UNIVERSITY OF CALIFORNIA
RIVERSIDE

Causal Inference using Boosting in IV Regression Models

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Economics

by

Hao Xu

June 2018

Dissertation Committee:

Dr. Tae-Hwy Lee, Chairperson
Dr. Michael Bates
Dr. Gloria Gonzalez-Rivera
Dr. Daniel Jeske
Dr. Aman Ullah

The Dissertation of Hao Xu is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I would like to express my deepest gratitude to my advisor Professor Tae-Hwy Lee for his excellent guidance, patience, and continuous support during my PhD study and research. He is not only a knowledge advisor that leads my way of research, but also a great mentor who show me how a good economist and responsible person should be.

For the rest of my committee members, I would like to thank Professor Aman Ullah and Professor Gloria Gonzalez-Rivera, who help me built up the foundation of Econometrics and give precious advise. I would like to thank Professor Daniel Jeske and Professor Michael Bates for their insightful feedback on my research.

My sincere thanks also goes to Professor Marcelle Chauvet, Professor Jang-Ting Guo, Professor David A. Malueg, and all my other instructors for their support and encouragement. I will not be able to reach this accomplishment without them. A special thanks to Gary Kuzas for his unfailing support and assistance.

I would like to thank my friends and fellow graduate students for the nice memories we share together. I want to thank Zhen Yu, Yi Shao, and Jenny Gao for all the encouragement and help they offered during my years as a PhD student.

Last but not the least, I would like to thank my family. The great examples that my parents and my grandparents set up for me make who I am today. They are always my strong support and inspire me to become a better person.

To my grandparents.

ABSTRACT OF THE DISSERTATION

Causal Inference using Boosting in IV Regression Models

by

Hao Xu

Doctor of Philosophy, Graduate Program in Economics
University of California, Riverside, June 2018
Dr. Tae-Hwy Lee, Chairperson

This dissertation focuses on using the machine learning technique, boosting, for causal inference in the instrumental variable (IV) regression models.

In Chapter 1, 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 even when there are more instruments than the sample size. We estimate the parameter of interest using generalized method of moments (GMM) with selected instruments. We refer this method as Double Boosting GMM (DB-GMM). We show that DB does not select weakly relevant or weakly valid instruments. In Monte Carlo, we compare DB-GMM with other methods such as GMM using Lasso penalty (penalized GMM). 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.

Extending from Chapter 1, Chapter 2 combines the DB selection algorithm from Chapter 1 with the multiple-layer neural networks (NN) for the first-stage IV estimation, where high dimensional sieve instrument variables are the activation functions at the last hidden layer of the neural networks.

Chapter 3 studies the panel data models with many instruments. When the regressors are endogenous in the panel data models, we employed the 2SLS approach for the FE estimator. We denote it as FE-2SLS. We find that the FE-2SLS estimator is sensitive to the number of instruments, where it is inconsistent when the number of instruments increases. We show that using the two regularization methods, SCAD and L_2Boosting, for instrument selection make the FE-2SLS estimator more robust and restore its consistency when there are many instruments. Furthermore, we consider a Stein-like combined estimator of the FE and FE-2SLS estimators and provide its asymptotic properties. A empirical study is conducted for the economics of real house price using the US state level panel data.

Contents

List of Figures **x**

List of Tables **xi**

1 Introduction **1**

2 Double Boosting GMM for High Dimensional IV Regression Models **4**
2.1 Introduction 4
2.2 Model 8
2.3 Boosting GMM (BGMM) 11
2.3.1 L_2Boosting Algorithm 13
2.3.2 Consistency of L_2Boosting 15
2.4 Double-Boosting GMM (DB-GMM) 17
2.4.1 Double-Boosting Algorithm 17
2.5 Monte Carlo 21
2.5.1 Simulation Results 24
2.6 Estimation of Automobile Demand Function 27
2.7 Conclusions 31

3 Double Boosting for Nonlinear IV Regression using Neural Network **37**
3.1 Introduction 37
3.2 Model 39
3.3 Double-Boosting Neural Network (DB-NN) 41
3.3.1 Instruments selection 43
3.4 Monte Carlo 45
3.4.1 Simulation Results 47
3.5 Estimation of Automobile Demand Function 50
3.6 Conclusions 53

4 Estimation of Panel Data Models for US State Level House Price with Many Instruments **64**
4.1 Introduction 64

4.2 Estimation of Panel Data Regression Models 66
4.2.1 FE, FE-2SLS, and Combined estimators 66
4.2.2 Asymptotic properties of FE, FE-2SLS, and Combined estimators . 69
4.3 Estimation of Panel Data Regression Models with Many Instruments 75
4.3.1 SCAD 77
4.3.2 Boosting 78
4.4 Monte Carlo 81
4.5 Estimation of House Price Panel Data Model in US States 85
4.6 Conclusions 88

5 Conclusions **93**

Bibliography **95**

A Appendix for Chapter 1 **98**
A.1 Proof of Theorem 1 98
A.2 Proof of Theorem 2 99

B Appendix for Chapter 3 **105**
B.1 Proof of Theorem 2: 105
B.2 Proof of the inconsistency of FE-2SLS 108

List of Figures

2.1 Price Elasticity of Cars by DB-GMM 36

3.1 Price Elasticity of Cars by DB-NN 63

4.1 Median Squared Error of FE, FE-2SLS and Combined Estimators, $n = 49$, $T = 29$, $q = 3$. 89
4.1 Median Squared Error of FE, FE-2SLS and Combined Estimators, $n = 49$, $T = 29$, $q = 3$. 90
4.2 Median Squared Error of FE, FE-2SLS, Combined, F2SLS-SCAD and Combined SCAD Estimators, $n = 49$, $T = 29$, $q = 3$, $p = 12$. 91
4.3 Median Squared Error of FE, FE-2SLS, Combined, F2SLS-Boosting and Combined Boosting Estimators, $n = 49$, $T = 29$, $q = 3$, $p = 12$. 91

List of Tables

2.1 Categories of instruments 32
2.2 Order of ω_j for each category of instruments 32
2.3 DGP 1 33
2.4 DGP 2 34
2.5 DGP 3 34
2.6 Estimation of the Automobile Demand 35
2.7 Number of Cars with Inelastic Demand 35

3.1 Categories of instruments 55
3.2 DGP 1 56
3.3 Instrument Selection (DGP 1) 57
3.4 DGP 2 58
3.5 Instrument Selection (DGP 2) 59
3.6 DGP 3 60
3.7 Instrument Selection (DGP 3) 61
3.8 Estimation of the Automobile Demand 62
3.9 Number of Cars with Inelastic Demand 62

4.1 Table 1. Economics of Real House Prices for 49 U.S. States, 1975–2003 . . . 92

Chapter 1

Introduction

Using instruments is one of the solution for estimation when regressors are endogenous. However, as discussed in Bekker (1994), the two-stage least squares (2SLS) estimation is inconsistent if the number of instruments is relatively large compared to the number of observations. Similar property is found in the generalized method of moments (GMM) by Stock and Wright (2000). Hence, a regularization method is necessary in order to maintain the consistency of the 2SLS (or GMM) estimation. Under the sparsity assumption, where only a few instruments are relevant to the endogenous regressors, the least absolute shrinkage and selection operator (Lasso) are commonly used for instruments selection. The use of Lasso can easily extend to GMM estimation by including an L_1 penalty to the objective function.

Chapter 1 studies the case when the functional form of the endogenous regressors is unknown. We apply the approximately sparse model, where a large set of sieve instruments are generated through polynomials, to approximate the true functional form. We relax

the assumptions on both the validity and relevancy of the instruments. An instrument is said to be valid if there is not correlation between the instrument and the structural error, and is said to be relevant if it is correlated with the endogenous regressors. Chapter 1 introduce a selection method, Double Boosting (DB), that will consistently select all the valid and relevant instrument. We compare the result of DB-GMM with another penalized GMM method, which includes a L_1 penalty that checks the validity and relevancy of the instruments. We find that DB-GMM is very robust even when the number of instruments is smaller than the number of observations.

Chapter 2 extends the use of DB to the multiple-layer neural network framework. When the approximately sparse model is used for approximating the nonlinear functional form of the endogenous regressors, it requires a strong assumption on the function form of the sieve instruments, where researchers need to choose the sieve functions that are closely related to the true functional form. Instead using polynomial as suggested in Chapter 1, the sieve instruments is generated through the last hidden layer of the multiple-hidden layer neural network in Chapter 2. Then we use DB to select only the valid and relevant instruments. Finally, we compute the first-stage IV estimation by neural network with selected sieve instruments. We apply the same empirical example as in Chapter 1, but different set of instruments are used.

Chapter 3 is a joint work with Bai Huang, Aman Ullah, and Tae-Hwy Lee. It extends the use of selection methods to the panel data models. When the regressors are endogenous due to simultaneity or measurement errors, the fixed effect (FE) estimator for a panel data model is inconsistent. In such cases the 2SLS approach using instrumental

variables may be employed for the FE estimator, which we denote as FE-2SLS. However, when the number of instrument is large, the inconsistency of 2SLS estimation discussed in Bekker (1996) for cross-sectional data models has carried over to the panel data models as well. Chapter 3 analytically shows that the FE-2SLS is inconsistent when the number of instruments increases. We also show that using regularization methods such as SCAD and L_2Boosting for the selection of instruments makes the FE-2SLS estimator more robust and restore its consistency when there are many instruments. Furthermore, extending Hansen (2017) to the structural panel data models, we consider a Stein-like combined estimator of the FE and FE-2SLS estimators and provide its asymptotic properties. We show that the combined estimator has the asymptotic risk strictly smaller than that of the FE-2SLS estimator when the FE-2SLS is consistent (which is ensured by the regularization of many instruments). Our Monte Carlo analysis shows that the asymptotic theory carries over to finite sample only when a small number of good instruments are carefully selected, while otherwise the combined estimator can be worse than the FE-2SLS estimator from using too many instruments. Guarded with these theoretical and numerical findings, a careful empirical study is conducted for the economics of real house price using the US state level panel data.

Chapter 2

Double Boosting GMM for High Dimensional IV Regression Models

2.1 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_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, 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 be-

cause 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 chapter, 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 chapter is organized as follows. In Section 2.2, 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 2.3, we review the instrument selection procedure by L_2Boosting 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 2.4, we propose a new instrument selection method, DB. In Section 2.5, Monte Carlo studies are presented to compare DB-GMM with other methods. Section 2.6 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 2.77 concludes. All proofs are gathered in Appendix A.

2.2 Model

Consider an IV model as

$$y_i = \beta' x_i + u_i \tag{2.1}$$

$$x_i = E(x_i|w_i) + v_i. \tag{2.2}$$

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

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \tag{2.3}$$

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

$$E(x_i|w_i) = \sum_{j=1}^{\ell_n} \gamma_j h_j(w_i) + r_i, \tag{2.4}$$

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

$$(z_{1,i} \ \dots \ z_{\ell_n,i})' \equiv (h_1(w_i) \ \dots \ h_{\ell_n}(w_i))'. \tag{2.5}$$

From Equations (2.2) and (2.4),

$$x_i = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + r_i + v_i. \tag{2.6}$$

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

$$g(z_{j,i}, \beta) = z_{j,i} u_i. \tag{2.7}$$

The validity of each instrument depends on the moment condition,

$$E\left(g(z_{j,i}, \beta)\right) = E\left(z_{j,i} u_i\right) = \frac{b_j}{n^{\delta_j}}. \tag{2.8}$$

And the relevancy of each instrument depends on the parameter,

$$\gamma_j = \frac{a_j}{n^{\alpha_j}}. \tag{2.9}$$

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: $\mathcal{V}_1 = \{j : b_j = 0\} \cup \left\{j : b_j \neq 0 \text{ and } \frac{1}{2} < \delta_j\right\}$, $\mathcal{V}_2 = \left\{j : b_j \neq 0 \text{ and } 0 < \delta_j \leq \frac{1}{2}\right\}$, 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 $\mathcal{S} = \{Z_1, \ldots, Z_{\ell_\mathcal{S}}\}$ 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 $\mathcal{B} \equiv \mathcal{B}_0 \cup \mathcal{B}_1$. Table 2.1 summarizes each subset of the instruments according to Definitions 1 and 2.

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_2Boosting. 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_2Boosting. At each step $m = 1, \ldots, \bar{M}$, where $\bar{M}$ is the maximum iteration of L_2Boosting, 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 leaner for $i = 1, \ldots, n$. The relationship between the weak learner and the strong learner is

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \tag{2.10}$$

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_2Boosting for each variable in x_i.

2.3.1 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 $X = (x_1 \ \dots \ x_n)'$ using instruments in $\mathcal{S}$ is

$$F_{0,i} = f_{0,i} = \hat{\gamma}_{0,\text{initial}} + \sum_{j=1}^{\ell_{\mathcal{S}}} \hat{\gamma}_{j,\text{initial}} z_{j,i}, \tag{2.11}$$

   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

$$\{\hat{\gamma}_{0,j}, \hat{\gamma}_j\} = \min_{\gamma_0, \gamma_j} \sum_{i=1}^{n} \left(\hat{v}_{m,i} - \gamma_0 - \gamma_j z_{j,i}\right)^2. \tag{2.12}$$

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

$$j_m = \arg \min_{j \in \{\ell_S + 1, \dots, \ell_n\}} \sum_{i=1}^{n} \left(\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i}\right)^2. \tag{2.13}$$

   (c) The weak learner is

$$f_{m,i} = \hat{\gamma}_{0,j_m} + \hat{\gamma}_{j_m} z_{j_m,i}, \tag{2.14}$$

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

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

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \tag{2.15}$$

   with $c_m > 0$.

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

A stopping rule is necessary in L_2Boosting 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 $\hat{V}_m = (\hat{v}_{m,1} \ \ldots \ \hat{v}_{m,n})'$, $f_m = (f_{m,1} \ \ldots \ f_{m,n})'$, $F_m = (F_{m,1} \ \ldots \ F_{m,n})'$, 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 (4.33),

$$\begin{aligned} \mathbf{1}\hat{\gamma}_{0,j_m} + Z_{j_m}\hat{\gamma}_{j_m} &= P_m\hat{V}_m \\ f_m &= P_m\left(X - F_{m-1}\right). \end{aligned} \tag{2.16}$$

Let $\mathbf{Z}_S = (Z_1 \ \ldots \ Z_{\ell_S})$. 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

$$AIC_c(m) = \log(\hat{\sigma}^2_{2,m}) + \frac{1 + \text{trace}(B_m)/n}{1 - (\text{trace}(B_m) + 2)/n}, \tag{2.17}$$

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 $\hat{M} = \arg\min_{m=1,\ldots,\bar{M}} AIC_c(m)$.

2.3.2 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: $\sup_{1\le j\le \ell_n, n\in\mathbb{N}} ||Z_j||_\infty < \infty$, *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 $\sum_{j=1}^{\ell_n} |\gamma_j| \to \infty$ as $n \to \infty$, 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_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 (2.6)

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

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* $E\left(r_i^2|w_i\right) \leq \sigma_2^2 \left(\frac{\log \ell_n}{n}\right)$, *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 sieves* $\{z_{j,i}\}$. *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 $W = (W_1 \ \ldots \ W_p)$.

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.

2.4 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.

2.4.1 Double-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). The difference is the new objective function (2.24) replacing (2.13). We now doubly minimize the invalidity (measured by (2.21)) and minimize the irrelevancy (measured by the inverse of (2.25)) of an instrument in each iteration, as we describe in details below.

First, we measure the invalidity based on the usual Lagrange Muliplier (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

$$\rho_j = \frac{E\left(z_{j,i}u_i\right)}{\sqrt{E\left(z_{j,i}^2\right)}\sqrt{E\left(u_i^2\right)}} = \frac{d_j}{n^{\delta_j}}. \tag{2.18}$$

where $d_i = \frac{b_j}{\sqrt{E\left(z_{j,i}^2\right)}\sqrt{E\left(u_i^2\right)}}$ and b_j is defined in Equation (2.8).

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,

$$\hat{u}_i \equiv y_i - \hat{\beta}_{\text{initial}} x_i, \tag{2.19}$$

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

$$\hat{\rho}_j = \frac{\frac{1}{n}\sum_{i=1}^{n} z_{j,i}\hat{u}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} z_{j,i}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2}}. \tag{2.20}$$

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

$$nR_{\mathcal{V},j}^2 = n\hat{\rho}_j^2. \tag{2.21}$$

Similarly, we also define the LM statistic measure for relevancy of z_j, $nR^2_{\mathcal{R},j}$, which we describe in (2.25) 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

$$F_{0,i} = f_{0,i} = \hat{\gamma}_{0,\text{initial}} + \sum_{j=1}^{\ell_{\mathcal{S}}} z_{j,i}\hat{\gamma}_{j,\text{initial}}, \tag{2.22}$$

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

$$\{\hat{\gamma}_{0,j},\ \hat{\gamma}_j\} = \min_{\gamma_0,\gamma_j} \sum_{i=1}^{n} \left(\hat{v}_{m,i} - \gamma_0 - \gamma_j z_{j,i}\right)^2. \tag{2.23}$$

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

$$j_m = \arg \min_{j \in \{\ell_S+1,\dots,\ell_n\}} \omega_j \equiv \frac{\left(nR^2_{\mathcal{V},j}\right)^{r_2}}{\left(nR^2_{\mathcal{R},j}\right)^{r_1}}, \tag{2.24}$$

   where

$$R^2_{\mathcal{R},j} = 1 - \frac{\sum_{i=1}^{n} \left(\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i}\right)^2}{\sum_{i=1}^{n} \left(\hat{v}_{m,i} - \bar{v}_m\right)^2}, \tag{2.25}$$

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

   (c) The weak learner is

$$f_{m,i} = \hat{\gamma}_{0,j_m} + \hat{\gamma}_{j_m} z_{j_m,i}, \tag{2.26}$$

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

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

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \tag{2.27}$$

   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_2Boosting.

