJR Instruments					
	Base Sample	Base Sample	Base Sample without Neo-Europes	Base Sample without Neo-Europes	Base Sample	Base Sample
	(1)	(2)	(3)	(4)	(5)	(6)
Average protection against expropriation risk 1985-1995	0.944***	0.995***	1.281***	1.211***	0.982***	1.107**
	(0.156)	(0.221)	(0.358)	(0.354)	(0.299)	(0.463)
Latitude		-0.647		0.938		-1.178
		(1.335)		(1.463)		(1.755)
Continent Dummies					✓	✓
Constant	1.909*	1.691	-0.141	0.144	2.032	1.440
	(1.026)	(1.293)	(2.265)	(2.183)	(2.011)	(2.839)
F	22.946	13.09	8.646	7.826	6.233	3.456
	ML Instruments					
Average protection against expropriation risk 1985-1995	0.725***	0.683***	0.850***	0.776***	0.770***	0.787***
	(0.118)	(0.143)	(0.220)	(0.213)	(0.175)	(0.215)
Latitude		0.668		1.445		-0.249
		(0.937)		(0.996)		(1.013)
Continent Dummies					✓	✓
Constant	3.336***	3.488***	2.572*	2.803**	3.448***	3.384***
	(0.775)	(0.842)	(1.393)	(1.316)	(1.184)	(1.330)
F	28.681	18.856	11.575	10.670	12.285	8.724
N	64	64	60	60	64	64

Standard errors in parentheses.

p-values in brackets.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Interestingly, for specification (3),(4),(5) and (6) where the AJR instrument falls below the commonly used F value threshold of 10, we see a significant increase in estimated effect of institutional structure, by as much as 36%. In contrast, MLIVs have higher F value. Further, for all other controls and robustness tests, we find MLIVs estimates to be much more consistent across specifications, compared to estimates using AJR instruments.

7. Discussion and Conclusions

In this paper, we propose that the problem of constructing instrumental variables from observational data can be cast as a machine learning problem. Simulations as well as an application to data from prior research demonstrate that our proposed algorithm is

a promising tool for researchers working on observational data. The approach can help address the problem of weak instruments, which is often the main concern, in practice, with the use of instrumental variables. There are some trade-offs with our approach. Since, our method entails nested levels of sample splitting, it might be computationally expensive to construct MLIVs for big datasets. However, we believe given the advancements in distributed computing it is not a big concern, as computations in each data fold can be parallelized. Further, our learning method might not approach the semi-parametric efficiency bound for many non-linear models, nonetheless, it can algorithmically construct strong instruments for most of them without requiring additional assumptions about the optimal instrument structure and it delivers good results in practice.

We note that the MLIV approach is focused on improving the relevance of instruments but does not explicitly address the exclusion restriction. MLIV assumes that the candidate instruments satisfy strong exogeneity, an assumption that is central to the broader optimal instruments literature because it enables the use of transformations, $H(z)$, as instruments.

References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001). The colonial origins of comparative development: An empirical investigation. American economic review, 91(5):1369–1401.

Ackerberg, D., Caves, K., and Frazer, G. (2006). Structural identification of production functions.

Amemiya, T. (1966). On the use of principal components of independent variables in two-stage least-squares estimation. International Economic Review, 7(3):283–303.

Amemiya, T. (1974). The nonlinear two-stage least-squares estimator. Journal of econometrics, 2(2):105–110.

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., Krueger, A. B., et al. (1993). Split sample instrumental variables. Hebrew University of Jerusalem, Department of Economics.

Arellano, M. and Bond, S. (1991). Some tests of specification for panel data: Monte carlo evidence and an application to employment equations. The review of economic studies, 58(2):277–297.

Arellano, M. and Bover, O. (1995). Another look at the instrumental variable estimation of error-components models. Journal of econometrics, 68(1):29–51.

Armstrong, T. B. (2016). Large market asymptotics for differentiated product demand estimators with economic models of supply. Econometrica, 84(5):1961–1980.

Ayyagari, R. (2010). Applications of influence functions to semiparametric regression models. PhD thesis, Harvard University.

Bai, J. and Ng, S. (2009). Boosting diffusion indices. Journal of Applied Econometrics, 24(4):607–629.

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.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, 81(2):608–650.

Bennett, A., Kallus, N., and Schnabel, T. (2019). Deep generalized method of moments for instrumental variable analysis. In Advances in Neural Information Processing Systems, pages 3559–3569.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. Econometrica: Journal of the Econometric Society, pages 841–890.

