Boosting GMM with Many Instruments When Some Are Invalid and/or Irrelevant *

HAO HAO† and TAE-HWY LEE‡

February 7, 2025

Abstract

When the endogenous variable is an unknown function of observable instruments, its conditional mean can be approximated using the sieve functions of observable instruments. We propose a novel instrument selection method, Double-criteria Boosting (DB), that consistently selects only valid and relevant instruments from a large set of candidate instruments. In the Monte Carlo simulation, we compare GMM using DB (DB-GMM) with other estimation methods and demonstrate that DB-GMM gives lower bias and RMSE. In the empirical application to the automobile demand, the DB-GMM estimator is suggesting a more elastic estimate of the price elasticity of demand than the standard 2SLS estimator.

JEL Classification: C1, C5

Key Words: Causal inference with high dimensional instruments, Irrelevant instruments, Invalid instruments, Instrument Selection, Machine Learning, Boosting.

Word count - abstract: 101
Word count - paper (10 tables included): about 10,000

*The author would like to thank the editor Debopam Bhattacharya and two anonymous referees for many useful comments. We thank seminar participants at NeurIPS Causal Inference Workshop, UCLA (Econ), UCR (Econ), UCR (Stat), USC (Econ), Ford Motor Company, Academia Sinica (Econ), BJTU (Econ), CAS (AMSS), CUFE (Econ) and PKU (Management) for many valuable comments.

Disclaimer: This paper is a part of the authors' academic work. The views expressed herein are solely those of the authors and do not reflect the opinions of Ford Motor Company.
† *Global Data Insight & Analytics, Ford Motor Company, Michigan USA 48124 (e-mail: hxu59@ford.com)*
‡ *Department of Economics, University of California, Riverside, California USA 92521 (e-mail: taelee@ucr.edu)*

I Introduction

According to Berry, Levinsohn, and Pakes (1995, BLP henceforth), the two-stage least squares (2SLS) estimators of the logit demand function are not aligned 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 incorporating high order polynomials and interaction terms of the instrumental variable (IV) and control variables. These additional instruments and control variables help capture the neglected nonlinearity.

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

In addition, Bekker (1994) shows that the 2SLS estimator becomes inconsistent when the number of instruments is too large relative to the number of observations. Thus, the consistency of 2SLS estimators fails if instruments are in high dimension.

Another challenge with high dimensional instruments is the potential presence of weakly relevant instruments (i.e., 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 issues arise in GMM estimation as proved in Stock and Wright (2000). The asymptotic distributions of point estimators of weakly identified parameters may not be asymptotically normal.

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

While Lasso has advantage in variable selection, its estimator is biased. Belloni and

Chernozhukov (2013) propose the Post-Lasso estimation, which can reduce the bias in the estimator. 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. Chernozhukov, Hansen, and Spindler (2015) further extended Lasso and Post-Lasso to both the first and second stages of the 2SLS estimation when both instruments and control variables are in high dimension. Additionally, Gillen, Moon, and Shum (2014) and Gillen, Montero, Moon, and Shum (2019) 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) propose a moment selection procedure by using an information criterion based on the asymptotic mean square error (MSE).

Unlike traditional variable selection methods, Hartford, et al. (2017) apply machine learning techniques to the IV regression model. In particular, Ng and Bai (2009) consider L_2Boosting for instrument selection. Bühlmann (2006) proves that L_2Boosting 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_2Boosting in Bühlmann (2006) shows that both methods share very similar properties. However, as discussed in Meinshausen (2007), Lasso may perform poorly in variable selection within a high-dimensional linear model that contains many irrelevant regressors.

However, the majority of these papers assume that instruments are "valid", meaning they are not correlated with the structural error, and thus do not question the validity of instruments but only focus on the relevancy of instruments for 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. This approach 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 into 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 adaptive Lasso and show that PGMM is asymptotically oracle in selecting valid and relevant moments.

When the number of instruments exceeds the number of observations, the LASSO-like or elastic net GMM methods may fail because its weighting matrix may not be invertible. In this paper, we propose an alternative selection algorithm based on boosting, which we refer to as "Double-criteria Boosting (DB)". It is a step-wise procedure for instrument selections and not constrained by the number of instruments. We demonstrate that DB is asymptotically oracle in selecting only strongly valid and strongly relevant instruments from a set of high dimensional instruments that may be either weakly valid, invalid, weakly relevant, or irrelevant. DB is based on a ratio of two criteria, which evaluate both the validity and relevancy of each candidate instrument. We prove that DB consistently selects only strongly valid and strongly relevant instruments. More importantly, we show that DB will not select a weakly valid instrument or a weakly relevant instrument (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 using a set of sieve functions, such as polynomials of observable instruments as in Chernozhukov, Hansen, and Spindler (2015). After DB selects instruments, we compute the GMM estimator using the selected instruments. This entire estimation process is referred to as DB-GMM.

This paper is organized as follows. In Section II, we set up the structural model for the high dimensional IV regression, define validity and relevancy of instruments, and classify instruments into different categories. In Section III, we review the L_2Boosting selection procedure introduced by Ng and Bai (2009). Since the estimator is computed by GMM after instrument selection, we refer to their method as Boosting GMM (BGMM). In Section IV, we propose a new instrument selection method, DB. Section V presents Monte Carlo studies that compare DB-GMM with other methods. Section VI presents an empirical application, following the design in Berry, Levinsohn, and Pakes (1995) and Chernozhukov, Hansen, and Spindler (2015), to demonstrate the merits of using the DB-GMM. Section VII concludes the paper. All proofs are gathered in Section VIII (Appendix).