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 $nR^2_{\mathcal{R},j}$ 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_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 $m = 1, \ldots, \bar{M}$. In addition, minimizing $\frac{1}{nR^2_{\mathcal{R},j}}$ is the same as maximizing $nR^2_{\mathcal{R},j}$.

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 (2.13) 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 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 $n \to \infty$. That is,*

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

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

Proof: Appendix B.

2.5 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):

$$y_i = \beta x_i + u_i,$$
$$x_i = \sum_{j=1}^{p} \gamma_j w_{j,i} + v_i = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + v_i, \tag{2.28}$$

where the endogenous variable x_i is a scaler ($k = 1$), 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

$$(z_{S,i}\ z_{A,i}\ z_{\mathcal{B}_0,i}\ z^*_{\mathcal{B}_1,i}) \sim N(0, \Sigma_Z) \tag{2.29}$$
$$(u_i\ v_i) \sim N(0, \Sigma), \tag{2.30}$$

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 $\Sigma_Z = \text{diag}(\Sigma_{\mathcal{S}\cup\mathcal{A}}, \Sigma_{\mathcal{B}})$. $\Sigma_{\mathcal{S}\cup\mathcal{A}}$ 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 2.3. 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

$$z_{j,i} = z^*_{j,i} + c_j u_i, \tag{2.31}$$

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

$$c_j = c_0 + \frac{(j-29)(\bar{c}-c_0)}{\ell_n/2 - 2}. \tag{2.32}$$

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):

$$\begin{aligned} y_i &= \beta x_i + u_i \\ x_i &= \sum_{j=1}^{p} \theta_j \left(w_{j,i} + w^2_{j,i} \right) + v_i, \end{aligned} \tag{2.33}$$

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

$$(w_{1,i}\ w_{2,i}\ w_{3,i}\ w_{4,i}\ w^*_{5,i}) \sim N(0, \Sigma_W), \tag{2.34}$$

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

$$(u_i\ v_i) \sim N(0, \Sigma), \tag{2.35}$$

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 (3.20), by adding the structural error u_i

$$w_{5,i} = w^*_{5,i} + u_i. \tag{2.36}$$

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,

$$x_i = \sum_{j=1}^{p} \theta_j \exp(w_{j,i}) + v_i. \tag{2.37}$$

In DGP 1, as $z_{j,i} = w_{j,i}$, all instruments are observable and the functional form of $h_j(\cdot)$ in (2.4) 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 $j = 1, \ldots, p$, and $z_{j,i} = h_j(w_i)$ for $j = p+1, \ldots, \ell_n$, where $h_j(w_i)$ is the polynomial of each instrument in w_i up to 4^{th} order. Let $z_{\mathcal{S},i} = (z_{1,i}\ z_{2,i})'$. We summarize the simulation results in Tables 2.3 to 2.5.

2.5.1 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 2.3 to 2.5, 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$^{\mathcal{SD}}$, 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. Especially when $a = 0.9$, the results of 2SLS$^{\mathcal{S}}$, 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$^{\mathcal{SD}}$ diverges, which confirms the theoretical result in Bekker (1994) . With only instruments in $\mathcal{S}$, 2SLS$^{\mathcal{S}}$ has improvement in the result. But both 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. 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.

2.6 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

$$\xi_{it} = \delta(w_{it}, x_{it}, u_{it}, \beta) + \varepsilon_{it}, \tag{2.38}$$

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

$$s_{it} = \frac{\exp(\delta_{it})}{1 + \sum_{\forall it} \exp(\delta_{it})}. \tag{2.39}$$

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

$$y_{it} = \beta_0 + \beta_{\text{price}} x_{it} + \beta_w' w_{it} + u_{it}, \tag{2.40}$$

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 (2.4),

we assume price is a linear combination of product characteristics and sieve functions of product characteristics such that

$$x_{it} = \gamma_0 + \gamma_w' w_{it} + \gamma_1' h_1(w_{it}) + \gamma_2' h_2(w_{it}, t) + \gamma_3' h_3(w_{it}) + v_{it}, \tag{2.41}$$

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 (2.40) and (2.41) 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 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 2.6. 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 $\text{2SLS}^{\mathcal{SD}}$ 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 $\text{2SLS}^{\mathcal{SD}}$ may be inconsistent. On the other hand, $\text{2SLS}^{\mathcal{S}}$ fails to capture all strongly valid and strongly relevant instruments among the nonlinear sieve instruments. Hence, the estimators in $\text{2SLS}^{\mathcal{S}}$ 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 2.6 and 2.7. 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$^{\mathcal{S}}$. The inefficiency of 2SLS$^{\mathcal{S}}$ 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

$$\hat{e}_{it} = \frac{\%\triangle s_{it}}{\%\triangle x_{it}} = \frac{\partial s_{it}/s_{it}}{\partial x_{it}/x_{it}} = \frac{x_{it}}{s_{it}}\frac{\partial s_{it}}{\partial x_{it}} = \hat{\beta}_{\text{price}}\ x_{it}\ (1 - s_{it}). \tag{2.42}$$

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 2.1, 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

$$\sum_{\forall it} \mathbf{1}(\hat{e}_{it} > -1), \tag{2.43}$$

and the counts are reported in Table 2.7. 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.

2.7 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.

Table 2.1: Categories of instruments

		Strongly Valid $\mathcal{V}_1$	Weakly Valid $\mathcal{V}_2$	Invalid $\mathcal{V}_3$
Irrelevant	$\mathcal{R}_1$			
Weakly Relevant	$\mathcal{R}_2$	$\mathcal{B}_0$	$\mathcal{B}_1$	
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 2.2: Order of ω_j for each category of instruments

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

Table 2.3: 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 $a =$ 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 2.4: DGP 2

n	ℓ_n	a	OLS	2SLS$^{\mathcal{SD}}$	2SLS$^{\mathcal{S}}$	2SLS$^{\mathcal{SA}}$	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 2.5: DGP 3

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

Table 2.6: Estimation of the Automobile Demand

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

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

Table 2.7: Number of Cars with Inelastic Demand

OLS	2SLS$^{\mathcal{SD}}$	2SLS$^{\mathcal{S}}$	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 (3.31) is larger than -1. We count how many cars have inelastic demand by (2.43). 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 $\left(\hat{\beta}_{\text{price}} \pm 1.96\text{se}(\hat{\beta}_{\text{price}})\right) x_{it}(1 - s_{it})$.

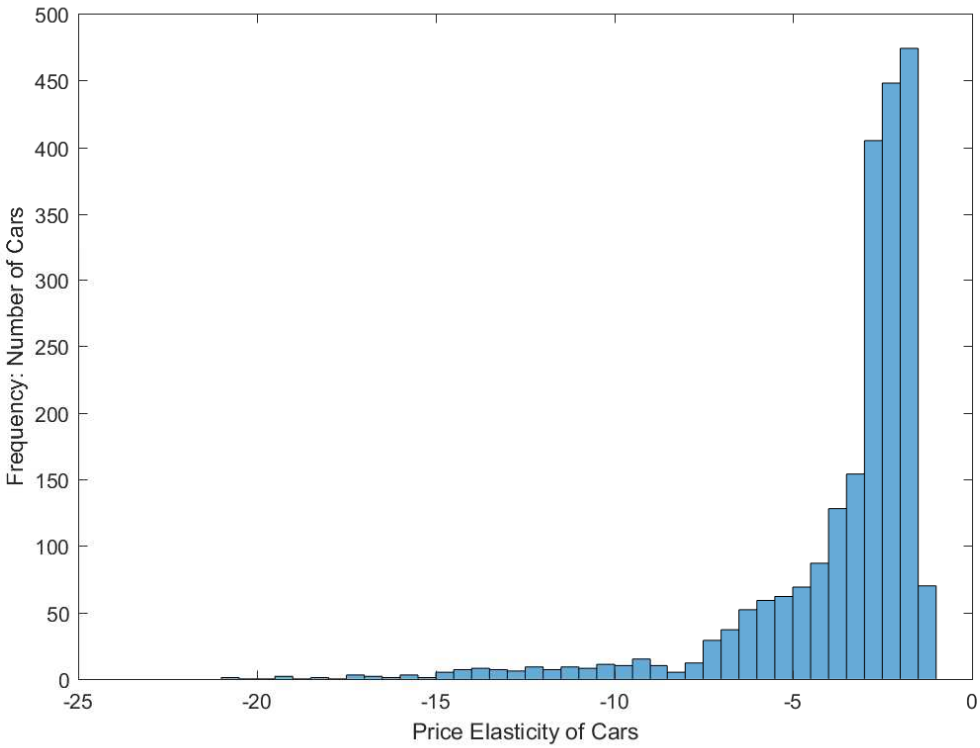


Figure 2.1: Price Elasticity of Cars by DB-GMM

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.

Chapter 3

Double Boosting for Nonlinear IV Regression using Neural Network

3.1 Introduction

A neural network is a nonlinear statistical model that can be applied to both regression or classification. Unlike other estimation procedure, the neural network includes a hidden layer, which is a set of activation functions with assigned weights on the input variables. Following the idea, White (2006) introduces a nonlinear model in the neural network architecture, where the activation functions are the logistic function with randomly assigned coefficients. He refers this method as QuickNet. However, compared to other forward-stagewise procedures such as boosting, neural networks have advantage neither in computation nor in accuracy.

With a significant improvement in the computation power, an innovation on the neural network by increasing the number of hidden layers has been introduced after 2010. The multiple-layer neural network is mainly applied to the image recognition. At the same time, with its capability on estimating the nonlinear function, the multiple-layer neural network can help to estimate any model even when the true functional form of the model is not observable. Thus, the application of the neural network can be easily extended to the causal inference analysis in a high dimensional instrumental variable (IV) regression model. Hartford, et al. (2016) and Chernozhukov, et al. (2016) apply the neural networks to study the causal inference, where the endogenous regressors are computed by an unknown nonlinear function. However, both papers assume that the instruments are not correlated with the structure error. Hence, all instruments are valid.

In this chapter, we use the activation functions in the neuron network (NN) to form the sieve instruments in the IV regression model. We relax the assumption on the validity of the instruments and apply the Double-Boosting (DB) selection algorithm (Algorithm 2) in Chapter 1 to check the validity and relevancy of the instruments. After the selection, we apply the generalized method of moments (GMM) with selected instruments to estimate the parameter of interest that examine the causal effect between the dependent variable and the endogenous variable.

This chapter is organized as follows. In Section 3.2, we set up a structural model for the IV regression, then we apply the definition of weak/strong validity or relevancy of instruments as in Chapter 1. In Section 3.3, we propose the DB neural network (DB-NN) procedure, where a set of sieve instruments is generated by a multiple-layer neural

network. Then we select valid and relevant instruments through Double Boosting. The final estimation is compute by two-stage least squares (2SLS), where the first stage estimation on the endogenous regressors is computed by a single layer neural network with selected instruments. In Section 3.4, Monte Carlo studies are presented to compare the new method with other methods. In Section 3.5, we apply the empirical application in Chapter 1 that follows the design in Berry, Levinsohn, and Pakes (1995). Section 3.6 concludes.

3.2 Model

Consider an IV model as

$$y_i = \beta' x_i + u_i \tag{3.1}$$

$$x_i = E(x_i|w_i) + v_i. \tag{3.2}$$

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

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \tag{3.3}$$

Let $z_{j,i} \equiv h_j(w_i)$ denotes the sieve instruments with a given functional form. The validity and the relevancy of instruments are defined in a local asymptotic framework. The moment function of each sieve instrument $z_{j,i}$ for $j = 1, \ldots, \ell_n$ is

$$g(z_{j,i}, \beta) = z_{j,i} u_i. \tag{3.4}$$

The validity of each sieve instrument depends on the moment condition,

$$E\left(g(z_{j,i}, \beta)\right) = E\left(z_{j,i} u_i\right) = \frac{b_j}{n^{\delta_j}}. \tag{3.5}$$

By Belloni, Chen, Chernozhukov, and Hansen (2012), the conditional mean of x_i can be approximated by the "approximately sparse model "with an approximation error r_i, such that

$$E(x_i|w_i) = \sum_{j=1}^{\ell_n} \gamma_j h_j(w_i) + r_i = \sum_{j=1}^{\ell_n} \gamma_j z_{j,i} + r_i. \tag{3.6}$$

The relevancy of each sieve instruments is depended on the coefficient between the instrument $z_{j,i}$ and the endogenous regressor x_i,

$$\gamma_j = \frac{a_j}{n^{\alpha_j}}. \tag{3.7}$$

Let $Z_j = (z_{j,1} \ \ldots \ z_{j,n})'$ for $j = 1, \ldots, \ell_n$. We follows the same definitions for validity and relevancy as in Chapter 1.

Definition 1 (Validity): The extent of validity depends on b_j and δ_j as follows: $\mathcal{V}_1 = \{j : b_j = 0\} \cup \left\{j : b_j \neq 0 \text{ and } \frac{1}{2} < \delta_j\right\}$, $\mathcal{V}_2 = \left\{j : b_j \neq 0 \text{ and } 0 < \delta_j \leq \frac{1}{2}\right\}$, 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 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, $z_{j,i}$ 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 $j \in \mathcal{R}_3$.

3.3 Double-Boosting Neural Network (DB-NN)

Instead using high order polynomials as described in Chapter 1, we generate the sieve instruments through the neural network procedure. Let ℓ_q denotes the number of activation functions generated at the q^{th} hidden layer. For $j = 1, \ldots, \ell_1$, the activation function at the first hidden layer is generated as

$$z_{j,i}^{(1)} = h_j(w_i) = h_j\left(\gamma_{j,0}^{(1)} + \sum_{j'=1}^{p} \gamma_{j,j'}^{(1)} w_{j',i}\right), \tag{3.8}$$

where $h_j(\cdot)$ is the given functional form of activation functions at each layer, $\gamma_{j,0}^{(1)}$ is the constant intercept, and $\gamma_{j,j'}^{(1)}$ is the weight assigned to the instrument $w_{j',i}$. For the q^{th} hidden layer, where $q > 1$,

$$z_{j,i}^{(q)} = h_j(z_i^{q-1}) = h_j\left(\gamma_{j,0}^{(q)} + \sum_{j=1}^{\ell_q} \gamma_{j,j'}^{(q)} z_{j',i}^{(q-1)}\right). \tag{3.9}$$

Let the maximum number of hidden layer is denoted as Q. We include the activation functions at the Q^{th} hidden layer, $z_i^{(Q)}$, together with the initial input variable, w_i, as an updated set of instruments, which include total $\ell_n = p + \ell_Q$ instruments. The new set of sieve instruments is an $n \times \ell_n$ matrix denoted as $Z = \left(W_1 \dots W_p \; Z_1^{(Q)} \dots Z_{\ell_Q}^{(Q)}\right)$. We partition Z into two subsets, the "sure" set $\mathcal{S}$ and the "doubt" set $\mathcal{D}$, following Cheng and Liao (2015). Let $\ell_{\mathcal{S}}$ denote the total number of instruments in $\mathcal{S}$, the "sure" set $\mathcal{S} = \{z_{1,i}, \dots, z_{\ell_{\mathcal{S}},i}\}$ includes the strongly valid and strongly relevant instruments that are initially selected. The remaining instruments are in the "doubt" set $\mathcal{D} = \{z_{\ell_{\mathcal{S}}+1,i}, \dots, z_{\ell_n,i}\}$, where 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 $\mathcal{A} \cup \mathcal{B}_0$. We summarizes each subset of the instruments according to Definition 1 and 2 in Table 3.1.

3.3.1 Instruments selection

In order to ensure only valid and relevant instruments are selected, we apply the Double Boosting algorithm as introduced in Chapter 1. Let

$$\rho_j = \frac{E\left(z_{j,i}u_i\right)}{\sqrt{E\left(z_{j,i}^2\right)}\sqrt{E\left(u_i^2\right)}} = \frac{d_j}{n^{\delta_j}}. \tag{3.10}$$

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 (3.5).

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,

$$\hat{u}_i \equiv y_i - \hat{\beta}_{\text{initial}}x_i, \tag{3.11}$$

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

$$\hat{\rho}_j = \frac{\frac{1}{n}\sum_{i=1}^{n} z_{j,i}\hat{u}_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} z_{j,i}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2}}. \tag{3.12}$$

We define the LM statistic measure for invalidity of z_j as

$$nR_{\mathcal{V},j}^2 = n\hat{\rho}_j^2. \tag{3.13}$$

Similarly, we also define the LM statistic measure for relevancy of z_j, $nR_{\mathcal{R},j}^2$, which we describe in (3.17) inside Double Boosting Algorithm.

Algorithm 2 Double Boosting

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

$$F_{0,i} = f_{0,i} = \hat{\gamma}_{0,\text{initial}} + \sum_{j=1}^{\ell_{\mathcal{S}}} z_{j,i}\hat{\gamma}_{j,\text{initial}}, \tag{3.14}$$

   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

$$\{\hat{\gamma}_{0,j},\ \hat{\gamma}_j\} = \min_{\gamma_0,\gamma_j} \sum_{i=1}^{n} \left(\hat{v}_{m,i} - \gamma_0 - \gamma_j z_{j,i}\right)^2. \tag{3.15}$$

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

$$j_m = \arg \min_{j \in \{\ell_S+1,\dots,\ell_n\}} \omega_j \equiv \frac{\left(nR^2_{\mathcal{V},j}\right)^{r_2}}{\left(nR^2_{\mathcal{R},j}\right)^{r_1}}, \tag{3.16}$$

   where

$$R^2_{\mathcal{R},j} = 1 - \frac{\sum_{i=1}^{n}\left(\hat{v}_{m,i} - \hat{\gamma}_{0,j} - \hat{\gamma}_j z_{j,i}\right)^2}{\sum_{i=1}^{n}\left(\hat{v}_{m,i} - \bar{v}_m\right)^2}, \tag{3.17}$$

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

   (c) The weak learner is

$$f_{m,i} = \hat{\gamma}_{0,j_m} + \hat{\gamma}_{j_m} z_{j_m,i}, \tag{3.18}$$

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

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

$$F_{m,i} = F_{m-1,i} + c_m f_{m,i}, \tag{3.19}$$

   with $c_m > 0$.