Berry, S. T. and Haile, P. A. (2014). Identification in differentiated products markets using market level data. Econometrica, 82(5):1749–1797.

Blundell, R. and Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel data models. Journal of econometrics, 87(1):115–143.

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.

Caner, M. (2009). Lasso-type gmm estimator. Econometric Theory, 25(1):270–290.

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2016). Double/debiased machine learning for treatment and causal parameters. arXiv preprint arXiv:1608.00060.

Chernozhukov, V. and Hansen, C. (2008). The reduced form: A simple approach to inference with weak instruments. Economics Letters, 100(1):68–71.

Chernozhukov, V. and Hansen, C. (2013). Econometrics of high-dimensional sparse models. Lecture, NBER, Cambridge, MA.

Chintagunta, P. K., Dubé, J.-P., and Singh, V. (2003). Balancing profitability and customer welfare in a supermarket chain. Quantitative Marketing and Economics, 1(1):111–147.

Chintagunta, P. K. and Nair, H. S. (2011). Structural workshop paper—discrete-choice models of consumer demand in marketing. Marketing Science, 30(6):977–996.

Deaner, B. (2019). On the sensitivity of nonparametric instrumental variables estimators to misspecification. arXiv preprint arXiv:1901.01241.

Dinner, I. M., Van Heerde, H. J., and Neslin, S. A. (2013). Driving online and offline sales: The cross-channel effects of traditional, online display, and paid search advertising. Journal of marketing research, 50(5):527–545.

Donald, S. G., Imbens, G. W., and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models. Journal of Econometrics, 152(1):28–36.

Fan, J., Guo, S., and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 74(1):37–65.

Fan, J. and Liao, Y. (2014). Endogeneity in high dimensions. Annals of statistics, 42(3):872.

Faugeras, O. and FAUGERAS, O. A. (1993). Three-dimensional computer vision: a geometric viewpoint. MIT press.

Gandhi, A. and Houde, J.-F. (2019). Measuring substitution patterns in differentiated products industries. Technical report, National Bureau of Economic Research.

Gautier, E. and Rose, C. (2011). High-dimensional instrumental variables regression and confidence sets. arXiv preprint arXiv:1105.2454.

Goldfarb, A., Lu, Q., and Moorthy, S. (2009). Measuring brand value in an equilibrium framework. Marketing Science, 28(1):69–86.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. pages 2672–2680.

Gould, S., Fernando, B., Cherian, A., Anderson, P., Cruz, R. S., and Guo, E. (2016). On differentiating parameterized argmin and argmax problems with application to bi-level optimization. arXiv preprint arXiv:1607.05447.

Hansen, C. and Kozbur, D. (2014). Instrumental variables estimation with many weak instruments using regularized jive. Journal of Econometrics, 182(2):290–308.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. Econometrica: Journal of the Econometric Society, pages 1029–1054.

Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. Journal of Econometrics, 30(1-2):203–238.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2016). Counterfactual prediction with deep instrumental variables networks. arXiv preprint arXiv:1612.09596.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pages 1414–1423. JMLR. org.

Hausman, J., Leonard, G., and Zona, J. D. (1994). Competitive analysis with differenciated products. Annales d'Economie et de Statistique, pages 159–180.

Hausman, J., Stock, J. H., and Yogo, M. (2005). Asymptotic properties of the hahn–hausman test for weak-instruments. Economics Letters, 89(3):333–342.

Hausman, J. A. (1996). Valuation of new goods under perfect and imperfect competition. In The economics of new goods, pages 207–248. University of Chicago Press.

Hendel, I. and Nevo, A. (2006). Measuring the implications of sales and consumer inventory behavior. Econometrica, 74(6):1637–1673.

Ichimura, H. and Newey, W. K. (2015). The influence function of semiparametric estimators. arXiv preprint arXiv:1508.01378.

Kapetanios, G. and Marcellino, M. (2010). Factor-gmm estimation with large sets of possibly weak instruments. Computational Statistics & Data Analysis, 54(11):2655–2675.

Kloek, T. and Mennes, L. (1960). Simultaneous equations estimation based on principal components of predetermined variables. Econometrica, Journal of the Econometric Society, pages 45–61.

Knittel, C. R. and Metaxoglou, K. (2014). Estimation of random-coefficient demand models: two empiricists' perspective. Review of Economics and Statistics, 96(1):34–59.

