

## ABSTRACT

FAN, QINGLIANG. The Adaptive Lasso Method for Instrumental Variable Selection. (Under the direction of Dr. Mehmet Caner.)

This paper develops theoretical adaptive lasso method to select instrumental variables. We recommend to use the k-class estimators such as two-stage least squares (TSLS) or limited information maximum likelihood (LIML) after first stage variable selection. It is important to use only the strong variables because finite sample properties of estimators are sensitive to the choice of instruments. In this paper we extend the technique of adaptive lasso to multivariate linear model framework. We address one important empirical question which is weak instruments under just identification. Adaptive lasso estimates irrelevant instruments as 0 asymptotically as if it were known. In Monte Carlo studies we show adaptive lasso can select the strong instrumental variable consistently and improve the accuracy of TSLS compared to non-selection and other selection methods such as Donald and Newey (2001), Kuersteiner and Okui (2010), Belloni et al. (2010) and conventional information criterion such as AIC and BIC. In application, we replicate the famous study of returns to education (Angrist and Krueger, 1991) and improve the important empirical results. Empirical researcher will find this method easy to implement in their own study. At last we also propose a new Adaptive Lasso type TSLS estimators which selects variables in both stages. We show that the procedure enjoys the same oracle properties as in Caner (2009). We show in simulation that with variable selection problem in the structural equation, adaptive lasso TSLS estimator performs better than other existing methods.

© Copyright 2012 by Qingliang Fan

All Rights Reserved

The Adaptive Lasso Method for Instrumental Variable Selection

by  
Qingliang Fan

A dissertation submitted to the Graduate Faculty of  
North Carolina State University  
in partial fulfillment of the  
requirements for the Degree of  
Doctor of Philosophy

Economics

Raleigh, North Carolina

2012

APPROVED BY:

---

Dr. Barry Goodwin

---

Dr. Charles Knoeber

---

Dr. Helen Zhang

---

Dr. Mehmet Caner  
Chair of Advisory Committee

## DEDICATION

To my parents: Peiquan Fan and Aizhen Li Fan.

## BIOGRAPHY

‘Michael’ Qingliang Fan grew up in Linfen City, Shanxi Province, China before attending Tianjin University of Finance and Economics for his undergraduate studies. In 2005, he received bachelor of arts degree with honors in economics. In his senior year at college, he studied at San Diego State University as an exchange student. Deeply attracted by the city and school, he chose to continue his education at SDSU and was offered a scholarship. In 2007 he graduated from SDSU with master’s degree in economics. He since developed strong interests in economics research while at college and decided to pursue a PhD degree. Dr. Barbara Bailey in Statistics Department and Dr. Tia Hilmer in Economics Department who were both professors at SDSU and North Carolina State University alumni helped him choosing econometrics as his main research field. As a result he came to study at North Carolina State University in 2007 on a scholarship.

## ACKNOWLEDGEMENTS

I am gratefully acknowledge Dr. Mehmet Caner for his enormous support and excellent guidance throughout my five years of studies at North Carolina State University. His great insight and enormous help made the dissertation possible and the whole writing process much more efficient and enjoyable. Since I started writing the dissertation in the fall of 2009, there are numerous times in the coffee shops near campus, we would work on the paper together while having refreshments. And those are the best memories of my research experience at NC State. I would like to thank Dr. Helen Hao Zhang for her great help on the simulation techniques, insightful comments on shrinkage method, and being very patient on my endless questions on R,  $\LaTeX$  among many other softwares that I need to learn along the way of writing the dissertation. Also special thanks to Helen for showing up in my preliminary oral examination in April 2010 when she was expecting a baby in about two months. I thank Dr. Charles Knoeber for helpful comments and his help in my Corporate Finance class which resulted in a term paper. I thank Dr. Barry Goodwin for encouragement and many inspiring conversations on my research. I thank Dr. Leonard Stefanski for his helpful comment on variance of IV estimator. I thank Dr. Lexin Li for excellent instruction of dimension reduction class. Furthermore, I acknowledge the help throughout my academic career by Drs. Edmund Balsdon, Babara Bailey, Cynthia Bansak, David Flath, Tia Hilmer, Tamah Morant, Douglas Pearce, Thitima Puttitanun, Douglas Stewart. I thank the support of my friends Robert, Yao, Yu-Chin. Finally, I thank my girlfriend Xuan, my wonderful parents for their support and love.

# TABLE OF CONTENTS

<b>List of Tables</b> . . . . .	<b>vii</b>
<b>Chapter 1 Introduction</b> . . . . .	<b>1</b>
1.1 Motivation of Research . . . . .	1
1.2 Contributions . . . . .	4
<b>Chapter 2 The Multivariate Adaptive Lasso Model</b> . . . . .	<b>6</b>
2.1 The Model with Irrelevant Instruments . . . . .	6
2.2 Assumptions . . . . .	8
2.3 The Limit Theory for Multivariate Adaptive Lasso . . . . .	8
<b>Chapter 3 Adaptive Lasso with Weak Instruments</b> . . . . .	<b>10</b>
3.1 The Model with Weak Instruments . . . . .	10
3.2 Assumptions . . . . .	11
3.3 Asymptotic Properties of Adaptive Lasso with Weak Instruments . . . . .	12
3.4 Computational Algorithm . . . . .	12
3.4.1 Computation of Adaptive Lasso Estimates . . . . .	12
3.4.2 Estimating Standard Error for Nonzero Parameter . . . . .	15
3.4.3 Post Selection Estimation . . . . .	15
<b>Chapter 4 Monte Carlo Study</b> . . . . .	<b>16</b>
4.1 Performance of Adaptive Lasso in First Stage Selection . . . . .	16
4.2 Effect of Adaptive Lasso Selection V.S. Non-selection on TSLS Estimator . . . . .	18
4.3 Adaptive Lasso V.S. Donald and Newey (2001) . . . . .	22
4.4 Adaptive Lasso V.S. Model Averaging . . . . .	25
4.5 Adaptive Lasso V.S. Post-Lasso . . . . .	31
4.6 Adaptive Lasso V.S. AIC and BIC . . . . .	36
4.7 Adaptive Lasso and Best Subset Selection . . . . .	40
<b>Chapter 5 Returns to Education Revisited</b> . . . . .	<b>42</b>
<b>Chapter 6 Conclusion</b> . . . . .	<b>47</b>
<b>References</b> . . . . .	<b>48</b>
<b>Appendices</b> . . . . .	<b>51</b>
Appendix A Proofs . . . . .	52
A.1 Proof of Theorem 1 . . . . .	52
A.2 Proof of Theorem 2 . . . . .	55

Appendix B	Adaptive Lasso type TSLS estimator . . . . .	57
B.1	The Basic Model . . . . .	57
B.2	Assumptions . . . . .	58
B.3	The Limit Theory for Adaptive Lasso IV estimator . . . . .	59
B.4	Standard Errors of Adaptive Lasso TSLS Estimator . . . . .	59
B.5	Proof of Theorems . . . . .	60
B.5.1	Proof of Estimation Consistency . . . . .	60
B.5.2	Proof of Oracle Property . . . . .	61
B.6	Simulation Results . . . . .	65
B.6.1	Adaptive lasso V.S. Post-Lasso and Model Averaging . . . . .	65
B.6.2	Adaptive Lasso V.S. bridge . . . . .	66
Appendix C	One Step Adaptive Lasso with No Structural Equation Variable Selection	69
C.1	The DGP . . . . .	69
C.2	Assumptions . . . . .	70
C.3	The Limit Theory for One Step Adaptive Lasso Estimator . . . . .	70
C.4	Proof of Theorems . . . . .	71
C.5	Simulation Results . . . . .	74
Appendix D	One Step Adaptive Lasso with Structural Equation Variable Selection .	77

## LIST OF TABLES

Table 4.1	Simulation results for Model 1 . . . . .	17
Table 4.2	Simulation results for Model 2 ( $t = 3$ ) . . . . .	17
Table 4.3	Summary statistics of TSLS for Model 3 . . . . .	20
Table 4.4	Summary statistics of TSLS for Model 4 ( $\gamma_2 = .1/\sqrt{n}$ ) . . . . .	20
Table 4.5	Summary statistics of TSLS for Model 4' . . . . .	21
Table 4.6	Summary statistics for Model 3: Donald and Newey v.s. adaptive lasso . . . . .	23
Table 4.7	Summary statistics for Model 4 ( $\gamma_2 = .1/\sqrt{n}$ ): Donald and Newey v.s. adaptive lasso . . . . .	23
Table 4.8	Summary statistics for Model 4': Donald and Newey v.s. adaptive lasso . . . . .	24
Table 4.9	Summary statistics for Model 5 ( $\rho = .1$ ): Model averaging v.s. adaptive lasso . . . . .	27
Table 4.10	Summary statistics for Model 5 ( $\rho = .5$ ): Model averaging v.s. adaptive lasso . . . . .	27
Table 4.11	Summary statistics for Model 5 ( $\rho = .9$ ): Model averaging v.s. adaptive lasso . . . . .	28
Table 4.12	Summary statistics for Model 6 ( $\rho = .1$ ): Model averaging v.s. adaptive lasso . . . . .	29
Table 4.13	Summary statistics for Model 6 ( $\rho = .5$ ): Model averaging v.s. adaptive lasso . . . . .	29
Table 4.14	Summary statistics for Model 6 ( $\rho = .9$ ): Model averaging v.s. adaptive lasso . . . . .	30
Table 4.15	Summary statistics of TSLS for Model 3 . . . . .	32
Table 4.16	Summary statistics of TSLS for Model 4 ( $\gamma_2 = .1/\sqrt{n}$ ) . . . . .	32
Table 4.17	Summary statistics of TSLS for Model 4' . . . . .	33
Table 4.18	Summary statistics of TSLS for Model 3 . . . . .	34
Table 4.19	Summary statistics of TSLS for Model 4 ( $\gamma_2 = .1/\sqrt{n}$ ) . . . . .	34
Table 4.20	Summary statistics of TSLS for Model 4' . . . . .	35
Table 4.21	Correct Model Selection Percentage for Just Identified Model (One weak $.1/\sqrt{n}$ ) . . . . .	36
Table 4.22	Summary statistics for Model 7: AIC / BIC v.s. adaptive lasso . . . . .	38
Table 4.23	Summary statistics for Model 8 ( $t = .1$ ): AIC / BIC v.s. adaptive lasso . . . . .	39
Table 4.24	Success rate of model selection: Model 9 . . . . .	40
Table 4.25	Success rate of model selection: Model 10 . . . . .	40
Table 5.1	TSLS And Adaptive Lasso TSLS Estimate of Return to Education . . . . .	45
Table 5.2	Cruz and Moreira (2005) replication (Table 3 of the paper) . . . . .	46
Table B.1	Summary statistics of TSLS for Design 1 . . . . .	66
Table B.2	Summary statistics of TSLS for Design 2 . . . . .	66
Table B.3	Summary statistics of TSLS for Design 1 . . . . .	67
Table B.4	Summary statistics of TSLS for Design 2 . . . . .	67