With the selected sieve instruments by Double Boosting, we estimate β by 2SLS, where the first stage estimation is computed by a single layer neural network.

3.4 Monte Carlo

To study the finite sample properties of different estimation methods under the IV regression model, where the function form of the endogenous regressor is unknown, we consider the three data generating processes (DGPs). For DGP 1, 2, and 3, let

$$y_i = \beta x_i + u_i,$$

where x_i is a scaler, $\beta = 0$, and $n \in \{250, 500\}$. Setting $p = 5$, we include 5 observable instruments initially. The observable strongly valid instruments are generated as

$$(w_{1,i}\ w_{2,i}\ w_{3,i}\ w_{4,i}\ w^*_{5,i}) \sim N(0, \Sigma_W), \tag{3.20}$$

where Σ_W is a $p \times p$ Toeplitz matrix with each (i, j) element $a^{|i-j|}$ and $a \in \{0, 0.5, 0.9\}$. The error terms u_i and v_i are generated as

$$(u_i\ v_i) \sim N(0, \Sigma), \tag{3.21}$$

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 (3.20), by adding the structural error u_i

$$w_{5,i} = w^*_{5,i} + u_i. \tag{3.22}$$

DGP 1 (**Polynomials**)**:**

$$x_i = \theta_0 + \sum_{j=1}^{p} \theta_j \left(w_{j,i} + w_{j,i}^2 \right) + v_i, \tag{3.23}$$

where $\theta_1 = \theta_2 = 0.1$, $\theta_3 = 0.5$, and $\theta_0 = \theta_4 = \theta_5 = 0$. So only the first three observable instruments are strongly relevant to x_i.

DGP 2 (**Exponential**)**:** The setting of parameters are the same as in DGP 1. But x_i is generated differently,

$$x_i = \theta_0 + \sum_{j=1}^{p} \theta_j \exp(w_{j,i}) + v_i, \tag{3.24}$$

where $\theta_1 = \theta_2 = 0.1$, $\theta_3 = 0.5$, and $\theta_0 = \theta_4 = \theta_5 = 0$.

DGP 3 (**Logistic**)**:**

$$x_i = \frac{\exp\left(\theta_0 + \sum_{j=1}^{p} \theta_j w_{j,i}\right)}{1 + \exp\left(\theta_0 + \sum_{j=1}^{p} \theta_j w_{j,i}\right)} + v_i, \tag{3.25}$$

where $\theta_1 = \theta_2 = 0.1$, $\theta_3 = 0.5$, and $\theta_0 = \theta_4 = \theta_5 = 0$.

Suppose the function forms of x_i in all DGPs are unknown. Let $\mathcal{S} = \{w_{1,i}\ w_{2,i}\}$ be initial selected. We compare the results when x_i is approximated (1) using polynomial up to 4^{th} order, (2) using neural network with selected sieve instruments. We summarize the simulation results in Tables 3.2 to 3.7.

3.4.1 Simulation Results

We compare DB-NN with OLS and five other methods that use up to 4^{th} order polynomial as sieve instruments. These methods include 2SLS$^{\mathcal{SD}}$ (2SLS with all instruments by polynomial), 2SLS$^{\mathcal{S}}$ (2SLS with only instruments in $\mathcal{S}$), 2SLS$^{\mathcal{SA}}$ (2SLS with all strongly valid and strongly relevant instruments in $\mathcal{S} \cup \mathcal{A}$), Penalized generalized method of moments (PGMM), and Double-Boosting GMM (DB-GMM). For DGP 1 and 2, 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 3.2, 3.4, and 3.6, the bias of the estimator $\hat{\beta}$ is in the first row and the root mean squared errors (RMSE) is in the second row.

The activation function used in DB-NN is the rectified linear unit (ReLU), where

$$h_j(z_i^{(q-1)}) = \max\left(0,\ \gamma_{j,0}^{(q)} + \sum_{j=1}^{\ell_n} \gamma_{j,j'}^{(q)} z_{j',i}^{(q-1)}\right). \tag{3.26}$$

The number of hidden layer $Q = 2$, and the number of activation functions in the first hidden layers is 10, and the number of activation functions in the second hidden layer is 30. We select from total 35 instruments in the selection procedure of DB-NN. We also try the simulation using different number of hidden layers in DB-NN.

DGP 1 and 2 are generated by additive separable functions of w_i, where no interaction term between w_i is involved. Then the sieve instruments with high order polynomials are sufficient for the estimation. In DGP 1, the bias and root mean squared errors (RMSE) of the OLS estimations for all cases are high because of the endogeneity. The results of 2SLS with all polynomial instruments (2SLS$^{\mathcal{SD}}$) are even worse when $n = 100$, where the

number of instruments is larger than the number of observations. When n gets larger, the bias and RMSE of 2SLS$^{\mathcal{SD}}$ is slightly smaller than OLS, but still larger than other methods. When only the instruments in $\mathcal{S}$ are selected for 2SLS, because the number of instruments is much smaller than n, the bias and RMSE have significant improvement compared to 2SLS$^{\mathcal{SD}}$. However, as 2SLS$^{\mathcal{S}}$ does not include all relevant instruments, its bias and RMSE jump significantly when there is not correlation between the observable instrument ($a = 0$). The estimations of 2SLS$^{\mathcal{SA}}$ provide the oracle result since it includes all the instrument is both $\mathcal{S}$ and $\mathcal{A}$. The standard solution of PGMM fails when $n = 100$ or $a = 0.9$ because the weighting matrix is not invertible. To solve these problem, we replace the weighting matrix with an identity matrix. The RMSE of PGMM remains high when $a = 0$, where all the observable instruments are uncorrelated with each others. As a and n increases, the result of PGMM gets better, and in some cases, the bias of PGMM is even the smallest among all other methods. DB-GMM gives the lowest RMSE in all the cases of DGP 1, and its bias is very close to the oracle result as provided under 2SLS$^{\mathcal{SA}}$. Even though the polynomial instruments are sufficient for DGP 1, the result of DB-NN is still very robust. In most of the cases, the RMSE of DB-NN is the second lowest among all other methods. When $m = 500$, the bias of DB-NN is smaller than the bias in DB-GMM for $a \in \{0.5, 0.9\}$.

Table 3.3 shows the summary of instrument selection by PGMM, DB-GMM, and DB-NN in DGP 1. For each cases, the first row is the average number of the instruments that are selected by each method. The second row is the standard deviation of the number of selected instruments. The variation of selection in PGMM is very large. When $a = 0$, the standard deviation on the number of selected instruments can be as high as 10.1740.

When $a = 0.9$, PGMM only selected very few addition instruments other than the 2 instruments that are initially selected by $\mathcal{S}$. The standard deviation on the number of selected instruments by DB-GMM are smoother compared to the PGMM. It ranges from 2.8093 to 4.8793. When a is the same, then umber of instruments that are selected by DB-GMM are similar at different cases of n. On average, DB-NN selects 5 to 6 instruments out of 35 instruments for all cases. The standard deviation ranges from 1.3142 to 1.8722.

The result of DGP 2 is similar to the result of DGP 1. Even though the endogenous variable x_i in DGP 2 is generated by an additively separable exponential function of w_i, the polynomial instruments are still sufficient for the final estimation since there is not interaction term within the exponential function. The selection summary of DGP 2 as reported in Table 3.5 also has similar pattern as the selection summary of DGP 1. The number of instrument selected is volatile in PGMM, and DB-NN selects around 6 instruments on average for all cases.

In DGP 3, we introduce a logistic function, where the interaction of w_i is included inside the exponential term. We find that all the linear estimation methods have failed under DGP 3, including $\text{2SLS}^{\mathcal{SA}}$, which is used as oracle result in DGP 1 and 2. On the other hand, DB-NN is able to remain a good estimation result that gives low bias. The RMSE of DB-NN is higher than the RMSE reported in DGP 1 and DGP 2, but it is still the lowest compared to other methods in DGP 3.

3.5 Estimation of Automobile Demand Function

We apply the same empirical application on the automobile market as used in Chapter 1. Consider a homogeneous individual log utility function

$$\xi_{it} = \delta(w_{it}, x_{it}, u_{it}, \beta) + \varepsilon_{it}, \tag{3.27}$$

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. Hence, each subscription it denotes a single observation. Let x_{it} be the price of car it, and it is endogenous because of the existence of some unobservable product characteristics u_{it}. The variable w_{it} is a vector of the observable market level product characteristics, and β is the parameters in $\delta(\cdot)$. Applying the simple logit model, the market share s_{it} for each car it is calculated as

$$s_{it} = \frac{\exp(\delta_{it})}{1 + \sum_{\forall it} \exp(\delta_{it})}. \tag{3.28}$$

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

$$y_{it} = \beta_0 + \beta_{\text{price}} x_{it} + \beta_w' w_{it} + u_{it}, \tag{3.29}$$

where $y_{it} = \log(s_{it}) - \log(s_{0t})$, and s_{0t} is the outside option in year t. The outside option includes the market share of those consumers who choose to buy a used car or to use alternative transportations.

Since price is endogenous, we assume price is a function of product characteristics and 10 additional instruments that has been described in BLP,

$$x_{it} = f\left(\gamma_0 + \gamma_1' h_1(w_{it})\right) + v_{it}, \tag{3.30}$$

where $h_1(w_{it})$ is generated 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. Instruments in $h_1(w_{it})$ should be included because the product characteristics of competitive cars and other cars in the same firm may also influence the price of the car.

The data used in BLP (1995) is obtained from annual issues of the *Automotive News Market Data Book* from 1971 to 1990. We use the data set given in Chernozhukov, Hansen, and Spindles (2015), who study the same automobile application as in BLP (1995). We include 4 observable market level product characteristics in the model, that are the dummy variable of air conditioning (AC), horsepower/weight (HPW), miles per dollar (MP$), and size of car (Size). Hence, $w_{it} = (\text{AC}_{it}\ \text{HPW}_{it}\ \text{MP\$}_{it}\ \text{Size}_{it})'$. Price is obtained from the listed retail price of the base model of each car in the unit of 1000 US dollars of year 1983. Given the price of gasoline, we calculate miles per dollar (MP$) by miles per gallon (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 (3.29) and (3.30) are estimated as if the data is cross-sectional (no time series) for $it = 1, \ldots, 2217$.

We assume the 4 observable product characteristics in w_{it} are valid. We select the constant and all variables in w_{it} for $\mathcal{S}$, where $\mathcal{S} = \{\mathbf{1}, \text{AC}, \text{HPW}, \text{MP\$}, \text{Size}\}$. The number of cars with inelastic demand should be small in order to be consistent with the profit maximization behavior of the firm. If the demand of a car is inelastic, firm would easily achieve a higher revenue by increasing the price.

The estimation results are reported in Table 3.8. We compare the result of DB-NN with 5 different methods as reported in Chapter 1. Other than OLS, the remaining 4 estimation methods use 63 sieve instruments, which include the quadratic and cubic terms of continuous variables in w_{it}, and the first order interactions of all variables in w_{it} and time t. The result of these 5 methods are the same as Table 2.6 in Chapter 1. For the result of DB-NN, all estimators are significant. DB-NN has different sign on the coefficients of HPW and MP\$ than DB-GMM because different number of instruments are selected. DB-NN only selects two additional instruments from the 35 potential sieve instruments that are generated at the last hidden layer. Its estimator of price is -0.4987, which implies that even more cars have elastic demand to price changes than the result in DB-GMM.

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

$$\hat{e}_{it} = \frac{\%\triangle s_{it}}{\%\triangle x_{it}} = \frac{\partial s_{it}/s_{it}}{\partial x_{it}/x_{it}} = \frac{x_{it}}{s_{it}}\frac{\partial s_{it}}{\partial x_{it}} = \hat{\beta}_{\text{price}}\, x_{it}\, (1 - s_{it}). \tag{3.31}$$

The price elasticity of demand using DB-NN ranges from -34.2084 to -1.6921 compared to the range of -20.5731 to -1.0176 in DB-GMM. The most elastic car is 1989 Porsche 911c, and the least elastic car is 1990 Yugo GV Plus. According to the histogram in Figure 3.1, the price elasticity of most of the cars is around -5.

A car is said to have an inelastic demand when its price elasticity of demand is larger than -1. We report the number of cars with inelastic demand in Table 3.9 by

$$\sum_{\forall it} \mathbf{1}(\hat{e}_{it} > -1). \tag{3.32}$$

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. Both of these two results are too large because they are inconsistent with pricing strategy that maximizes profit. In DB-GMM, none of cars has inelastic demand. Similarly, in DB-NN, none of cars has inelastic demand. With the standard error equals to 0.1112 in DB-NN, the lower bound for the number of cars with inelastic demand is 0 and the upper bound is 5.

3.6 Conclusions

We extend DB-GMM in Chapter 1 to the neural network, where Double Boosting is used to select the sieve instruments that are generated at the last hidden layers of the neural network procedure. Then we perform the two-stage least squares, where the first stage estimation is computed by using a single layer neural network with selected sieve instruments. Given the property of Double Boosting, we can consistently select strongly valid and strongly relevant sieve instrument. We show that DB-NN gives robust results when X is generated by an additive separable functions. And it will outperform all other linear estimation methods when X is generated by a non-separable function. In the application from BLP (1995) where instruments are the same as in the original BLP paper, none of the cars have inelastic demand of price by using DB-NN. Hence, DB-NN can fully capture the

nonlinearity in the model and select only the valid and relevant sieve instruments that are provided by the neural network.

Table 3.1: 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 3.2: DGP 1

			$\ell_n = 125$					$\ell_n = 35$
n	a	OLS	2SLS$^{\mathcal{SD}}$	2SLS$^{\mathcal{S}}$	2SLS$^{\mathcal{SA}}$	PGMM	DB-GMM	DB-NN
100	0	0.3007	1.6007	0.2014	0.0284	0.2114	0.0400	0.0799
		0.3130	11.1710	0.6846	0.1033	0.6257	0.1319	0.1940
100	0.5	0.2681	0.9357	0.0286	0.0262	0.0145	0.0318	0.0624
		0.2786	8.1831	0.6526	0.0777	0.5919	0.1041	0.1507
100	0.9	0.2207	0.2051	-0.0187	0.0051	-0.0181	0.0185	0.0466
		0.2341	1.9076	0.2006	0.0769	0.1847	0.1107	0.1235
250	0	0.2901	0.2449	0.1440	0.0128	0.1289	0.0160	0.0209
		0.2948	0.2510	0.8158	0.0569	1.1071	0.0744	0.1062
250	0.5	0.2548	0.2182	-0.0071	0.0026	-0.0062	0.0057	0.0344
		0.2604	0.2250	0.1609	0.0488	0.1495	0.0619	0.0989
250	0.9	0.2396	0.2370	0.0132	0.0150	0.0140	0.0329	0.0512
		0.2450	0.2431	0.1221	0.0551	0.1203	0.0934	0.1062
500	0	0.2797	0.2084	-0.0236	0.0055	0.0133	0.0057	0.0033
		0.2824	0.2121	0.3426	0.0321	0.3265	0.0418	0.0909
500	0.5	0.2518	0.1975	-0.0151	-0.0050	-0.0150	-0.0059	0.0034
		0.2548	0.2018	0.1076	0.0406	0.1071	0.0548	0.0808
500	0.9	0.2143	0.2085	0.0071	-0.0032	0.0062	0.0111	0.0242
		0.2170	0.2118	0.0694	0.0336	0.0703	0.0580	0.0917

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. When $\ell_n = 125$, instruments are generated by 4^{th} order polynomials. 2SLS$^{\mathcal{S}}$ denotes 2SLS with instruments in $\mathcal{S}$. When $\ell_n = 35$, the instruments are the 5 observable instruments together with the 30 activation functions at the last layer of DNN. 2SLS$^{\mathcal{SA}}$ denotes 2SLS with instruments in $\mathcal{S} \cup \mathcal{A}$, which demonstrates the oracle results in DGPs 1 and 2.