II Model

Consider an IV model as

$$y_i = \beta' x_i + u_i \tag{1}$$

$$x_i = E(x_i|w_i) + v_i. \tag{2}$$

For $i = 1, \ldots, n$, 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 $w_i = (w_{1,i} \ \ldots \ w_{p,i})'$ 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

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

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 $h(w_i) = (h_1(w_i) \ \ldots \ h_{\ell_n}(w_i))'$ such that

$$E(x_i|w_i) = \sum_{j=1}^{\ell_n} \gamma_j h_j(w_i) + r_i, \tag{3}$$

where the parameter γ_j is a $k \times 1$ vector for each $j = 1, \ldots, \ell_n$, and $r_i = (r_{1,i} \ \ldots \ r_{k,i})'$ 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

$$(z_{1,i} \ \ldots \ z_{\ell_n,i})' \equiv (h_1(w_i) \ \ldots \ h_{\ell_n}(w_i))'. \tag{4}$$

From Equations (2) and (3),

$$x_i = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + r_i + v_i. \tag{5}$$

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

$$g(z_{j,i}, \beta) = z_{j,i} u_i. \tag{6}$$

The validity of each instrument depends on the moment condition,

$$E\left(g(z_{j,i}, \beta)\right) = E\left(z_{j,i} u_i\right) = \frac{b_j}{n^{\delta_j}}. \tag{7}$$

And the relevancy of each instrument depends on the parameter,

$$\gamma_j = \frac{a_j}{n^{\alpha_j}}. \tag{8}$$

Let $Z_j = (z_{j,1} \ \ldots \ z_{j,n})'$ for $j = 1, \ldots, \ell_n$. 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: $\mathcal{V}_1 = \{j : b_j = 0\} \cup \left\{j : b_j \neq 0 \text{ and } \delta_j > \frac{1}{2}\right\}$, and $\mathcal{V}_2 = \left\{j : b_j \neq 0 \text{ and } 0 \leq \delta_j \leq \frac{1}{2}\right\}$. Z_j is strongly valid if $j \in \mathcal{V}_1$, and weakly valid or invalid if $j \in \mathcal{V}_2$.

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

We partition the set of instruments into two subsets, $\mathcal{S}$ and $\mathcal{D}$, following Cheng and Liao (2015). The "sure" set $\mathcal{S} = \{Z_1, \ldots, Z_{\ell_\mathcal{S}}\}$ includes the strongly valid and strongly relevant instruments that are initially selected, and $\ell_\mathcal{S}$ denotes the total number of instruments in $\mathcal{S}$. The "doubt" set $\mathcal{D} = \{Z_{\ell_\mathcal{S}+1}, \ldots, Z_{\ell_n}\}$ 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 $\mathcal{B}_0 \cup \mathcal{B}_1$. TABLE 1 summarizes each subset of the instruments according to Definitions 1 and 2.

III Boosting GMM (BGMM)

Ng and Bai (2009) propose a two-stage procedure for the high dimensional IV regression model, which we refer to as Boosting GMM (BGMM). In the first stage, instruments are selected through L_2Boosting. In the second stage, the parameter of interest β is estimated by GMM with the selected instruments.

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 are considered by L_2Boosting. At each step $m = 1, \ldots, \bar{M}$, where $\bar{M}$ is the maximum iteration of L_2Boosting, we first compute the residual from the difference between x_i and its fitted value from the previous steps. This is referred to as "current residual", representing the remaining unexplained factors from the previous step. To identify the next most relevant instrument, we regress the "current residual" on each instrument in $\mathcal{D}$ and select the instrument that minimizes the loss function. With the selected instrument, we obtain an estimate of the "current residual". We refer this estimate as a weak learner because it captures only a small portion of the overall picture and may not be highly accurate on its own. When aggregating all weak learners up to the current step, this forms a strong learner, which provides better accuracy. 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 leaner for $i = 1, \ldots, n$. The relationship between the weak learner and the strong learner is

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \tag{9}$$

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_2Boosting for each variable in x_i.

L_2Boosting algorithm

The detail description of L_2Boosting is listed in Algorithm 1.

Algorithm 1 BGMM

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

$$F_{0,i} = f_{0,i} = \hat{\gamma}_{0,\text{initial}} + \sum_{j=1}^{\ell_{\mathcal{S}}} \hat{\gamma}_{j,\text{initial}} z_{j,i}, \tag{10}$$

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

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

   (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 $j = \ell_{\mathcal{S}} + 1, \ldots, \ell_n$. The estimators $\hat{\gamma}_0$ and $\hat{\gamma}_j$ are solved as

$$\{\hat{\gamma}_{0,j}, \hat{\gamma}_j\} = \min_{\gamma_0, \gamma_j} \sum_{i=1}^{n} \left(\hat{v}_{m,i} - \gamma_0 - \gamma_j z_{j,i}\right)^2. \tag{11}$$

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

$$j_m = \arg \min_{j \in \{\ell_{\mathcal{S}}+1, \ldots, \ell_n\}} \sum_{i=1}^{n} \left(\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i}\right)^2. \tag{12}$$

   (c) The weak learner is

$$f_{m,i} = \hat{\gamma}_{0,j_m} + \hat{\gamma}_{j_m} z_{j_m,i}, \tag{13}$$

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

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

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \tag{14}$$

   with learning rate $c_m > 0$.

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

L_2Boosting controls over-fitting in two ways. First, it applies the learning rate c_m to the weak learner $f_{m,i}$ at each step. The learning rate controls the influence of each weak learner when adding up to the strong learner. A smaller learning rate implies greater regularization in the L_2Boosting, which reduces the impact of weak learners and requires a larger number of iterations. Second, an early stopping rule is used to determine the number of steps required in L_2Boosting before over-fitting. The stopping rule we used in this paper is 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 (13),

$$\begin{aligned} \mathbf{1}\hat{\gamma}_{0,j_m} + Z_{j_m}\hat{\gamma}_{j_m} &= P_m\hat{V}_m \\ f_m &= P_m\left(X - F_{m-1}\right). \end{aligned} \tag{15}$$

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

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

AIC is computed as

$$AIC_c(m) = \log(\hat{\sigma}^2_{2,m}) + \frac{1 + \text{trace}(B_m)/n}{1 - (\text{trace}(B_m) + 2)/n}, \tag{16}$$

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

Consistency of L_2Boosting

Consider the following assumptions from Bühlmann (2006).