Table B.5	Summary statistics for Design 1 . . . . .	68
Table B.6	Summary statistics for Design 2 . . . . .	68
Table C.1	Summary statistics for Model 7 . . . . .	76

# Chapter 1

## Introduction

### 1.1 Motivation of Research

Instrumental variable selection has become the focus of much research in areas of application for which datasets with both strong and weak instruments are available. Empirical researchers who use IV methods often cannot decide which instrumental variables are valid. E.g. the  $k^{th}$  lag of the dependent variable in time series such as in VAR models and dynamic panel data models, or using the dummy variables as IV such as in the returns to education research by Angrist and Krueger (1991). Sometimes even intuitive instruments may not be valid under specification test (Hahn and Hausman, 2001). When the instruments are all relevant but some are weak, which means they are weakly correlated with endogenous structural variables, the finite sample properties of estimators are sensitive to the choice of instruments. In Bound et al. (1995), first stage partial  $R^2$  and F test is suggested to examine the quality of IV estimates. Instruments selection can improve the precision of IV estimator in the case of many constructed instruments (dummies, polynomials and the various interactions of those variables). It is also necessary when there are not sufficient observations to obtain the standard estimates.

Adaptive lasso (Zou, 2006) is a weighted  $\ell_1$  penalization method for simultaneous estimation and variable selection. The degree of penalty is adaptively chosen (data dependent). Adaptive lasso has the celebrated ‘oracle’ properties (Fan and Li, 2001) of variable selection and estimation consistency and efficiency. When the true parameter is nonzero (zero), it is estimated as nonzero (zero) as if it were known. Adaptive lasso solution is continuous and essentially a convex optimization problem, so that it can avoid the local optimization trap and often improve the prediction accuracy due to the bias-variance trade-off. Computation-wise, we can use an efficient algorithm which is proposed for optimizing penalized functions. Thus adaptive lasso provides us both desired theoretical properties and computational convenience to screen the instrumental variables. Using adaptive lasso we can cull the weak instrumental variables from

the strong en route getting good estimation and inference properties on IV estimators. In the main body of this paper we focus on the reduced form equation. After first stage selection the conventional TSLS or LIML method is recommended. We assume the instruments are strictly exogenous but some instruments are weakly correlated to included endogenous variables. We also assume rank condition is satisfied after correct variable selection (the number of strong instruments is at least equal to the number of endogenous variables). If the variable selection results in an unidentified model, then it is a warning sign that the researcher needs to find other valid instruments. Our task is to select only the strong instrumental variables with probability tending to one. We wish to show this property in two weak IV model setups: irrelevant (unidentified, the instrument's parameter is exactly 0) instruments and the weak instruments (parameter is local to zero, see Staiger and Stock, 1997). In the Appendices we introduce adaptive lasso method to select variables in the structural equation too.

A large amount of literature on weak instruments shows that our theoretically optimal variable selection work is necessary. IV estimator can be nested in a more comprehensive GMM framework introduced by Hansen (1982). GMM estimator is known to be inaccurate when it has weak identification problems. IV regression with weak instruments has consistency and identification problems (Nelson and Startz, 1990; Stock and Wright, 2000), namely, the IV estimator could be biased toward  $\text{plim}(\hat{\beta}_{\text{OLS}})$  (the direction of OLS estimator) and confidence intervals be severely distorted. In empirical studies, F test is commonly used for testing collective strength of instruments. Staiger and Stock (1997) suggests that instruments do not enter the first stage equation if F statistic is less than 10. Cragg and Donald (1997) develops tests for the identifiability of IV model based on moment specifications. Wright (2003) constructs a generalized test on nonlinear GMM model. The null hypothesis is underidentification with hypothesized rank less than full rank. F test can be treated as a special case of Wright (2003) test when we have linear IV model and the hypothesized rank is zero. Wright (2003) test is a conservative one, which can provide useful diagnostics for the validity of conventional asymptotic theory. But Wright (2003) does not select instruments.

In existing literature many methods have been developed to solve the weak instruments problem. First, developing more robust test on IV estimator, which is nearly optimal whether the IVs are weak or strong. Such tests include Anderson-Rubin (1949), Lagrange multiplier (Kleibergen, 2002), conditional likelihood ratio (CLR) (Andrews, et al., 2006). Inference on weak instrumental variables is also addressed by Caner (2007, 2010a, 2010b). Anderson-Rubin test uses a grid search method which can be extremely time consuming or even infeasible for many weak instruments. On the other hand, adaptive lasso is computationally efficient. Anderson-Rubin test will also give a very wide confidence interval which makes it useless if the underlying model is sparse. If the structural equation has 'no effect' parameters, adaptive lasso is straight-forward in model selection but Anderson-Rubin test can not select models. Second,

developing robust estimators e.g. k-Class (LIML as a special case), BTSLS (Donald and Newey, 2001), Jackknife IV estimator (Angrist, Imbens and Krueger, 1999), empirical likelihood estimator (exponential tilting, Caner, 2010c). Third, selecting moment conditions under GMM framework. The model setup of moment selection is different from the first two categories that both strong and weak instruments are present. But there is no a priori reason to prefer one moment condition over the other. Andrews (1999) proposes a consistent moment condition selection method using GMM-BIC amid overidentification of the model to select the correct moments (largest set of valid instruments). Hall and Peixe (2003)s canonical correlations information criterion (CCIC) approach is designed to eliminate the redundant moment conditions. Liao (2010) proposes a GMM shrinkage method to select moment conditions. Okui (2009) selects moment conditions based on the bias-variance trade-off of Nagar (1959) approximation of mean squared error (MSE). Fourth, selecting the number of instruments which minimizes MSE. Donald and Newey (2001) (hereafter referred to as DN (2001)) choose the number of instruments by minimizing Nagar (1959) type approximation of MSE. They showed that the optimal set of instruments that minimizes structural equation MSE of the estimator is converging in probability to the minimization of the MSE criteria as the dominant term in an Edgeworth expansion. DN (2001) method also applies to the scenarios where the number of instruments grows with sample size, but at a slower rate. DN (2001) does not choose strong instruments against weak ones. But only focusing on minimizing MSE results in selecting the irrelevant model more often therefore the bias of DN (2001) is inevitably large compared to adaptive lasso selected model. Fifth, construct optimal set of instruments using model averaging. Kuersteiner and Okui (2010) (hereafter referred to as KO (2010)) propose a model averaging approach of Hansen (2007) on first stage equation of TSLS estimator as well as modified LIML and Fuller's (1997) estimator. In their method, weights for averaging are chosen to minimize the asymptotic MSE of the estimators. The selection does not depend on prior knowledge about the strength of the instruments. Stronger instruments get more weights in their direction. KO (2010) controls for higher order bias term of MSE therefore it has theoretical advantage of more favorable trade off between bias and efficiency compared to DN (2001). But in KO (2010), nonzero and even negative weights are assigned to the irrelevant variables which can further increase the finite sample bias. Sixth, Belloni et al. (2010) proposes a lasso/post-lasso procedure to estimate first stage instruments (and more generally, nonparametric conditional expectation functions) and use the predicted value of IVs in the conventional IV regression. The key assumptions for the effectiveness of the method is the sparsity of the instruments set, and strong instruments are available (necessary for the estimation of conditional expectation functions). They showed that lasso/post-lasso procedure can estimate the near-oracle first stage regression function (the IV estimator and its variance matrix), therefore the IV estimator will be efficient albeit bias in lasso estimation and model selection. Furthermore their method is robust to (i), the number of

instruments  $p \gg n$ , (ii), heteroskedasticity and non-Gaussian structural equation disturbances. While in this paper we select the relevant instruments using adaptive lasso instead of estimate the optimal IV set. And we select variable in the structural equation which is not addressed by any of the above methods.

There has been very few works on simultaneous estimation and variable selection under GMM framework. Caner (2009) is the first to introduce the lasso type GMM method and showed the oracle property. Lasso type GMM dominates BIC and DT in selecting the right model. Also, lasso type GMM estimator with  $\ell_q(0 < q < 1)$  penalty is consistent. But Caner (2009) does not select instrumental variables. And the bridge estimator with  $\ell_q(0 < q < 1)$  penalty is neither continuous nor convex. Because the discontinuity results in instability in model prediction, the  $\ell_q(0 < q < 1)$  penalty is considered less favorable than the  $\ell_1$  penalty (Fan and Li, 2001). In this paper, we choose adaptive lasso over lasso because of the advantage of adaptive lasso in variable selection consistency. Lasso could have bias in estimation and is inconsistent on variable selection under certain circumstances (Zou, 2006). But when the dimension is high and possible severe collinearity, adaptive lasso may fail too (Zou and Zhang, 2009).