Table 3.3: Instrument Selection (DGP 1)

		$\ell_n = 125$		$\ell_n = 35$
n	a	PGMM	DB-GMM	DB-NN
100	0	8.8300	12.4600	5.7600
		4.7843	4.5135	1.7004
100	0.5	4.4100	11.1800	5.9200
		3.0521	4.4116	1.4473
100	0.9	2.2800	7.3700	5.4900
		2.2878	2.8093	1.4177
250	0	16.6200	14.0800	6.4100
		10.1740	4.3268	1.6943
250	0.5	5.6500	11.5900	6.5100
		3.3616	3.9647	1.8722
250	0.9	3.3100	8.4300	5.6600
		3.2340	3.6272	1.4370
500	0	14.6500	14.5100	6.0100
		7.4663	4.8793	1.3142
500	0.5	6.9400	11.1500	5.7800
		3.7788	3.1475	1.4112
500	0.9	4.3800	8.4100	5.4300
		4.0546	3.3001	1.3799

Table 3.4: DGP 2

			$\ell_n = 125$					$\ell_n = 35$
n	a	OLS	2SLS$^{\mathcal{SD}}$	2SLS$^{\mathcal{S}}$	2SLS$^{\mathcal{SA}}$	PGMM	DB-GMM	DB-NN
100	0	0.1723	0.2091	0.0969	-0.0032	0.0906	0.0152	0.0106
		0.1841	1.4646	0.7989	0.0630	0.6899	0.0990	0.1065
100	0.5	0.1621	0.1029	-0.0066	0.0104	-0.0154	0.0116	0.0227
		0.1820	1.3411	0.2081	0.0651	0.2109	0.0991	0.0982
100	0.9	0.1552	0.2393	0.0144	0.0131	0.0094	0.0280	0.0326
		0.1669	1.3275	0.0973	0.0565	0.0941	0.0751	0.0742
250	0	0.1659	0.1345	-0.0520	-0.0040	-0.0272	0.0052	0.0052
		0.1743	0.1444	0.3858	0.0413	0.3029	0.0474	0.0547
250	0.5	0.1527	0.1282	0.0076	0.0047	0.0074	0.0067	0.0144
		0.1600	0.1365	0.1082	0.0397	0.1116	0.0480	0.0622
250	0.9	0.1333	0.1285	-0.0172	-0.0038	-0.0185	0.0028	0.0145
		0.1408	0.1368	0.0657	0.0354	0.0667	0.0458	0.0734
500	0	0.1629	0.1129	0.0154	-0.0016	0.0271	0.0023	0.0001
		0.1656	0.1162	0.2258	0.0263	0.2080	0.0306	0.0474
500	0.5	0.1494	0.1118	-0.0031	0.0002	-0.0034	0.0009	0.0061
		0.1527	0.1161	0.0706	0.0243	0.0715	0.0271	0.0363
500	0.9	0.1313	0.1238	0.0025	-0.0027	0.0022	0.0040	0.0038
		0.1353	0.1284	0.0407	0.0263	0.0406	0.0409	0.0376

Table 3.5: Instrument Selection (DGP 2)

		$\ell_n = 125$		$\ell_n = 35$
n	a	PGMM	DB-GMM	DB-NN
100	0	7.4900	12.2500	6.0500
		5.1591	4.4955	1.6291
100	0.5	3.8100	10.2400	6.2100
		2.6882	3.6157	1.7191
100	0.9	2.3300	7.9500	5.7500
		2.4784	3.3526	1.6167
250	0	13.2600	14.3900	5.5000
		8.1694	4.6945	1.5008
250	0.5	4.7500	11.0200	5.3700
		2.7206	3.4640	1.5219
250	0.9	3.2200	7.8000	5.6000
		2.6877	3.2998	1.3999
500	0	11.0200	14.6700	5.9200
		5.3352	4.5860	1.4749
500	0.5	6.2000	11.7100	5.6400
		3.3090	3.3853	1.4875
500	0.9	5.6100	7.9600	5.3200
		4.8198	2.9471	1.1090

Table 3.6: DGP 3

			$\ell_n = 125$					$\ell_n = 35$
n	a	OLS	2SLS$^{\mathcal{SD}}$	2SLS$^{\mathcal{S}}$	2SLS$^{\mathcal{SA}}$	PGMM	DB-GMM	DB-NN
100	0	0.6641	10.8145	0.8487	0.4367	0.7685	0.4502	0.1060
		0.6707	99.5346	2.6133	0.8046	1.5034	1.0764	0.3538
100	0.5	0.6537	0.6757	0.4540	0.2861	0.5533	0.3959	0.1418
		0.6581	0.8887	2.1114	0.7634	2.8928	1.7859	0.3084
100	0.9	0.6352	0.0738	0.0933	0.2505	0.0731	0.2475	0.1296
		0.6427	3.7400	0.9054	0.6080	0.9165	0.6307	0.2505
250	0	0.6444	0.6509	0.5659	0.1867	0.5690	0.2580	0.0852
		0.6470	0.6548	1.2254	0.5536	3.2049	0.6288	0.4178
250	0.5	0.6430	0.6674	0.2225	0.0620	0.2289	0.1626	0.0593
		0.6453	0.6713	0.8011	0.4944	1.2008	0.5499	0.2061
250	0.9	0.6556	0.7585	-0.0395	0.0388	-0.4151	0.1197	0.0633
		0.6584	0.7619	0.6839	0.5093	2.6467	0.6702	0.1980
500	0	0.6552	0.6709	0.3782	0.0525	0.2207	0.1255	0.0550
		0.6564	0.6733	1.0108	0.3612	2.1402	0.4655	0.1370
500	0.5	0.6527	0.7103	0.0702	0.0572	0.1171	0.0788	0.0246
		0.6538	0.7126	0.5253	0.3814	0.6540	0.3170	0.1110
500	0.9	0.6519	0.8272	0.0083	0.0357	0.1099	0.1662	0.0292
		0.6529	0.8285	0.3715	0.3216	0.8053	0.5199	0.1333

Table 3.7: Instrument Selection (DGP 3)

		$\ell_n = 125$		$\ell_n = 35$
n	a	PGMM	DB-GMM	DB-NN
100	0	12.9600	4.2800	6.9300
		6.3832	3.1337	2.5634
100	0.5	10.9300	3.5500	6.8000
		6.2672	2.7721	2.4163
100	0.9	6.8600	3.7900	6.6600
		5.1070	3.3674	2.5907
250	0	28.0300	6.4100	7.7800
		11.1251	5.3938	2.2454
250	0.5	19.2900	4.7100	8.0300
		10.2082	4.0956	2.7467
250	0.9	9.3300	4.0100	7.9000
		7.4888	3.6529	2.5643
500	0	31.0100	9.2100	8.0400
		9.9071	5.7758	2.4159
500	0.5	15.3600	7.0800	8.2800
		9.4607	5.7624	2.7637
500	0.9	8.2500	4.5800	7.8800
		6.4766	4.4294	2.4915

Table 3.8: Estimation of the Automobile Demand

		$\ell_n = 125$				$\ell_n = 35$
	OLS	2SLS$^{\mathcal{SD}}$	2SLS$^{\mathcal{S}}$	PGMM*	DB-GMM	DB-NN
constant	-10.0716	-10.0438	-11.4900	-11.2749	-9.3702	-5.3996
	(0.2576)	(0.2608)	(0.6114)	(4.5630)	(0.3800)	(1.1150)
HPW	-0.1243	0.1161	-12.3812	-44.6293	5.9361	-2.2784
	(0.2790)	(0.3179)	(0.5440)	(2.7079)	(1.0240)	(0.2984)
AC	-0.0343	0.0584	-4.7606	3.7666	2.3026	1.5013
	(0.0710)	(0.0880)	(0.2894)	(1.6090)	(0.3631)	(0.5648)
MP$	0.2650	0.2484	1.1134	-1.5585	-0.1544	0.4495
	(0.0425)	(0.0433)	(0.1048)	(0.7671)	(0.0833)	(0.0468)
Size	2.3421	2.3331	2.8004	2.6639	2.1155	2.4900
	(0.1246)	(0.1265)	(0.2623)	(2.0612)	(0.1770)	(0.1336)
Price	-0.0886	-0.0970	0.3387	0.3968	-0.2999	-0.4987
	(0.0043)	(0.0063)	(0.0171)	(0.1014)	(0.0328)	(0.1112)

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

Table 3.9: Number of Cars with Inelastic Demand

	$\ell_n = 125$				$\ell_n = 35$
OLS	2SLS$^{\mathcal{SD}}$	2SLS$^{\mathcal{S}}$	PGMM*	DB-GMM	DB-NN
1502	1427	2217	2217	0	0
(1425,1626)	(1230,1563)	(2217,2217)	(2217, 2217)	(0, 7)	(0,5)

Note: The demand is said to be inelastic when the price elasticity of demand in (3.31) is larger than -1. We count how many cars have inelastic demand by (3.32). 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 $\left(\hat{\beta}_{\text{price}} \pm 1.96\text{se}(\hat{\beta}_{\text{price}})\right) x_{it}(1 - s_{it})$.

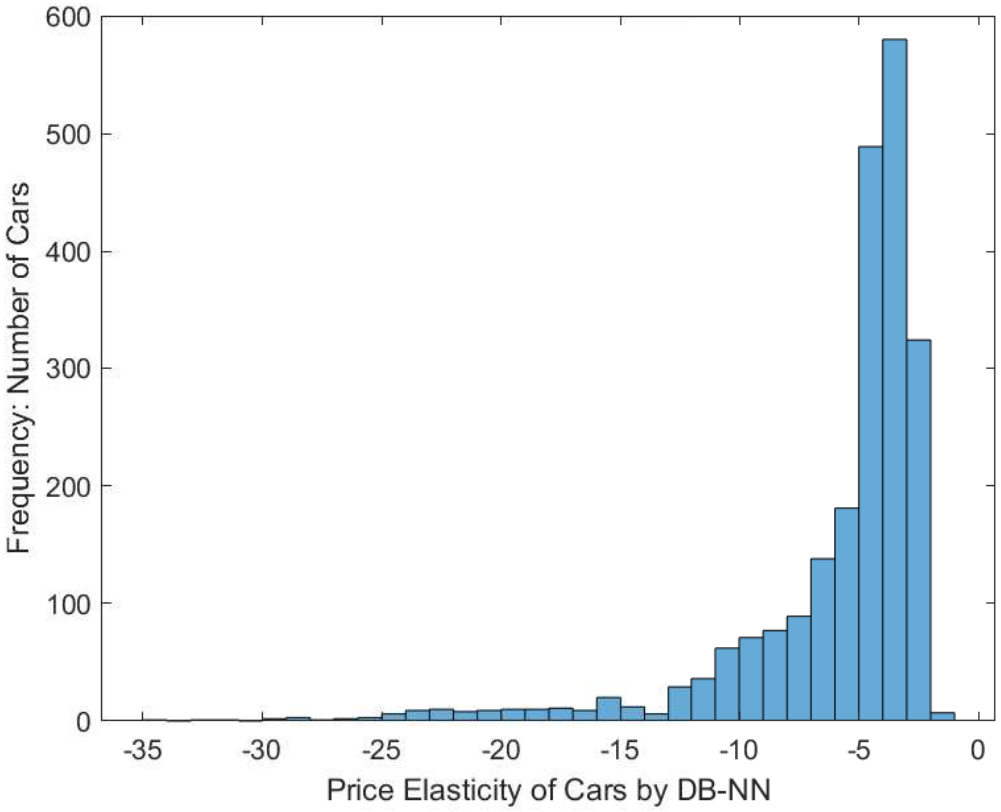


Figure 3.1: Price Elasticity of Cars by DB-NN

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

Chapter 4

Estimation of Panel Data Models for US State Level House Price with Many Instruments

4.1 Introduction

When the regressors are endogenous due to simultaneity or measurement errors, the fixed effect (FE) estimator for a panel data model is inconsistent. We apply the 2SLS estimation approach to the FE estimator using instrumental variables (which will be called FE-2SLS) to analyze the US house prices using state level panel data. However, we find that the FE-2SLS estimator is very sensitive to the number of selected instruments when there are many available instruments. An example is shown as in the house price panel data model, where the instruments are taken from the lagged endogenous regressors.

We have conducted an extensive theoretical, numerical and empirical analysis of the FE-2SLS estimators and demonstrate the needs and benefits of using regularization methods such as SCAD and L_2Boosting for the selection of relevant instruments when there are many instruments. The regularization makes the 2SLS estimator more robust in using many instruments.

Also, extending Hansen (2017) to the structural panel data models, we propose a combined (model averaging) estimator of the FE and FE-2SLS estimators and provide the asymptotic properties of these estimators. It is shown that the combined estimator has the asymptotic risk strictly smaller than that of the FE-2SLS estimator when FE-2SLS is consistent. While our Monte Carlo simulation confirms the asymptotic theory, it also shows the asymptotic theory carries over to finite sample only when the small number of good instruments are carefully selected. Using too many instruments, FE-2SLS can be as bad as FE even when endogeneity is strong, and the combined estimator can be worse than FE-2SLS. Guarded with these theoretical and numerical findings, a careful empirical study is conducted for the economics of real house price using the US state level panel data.

This chapter is organized as follows. In Section 4.2, we first present the asymptotic properties of FE, FE-2SLS, and the combined estimator of FE and FE-2SLS in the panel data regression model with endogenous regressors. In Section 4.3, we examine issues that FE-2SLS can go bad if too many instruments are used, and regularization methods are called for to guard FE-2SLS against the problem. We extend the two regularization methods, SCAD by Fan and Li (2001) and L_2Boosting by Bhlmann (2006), to the panel data model. We modify the stopping rule of L_2Boosting to ensure the consistency of the estimator for

the panel data models. Section 4.4 presents Monte Carlo simulation to demonstrate how the FE-2SLS and combined estimators are affected by the number of selected instruments, and how the FE-2SLS estimator can be regularized. Section 4.5 presents the empirical analysis of the US house price using the state level panel data of Holly, Pesaran, and Yamagata (2010). Concluding remarks are given in Section 4.6. All proofs are collected in Appendix.

4.2 Estimation of Panel Data Regression Models

4.2.1 FE, FE-2SLS, and Combined estimators

In a panel data model, the fixed effects estimator helps us to resolve the endogeneity issues that arise because of the correlated unobserved effects. Endogeneity issues may also arise due to a nonzero correlation between explanatory variables and idiosyncratic errors. In the presence of such correlations, both fixed effects (FE) and random effects (RE) estimators yield biased and inconsistent estimates of the parameter. The resulting biases can not be removed via differencing estimation. The traditional technique to overcome this problem is to find instruments for those explanatory variables which are potentially correlated with idiosyncratic errors. For example see Hausman and Taylor (1981), Amemiya and MaCurdy (1986) and Breusch et al (1989). These papers consider the application of instrumental-variable procedures to estimate the parameters of the model with endogenous regressors, with the error structure implied by random effects. See Baltagi (2008) for the commonly used fixed effects 2SLS estimator.

It is well known that the finite sample properties of the 2SLS estimator are often problematic. Thus, most of the justification for the use of 2SLS estimator is asymptotic. Its

performance in small samples may be poor. In the presence of weak instruments, the loss of precision will be severe, and 2SLS estimates may be no improvement over the individual effects estimators.

Consider the following panel regression model with fixed effects:

$$y_{it} = x'_{it}\beta + \alpha_i + u_{it}, \quad i = 1, \ldots, n, \ t = 1, \ldots, T, \tag{4.1}$$

where x_{it} is $q \times 1$, and β is a $q \times 1$ vector of unknown parameters. α_i's are fixed effects and u_{it}'s are the random disturbances. In matrix notation, Eq (4.1) can be written as

$$y = X\beta + D\alpha + u, \tag{4.2}$$

$D \equiv I_n \otimes \iota_T$ is $nT \times n$ where ι_T is a vector of ones, α is $n \times 1$, and $u \sim \left(0, \ \sigma_u^2 I_{nT}\right)$. Pre-multiplying the model (4.2) by $Q \equiv I_{nT} - D\left(D'D\right)^{-1} D'$ and performing OLS on the resulting transformed model:

$$Qy = QX\beta + QD\alpha + Qu, \tag{4.3}$$

where $QD = 0$. Noting that Q is idempotent, the $\hat{\beta}_{\text{FE}}$ can be obtained as

$$\hat{\beta}_{\text{FE}} = \left(X'QX\right)^{-1} X'Qy. \tag{4.4}$$

The asymptotic distribution of $\hat{\beta}_{\text{FE}}$ is

$$\sqrt{n}\left(\hat{\beta}_{\text{FE}}-\beta\right)\xrightarrow{d}N\left(0,V_1\right), \tag{4.5}$$

where $V_1=\sigma_u^2\left(\text{plim}\frac{X'QX}{n}\right)^{-1}$.

The endogeneity occurs due to the (i) correlation of α_i with x_{it} or (ii) correlation of u_{it} with x_{it}. We consider the latter case here for which the FE estimator becomes inconsistent. With u_{it} and x_{it} correlated, the vector x_{it} is endogenous. Performing 2SLS on (4.3) with QZ as the set of instruments

$$Z'QQy=Z'QQX\beta+Z'QQu, \tag{4.6}$$

one gets the FE-2SLS estimator

$$\hat{\beta}_{\text{FE-2SLS}}=\left(X'H_ZX\right)^{-1}X'H_Zy, \tag{4.7}$$

where $H_Z=QZ\left(Z'QZ\right)^{-1}Z'Q$. The asymptotic distribution of $\hat{\beta}_{\text{FE-2SLS}}$ follows

$$\sqrt{n}\left(\hat{\beta}_{\text{FE-2SLS}}-\beta\right)\xrightarrow{d}N\left(0,V_2\right), \tag{4.8}$$

where $V_2=\sigma_u^2\left(\text{plim}\frac{X'H_ZX}{n}\right)^{-1}$.

The FE-2SLS estimator is preferred to the FE estimator as it is consistent under endogeneity, while the FE estimator is inconsistent. However in small samples, FE-2SLS can have much larger variance so FE can have better MSE precision especially when the

extent of endogeneity is not severe. To consider this scenario, the endogeneity is set to be local to zero as explained below, and we propose the following combined estimators $\hat{\beta}_c$, which is weighted average of FE and FE-2SLS estimators with the weights depending on the Hausman (1978) statistic

$$\hat{\beta}_{\mathrm{c}} = w\hat{\beta}_{\mathrm{FE}} + (1-w)\hat{\beta}_{\text{FE-2SLS}}, \tag{4.9}$$

where

$$w = \begin{cases} \frac{\tau}{H_n} & \text{if } H_n \geq \tau \\ 1 & \text{if } H_n < \tau \end{cases}, \tag{4.10}$$

$$H_n = (\hat{\beta}_{\text{FE-2SLS}} - \hat{\beta}_{\mathrm{FE}})' \left(\hat{V}_2 - \hat{V}_1\right)^{-1} (\hat{\beta}_{\text{FE-2SLS}} - \hat{\beta}_{\mathrm{FE}}), \tag{4.11}$$

and τ is a shrinkage parameter. The degree of shrinkage depends on the ratio τ/H_n.