Assumption 1: *The dimension of instruments satisfies* $\ell_n = O(\exp(Cn^{1-\eta}))$, $n \to \infty$, *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. 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_2Boosting 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 (5)

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

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 sieve instruments $\{z_{j,i}\}$, the sparse approximation error r_i satisfies that $E\left(r_i^2|w_i\right) \leq \sigma_2^2 \left(\frac{\log \ell_n}{n}\right)$, 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_2Boosting 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 sieve instruments* $\{z_{j,i}\}$. *Under Assumptions 1-5, for some sequence* $(m_n)_{n\in\mathbb{N}}$ *with* $m_n \to \infty$ *sufficiently slowly as* $n \to \infty$, *the* L_2*Boosting estimation converges to the conditional mean of* x_i,

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

where $W = (w_1 \ \dots \ w_n)'$ *is an* $n \times p$ *matrix and* $w_i = (w_{1,i} \ \dots \ w_{p,i})'$.

Proof: Appendix A.

However, L_2Boosting 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_2Boosting 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.

IV Double-criteria 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 DB-GMM estimator.

Double-criteria Boosting algorithm

The DB algorithm is described in Algorithm 2. The new selection algorithm (Algorithm 2) is similar to L_2Boosting (Algorithm 1) in the previous section, except Step 2(b), where the new objective function Equation (23) is replacing Equation (12) in Algorithm 1. We now doubly minimize the invalidity (measured by Equation (20)) and minimize the irrelevancy (measured by the inverse of Equation (24)) 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

$$\rho_j = \frac{E\left(z_{j,i}u_i\right)}{\sqrt{E\left(z_{j,i}^2\right)}\sqrt{E\left(u_i^2\right)}} = \frac{d_j}{n^{\delta_j}}. \tag{17}$$

where $d_i = \frac{b_j}{\sqrt{E\left(z_{j,i}^2\right)}\sqrt{E\left(u_i^2\right)}}$ and b_j defined in Equation (7).

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,

$$\hat{u}_i \equiv y_i - \hat{\beta}_{\text{initial}} x_i, \tag{18}$$

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

$$\hat{\rho}_j = \frac{\frac{1}{n}\sum_{i=1}^n z_{j,i}\hat{u}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^n z_{j,i}^2}\sqrt{\frac{1}{n}\sum_{i=1}^n \hat{u}_i^2}}. \tag{19}$$

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

$$nR_{\mathcal{V},j}^2 = n\hat{\rho}_j^2. \tag{20}$$

Similarly, we also define the LM statistic measure for relevancy of z_j, $nR_{\mathcal{R},j}^2$, which we describe in Equation (24) 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

$$F_{0,i} = f_{0,i} = \hat{\gamma}_{0,\text{initial}} + \sum_{j=1}^{\ell_{\mathcal{S}}} z_{j,i}\hat{\gamma}_{j,\text{initial}}, \tag{21}$$

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 $j \in \{\ell_{\mathcal{S}} + 1, \dots, \ell_n\}$. The estimators $\hat{\gamma}_{0,j}$ and $\hat{\gamma}_j$ are solved as

$$\{\hat{\gamma}_{0,j}, \ \hat{\gamma}_j\} = \min_{\gamma_0, \gamma_j} \sum_{i=1}^{n} \left(\hat{v}_{m,i} - \gamma_0 - \gamma_j z_{j,i}\right)^2. \tag{22}$$

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

$$j_m = \arg \min_{j \in \{\ell_{\mathcal{S}}+1, \dots, \ell_n\}} \omega_j \equiv \frac{\left(nR^2_{\mathcal{V},j}\right)^{r_2}}{\left(nR^2_{\mathcal{R},j}\right)^{r_1}}, \tag{23}$$

   where

$$R^2_{\mathcal{R},j} = 1 - \frac{\sum_{i=1}^{n} \left(\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i}\right)^2}{\sum_{i=1}^{n} \left(\hat{v}_{m,i} - \bar{v}_m\right)^2}, \tag{24}$$

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

   (c) The weak learner is

$$f_{m,i} = \hat{\gamma}_{0,j_m} + \hat{\gamma}_{j_m} z_{j_m,i}, \tag{25}$$

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

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

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \tag{26}$$

   with $c_m > 0$.

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

Remark 1: We introduce the selection criterion ω_j to evaluate the validity and relevancy of each instrument Z_j. The user selected constants r_1 and r_2 control the penalties associated with the validity and relevancy. A higher value of r_2 results in a greater penalty for the invalid instrument, as it increases the numerator in ω_j. On the other hand, when the instrument has high relevancy score, a higher value of r_1 leads to a larger denominator in ω_j and an overall smaller ω_j. In simulations and application of this paper, we report 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) 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_2Boosting. (iii) If r_2 is too small compare to r_1, DB may select weakly valid instrument in finite samples. But asymptotically, the probability of selecting weakly valid instrument will go to 0.

Remark 2: Our selection criterion ω_j serves the same purpose as the information based adjustment in PGMM, as proposed by Cheng and Liao (2015). However, in PGMM, the selection criterion for each instrument Z_j is calculated independently of other instruments in $\mathcal{D}$. As a result, the PGMM selection criterion does not consider the selection of other instruments. In contrast, Double-criteria Boosting updates the current residual $\hat{v}_{m,i}$ at each DB iteration. Then the relevancy criterion $nR^2_{\mathcal{R},j}$ not only depends on $\mathcal{S}$ but also on all previously selected instruments. In both methods, the instruments selected from $\mathcal{D}$ are not necessarily weaker than those in $\mathcal{S}$. But instruments in $\mathcal{S}$ must be valid and relevant before entering the selection process.