## 1.2 Contributions

This paper makes two main contributions to the econometric theory of IV regression. First, we extend the literature on IV selection. We select instruments in the reduced form equation that some instruments are weak or irrelevant. We show that irrelevant instruments will be estimated as 0 as if we know they were. Adaptive lasso has optimal convergence rate of  $\sqrt{n}$ . Adaptive lasso selects the correct model with probability converging to one. These are the celebrated oracle property of variable selection. We also extend the literature on  $\ell_1$  – related methods with univariate response variable to multiple response variables. In empirical instruments selection problems, multiple endogenous variables model is more common compared to single endogenous variable model. In this situation we seek a subset of instruments that is useful for several included endogenous variables simultaneously. We introduce the multivariate adaptive lasso and show the same oracle properties hold. Second, this paper develops asymptotic theory for adaptive lasso when the partial correlations between some instruments and included endogenous variables are weak. The issue with selecting weak instruments is they are not exactly 0 in finite sample but the role these variables play in the model is diminishing as sample size increases. In the model setup where we have enough number of available strong instruments for model identification, the ideal solution is to shrink all weak instruments parameter to zero and use the strong ones. Knight and Fu (2001) discussed the idea of weak instruments model within the context of small parameters and the shrinkage of lasso estimator to zero with pos-

itive probability. In this paper, the small parameter is estimated as 0 with probability one asymptotically. Adaptive lasso gives us the sparsity property that we need, even if the true parameter is not exactly zero locally. The inference and prediction of the TSLS model improve if we can consistently select the strong variables.

Several extensions are also studied. First, we can also proceed to the second stage with the selected instruments and use adaptive lasso again (adaptive lasso type TSLS estimator<sup>1</sup>). Second, we introduce the one step adaptive lasso in both stages. We use the predicted value of endogenous regressors by first stage adaptive lasso, either in the conventional IV method or a second stage adaptive lasso estimation. We show the oracle properties of these new estimators in Appendices C and D.

Simulation studies are conducted to assess the performance of the above adaptive lasso methods. We find that adaptive lasso can select the correct model consistently and it beats the empirical ‘rule of thumb’ F-test in second stage inference. Adaptive lasso in first stage reduces bias in the second stage at the cost of reduced finite sample variance. Adaptive lasso selected IV estimator has much smaller bias than DN (2001) and KO (2010) which chooses irrelevant parameters more often. Our results are similar to the lasso estimation by Belloni et al. (2010). We also find that adaptive lasso method selects the correct model much more often than BIC which could be locally optimal rather than global optimal.

One major contribution of this dissertation to the applied economics research is that our method can be easily implemented to any IV regression. Instead of using all instruments (lags, dummies, interactions, etc), our method gives a guideline of which instrument is to keep and which one is to drop. We use the famous returns to education (Angrist and Krueger, 1991) as an example, where we replicate the study using their set of instruments and data. We find significant improvement of the TSLS estimator by selecting a subset of instruments.

The remainder of the paper is organized as follows. Chapter 2 introduces the basic multivariate adaptive lasso model with irrelevant instruments (strictly 0). Chapter 3 introduces weak instruments (small but not 0) and the limit theory. Chapter 4 provides a Monte Carlo study. In Chapter 5, we conduct an empirical study on returns to education. Chapter 6 concludes. The Appendix A contains all proofs and in Appendix B we propose an adaptive lasso type TSLS estimator. In Appendix C and Appendix D, we introduce a new one step adaptive lasso estimator using predicted value of dependent variables.

---

<sup>1</sup>In Appendix B we will discuss selection in structural equation. None of the existing IV selection methods select the structural equation variables

## Chapter 2

# The Multivariate Adaptive Lasso Model

### 2.1 The Model with Irrelevant Instruments

The structural model is

$$y = X\beta^* + \epsilon$$

where  $X$  is an  $n \times p$  matrix of endogenous variables,  $y$  is  $n \times 1$  vector of dependent variable,  $\epsilon$  is  $n \times 1$  random disturbances,  $\beta^*$  is  $p \times 1$  coefficients.

In this Chapter we focus on the reduced form equation

$$X = Z\gamma^* + \nu \tag{2.1}$$

where  $Z$  is an  $n \times q$  matrix of instrumental variables. Assume that  $q > p$ .  $\nu$  is  $n \times p$  vector of unobserved errors.

$$\gamma^* = \begin{bmatrix} \gamma_{11}^* & \gamma_{12}^* & \cdots & \gamma_{1p}^* \\ \gamma_{21}^* & \gamma_{22}^* & \cdots & \gamma_{2p}^* \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{q1}^* & \gamma_{q2}^* & \cdots & \gamma_{qp}^* \end{bmatrix}$$

which is  $q \times p$  coefficient matrix, each  $\gamma_j^{*'} = (\gamma_{j1}^*, \gamma_{j2}^*, \dots, \gamma_{jp}^*)$ ,  $j = 1, 2, \dots, q$ , is a  $1 \times p$  row vector. Without loss of generality, we assume the data are centered, so the intercept is not included in (2.1). The model of interest here is the sparse one which can be interpreted as both strong and weak instruments are included (in this section the weak instrument's parameters are exactly 0, non-zero small parameters will be discussed in Chapter 3 ). We now define the multivariate version of coefficient sparsity. If one of the vectors in  $\gamma^*$  matrix is zero such as

$\gamma_j^* = 0$ , then all the elements of the coefficient vector are zero. And we define  $\gamma_j^* \neq 0$  as all the elements of the vector are non-zero for simplicity. Define integer  $q_0$  as the number of nonzero row vectors in the coefficient matrix. We further assume that  $q_0 \geq p$ , this is equivalent to say the number of valid instrumental variables is at least equal to the number of endogenous variables. Thus the structural model satisfies rank condition given correct variable selection<sup>2</sup>. Let  $\mathcal{A} = \{ \text{all the vectors in } \gamma^* \text{ except for the zero vectors} \}$ .  $\gamma_{\mathcal{A}}$  denote the block of partitioned matrix  $\gamma^*$  such that  $\gamma_j^* \neq 0$ . Let  $\mathcal{A}^c = \{ \text{all the zero vectors in } \gamma^* \}$ .

In order to solve the multivariate adaptive lasso, we vectorize (2.1):

$$X_V \equiv \text{vec}(X) = \text{vec}(Z\gamma^*) + \text{vec}(\nu) = (I_p \otimes Z)\text{vec}(\gamma^*) + \text{vec}(\nu) = Z_K \text{vec}(\gamma^*) + \nu_V \quad (2.2)$$

where  $X_V$  is  $np \times 1$ ,  $Z_K$  is the Kronecker products of identity matrix  $I_p$  and  $Z$ , which is  $np \times qp$ ,  $\text{vec}(\gamma^*)$  is  $qp \times 1$ ,  $\nu_V = \text{vec}(\nu)$ , which is  $np \times 1$ .

The adaptive lasso is first introduced by Zou (2006). In our multivariate model framework it is defined as:

$$\hat{\gamma}_n = \arg \min_{\gamma} \left\{ (X_V - Z_K \gamma_V)' (X_V - Z_K \gamma_V) + \lambda_n \sum_{k=1}^p \sum_{j=1}^q \hat{w}_{jk} |\gamma_{jk}| \right\} \quad (2.3)$$

where  $\lambda_n$  is a nonnegative regularization parameter. When  $\lambda_n$  is appropriately chosen, some estimators which minimize (2.3) could be exactly zero.  $\gamma_V = \text{vec}(\gamma)$  ( $qp \times 1$ , arbitrary  $\gamma$ ). Suppose  $\tilde{\gamma}$  is a  $\sqrt{n}$  consistent estimator of  $\gamma^*$ , e.g. the Lasso type GMM estimator by Caner (2009) which is also immune to collinearity. The data-dependent weight  $\hat{w}_{jk} = 1/|\tilde{\gamma}_{jk}|^{\tau}$ , where  $0 < \tau \leq 1$ ,  $j = 1, 2, \dots, q$ ;  $k = 1, 2, \dots, p$ . In Chapter 4 simulations, we will use the OLS estimator  $\tilde{\gamma}_{\text{OLS}}$  and  $\tau = 1$ . Denote each row vector of  $\hat{\gamma}_n$  as  $\hat{\gamma}_j$ . Let  $\mathcal{A}_n = \{ \text{the nonzero vectors in } \hat{\gamma}_n \text{ matrix} \}$ .  $\hat{\gamma}_{\mathcal{A}_n}$ ,  $\hat{\gamma}_{\mathcal{A}_n^c}$  denote the block of  $\hat{\gamma}_n$  matrix such that  $\hat{\gamma}_j \neq 0$  and  $\hat{\gamma}_j = 0$  respectively.

Adaptive lasso is essentially a convex optimization problem, which means we can solve for the global minimization instead of possible multiple local minimizations. Also adaptive lasso is an  $\ell_1$  penalization method, the second term in (2.3) is a general form (with adaptive weight) of the so-called “ $\ell_1$  penalty”. We can use efficient algorithm such as LARS (least angle regression, Efron et al., 2004) to compute the estimates. Most importantly, adaptive lasso has oracle properties which are pivotal to instrumental variable selection.

We discuss the implementation algorithm and a consistent method to select the tuning parameter  $\lambda$  in Chapter 4.

---

<sup>2</sup>As we discussed in Chapter 1, if in empirical studies the selection resulted in non-identification, then it is a warning sign that the researcher needs to find other instruments.

## 2.2 Assumptions

We then present the assumptions that are useful in providing the oracle properties of adaptive lasso estimator in the linear IV model.

**Assumption 2.1** : the unobserved errors  $\nu_i$ 's are independent identically distributed (i.i.d.) with mean 0 and covariance matrix  $\Omega = \sigma_\nu^2 \mathbf{I}_p$

**Assumption 2.2** :  $C_n = \frac{1}{n} Z'_K Z_K \rightarrow C$  where  $C$  is a  $qp \times qp$  positive definite matrix. Without loss of generality, we assume the first  $q_0$  instrumental variables are nonzero, so that  $\mathcal{A} = \{\gamma_1^*, \gamma_2^*, \dots, \gamma_{q_0}^*\}$ . Let  $C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$ , where  $C_{11}$  is  $q_0 p \times q_0 p$  matrix,  $C_{22}$  is  $(q - q_0)p \times (q - q_0)p$  matrix and  $C_{12} = C'_{21}$ .