4.2.2 Asymptotic properties of FE, FE-2SLS, and Combined estimators

Write the reduced form equation for the endogenous variable x_{it} as

$$x_{it} = \Pi' z_{it} + v_{it} \tag{4.12}$$

with $E(z_{it}v_{it}) = 0$. Instruments z_{it} is $\ell \times 1$ and Π is $\ell \times q$. Next, we write the structural equation error u_{it} as a linear function of the reduced form error v_{it} and an orthogonal error ε_{it}

$$u_{it} = v_{it}\rho + \varepsilon_{it} \tag{4.13}$$

with $E(v_{it}\varepsilon_{it}) = 0$. The variable x_{it} is exogenous if u_{it} and v_{it} are uncorrelated, or equivalently that the coefficient ρ is zero. We use the local asymptotic approach. For fixed T, ρ is local to zero

$$\rho = \frac{1}{\sqrt{n}}\delta \tag{4.14}$$

δ is a $q \times 1$ localizing parameter, which indexes the degree of correlation between u_{it} and e_{it}. When $\delta = 0$, x_{it} are exogenous. When $\delta \neq 0$, x_{it} are endogenous in finite sample. δ (and thus ρ) controls the degree of endogeneity.

We make the following assumptions:

Assumption 1. $(x_i,\ \alpha_i,\ u_i)$ are i.i.d. over i, u_{it} is i.i.d. over t; $E\left(u_{it}^4 | x_{it},\ \alpha_i\right) < \infty$, and $E(\varepsilon_{it}|v_{it})^4 < \infty$.

Assumption 2. $E\|x_{it}\|^{2+k} < \infty$ and $E|u_{it}|^{2+k} < \infty$ for some $k > 0$.

Assumption 3. $E\|x_{it}\|^4 < \infty$, $E\|z_{it}\|^4 < \infty$, $E\|e_{it}\|^4 < \infty$; $\sigma_u^2\left(\text{plim}_{n\to\infty}\frac{1}{n}X'QX\right)^{-1} = V_1$,

$\sigma_u^2\left(\text{plim}_{n\to\infty}\frac{1}{n}X'H_ZX\right)^{-1} = V_2$.

Assumption 4. $\hat{\sigma}_u^2 = \sigma_u^2 + o_p(1)$.

Assumption 5. $\text{rank}(\Pi) = q$.

Assumptions 1-3 specify that the variables have finite fourth moments so that a central limit theory applies. Assumption 5 is the rank condition on Π to ensure that the coefficient β is identified. Denote $\Sigma = E(v_{it}v_{it}')$.

Theorem 1: *Under Assumptions 1-5*,

$$\sqrt{n}\begin{pmatrix}\hat{\beta}_{\text{FE}}-\beta\\ \hat{\beta}_{\text{FE-2SLS}}-\beta\end{pmatrix}\xrightarrow{d} h+\xi, \tag{4.15}$$

where

$$h=\begin{pmatrix}\sigma_u^{-2}V_1\text{tr}\left(Q\Sigma\right)\delta\\ 0\end{pmatrix}, \tag{4.16}$$

$$\xi\sim N(0,V), \tag{4.17}$$

$$V=\begin{pmatrix}V_1 & V_1\\ V_1 & V_2\end{pmatrix}. \tag{4.18}$$

Furthermore,

$$H_n\xrightarrow{d}(h+\xi)'B(h+\xi), \tag{4.19}$$

$$\sqrt{n}\left(\hat{\beta}_{\text{c}}-\beta\right)\xrightarrow{d}\Psi=G_2'\xi-\left(\frac{\tau}{(h+\xi)'B(h+\xi)}\right)_1 G'\left(h+\xi\right), \tag{4.20}$$

where $B=G\left(V_2-V_1\right)^{-1}G'$, $G=\begin{pmatrix}-I & I\end{pmatrix}'$, $G_2=\begin{pmatrix}0 & I\end{pmatrix}'$, *and* $\left(a\right)_1=\min\left[1,a\right]$.

The theorem is similar to Theorem 1 in Hansen (2017) except for the expressions of h, V_1, and V_2, and its proof is straightforward and thus omitted. This theorem gives expression for the joint asymptotic distribution of $\hat{\beta}_{\text{FE}}$ and $\hat{\beta}_{\text{FE-2SLS}}$ estimators, the Hausman statistic, and the combined estimators under the local exogeneity assumption in (4.14). The joint asymptotic distributions are normal. $\hat{\beta}_{\text{FE}}$ has asymptotic bias when $\delta\neq 0$ but not the $\hat{\beta}_{\text{FE-2SLS}}$ estimator. The Hausman statistic controls the weight and thus the degree of shrinkage. It is an asymptotic non-central chi-square random variable with non-centrality parameter depending on the local endogeneity parameter δ. The asymptotic distribution of

the combined estimator is nonlinear functions of the normal random vector and functions of the non-centrality parameter. The asymptotic distribution of $\hat{\beta}_{\mathrm{c}}$ in (4.9) is written as a random weighted average of the asymptotic distributions of $\hat{\beta}_{\mathrm{FE}}$ and $\hat{\beta}_{\text{FE-2SLS}}$.

These alternative estimators are compared in the asymptotic risk. The asymptotic risk of any sequence of estimators β_n of β is defined as

$$R\left(\beta_n, \beta, W\right) = \lim_{n\to\infty} E\left[n\left(\beta_n - \beta\right)' W\left(\beta_n - \beta\right)\right] = R\left(\beta_n\right), \tag{4.21}$$

so long as the estimator has an asymptotic distribution

$$\sqrt{n}\left(\beta_n - \beta\right) \xrightarrow{d} \psi, \tag{4.22}$$

for some random variable ψ. The asymptotic risk of the estimator β_n can be calculated using

$$R\left(\beta_n\right) = E\left(\psi' W \psi\right) = \operatorname{tr}\left(W E\left(\psi\psi'\right)\right). \tag{4.23}$$

Denote the largest eigenvalue $\lambda_1 \equiv \lambda_{\max}\left(W\left(V_2 - V_1\right)\right)$ of the matrix $W\left(V_2 - V_1\right)$ and the ratio $d \equiv \operatorname{tr}(W\left(V_2 - V_1\right))/\lambda_1$. The following theorem is an extension of Theorem 2 of Hansen (2017) for the panel data model.

Theorem 2: *Under Assumptions 1-5, if*

$$d > 2 \text{ and } 0 < \tau \leq 2\left(d-2\right), \tag{4.24}$$

then

$$\begin{aligned} R\left(\hat{\beta}_{\text{FE-2SLS}}\right) &= \text{tr}\left(WV_2\right), \\ R\left(\hat{\beta}_{\text{c}}\right) &< R\left(\hat{\beta}_{\text{FE-2SLS}}\right) - \frac{\tau\lambda_1\left(2\left(d-2\right)-\tau\right)}{\sigma_u^{-4}\delta'\text{tr}\left(Q\Sigma\right)V_1\left(V_2-V_1\right)^{-1}V_1\text{tr}\left(Q\Sigma\right)\delta+q}. \end{aligned} \tag{4.25}$$

Its proof is provided in Appendix B.1. Equation (4.25) shows that the asymptotic risk of the combined FE and FE-2SLS estimator is strictly less than that of the FE-2SLS estimator, so long as the shrinkage parameter τ satisfies the condition (4.24). The assumption $d > 2$ is necessary in order for the right-hand-side of the inequality equation in (4.24) to be positive, which is necessary for the existence of τ. τ appears in the risk bound (4.25) as a quadratic expression, so there is an optimal choice $\tau_{opt} = \frac{\text{tr}(W(V_2-V_1))}{\lambda_1} - 2$ which minimizes this bound. In the special case $W = \left(V_2 - V_1\right)^{-1}$, we find that condition (4.24) simplifies to $q > 2$ and $0 < \tau \leq 2\left(q-2\right)$. The assumption $q > 2$ is Stein's (1956) classic condition for shrinkage. Stein (1956) shows that the shrinkage dimension must exceed 2 in order for shrinkage to achieve global reductions in risk relative to unrestricted estimation.

Corollary 3: *If* $d > 2$ *and* $0 < \tau \leq 2(d-2)$, $R\left(\hat{\beta}_{\mathrm{c}}\right) - R\left(\hat{\beta}_{\text{FE-2SLS}}\right) < 0$.

The following two corollaries are obtained with $W = (V_2 - V_1)^{-1}$.

Corollary 4: *Under the local exogeneity assumption,*

$$R\left(\hat{\beta}_{\mathrm{FE}}\right) = tr(WV_1) + \sigma_u^{-4}\delta' \mathrm{tr}(Q\Sigma) V_1 W V_1 \mathrm{tr}(Q\Sigma)\delta,$$

and

$$\begin{cases} R\left(\hat{\beta}_{\mathrm{FE}}\right) \leq R\left(\hat{\beta}_{\text{FE-2SLS}}\right) & \text{if } \sigma_u^{-4}\delta' \mathrm{tr}(Q\Sigma) V_1 W V_1 \mathrm{tr}(Q\Sigma)\delta \leq q \\ R\left(\hat{\beta}_{\mathrm{FE}}\right) > R\left(\hat{\beta}_{\text{FE-2SLS}}\right) & \text{otherwise.} \end{cases}$$

Corollary 4 indicates that when endogeneity is weak (ρ and hence δ is close to zero) the FE estimator may perform better than the FE-2SLS estimator.

Corollary 5: *If* $q < \sigma_u^{-4}\delta' \mathrm{tr}(Q\Sigma) V_1 W V_1 \mathrm{tr}(Q\Sigma)\delta$, $d > 2$, $0 < \tau \leq 2(d-2)$, *then* $R\left(\hat{\beta}_{\mathrm{c}}\right) - R\left(\hat{\beta}_{\mathrm{FE}}\right) < 0$.

Corollary 5 indicates that when endogeneity is strong, $d > 2$, $0 < \tau \leq 2(d-2)$, the combined estimator performs better than both the FE and FE-2SLS estimators.

Remark 1: If a subset of regressors is treated as endogenous, consider the following structural equation of a panel data model:

$$y = X\beta + D\alpha + u \tag{4.26}$$

where $X = (X_1 \; Z_1)$ and $\beta = (\beta_1 \; \beta_2)$. Let X_1 be q_1 endogenous variables, Z_1 be ℓ_1 included

exogenous variables, and $q = q_1 + \ell_1$. Let $Z = (Z_1\ Z_2)$ be the set of ℓ $(= \ell_1 + \ell_2)$ exogenous variables (instrumental variables). This equation is identified if $\ell_2 \geq q_1$. In this case, one can use QZ as the set of instruments to get the FE-2SLS estimator as

$$\hat{\beta}_{\text{FE-2SLS}} = \left(Z' H_Z Z\right)^{-1} Z' H_Z y$$

with $H_Z = QZ \left(Z'QZ\right)^{-1} Z'Q$.

4.3 Estimation of Panel Data Regression Models with Many Instruments

Under the cross-sectional data, Bekker (1994) has shown that OLS will be inconsistent when there exists an endogenous variables, and 2SLS will be inconsistent when the number of instruments is large. Then selecting proper instruments is important for a consistent result in the IV model. When there is a large set of potential instruments and only a small subset of the instruments are relevant, an instrument selection procedure is needed to reduce the dimension of instruments and obtain a consistent estimation. Given its capability to shrink some of the estimators toward zero, the least absolute shrinkage and selection operator (Lasso) has been applied widely for variable selection. In Belloni, Chen, Chernozhukov, and Hansen (2012) and Belloni and Chernozhukov (2013), Lasso is used for the instrument selection. Caner (2009), Fan and Liao (2014), Liao (2013) and Cheng and Liao (2015) extend the penalized least square estimation to GMM. In addition, other penalized estimation methods include adaptive elastic net and smoothly clipped ab-

solute deviation penalty (SCAD). Instead of these regularized (penalized) methods, Ng and Bai (2008) and Chapter 1 use boosting for variable selection. Donald, Imbens, and Newey (2009) use information criteria for moment selection.

According to Bekker (1994), the 2SLS estimator is inconsistent when the number of instruments ℓ is large, i.e., the 2SLS is inconsistent unless $\frac{\ell}{n} \to 0$. We extend the Bekker results for the fixed effect panel data model. By replacing X with $X^* = QX$ and Z with $Z^* = QZ$, the Bekker theorem on the inconsistency result of the 2SLS estimator can be extended to the FE-2SLS estimators using large number of instruments in the panel data models. While $\hat{\beta}_{\text{FE}}$ is inconsistent due to the endogeneity, $\hat{\beta}_{\text{FE-2SLS}}$ may also be inconsistent due to the large number of instruments (see Appendix B.2).

In order to ensure the consistency of the FE-2SLS estimator, we extend the two regularization methods, SCAD by Fan and Li (2001) and L_2Boosting by Bhlmann (2006), to the panel data model. Let $X^* = (X_1^* \dots X_q^*)$, $Z^* = (Z_1^* \dots Z_\ell^*)$, and Π_k $(k = 1, \dots, q)$ be the kth column of Π. If some elements in Π are zeros, only a subset of z_{it} is relevant to x_{it}. Then we use SCAD and L_2Boosting to select only the relevant instruments and compute the FE-2SLS estimator based on the selected instruments. We call the estimator using SCAD as FE-2SLS-SCAD, and the estimator using L_2Boosting as FE-2SLS-Boosting. With the regularized FE-2SLS estimators, we call the combined estimator of the FE and FE-2SLS-SCAD estimators as Combined-SCAD, and the combined estimator of the FE and FE-2SLS-Boosting as Combined-Boosting.

Since both SCAD and L_2Boosting are able to shrink some elements of the coefficient matrix $\hat{\Pi}$ to zero corresponding to weak instruments, the subsequent application of

2SLS will consider those removed instruments as irrelevant to the endogenous variables. Once the instruments are selected, we use the selected instruments to compute the regularized FE-2SLS estimators (i.e., FE-2SLS-SCAD and FE-2SLS-Boosting).

4.3.1 SCAD

SCAD is used to ensure only relevant instruments will enter the estimation procedure. For $k = 1, \ldots, q$ and $j = 1, \ldots, \ell$, consider an objective function

$$\hat{\Pi}_k = \min_{\Pi_k} \left\{ \frac{1}{2nT} \sum_{i=1}^{n} \sum_{t=1}^{T} (x_{it,k}^* - \Pi_k' z_{it}^*)^2 + \lambda_k \sum_{j=1}^{\ell} p_j \left(\Pi_{kj} \right) \right\}, \tag{4.27}$$

where $p_j(\cdot)$ is the penalty function for $j = 1, \ldots, \ell$. For simplicity, let the penalty function be the same for each j, then $p_{\lambda_k}(\cdot) = \lambda_k p_j(\Pi_{kj})$ with tuning parameter λ_k. The objective function of SCAD uses a penalty function that is continuously differentiable, and it is able to assign different penalty to different instruments. Hence,

$$\lambda_k p_j \left(\Pi_{kj} \right) = \begin{cases} \lambda_k \left| \Pi_{kj} \right| & \text{if } \left| \Pi_{kj} \right| \leq \lambda_k \\ -\frac{\left| \Pi_{kj} \right|^2 - 2a\lambda_k \left| \Pi_{kj} \right|}{2(a-1)} & \text{if } \lambda_k < \left| \Pi_{kj} \right| \leq a\lambda_k \\ \frac{(a+1)\lambda_k^2}{2} & \text{if } \left| \Pi_{kj} \right| > a\lambda_k \end{cases}, \tag{4.28}$$

where $a = 3.7$. To optimize the objective function, Fan and Li (2001) derive the first order derivative of SCAD penalty function with respect to Π_k as

$$p_{\lambda_k}' \left(\Pi_{kj} \right) = \lambda_k \left\{ 1 \left(\Pi_{kj} \leq \lambda_k \right) + \frac{\left(a\lambda_k - \Pi_{kj} \right)_+}{(a-1)\lambda_k} 1 \left(\Pi_{kj} > \lambda_k \right) \right\}. \tag{4.29}$$

So the estimator $\hat{\Pi}_{kj}$ can be solved as

$$\hat{\Pi}_{kj} = \begin{cases} \text{sign}\left(\tilde{\Pi}_{kj}\right)\left|\tilde{\Pi}_{kj} - \lambda_k\right| & \text{if } \left|\tilde{\Pi}_{kj}\right| \leq 2\lambda_k \\ -\frac{(a-1)\left|\tilde{\Pi}_{kj}\right| - \text{sign}\left(\tilde{\Pi}_{kj}\right)a\lambda_k}{a-2} & \text{if } 2\lambda_k < \left|\tilde{\Pi}_{kj}\right| \leq a\lambda_k \\ \tilde{\Pi}_{kj} & \text{if } \left|\tilde{\Pi}_{kj}\right| > a\lambda_k \end{cases},$$

where $\tilde{\Pi}_{kj}$ is the OLS estimator of $x^*_{it,k}$ regressing on instruments z^*_{it}. The optimal tunning parameter λ_k for SCAD is estimated using the 10-fold cross-validated mean squared errors.

4.3.2 Boosting

Let m denotes the m^{th} iteration in the boosting procedure, and $\bar{M}$ denotes the maximum number of iteration. L_2Boosting performs an instrument selection for each X^*_k using the following procedure.

1. When $m = 0$, the initial weak learner for $x^*_{it,k}$ is

$$F_{0,it} = f_{0,it} = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} x^*_{it,k} \tag{4.30}$$

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

   (a) We compute the "current residual", $\hat{v}_{m,it} = x^*_{it,k} - F_{m-1,it}$.

   (b) Next, we regress the current residual $\hat{v}_{m,i}$ on each instrument $z^*_{j,it}$, for $j = 1, \ldots, \ell$.