Lewis, G. and Syrgkanis, V. (2018). Adversarial generalized method of moments. arXiv preprint arXiv:1803.07164.

Li, Y., Song, L., Wu, X., He, R., and Tan, T. (2018). Anti-makeup: Learning a bi-level adversarial network for makeup-invariant face verification.

Melitz, M. J. and Polanec, S. (2015). Dynamic olley-pakes productivity decomposition with entry and exit. The Rand journal of economics, 46(2):362–375.

Muandet, K., Mehrjou, A., Lee, S. K., and Raj, A. (2019). Dual iv: A single stage instrumental variable regression. arXiv preprint arXiv:1910.12358.

Nair, H., Chintagunta, P., and Dubé, J.-P. (2004). Empirical analysis of indirect network effects in the market for personal digital assistants. Quantitative Marketing and Economics, 2(1):23–58.

Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. Econometrica, 69(2):307–342.

Newey, W. K. (1990). Semiparametric efficiency bounds. Journal of applied econometrics, 5(2):99–135.

Newey, W. K. (1993). 16 efficient estimation of models with conditional moment restrictions.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. Handbook of econometrics, 4:2111–2245.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. Econometrica, 71(5):1565–1578.

Newey, W. K. and Robins, J. R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. arXiv preprint arXiv:1801.09138.

Phillips, P. C., Park, J. Y., and Chang, Y. (2004). Nonlinear instrumental variable estimation of an autoregression. Journal of Econometrics, 118(1-2):219–246.

Reynaert, M. and Verboven, F. (2014). Improving the performance of random coefficients demand models: the role of optimal instruments. Journal of Econometrics, 179(1):83–98.

Rossi, P. E. (2014). Even the rich can make themselves poor: A critical examination of iv methods in marketing applications. Marketing Science, 33(5):655–672.

Rutz, O. J. and Watson, G. F. (2019). Endogeneity and marketing strategy research: an overview. Journal of the Academy of Marketing Science, 47(3):479–498.

Samuel, K. G. and Tappen, M. F. (2009). Learning optimized map estimates in continuously-valued mrf models. pages 477–484.

Scornet, E., Biau, G., Vert, J.-P., et al. (2015). Consistency of random forests. The Annals of Statistics, 43(4):1716–1741.

Singh, R., Sahani, M., and Gretton, A. (2019). Kernel instrumental variable regression. In Advances in Neural Information Processing Systems, pages 4595–4607.

Staiger, D. and Stock, J. H. (1994). Instrumental variables regression with weak instruments. Technical report, National Bureau of Economic Research.

Steinwart, I. (2005). Consistency of support vector machines and other regularized kernel classifiers. IEEE Transactions on Information Theory, 51(1):128–142.

Stock, J. H. and Wright, J. H. (2000). Gmm with weak identification. Econometrica, 68(5):1055–1096.

Stock, J. H. and Yogo, M. (2002). Testing for weak instruments in linear iv regression.

Van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6(1).

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242.

Wang, Z. and Yu, Z. (2012). Trading partners, traded products and firm performances of china's exporter-importers: does processing trade make a difference? The World Economy, 35(12):1795–1824.

Webb, A. R. (2003). Statistical pattern recognition. John Wiley & Sons.

Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press.

Yogo, M. (2004). Estimating the elasticity of intertemporal substitution when instruments are weak. Review of Economics and Statistics, 86(3):797–810.

Young, A. (2017). Consistency without inference: Instrumental variables in practical application. Unpublished manuscript, London: London School of Economics and Political Science. Retrieved from: http://personal. lse. ac. uk/YoungA.

Zoph, B. and Le, Q. V. (2016). Neural architecture search with reinforcement learning. arXiv preprint arXiv:1611.01578.

Appendices

A. Technical and Implementation Details

LEMMA 6. **(Exclusion Restriction)**: *Suppose $\hat{\eta}_i$ be the estimated hyperparameters for some function H corresponding to data D_i. Then exclusion restrictions hold if $\hat{\eta}_i$ was estimated using data $D_i^c \subseteq D \setminus D_i$.*

$$\mathbb{E}\left[\xi_i \cdot H(z_i; \hat{\eta}_i)\right] = 0 \tag{39}$$

Using law of iterated expectations and assumption (2), we have:

$$\begin{aligned}\mathbb{E}\left[\xi_i \cdot H(z_i; \hat{\eta}_i)\right] &= \mathbb{E}\left[\mathbb{E}\left[\xi_i \cdot H(z_i; \hat{\eta}_i)|z_i, D_i^c\right]\right] \\ &= \mathbb{E}\left[H(z_i; \hat{\eta}_i) \cdot \mathbb{E}\left[\xi_i|z_i, D_i^c\right]\right] \\ &= 0,\end{aligned} \tag{40}$$

which completes the proof ■

LEMMA 1. *If assumption 1 is satisfied and for every $k = 1, 2 \ldots, K$ $\mathbb{E}[b(w, \theta, \hat{\eta}_k)] < \infty$, where $b(w_i, \theta, \hat{\eta}_k) = \xi(x_i, \theta)^4 (h(z_i, \eta^*) - h(z_i, \hat{\eta}_k))^2$ and $\theta \in \Theta$ such that Θ is compact; Then,*

$$\hat{f}(\theta) \xrightarrow{p} \mathbb{E}[f(w_i, \theta, \eta^*)]$$

where,

$$\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} f(w_i, \theta, \hat{\eta}_k)$$

We first show that $\hat{f}(\theta) \xrightarrow{p} \frac{1}{n} \sum_{i=1}^{i=n} f(w_i, \theta, \eta^*)$ where $\hat{f}(\theta) = \frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} f(w_i, \theta, \hat{\eta}_k)$. Let

$$\hat{\Gamma}_{ik} \triangleq f(w_i, \theta, \hat{\eta}_k) - f(w_i, \theta, \eta^*)$$

We first show that $\mathbb{E}\left[|\frac{1}{n} \sum_{i \in I_k} \hat{\Gamma}_{ik}| \middle| D_k^c\right] \xrightarrow{p} 0$. Consider the following,

$$\begin{aligned}\mathbb{E}\left[|\frac{1}{n} \sum_{i \in I_k} \hat{\Gamma}_{ik}| \middle| D_k^c\right] &\leq \frac{1}{n} \sum_{i \in I_k} \mathbb{E}\left[|\hat{\Gamma}_{ik}| \middle| D_k^c\right] \\ &\leq \int \left| f(w, \theta, \eta^*) - f(w, \theta, \hat{\eta}_k) \right| d\mathbb{P}_0(w) \\ &\leq \left[\int \left| f(w, \theta, \eta^*) - f(w, \theta, \hat{\eta}_k) \right|^2 d\mathbb{P}_0(w) \right]^{\frac{1}{2}}\end{aligned}$$

$$\leq \Big[\int \xi(x,\theta)^2 (h(z,\eta^*) - h(z,\hat{\eta}_k))^2 d\mathbb{P}_0(w) \Big]^{\frac{1}{2}}$$
$$\leq \Big[\int \xi(x,\theta)^4 (h(z,\eta^*) - h(z,\hat{\eta}_k))^2 d\mathbb{P}_0(w) \int (h(z,\eta^*) - h(z,\hat{\eta}_k))^2 d\mathbb{P}_0(w) \Big]^{\frac{1}{4}}$$
$$= O_p(1) o_p(1)$$

Thus it follows from conditional markov inequality

$$\frac{1}{n} \sum_{i \in I_k} \hat{\Gamma}_{ik} \xrightarrow{p} 0$$

Hence, we have,

$$\frac{1}{n} \sum_{k=1}^{k=K} \sum_{i \in I_k} \hat{\Gamma}_{ik} \xrightarrow{p} 0$$

Thus we have, $\hat{f}(\theta) \xrightarrow{p} \frac{1}{n} \sum_{i=1}^{i=n} f(w_i, \theta, \eta^*)$ Using law of large numbers we also know that,

$$\frac{1}{n} \sum_{i=1}^{i=n} f(w_i, \theta, \eta^*) \xrightarrow{p} \mathbb{E}[f(w_i, \theta, \eta^*)]$$

Thus, using triangle inequality we have,

$$\hat{f}(\theta) \xrightarrow{p} \mathbb{E}[f(w_i, \theta, \eta^*)]$$

This completes the proof ■

Lemma 2. *If the instruments are estimated in a mean square consistent manner for each* $k = 1, 2 \ldots K$ *i.e* $\int \Big[h(z_i, \eta_0) - h(z_i, \hat{\eta}_k) \Big]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0$ *then for each* $k = 1, 2 \ldots K$*;*

$$\int \Big[f(w, \theta_0, \eta_0) - f(w, \theta_0, \hat{\eta}_k) \Big]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0 \tag{10}$$

