

Double Boosting GMM for High Dimensional IV Regression Models*

Tae-Hwy Lee[†] and Hao Xu[‡]

March 2018

Abstract

When endogenous variables are approximated by sieve functions of observable instruments the number of instruments increases rapidly and many may be invalid or irrelevant. We introduce Double Boosting (DB) which consistently selects only valid ‘and’ relevant instruments. Also we show DB does not select weakly relevant or weakly valid instruments. DB works even when there are more instruments than the sample size. Monte Carlo compares DB-GMM with other methods such as GMM using Lasso. In the application of estimating the BLP-type automobile demand function, where price is endogenous and instruments are high dimensional functions of product characteristics, we find the DB-GMM estimator of the price elasticity of demand is more elastic than other estimators.

Key Words: Causal inference with high dimensional instruments, Irrelevant instruments, Invalid instruments, Weakly relevant instruments, Weakly valid instruments, Boosting, Price elasticity of automobile demand.

JEL Classification: C1, C5, D2, L6

*We thank seminar participants at Academia Sinica, BJTU, CAS/AMSS, CUF, PKU, Ford Motor Company, UCLA, UCR (Econ), UCR (Stat), USC, and NIPS2017.

[†]Department of Economics, University of California, Riverside, CA 92521. E-mail: taelee@ucr.edu

[‡]Department of Economics, University of California, Riverside, CA 92521. E-mail: hao.xu@email.ucr.edu

1 Introduction

According to Berry, Levinsohn, and Pakes (1995, BLP henceforth), the two-stage least squares (2SLS) estimators of the logit demand function are inconsistent with the profit maximization behavior of firms because the estimated price elasticities of demand for a large number of cars are too small to make sense. Later, Chernozhukov, Hansen, and Spindler (2015) show that the inconsistency in the 2SLS estimation can be resolved by including high order polynomials and interaction terms of the instruments and control variables. These additional instruments and control variables help capturing the neglected nonlinearity.

However, the resulting high dimensionality of the instruments and control variables may cause collinearity problem. In the generalized method of moments (GMM) estimation, highly correlated instruments can make a singular weighting matrix.

In addition, Bekker (1994) states that the 2SLS estimator is inconsistent because the number of instruments is too many relative to the number of observations. Hence, the consistency of 2SLS estimators fails if instruments are in high dimension.

Another challenge with high dimensional instruments is the possible existence of weakly relevant instruments (weak instruments). According to Phillips (1989) and Staiger and Stock (1997), when instruments are weakly correlated with the endogenous variable, the 2SLS estimator fails the consistency because the asymptotic distribution of the estimator will be Cauchy-like (not normally distributed and has no moments), and the inference will be invalid. Similar problem exists in GMM estimation, as shown in Stock and Wright (2000), the asymptotic distribution of weakly identified parameters is not asymptotically normal.

Hence, an instrument selection procedure is necessary in order to ensure the consistency of these estimators, for which different approaches have been developed, such as the least absolute shrinkage and selection operator (Lasso), multiple testing, and information criteria.

While Lasso performs the variable selection, it produces bias in estimation. Belloni and Chernozhukov (2013) suggested the Post-Lasso estimation which can reduce the bias. Belloni, Chen, Chernozhukov, and Hansen (2012) apply Lasso and Post-Lasso for the first-stage prediction and instrument selection in a high dimensional IV regression model. Later in Chernozhukov, Hansen, and

Spindler (2015), they apply Lasso and Post-Lasso to both the first stage and the second stage of the 2SLS estimation when both instruments and control variables are in high dimension. Gillen, Moon, and Shum (2014) and Gillen, Montero, Moon, and Shum (2015) apply Lasso to select instruments and control variables for the BLP-type model.

Caner (2009), Caner and Zhang (2014) and Fan and Liao (2014) discuss the use of penalty for moment selection in GMM. Donald, Imbens, and Newey (2009) suggest a moment selection procedure by using an information criterion based on the asymptotic mean square error (MSE).

Different than using a Lasso type approach or information criteria, Bajari, Nekipelov, Ryan, and Yang (2015), Hartford, et al. (2016), and Chernozhukov, et al. (2016) apply machine learning techniques to the IV regression model. In particular, Ng and Bai (2008) consider L_2 Boosting for instrument selection. Bühlmann (2006) proves that L_2 Boosting achieves a consistent estimation on the regression function even when the number of regressors increases exponentially with the sample size. A simulation comparison between Lasso and L_2 Boosting in Bühlmann (2006) shows that both methods share very similar properties, although, as discussed in Meinshausen (2007), Lasso may have poor performance on variable selection in a high-dimensional linear model with many irrelevant regressors.

However, we note that majority of these papers assume that instruments are “valid”, such that instruments are not correlated with the structural error, and thus do not question the validity of instruments but only focus on the relevancy between instruments and the corresponding endogenous variables.

Only a few recent papers have relaxed the validity assumption on the instruments. DiTraglia (2016) allows highly relevant but somewhat invalid moments to be selected because of the benefit in reducing the MSE even at the cost of bias. That may be reasonable for prediction but not for inference. To make correct statistical inference, the bias should be the first priority before improving the overall efficiency measured by the MSE. Hence, it is important to remove all invalid moments to avoid bias. By adding different types of penalties to the GMM objective function, Liao (2013) illustrates how to perform moment selection when some of the moments are invalid. Similarly, Caner, Han, and Lee (2017) extend the adaptive elastic net GMM estimation by allowing many invalid moments. Cheng and Liao (CL, 2015) introduce the “Penalized GMM (PGMM)”

method with a cleverly modified Lasso and show that PGMM is asymptotically oracle in selecting valid and relevant moments.

In this paper, we propose another selection algorithm based on boosting, which we call “Double-Boosting (DB)”. We show that DB is asymptotically oracle in selecting valid and relevant instruments from a set of high dimensional instruments that may be either (weakly) invalid or/and (weakly) irrelevant. DB is based on a new criterion, which will check both the validity and the relevancy of each potential instrument. We prove that DB consistently selects valid and relevant instruments simultaneously. More importantly, we show that DB will not select weakly valid instruments or weakly relevant instruments (with the extent of ‘weakness’ being defined for the local-to-zero asymptotics). Furthermore, in proving the consistency of DB, we allow the endogenous variable to be an unknown nonlinear function of instruments, which we approximate by a set of sieve functions, e.g., polynomials of observable instruments as in Chernozhukov, Hansen, and Spindler (2015). Once DB selects instruments, we compute the GMM estimator using the selected instruments. The estimator will be called the “DB-GMM” estimator.

This paper is organized as follows. In Section 2, we set up a structural model for the high dimensional IV regression, then define weak/strong validity or relevancy of instruments and classify them in several different categories. In Section 3, we review the instrument selection procedure by L_2 Boosting in the literature. Since the estimator is computed by GMM with the selected instruments, we refer to this method as “Boosting GMM (BGMM)”. In Section 4, we propose a new instrument selection method, DB. In Section 5, Monte Carlo studies are presented to compare DB-GMM with other methods. Section 6 is the empirical application that follows the design in Berry, Levinsohn, and Pakes (1995) and Chernozhukov, Hansen, and Spindles (2015) to demonstrate the merits of using the Double-Boosting algorithm. Section 7 concludes. All proofs are gathered in Section 8 (Appendix).

2 Model

Consider an IV model as

$$y_i = \beta' x_i + u_i \quad (1)$$

$$x_i = E(x_i|w_i) + v_i. \quad (2)$$

For $i = 1, \dots, n$, y_i is the scalar dependent variable, x_i is a $k \times 1$ vector of endogenous variables, and β is a $k \times 1$ vector of parameters. The conditional mean $E(x_i|w_i)$ is an unknown function of observable instruments w_i , where $w_i = (w_{1,i} \dots w_{p,i})'$ is a $p \times 1$ vector. The two error terms u_i and v_i have dimensions of 1×1 and $k \times 1$ respectively and have the $(k+1) \times (k+1)$ variance-covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \quad (3)$$

According to Belloni, Chen, Chernozhukov, and Hansen (2012), the exact sparse model can be estimated by the “approximately sparse model” with an approximation error r_i . $E(x_i|w_i)$ can be approximated by a linear combination of sieve functions $h(w_i) = (h_1(w_i) \dots h_{\ell_n}(w_i))'$ such that

$$E(x_i|w_i) = \sum_{j=1}^{\ell_n} \gamma_j h_j(w_i) + r_i, \quad (4)$$

where the parameter γ_j is a $k \times 1$ vector for each $j = 1, \dots, \ell_n$, and $r_i = (r_{1,i} \dots r_{k,i})'$ is a $k \times 1$ vector of the approximation error. Since the functional form of $h_j(\cdot)$ is known, we define a sieve instrument $z_{j,i} \equiv h_j(w_i)$ and

$$(z_{1,i} \dots z_{\ell_n,i})' \equiv (h_1(w_i) \dots h_{\ell_n}(w_i))'. \quad (5)$$

From Equations (2) and (4),

$$x_i = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + r_i + v_i. \quad (6)$$

The validity and the relevancy of instruments are defined in a local asymptotic framework. The moment function of each instrument $z_{j,i}$ for $j = 1, \dots, \ell_n$ is

$$g(z_{j,i}, \beta) = z_{j,i} u_i. \quad (7)$$

The validity of each instrument depends on the moment condition,

$$E(g(z_{j,i}, \beta)) = E(z_{j,i}u_i) = \frac{b_j}{n^{\delta_j}}. \quad (8)$$

And the relevancy of each instrument depends on the parameter,

$$\gamma_j = \frac{a_j}{n^{\alpha_j}}. \quad (9)$$

Let $Z_j = (z_{j,1} \dots z_{j,n})'$ for $j = 1, \dots, \ell_n$. We define deferent degrees of validity and relevancy as stated below.