   The estimators $\hat{\Pi}_{k0}$ and $\hat{\Pi}_{kj}$ are solved as

$$\left\{\hat{\Pi}_{k0,j}, \hat{\Pi}_{kj}\right\} = \min_{\Pi_{k0},\Pi_{kj}} \sum_{i=1}^{n}\sum_{t=1}^{T}\left(\hat{v}_{m,it} - \Pi_{k0} - \Pi_{kj} z^*_{j,it}\right)^2. \tag{4.31}$$

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

$$j_m = \arg \min_{j \in \{\ell_S+1,\ldots,\ell_n\}} \sum_{i=1}^{n} \sum_{t=1}^{T} \left(\hat{v}_{m,it} - \hat{\Pi}_{k0,j} - \hat{\Pi}_{kj} z^*_{j,it} \right)^2. \tag{4.32}$$

(c) The weak learner is

$$f_{m,i} = \hat{\Pi}_{k0,j_m} + \hat{\Pi}_{kj_m} z^*_{j_m,it}, \tag{4.33}$$

where $z^*_{j_m,it}$ is the instrument that is selected.

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

$$F_{m,it} = F_{m-1,it} + c_m f_{m,it}, \tag{4.34}$$

with $c_m > 0$.

3. We repeat 1 and 2 for $k = 1, \ldots, q$.

A stopping rule is necessary in L_2Boosting in order to avoid over-fitting. In Bhlmann (2006), the optimal number of iteration $\hat{M}$ is chosen by a suggested version of AIC, denoted as AIC_c, in a cross-sectional data model. Let $\hat{V}_m = (\hat{v}_{m,11} \ \ldots \ \hat{v}_{m,nT})'$, $f_m = (f_{m,11} \ \ldots \ f_{m,nT})'$, $F_m = (F_{m,11} \ \ldots \ F_{m,nT})'$, and $\mathbf{1}$ be an $nT \times 1$ vector of ones. We define $Q\mathbf{Z}_{j_m} = (\mathbf{1} \ QZ_{j_m})$, and $P_m = Q\mathbf{Z}_{j_m}(\mathbf{Z}'_{j_m} Q \mathbf{Z}_{j_m})^{-1} \mathbf{Z}'_{j_m} Q$ to be an $nT \times nT$ matrix. From Equation (4.33),

$$\mathbf{1}\hat{\Pi}_{k0,j_m} + QZ_{j_m}\hat{\Pi}_{kj_m} = P_m \hat{V}_m$$

$$f_m = P_m \left(X^*_k - F_{m-1} \right). \tag{4.35}$$

When $m = 0$, P_{j_0} is an $nT \times nT$ matrix of $\frac{1}{nT}$. Then the strong learner at each step m is

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

AIC is computed as

$$AIC_c(m) = \log(\hat{\sigma}_{k,m}^2) + \frac{1 + \text{trace}(B_m)/nT}{1 - (\text{trace}(B_m) + 2)/nT}, \tag{4.36}$$

where $\log(\hat{\sigma}_{k,m}^2) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(\hat{v}_{m,i} - c_m f_{m,it}\right)^2$. Then $\hat{M} = \arg\min_{m=1,\ldots,\bar{M}} AIC_c(m)$. However, AIC_c in Bhlmann (2006) does not provide a proper penalty in the panel data model. Hence, we propose a new stopping rule based on the modified AIC_c, denoted as BIC_c, following Bai (2003) that is

$$BIC_c(m) = \log(\hat{\sigma}_{k,m}^2) + \frac{1 + \text{trace}(B_m)\log(\frac{nT}{n+T})/\frac{nT}{n+T}}{1 - (\text{trace}(B_m) + 2)\log(\frac{nT}{n+T})/\frac{nT}{n+T}}. \tag{4.37}$$

Remark 2: In the simulation (not reported) using the same DGP as in the next section, we have compared the above two stopping rules in (4.36) and (4.37). We find that AIC_c in (2.17) selects too many instruments while BIC_c in (4.37) provides a good stopping rule for L_2Boosting in the panel data models to select only the relevant instruments.

4.4 Monte Carlo

We consider the following data generating progress (DGP)

$$y_{it} = x'_{it}\beta + \alpha_i + u_{it} \tag{4.38a}$$

$$x_{it} = \theta\ v_{it-1} + v_{it} \tag{4.38b}$$

$$u_{it} = \frac{\rho}{\sqrt{q}} v_{it} + \sqrt{1-\rho^2}\varepsilon_{it} \tag{4.38c}$$

Recall that x_{it} is an $q \times 1$ vector. The ε_{it} and all elements of v_{it} are i.i.d $N(0,\ 1)$ across i, t, for each elements of errors u_{it} and v_{it} having covariance $\frac{\rho}{\sqrt{q}}$, but all other correlation zero. α_i are i.i.d. $N(0,\ 1)$ independent of $\{x_{it},\ u_{it}\}$. The parameter ρ controls the extent of endogeneity of x_{it}. The endogenous variable vector x_{it} follows an invertible vector moving average process of order 1, VMA(1). For simplicity we consider the case when all elements of the $q \times 1$ vector MA parameter θ are the same and set at the value 0.3. The results are not quantitatively sensitive to the value of β so we set β to be zero.

Our goal is the consistent and efficient estimation of the structural parameter β. In the DGP, the variable x_{it} is endogenous following a VMA(1) process in (4.38b), which we approximate by the VAR(p) model of order $p \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. We consider the lagged variables of x_{it} as instruments, i.e., $z_{it} \equiv (x_{i,t-1} \ldots x_{i,t-p})$. The parameter θ controls the strength of the instruments z_{it} as they are taken from the lagged x_{it}. The number of instruments equals to $\ell = q \times p$. To mimic the situation in the empirical application for estimation of the US house prices using three endogenous variables in 49 states over 29 years in 1975-2003, we consider $n = 49$, $T = 29$, $q = 3$. We consider a range of

ρ on a 20-point grid on $[0, 0.975]$. We generated $2,000$ samples on each calculated $\hat{\beta}_{\text{FE-2SLS}}$, $\hat{\beta}_{\text{FE}}$, $\hat{\beta}_{\text{c}}$. To compare these three estimators, we calculate the median squared error of each estimator

$$R\left(\hat{\beta}\right) = median\left(\left(\hat{\beta}-\beta\right)'\left(\hat{\beta}-\beta\right)\right). \tag{4.39}$$

We present the results graphically. The same length of period ($t = 1, \ldots, 17$) is used for each subfigure in Figure 4.1. In Figure 4.1: (a) plots $R\left(\hat{\beta}\right)$ for $p = 1$, $\ell = 3$. This context is with just-identified instruments. Figure 4.1(a) shows that the combined estimator has lower $R\left(\hat{\beta}\right)$ than FE-2SLS, regardless of the degree of endogeneity. It also shows that the classical 2SLS estimator problem that its moments do not exist for the just-identified case, cf. Sawa (1969, 1972). Observe that $R\left(\hat{\beta}\right)$ of FE-2SLS in plot (a) is quite large compared to that in plot (b) for all degrees of endogeneity ρ. It shows that the number of instruments in (a) is too small and it would be important to increase the number of instruments. (b) plots the MSE for $p = 2$, $\ell = 6$. Now the model is over-identified and FE-2SLS is well behaved. Figure 4.1(b) shows that the combined estimator has similar $R\left(\hat{\beta}\right)$ to FE for the small values of ρ where FE has small $R\left(\hat{\beta}\right)$. The reduction in risk achieved by combined estimator is large unless ρ is large. (c) plots the $R\left(\hat{\beta}\right)$ for $p = 3$, $\ell = 9$. (d) plots the $R\left(\hat{\beta}\right)$ for $p = 4$, $\ell = 12$. (e) plots the median squared error for $p = 5$, $\ell = 15$. The four plots in (b), (c), (d), (e) look similar. The combined estimator achieves some reduction in $R\left(\hat{\beta}\right)$ relative to FE-2SLS for small values of ρ. However, when the number of instruments becomes larger in plot (f) with $p = 6$, $\ell = 18$, it starts to show that for large ρ FE-2SLS becomes biased towards FE. It becomes more apparent as ℓ becomes even larger as shown in the subsequent plots. Continuing to plot (g) for $p = 7$, $\ell = 21$, plot

(h) for $p = 8$, $\ell = 24$, plot (i) for $p = 9$, $\ell = 27$, plot (j) for $p = 10$, $\ell = 30$, plot (k) $p = 11$, $\ell = 33$, and plot (l) for $p = 12$, $\ell = 36$, it is easy to see that the FE-2SLS is biased and the bias in FE-2SLS tends to get worse as more instruments are used.

To fix this bias problem in FE-2SLS, we use the SCAD [1] and L_2Boosting for the selection of instruments, which makes the FE-2SLS estimator and the combined estimator more robust and restore its consistency when there are many instruments. To demonstrate this, we consider the setup in the last plot of Figure 4.1 with $p = 12$, $\ell = 36$. It is the most apparent case for the inconsistency due to using too many instrument variables. In Figure 4.2 we zoom in the plot (l) of Figure 4.1 and add two SCAD regularized estimators (FE-2SLS-SCAD and Combined-SCAD), and in Figure 4.3, we add two boosting regularized estimators (FE-2SLS-Boosting and Combined-Boosting). Notice the vertical scale of Figure 4.2 and Figure 4.3 are between 0 and 0.2. By using the SCAD to regularize a panel data model with many instruments, the FE-2SLS-SCAD estimator restores its consistency. The FE-2SLS-SCAD is the post-selection FE-2SLS estimator. The Combined-SCAD estimator is the combined estimator of FE and FE-2SLS-SCAD estimators. The risk of FE-2SLS-SCAD estimator and the Combined-SCAD estimator are significantly reduced when compared to the risk of the FE-2SLS estimator. When the endogeneity is weak (small values of ρ), the Combined-SCAD estimator (cyan-colored long-dashed) dominates the FE-2SLS-SCAD (green-colored dotted). As the endogeneity gets stronger, the combining weight in the Combined-SCAD goes toward to the FE-2SLS-SCAD.

[1] A third party MATLAB toolbox by Zhou, Armagan and Dunson (2012) and Zhou and Gaines (2017) is used for SCAD.

Similar pattern also shows for the FE-2SLS-Boosting estimator and Combined-Boosting estimator in Figure 4.3. The risk of both estimators are significantly reduced when compared to the risk of the estimators without instrument selection. The Combined-Boosting estimator dominates the FE-2SLS-Boosting estimator when the endogeneity is week. However, the weight in the Combined-Boosting goes toward to the FE-2SLS-Boosting as the endogeneity gets stronger.

Remark 3: In the presence of many instruments, the combined estimator may behave poorly, and Theorem 2 may not hold for moderate to large values of ρ. Theorem 2 says the combined estimator is always better than the FE-2SLS estimator in the asymptotic risk. Figures 4.1(g)-(l), Figure 4.2, and Figure 4.3 show, however, that the theorem does not hold when p is large. This is because Theorem 2 holds only when FE-2SLS is consistent. When p is large, ℓ is large, FE-2SLS is inconsistent and therefore Theorem 2 does not hold when there are many instruments. After selecting instruments through SCAD or L_2Boosting, we reduce the dimension of instruments, and we only use relevant instruments to estimate the FE-2SLS estimator. This will restore Theorem 2 and ensure the consistency of the FE-2SLS-SCAD (FE-2SLS-Boosting) and the Combined-SCAD (Combined-Boosting) estimators even with many instruments.

Remark 4: Since the endogenous variables x_{it} are generated following a VMA(1) process in (4.38b), the relevancy of instruments to x_{it} is dependent on the strength of the parameter θ. When the value of θ is very low (say, $\theta = 0.1$), the relevancy of all instruments becomes weak. Thus, the FE-2SLS, FE-2SLS-SCAD and FE-2SLS-Boosting are inconsistent because of the weak instruments. On the other hand, if the value of θ is very high (say, $\theta = 0.5$),

all instruments from the lagged x_{it} are strongly relevant. Then the sparsity assumption required for SCAD and L_2Boosting may not hold. Thus, SCAD and L_2Boosting may fail to 'select' the relevant instruments since all instruments are relevant.

4.5 Estimation of House Price Panel Data Model in US States

Real house prices can vary between States because real incomes differ, they can also differ because of scarcity of other idiosyncratic factors. The effects of common shocks on house prices such as changes in interest rates, could also differ across States. Holly, Pesaran, and Yamagata (HPY 2010) examine the extent to which real house prices at the State level are driven by fundamentals such as real per capita disposable income, as well as by common shocks. Baltagi and Li (2014) replicate the results of HPY, using a slightly different data set. They extend the period of study to 2011, incorporating the information reflected by the housing market crash in 2007. Using housing price indexes for 381 metropolitan statistical areas and over the period 1975–2011, they find that the HPY results are fairly robust. As noted in Baltagi and Li (2014, p. 515), "The US housing price indexes, published by the Federal Housing Finance Agency (FHFA), ran up by almost 40 percent from January 2003 to June 2006, followed by a 28 percent drop, unprecedented in US history", the extended period from 2004-2011 covers quite unusual housing market data of boom, crash, and slow recovery. Therefore in this chapter, we will use the original HPY data for 1975-2003. In this section, we use the panel of 49 states over the 29 year period of 1975–2003. We consider

the following panel data model for US States

$$p_{it} = \beta_0 + \beta_y y_{it} + \beta_g g_{it} + \beta_c c_{it} + \alpha_i + u_{it} \tag{4.40}$$

where $i = 1, \ldots, 49, t = 1, \ldots, 29$, p_{it} is the logarithm of the real price of housing in the ith State during year t, and y_{it} is the logarithm of the real per capita personal disposable income. The net cost of borrowing defined by $c_{it} = r_{it} - \Delta p_{it}$, where r_{it} represents the long-term real interest rate and g_{it} represents the population growth rate. The state-specific effects can be treated as the endowment of climate, location and culture. A more detailed description can be found in HPY. We would expect a rise in c_{it} to be associated with a fall in the price income ratio, and hence a negative coefficient for c_{it}. The effect of population growth on real house prices is expected to be positive.

Let $x_{it} \equiv (y_{it}\ g_{it}\ c_{it})'$. We consider the lagged variables of x_{it} as instruments, i.e., $z_{it} \equiv (x_{i,t-1} \ldots x_{i,t-p})$. Hence we write an VAR($p$) model for x_{it} in State i as

$$\begin{aligned} x_{it} &= \Pi' z_{it} + v_{it} \\ x_{it} &= A_1 x_{i,t-1} + \cdots + A_p x_{i,t-p} + v_{it}, \end{aligned} \tag{4.41}$$

where $\Pi' = (A_1 \ldots A_p)$ and $z_{it} = (x_{i,t-1} \ldots x_{i,t-p})$. Note that if the lag order $p = 6$ in VAR(p) is used, the number of instruments is $\ell = q \times p = 18$. In Table 4.1, for all values of $p = 1, \ldots, 6$, we discard the first 6 years of the sample of 1975-1980 and the panel data model is estimated for the remaining 23 years for 1981-2003, so that the parameter estimates

using different lags in VAR(p) model are estimated over the same sample period and the results can be comparable. SCAD and L_2Boosting are used for $p = 6$.

Table 4.1 shows that the income elasticity of real house prices for the combined estimator is significant and positive but widely changing in the range from 0.24 to 0.52 depending on the number of instruments used. The estimates of the coefficients on the population growth and the net cost of borrowing also exhibit wild variation over different lag orders p and different number of instruments ℓ. By computing the Hausman statistics with different lag orders p, we find there is strong endogeneity in all variables of x_{it}. The Hausman statistics ranges from 127.0993 ($p = 1$) to 29.4676 ($p = 5$), so exogeneity is rejected at one percent significant level for different value of p. Due to the presence of severe endogeneity in the model, the FE estimator is different from the FE-2SLS, FE-2SLS-SCAD, and FE-2SLS-Boosting estimators. In addition, because of the large value in the Hausman statistics, the combined estimators are weighted heavily toward to the FE-2SLS estimators. Hence, the FE-2SLS and combined estimators are similar in all cases.