Using cauchy-schwarz inqequality and equation 1, we have:

$$\int \Big[f(w, \theta_0, \eta_0) - f(w, \theta_0, \hat{\eta}_k) \Big]^2 d\mathbb{P}_0(w) = \int \xi(x_i, \theta_0)^2 \Big[h(z_i, \eta_0) - h(z_i, \hat{\eta}_k) \Big]^2 d\mathbb{P}_0(w)$$
$$\leq \int \xi(x_i, \theta_0)^2 \Big[h(z_i, \eta_0) - h(z_i, \hat{\eta}_k) \Big]^2 d\mathbb{P}_0(w)$$
$$\leq \Big[\int \xi(x_i, \theta_0)^4 \Big[h(z_i, \eta_0) - h(z_i, \hat{\eta}_k) \Big]^2 d\mathbb{P}_0(w)$$

$$\left[\int \Big[h(z_i,\eta_0)-h(z_i,\hat{\eta}_k)\Big]^2 d\mathbb{P}_0(w)\right]^{\frac{1}{2}}$$
$$\xrightarrow{p} O_p(1)o_p(1)$$

which completes the proof ■

LEMMA 3. *If assumption 2 is satisfied then estimation of $\hat{\eta}$ does not have any asymptotic effect on the moment conditions for a K fold MLIV estimator i.e*

$$\sqrt{n}\hat{f}(\theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f(x_i,z_i,\theta_0,\eta_0) + o_p(1) \tag{11}$$

where $\hat{f}$ is given as

$$\hat{f}(\theta_0) = \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I_k} f(x_i,z_i,\theta_0,\hat{\eta}_k) \tag{12}$$

η_k is the estimator for the k^{th} fold estimated using data excluding the k^{th} fold (i.e., D_k^c).

Consider $\hat{\zeta}_{ik} \triangleq f(x_i,z_i,\theta_0,\hat{\eta}_i) - f(x_i,z_i,\theta_0,\eta_0) \forall i \in I_l$
Further by construction of MLIV recipe and independence of data samples. $E\Big[\hat{\zeta}_{ik}|D_k^c\Big] = 0$. Also, due to independence of samples within the fold $E\Big[\hat{\zeta}_{ik}\hat{\zeta}_{jk}|D_k^c\Big] = 0 \ \forall i,j \in I_k$. First we show that $\mathbb{E}[(\frac{1}{\sqrt{n}}\sum_{i\in I_k}\hat{\zeta}_{ik})^2|D_k^c] \xrightarrow{p} 0$. Consider the following

$$\begin{aligned}
\mathbb{E}[(\frac{1}{\sqrt{n}}\sum_{i\in I_k}\hat{\zeta}_{ik})^2|D_k^c] &= \frac{1}{n}\Big[\sum_{i\in I_k}\mathbb{E}[(\hat{\zeta}_{ik})^2|D_k^c] + \sum_{i\neq j}\mathbb{E}[(\hat{\zeta}_{ik}\hat{\zeta}_{jk})|D_k^c]\Big] \\
&= \frac{1}{n}\Big[\sum_{i\in I_k}\mathbb{E}[(\hat{\zeta}_{ik})^2|D_k^c]\Big] \\
&= \frac{1}{K}\Big[\int \Big[f(w,\theta_0,\eta_0) - f(w,\theta_0,\hat{\eta}_k)\Big]^2 d\mathbb{P}_0(w)\Big]
\end{aligned}$$

Thus from assumption 2 we have $\mathbb{E}[(\frac{1}{\sqrt{n}}\sum_{i\in I_k}\hat{\zeta}_{ik})^2|D_k^c] \xrightarrow{p} 0$. Using the conditional markov inequality it follows

$$\frac{1}{\sqrt{n}}\sum_{i\in I_k}\hat{\zeta}_{ik} \xrightarrow{p} 0$$

Thus,

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{k=K}\sum_{i\in I_k}\hat{\zeta}_{ik} \xrightarrow{p} 0$$