Remark 3: When the validity and relevancy of the instruments are uncertain, DB-GMM is capable of starting the selection process with an empty "sure"set, $\mathcal{S}$. We demonstrate this property through a Monte Carlo simulation reported in Section V (TABLE 9). We thank a referee who asked us to consider this case. The results show that DB-GMM performs well even when $\mathcal{S}$ is empty. In contrast, PGMM requires $\mathcal{S}$ to be a non-empty set.

Remark 4: As DB follows a forward selection procedure, the statistic $R^2_{\mathcal{R},j}$, which is the R^2 of the regression at each step, reflects the relevancy strength of each instruments to the endogenous variable. It is essentially related to the F-statistics or concentration parameters. As defined in Equation (23), minimizing $\frac{1}{nR^2_{\mathcal{R},j}}$ is equivalent to maximizing $nR^2_{\mathcal{R},j}$. Therefore, instrument that explain the most variation will have higher probability of being selected by DB.

Remark 5: The stopping rule in DB is the same as in L_2Boosting. As $R^2_{\mathcal{V},j}$ is computed based on the 2SLS estimation using only instruments in $\mathcal{S}$, $nR^2_{\mathcal{V},j}$ is fixed at any iteration $m = 1, \ldots, \bar{M}$. According to the definition, the maximization of $R^2_{\mathcal{R},j}$ 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 in Equation (12) for L_2Boosting. 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 1 asymptotically. In other words, DB ensures 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 $n \to \infty$. That is,*

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

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

Proof: Appendix B.

Compare to other methods that minimize estimation risk, DB selection is better for inference as it only selects strongly valid and strongly relevant instruments.

V 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):

$$y_i = \beta x_i + u_i,$$
$$x_i = \sum_{j=1}^{p} \gamma_j w_{j,i} + v_i = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + v_i, \tag{27}$$

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 $z_{\mathcal{A},i} = (z_{3,i}\ z_{4,i})'$, $z_{\mathcal{B}_0,i} = (z_{5,i}\ \ldots\ z_{28,i})'$ and $z_{\mathcal{B}_1,i} = (z_{29,i}\ \ldots\ z_{52,i})'$ 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 $z_{\mathcal{B}_1,i}$, we first need to generate a strongly valid instrument $z^*_{\mathcal{B}_1,i}$. The strongly valid instruments and error terms follow the normal distribution where

$$(z_{S,i}\ z_{A,i}\ z_{\mathcal{B}_0,i}\ z^*_{\mathcal{B}_1,i}) \sim N(0, \Sigma_Z)$$
$$(u_i\ v_i) \sim N(0, \Sigma),$$

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 $\Sigma_Z = \text{diag}(\Sigma_{\mathcal{S}\cup\mathcal{A}}, \Sigma_{\mathcal{B}})$. $\Sigma_{\mathcal{S}\cup\mathcal{A}}$ 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. 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 $j = 29, \ldots, 52$, the invalid instrument $z_{j,i}$ is generated as

$$z_{j,i} = z^*_{j,i} + c_j u_i, \tag{28}$$

where $z^*_{j,i}$ is the strongly valid instrument in $z^*_{\mathcal{B}_1,i}$, and

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

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):

$$\begin{aligned} y_i &= \beta x_i + u_i \\ x_i &= \sum_{j=1}^{p} \theta_j \left(w_{j,i} + w_{j,i}^2\right) + v_i, \end{aligned} \tag{30}$$

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

$$(w_{1,i}\ w_{2,i}\ w_{3,i}\ w_{4,i}\ w^*_{5,i}) \sim N(0, \Sigma_W), \tag{31}$$

where Σ_W is a 5×5 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

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

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 Equation (31), by adding the structural error u_i

$$w_{5,i} = w^*_{5,i} + u_i. \tag{32}$$

Suppose 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 $j = 1, \ldots, 5$, and $z_{j,i} = h_j(w_i)$ for $j = 6, \ldots, \ell_n$, where $h_j(w_i)$ is the polynomial of each instrument in w_i up to the 4^{th} order. This leads to a total of 125 instruments: 5 from the observable instruments w_i, 15 from the 2^{nd} order, 35 from the 3^{rd} order, and 70 from the 4^{th} order. The pre-selected "sure " set is $z_{\mathcal{S},i} = (z_{1,i}\ z_{2,i})'$.

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,

$$x_i = \sum_{j=1}^{p} \theta_j \exp(w_{j,i}) + v_i. \tag{33}$$

In DGP 1, as $z_{j,i} = w_{j,i}$, all instruments are observable and the functional form of $h_j(\cdot)$ is known. The functional form of x_i in DGP 3 is unknown as in DGP 2, and we use sieve instruments to approximate as in DGP 2. The total number of instruments in DGP 3 is also 125, and we set $z_{\mathcal{S},i} = (z_{1,i}\ z_{2,i})'$.

Simulation results

In TABLEs 3 to 5, we compare the bias and root mean squared errors (RMSE) of DB-GMM with 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. The user selected parameters in PGMM are the same as in Cheng and Liao (2015). We choose the learning rate for both boosting algorithms to be $c_m = 0.01$ for all m.