Definition 1 (Validity): The extent of validity depends on b_j and δ_j as follows: $\mathcal{V}_1 = \{j : b_j = 0\} \cup \{j : b_j \neq 0 \text{ and } \frac{1}{2} < \delta_j\}$, $\mathcal{V}_2 = \{j : b_j \neq 0 \text{ and } 0 < \delta_j \leq \frac{1}{2}\}$, and $\mathcal{V}_3 = \{j : b_j \neq 0 \text{ and } \delta_j = 0\}$. Then, Z_j is said to be a strongly valid instrument if $j \in \mathcal{V}_1$, a weakly valid instrument if $j \in \mathcal{V}_2$, and Z_j an invalid instrument if $j \in \mathcal{V}_3$.

Definition 2 (Relevancy): The extent of relevancy depends on a_j and α_j as follows: $\mathcal{R}_1 = \{j : a_j = 0\}$, $\mathcal{R}_2 = \{j : a_j \neq 0 \text{ and } \alpha_j > 0\}$, and $\mathcal{R}_3 = \{j : a_j \neq 0 \text{ and } \alpha_j = 0\}$. Then, Z_j is said to be an irrelevant instrument if $j \in \mathcal{R}_1$, a weakly relevant instrument if $j \in \mathcal{R}_2$, and a strongly relevant instrument if $j \in \mathcal{R}_3$.

We partition the set of instruments into two subsets, \mathcal{S} and \mathcal{D} , following Cheng and Liao (2015). The “sure” set $\mathcal{S} = \{Z_1, \dots, Z_{\ell_S}\}$ includes the strongly valid and strongly relevant instruments that are initially selected, and ℓ_S denotes the total number of instruments in \mathcal{S} . The “doubt” set $\mathcal{D} = \{Z_{\ell_S+1}, \dots, Z_{\ell_n}\}$ is the set of instruments that are not in \mathcal{S} , and we do not know the validity and relevancy of these instruments in \mathcal{D} . Hence, an instrument selection is needed for instruments in \mathcal{D} . We further partition \mathcal{D} into three subsets, $\mathcal{D} = \mathcal{A} \cup \mathcal{B}_0 \cup \mathcal{B}_1$. The subset \mathcal{A} is a set of strongly valid and strongly relevant instruments that share the same properties as instruments in \mathcal{S} . The subset \mathcal{B}_0 is a set of strongly valid but irrelevant or weakly relevant instruments, and the subset \mathcal{B}_1 is a set of invalid or weakly valid instruments that are not in $\mathcal{A} \cup \mathcal{B}_0$. Our goal is to select only instruments in \mathcal{A} but none from $\mathcal{B} \equiv \mathcal{B}_0 \cup \mathcal{B}_1$. Table 1 summarizes each subset of the instruments according to Definitions 1 and 2.

3 Boosting GMM (BGMM)

Ng and Bai (2008) propose a two-stage procedure for the high dimensional IV regression model, which we refer to as Boosting GMM (BGMM). At the first stage, instruments are selected through L_2 Boosting. Then, with the selected instruments, the parameter of interest β is estimated by GMM at the second stage.

Referring to the model described in Section 2, \mathcal{S} includes all the strongly valid and strongly relevant instruments that are initially selected. The instruments in \mathcal{D} are the potential instruments that will be selected by L_2 Boosting. At each step $m = 1, \dots, \bar{M}$, where \bar{M} is the maximum iteration of L_2 Boosting, we regress the “current residual” on each instrument in \mathcal{D} , and select the instrument that is most relevant to the “current residual”. We denote $F_{m,i} = F_{m,i}(z_i)$ as the strong learner and $f_{m,i} = f_{m,i}(z_i)$ as the weak learner for $i = 1, \dots, n$. The relationship between the weak learner and the strong learner is

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \quad (10)$$

where $c_m > 0$ is a learning rate. For simplicity, we assume the dimension of x_i to be $k = 1$ and $\sigma_2^2 = \Sigma_{22}$. If $k > 1$, we repeat L_2 Boosting for each variable in x_i .

3.1 L_2 Boosting Algorithm

The detail description of L_2 Boosting is listed in Algorithm 1.

Algorithm 1 BGMM

1. When $m = 0$, the initial weak learner of $X = (x_1 \dots x_n)'$ using instruments in \mathcal{S} is

$$F_{0,i} = f_{0,i} = \hat{\gamma}_{0,\text{initial}} + \sum_{j=1}^{\ell_{\mathcal{S}}} \hat{\gamma}_{j,\text{initial}} z_{j,i}, \quad (11)$$

where $\hat{\gamma}_{0,\text{initial}}$ and $\hat{\gamma}_{j,\text{initial}}$ are the OLS estimators.

2. For each step $m = 1, \dots, \bar{M}$
- (a) We compute the “current residual”, $\hat{v}_{m,i} = x_i - F_{m-1,i}$.
 - (b) Next, we regress the current residual $\hat{v}_{m,i}$ on each instrument $z_{j,i}$, for $j = \ell_{\mathcal{S}} + 1, \dots, \ell_n$. The estimators $\hat{\gamma}_0$ and $\hat{\gamma}_j$ are solved as

$$\{\hat{\gamma}_{0,j}, \hat{\gamma}_j\} = \min_{\gamma_0, \gamma_j} \sum_{i=1}^n (\hat{v}_{m,i} - \gamma_0 - \gamma_j z_{j,i})^2. \quad (12)$$

We select the instrument that has the minimum sum of squared residuals, such that

$$j_m = \arg \min_{j \in \{\ell_{\mathcal{S}}+1, \dots, \ell_n\}} \sum_{i=1}^n (\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i})^2. \quad (13)$$

- (c) The weak learner is

$$f_{m,i} = \hat{\gamma}_{0,j_m} + \hat{\gamma}_{j_m} z_{j_m,i}, \quad (14)$$

where $z_{j_m,i}$ is the instrument that is selected.

- (d) The strong learner $F_{m,i}$ is updated as

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \quad (15)$$

with $c_m > 0$.

3. We compute the GMM estimator using the selected instruments.
-

A stopping rule is necessary in L_2 Boosting in order to avoid over-fitting. The optimal number of iteration \hat{M} is chosen by a version of AIC suggested in Bühlmann (2006). Let $\hat{V}_m = (\hat{v}_{m,1} \dots \hat{v}_{m,n})'$, $f_m = (f_{m,1} \dots f_{m,n})'$, $F_m = (F_{m,1} \dots F_{m,n})'$, and $\mathbf{1}$ be an $n \times 1$ vector of ones. We define $\mathbf{Z}_{j_m} = [\mathbf{1} \ Z_{j_m}]$, and $P_m = \mathbf{Z}_{j_m} (\mathbf{Z}_{j_m}' \mathbf{Z}_{j_m})^{-1} \mathbf{Z}_{j_m}'$ to be an $n \times n$ matrix. From Equation (14),

$$\begin{aligned} \mathbf{1} \hat{\gamma}_{0,j_m} + \mathbf{Z}_{j_m} \hat{\gamma}_{j_m} &= P_m \hat{V}_m \\ f_m &= P_m (X - F_{m-1}). \end{aligned} \quad (16)$$

Let $\mathbf{Z}_S = (Z_1 \dots Z_{\ell_S})$. When $m = 0$, $P_{j_0} = \mathbf{Z}_S(\mathbf{Z}_S' \mathbf{Z}_S)^{-1} \mathbf{Z}_S'$. Then the strong learner at each step m is

$$\begin{aligned} F_m &= F_{m-1} + c_m P_m (X - F_{m-1}) \\ &= \left[I_{n \times n} - \prod_{a=0}^m (I_{n \times n} - c_{j_a} P_{j_a}) \right] X =: B_m X. \end{aligned}$$

AIC is computed as

$$AIC_c(m) = \log(\hat{\sigma}_{2,m}^2) + \frac{1 + \text{trace}(B_m)/n}{1 - (\text{trace}(B_m) + 2)/n}, \quad (17)$$

where $\log(\hat{\sigma}_{2,m}^2) = \frac{1}{n} \sum_{i=1}^n (\hat{v}_{m,i} - c_m f_{m,i})^2$. Then $\hat{M} = \arg \min_{m=1, \dots, \bar{M}} AIC_c(m)$.

3.2 Consistency of L_2 Boosting

Consider the following assumptions from Bühlmann (2006).

Assumption 1: *The dimension of instruments satisfies $\ell_n = O(\exp(Cn^{1-\eta}))$, $n \rightarrow \infty$, for some $0 < \eta < 1$, $0 < C < \infty$.*

Assumption 2: $\sup_{n \in \mathbb{N}} \sum_{j=1}^{\ell_n} |\gamma_j| < \infty$.

Assumption 3: $\sup_{1 \leq j \leq \ell_n, n \in \mathbb{N}} \|Z_j\|_\infty < \infty$, where $\|Z_j\|_\infty = \sup_{\omega \in \Omega} |Z_j(\omega)|$ and Ω denotes the underlying probability space.

Assumption 4: $E|v_i|^s < \infty$ for some $s > 4/\eta$ with η in Assumption 1.

In Assumption 1, the dimension of instruments is allowed to grow exponentially with respect to the number of observations. So instruments can be in a high dimension. Assumption 2 gives an L_1 -norm sparseness condition that the sum of the coefficient γ_j for all j is bounded. Hence, only finite number of instruments are strongly relevant. In addition, Assumption 2 can be generalized to $\sum_{j=1}^{\ell_n} |\gamma_j| \rightarrow \infty$ as $n \rightarrow \infty$, but additional restriction on ℓ_n is needed. In this case, all instruments may be relevant, but the contribution of very large proportion of instruments is small. Hence weakly relevant instruments are allowed in the model. Assumption 3 states that by restricting the growth rate of ℓ_n , the maximum realization of random variable Z_j under sample space Ω needs to be bounded. In Assumption 4, the existence of some higher moments of the error term v_i is needed, and the number of existing moments depends on η from Assumption 1. Thus the number of existing moments and the growth rate of ℓ_n are related.