This completes the proof ■

LEMMA 4. *If the instruments are estimated in a mean square consistent manner for each* $k = 1, 2 \dots K$ *i.e* $\int \Big[h(z_i, \eta_0) - h(z_i, \hat{\eta}_k)\Big]^2 d\mathbb{P}_0(w) \xrightarrow{p} 0$ *and* $\mathbb{E}\left[\left(\frac{\partial \xi_i(\theta_0)}{\partial \theta_j}\right)^2\right] < \infty$ *then for each* $k = 1, 2 \dots K$*;*

$$\int \Big|\frac{\partial f(w, \theta_0, \eta_0)_i}{\partial \theta_j} - \frac{f(w, \theta_0, \hat{\eta}_k)_i}{\partial \theta_j}\Big| d\mathbb{P}_0(w) \xrightarrow{p} 0 \tag{13}$$

Using cauchy-schwarz inqequality and equation 1, we have:

$$\begin{aligned}\int |\frac{\partial f(w, \theta_0, \eta_0)_i}{\partial \theta_j} - \frac{f(w, \theta_0, \hat{\eta}_k)_i}{\partial \theta_j}| d\mathbb{P}_0(w) &= \int \Big|\frac{\partial \xi(x_s, \theta_0)}{\partial \theta_j}\Big|\Big|[h(z_s, \eta_0) - h(z_s, \hat{\eta}_k)]_i\Big| d\mathbb{P}_0(w) \\ &\leq \Big[\int \Big|\frac{\partial \xi(x_s, \theta_0)}{\partial \theta_j}\Big|^2 d\mathbb{P}_0(w) \int \Big[h(z_s, \eta_0) - h(z_s, \hat{\eta}_k)\Big]_i^2 d\mathbb{P}_0(w)\Big]^{\frac{1}{2}} \\ &\xrightarrow{p} O_p(1) o_p(1)\end{aligned}$$

This completes the proof ■

LEMMA 5. *If assumption 3 is satisfied and for any* $\bar{\theta} \xrightarrow{p} \theta$*,* $\hat{f}$ *is differentiable at* $\bar{\theta}$ *with probability approaching one, then* $\hat{G}(\bar{\theta}) \xrightarrow{p} G$*. where* $\hat{G}(\bar{\theta}) = \frac{\partial \hat{f}(\bar{\theta})}{\partial \theta}$ *and* $G = \mathbb{E}\left[\frac{\partial f(w_i, \theta, \eta_0)}{\partial \theta} | \theta = \theta_0\right]$*.*

Let $\hat{f}(\theta) \triangleq \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in I_k} f(w_i, \theta, \hat{\eta}_k)$. Let $\hat{G}(\theta) \triangleq \frac{\partial \hat{f}(\theta)}{\partial \theta}$ when it exists. $\tilde{G}_k \triangleq \frac{1}{n}\sum_{i \in I_k} \frac{\partial f(w_i, \theta_0, \eta_0)}{\partial \theta}$ and let $\bar{G}_k \triangleq \frac{1}{n}\sum_{i \in I_k} \frac{\partial f(w_i, \theta_0, \hat{\eta}_k)}{\partial \theta}$. By law of large numbers:

$$\sum_{k=1}^{k=K} \tilde{G}_k \xrightarrow{p} \mathbb{E}[\frac{\partial f(w_i, \theta_0, \eta_0)}{\partial \theta}]$$

$$\sum_{k=1}^{k=K} \tilde{G}_k \xrightarrow{p} G$$

Also using triangle inequality we know,

$$\mathbb{E}[|\bar{G}_k^{ij} - \tilde{G}_k^{ij}| D_k^c] \leq \int \left|\frac{\partial f(w, \theta_0, \eta_0)_i}{\partial \theta_j} - \frac{f(w, \theta_0, \hat{\eta}_k)_i}{\partial \theta_j}\right| d\mathbb{P}_0(w) \xrightarrow{p} 0 \tag{41}$$

Then by using conditional markov inequality

$$P[|\bar{G}_k^{ij} - \tilde{G}_k^{ij}| > \epsilon | D_k^c] \leq \mathbb{E}[|\bar{G}_k^{ij} - \tilde{G}_k^{ij}| D_k^c]/\epsilon$$

$$P[|\bar{G}_k^{ij} - \tilde{G}_k^{ij}| > \epsilon | D_k^c] \leq o_p(1)$$

Thus it directly follows,

$$\bar{G}_k^{ij} - \tilde{G}_k^{ij} \xrightarrow{p} 0$$

Using the triangle inequality, we get

$$|\sum_{k=1}^{k=K} \bar{G}_k - G| = |\sum_{k=1}^{k=K} \bar{G}_k - \sum_{k=1}^{k=K} \tilde{G}_k + \sum_{k=1}^{k=K} \tilde{G}_k - G|$$