**Assumption 2.3** :  $E(\nu | Z) = 0$ ,  $E(\varepsilon | Z) = 0$  where  $\varepsilon$  ( $p \times 1$ ) is the structural equation error term

**Assumption 2.4** :  $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$  and  $\lambda_n n^{\frac{\tau-1}{2}} \rightarrow \infty$

The first two assumptions are adopted from the setup of Knight and Fu (2000) for the analysis of large sample theory. Assumption 2.3 is the exogeneity of instrumental variables. Since we focus on the reduced form equation, we do not give the specific formula of the structural form equation. But to make sure the endogeneity problem exists, we have  $E(\varepsilon_i \nu_{ik} | Z_i) = \sigma_{\nu\varepsilon} \neq 0$ ,  $i = 1, 2, \dots, n$ ;  $k = 1, 2, \dots, p$ . And no serial correlation  $E(\nu_i \varepsilon_{i'k} | Z_i) = 0$ ,  $i, i' = 1, 2, \dots, n$ ,  $i \neq i'$ . By Assumption 2.4 we can derive  $\lambda_n n^{\frac{\tau-1}{2}} \rightarrow 0$ . In the proofs, this result is necessary for the estimation consistency which is not explicitly shown in Zou (2006). The two components of Assumption 2.4 are stronger than the assumptions in Knight and Fu (2000). Under Assumption 4 we will have stronger results, namely shrinkage of irrelevant variables to 0 with probability 1 asymptotically.

## 2.3 The Limit Theory for Multivariate Adaptive Lasso

In this section we present the asymptotic properties of adaptive lasso estimator. Adaptive lasso has ideal asymptotic properties of estimation and selection consistency. Adaptive lasso can select the correct model as if it were known, also known as the oracle properties.

**Theorem 1** (The oracle properties of multivariate adaptive lasso)

1. Asymptotic normality:  $\sqrt{n}[\text{vec}(\hat{\gamma}_{\mathcal{A}}) - \text{vec}(\gamma_{\mathcal{A}})] \rightarrow_d N(0, \sigma_\nu^2 C_{11}^{-1})$
2. Consistency in variable selection:  $\lim_{n \rightarrow \infty} P(\mathcal{A}_n = \mathcal{A}) = 1$

**Remarks.** In adaptive lasso, the zero parameter is penalized more severely than nonzero parameter. The weight of zero parameter converges to infinity while the weight of nonzero parameter converges to a constant in large samples. The asymptotic normality applies to

nonzero coefficients. Zero coefficients converge to 0. It should be noticed that adaptive lasso estimator for nonzero coefficients behaves just like LS estimator in large sample. We suggest  $\tilde{\gamma}_{OLS}$  for adaptive weight in applications because of its popularity in linear regression models. If collinearity is a concern, we can try more stable estimators such as ridge estimator. When we set  $\tau = 1$  with  $\tilde{\gamma}_{OLS}$ , the adaptive lasso solution is identical to non-negative garrotte (Breiman, 1995) with additional sign constraints. In fact, Zou (2006) points out  $\tilde{\gamma}$  can converge to true  $\gamma^*$  in a slower rate than  $\sqrt{n}$ . The oracle properties can still hold with minor modifications in the assumptions. Zou (2006) shows adaptive lasso risk function has oracle inequality that attained performance differs from ideal performance by at most a factor of  $2 \log n$ . Therefore adaptive lasso is near-minimax optimal (Donoho and Johnstone, 1994). To choose  $\lambda_n$ , we use the BIC tuning parameter selector which can consistently select the true model (Wang and Leng, 2007). Furthermore, the adaptive lasso also applies in penalized log-likelihood estimation (with the adaptively weighted  $\ell_1$  penalty) and the oracle properties hold (Zou, 2006, Zhang and Lu, 2007).

## Chapter 3

# Adaptive Lasso with Weak Instruments

In Chapter 2, we showed that the limiting distributions of adaptive lasso estimator converge to a degenerate point at 0 when the true coefficient vector  $\gamma^* = 0$ . In many empirical applications, the researchers often confront weak instrumental variables which translate into small but not exactly 0 parameters in our model. Weak instruments causes conventional inference to be misleading, namely the inconsistency and nonstandard distributions of the IV estimator even in large sample. In this section we consider the model with both strong and weak instruments. It is thus desirable in large sample to select only the strong instruments since the weak instruments provide no useful information. In this chapter we show adaptive lasso estimator for weak instrumental variable can converge to 0 with probability 1. In simulations we find that weak instruments could be estimated as exactly 0 in finite samples when strong instruments are present.

### 3.1 The Model with Weak Instruments

We adopt the setup of Knight and Fu (2000) for local asymptotic analysis. Assume that we have a triangular array of observations (a device for asymptotic theory where the parameters change with sample size)

$$X_n = Z_n \gamma_n + \nu_n \tag{3.1}$$

where for each  $n$ ,  $X_n$  is  $n \times p$ ,  $Z_n$  is  $n \times q$ ,  $\gamma_n = \begin{bmatrix} \gamma'_{n1} \\ \gamma'_{n2} \\ \vdots \\ \gamma'_{nq} \end{bmatrix}$  is  $q \times p$ , each coefficient  $\gamma'_{nj}$ ,  $j = 1, 2, \dots, q$  is a row vector  $1 \times p$ ,  $\nu_n$  is  $n \times p$  unobserved errors. Assume that  $q > p$ . Note

that in triangular array format the parameter  $\gamma_n$  have a subscript  $n$ , which is not to be confused with the subscript  $n$  of adaptive lasso estimator.

Suppose  $\gamma_{nj} = \gamma_j^* + \frac{t_j}{\sqrt{n}}$ ,  $j = 1, 2, \dots, q$ , where  $t_j \in \mathbb{R}^p$  is a vector of some constant real numbers. If  $\gamma_j^* = 0$ , it means all the elements of the coefficient vector are zero, and the corresponding instrumental variable is weak. If  $\gamma_j^* \neq 0$ , it means all the elements of the vector are non-zero. And the corresponding instrumental variable is strong. In finite sample the coefficients of weak instrumental variables are small but not zero. We assume that the number of weak instruments  $q_w$  satisfies that  $0 < q_w \leq q - p$  so that the structural model can be identified after selection. Let  $\mathcal{A} = \{\text{all parameter vectors of strong instruments}\}$ .  $\gamma_{\mathcal{A}}$  denote the block of  $\gamma_n$  matrix which  $\gamma_j^* \neq 0$ .

Keeping the same notation as in Chapter 2, let  $X_V = \text{vec}(X_n)$  which is  $np \times 1$ ,  $Z_K$  is the Kronecker products of  $I_p$  and  $Z_n$ , which is  $np \times qp$ .

The adaptive lasso estimator is:

$$\hat{\gamma}_n = \arg \min_{\gamma} \left\{ (X_V - Z_K \gamma_V)' (X_V - Z_K \gamma_V) + \lambda_n \sum_{k=1}^p \sum_{j=1}^q \hat{w}_{jk} |\gamma_{njk}| \right\} \quad (3.2)$$

$\gamma_V$  ( $qp \times 1$ ) is the vectorized matrix of any  $\gamma$  ( $q \times p$ ). The weight vector  $\hat{w}_{jk} = 1/|\tilde{\gamma}_{jk}|^{\tau}$ , where  $\tilde{\gamma}$  is a  $\sqrt{n}$  consistent estimator of  $\gamma^*$  and  $0 < \tau \leq 1$ . Denote each row vector of  $\hat{\gamma}_n$  as  $\hat{\gamma}_{nj}$ . Let  $\mathcal{A}_n = \{\text{the nonzero vectors in } \hat{\gamma}_n \text{ matrix}\}$ .  $\hat{\gamma}_{\mathcal{A}_n}$  denotes the block of  $\hat{\gamma}_n$  matrix which  $\hat{\gamma}_j \neq 0$ .

## 3.2 Assumptions

Similar to Chapter 2, we need the following assumptions for the limit theory of adaptive lasso with weak instruments.

**Assumption 3.1** : for each  $n$ , the unobserved errors  $\nu_n$  are i.i.d. with mean 0 and covariance matrix  $\Omega = \sigma_{\nu}^2 I_p$

**Assumption 3.2** :  $\frac{1}{n} Z_K^T Z_K \rightarrow C$  where  $C$  is a positive definite matrix. Without loss of generality, we assume the first  $q_w$  instrumental variables are weak. Let  $C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$ , where

$C_{22}$  is  $q_w p \times q_w p$  matrix. And  $\frac{1}{n} \max_{1 \leq i \leq n} Z_{ni}^T Z_{ni} \rightarrow 0$  where  $Z_{ni}$  is the  $i^{\text{th}}$  row of  $Z_n$

**Assumption 3.3** :  $E(\nu, \varepsilon | Z_n) = 0$  where  $\varepsilon$  is the structural equation error term

**Assumption 3.4** :  $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$  and  $\lambda_n n^{\frac{\tau-1}{2}} \rightarrow \infty$

This model allows us to illustrate the asymptotics of adaptive lasso estimator when both strong and weak instruments are present. The goal is to treat weak instruments as 0 asymptot-

ically as if we knew they were weak instruments. It has the flavor of oracle property of Theorem 1, but the difference is in weak IV case, we only need the convergence to 0, not the variable selection consistency.

### 3.3 Asymptotic Properties of Adaptive Lasso with Weak Instruments

**Theorem 2**  $\sqrt{n}(\hat{\gamma}_n - \gamma_n) \rightarrow_d \arg \min(V)$ , where

$$V(u) = u^T C u - 2u^T W + \frac{\lambda_n}{\sqrt{n}} \sum_{k=1}^p \sum_{j=1}^q \hat{w}_{jk} \sqrt{n} (|\gamma_{nj} + \frac{u_{jk}}{\sqrt{n}}| - |\gamma_{nj}|) \quad (3.3)$$

$W$  has a  $N(0, \sigma_\nu^2 C)$  distribution. Theorem 2 implies that  $\sqrt{n}[\text{vec}(\hat{\gamma}_{\mathcal{A}_n}) - \text{vec}(\gamma_{\mathcal{A}})] \rightarrow_d N(0, \sigma_\nu^2 C_{11}^{-1})$  if  $\gamma_j^* \neq 0$ . And  $\sqrt{n}(\hat{\gamma}_{nj} - \gamma_{nj}) \rightarrow_d -t_j$  if  $\gamma_j^* = 0$ , which is equivalent to  $\sqrt{n}\hat{\gamma}_{nj} \rightarrow_d 0$ .