According to Bühlmann (2006 Theorem 1), the L_2 Boosting estimation converges to the conditional mean of x_i in quadratic mean under a linear model. We extend this result of Bühlmann (2006) to the case when $E(x_i|w_i)$ is nonlinear and is approximated by the approximately sparse model in Belloni, Chen, Chernozhukov, and Hansen (2012). Recall Equation (6)

$$x_i = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + r_i + v_i,$$

where $\{z_{j,i}\}$ is a set of sieve instruments such as polynomials of instruments in w_i , and r_i is the approximation error. Here we make an additional assumption to control the relative size of the sparse approximation error r_i with respect to the size of the error term v_i and number of sieve instruments ℓ_n .

Assumption 5: When $E(x_i|w_i) = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + r_i$ is approximated by a linear function of sieves $\{z_{j,i}\}$, the sparse approximation error r_i satisfies that $E(r_i^2|w_i) \leq \sigma_2^2 \left(\frac{\log \ell_n}{n} \right)$, where $\sigma_2^2 = E(v_i^2)$.

Assumption 5 requires that the mean squared approximation error needs to be bounded by the product of the variance of v_i and $\frac{\log(\ell_n)}{n}$. We now state a theorem that L_2 Boosting still works in the sense that $F_{m_n,i}$ converges to $E(x_i|w_i)$ in quadratic mean.

Theorem 1: Let $E(x_i|w_i) = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + r_i$ be approximated by a linear function of sieves $\{z_{j,i}\}$. Under Assumptions 1-5, for some sequence $(m_n)_{n \in \mathbb{N}}$ with $m_n \rightarrow \infty$ sufficiently slowly as $n \rightarrow \infty$, the L_2 Boosting estimation converges to the conditional mean of x_i ,

$$E \left[\frac{1}{n} \sum_{i=1}^n (F_{m_n,i} - E(x_i|w_i))^2 \middle| W \right] = o_p(1) \text{ as } n \rightarrow \infty.$$

where $W = (W_1 \dots W_p)$.

Proof: Appendix A.

However, L_2 Boosting can only check for the relevancy of instruments but not the validity of instruments. Theorem 1 may still hold with the existence of invalid instruments. But a possible selection of weakly valid or invalid instruments by L_2 Boosting will cause the BGMM estimators to be inconsistent for β . Hence, we develop a new boosting algorithm to select only relevant and valid instruments, which we discuss next.

4 Double-Boosting GMM (DB-GMM)

We propose a new selection procedure, DB, that checks for both the relevancy and the validity of instruments. After the selection, we use GMM to compute the estimators, DB-GMM.

4.1 Double-Boosting Algorithm

The DB algorithm is described in Algorithm 2. The new selection algorithm (Algorithm 2) is similar to L_2 Boosting (Algorithm 1) in the previous section, except Step 2(b). The difference is the new objective function (24) replacing (13). We now doubly minimize the invalidity (measured by (21)) and minimize the irrelevancy (measured by the inverse of (25)) of an instrument in each iteration, as we describe in details below.

First, we measure the invalidity based on the usual Lagrange Multiplier (LM) test statistic. It is now more convenient to use the correlation coefficient instead of using the covariance between Z_j and U as in the moment condition for Algorithm 1. Let

$$\rho_j = \frac{E(z_{j,i}u_i)}{\sqrt{E(z_{j,i}^2)}\sqrt{E(u_i^2)}} = \frac{d_j}{n^{\delta_j}}. \quad (18)$$

where $d_i = \frac{b_j}{\sqrt{E(z_{j,i}^2)}\sqrt{E(u_i^2)}}$ and b_j defined in Equation (8).

We estimate ρ_j by using the initial 2SLS estimator $\hat{\beta}_{\text{initial}}$, which is computed using the instruments in set \mathcal{S} . Then the residual with the initial 2SLS estimators,

$$\hat{u}_i \equiv y_i - \hat{\beta}_{\text{initial}}x_i, \quad (19)$$

is used to obtain the sample correlation coefficient between \hat{U} and each $Z_j \in \mathcal{D}$, that is

$$\hat{\rho}_j = \frac{\frac{1}{n} \sum_{i=1}^n z_{j,i} \hat{u}_i}{\sqrt{\frac{1}{n} \sum_{i=1}^n z_{j,i}^2} \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2}}. \quad (20)$$

Then we define the LM statistic measure for invalidity of z_j as

$$nR_{\mathcal{V},j}^2 = n\hat{\rho}_j^2. \quad (21)$$

Similarly, we also define the LM statistic measure for relevancy of z_j , $nR_{\mathcal{R},j}^2$, which we describe in (25) inside Algorithm 2.

Algorithm 2 DB-GMM

1. When $m = 0$, the initial weak learner of $X = (x_1 \dots x_n)'$ using instruments in \mathcal{S} is

$$F_{0,i} = f_{0,i} = \hat{\gamma}_{0,\text{initial}} + \sum_{j=1}^{\ell_S} z_{j,i} \hat{\gamma}_{j,\text{initial}}, \quad (22)$$

where $\hat{\gamma}_{0,\text{initial}}$ and $\hat{\gamma}_{j,\text{initial}}$ are the OLS estimators.

2. For each step $m = 1, \dots, \bar{M}$

- (a) The “current residual” is defined as $\hat{v}_{m,i} = x_i - F_{m-1,i}$.
(b) Next, we regress the current residual $\hat{v}_{m,i}$ on each instrument $z_{j,i}$, for $j \in \{\ell_S + 1, \dots, \ell_n\}$. The estimators $\hat{\gamma}_{0,j}$ and $\hat{\gamma}_j$ are solved as

$$\{\hat{\gamma}_{0,j}, \hat{\gamma}_j\} = \min_{\gamma_0, \gamma_j} \sum_{i=1}^n (\hat{v}_{m,i} - \gamma_0 - \gamma_j z_{j,i})^2. \quad (23)$$

We select the instrument $z_{j_m,i}$ that gives the minimum ω_j , i.e.,

$$j_m = \arg \min_{j \in \{\ell_S + 1, \dots, \ell_n\}} \omega_j \equiv \frac{\left(n R_{\mathcal{V},j}^2\right)^{r_2}}{\left(n R_{\mathcal{R},j}^2\right)^{r_1}}, \quad (24)$$

where

$$R_{\mathcal{R},j}^2 = 1 - \frac{\sum_{i=1}^n (\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i})^2}{\sum_{i=1}^n (\hat{v}_{m,i} - \bar{v}_m)^2}, \quad (25)$$

$\bar{v}_m = \frac{1}{n} \sum_{i=1}^n \hat{v}_{m,i}$, and r_1 and r_2 are the user selected constants such that $r_1, r_2 > 0$.

- (c) The weak learner is

$$f_{m,i} = \hat{\gamma}_{0,j_m} + \hat{\gamma}_{j_m} z_{j_m,i}, \quad (26)$$

where $z_{j_m,i}$ is the instrument that is selected.

- (d) The strong learner $F_{m,i}$ is updated as,

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \quad (27)$$

with $c_m > 0$.

3. We compute the GMM estimator using the selected instruments.
-

Remark 1: We introduce the selection criterion ω_j to check the validity and relevancy of each instrument Z_j . The user selected constants r_1 and r_2 are used to control the penalty on the validity and relevancy. With a higher value in r_2 , the invalid instrument will be punished more with a higher numerator in ω_j . On the other hand, a higher value in r_1 will ensure the relevant

instrument to obtain a higher denominator in ω_j , which leads to a smaller value in ω_j . In simulation and application, we report the results using $r_1 = r_2 = 1$. We have experimented by simulation with different values of r_2 with fixing $r_1 = 1$. (i) When $r_1 > r_2$, the penalty on invalid instruments is weaker. The probability of selecting invalid instruments will be higher. Then the DB-GMM estimation may become more biased. Our simulation results confirm that the bias is larger when $r_1 > r_2$ than when $r_1 = r_2$. (ii) When $r_1 < r_2$, the penalty on invalid instruments is stronger. The simulation results shows that the bias and the mean squared error (MSE) when $r_1 < r_2$ are not significantly different from the default setting with $r_1 = r_2$. This highlights the importance of removing invalid instruments by choosing r_2 such that $r_1 \leq r_2$, a feature of DB, that is absent in L_2 Boosting.

Remark 2: Our selection criterion ω_j shares similar property as the information based adjustment in PGMM of Cheng and Liao (2015). However, for each j , PGMM only adds one Z_j in \mathcal{D} to \mathcal{S} to check the relevancy of the corresponding instrument. In Double-Boosting, we update the current residual $\hat{v}_{m,i}$ at each DB iteration. Then the relevancy criterion $nR_{\mathcal{R},j}^2$ is not only depended on \mathcal{S} but also on all the previously selected instruments.

Remark 3: The stopping rule in DB is the same as in L_2 Boosting. As $R_{\mathcal{V},j}^2$ is computed based on the 2SLS estimation using only instruments in \mathcal{S} , $nR_{\mathcal{V},j}^2$ is fixed at any iteration $m = 1, \dots, \bar{M}$. In addition, minimizing $\frac{1}{nR_{\mathcal{R},j}^2}$ is the same as maximizing $nR_{\mathcal{R},j}^2$. According to the definition, the maximization of $R_{\mathcal{R},j}^2$ can be achieved by minimizing the ratio $\frac{\sum_{i=1}^n (\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i})^2}{\sum_{i=1}^n (\hat{v}_{m,i} - \bar{v}_m)^2}$. Since $\sum_{i=1}^n (\hat{v}_{m,i} - \bar{v}_m)^2$ is the same for all j at each m , $\frac{\sum_{i=1}^n (\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i})^2}{\sum_{i=1}^n (\hat{v}_{m,i} - \bar{v}_m)^2} \propto \sum_{i=1}^n (\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i})^2$. Note that $\sum_{i=1}^n (\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i})^2$ is the criterion (13) for L_2 Boosting. Hence, the same stopping rule is applied to DB.

Next, in Theorem 2, we prove that DB will only select the strongly valid and strongly relevant instruments in \mathcal{A} , and will not select any instrument in \mathcal{B}_0 or \mathcal{B}_1 , with probability one asymptotically. In other words, DB will ensure that ω_{j_m} for all $Z_{j_m} \in \mathcal{A}$ will be smaller than ω_j for $Z_j \in \mathcal{B} \equiv \mathcal{B}_0 \cup \mathcal{B}_1$, with probability approaching 1 (w.p.a.1) in each iteration m .