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 higher because x_i is endogenous. As the correlation between instruments becomes stronger, the OLS estimation has slight improvement in its bias and RMSE. The $2SLS^{\mathcal{SD}}$ estimation is also inconsistent because of the existence of invalid and irrelevant instruments. $2SLS^{\mathcal{S}}$ has lower bias but higher RMSE compared to $2SLS^{\mathcal{SA}}$ because 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 due to existence of 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, $2SLS^{\mathcal{SD}}$, 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$^{\mathcal{SA}}$). 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. When $a = 0.9$, the results of 2SLS$^{\mathcal{S}}$, PGMM, and DB-GMM are very close to the oracle result. 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 is inconsistent due to the endogeneity. When $\ell_n > n$, the RMSE of 2SLS$^{\mathcal{SD}}$ diverges, which confirms the theoretical result in Bekker (1994) . If only instruments in $\mathcal{S}$ are selected, the bias and RMSE of 2SLS$^{\mathcal{S}}$ remain high because 2SLS$^{\mathcal{S}}$ 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, which will cause the RMSE of PGMM to be 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.

To control the over-fitting problem, a smaller learning rate in boosting leads to more regularization. In TABLE 6, we compare the estimation results of BGMM and DB-GMM with a set of learning rates, $c_m \in (0.01, 0.05, 0.1, 1)'$. In the cases of $\ell_n = 250$, the DB-GMM estimations with $c_m = 0.01$ are always less bias comparing to other learning rates. TABLE

7 reports the average number of steps iterated and average number of instruments selected in BGMM and DB-GMM across different learning rates. Even though the number of steps is the highest in most cases when $c_m = 0.01$, the number of instruments selected are the smallest compared to other learning rates. These confirms our selection of $c_m = 0.01$ in estimation of the above simulations.

We also conducted two additional experiments on DGP 2, with results presented in TABLEs 8 and 9. Panel A in TABLE 8 reports the 90% Monte Carlo confidence interval of the estimators. The results confirm that DB-GMM performs well for inference, delivering the smallest RMSE and narrowest confidence intervals. Panel B reports the empirical coverage probability that the true value of zero is included in the 90% asymptotic confidence intervals, which is estimated by the point estimates $\hat{\beta} \pm 1.645\times$ the asymptotic standard error.

In view of Remark 3 above, TABLE 9 reports the results when no instruments are pre-selected in the $\mathcal{S}$ (i.e., $\mathcal{S}$ is an empty set). Still, DB-GMM continues to outperform other methods. TABLE 9 does not include the results of PGMM, as $\mathcal{S}$ cannot be empty in PGMM.

VI Empirical Application

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

$$\xi_{it} = \varphi(w_{it}, x_{it}, u_{it}, \beta) + \varepsilon_{it}, \tag{34}$$

where $\varphi(w_{it}, x_{it}, u_{it}, \beta) = \varphi_{it}$ 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 $\varphi(\cdot)$. Applying the simple logit model, the market share s_{it} for each car it is calculated as

$$s_{it} = \frac{\exp(\varphi_{it})}{1 + \sum_{\forall it} \exp(\varphi_{it})}. \tag{35}$$

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

$$y_{it} = \beta_0 + \beta_{\text{price}} x_{it} + \beta_w' w_{it} + u_{it}, \tag{36}$$

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 Equation (3), we assume price is a linear combination of product characteristics and sieve functions of product characteristics such that

$$x_{it} = \gamma_0 + \gamma_w' w_{it} + \gamma_1' h_1(w_{it}) + \gamma_2' h_2(w_{it}, t) + \gamma_3' h_3(w_{it}) + v_{it}, \tag{37}$$

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 Equations (36) and (37) are estimated as if the data is cross-sectional (no time series) for $it = 1, \ldots, 2217$.

We use the data set in Chernozhukov, Hansen, and Spindler (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. As demonstrated by the simulation results in Table 9, DB-GMM can perform both selection and estimation without any pre-selected instrument in $\mathcal{S}$. Given this property, we set $\mathcal{S}$ as an empty set. After excluding the constant, all 62 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.

We compare the estimation results of 4 different methods in TABLE 10. Due to the characteristics of empirical date, we assign lower penalty to invalidly than irrelevancy, with $r_1 = 1.5$ and $r_2 = 1$. Because no instrument is pre-selected in $\mathcal{S}$, results for $2\text{SLS}^{\mathcal{S}}$ and PGMM are not reported.

In TABLE 10, we find some estimators are insignificant at 5% significant level across different methods. These include HPW in OLS, $2\text{SLS}^{\mathcal{SD}}$ and BGMM, and AC in OLS and $2\text{SLS}^{\mathcal{SD}}$. The other estimators are very significant regardless of the estimation methods. The estimators of the Price coefficient are negative for all methods and ranges from -0.0688 to -0.3790. The signs of the estimators of HPW, AC, MP\$, and Size vary across the methods due to the instruments selection.

Because of the endogeneity in Price, possible high collinearity and high dimensionality of instruments, estimators in OLS and $2\text{SLS}^{\mathcal{SD}}$ may be inconsistent. BGMM selects 15 instruments in total. Among these 15 instruments, MP\$ is selected from w_{it}, 7 of them are from $h_1(w_{it})$ and $h_2(w_{it}, t)$, and 7 of them are from $h_3(w_{it})$. As BGMM only checks relevancy, it selects too many instruments, where some of the instruments may be invalid.

In comparison, we find that DB-GMM selects 3 instruments, 2 are from $h_1(w_{it})$ and $h_2(w_{it}, t)$, and one is from $h_3(w_{it}, t)$. Compared with other methods, the estimator of Price in DB-GMM is the smallest and thus suggesting the most elastic demand to price changes.

VII Conclusions

We propose the Double-criteria Boosting algorithm that consistently selects strongly valid and strongly relevant instruments in a high dimensional IV regression model. We theoretically prove that DB will not select a weakly valid instrument nor a weakly irrelevant instrument. The simulation results illustrate that DB-GMM estimation has smaller RMSE compared to other methods even in the extreme case where no "sure" instrument is pre-selected. In addition to the simulation results presented in the paper, we also explore different learning rates, such as 0.01, and find similar conclusions. Moreover, in the application based on BLP (1995), where instruments are generated from polynomials of the product characteristics, the DB-GMM result suggests that the price elasticity of demand estimated by DB-GMM is more elastic than the other methods.