$$|\sum_{k=1}^{k=K} \tilde{G}_k - G| \leq |\sum_{k=1}^{k=K} \bar{G}_k - \sum_{k=1}^{k=K} \tilde{G}_k| + |\sum_{k=1}^{k=K} \tilde{G}_k - G|$$

$$|\sum_{k=1}^{k=K} \bar{G}_k - G| \leq \sum_{k=1}^{k=K} |\tilde{G}_k - \bar{G}_k| + |\sum_{k=1}^{k=K} \tilde{G}_k - G|$$

$$|\sum_{k=1}^{k=K} \bar{G}_k - G| \leq o_p(1) + o_p(1)$$

Thus,

$$\sum_{k=1}^{k=K} \bar{G}_k \xrightarrow{p} G \tag{42}$$

Also, we know for any $\bar{\theta} \xrightarrow{p} \theta_0$ and using triangle inequality we get

$$\Big|\Big|\hat{G}(\bar{\theta})^{ij} - \sum_{k=1}^{k=K} \bar{G}_k^{ij}\Big|\Big| \leq \frac{1}{n} \sum_{i=1}^{i=n} \Big|\Big|\frac{\partial f(w_i, \bar{\theta}, \hat{\eta}_k)_i}{\partial \theta_j} - \frac{f(w_i, \theta_0, \hat{\eta}_k)_i}{\partial \theta_j}\Big|\Big|$$

$$\leq \frac{1}{n} \sum_{i=1}^{i=n} u(w_i) ||\bar{\theta} - \theta_0||^{\xi}$$

$$= O_p(1) o_p(1) \xrightarrow{p} 0$$

Thus, we have for any $\bar{\theta} \xrightarrow{p} \theta_0$

$$\hat{G}(\bar{\theta}) \xrightarrow{p} \sum_{k=1}^{k=K} \bar{G}_k \tag{43}$$

Finally, using triangle inequality

$$\Big|\Big|\hat{G}(\bar{\theta}) - G\Big|\Big| = \Big|\Big|\hat{G}(\bar{\theta}) - \sum_{k=1}^{k=K} \bar{G}_k + \sum_{k=1}^{k=K} \bar{G}_k - G\Big|\Big|$$

$$\leq \Big|\Big|\hat{G}(\bar{\theta}) - \sum_{k=1}^{k=K} \bar{G}_k\Big|\Big| + \Big|\Big| \sum_{k=1}^{k=K} \bar{G}_k - G\Big|\Big|$$

$$\leq \Big|\Big|\hat{G}(\bar{\theta}) - \sum_{k=1}^{k=K} \bar{G}_k\Big|\Big| + \Big|\Big| \sum_{k=1}^{k=K} \bar{G}_k - G\Big|\Big|$$
$$\leq o_p(1) + o_p(1)$$

Thus, $\hat{G}(\bar{\theta}) \xrightarrow{p} G$

This completes the proof ■

THEOREM 3. *If assumption 3 and 2 are satisfied and $G'WG$ is nonsingular then $\hat{S} \xrightarrow{p} S$ and $\hat{V} \xrightarrow{p} V_0$*

Let $\hat{f}_i \triangleq f(w_i, \hat{\theta}, \hat{\eta}_k)$ and $f_i \triangleq f(w_i, \theta_0, \eta_0)$. We first show that $\frac{1}{n}\sum_{i=1}^{i=n} ||\hat{f}_i - f_i||^2 \xrightarrow{p} 0$. Then we use that to show $\hat{\Omega} \xrightarrow{p} \Omega$ and complete the proof. Consider $\hat{f}_i - f_i = \underbrace{f(w_i, \hat{\theta}, \hat{\eta}_k) - f(w_i, \theta_0, \hat{\eta}_k)}_{\hat{A}_i} + \underbrace{f(w_i, \theta_0, \hat{\eta}_k) - f(w_i, \theta_0, \eta_0)}_{\hat{B}_i}$. Thus, it is suffice to show that $\frac{1}{n}\sum_{i=1}^{i=n} ||\hat{A}_i||^2 \xrightarrow{p} 0$, and $\frac{1}{n}\sum_{i=1}^{i=n} ||\hat{B}_i||^2 \xrightarrow{p} 0$

$$\frac{1}{n}\sum_{i=1}^{i=n} ||\hat{f}_i - f_i||^2 = \frac{1}{n}\sum_{i=1}^{i=n} ||\hat{A}_i + \hat{B}_i||^2$$
$$\leq \frac{2}{n}\sum_{i=1}^{i=n} ||\hat{A}_i||^2 + \frac{2}{n}\sum_{i=1}^{i=n} ||\hat{B}_i||^2$$