Theorem 2: Under Assumptions 1-5, in each iteration m , the selected instrument Z_{j_m} is strongly valid and strongly relevant w.p.a.1 as $n \rightarrow \infty$. That is,

$$\Pr(\omega_{j_m} < \omega_j) \rightarrow 1 \text{ for all } Z_j \in \mathcal{B}, \text{ as } n \rightarrow \infty,$$

and thus, the selected instrument $Z_{j_m} \in \mathcal{A}$.

Proof: Appendix B.

5 Monte Carlo

To study the finite sample properties of different estimation methods under the high dimensional IV regression model, we consider the following three data generating processes (DGPs).

DGP 1 (Linear):

$$\begin{aligned} y_i &= \beta x_i + u_i, \\ x_i &= \sum_{j=1}^p \gamma_j w_{j,i} + v_i = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + v_i, \end{aligned} \quad (28)$$

where the endogenous variable x_i is a scalar ($k = 1$), and $z_{j,i} = w_{j,i}$. DGP 1 follows the design of DGP in Cheng and Liao (2015). We set $\beta = 0$ as the true value, $n \in \{100, 250\}$, and $p = \ell_n = 52$. Let $z_{S,i} = (z_{1,i} \ z_{2,i})'$ be the strongly valid and strongly relevant instruments in \mathcal{S} . Let $z_{A,i} = (z_{3,i} \ z_{4,i})'$, $z_{B_0,i} = (z_{5,i} \ \dots \ z_{28,i})'$ and $z_{B_1,i} = (z_{29,i} \ \dots \ z_{52,i})'$ be the “doubt” instruments in \mathcal{D} . We set $\gamma_1 = 0.1$, $\gamma_2 = 0.3$, $\gamma_3 = 0.5$, $\gamma_4 \in \{0.5, 0.01\}$, and $\gamma_j = 0$ for any $j \geq 5$. Then $z_{4,i}$ is a weakly relevant instrument if $\gamma_4 = 0.01$. In order to compute the invalid instrument $z_{B_1,i}$, we first need to generate a strongly valid instrument $z_{B_1,i}^*$. The strongly valid instruments and error terms follow the normal distribution where

$$(z_{S,i} \ z_{A,i} \ z_{B_0,i} \ z_{B_1,i}^*) \sim N(0, \Sigma_Z) \quad (29)$$

$$(u_i \ v_i) \sim N(0, \Sigma), \quad (30)$$

and $\Sigma = \begin{pmatrix} 0.5 & 0.6 \\ 0.6 & 1 \end{pmatrix}$. For Σ_Z , we consider two different cases. In the first case, it is exactly the same as in Cheng and Liao (2015), where $\Sigma_Z = \text{diag}(\Sigma_{S \cup A}, \Sigma_B)$. $\Sigma_{S \cup A}$ is a 4×4 Toeplitz matrix that each (i, j) element equals to $0.2^{|i-j|}$, and Σ_B is an $(\ell_n - 4) \times (\ell_n - 4)$ identity matrix. We

denote the first case as “CL” in Table 3. In the second case, Σ_Z is an $\ell_n \times \ell_n$ Toeplitz matrix, where each (i, j) element equals to $a^{|i-j|}$ with $a \in \{0.5, 0.9\}$. Lastly, following Cheng and Liao (2015), for $j = 29, \dots, 52$, the invalid instrument $z_{j,i}$ is generated as

$$z_{j,i} = z_{j,i}^* + c_j u_i, \quad (31)$$

where $z_{j,i}^*$ is the strongly valid instrument in $z_{B_1,i}^*$, and

$$c_j = c_0 + \frac{(j - 29)(\bar{c} - c_0)}{\ell_n/2 - 2}. \quad (32)$$

So c_j increases from c_0 to \bar{c} as j increases. We choose $c_0 = 0.2$, $\bar{c} = 2.4$.

DGP 2 (Polynomials):

$$\begin{aligned} y_i &= \beta x_i + u_i \\ x_i &= \sum_{j=1}^p \theta_j (w_{j,i} + w_{j,i}^2) + v_i, \end{aligned} \quad (33)$$

where x_i is a scalar, $\beta = 0$, and $n \in \{100, 250\}$ as in DGP 1. Let $p = 5$, then the observable strongly valid instruments are generated as

$$(w_{1,i} \ w_{2,i} \ w_{3,i} \ w_{4,i} \ w_{5,i}^*) \sim N(0, \Sigma_W), \quad (34)$$

where Σ_W is a $p \times p$ Toeplitz matrix with each (i, j) element $a^{|i-j|}$ and $a \in \{0, 0.5, 0.9\}$. We set $\theta_1 = \theta_2 = 0.1$, $\theta_3 = 0.5$, and $\theta_4 = \theta_5 = 0$. So only the first three observable instruments are strongly relevant to x_i . The error terms u_i and v_i are generated as

$$(u_i \ v_i) \sim N(0, \Sigma), \quad (35)$$

where $\Sigma = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 1 \end{pmatrix}$. To generate an invalid instrument, we contaminate $w_{5,i}^*$, which was constructed as a valid instrument in (34), by adding the structural error u_i

$$w_{5,i} = w_{5,i}^* + u_i. \quad (36)$$

DGP 3 (Exponential): The generation of variables in DGP 3 is similar as in DGP 2. The only difference is that x_i is generated as an additively separable exponential function of w_i ,

$$x_i = \sum_{j=1}^p \theta_j \exp(w_{j,i}) + v_i. \quad (37)$$

In DGP 1, as $z_{j,i} = w_{j,i}$, all instruments are observable and the functional form of $h_j(\cdot)$ in (4) is known. In DGP 2 and DGP 3, the functional form of x_i is unknown. We approximate x_i using sieve instruments $\{z_{j,i}\}$. We set $z_{j,i} = w_{j,i}$ for $j = 1, \dots, p$, and $z_{j,i} = h_j(w_i)$ for $j = p+1, \dots, \ell_n$, where $h_j(w_i)$ is the polynomial of each instrument in w_i up to 4th order. Let $z_{\mathcal{S},i} = (z_{1,i} \ z_{2,i})'$. We summarize the simulation results in Tables 3 to 5.

5.1 Simulation Results

We compare DB-GMM with six different methods that are OLS, 2SLS ^{\mathcal{SD}} (2SLS with all instruments in $\mathcal{S} \cup \mathcal{D}$), 2SLS ^{\mathcal{S}} (2SLS with only instruments in \mathcal{S}), 2SLS ^{\mathcal{SA}} (2SLS with all strongly valid and strongly relevant instruments in $\mathcal{S} \cup \mathcal{A}$), BGMM, and PGMM. 2SLS ^{\mathcal{SA}} gives the oracle result. We choose the learning rate for boosting to be $c_m = 0.01$ for all m . The user selected parameters in PGMM are the same as in Cheng and Liao (2015). For each result reported in Tables 3 to 5, the bias of the estimator $\hat{\beta}$ is in the first row and the root mean squared errors (RMSE) is in the second row.

DGP 1 is linear where all instruments are observable. Compared to the oracle result in the column of 2SLS ^{\mathcal{SA}} , the bias and the RMSE of the OLS estimation are large because x_i is endogenous. As the correlation between instruments becomes stronger, the OLS estimation has slight improvement in its bias and RMSE. The estimation in 2SLS ^{\mathcal{SD}} is also inconsistent because of the existence of invalid and irrelevant instruments. The bias and the RMSE are even higher in 2SLS ^{\mathcal{SD}} than in OLS for most cases. 2SLS ^{\mathcal{S}} is able to maintain a low bias but a higher RMSE compared to 2SLS ^{\mathcal{SA}} . The reason is that DGP 1 has only four strongly valid and strongly relevant instruments, and only two of them are included in \mathcal{S} . When the coefficient of the fourth instrument ($Z_4 \in \mathcal{A}$) reduces from 0.5 to 0.01, the bias of 2SLS ^{\mathcal{S}} is similar to the case when $\gamma_4 = 0.5$, but the RMSE is slightly higher with the weak instrument ($\gamma_4 = 0.01$). BGMM has similar problem as in 2SLS ^{\mathcal{SD}} .

Due to the inclusion of invalid instruments, BGMM has a higher bias and RMSE than OLS in most of cases. The bias and RMSE of OLS, $2SLS^{S^D}$, and BGMM become significantly worse when γ_4 reduces to 0.01. Both of the last two methods, PGMM and DB-GMM, are able to check the validity and relevancy of the instruments. When $\gamma_4 = 0.5$ (strong instrument), PGMM has a lower bias than DB-GMM, but the RMSE of DB-GMM is always the smallest among all other methods (excluding the oracle $2SLS^{S^A}$). When γ_4 decreases to 0.01, PGMM still has a lower bias than DB-GMM. However, the RMSE of PGMM now is lower than the RMSE of DB-GMM in 3 out of 6 cases. In general, when the correlation between instruments increases (a increases), the results of all methods are improving. Especially when $a = 0.9$, the results of $2SLS^S$, PGMM, and DB-GMM are close to the oracle result. This happens because when instruments are highly correlated, selecting a few strongly valid and strongly relevant instruments will be as efficient as selecting all instruments in $\mathcal{S} \cup \mathcal{A}$.

In both DGP 2 and DGP 3, there are total of 125 sieve instruments. Because the sieve instruments Z are generated from the polynomial of w_i , high collinearity between instruments exists even when there has not been correlation between w_i ($a = 0$). In DGP 2, OLS has large bias and large RMSE because of the endogeneity. When $\ell_n > n$, the RMSE of $2SLS^{S^D}$ diverges, which confirms the theoretical result in Bekker (1994). With only instruments in \mathcal{S} , $2SLS^S$ has improvement in the result. But both the bias and RMSE of $2SLS^S$ remain high because $2SLS^S$ fails to capture any nonlinearity in the endogenous variable. The performance of BGMM is very stable across all cases even when $\ell_n > n$. PGMM fails for $\ell_n > n$, where the weighting matrix is not invertible during the estimation. It also fails when $a = 0.9$ and $n = 250$ because of the high collinearity among all sieve instruments. These problems can be solved by replacing the weighting matrix with an identity matrix. However, the RMSE of PGMM using the identity weighting matrix is strictly higher than the RMSE of DG-GMM. DB-GMM has the lowest bias and RMSE for most of the cases. The results in DGP 3 are very similar to DGP 2. Hence, we conclude that DB-GMM has the best performance in the nonlinear cases as demonstrated in the results of DGP 2 and DGP 3.