VIII Appendix

This appendix includes the proofs on Theorem 1 and Theorem 2.

A. Proof of Theorem 1

Under the approximately sparse model in Equation (3), the conditional quadratic mean of regression error using L_2Boosting is,

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

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

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

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

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

Hence,

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

□

B. Proof of Theorem 2

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

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

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

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

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:

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

□

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,

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

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

$$\frac{1}{n}\sum_{i=1}^{n} z_{j,i}\hat{v}_{m,i} = \frac{1}{n}\sum_{i=1}^{n} z_{j,i}v_i + o_p(1) \overset{p}{\to} E\left(z_{j,i}v_i\right). \qquad \square$$

Proof of Theorem 2:

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

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} z_{j,i}\hat{u}_i\right) = O_p(b_j n^{\frac{1}{2}-\delta_j}) = b_j O_p(n^{\frac{1}{2}-\delta_j}).$$

Then

$$\begin{aligned} nR^2_{\mathcal{V},j} &= n\hat{\rho}_j^2\\ &= \frac{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_{j,i}\hat{u}_i\right)^2}{\left(\frac{1}{n}\sum_{i=1}^{n} z_{j,i}^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2\right)}\\ &= b_j^2 O_p\left(n^{1-2\delta_j}\right).\end{aligned}$$

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

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} z_{j,i}\hat{v}_{m,i}\right) = O_p(a_j n^{\frac{1}{2}-\alpha_j}) = a_j O_p(n^{\frac{1}{2}-\alpha_j}).$$

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

$$\begin{aligned} nR^2_{\mathcal{R},j} &\propto n\hat{\gamma}_j^2 \\ &= \left(\frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n z^*_{j,i} \hat{v}^*_{m,i}}{\frac{1}{n} \sum_{i=1}^n z^{*2}_{j,i}} \right)^2 \\ &= a_j^2 O_p\left(n^{1-2\alpha_j}\right). \end{aligned}$$

Notice that $\mathcal{A} = \mathcal{V}_1 \cap \mathcal{R}_2$ is the set of strongly valid and strongly relevant instruments, $\mathcal{B}_0 = \mathcal{V}_1 \cap \mathcal{R}_1$ is the set of strongly valid and weakly relevant or irrelevant instruments, and $\mathcal{B}_1 = \mathcal{V}_2$ is the set of weakly valid or invalid instruments. For simplicity, consider the cases where both a_j and b_j are not equal to 0. Then the orders of $nR^2_{\mathcal{V},j}$ and $nR^2_{\mathcal{R},j}$ depend on δ_j and α_j respectively. For instrument in each of $\mathcal{A}$, $\mathcal{B}_0$, and $\mathcal{B}_1$, ω_j has the following orders in probability:

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

We summarize the above results in TABLE 2, which adds the orders of ω_j to TABLE 1. By definition, $\alpha_j > 0$ for irrelevant and weakly relevant instruments. Hence, when comparing instruments between $\mathcal{A}$ and $\mathcal{B}_0$, $o_p(n^{-r_1}) < o_p(n^{-r_1(1-2\alpha_j)})$ for any instruments $Z_{j_m} \in \mathcal{B}_0$. Similarly, $0 \leq \delta_j \leq \frac{1}{2}$ for invalid and weakly valid instrument. Then $o_p(n^{-r_1}) < O_p\left(n^{r_2(1-2\delta_j)-r_1(1-2\alpha_j)}\right)$ for $Z_{j_m} \in \mathcal{B}_1$. Therefore, for any selected instrument Z_{j_m} by the DB algorithm,

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

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

References

Bekker, P. (1994). 'Approximations to the Distributions of Instrumental Variable Estimators', *Econometrica* Vol. 62, pp. 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* Vol. 80, pp. 2369-2429.

Belloni, A., Chernozhukov, V. (2013). 'Least Squares After Model Selection in High-dimensional Sparse Models', *Bernoulli* Vol. 19, pp. 521-547.

Berry, S., Levinsohn, J., and Pakes, A. (1995). 'Automobile Prices in Market Equilibrium', *Econometrica* Vol. 63, pp. 841 - 890.

Bühlmann, P. (2006). 'Boosting for High-dimensional Linear Models', *Annals of Statistics* Vol. 34, pp. 559-583.

Caner, M. (2009). 'Lasso-type GMM Estimator', *Econometric Theory* Vol. 25, pp. 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* Vol. 36, pp. 24-46.

Caner, M., and Zhang, H. H. (2014). 'Adaptive Elastic Net for Generalized Methods of Moments', *Journal of Business and Economic Statistics* Vol. 32, pp. 30-47.

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* Vol. 105, pp. 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* Vol.186, pp. 443-464.

DiTraglia, F. (2016). 'Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM', *Journal of Econometrics* Vol. 195, pp. 187-208.

Donald., S., Imbens, G., and Newey, W. (2009). 'Choosing Instrumental Variables in Conditional Moment Restriction Models', *Journal of Econometrics* Vol. 152, pp. 28-36.

Fan, J. and Liao, Y. (2014). 'Endogeneity in High Dimensions', *Annals of Statistics* Vol. 42, pp. 872-917.

Gillen, B. J., Montero, S., Moon, H. R., and Shum, M. (2019). 'BLP-Lasso for Aggregate Discrete Choice Models of Elections with Rich Demographic Covariates', *Econometrics Journal* Vol.22, pp. 262-281.

Gillen, B. J., Moon, H. R., and Shum, M. (2014). 'Demand Estimation with High-dimensional Product Characteristics', *Advances in Econometrics*Vol. 34, pp. 301-324.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). 'Counterfactual Prediction with Deep Instrumental Variables Networks', *ICML 2017*.