It suggests that weak instruments will be estimated as exact 0 with probability tending to 1 in large samples even when strong instruments are present. For example, suppose that  $\gamma_{nj} = \frac{t_j}{\sqrt{n}}$ , then the limiting distribution of  $\sqrt{n}(\hat{\gamma}_{nj} - \frac{t_j}{\sqrt{n}})$  given by Theorem 2 is  $-t_j$ , irrespective of the values of other coefficients. So the limiting distribution of  $\sqrt{n}\hat{\gamma}_{nj} = o_p(1)$ . Thus, Theorem 2 indicates that weak instruments will be estimated as exactly 0 asymptotically.

We see adaptive lasso can treat weak instruments as zero with positive possibility even in finite sample. In empirical studies this method can provide a conclusive solution for instrumental variable selection. Compared to F-test ‘rule of thumb’, Donald and Newey (2001), or BIC, adaptive lasso can identify the weak instruments instead of just sending a signal of suspicion without identifying the exact ones.

### 3.4 Computational Algorithm

#### 3.4.1 Computation of Adaptive Lasso Estimates

After the main theorems presented, we now briefly discuss the computation algorithms of adaptive lasso. Zhang and Lu (2007) provides a modified shooting algorithm by Fu (1998) which is shown to converge to the global minimizer. We can also use a simple modification of LARS algorithm to solve the adaptive lasso estimates efficiently. The LARS algorithm is implemented to compute the entire solution path of the lasso. The efficient path algorithm makes the adaptive lasso an attractive method for real applications. Friedman et al. (2007) develops an extremely efficient procedure known as coordinate descent for fitting the entire lasso or elastic-net regularization path. In the simulations and empirical studies of this paper we will use LARS algorithm

to solve adaptive lasso estimates.

Adaptive Lasso estimates can be computed efficiently by a modification of LARS (Least Angle Regression, and S suggesting ‘LASSO’ and ‘Stagewise’) algorithm (Efron et al., 2004). The computational efficiency is an advantage of adaptive Lasso in practice compared to other Oracle methods such as SCAD (Fan and Li, 2001) and bridge estimator. In this section, we briefly discuss the implementation of LARS in adaptive Lasso.

In Zou (2006), a simple modified version of LARS can be adopted for the adaptive Lasso estimation. It works as follows.

To illustrate the method, we set up a basic linear model, for  $i = 1, \dots, n$

$$x_i = z_i' \gamma + \nu_i \quad (3.4)$$

where  $x_i$  is the univariate endogenous variable,  $z_i = (z_{i1}, \dots, z_{id})'$  is the associated  $d$ -dimensional instruments,  $\gamma = (\gamma_1, \dots, \gamma_d)'$  is the coefficient vector and  $\nu_i$  is the random error with mean 0 and variance  $\sigma_\nu^2$ . For multivariate model which we presented in Section 2, we vectorize  $x$ ,  $z$ ,  $\nu$  etc., the algorithm works as in the univariate case. Assume that we fit  $d$  predictors and the true model has  $d_0$  variables ( $1 \leq d_0 \leq d$ ).

#### Adaptive Lasso Algorithm

1. Create new covariates  $z_j^* = z_j / \hat{w}_j$ ,  $j = 1, 2, \dots, d$ , where  $\hat{w}_j$  is the adaptive weight as defined in Section 2.
2. Solve the LASSO via LARS algorithm for given  $\lambda$ .

$$\tilde{\gamma} = \arg \min_{\gamma} \|X - Z^* \gamma\|^2 + \lambda \sum_{j=1}^d |\gamma_j| \quad (3.5)$$

where  $X = (x_1, \dots, x_n)'$ ,  $Z^* = (z_1^*, \dots, z_n^*)'$ .

3. Output  $\hat{\gamma}_j = \tilde{\gamma}_j / \hat{w}_j$ ,  $j = 1, 2, \dots, d$ . This is the adaptive Lasso estimate.
4. Then we put  $\hat{\gamma}_j$ ,  $j = 1, 2, \dots, d$ , adaptive lasso estimate from Step 3 in equation (3.6) below. This provides us a BIC number for a given  $\lambda$ . Note that tuning is explained in details after the Algorithm.
5. Repeat Step 2-4 for each remaining  $\lambda$  in a set of  $\Lambda$  (e.g.,  $\Lambda = \{\lambda_1, \dots, \lambda_{100}\}$ ) and record each  $BIC_\lambda$  for given  $\hat{\gamma}(\lambda)$ .
6. Choose the pair of  $\hat{\gamma}(\lambda)$ , which minimizes BIC over  $\lambda$ .

As one can see Step 2 is LASSO. But we use LARS algorithm to compute LASSO. Then in Step 3, we convert this to adaptive lasso solution. Now we explain the simple intuition behind

LARS.

In Step 2 of the Algorithm, we use LARS algorithm. LARS procedure works as follows (see Efron et al., 2004 for more details). Assume for simplicity that we have standardized our explanatory variables to have zero mean and unit variance, and that our response variable also has zero mean. We start with all coefficients being equal to zero (no variables in the model), and find the predictor most correlated with endogenous variable  $x$ , say it's the first covariate  $z_1$  (as we can always switch the 'position' of the covariates). The reasoning here is the largest correlation possibly shows significance of the variable, so we take out that variable and include in our model. Since the covariate  $z$  most correlated with the residual is equivalently the one that makes the least angle with the residual, the name of the method is called 'least angle regression'. We take the largest step possible in the direction of this predictor until some other predictor, say  $z_2$ , has as much correlation with the current residual. Different from classic Forward Selection which takes a 'full step' with  $z_1$ , LARS now proceeds in a direction equiangular between the two predictors  $z_1$  and  $z_2$  (so that the residual makes equal angles with both covariates) until a third variable  $z_3$  becomes equally correlated with the current residual. LARS then proceeds equiangularly between  $z_1$ ,  $z_2$  and  $z_3$ , that is, along the 'least angle direction', until a fourth variable enters, and so on. The result of  $\hat{\gamma}$  in Step 2 depends on  $\lambda$  choice. We explain this in detail next in choice of  $\lambda$ . As shown in Theorem 1 of Efron et al. (2004), a slight modification to LARS can get us the full solution path of LASSO.

In this part we explain the method to select tuning parameter  $\lambda$  which we used in the adaptive Lasso simulations. Recall that the tuning parameter  $\lambda$  controls the penalty level and therefore the model complexity. We follow the tuning parameter selector by Wang and Leng (2007). In their paper, it is showed that the BIC method tuning parameter can achieve the Oracle properties of adaptive Lasso. BIC method is selection consistent for adaptive LASSO under fixed predictor dimension and a slight modification of the BIC method is also consistent under diverging number of parameters (Wang et al., 2009). In Wang and Leng (2007), they showed that the tuning parameter by BIC method can select the correct model with probability approaching 1.

The BIC method for  $\lambda$  selector is to minimize

$$BIC_\lambda = \hat{\sigma}_\lambda^2 + DF_\lambda \log(n)/n \tag{3.6}$$

where  $\hat{\sigma}_\lambda^2 = n^{-1} \|X - Z\hat{\gamma}\|^2$ ,  $DF_\lambda$  is the number of nonzero coefficients in  $\hat{\gamma}$  which is described in Step 3 of adaptive Lasso Algorithm. In Wang and Leng (2007),  $\hat{\sigma}_\lambda^2$  is approximated by the first two order of Taylor series expansion, but the asymptotic results holds for both.

The reason we use this method is that BIC method can get us the model selection consistency when the sample size  $n$  is approaching  $\infty$ . But other methods such as AIC or GCV

(generalized cross validation) can not get us there (see Wang et al., 2007).

### 3.4.2 Estimating Standard Error for Nonzero Parameter

We follow the standard adaptive lasso standard error sandwich formula by Zou (2006). Efron et al. (2004) Tibshirani (1996), Fan and Li (2001) show that covariance of nonzero penalized estimates can be approximated by iteratively computing the ridge solution. Zou's (2006) standard error formula follows local quadratic approximation (LQA) approach which can provide a consistent sandwich formula for computing the covariance of the estimates of nonzero parameters (Fan and Peng, 2004). In Chapter 4, we will use the same sandwich formula in Zou (2006).

### 3.4.3 Post Selection Estimation

We proceed with conventional class of estimators estimation (assume rank condition holds):

$$\hat{\beta}_{\alpha,n} = (X'P_{\mathcal{A}_n}X - \hat{\alpha}X'X)^{-1}(X'P_{\mathcal{A}_n}y - \hat{\alpha}X') \quad (3.7)$$

where  $P_{\mathcal{A}_n}$  is the projection matrix of  $Z_{\mathcal{A}_n}$  (the matrix of selected instruments),  $y$  is  $n \times 1$  dependent variable in the structural model. As discussed in in Hansen, Hausman, and Newey (2004), the class of estimators defined by (3.7) includes all of the well-known k-class estimators (except OLS estimator) such as TSLS (set  $\hat{\alpha} = 0$ ) and LIML (set  $\hat{\alpha} = \min_{\|\theta\|=1} \theta'Y'P_{\mathcal{A}_n}Y\theta / \theta'Y'Y\theta$ , where  $Y = [y, X]$ ). In Chapter 4 simulations we use TSLS estimator after first stage selection.

# Chapter 4

## Monte Carlo Study

In this chapter we devise Monte Carlo simulations to address the following issues that are related to the applied research using IV regression.

### 4.1 Performance of Adaptive Lasso in First Stage Selection

First, we look at the performance of adaptive lasso in reduced form equation. We report (i) the percentage of correct model selection, when the nonzero (strong) instruments are estimated as nonzero and zero (or local to zero) instruments are estimated as zero, (ii) standard error of adaptive lasso estimates, we use the sandwich formula in Zou (2006) and (iii) prediction performance measured by MSE. We adopt the simple modification in Zou (2006) to solve adaptive lasso estimates using LARS algorithm.