6 Estimation of Automobile Demand Function

We apply DB-GMM to estimate the automobile demand function of BLP (1995). For simplicity, consider a homogeneous individual log utility function

$$\xi_{it} = \delta(w_{it}, x_{it}, u_{it}, \beta) + \varepsilon_{it}, \quad (38)$$

where $\delta(w_{it}, x_{it}, u_{it}, \beta) = \delta_{it}$ is a function that includes all information on the product characteristics of car i in year t . The subscription it together denotes one car. Let x_{it} denote the price of each car it , w_{it} be a vector of the observable market level product characteristics of a car it , u_{it} be the unobservable product characteristics of a car it which cause the endogeneity in the price, and β be the parameters in $\delta(\cdot)$. Applying the simple logit model, the market share s_{it} for each car it is calculated as

$$s_{it} = \frac{\exp(\delta_{it})}{1 + \sum_{\forall it} \exp(\delta_{it})}. \quad (39)$$

Suppose δ_{it} is linearized in all of its components. The demand equation in terms of market share can be calculated as

$$y_{it} = \beta_0 + \beta_{\text{price}} x_{it} + \beta'_w w_{it} + u_{it}, \quad (40)$$

where $y_{it} = \log(s_{it}) - \log(s_{0t})$, and s_{0t} is the outside option in year t . The outside option refers to consumers' choosing to buy a used car or to use alternative transportations.

Since price is endogenous, by applying the “approximately sparse model” in (4), we assume price is a linear combination of product characteristics and sieve functions of product characteristics such that

$$x_{it} = \gamma_0 + \gamma'_w w_{it} + \gamma'_1 h_1(w_{it}) + \gamma'_2 h_2(w_{it}, t) + \gamma'_3 h_3(w_{it}) + v_{it}, \quad (41)$$

where $h_1(w_{it})$ is the set of quadratic and cubic terms of continuous variables in w_{it} , and $h_2(w_{it}, t)$ is the set of the first order interactions of all variables in w_{it} and time t . We generate additional instruments in $h_3(w_{it})$ as follows: 1) the sum of each characteristics of other cars that are produced by the same firm in the same year as car it , and the count of these cars; 2) the sum of each characteristics of cars that are produced by other firms in the same year as car it , and the count of these cars. It is necessary to include instruments in $h_3(w_{it})$ because the product characteristics of competitive cars also influence the price.

The data used in BLP (1995) is obtained from annual issues of the *Automotive News Market Data Book* from 1971 to 1990. The product characteristics in the data set are weight, horsepower, length, width, miles per gallon ratio (MPG), and a dummy variable for air condition as a standard equipment. Price is obtained from the listed retail price of the base model in the unit of 1000 dollars of year 1983. In addition, the price of gasoline is also included in the data. With the given information, we calculate miles per dollar (MP\$) by MPG divided by the price per gallon. With treating each model of a car in each year as one car, there are total of 2217 cars included in the data set. Hence, the model in (40) and (41) are estimated as if the data is cross-sectional (no time series) for $it = 1, \dots, 2217$.

We use the data set in Chernozhukov, Hansen, and Spindles (2015), who also study the automobile application in BLP (1995). We include 4 control variables in the model - namely, the dummy variable of air conditioning (AC), horsepower/weight (HPW), miles per dollar (MP\$), and size of car (Size). We denote these control variables as $w_{it} = (\text{AC}_{it} \text{ HPW}_{it} \text{ MP\$}_{it} \text{ Size}_{it})'$. There are total 63 instruments, including the constant. Since the first 4 instruments are control variables, we assume all these 4 variables are valid. We select the constant and all four control variables for \mathcal{S} . The rest of instruments are in \mathcal{D} . In order to be consistent with the profit maximization behavior of the firm, the number of cars that have inelastic demand need to be small, because if the demand were inelastic to price changes, firm would easily make higher revenue by increasing the price.

The estimation results are reported in Table 6. With $\mathcal{S} = \{\mathbf{1}, \text{AC}, \text{HPW}, \text{MP\$}, \text{Size}\}$, the estimates of the constant term are ranged from -11.2749 to -9.3702 across different methods. The estimators of HPW and AC using OLS and $2\text{SLS}^{\mathcal{D}}$ are insignificant at 5% significant level. The other coefficient estimates are very significant regardless of the estimation methods. The signs of the coefficient estimates for HPW, AC, and MP\$ vary across the methods due to the different instruments selected. However, the coefficient estimates of Size are positive from all estimation methods. Because of the endogeneity in price, possible high collinearity and high dimensionality of instruments, estimators in OLS and $2\text{SLS}^{\mathcal{D}}$ may be inconsistent. On the other hand, $2\text{SLS}^{\mathcal{S}}$ fails to capture all strongly valid and strongly relevant instruments among the nonlinear sieve instruments. Hence, the estimators in $2\text{SLS}^{\mathcal{S}}$ may be inefficient.

When using PGMM, we re-estimate the coefficients using GMM with the selected instruments.

We refer to this method as Post-PGMM. It is just like Post-Lasso (Belloni and Chernozhukov 2013) for estimation using the variables selected by Lasso. We denote Post-PGMM as PGMM* in Tables 6 and 7. PGMM selects only the linear instruments and fails to select any nonlinear sieve instruments. Then the Post-PGMM estimators using the selected instruments only in \mathcal{S} are inefficient as in 2SLS^S. The inefficiency of 2SLS^S and Post-GMM may lead to a wrong sign in the demand function as shown with their positive coefficient estimates for Price.

In the mean time, BGMM selects too many instruments as it checks only for the relevancy, with 16 additional instruments from \mathcal{D} . And, among these 16 instruments, 6 of them are in $h_1(w_{it})$ and $h_2(w_{it}, t)$, and 10 of them are in $h_3(w_{it})$.

In comparison, we find that DB-GMM selects only one additional instrument from $h_3(w_{it}, t)$, which is the sum of the first order interaction between HPW and Size of all other cars under the same firm. The estimator of price in DB-GMM is -0.2999 , the smallest and thus suggesting the most elastic demand to price changes.

With the estimator $\hat{\beta}_{\text{price}}$, the price elasticity of demand \hat{e}_{it} for car it is

$$\hat{e}_{it} = \frac{\% \Delta s_{it}}{\% \Delta x_{it}} = \frac{\partial s_{it} / s_{it}}{\partial x_{it} / x_{it}} = \frac{x_{it}}{s_{it}} \frac{\partial s_{it}}{\partial x_{it}} = \hat{\beta}_{\text{price}} x_{it} (1 - s_{it}). \quad (42)$$

The price elasticity of demand using DB-GMM is ranged from -20.5731 to -1.0176 . The most elastic car is 1989 Nissan Maxima, and the least elastic car is 1990 Yugo GV Plus. According to the histogram in Figure 1, the price elasticity of most of the cars is between -7 and -1 .

With the demand being said to be inelastic when the price elasticity of demand is larger than -1 , i.e., $\hat{e}_{it} > -1$, we count how many cars have inelastic demand from

$$\sum_{\forall it} \mathbf{1}(\hat{e}_{it} > -1), \quad (43)$$

and the counts are reported in Table 7. Out of 2217 cars, the OLS estimates suggests that there are 1502 cars with inelastic demand. The 2SLS estimates indicates 1427 cars have inelastic demand. However, as mentioned in BLP (1995), when the number of cars with inelastic demand is large, the result does not make sense as it is inconsistent with pricing strategy that maximizes profit. In DB-GMM, none of cars has inelastic demand. With the standard error equals to 0.0328, the lower bound for the number of cars with inelastic demand is 0, and the upper bound of the number of

cars with inelastic demand is 7. Hence, the price elasticity of demand estimated by DB-GMM is more elastic than that estimated by other methods.

7 Conclusions

We propose the Double-Boosting algorithm that will consistently select strongly valid and strongly relevant instruments in a high dimensional IV regression model. We show that DB-GMM will give smaller MSE than PGMM in the simulations. In the application from BLP (1995) where instruments are generated from polynomials of the product characteristics, DB-GMM indicates that none of the cars have inelastic demand of price. Comparing to the estimation results from other methods, the price elasticity of demand estimated by DB-GMM is more elastic.

8 Appendix

A. Proof of Theorem 1

Under the approximately sparse model in (4), the conditional quadratic mean of regression error using L_2 Boosting is,

$$\begin{aligned} & \left\{ E \left[\frac{1}{n} \sum_{i=1}^n (F_{m_n,i} - E(x_i|w_i))^2 \middle| W \right] \right\}^{1/2} \\ &= \left\{ E \left[\frac{1}{n} \sum_{i=1}^n \left(F_{m_n,i} - \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} - r_i \right)^2 \middle| W \right] \right\}^{1/2} \\ &\leq \left\{ E \left[\frac{1}{n} \sum_{i=1}^n \left(F_{m_n,i} - \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} \right)^2 \middle| W \right] \right\}^{1/2} + \left\{ E \left[\frac{1}{n} \sum_{i=1}^n r_i^2 \middle| W \right] \right\}^{1/2} \end{aligned}$$

by Minkowski's inequality. By Bühlmann (2006) Theorem 1, the first term is $o_p(1)$. By Assumptions 1 and 5, the second term is

$$E \left(\frac{1}{n} \sum_{i=1}^n r_i^2 \right) \leq \sigma_2^2 \left(\frac{\log \ell_n}{n} \right) = O_p(Cn^{-\eta}) = o_p(1).$$

Hence,

$$E \left[\frac{1}{n} \sum_{i=1}^n (F_{m_n,i} - E(x_i|w_i))^2 \middle| W \right] = o_p(1).$$

□

B. Proof of Theorem 2

Lemma 1: Under Assumptions 3 and 4, $R_{\mathcal{R},j}^2 = O_p(\hat{\gamma}_j^2)$.