Liao, Z. (2013). 'Adaptive GMM Shrinkage Estimation with Consistent Moment Selection', *Econometric Theory* Vol. 29, pp. 857-904.

Meinshausen, N. (2007). 'Relaxed Lasso', *Computational Statistics & Data Analysis* Vol. 52, pp. 374-393.

Ng, S. and Bai, J. (2009). 'Selecting Instrumental Variables in a Data Rich Environment', *Journal of Time Series Econometrics* Vol. 1, pp. 1-34.

Phillips, P. (1989). 'Partially Identified Econometric Models', *Econometric Theory* Vol. 5, pp. 181-240.

Staiger, D. and Stock, J (1997). 'Instrumental Variables Regression with Weak Instruments', *Econometrica* Vol. 65, pp. 557-586.

Stock, J.H. and Wright, J. (2000). 'GMM with Weak Identification', *Econometrica* Vol. 68, pp. 1055-1096.

TABLE 1:
Categories of instruments

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

Notes: 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 strongly valid but not strongly relevant, and those in $\mathcal{B}_1$ are not strongly valid.

TABLE 2:
Order of ω_j for each category of instruments

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

Notes: The notations are the same as TABLE 1

TABLE 3:
DGP 1 - Estimates and RMSEs

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

Notes: 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 $a =$ 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 - Estimates and RMSEs

n	ℓ_n	a	*OLS*	*2SLS*$^{\mathcal{SD}}$	*2SLS*$^{\mathcal{S}}$	*2SLS*$^{\mathcal{SA}}$	*BGMM*	*PGMM*	*DB-GMM*
100	125	0	0.3103	4.9391	0.2218	0.0181	0.2363	0.1975	0.0216
			0.3205	46.5148	0.8011	0.0930	0.2621	0.6628	0.1848
100	125	0.5	0.2707	0.2649	-0.0010	0.0132	0.2011	0.0228	0.0196
			0.2825	0.7593	0.4671	0.0775	0.2286	0.4410	0.1364
100	125	0.9	0.2196	-0.5400	-0.0176	0.0009	0.2233	-0.0264	0.0096
			0.2327	6.9021	0.3236	0.0731	0.2439	0.4009	0.1004
250	125	0	0.2792	0.2329	0.0771	0.0036	0.1993	0.1267	0.0043
			0.2843	0.2398	0.6743	0.0581	0.2130	1.0879	0.1588
250	125	0.5	0.2554	0.2218	0.0130	0.0069	0.1843	0.0120	0.0039
			0.2609	0.2281	0.1432	0.0474	0.1971	0.1415	0.0653
250	125	0.9	0.2158	0.2121	-0.0102	0.0013	0.2213	-0.0106	-0.0024
			0.2207	0.2177	0.1032	0.0480	0.2313	0.1023	0.0658

Notes: See TABLE 3.

TABLE 5:
DGP 3 - Estimates and RMSEs

n	ℓ_n	a	*OLS*	*2SLS*$^{\mathcal{SD}}$	*2SLS*$^{\mathcal{S}}$	*2SLS*$^{\mathcal{SA}}$	*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

Notes: See TABLE 3.

TABLE 6:
DGP 2 - Estimation with Different Learning Rate

			BGMM				*DB-GMM*			
n	ℓ_n	a	0.01	0.05	0.1	1	0.01	0.05	0.1	1
100	125	0	0.2363	0.2581	0.2573	0.2558	0.0216	0.0245	0.0187	0.0139
			0.2621	0.2776	0.2774	0.2770	0.1848	0.1526	0.1456	0.1434
100	125	0.5	0.2011	0.2214	0.2213	0.2181	0.0196	0.0179	0.0164	0.0149
			0.2286	0.2424	0.2426	0.2417	0.1364	0.1067	0.1067	0.1225
100	125	0.9	0.2233	0.2204	0.2208	0.2189	0.0096	0.0279	0.0275	0.0189
			0.2439	0.2399	0.2405	0.2398	0.1004	0.1143	0.1137	0.1046
250	125	0	0.1993	0.2079	0.2073	0.2045	0.0043	0.0106	0.0111	0.0080
			0.2130	0.2181	0.2176	0.2163	0.1588	0.0893	0.0892	0.0914
250	125	0.5	0.1843	0.2017	0.2008	0.1969	0.0039	0.0060	0.0062	0.0055
			0.1971	0.2111	0.2101	0.2065	0.0653	0.0605	0.0605	0.0620
250	125	0.9	0.2213	0.2141	0.2128	0.2120	-0.0024	0.0123	0.0133	0.0073
			0.2313	0.2229	0.2218	0.2212	0.0658	0.0783	0.0794	0.0671

Notes: For each different case, the first row is the bias of $\hat{\beta}$, and the second row is the RMSE of $\hat{\beta}$. Each column represents the result of different learning rate (c_m)

TABLE 7:
DGP 2 - Instrument Selection Count

			BGMM				*DB-GMM*			
n	ℓ_n	a	0.01	0.05	0.1	1	0.01	0.05	0.1	1
100	125	0	492.00	461.98	229.19	20.96	414.21	435.82	217.48	19.57
			9.39	28.46	27.92	18.79	5.12	11.38	11.49	8.78
100	125	0.5	492.00	473.93	236.24	22.59	448.15	464.31	232.11	22.19
			10.03	25.57	25.13	18.60	4.56	9.50	9.48	8.22
100	125	0.9	492.00	465.74	234.74	23.79	482.23	475.75	235.53	23.34
			8.45	16.26	16.08	14.20	3.81	6.39	6.39	5.90
250	125	0	492.00	488.04	243.12	22.03	459.63	485.69	241.68	23.42
			7.30	33.10	32.80	20.60	5.61	12.95	13.12	10.76
250	125	0.5	492.00	491.45	245.35	24.28	490.38	488.23	243.59	23.90
			9.22	27.02	26.78	20.34	4.48	9.43	9.60	8.86
250	125	0.9	492.00	489.14	244.46	24.92	491.46	491.36	245.41	24.39
			8.24	17.03	16.77	14.49	3.34	5.90	5.96	5.85