We use the following design for the reduced form equation. The reduced form equation is:

$$X = Z\gamma + \nu$$

where  $X$  is  $n \times 1$  endogenous variables,  $Z$  is  $n \times 2$  matrix of instruments.  $Z_i \sim N(0, I_2)$  i.i.d. and  $E(\nu_i|Z_i) = 0$  for  $i = 1, 2, \dots, n$ .  $\nu_i \sim N(0, \sigma^2)$  i.i.d.  $\gamma = (\gamma_1, \gamma_2)$  is  $2 \times 1$  true parameter vector. The IV model has two settings of parameter values:

**Model 1** : one strong and one irrelevant instrument.  $\gamma = (2, 0)'$

**Model 2**: one strong and one weak (local to zero) instrument.  $\gamma = (2, \frac{t}{\sqrt{n}})'$ , where  $t$  is a constant real number

**Note:** For Model 2, the ‘correct model selection’ we report is actually the scenario where the selection method picks up only  $z_1$ , the strong instrument. It is a little misleading in a strict correct model selection sense because  $z_2$  is not exact zero in finite sample. But with large sample size it would be desirable to use the strong instruments only. Given all the models presented here are identifiable with only the strong instruments, we define the correct model

selection to be selection of only strong instruments.

In both models we simulated 100 datasets for each combination of  $(\sigma, n)$ . We use three sample sizes,  $n = 60, 120$  and  $300$  and  $\sigma$  takes on values  $6, 3$  in corresponding model setup. We set  $t = 3$ .

Second, we use F test (joint test on both instruments) on the reduced form equation and report the percentage of when F-statistic is greater than 10. It is common in applied studies that diagnose instruments to be weak if F-statistic is less than 10 (Staiger and Stock, 1997). F-statistic is also approximately increasing with concentration parameter which is a unit-less measure of instruments strength. (Stock et al., 2002) We'll compare the model selection performance of adaptive lasso and first stage F-test 'rule of thumb' (use all instruments whenever F-statistic is greater than 10).

The results are shown in Table 4.1 and Table 4.2.

Table 4.1: Simulation results for Model 1

	MSP	MSE	s.e.	FMSP
$n = 60, \sigma = 6$	.65	1.581	.335	.96
$n = 60, \sigma = 3$	.98	.218	.127	.26
$n = 120, \sigma = 6$	.87	.599	.217	.76
$n = 120, \sigma = 3$	.98	.089	.068	.00
$n = 300, \sigma = 6$	.97	.155	.106	.06
$n = 300, \sigma = 3$	.99	.034	.029	.00

MSP is the rate of correct model selection, FMSP is 1 minus the rate of F-statistics greater than 10

Table 4.2: Simulation results for Model 2 ( $t = 3$ )

	MSP	MSE	s.e.	FMSP
$n = 60, \sigma = 6$	.41	1.218	.223	.97
$n = 60, \sigma = 3$	.81	.385	.130	.21
$n = 120, \sigma = 6$	.77	.763	.225	.81
$n = 120, \sigma = 3$	.87	.171	.070	.00
$n = 300, \sigma = 6$	.95	.177	.105	.09
$n = 300, \sigma = 3$	.92	.075	.028	.00

MSP is the rate of correct model selection (treats  $\gamma_2$  as 0), FMSP is 1 minus the rate of F-statistics greater than 10

In our simulations, when sample size is bigger than 120 or  $\sigma$  is less than 3, F-test ‘rule of thumb’ often fails to detect whether weak instruments are in the model. F-test ‘rule of thumb’ tends to miss the mark of model selection since it usually cannot reject the  $H_0$  that both coefficients are zero. It is known that rejection of the null hypothesis by no means implies there is no weak instrument (Staiger and Stock, 1997). One issue of using of F-test is, as it is illustrated in our simulation, the researchers could not decide which instrument is weak based on the result of F-test unless they do subset selection (t test wouldn’t work) which we know is prone to instability (similar to BIC) plus size distortion. On the other hand, adaptive lasso not only shows there are weak instruments in the model, it specifically tells which ones are (by shrinking them to 0). We see as the sample size increase the correct model selection rate is getting closer to 1. Most importantly, in the case of weak instrument where the parameter is small but not zero, adaptive lasso can also estimate it as 0 in finite sample with positive probability.

## 4.2 Effect of Adaptive Lasso Selection V.S. Non-selection on TSLS Estimator

Recall the original research problem of this paper is to select the strong instruments for TSLS regression. Weak instruments lead to a nonstandard distribution and bias toward  $plim(\tilde{\beta}_{OLS})$ . TSLS is consistent because it only picks up information from the strong instruments as  $n \rightarrow \infty$ . But finite sample property such as bias of IV estimator can be affected by the irrelevant instruments. It is crucial to the study how IV selection affect the finite sample property of TSLS estimator in the aspect statistical inference. In simulation we find out that the instrumental variable selection in reduced form equation improves the inference of TSLS estimator.

We conduct the study in the following two steps. (i) Adaptive lasso selection was implemented in the first stage reduced form equation. (ii) Use the selected instruments in conventional TSLS. We report the median bias (Bias), MSE and standard error (s.e.) of the TSLS estimator for the full model (with all instruments), the strong/weak instruments only model and the adaptive lasso selected model.

The linear IV regression model with a single endogenous regressor and no included exogenous variable is:

$$y_i = \beta x_i + \epsilon_i \tag{4.1}$$

$$x_i = \gamma_1 z_{1i} + \gamma_2 z_{2i} + \nu_i \tag{4.2}$$

$\beta$  is the scalar parameter of interest, fix the true value  $\beta = 1$ .  $z_i = (z_{1i}, z_{2i}) \sim \text{i.i.d. } N(0, I_2)$ ,  $i = 1, 2, \dots, n$ . The errors  $(\epsilon_i, \nu_i) \sim \text{i.i.d. } N(0, \Sigma)$  ( $i = 1, 2, \dots, n$ ), where  $\Sigma = \begin{bmatrix} \sigma_\epsilon^2 & \sigma_{\epsilon\nu} \\ \sigma_{\epsilon\nu} & \sigma_\nu^2 \end{bmatrix}$ .  $\rho = \sigma_{\epsilon\nu}/(\sigma_\epsilon\sigma_\nu)$  is correlation between the two error terms  $\epsilon$  and  $\nu$ . Let  $\sigma_\epsilon = \sigma_\nu = \sigma$ . The degree of endogeneity is controlled by the covariance  $\rho$ . The closer  $\rho$  is to 1, the stronger endogeneity  $x$  is. We use two values for  $\rho$ , .5 and .99.  $n = 60, 120, 300$ . Each model is replicated 500 times. We use the following three settings of  $\gamma$ . Notice we model weak instrument by using local to 0 (Model 4) and small but nonzero (Model 4'), the former one is more interesting in theory.

**Model 3** : one strong and one irrelevant instruments.  $\gamma = (1, 0)'$

**Model 4**: one strong and one weak instruments.  $\gamma = (1, \frac{t}{\sqrt{n}})'$ , where  $t$  is a constant real number

**Model 4'**: one strong and one weak instruments.  $\gamma = (1, .1)'$

We show in Table 4.3 and Table 4.4 that first stage adaptive lasso selection improves the accuracy and inference of TSLS estimator. We have three comparison models, (i) the full model which uses all available instruments, (ii) the model which uses only the strong instrument (SO), and (iii) the model which uses only the weak instrument (WO). The bias of adaptive lasso selected model is generally smaller than the full model which uses all available instruments. In Table 4.4 we can see the effect of instruments selection in weak instruments case. When the sample size is 300 and 1000, adaptive lasso selected IV estimator has much smaller bias than the full model.

We recommend first stage adaptive lasso selection for empirical studies especially when an empirical researcher does not know which instruments to include in the model. It is better (in the sense of finite sample bias) to apply adaptive lasso selection in the reduced form equation than to use all available instruments. Similar to the results of Bound et al. (1995), we find that adding independent but irrelevant instruments can reduce the standard error of IV estimator but introduces more bias. In our simulations, TSLS estimator of adaptive lasso selected model has smaller bias than non-selection full model. Bias decreases when sample size increases. If we would have known the true model, e.g. Model 3, we know it is just identified and achieves asymptotic efficiency. But in finite sample, the standard error of IV estimator under true model is higher than the one under the full model. Adaptive lasso model is close to SO in bias, since it can select the correct model most of times. And we know that unless the model satisfies the lower bound condition of Kim and Park (1998), adding more irrelevant instruments can reduce the finite sample variance of IV estimator. In empirical studies the bias could be very high if many weak instruments are used. We recommend to use the correct model other than reduce the standard error, since the inference is affected more by bias-variance trade-off. Therefore correct model selection is important in the estimation and inference of TSLS estimator.

Table 4.3: Summary statistics of TSLS for Model 3

	SO			WO			FM			AD		
	Bias	MSE	s.e.	Bias	MSE	s.e.	Bias	MSE	s.e.	Bias	MSE	s.e.
$n = 60, \sigma = 2, \rho = .5$	-.007	.100	.317	.385	3.224	1.754	.018	.087	.295	.032	.104	.321
$n = 60, \sigma = 2, \rho = .99$	-.016	.139	.372	.843	1.684	.987	.050	.096	.305	.036	.105	.328
$n = 120, \sigma = 2, \rho = .5$	.004	.037	.193	.376	4.491	2.086	.017	.036	.187	.007	.037	.193
$n = 120, \sigma = 2, \rho = .99$	-.003	.039	.199	.763	1.440	.926	.033	.037	.190	.003	.039	.198
$n = 300, \sigma = 3, \rho = .5$	-.001	.034	.185	.471	2.687	1.570	.022	.033	.180	.006	.034	.184
$n = 300, \sigma = 3, \rho = .99$	.018	.035	.185	.877	1.199	.656	.049	.033	.176	.024	.034	.184
$n = 1000, \sigma = 3, \rho = .5$	.006	.009	.096	.553	3.276	1.723	.010	.009	.096	.007	.009	.096
$n = 1000, \sigma = 3, \rho = .99$	.002	.009	.098	.910	1.312	.696	.013	.009	.097	.004	.009	.098

Table 4.4: Summary statistics of TSLS for Model 4 ( $\gamma_2 = .1/\sqrt{n}$ )

	SO			WO			FM			AD		
	Bias	MSE	s.e.	Bias	MSE	s.e.	Bias	MSE	s.e.	Bias	MSE	s.e.
$n = 60, \sigma = 2, \rho = .5$	-.009	.092	.304	.433	2.736	1.596	.026	.077	.277	.026	.091	.301
$n = 60, \sigma = 2, \rho = .99$	.013	.141	.375	.802	1.324	.824	.076	.089	.288	.070	.103	.313
$n = 120, \sigma = 2, \rho = .5$	.004	.039	.199	.343	3.054	1.714	.017	.038	.193	.009	.039	.198
$n = 120, \sigma = 2, \rho = .99$	.023	.041	.202	.796	1.418	.886	.052	.038	.189	.030	.040	.198
$n = 300, \sigma = 3, \rho = .5$	-.003	.034	.184	.375	4.148	2.002	.014	.032	.179	.003	.033	.183
$n = 300, \sigma = 3, \rho = .99$	-.003	.036	.191	.881	1.507	.855	.032	.034	.180	.007	.036	.188
$n = 1000, \sigma = 3, \rho = .5$	.004	.009	.096	.418	4.199	2.006	.008	.009	.096	.004	.009	.096
$n = 1000, \sigma = 3, \rho = .99$	.006	.009	.097	.897	1.237	.657	.017	.009	.096	.007	.009	.097

SO is the model which we only use the strong instrument, WO is the the model which we only use the weak instrument, FM is the full model which we use all instruments, AD is the model which TSLS uses adaptive lasso selected (first stage) instruments.