Proof: Denote $\hat{v}_{m,i}^* = \hat{v}_{m,i} - \bar{v}_m$, and $z_{j,i}^* = z_{j,i} - \bar{z}_j$. Then

$$\begin{aligned}
R_{\mathcal{R},j}^2 &= 1 - \frac{\sum_{i=1}^n \left(\hat{v}_{m,i}^* - \hat{\gamma}_j z_{j,i}^* \right)^2}{\sum_{i=1}^n \hat{v}_{m,i}^{*2}} \\
&= \frac{\sum_{i=1}^n \hat{v}_{m,i}^{*2} - \sum_{i=1}^n \left(\hat{v}_{m,i}^* - \hat{\gamma}_j z_{j,i}^* \right)^2}{\sum_{i=1}^n \hat{v}_{m,i}^{*2}} \\
&= \frac{\sum_{i=1}^n \left(2\hat{v}_{m,i}^* \hat{\gamma}_j z_{j,i}^* - \hat{\gamma}_j^2 z_{j,i}^{*2} \right)}{\sum_{i=1}^n \hat{v}_{m,i}^{*2}} \\
&= 2\hat{\gamma}_j \frac{\sum_{i=1}^n \hat{v}_{m,i}^* z_{j,i}^*}{\sum_{i=1}^n \hat{v}_{m,i}^{*2}} - \hat{\gamma}_j^2 \frac{\sum_{i=1}^n z_{j,i}^{*2}}{\sum_{i=1}^n \hat{v}_{m,i}^{*2}} \\
&= 2\hat{\gamma}_j^2 \left(\frac{\sum_{i=1}^n z_{j,i}^{*2}}{\sum_{i=1}^n \hat{v}_{m,i}^* z_{j,i}^*} \right) \frac{\sum_{i=1}^n \hat{v}_{m,i}^* z_{j,i}^*}{\sum_{i=1}^n \hat{v}_{m,i}^{*2}} - \hat{\gamma}_j^2 \frac{\sum_{i=1}^n z_{j,i}^{*2}}{\sum_{i=1}^n \hat{v}_{m,i}^{*2}} \\
&= \hat{\gamma}_j^2 \frac{\sum_{i=1}^n z_{j,i}^{*2}}{\sum_{i=1}^n \hat{v}_{m,i}^{*2}}.
\end{aligned}$$

Under Assumptions 3 and 4, $\frac{1}{n} \sum_{i=1}^n z_{j,i}^{*2} = O_p(1)$ and $\frac{1}{n} \sum_{i=1}^n \hat{v}_{m,i}^{*2} = O_p(1)$. Then, $\frac{\sum_{i=1}^n z_{j,i}^{*2}}{\sum_{i=1}^n \hat{v}_{m,i}^{*2}} = O_p(1)$. Hence, $R_{\mathcal{R},j}^2 = O_p(\hat{\gamma}_j^2)$. \square

Lemma 2: Under Assumption 3, $\frac{1}{n} \sum_{i=1}^n z_{j,i} \hat{u}_i \xrightarrow{p} E(z_{j,i} u_i)$.

Proof:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n z_{j,i} \hat{u}_i &= \frac{1}{n} \sum_{i=1}^n z_{j,i} \left(y_i - x_i \hat{\beta}_{2\text{SLS}} \right) \\
&= \frac{1}{n} \sum_{i=1}^n z_{j,i} \left[\left(y_i - x_i \left(x_i' z_{S,i} (z_{S,i}' z_{S,i})^{-1} z_{S,i}' x_i \right)^{-1} \left(x_i' z_{S,i} (z_{S,i}' z_{S,i})^{-1} z_{S,i}' y_i \right) \right) \right] \\
&= \frac{1}{n} \sum_{i=1}^n z_{j,i} \left[\left(y_i - x_i \beta - x_i \left(x_i' z_{S,i} (z_{S,i}' z_{S,i})^{-1} z_{S,i}' x_i \right)^{-1} \left(x_i' z_{S,i} (z_{S,i}' z_{S,i})^{-1} z_{S,i}' u_i \right) \right) \right] \\
&= \frac{1}{n} \sum_{i=1}^n z_{j,i} \left[u_i - x_i \left(x_i' z_{S,i} (z_{S,i}' z_{S,i})^{-1} z_{S,i}' x_i \right)^{-1} \left(x_i' z_{S,i} (z_{S,i}' z_{S,i})^{-1} z_{S,i}' u_i \right) \right] \\
&= \frac{1}{n} \sum_{i=1}^n z_{j,i} u_i - \frac{1}{n} \sum_{i=1}^n z_{j,i} x_i \left(x_i' z_{S,i} (z_{S,i}' z_{S,i})^{-1} z_{S,i}' x_i \right)^{-1} \left(x_i' z_{S,i} (z_{S,i}' z_{S,i})^{-1} z_{S,i}' u_i \right) \\
&= \frac{1}{n} \sum_{i=1}^n z_{j,i} u_i + o_p(1) \xrightarrow{p} E(z_{j,i} u_i). \quad \square
\end{aligned}$$

Lemma 3: Under Assumptions 1 to 5, $\frac{1}{n} \sum_{i=1}^n z_{j,i} \hat{v}_{m,i} \xrightarrow{p} E(z_{j,i} v_i)$.

Proof: First, we rewrite $\frac{1}{n} \sum_{i=1}^n z_{j,i} \hat{v}_{m,i}$ in terms of the strong learner $F_{m-1,i}$ and the error term v_i .

We obtain,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n z_{j,i} \hat{v}_{m,i} &= \frac{1}{n} \sum_{i=1}^n z_{j,i} (x_i - F_{m-1,i}) \\
&= \frac{1}{n} \sum_{i=1}^n z_{j,i} \left(x_i - \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} - F_{m-1,i} \right) \\
&= \frac{1}{n} \sum_{i=1}^n z_{j,i} \left(v_i + \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} - F_{m-1,i} \right) \\
&= \frac{1}{n} \sum_{i=1}^n z_{j,i} v_i - \frac{1}{n} \sum_{i=1}^n z_{j,i} \left(F_{m-1,i} - \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} \right)
\end{aligned}$$

By Theorem 1, $F_{m-1,i} \xrightarrow{q.m.} \sum_{j=1}^{\ell_n} \gamma_j z_{j,i}$ implies $F_{m-1,i} \xrightarrow{p} \sum_{j=1}^{\ell_n} \gamma_j z_{j,i}$. Hence

$$\frac{1}{n} \sum_{i=1}^n z_{j,i} \hat{v}_{m,i} = \frac{1}{n} \sum_{i=1}^n z_{j,i} v_i + o_p(1) \xrightarrow{p} E(z_{j,i} v_i). \quad \square$$

Proof of Theorem 2:

For validity, $\rho_j \propto \frac{b_j}{n^{\delta_j}}$. By Lemma 2,

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n z_{j,i} \hat{u}_i \right) = \begin{cases} O_p(b_j n^{\frac{1}{2}-\delta_j}) & = o_p(1) & \text{if } \delta_j > \frac{1}{2} \\ O_p(b_j n^0) & = O_p(1) & \text{if } \delta_j = \frac{1}{2} \\ O_p(b_j n^{\frac{1}{2}-\delta_j}) & = O_p(n^{\frac{1}{2}-\delta_j}) & \text{if } \delta_j < \frac{1}{2}. \end{cases}$$

Then

$$\begin{aligned}
nR_{\mathcal{V},j}^2 &= n\hat{\rho}_j^2 \\
&= \frac{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n z_{j,i} \hat{u}_i \right)^2}{\left(\frac{1}{n} \sum_{i=1}^n z_{j,i}^2 \right) \left(\frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 \right)} \\
&= \begin{cases} o_p(1) & \text{if } b_j = 0 & (\mathcal{V}_1) \\ o_p(1) & \text{if } b_j \neq 0 \text{ or } \delta_j > \frac{1}{2} & (\mathcal{V}_1) \\ O_p(1) & \text{if } b_j \neq 0 \text{ and } \delta_j = \frac{1}{2} & (\mathcal{V}_2) \\ O_p(n^{1-2\delta_j}) & \text{if } b_j \neq 0 \text{ and } 0 < \delta_j < \frac{1}{2} & (\mathcal{V}_2) \\ O_p(n) & \text{if } b_j \neq 0 \text{ and } \delta_j = 0 & (\mathcal{V}_3). \end{cases}
\end{aligned}$$

For relevancy, $\gamma_j = \frac{a_j}{n^{\alpha_j}}$, and $nR_{\mathcal{R},j}^2 = O_p\left(n\hat{\gamma}_j^2\right)$ by Lemma 1. From Lemma 3,

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n z_{j,i} \hat{v}_{m,i} \right) = \begin{cases} O_p(a_j n^{\frac{1}{2}-\alpha_j}) & = o_p(1) & \text{if } \alpha_j > \frac{1}{2} \\ O_p(a_j n^0) & = O_p(1) & \text{if } \alpha_j = \frac{1}{2} \\ O_p(a_j n^{\frac{1}{2}-\alpha_j}) & = O_p(n^{\frac{1}{2}-\alpha_j}) & \text{if } \alpha_j < \frac{1}{2}. \end{cases}$$