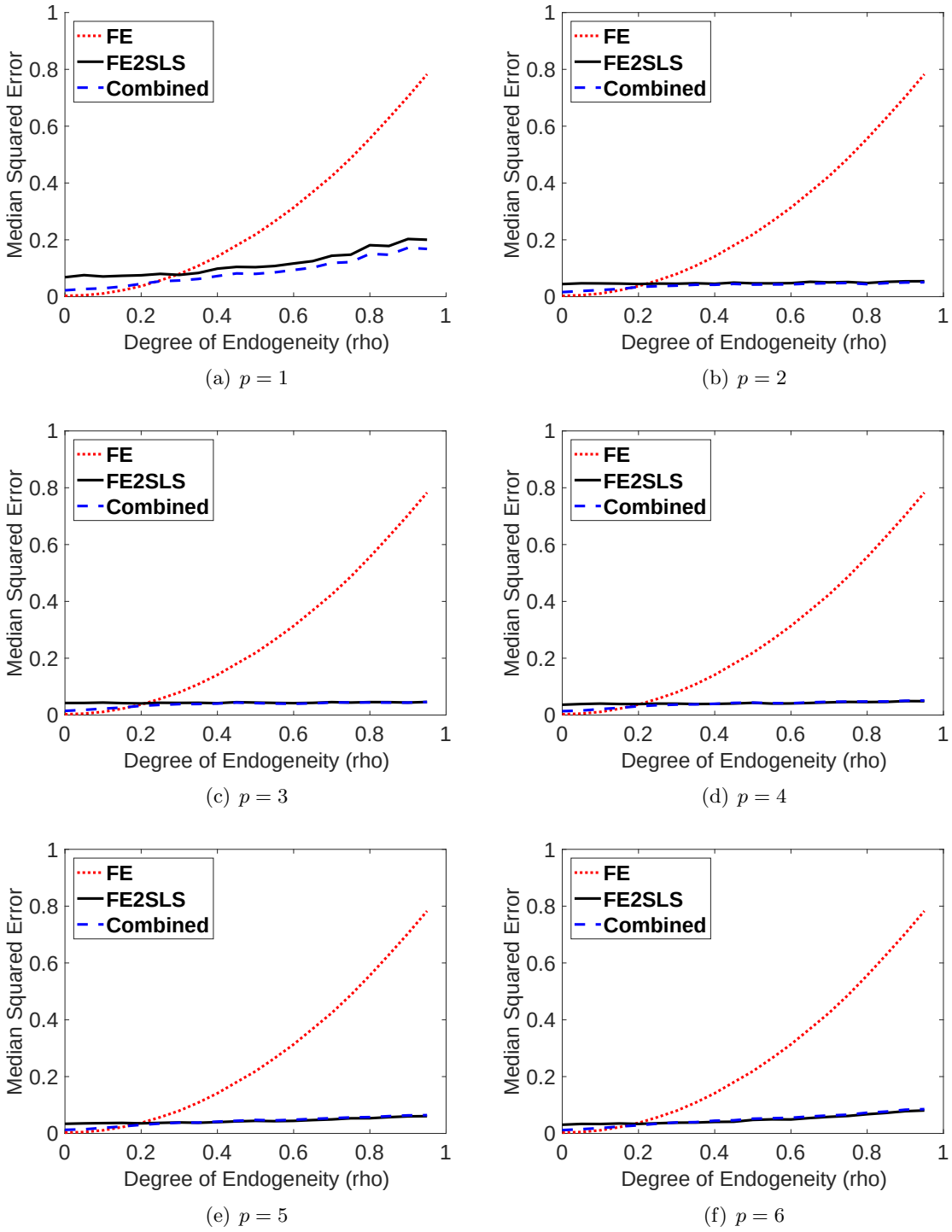
The standard error of estimators are computed using bootstrap. As the order p increases, the standard error of estimators is decreasing. Although both FE-2SLS, FE-2SLS-SCAD and FE-2SLS-Boosting estimators are using lags up to 6, we find that FE-2SLS-SCAD estimators are having smaller standard error than FE-2SLS estimators on the coefficients of the population growth and the net cost of borrowing, and larger standard error on the income elasticity of real house price. But FE-2SLS-Boosting estimators are having larger standard error than FE-2SLS estimators on the coefficients of the population growth and the net cost of borrowing, and smaller standard error on the income elasticity

of real house price. In addition, when $p = 6$, $R\left(\hat{\beta}\right)$ for FE-2SLS estimators is 0.1607, and for combined estimators is 0.1569. After applying SCAD, $R\left(\hat{\beta}\right)$ has been reduced to 0.1334 for FE-2SLS-SCAD estimators and 0.1265 for Combined-SCAD estimators. The better performance of SCAD estimators implies that the instrument variable selection is necessary when the number of instruments is large.

It is interesting to note that the combined estimates have smaller standard errors than FE-2SLS although the difference is often small. It is due to the combination with FE which is biased but has smaller variance. The gains from the combined estimator arise from this bias and variance trade-off.

4.6 Conclusions

When the regressors are endogenous due to simultaneity or measurement errors, the fixed effect estimator is inconsistent and FE-2SLS can be used. However the FE-2SLS estimator is very sensitive to the number of selected instruments and can be inconsistent when many instruments are used even when all the instruments are relevant and valid. In this chapter, we have examined the regularized FE-2SLS estimator and the regularized combined estimator under the presence of many instruments. It is clearly demonstrated that the selection of relevant instruments using regularization methods is important, which restores the consistency of the FE-2SLS estimator and the robustness of the combined estimator.

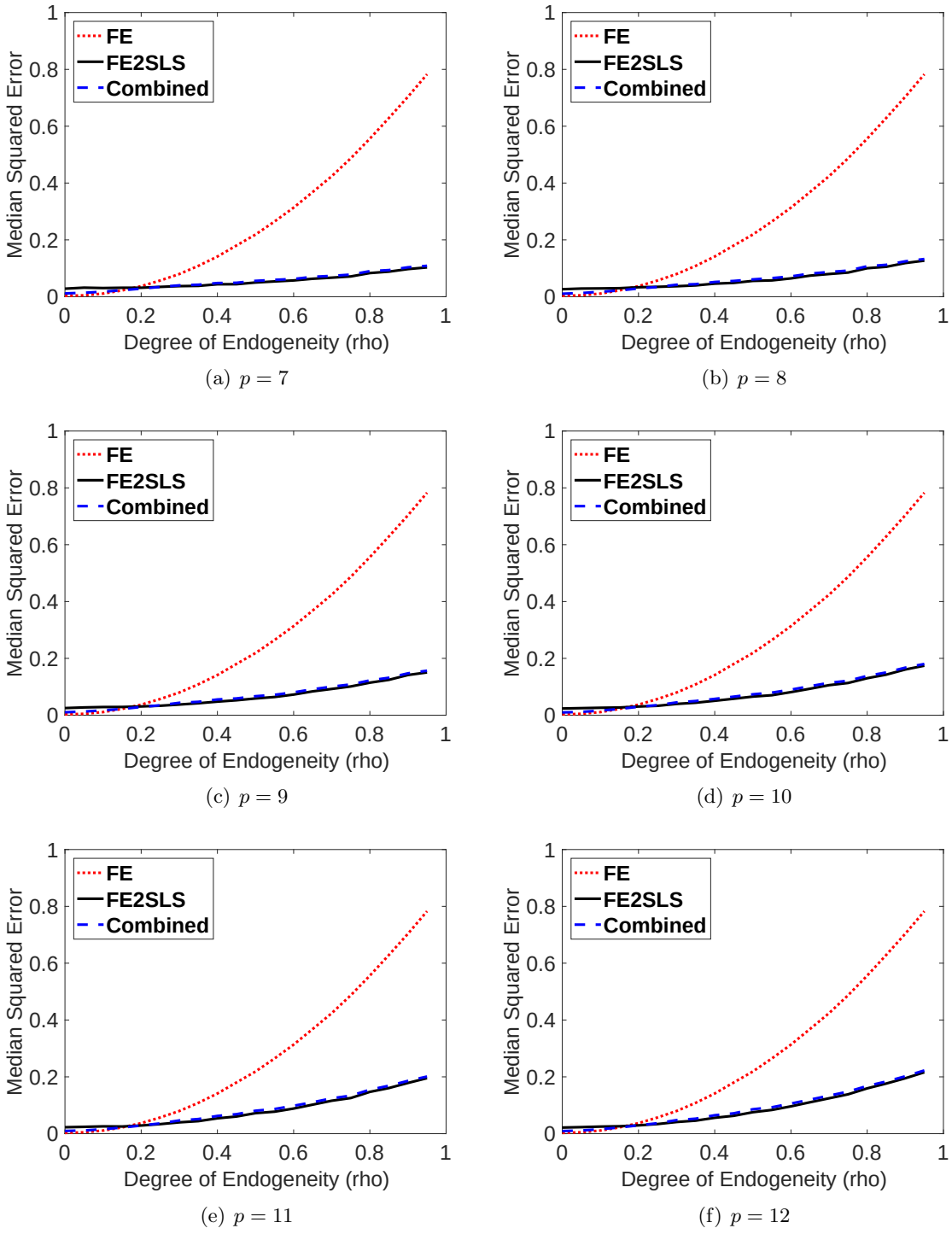


(a) $p = 1$ (b) $p = 2$

(c) $p = 3$ (d) $p = 4$

(e) $p = 5$ (f) $p = 6$

Figure 4.1: Median Squared Error of FE, FE-2SLS and Combined Estimators, $n = 49$, $T = 29$, $q = 3$.



(a) $p = 7$ (b) $p = 8$

(c) $p = 9$ (d) $p = 10$

(e) $p = 11$ (f) $p = 12$

Figure 4.1: Median Squared Error of FE, FE-2SLS and Combined Estimators, $n = 49$, $T = 29$, $q = 3$.

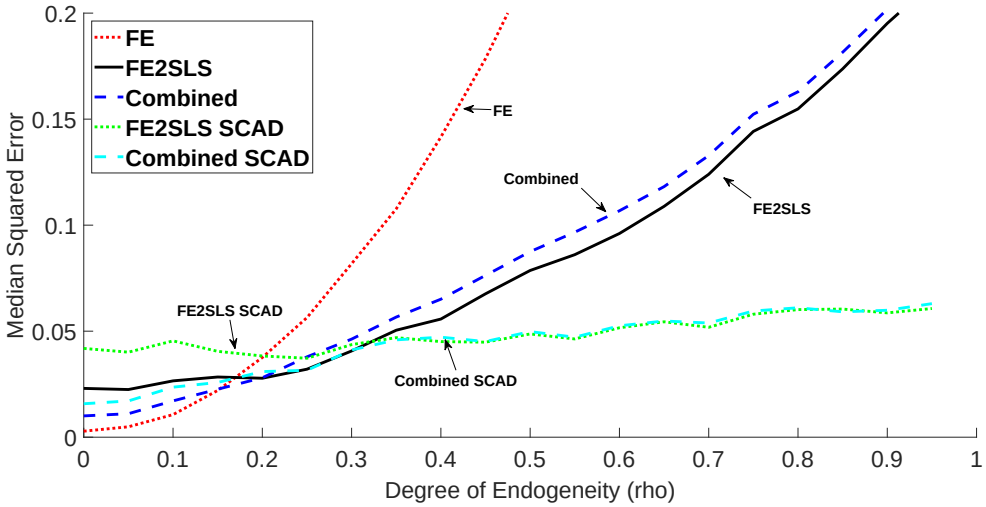


Figure 4.2: Median Squared Error of FE, FE-2SLS, Combined, F2SLS-SCAD and Combined SCAD Estimators, $n = 49$, $T = 29$, $q = 3$, $p = 12$.

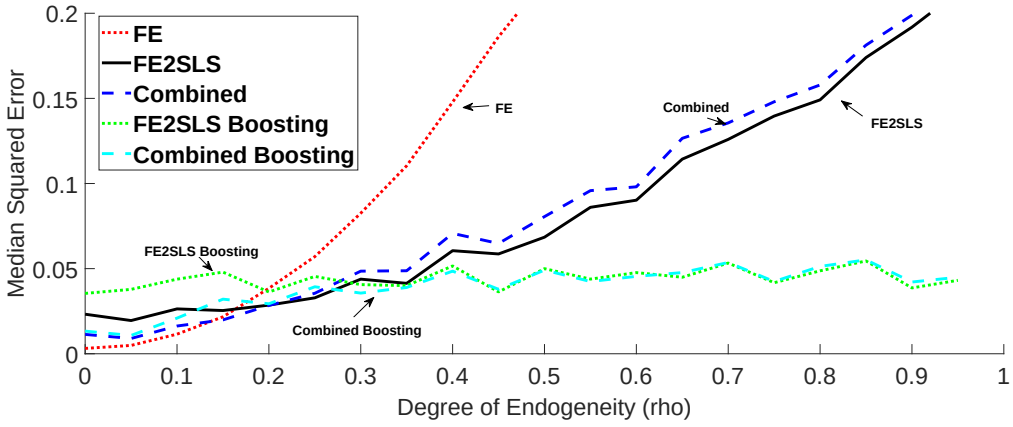


Figure 4.3: Median Squared Error of FE, FE-2SLS, Combined, F2SLS-Boosting and Combined Boosting Estimators, $n = 49$, $T = 29$, $q = 3$, $p = 12$.

Table 4.1: Table 1. Economics of Real House Prices for 49 U.S. States, 1975–2003

		p	$\hat{\beta}_y$	$\hat{\beta}_g$	$\hat{\beta}_c$
(a)	FE		0.4726	0.1520	-0.2286
			(0.0286)	(0.4156)	(0.0694)
	FE-2SLS	1	0.2435	2.5257	-1.0777
(b)			(0.0416)	(0.6214)	(0.1462)
	Combined	1	0.2453	2.5070	-1.1071
			(0.0414)	(0.6184)	(0.1456)
	FE-2SLS	2	0.3178	0.6890	-0.8590
(c)			(0.0420)	(0.6903)	(0.1219)
	Combined	2	0.3204	0.6802	-0.8488
			(0.0416)	(0.6821)	(0.1216)
	FE-2SLS	3	0.4144	1.4687	-0.4588
(d)			(0.0361)	(0.6471)	(0.1132)
	Combined	3	0.4157	1.4379	-0.4534
			(0.0356)	(0.6443)	(0.1103)
	FE-2SLS	4	0.4546	1.1599	-0.3319
(e)			(0.0370)	(0.5442)	(0.1075)
	Combined	4	0.4552	1.1257	-0.3284
			(0.0363)	(0.5413)	(0.1062)
	FE-2SLS	5	0.4826	1.1627	-0.2445
(f)			(0.0316)	(0.5217)	(0.0961)
	Combined	5	0.4823	1.1290	-0.2440
			(0.0305)	(0.5119)	(0.0912)
	FE-2SLS	6	0.5037	1.1994	-0.1795
(g)			(0.0307)	(0.5512)	(0.1112)
	Combined	6	0.5028	1.1681	-0.1809
			(0.0302)	(0.5429)	(0.1086)
	FE-2SLS-SCAD	6	0.5055	1.2125	-0.1733
(h)			(0.0364)	(0.4990)	(0.1044)
	Combined SCAD	6	0.5045	1.1807	-0.1750
			(0.0357)	(0.4850)	(0.1008)
	FE-2SLS-Boosting	6	0.5180	2.2028	-0.0924
(i)			(0.0357)	(0.5330)	(0.1183)
	Combined Boosting	6	0.5170	2.1564	-0.0955
			(0.0344)	(0.5309)	(0.1130)

Note: Standard errors are in parentheses. "p" is the order of VAR(p) for the reduced form equation. (a) reports FE, (b) to (g) reports FE-2SLS and the Combined estimator of FE and FE-2SLS when lag $p = 1, \ldots, 6$, (h) reports FE-2SLS-SCAD estimator which is the post-selection FE-2SLS estimator using the instruments selected by SCAD, and Combined SCAD combines FE and FE-2SLS-SCAD when lag $p = 6$, and (i) reports FE-2SLS-Boosting estimator which is the post-selection FE-2SLS estimator using the instruments selected by L_2Boosting, and Combined Boosting combines FE and FE-2SLS-Boosting when lag $p = 6$.

Chapter 5

Conclusions

When the number of instruments is large relative to the number of observations, a regularization method need to be imposed for both cross-sectional data and panel data. Under the cross-sectional data, we relax both the validity and relevancy assumption on the instruments. We propose Double Boosting (DB) selection that will consistently select both valid and relevant instruments. In Chapter 1, we show that DB-GMM will give smaller bias and MSE especially when the function form is unknown and nonlinear. In Chapter 2, we found that the robustness of DB-GMM is highly correlated with the sieve functions that are used to approximate the true functional form of the endogenous regressors. By applying the multiple-layer neural network, we find that DB-NN, where sieve instruments are generated through the neural network, is robust even when the true functional form of the endogenous regressors are very complicate. In the empirical application of the automobile market, the price elasticity of demand estimated by both DB-GMM (in Chapter 1) and DB-NN (in Chapter 2) are very elastic.

In Chapter 3, we extend the stein-like combined estimator by Hansen (2017) to the panel data models, where the regressors are endogenous. We have proved that the asymptotic risk of the combined estimator is strictly smaller than FE-2SLS when FE-2SLS is consistent. Both the analytic and simulation results show that FE-2SLS is inconsistent when the number of instruments is large. We extend the two regularization methods under cross-section data models, SCAD and L_2Boosting, to the panel data models. We also propose an new stopping rule for L_2Boosting to ensure the consistency of the FE-2SLS estimator. The simulation results confirm that by applying the regularization methods, FE-2SLS is consistent even when the number of instruments is large. Thus the asymptotic property of the combined estimator carries over to the finite sample case even when there are many instruments.

Bibliography

[1] Amemiya, T. and MaCurdy, T. E. (1986). "Instrumental-variable estimation of an error-components model". *Econometrica* 54(4), 869-880.

[2] Bai, J. (2003). "Inferential Theory for Factor Models of Large Dimensions ". *Econometrica* 71(1), 135-171.

[3] 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.

[4] Baltagi, B. (2008). *Econometric Analysis of Panel Data*. John Wiley & Sons.

[5] Baltagi, B. H. and Li, J. (2014). "Further evidence on the spatio-temporal model of house prices in the united states." *Journal of Applied Econometrics*, 29(3), 515-522.

[6] Bekker, P. (1994). "Approximations to the Distributions of Instrumental Variable Estimators". *Econometrica* 62(3), 657-681.

[7] 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, 2369-2429.

[8] Belloni, A. and Chernozhukov, V. (2013). "Least Squares After Model Selection in High-dimensional Sparse Models", *Bernoulli* 19(2), 521-547.

[9] Berry, S., Levinsohn, J., and Pakes, A. (1995). "Automobile Prices in Market Equilibrium". *Econometrica* 63(4), 841 - 890.

[10] Breusch, T. S., Mizon, G. E., and Schmidt, P. (1989). Efficient estimation using panel data. *Econometrica* 57(3), 695-700.

[11] Bhlmann, P. (2006). "Boosting for High-dimensional Linear Models". *Annals of Statistics* 34(2), 559-583.

[12] Caner, M. (2009). "Lasso-type GMM Estimator". *Econometric Theory* 25(1), 270-290.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] DiTraglia, F. (2016). "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM". *Journal of Econometrics* 195(2),2 187-208.

[19] Donald., S., Imbens, G., and Newey, W. (2009). "Choosing Instrumental Variables in Conditional Moment Restriction Models ". *Journal of Econometrics* 152(1), 28-36.

[20] Fan, J. and Li, R. (2001). "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties". *Journal of the American Statistical Association* 96(456): 1348-1360

[21] Fan, J. and Liao, Y. (2014). "Endogeneity in High Dimensions". *Annals of Statistics* 42(3), 872-917.

[22] 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.

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

[24] Hansen, B.E. (2017). "Stein-Like 2SLS Estimator." *Econometric Reviews* 36, 840-852.

[25] Hausman, J.A. (1978). "Specification tests in econometrics." *Econometrica* 46(6), 1251-1271.