We first show $\frac{1}{n}\sum_{i=1}^{i=n} ||\hat{A}_i||^2 \xrightarrow{p} 0$. Note, this is the same as showing $\frac{1}{n}\sum_{i\in I_k} ||\hat{A}_i||^2 \xrightarrow{p} 0$. Using the mean value expansion it follows that

$$\frac{1}{n}\sum_{i\in I_k} \Big|\Big| f(w_i, \hat{\theta}, \hat{\eta}_k) - f(w_i, \theta_0, \hat{\eta}_k)\Big|\Big|^2 = \frac{1}{n}\sum_{i\in I_k} \Big|\Big| \frac{\partial f(w_i, \bar{\theta}, \hat{\eta}_k)}{\partial \theta}(\hat{\theta} - \theta_0)\Big|\Big|^2$$
$$\leq \frac{1}{n}(\sum_{i\in I_k} d(w_i)^2)\Big|\Big|(\hat{\theta} - \theta_0)\Big|\Big|^2$$
$$\leq O_p(1)o_p(1)$$

Next, we show that $\frac{1}{n}\sum_{i=1}^{i=n} ||\hat{B}_i||^2 \xrightarrow{p} 0$. Note, this is the same as showing $\frac{1}{n}\sum_{i\in I_k} ||\hat{B}_i||^2 \xrightarrow{p} 0$. Consider $\mathbb{E}\left[\frac{1}{n}\sum_{i\in I_k} ||\hat{B}_i||^2 \Big| D_k^c\right]$ and if $K \ll n$, we have

$$\mathbb{E}\left[\frac{1}{n}\sum_{i\in I_k} ||\hat{B}_i||^2 \Big| D_k^c\right] \leq \int \Big[f(w, \theta_0, \eta_0) - f(w, \theta_0, \hat{\eta}_k)\Big]^2 d\mathbb{P}_0(w)$$
$$\xrightarrow{p} 0$$

Next using conditional markov inequality we have

$$\frac{1}{n}\sum_{i\in I_k}||\hat{B}_i||^2 \xrightarrow{p} 0$$

Thus, we have

$$\frac{1}{n}\sum_{k=1}^{k=K}\sum_{i\in I_k}||\hat{B}_i||^2 \xrightarrow{p} 0$$

We finally, show that $\hat{\Omega} \xrightarrow{p} \Omega$, where $\Omega = \mathbb{E}_{\mathbb{P}_0}[f_i f_i']$ and $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{i=n}\hat{f}_i\hat{f}_i'$. By law of large numbers we know that $\frac{1}{n}\sum_{i=1}^{i=n} f_i f_i' \xrightarrow{p} \mathbb{E}_{\mathbb{P}_0}[f_i f_i']$. Next, consider

$$\begin{aligned}
\Big|\Big|\frac{1}{n}\sum_{i=1}^{i=n}\hat{f}_i\hat{f}_i' - \frac{1}{n}\sum_{i=1}^{i=n} f_i f_i'\Big|\Big|^2 &\leq \frac{1}{n}\sum_{i=1}^{i=n}\Big|\Big|\hat{f}_i\hat{f}_i' - f_i f_i'\Big|\Big|^2 \\
&\leq \frac{1}{n}\sum_{i=1}^{i=n}\Big|\Big|\hat{f}_i' - f_i\Big|\Big|^2 + \frac{2}{n}\sum_{i=1}^{i=n}\Big|\Big|f_i\Big|\Big|\Big|\Big|\hat{f}_i' - f_i\Big|\Big| \\
&\leq o_p(1) + 2(\frac{1}{n}\sum_{i=1}^{i=n}\Big|\Big|f_i\Big|\Big|^2)^{\frac{1}{2}}(\frac{1}{n}\sum_{i=1}^{i=n}\Big|\Big|\hat{f}_i' - f_i\Big|\Big|^2)^{\frac{1}{2}} \\
&\leq o_p(1) + 2(O_p(1))^{\frac{1}{2}}(o_p(1))^{\frac{1}{2}}
\end{aligned}$$

Thus $\hat{\Omega} \xrightarrow{p} \Omega$.

This completes the proof ■