As $\hat{v}_{m,i}^* = \hat{v}_{m,i} - \bar{v}_m$ and $z_{j,i}^* = z_{j,i} - \bar{z}_j$, $\hat{v}_{m,i}^*$, and $z_{j,i}^*$ will have the same order as $\hat{v}_{m,i}$ and $z_{j,i}$. Then

$$\begin{aligned} nR_{\mathcal{R},j}^2 &\propto n\hat{\gamma}_j^2 \\ &= \left(\frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n z_{j,i}^* \hat{v}_{m,i}^*}{\frac{1}{n} \sum_{i=1}^n z_{j,i}^{*2}} \right)^2 \\ &= \begin{cases} o_p(1) & \text{if } a_j = 0 & (\mathcal{R}_1) \\ o_p(1) & \text{if } a_j \neq 0 \text{ and } \alpha_j > \frac{1}{2} & (\mathcal{R}_2) \\ O_p(1) & \text{if } a_j \neq 0 \text{ and } \alpha_j = \frac{1}{2} & (\mathcal{R}_2) \\ O_p(n^{1-2\alpha_j}) & \text{if } a_j \neq 0 \text{ and } 0 < \alpha_j < \frac{1}{2} & (\mathcal{R}_2) \\ O_p(n) & \text{if } a_j \neq 0 \text{ and } \alpha_j = 0 & (\mathcal{R}_3). \end{cases} \end{aligned}$$

Notice that $\mathcal{A} = \mathcal{V}_1 \cap \mathcal{R}_3$ is the set of strongly valid and strongly relevant instruments, $\mathcal{B}_0 = \mathcal{V}_1 \cap (\mathcal{R}_1 \cup \mathcal{R}_2)$ is the set of strongly valid and weakly relevant or irrelevant instruments, and $\mathcal{B}_1 = (\mathcal{V}_2 \cup \mathcal{V}_3) \cap (\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_3)$ is the set of weakly valid or invalid instruments that are not in $\mathcal{A} \cup \mathcal{B}_0$. For instrument in each of \mathcal{A} , \mathcal{B}_0 , and \mathcal{B}_1 , ω_j has the following orders in probability:

$$\begin{aligned} \omega_j &= \frac{\left(nR_{\mathcal{V},j}^2\right)^{r_2}}{\left(nR_{\mathcal{R},j}^2\right)^{r_1}} = \frac{o_p(1)}{O_p(n^{r_1})} = o_p(n^{-r_1}), & Z_j \in \mathcal{A}, \\ \omega_j &= \frac{\left(nR_{\mathcal{V},j}^2\right)^{r_2}}{\left(nR_{\mathcal{R},j}^2\right)^{r_1}} = \frac{o_p(1)}{O_p(n^{r_1(1-2\alpha_j)})} = o_p(n^{r_1(2\alpha_j-1)}), & Z_j \in \mathcal{B}_0, \\ \omega_j &= \frac{\left(nR_{\mathcal{V},j}^2\right)^{r_2}}{\left(nR_{\mathcal{R},j}^2\right)^{r_1}} = \frac{O_p(n^{r_2})}{O_p(n^{r_1})} = O_p(n^{r_2-r_1}), & Z_j \in \mathcal{B}_1. \end{aligned}$$

We summarize the above results in Table 2, which adds the orders of ω_j to Table 1. Because $\alpha_j > 0$ for $Z_j \in \mathcal{B}_0$, we have $o_p(n^{r_1(2\alpha_j-1)}) \leq \min\{o_p(n^{r_1(2\alpha_j-1)}), O_p(n^{r_2-r_1})\}$. Therefore, for any selected instrument Z_{j_m} by the DB algorithm,

$$\Pr(\omega_{j_m} < \omega_j) \rightarrow 1 \text{ for all } Z_j \in \mathcal{B}_0 \cup \mathcal{B}_1, \text{ as } n \rightarrow \infty,$$

so that $Z_j \in \mathcal{B}_0 \cup \mathcal{B}_1$ will not be selected w.p.a.1. □

References

- Bajari, P., Nekipelov, D., Ryan, S. P., and Yang, M. (2015). “Demand Estimation with Machine Learning and Model Combination”. *National Bureau of Economic Research Paper* No. 20955.
- Bekker, P. (1994). “Approximations to the Distributions of Instrumental Variable Estimators”. *Econometrica* 62(3), 657-681.
- 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. (2013). “Least Squares After Model Selection in High-dimensional Sparse Models”. *Bernoulli* 19(2), 521-547.
- Berry, S., Levinsohn, J., and Pakes, A. (1995). “Automobile Prices in Market Equilibrium”. *Econometrica* 63(4), 841 - 890.
- Bühlmann, P. (2006). “Boosting for High-dimensional Linear Models”. *Annals of Statistics* 34(2), 559-583.
- Caner, M. (2009). “Lasso-type GMM Estimator”. *Econometric Theory* 25(1), 270-290.
- Caner, M., Han, X., and Lee, Y. (2017). “Adaptive Elastic Net GMM Estimation with Many Invalid Moment Conditions: Simultaneous Model and Moment Selection”. *Journal of Business and Economic Statistics* 36(1), 24-46.
- Caner, M., and Zhang, H. H. (2014). “Adaptive Elastic Net for Generalized Methods of Moments”. *Journal of Business and Economic Statistics* 32(1), 30-47.
- 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:1608.00060.
- Chernozhukov, V., Hansen, C., and Spindler, M. (2015). “Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments”. *American Economic Review, Papers and Proceedings* 105(5), 486-490.
- Cheng, X. and Liao, Z. (2015). “Select the Valid and Relevant Moments: An Information-based LASSO for GMM with Many Moments”. *Journal of Econometrics* 186(2), 443-464.
- DiTraglia, F. (2016). “Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM”. *Journal of Econometrics* 195(2), 187-208.

- Donald, S., Imbens, G., and Newey, W. (2009). “Choosing Instrumental Variables in Conditional Moment Restriction Models”. *Journal of Econometrics* 152(1), 28-36.
- Fan, J. and Liao, Y. (2014). “Endogeneity in High Dimensions”. *Annals of Statistics* 42(3), 872-917.
- Gillen, B. J., Montero, S., Moon, H. R., and Shum, M. (2015). “BLP-Lasso for Aggregate Discrete Choice Models of Elections with Rich Demographic Covariates”. *USC-INET Research Paper* No. 15-27.
- Gillen, B. J., Moon, H. R., and Shum, M. (2014). “Demand Estimation with High-dimensional Product Characteristics”. *Advances in Econometrics* 34, 301-324.
- Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2016). “Counterfactual Prediction with Deep Instrumental Variables Networks”. arXiv:1612.09596.
- Liao, Z. (2013). “Adaptive GMM Shrinkage Estimation with Consistent Moment Selection”. *Econometric Theory* 29, 857-904.
- Meinshausen, N. (2007). “Relaxed Lasso”. *Computational Statistics & Data Analysis* 52(1), 374-393.
- Ng, S. and Bai, J. (2008). “Selecting Instrumental Variables in a Data Rich Environment”. *Journal of Time Series Econometrics* 1(1), 1-34.
- Phillips, P. (1989). “Partially Identified Econometric Models”. *Econometric Theory* 5(2), 181-240.
- Staiger, D. and Stock, J. (1997). “Instrumental Variables Regression with Weak Instruments”. *Econometrica* 65(3), 557-586.
- Stock, J.H. and Wright, J. (2000). “GMM with Weak Identification”. *Econometrica* 68(5), 1055-1096.

Table 1: Categories of instruments

		Strongly Valid \mathcal{V}_1	Weakly Valid \mathcal{V}_2	Invalid \mathcal{V}_3
Irrelevant	\mathcal{R}_1		\mathcal{B}_1	
Weakly Relevant	\mathcal{R}_2	\mathcal{B}_0		
Strongly Relevant	\mathcal{R}_3	\mathcal{S}, \mathcal{A}		

Note: The notation for each subset of instruments follows Cheng and Liao (2015, p. 446, Table 2.1). Instruments in \mathcal{S} are sure to be valid and relevant. Instruments in \mathcal{A} are valid and relevant, those in \mathcal{B}_0 are valid but redundant, and those in \mathcal{B}_1 are invalid.

Table 2: Order of ω_j for each category of instruments

		Strongly Valid \mathcal{V}_1	Weakly Valid \mathcal{V}_2	Invalid \mathcal{V}_3
Irrelevant	\mathcal{R}_1		$\mathcal{B}_1 : \omega_j = O_p(n^{r_2-r_1})$	
Weakly Relevant	\mathcal{R}_2	$\mathcal{B}_0 : \omega_j = o_p(n^{r_1(2\alpha_j-1)})$		
Strongly Relevant	\mathcal{R}_3	$\mathcal{A} : \omega_j = o_p(n^{-r_1})$		

Table 3: DGP 1

n	ℓ_n	a	OLS	2SLS \mathcal{SD}	2SLS \mathcal{S}	2SLS \mathcal{SA}	BGMM	PGMM	DB-GMM
Panel A: strong instrument with $\gamma_4 = 0.5$									
100	52	CL	0.3363	0.3604	0.0020	0.0162	0.3706	0.0068	0.0288
			0.3388	0.3632	0.1979	0.0786	0.3743	0.2980	0.1746
100	52	0.5	0.2816	0.2911	-0.0024	0.0088	0.2970	-0.0021	0.0116
			0.2841	0.2941	0.1048	0.0686	0.3013	0.1045	0.0917
100	52	0.9	0.2172	0.2079	-0.0005	0.0076	0.2020	-0.0001	0.0057
			0.2204	0.2118	0.0593	0.0535	0.2078	0.0598	0.0591
250	52	CL	0.3329	0.3736	-0.0002	0.0054	0.3777	0.0005	0.0121
			0.3339	0.3748	0.1058	0.0493	0.3795	0.1044	0.0889
250	52	0.5	0.2804	0.2968	0.0014	0.0050	0.3003	0.0016	0.0064
			0.2815	0.2983	0.0601	0.0425	0.3023	0.0603	0.0538
250	52	0.9	0.2166	0.2002	-0.0010	0.0019	0.1979	-0.0009	0.0017
			0.2179	0.2020	0.0358	0.0329	0.2005	0.0358	0.0356
Panel B: weak instrument with $\gamma_4 = 0.01$									
100	52	CL	0.4210	0.4619	0.0026	0.0290	0.4780	0.0261	0.0348
			0.4231	0.4643	0.1970	0.1110	0.4811	0.1573	0.1600
100	52	0.5	0.3846	0.4112	0.0015	0.0216	0.4256	0.0028	0.0164
			0.3868	0.4138	0.1316	0.0990	0.4292	0.1267	0.1245
100	52	0.9	0.3405	0.3428	-0.0014	0.0135	0.3478	-0.0017	0.0136
			0.3431	0.3460	0.0850	0.0799	0.3529	0.0865	0.1001
250	52	CL	0.4178	0.4890	-0.0004	0.0120	0.4952	0.0002	0.0144
			0.4186	0.4901	0.1081	0.0654	0.4967	0.1083	0.0923
250	52	0.5	0.3842	0.4310	0.0009	0.0093	0.4370	0.0010	0.0087
			0.3851	0.4322	0.0728	0.0575	0.4387	0.0729	0.0667
250	52	0.9	0.3392	0.3440	-0.0007	0.0061	0.3473	-0.0010	0.0079
			0.3402	0.3456	0.0531	0.0505	0.3496	0.0533	0.0630