Table 4.5: Summary statistics of TSLS for Model 4'

	SO			WO			FM			AD		
	Bias	MSE	s.e.	Bias	MSE	s.e.	Bias	MSE	s.e.	Bias	MSE	s.e.
$n = 60, \sigma = 2, \rho = .5$	.009	.476	.690	.340	3.168	1.747	.035	.080	.281	.036	.069	.260
$n = 60, \sigma = 2, \rho = .99$	.012	.178	.422	.735	1.249	.841	.082	.093	.295	.076	.075	.263
$n = 120, \sigma = 2, \rho = .5$	.006	.037	.193	.317	2.340	1.497	.010	.035	.187	.010	.036	.190
$n = 120, \sigma = 2, \rho = .99$	-.003	.039	.198	.570	1.083	.871	.033	.036	.186	.009	.037	.193
$n = 300, \sigma = 3, \rho = .5$	-.002	.035	.175	.352	2.373	1.500	.013	.033	.181	.002	.034	.186
$n = 300, \sigma = 3, \rho = .99$	.012	.034	.183	.579	1.143	.899	.041	.032	.174	.023	.033	.181
$n = 1000, \sigma = 3, \rho = .5$	.009	.009	.096	.152	1.095	1.035	.011	.009	.095	.011	.009	.096
$n = 1000, \sigma = 3, \rho = .99$	-.003	.010	.278	.953	.936	.006	.009	.096	.096	-.001	.010	.098

SO is the model which we only use the strong instrument, WO is the the model which we only use the weak instrument, FM is the full model which we use all instruments, AD is the model which TSLS uses adaptive lasso selected (first stage) instruments.

### 4.3 Adaptive Lasso V.S. Donald and Newey (2001)

In empirical studies, DN (2001) is most commonly used method in selection of instruments. We now compare TSLS estimator using adaptive lasso with that of DN (2001) which chooses the number of instruments to minimize the leading term of Nagar (1959) type MSE. DN (2001)'s method integrates bias-variance trade-off but it does not target on correct model selection of the reduced form equation. We could improve the bias performance of TSLS using adaptive lasso, since adaptive lasso can give us the correct model. E.g., if the true reduced form equation is Model 5 below, including the irrelevant instrument could reduce the MSE, but as we know, the irrelevant instrument should not be in the model. So DN (2001) only focus on minimizing MSE, but the cost is including some weak instruments<sup>3</sup> in the model which could aggravate the bias. We show the advantage of adaptive lasso in bias and inference given the model has both strong and weak instruments. In Table 4.6 and Table 4.7 we report the median bias, correct model selection rate, MSE and prediction error (or sampling error) of TSLS estimator over 500 replicates.

---

<sup>3</sup>In their paper the authors do not discuss redundant instruments directly, and in their simulation, weak identification is modeled as small first stage R squares, and all coefficients are small but nonzero

Table 4.6: Summary statistics for Model 3: Donald and Newey v.s. adaptive lasso

	Donald & Newey				Adaptive Lasso			
	Bias	MSP	MSE	$\hat{S}$	Bias	MSP	MSE	$\hat{S}$
$n = 60, \sigma = 2, \rho = .5$	.038	.320	.082	29.129	.008	.908	.107	32.731
$n = 60, \sigma = 2, \rho = .99$	.063	.524	.086	32.201	.024	.906	.101	35.912
$n = 120, \sigma = 2, \rho = .5$	.018	.312	.032	23.714	.010	.950	.033	25.321
$n = 120, \sigma = 2, \rho = .99$	.040	.546	.035	33.088	.027	.954	.042	35.960
$n = 300, \sigma = 4, \rho = .5$	.055	.152	.059	382.778	.021	.930	.081	422.198
$n = 300, \sigma = 4, \rho = .99$	.067	.288	.072	505.455	.015	.920	.086	551.779
$n = 1000, \sigma = 6, \rho = .5$	.031	.110	.038	1956.240	.016	.984	.046	2097.839
$n = 1000, \sigma = 6, \rho = .99$	.021	.192	.038	2714.484	-.011	.968	.047	2939.944

Table 4.7: Summary statistics for Model 4 ( $\gamma_2 = .1/\sqrt{n}$ ): Donald and Newey v.s. adaptive lasso

	Donald & Newey				Adaptive Lasso			
	Bias	MSP	MSE	$\hat{S}$	Bias	MSP	MSE	$\hat{S}$
$n = 60, \sigma = 2, \rho = .5$	.048	.298	.109	24.833	.018	.876	.132	27.706
$n = 60, \sigma = 2, \rho = .99$	.038	.536	.096	35.513	.003	.888	.107	38.545
$n = 120, \sigma = 2, \rho = .5$	.040	.316	.038	26.867	.036	.950	.039	28.234
$n = 120, \sigma = 2, \rho = .99$	.040	.548	.035	33.062	.027	.954	.042	35.970
$n = 300, \sigma = 4, \rho = .5$	.055	.152	.059	380.246	.021	.922	.083	417.718
$n = 300, \sigma = 4, \rho = .99$	.066	.288	.072	505.088	.014	.922	.086	551.963
$n = 1000, \sigma = 6, \rho = .5$	.031	.104	.038	1955.731	.016	.984	.046	2097.958
$n = 1000, \sigma = 6, \rho = .99$	.035	.170	.042	2388.511	.003	.974	.047	2540.909

MSP is the rate of correct model selection, PE is the prediction error of structural equation

Table 4.8: Summary statistics for Model 4': Donald and Newey v.s. adaptive lasso

	Donald & Newey				Adaptive Lasso			
	Bias	MSP	MSE	$\hat{S}$	Bias	MSP	MSE	$\hat{S}$
$n = 60, \sigma = 2, \rho = .5$	.018	.296	.087	27.065	.010	.880	.108	30.224
$n = 60, \sigma = 2, \rho = .99$	.051	.428	.098	38.479	-.003	.892	.108	42.613
$n = 120, \sigma = 2, \rho = .5$	.011	.258	.035	25.659	.004	.922	.037	27.509
$n = 120, \sigma = 2, \rho = .99$	.036	.474	.035	44.176	.005	.930	.043	37.735
$n = 300, \sigma = 4, \rho = .5$	.032	.142	.063	406.162	.008	.898	.087	445.260
$n = 300, \sigma = 4, \rho = .99$	.058	.252	.063	482.168	.023	.910	.079	532.032
$n = 1000, \sigma = 6, \rho = .5$	.023	.096	.050	2297.677	-.004	.952	.058	2525.040
$n = 1000, \sigma = 6, \rho = .99$	.041	.178	.042	2465.182	.011	.960	.052	2641.615

MSP is the rate of correct model selection, PE is the prediction error of structural equation

We now discuss briefly how the DN (2001) method work. Use the notation in (4.1) and (4.2), the 2SLS estimator is

$$\hat{\beta} = (X'P^K X)^{-1} X'P^K Y$$

where  $X = (x_1, \dots, x_n)'$ ,  $Y = (y_1, \dots, y_n)'$  and  $P^K = Z^K(Z^{K'}Z^K)^{-1}Z^K$ , where  $K$  is the index for the number of instruments included in the regression. In Model 5, since we have 2 instruments,  $K = 1, 2$ . And for each  $K$ , we can have different choice of instruments, e.g.,  $Z^1 = (z_{1,1}, \dots, z_{1,n})$  or  $Z^1 = (z_{2,1}, \dots, z_{2,n})$ ,  $Z^1$  is  $n \times 1$ . And  $Z^2 = (z_1, \dots, z_n)$ , where  $z = (z_1, z_2)$ , and  $Z^2$  is  $n \times 2$ . As pointed out in DN (2001),  $Z^K$  corresponds to a series of instruments. But when  $K$  is large, we don't know which instruments we should use, and it could be very computationally consuming when  $K$  gets large.

Now we define the necessary variables to minimize MSE with respect to  $K$  as described in DN (2001). Let  $\tilde{\beta}$  be some preliminary estimator of  $\beta$ , e.g., it can be the regular 2SLS estimator.  $\tilde{\epsilon} = Y - X\tilde{\beta}$ .  $\tilde{H} = X'P^K X/n$ .  $\tilde{u} = (I - P^K)X$ .  $\tilde{u}_\lambda = \tilde{u}\tilde{H}^{-1}\tilde{\lambda}$ , where we let  $\tilde{\lambda} = 1$  (see details in DN 2001). We have the following variables:  $\hat{\sigma}_\epsilon^2 = \tilde{\epsilon}'\tilde{\epsilon}/n$ ,  $\hat{\sigma}_\lambda^2 = \tilde{u}'_\lambda\tilde{u}_\lambda/n$ ,  $\hat{\sigma}_{\lambda\epsilon} = \tilde{u}'_\lambda\tilde{\epsilon}/n$ . These preliminary estimators are not depend on  $K$ , they remain as constants as the approximate of MSE for different instruments are calculated. Define a Mallor's creteria. First, let  $\hat{u}^K = (I - P^K)X$ ,  $\hat{u}_\lambda^K = \hat{u}^K\tilde{H}\hat{\lambda}$ . So the Mallor's creteria is  $\hat{R}_\lambda^m(K) = \frac{\hat{u}_\lambda^{K'}\hat{u}_\lambda^K}{n} + \hat{\sigma}_\lambda^2(2K/n)$ . Finally, the approximate MSE of the 2SLS estimator is  $\hat{S}_\lambda(K) = \hat{\sigma}_{\lambda\epsilon}^2 \frac{K^2}{n} + \hat{\sigma}_\epsilon^2 \left( \hat{R}_\lambda^m(K) - \sigma_\lambda^2 \frac{K}{n} \right)$ .