Notes: For each different case, the first row is the number of steps in boosting algorithm. The second row is the number of instruments selected by boosting algorithms from $\mathcal{D}$. Each column represents the result of different learning rate (c_m)

TABLE 8:
DGP 2 - Coverage Performance

n	ℓ_n	a		*OLS*	*2SLS*$^{\mathcal{SD}}$	*2SLS*$^{\mathcal{S}}$	*2SLS*$^{\mathcal{SA}}$	*BGMM*	*PGMM*	*DB-GMM*
				Panel A: 90% Monte Carlo Confidence Intervals						
250	125	0	5%-Quantile	0.1971	0.1483	-0.9596	-0.0870	0.1001	-0.6770	-0.1560
			95%-Quantile	0.3713	0.3279	0.9220	0.0949	0.3160	0.8526	0.1528
250	125	0.5	5%-Quantile	0.1751	0.1385	-0.2747	-0.0818	0.0947	-0.2720	-0.1233
			95%-Quantile	0.3499	0.3229	0.2779	0.0933	0.3136	0.2786	0.1303
250	125	0.9	5%-Quantile	0.1385	0.1305	-0.1776	-0.0695	0.1122	-0.1753	-0.0954
			95%-Quantile	0.3066	0.3078	0.1706	0.0807	0.3400	0.1724	0.1018
			Panel B: Coverage Probability of the 90% Asymptotic Confidence Intervals							
250	125	0	Coverage	.000	.025	.930	.889	.043	.273	.901
250	125	0.5	Coverage	.001	.029	.910	.882	.031	.652	.860
250	125	0.9	Coverage	.004	.026	.918	.869	.009	.861	.856

Notes: We re-ran DGP 2 in TABLE 4 with 1,000 Monte Carlo simulations. Panel A reports the 90% Monte Carlo confidence intervals. They are the 5% and 95% quantiles of the Monte Carlo distributions of the estimators. Panel B reports the empirical coverage probability (counts / 1000) that the true value of zero under the null hypothesis is included in the 90% asymptotic confidence intervals. The asymptotic confidence interval is the point estimate $\pm 1.645 \times$SE, where SE is the estimated asymptotic standard error of each estimator.

TABLE 9:
DGP 2 - Estimates When $\mathcal{S}$ Is Empty

n	ℓ_n	a		*OLS*	*2SLS*$^{\mathcal{SD}}$	*2SLS*$^{\mathcal{SA}}$	*BGMM*	*DB-GMM*
100	125	0	Bias	0.2853	1.2577	0.0184	0.2378	0.0050
			RMSE	0.2981	19.4112	0.0868	0.2582	0.1386
100	125	0.5	Bias	0.2620	0.1175	0.0122	0.2186	-0.0123
			RMSE	0.2750	3.6345	0.0855	0.2392	0.1430
100	125	0.9	Bias	0.2301	0.2144	0.0119	0.2261	-0.0020
			RMSE	0.2435	2.0268	0.0755	0.2477	0.1281
250	125	0	Bias	0.2832	0.2386	0.0096	0.2129	-0.0011
			RMSE	0.2881	0.2449	0.0572	0.2228	0.0876
250	125	0.5	Bias	0.2604	0.2255	0.0069	0.2004	0.0044
			RMSE	0.2657	0.2321	0.0522	0.2109	0.0905
250	125	0.9	Bias	0.2194	0.2159	0.0049	0.2182	0.0007
			RMSE	0.2251	0.2224	0.0479	0.2283	0.0825

Notes: We re-ran DGP 2 in TABLE 4 when $\mathcal{S}$ is an empty set. PGMM is not reported because it requires at least one instrument in $\mathcal{S}$. Note that*2SLS*$^{\mathcal{SD}}$ denotes the 2SLS using all instruments in the doubt set since $\mathcal{S} \cup \mathcal{D} = \mathcal{D}$ as $\mathcal{S}$ is an empty set. Also, the oracle estimator *2SLS*$^{\mathcal{SA}}$ denotes the 2SLS using all the good instruments in $\mathcal{A}$ since $\mathcal{S} \cup \mathcal{A} = \mathcal{A}$ and therefore *2SLS*$^{\mathcal{SA}}$ = *2SLS*$^{\mathcal{A}}$.

TABLE 10:
Estimation of the Automobile Demand

	OLS	$2SLS^{\mathcal{SD}}$	BGMM	DB-GMM
constant	-10.0716	-10.0438	-10.1862	-5.7703
	(0.2576)	(0.2608)	(0.2754)	(0.0473)
HPW	-0.1243	0.1161	0.5607	8.8917
	(0.2790)	(0.3179)	(0.3601)	(0.5010)
AC	-0.0343	0.0584	0.3226	3.2981
	(0.0710)	(0.0880)	(0.1106)	(0.6078)
MP$	0.2650	0.2484	0.2633	0.5145
	(0.0425)	(0.0433)	(0.0469)	(0.0762)
Size	2.3421	2.3331	2.4378	-1.3084
	(0.1246)	(0.1265)	(0.1353)	(0.1825)
Price	-0.0886	-0.097	-0.1138	-0.3790
	(0.0043)	(0.0063)	(0.0085)	(0.0436)

Notes: DB-GMM is capable of estimating without any pre-selected instrument. We ran the empirical application with $\mathcal{S}$ as an empty set. $2SLS^{\mathcal{S}}$ and PGMM are not reported because both of them require at least one instrument in $\mathcal{S}$.