Note: For each different case, the first row is the bias of $\hat{\beta}$, and the second row is the RMSE of $\hat{\beta}$. 2SLS ^{\mathcal{SD}} denotes 2SLS with all instruments. 2SLS ^{\mathcal{S}} denotes 2SLS with instruments in \mathcal{S} . 2SLS ^{\mathcal{SA}} denotes 2SLS with instruments in $\mathcal{S} \cup \mathcal{A}$, which demonstrates the oracle result. Column 3 indicates different variance-covariance matrix of Z . When $a = \text{CL}$, Σ_Z is the same as in Cheng and Liao (2015), where $\Sigma_Z = \text{diag}(\Sigma_{\mathcal{S} \cup \mathcal{A}}, \Sigma_{\mathcal{B}})$. $\Sigma_{\mathcal{S} \cup \mathcal{A}}$ is a 4×4 Toeplitz matrix that each (i, j) element equals to $0.2^{|i-j|}$, and $\Sigma_{\mathcal{B}}$ is an $(\ell_n - 4) \times (\ell_n - 4)$ identity matrix. When $a \in \{0.5, 0.9\}$, Σ_Z is an $\ell_n \times \ell_n$ Toeplitz matrix, where each (i, j) element equals to $a^{|i-j|}$.

Table 4: DGP 2

n	ℓ_n	a	OLS	2SLS SD	2SLS S	2SLS SA	BGMM	PGMM	DB-GMM
100	125	0	0.2854	0.1849	-0.1153	0.0152	0.2345	-0.0042	0.0548
			0.2987	1.7392	3.7238	0.0901	0.2586	2.7515	0.1823
100	125	0.5	0.2696	0.2897	0.1021	0.0146	0.2346	0.0740	0.0284
			0.2850	3.6019	0.7206	0.0886	0.2620	0.4368	0.1338
100	125	0.9	0.2295	0.1853	-0.0100	0.0154	0.2264	-0.0238	-0.0003
			0.2427	0.8875	0.1855	0.0753	0.2481	0.1842	0.0906
250	125	0	0.2878	0.2409	0.0651	0.0115	0.2163	-0.0279	0.0011
			0.2926	0.2471	0.5581	0.0662	0.2268	1.0748	0.1713
250	125	0.5	0.2535	0.2186	0.0110	0.0069	0.1937	0.0121	0.0192
			0.2578	0.2241	0.1624	0.0527	0.2016	0.1638	0.0798
250	125	0.9	0.2206	0.2175	0.0214	0.0070	0.2208	0.0203	0.0122
			0.2260	0.2238	0.1172	0.0517	0.2312	0.1196	0.0629

Table 5: DGP 3

n	ℓ_n	a	OLS	2SLS SD	2SLS S	2SLS SA	BGMM	PGMM	DB-GMM
100	125	0	0.1659	0.1329	0.3219	-0.0010	0.1197	0.5625	0.0070
			0.1712	0.1388	2.1693	0.0366	0.1282	3.3374	0.0925
100	125	0.5	0.1611	0.1355	0.0155	0.0054	0.1222	0.0151	0.0135
			0.1677	0.1434	0.0954	0.0387	0.1341	0.0960	0.0496
100	125	0.9	0.1372	0.1323	0.0100	0.0012	0.1396	0.0090	0.0069
			0.1429	0.1385	0.0619	0.0329	0.1496	0.0618	0.0353
250	125	0	0.1740	0.1420	0.0409	0.0050	0.1282	0.3717	0.0149
			0.1796	0.1484	0.3838	0.0426	0.1380	3.0727	0.0668
250	125	0.5	0.1536	0.1286	-0.0024	-0.0016	0.1156	-0.0001	0.0033
			0.1588	0.1345	0.1074	0.0376	0.1258	0.1029	0.0504
250	125	0.9	0.1320	0.1273	-0.0078	-0.0052	0.1327	-0.0072	-0.0012
			0.1404	0.1366	0.0681	0.0332	0.1471	0.0687	0.0398

Table 6: Estimation of the Automobile Demand						
	OLS	2SLS ^{<i>S^D</i>}	2SLS ^{<i>S</i>}	BGMM	PGMM*	DB-GMM
constant	-10.0716 (0.2576)	-10.0438 (0.2608)	-11.4900 (0.6114)	-9.7926 (0.2642)	-11.2749 (4.5630)	-9.3702 (0.3800)
HPW	-0.1243 (0.2790)	0.1161 (0.3179)	-12.3812 (0.5440)	0.8962 (0.3518)	-44.6293 (2.7079)	5.9361 (1.0240)
AC	-0.0343 (0.0710)	0.0584 (0.0880)	-4.7606 (0.2894)	0.4778 (0.1139)	3.7666 (1.6090)	2.3026 (0.3631)
MP\$	0.2650 (0.0425)	0.2484 (0.0433)	1.1134 (0.1048)	0.1730 (0.0443)	-1.5585 (0.7671)	-0.1544 (0.0833)
Size	2.3421 (0.1246)	2.3331 (0.1265)	2.8004 (0.2623)	2.2783 (0.1283)	2.6639 (2.0612)	2.1155 (0.1770)
Price	-0.0886 (0.0043)	-0.0970 (0.0063)	0.3387 (0.0171)	-0.1277 (0.0086)	0.3968 (0.1014)	-0.2999 (0.0328)

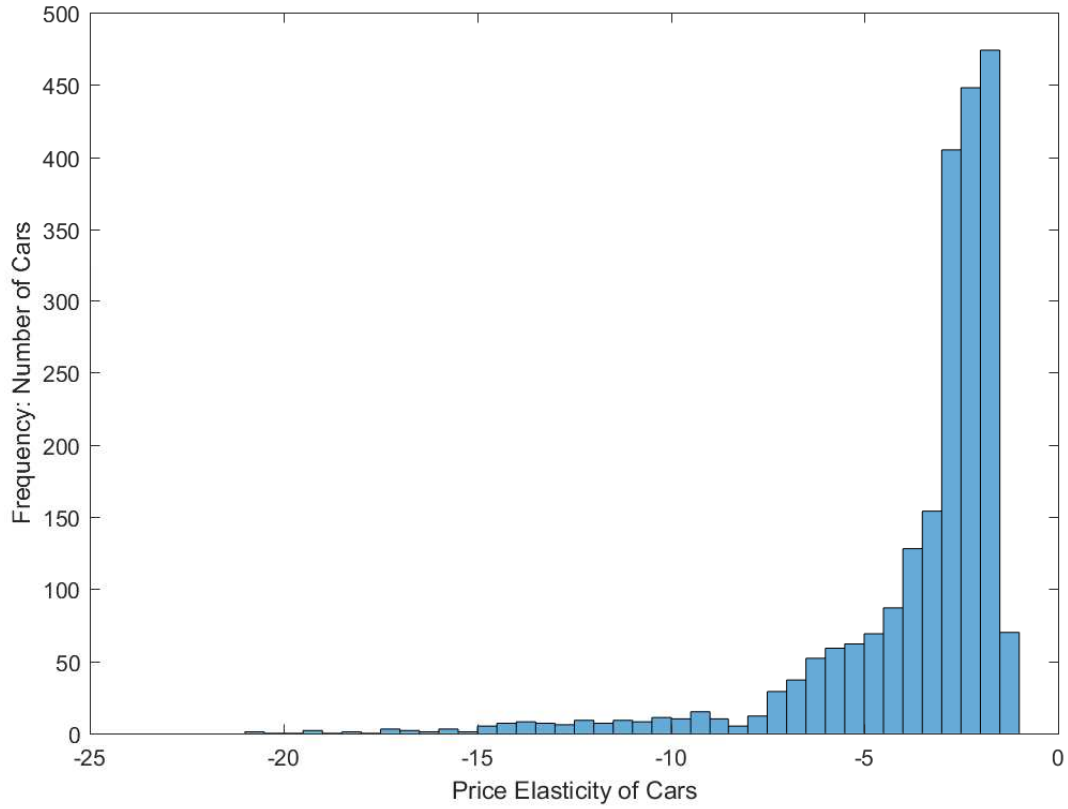
Note: PGMM* is the Post-PGMM. The values inside the parentheses are the standard error of the corresponding estimators.

Table 7: Number of Cars with Inelastic Demand					
OLS	2SLS ^{<i>S^D</i>}	2SLS ^{<i>S</i>}	BGMM	PGMM*	DB-GMM
1502	1427	2217	868	2217	0
(1425,1626)	(1230,1563)	(2217,2217)	(641,1207)	(2217, 2217)	(0, 7)

Note: The demand is said to be inelastic when the price elasticity of demand in (42) is larger than -1 . We count how many cars have inelastic demand by (43). The numbers inside the parentheses in the second row are the 95% confidence interval following the normal distribution. The upper bound and lower bound of price elasticity of demand are calculated as

$$\left(\hat{\beta}_{\text{price}} \pm 1.96\text{se}(\hat{\beta}_{\text{price}}) \right) x_{it}(1 - s_{it}).$$

Figure 1: Price Elasticity of Cars by DB-GMM



Note: The empirical distribution of the estimated price elasticity of demand using DB-GMM is shown. Demand for all cars are elastic with the estimated price elasticity of demand lower than -1 .