[26] Hausman, J. A., and Taylor, W. E. (1981). "Panel data and unobservable individual effects". *Journal of the Econometrics*, 1377-1398.

[27] Holly, S., Pesaran, M. H., and Yamagata, T. (2010). "A spatio-temporal model of house prices in the USA." *Journal of Econometrics* 158(1), 160-173.

[28] Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2016). "Counterfactual Prediction with Deep Instrumental Variables Networks". arXiv:1612.09596.

[29] Liao, Z. (2013). "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection". *Econometric Theory* 29, 857-904.

[30] Meinshausen, N. (2007). "Relaxed Lasso". *Computational Statistics & Data Analysis* 52(1), 374-393.

[31] Ng, S. and Bai, J. (2008). "Selecting Instrumental Variables in a Data Rich Environment". *Journal of Time Series Econometrics* 1(1), 1-34.

[32] Phillips, P. (1989). "Partially Identified Econometric Models". *Econometric Theory* 5(2), 181-240.

[33] Sawa, T. (1969). "The Exact Sampling Distributions of Ordinary Least Squares and Two Stage Least Squares Estimates." *Journal of the American Statistical Association* 64, 923-980.

[34] Sawa, T. (1972). "Finite Sample Properties of k-Class Estimators." *Econometrica* 40, 653-680.

[35] Staiger, D. and Stock, J (1997). "Instrumental Variables Regression with Weak Instruments". *Econometrica* 65(3), 557-586.

[36] Stein, C. (1956). "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." *Proceedings of the Third Berkeley symposium on mathematical statistics and probability* Vol.1, 399.

[37] Stock, J.H. and Wright, J. (2000). "GMM with Weak Identification". *Econometrica* 68(5), 1055-1096.

[38] White, H. (2006). "Approximate Nonlinear Forecasting Methods". *Handbook of Economic Forecasting* 1, 460 - 512.

[39] Zhou, H., Armagan, A. and Dunson, D (2012). "Path Following and Empirical Bayes Model Selection for Sparse Regressions". arXiv:1201.3528

[40] Zhou, H. and Gaines, B. (2017). Matlab SparseReg Toolbox Version 1.0.0. http://huazhou.github.io/SparseReg/

Appendix A

Appendix for Chapter 1

A.1 Proof of Theorem 1

Under the approximately sparse model in (2.4), the conditional quadratic mean of regression error using L_2Boosting is,

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

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

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

Hence,

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

□

A.2 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

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

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

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:

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

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,

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

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,

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

Then

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

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

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

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

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

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:

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

We summarize the above results in Table 2, which adds the orders of ω_j to Table 1. Because $\alpha_j > 0$ for $Z_j \in \mathcal{B}_0$, we have $o_p(n^{r_1(2\alpha_j-1)}) \leq \min\left\{o_p(n^{r_1(2\alpha_j-1)}), O_p\left(n^{r_2-r_1}\right)\right\}$.

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. □

Appendix B

Appendix for Chapter 3

B.1 Proof of Theorem 2:

Noting that $\sqrt{n}\left(\hat{\beta}_{\text{FE-2SLS}}-\beta\right) \xrightarrow{d} G_2'\xi \sim N\left(0,\ V_2\right)$, then

$$R\left(\hat{\beta}_{\text{FE-2SLS}}\right) = E\left(\xi' G_2' W G_2' \xi\right) = \operatorname{tr}\left(W V_2\right).$$

Define Ψ^* as a random variable without positive part trimming

$$\Psi^* = G_2'\xi - \left(\frac{\tau}{(h+\xi)' B (h+\xi)}\right) G'\left(h+\xi\right).$$

Then using the fact that the pointwise quadric risk of Ψ is strictly smaller than that of Ψ^*, then we have

$$R\left(\hat{\beta}_{\text{c}}\right) = E\left(\Psi' W \Psi\right) < E\left(\Psi^{*\prime} W \Psi^*\right).$$

We can calculate that

$$E\left(\Psi^{*\prime}W\Psi^{*}\right)=R\left(\hat{\beta}_{\text{FE-2SLS}}\right)+\tau^{2}E\left(\frac{(h+\xi)^{\prime}GWG^{\prime}(h+\xi)}{\left((h+\xi)^{\prime}B(h+\xi)\right)^{2}}\right)-2\tau E\left(\frac{(h+\xi)^{\prime}GWG_{2}^{\prime}\xi}{(h+\xi)^{\prime}B(h+\xi)}\right).$$

If $Z\sim N(0,V)$ is $q\times 1$, K is $q\times q$, and $\eta(x)$ is absolutely continuous, then by Stein's Lemma

$$E\left(\eta(Z+h)^{\prime}KZ\right)=E\ \text{tr}\left(\frac{\partial}{\partial x}\eta(Z+h)^{\prime}KV\right),$$

$\eta(x)=x/\left(x^{\prime}Bx\right)$, and

$$\frac{\partial}{\partial x}\eta(x)=\frac{1}{x^{\prime}Bx}I-\frac{2}{\left(x^{\prime}Bx\right)^{2}}Bxx^{\prime}.$$

Therefore,

$$\begin{aligned}
&E\left(\frac{(h+\xi)^{\prime}GWG_{2}^{\prime}\xi}{(h+\xi)^{\prime}B(h+\xi)}\right)\\
&=E\text{tr}\left(\frac{GWG_{2}^{\prime}V}{(h+\xi)^{\prime}B(h+\xi)}-\frac{2GWG_{2}^{\prime}V}{\left((h+\xi)^{\prime}B(h+\xi)\right)^{2}}B(h+\xi)(h+\xi)^{\prime}\right)\\
&=E\left(\frac{\text{tr}\left(GWG_{2}^{\prime}V\right)}{(h+\xi)^{\prime}B(h+\xi)}\right)-2E\text{tr}\left(\frac{GWG_{2}^{\prime}V}{\left((h+\xi)^{\prime}B(h+\xi)\right)^{2}}B(h+\xi)(h+\xi)^{\prime}\right).
\end{aligned}$$

Since

$$GWG_{2}^{\prime}V=WG_{2}^{\prime}VG=W\left(V_{2}-V_{1}\right),$$

$$GWG_{2}^{\prime}VB=GWG_{2}^{\prime}VG\left(V_{2}-V_{1}\right)^{-1}G^{\prime}=GWG^{\prime},$$

we have

$$E\text{tr}\left(\frac{GWG_2'V}{\left((h+\xi)'B(h+\xi)\right)^2}B(h+\xi)(h+\xi)'\right)=E\text{tr}\left(\frac{(h+\xi)'GWG'(h+\xi)}{\left((h+\xi)'B(h+\xi)\right)^2}\right).$$

Thus

$$\begin{aligned}E\left(\psi^{*\prime}W\psi^{*}\right)=R\left(\hat{\beta}_{\text{FE-2SLS}}\right)&+\tau^2E\left(\frac{(h+\xi)'GWG'(h+\xi)}{\left((h+\xi)'B(h+\xi)\right)^2}\right)\\&+4\tau E\text{tr}\left(\frac{(h+\xi)'GWG'(h+\xi)}{\left((h+\xi)'B(h+\xi)\right)^2}\right)-2\tau E\text{tr}\left(\frac{(W(V_2-V_1))}{(h+\xi)'B(h+\xi)}\right).\end{aligned}$$

Define

$$B_1=(V_2-V_1)^{-\frac{1}{2}}G'$$

$$A^*=(V_2-V_1)^{\frac{1}{2}}W(V_2-V_1)^{\frac{1}{2}}.$$

Note that $GWG_2'VP=GWG'=B_1'A^*B_1$, $B_1'B_1=B$.

Using the inequality $b'ab\leq(b'b)\lambda_{\max}(a)$ for symmetric a, and let

$$\lambda_{\max}(a)=\lambda_{\max}(W(V_2-V_1))=\lambda_1.$$

Then

$$\begin{aligned}\text{tr}\left(B(h+\xi)(h+\xi)'GWG_2'V\right)&=(h+\xi)'B_1'A^*B_1(h+\xi)\\&\leq(h+\xi)'B(h+\xi)\lambda_1.\end{aligned}\tag{B.1}$$

Using equation (B.1) and Jensen's inequality, we have

$$\begin{aligned} E\left(\psi^{*\prime}W\psi^{*}\right) &\leq R\left(\hat{\beta}_{\text{FE-2SLS}}\right)+\left(\tau^{2}+4\tau\right)E\left(\frac{\lambda_{1}}{(h+\xi)^{\prime}B(h+\xi)}\right) \\ &\quad -2\tau E\text{tr}\left(\frac{\left(W\left(V_{2}-V_{1}\right)\right)}{(h+\xi)^{\prime}B(h+\xi)}\right) \\ &= R\left(\hat{\beta}_{\text{FE-2SLS}}\right)-E\left(\frac{\tau\left(2\left(\text{tr}\left(W\left(V_{2}-V_{1}\right)\right)-2\lambda_{1}\right)-\tau\lambda_{1}\right)}{(h+\xi)^{\prime}B(h+\xi)}\right) \\ &\leq R\left(\hat{\beta}_{\text{FE-2SLS}}\right)-\frac{\tau\left(2\left(\text{tr}\left(W\left(V_{2}-V_{1}\right)\right)-2\lambda_{1}\right)-\tau\lambda_{1}\right)}{E\left((h+\xi)^{\prime}B(h+\xi)\right)}. \end{aligned} \tag{B.2}$$

Since $\text{tr}(BV)=\text{tr}\Big(G\left(V_{2}-V_{1}\right)^{-1}G^{\prime}V\Big)=q$, we have

$$\begin{aligned} E\left((h+\xi)^{\prime}B(h+\xi)\right) &= h^{\prime}Bh+\text{tr}\left(BV\right) \\ &= \sigma_{u}^{-4}\delta^{\prime}\text{tr}\left(Q\Sigma\right)V_{1}\left(V_{2}-V_{1}\right)^{-1}V_{1}\text{tr}\left(Q\Sigma\right)\delta+q. \end{aligned}$$

Substituted into (B.2) we have

$$\begin{aligned} R\left(\hat{\beta}_{\text{c}}\right) &< R\left(\hat{\beta}_{\text{FE-2SLS}}\right)-\frac{\tau\left(2\left(\text{tr}\left(W\left(V_{2}-V_{1}\right)\right)-2\lambda_{1}\right)-\tau\lambda_{1}\right)}{\sigma_{u}^{-4}\delta^{\prime}\text{tr}\left(Q\Sigma\right)V_{1}\left(V_{2}-V_{1}\right)^{-1}V_{1}\text{tr}\left(Q\Sigma\right)\delta+q} \\ &= R\left(\hat{\beta}_{\text{FE-2SLS}}\right)-\frac{\tau\lambda_{1}\left(2\left(d-2\right)-\tau\right)}{\sigma_{u}^{-4}\delta^{\prime}\text{tr}\left(Q\Sigma\right)V_{1}\left(V_{2}-V_{1}\right)^{-1}V_{1}\text{tr}\left(Q\Sigma\right)\delta+q} \end{aligned}$$

with $0<\tau\leq 2\left(\frac{\text{tr}(W(V_{2}-V_{1}))}{\lambda_{1}}-2\right)$.

B.2 Proof of the inconsistency of FE-2SLS

In this appendix we show the inconsistency of the FE and the FE-2SLS estimators in the panel data models.

First, consider the FE estimator under endogeneity. From eq (4.4), the $\hat{\beta}_{\text{FE}}$ is obtained as

$$\hat{\beta}_{\text{FE}} = \left(X'QX\right)^{-1} X'Qy = \beta + \left(X'QX\right)^{-1} X'Qu,$$

Premultiplying X by Q,

$$QX = QZ\Pi + Qv. \tag{B.3}$$

Let $y^* = Qy$, $X^* = QX$, $Z^* = QZ$, $u^* = Qu$, and $v^* = Qv$, the $\hat{\beta}_{\text{FE}}$ can be rewritten as

$$\hat{\beta}_{\text{FE}} = \beta + \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}u^*.$$

Define $\xi = [Qu, Qv]$, and the variance-covariance matrix of ξ is

$$\Omega = \begin{bmatrix} \Omega_u & \Omega_{uv} \\ \Omega_{vu} & \Omega_v \end{bmatrix}. \tag{B.4}$$

Then

$$\begin{aligned} E\left(\frac{1}{n}X'Qu\right) &= \frac{1}{n}E\left(\Pi'Z^{*\prime}u^* + v^{*\prime}u^*\right) \\ &= \frac{1}{n}E\left(v^{*\prime}u^*\right) = \Omega_{uv}, \end{aligned}$$

and

$$
\begin{aligned}
E\left[\frac{1}{n}X'QX\right] &= \frac{1}{n}E\left[(Z\Pi+v)'\,Q\,(Z\Pi+v)\right] \\
&= \frac{1}{n}E\left[\Pi'Z^{*\prime}Z^{*}\Pi+\Pi'Z^{*\prime}v^{*}+v^{*\prime}Z^{*}\Pi+v^{*\prime}v^{*}\right] \\
&= \frac{1}{n}\Pi'E\left(Z^{*\prime}Z^{*}\right)\Pi+\Omega_{v}.
\end{aligned}
$$

If X is endogenous, $\Omega_{uv} \neq 0$. Then the OLS estimator $\hat{\beta}_{\mathrm{FE}}$ is inconsistent because

$$
\hat{\beta}_{\mathrm{FE}}-\beta \xrightarrow{p} \left[\frac{1}{n}\Pi'E\left(Z^{*\prime}Z^{*}\right)\Pi+\Omega_{v}\right]^{-1}\Omega_{uv} \neq 0. \tag{B.5}
$$

Next, we show that the FE-2SLS estimator is inconsistent when there are too many instruments. From (4.7), the $\hat{\beta}_{\text{FE-2SLS}}$ is

$$
\hat{\beta}_{\text{FE-2SLS}} = \left(X'H_{Z}X\right)^{-1}X'H_{Z}y = \beta+\left(X^{*\prime}H_{Z}X^{*}\right)^{-1}X^{*\prime}H_{Z}u^{*}, \tag{B.6}
$$

where $H_{Z} = QZ\left(Z'QZ\right)^{-1}Z'Q = Z^{*}\left(Z^{*\prime}Z^{*}\right)^{-1}Z^{*\prime}$. Let the idempotent matrix H_{Z} decompose into $H_{Z} = \Psi\Lambda\Psi$, where $\Lambda = \mathrm{diag}(I_{\ell}, 0_{nT-\ell})$, and Ψ is a $nT \times nT$ matrix that has orthonormal properties. There exist an $nT \times (q+1)$ matrix $R = \Psi'\xi\Omega^{-1/2}$ that satisfies

$$
E(R'R) = E(\Omega^{-1/2}\xi'\Psi\Psi'\xi\Omega^{-1/2}) = E(\Omega^{-1/2}\xi'\xi\Omega^{-1/2}) = I_{k+1}. \tag{B.7}
$$

We partition $R = [r_1, r_2]'$, where r_1 is $\ell \times q + 1$. Then,

$$
\begin{aligned}
E\left(\frac{1}{n}\xi' H_Z \xi\right) &= \frac{1}{n} E\left(\xi' \Psi \Lambda \Psi \xi\right) \\
&= \frac{1}{n}\Omega^{1/2} E\left(\Omega^{-1/2}\xi' \Psi \Lambda \Psi \xi \Omega^{-1/2}\right)\Omega^{1/2} \\
&= \frac{1}{n}\Omega^{1/2} E\left(R' \Lambda R\right)\Omega^{1/2} \\
&= \frac{1}{n}\Omega^{1/2} E\left([r_1', r_2'] \begin{bmatrix} I_\ell & 0 \\ 0 & 0_{nT-\ell} \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}\right)\Omega^{1/2} \\
&= \frac{1}{n}\Omega^{1/2} E\left(r_1' r_1\right)\Omega^{1/2} \\
&= \frac{\ell}{n}\Omega,
\end{aligned}
\tag{B.8}
$$

which implies that $E\left(\frac{1}{n}v' H_Z v\right) = \frac{\ell}{n}\Omega_v$ and $E\left(\frac{1}{n}v' H_Z u\right) = \frac{\ell}{n}\Omega_{vu}$. So,

$$
\begin{aligned}
E\left(\frac{1}{n}X^{*\prime} H_Z X^*\right) &= \frac{1}{n} E\left((Z^*\Pi + v^*)' H_Z (Z^*\Pi + v^*)\right) \\
&= \frac{1}{n} E\left(\Pi' Z^{*\prime} H_Z Z^* \Pi + \Pi' Z^{*\prime} H_Z v^* + v^{*\prime} H_Z Z^* \Pi + v^{*\prime} H_Z v^*\right) \\
&= \frac{1}{n} E\left(\Pi' Z^{*\prime} Z^* \Pi\right) + \frac{1}{n} E\left(v^{*\prime} H_Z v^*\right) \\
&= \frac{1}{n} E\left(\Pi' Z^{*\prime} Z^* \Pi\right) + \frac{\ell}{n}\Omega_v,
\end{aligned}
\tag{B.9}
$$

and

$$
\begin{aligned}
E\left(\frac{1}{n}X^{*\prime} H_Z u^*\right) &= \frac{1}{n} E\left((Z^*\Pi + v^*)' H_Z u^*\right) \\
&= \frac{1}{n} E\left(v' H_Z u\right) = \frac{\ell}{n}\Omega_{vu}.
\end{aligned}
\tag{B.10}
$$

Hence,

$$\hat{\beta}_{\text{FE-2SLS}} - \beta \xrightarrow{p} \frac{\ell}{n}\left[\frac{1}{n}\Pi' E\left(Z^{*\prime}Z^{*}\right)\Pi + \frac{\ell}{n}\Omega_v\right]^{-1}\Omega_{uv} \neq 0, \tag{B.11}$$

and $\hat{\beta}_{\text{FE-2SLS}}$ is inconsistent unless $\frac{\ell}{n} \to 0$.