Table 4.6 and 4.7 show that DN (2001) TSLS estimator has smaller MSE but higher bias than that of adaptive lasso. We can also see that DN (2001) performs sub par in model selection. This is an evidence that only minimizing MSE does not guarantee consistent model selection. So in the aspect of using the correct model in TSLS, adaptive lasso is better. In the case of identifiable model with both strong and weak instruments, adaptive lasso gives us asymptotic efficiency on TSLS estimator. Also notice that bias and MSE is higher as the degree of endogeneity increases.

## 4.4 Adaptive Lasso V.S. Model Averaging

Kuersteiner and Okui (2010) proposed a first-stage model averaging method to construct optimal instruments. In the paper they showed that model averaging is more favorable in the aspect of trade-off between estimator bias and efficiency relative to DN (2001) and other procedures that rely on a single set of instruments. In KO (2010), the model averaging weights are not restricted to be positive or bounded within one (though they provided different search approach under corresponding restrictions). We compare the bias and MSE of adaptive lasso with both

KO (2010)<sup>4</sup> and DN (2001). For comparison of adaptive lasso and DN (2001), we also report correct model selection rate (MSP) and the MSE as defined in DN (2001)<sup>5</sup>. Table 4.9 to 4.14 report the results based on 500 repeats.

We use the following two setup of  $\gamma$  :

**Model 5** : one nonzero (strong) and one exact zero coefficients  $\gamma = (.8, 0)'$

**Model 6**: one nonzero (strong) and one local to zero (weak) coefficients  $\gamma = (.8, \frac{t}{\sqrt{n}})'$ , where  $t$  is a constant real number

---

<sup>4</sup>We report the minimum MSE of KO (2010) among the restrictions U, C, P, Ps.

<sup>5</sup>Calculated using the formula in DN (2001) p. 1165.

Table 4.9: Summary statistics for Model 5 ( $\rho = .1$ ): Model averaging v.s. adaptive lasso

	Adaptive Lasso				Donald & Newey				Model Averaging	
	Bias	MSP	MSE	$\hat{S}$	Bias	MSP	MSE	$\hat{S}$	Bias	MSE
$n = 60, \sigma = 1.5$	-.010	.916	.083	22.678	-.008	.690	.071	22.642	-.005	.069
$n = 120, \sigma = 1.5$	.001	.952	.032	18.779	.004	.722	.031	18.752	.006	.031
$n = 300, \sigma = 3$	-.011	.956	.063	279.272	-.001	.706	.052	279.073	-.001	.052
$n = 300, \sigma = 1.5$	.003	.980	.011	13.632	.004	.726	.011	13.620	.003	.011
$n = 500, \sigma = 3$	.010	.992	.032	267.588	.009	.698	.033	267.428	.010	.032
$n = 1000, \sigma = 3$	.007	.990	.015	222.888	.008	.654	.014	222.795	.009	.015

Table 4.10: Summary statistics for Model 5 ( $\rho = .5$ ): Model averaging v.s. adaptive lasso

	Adaptive Lasso				Donald & Newey				Model Averaging	
	Bias	MSP	MSE	$\hat{S}$	Bias	MSP	MSE	$\hat{S}$	Bias	MSE
$n = 60, \sigma = 1.5$	.019	.908	.078	21.651	.033	.750	.073	21.611	.038	.067
$n = 120, \sigma = 1.5$	.010	.944	.037	18.601	.017	.742	.034	18.579	.022	.033
$n = 300, \sigma = 3$	-.001	.970	.071	337.072	.010	.720	.069	336.861	.025	.062
$n = 300, \sigma = 1.5$	.004	.962	.012	14.810	.007	.748	.012	14.799	.011	.011
$n = 500, \sigma = 3$	.004	.970	.031	303.306	.018	.716	.030	303.169	.022	.029
$n = 1000, \sigma = 3$	.006	.984	.014	228.485	.012	.700	.013	228.399	.013	.013

MSP is the success percentage of correct model selection, MSE is regular MSE,  $\hat{S}$  is DN type MSE

Table 4.11: Summary statistics for Model 5 ( $\rho = .9$ ): Model averaging v.s. adaptive lasso

	Adaptive Lasso				Donald & Newey				Model Averaging	
	Bias	MSP	MSE	$\hat{S}$	Bias	MSP	MSE	$\hat{S}$	Bias	MSE
$n = 60, \sigma = 1.5$	.016	.888	.093	25.174	.033	.790	.129	25.151	.071	.095
$n = 120, \sigma = 1.5$	-.006	.962	.036	22.865	.004	.844	.034	22.845	.021	.032
$n = 300, \sigma = 3$	.003	.944	.078	346.590	.022	.698	.062	346.418	.041	.052
$n = 300, \sigma = 1.5$	.009	.976	.013	14.950	.020	.798	.012	14.939	.026	.012
$n = 500, \sigma = 3$	.010	.982	.039	351.957	.028	.744	.034	351.792	.035	.032
$n = 1000, \sigma = 3$	.014	.986	.015	245.571	.020	.748	.014	245.513	.026	.014

MSP is the success percentage of correct model selection, MSE is regular MSE,  $\hat{S}$  is DN type MSE

Table 4.12: Summary statistics for Model 6 ( $\rho = .1$ ): Model averaging v.s. adaptive lasso

	Adaptive Lasso				Donald & Newey				Model Averaging	
	Bias	MSP	MSE	$\hat{S}$	Bias	MSP	MSE	$\hat{S}$	Bias	MSE
$n = 60, \sigma = 1.5$	-.020	.918	.085	22.013	-.020	.674	.074	21.960	-.020	.073
$n = 120, \sigma = 1.5$	.010	.966	.033	17.656	.009	.702	.032	17.629	.009	.031
$n = 300, \sigma = 3$	-.009	.954	.067	267.702	.006	.686	.048	267.471	.003	.047
$n = 300, \sigma = 1.5$	.005	.982	.012	13.944	.006	.684	.012	13.929	.006	.012
$n = 500, \sigma = 3$	.002	.982	.029	250.872	.003	.666	.029	250.704	.004	.028
$n = 1000, \sigma = 3$	-.001	.988	.014	235.262	-.001	.732	.014	235.187	-.003	.014

Table 4.13: Summary statistics for Model 6 ( $\rho = .5$ ): Model averaging v.s. adaptive lasso

	Adaptive Lasso				Donald & Newey				Model Averaging	
	Bias	MSP	MSE	$\hat{S}$	Bias	MSP	MSE	$\hat{S}$	Bias	MSE
$n = 60, \sigma = 1.5$	.003	.904	.095	21.767	.020	.752	.082	21.734	.037	.076
$n = 120, \sigma = 1.5$	-.007	.930	.035	22.116	.001	.742	.034	22.089	.003	.032
$n = 300, \sigma = 3$	.010	.960	.063	310.169	.028	.672	.059	309.944	.034	.055
$n = 300, \sigma = 1.5$	-.004	.978	.012	14.975	.001	.744	.012	14.960	.001	.012
$n = 500, \sigma = 3$	.001	.968	.029	273.088	.016	.676	.028	272.935	.024	.028
$n = 1000, \sigma = 3$	.012	.988	.015	231.736	.014	.734	.014	231.672	.019	.014

MSP is the success percentage of correct model selection, MSE is regular MSE,  $\hat{S}$  is DN type MSE

Table 4.14: Summary statistics for Model 6 ( $\rho = .9$ ): Model averaging v.s. adaptive lasso

	Adaptive Lasso				Donald & Newey				Model Averaging	
	Bias	MSP	MSE	$\hat{S}$	Bias	MSP	MSE	$\hat{S}$	Bias	MSE
$n = 60, \sigma = 1.5$	.029	.900	.095	28.392	.044	.812	.097	28.372	.072	.080
$n = 120, \sigma = 1.5$	.006	.968	.038	24.444	.014	.838	.037	24.430	.033	.033
$n = 300, \sigma = 3$	.023	.938	.068	387.483	.041	.738	.064	387.314	.063	.057
$n = 300, \sigma = 1.5$	-.003	.978	.014	15.870	.006	.786	.013	15.857	.009	.012
$n = 500, \sigma = 3$	-.007	.988	.036	366.645	.005	.766	.034	366.516	.019	.032
$n = 1000, \sigma = 3$	.014	.988	.016	239.523	.020	.718	.015	239.446	.025	.015

MSP is the success percentage of correct model selection, MSE is regular MSE,  $\hat{S}$  is DN type MSE

In both Model 5 and Model 6, KO (2010) weighted TSLS estimator has the least regular MSE among the three methods. Adaptive lasso selected TSLS estimator has greatest MSE among the three methods. Also adaptive lasso selected TSLS estimator has higher DN type MSE than DN (2001) selected TSLS estimator. When endogeneity is weak,  $\rho = .1$ , the adaptive lasso selected TSLS is outperformed by both DN (2001) and KO (2010) in bias and MSE in small sample sizes. For  $\rho = .5$  and  $\rho = .9$ , adaptive lasso selected TSLS estimator has much smaller bias than both DN (2001) and KO (2010) most of the time (except for Model 6,  $n = 300$ ,  $\sigma = 1.5$ ). As the degree of endogeneity ( $\rho$ ) increase, the bias of the three TSLS estimators increase compared to collateral group. For DN (2001) and KO (2010), bias decreases as the sample size increases. Adaptive lasso model selection is converging when sample size increases, but DN (2001) does not.

## 4.5 Adaptive Lasso V.S. Post-Lasso

In this section we compare the performance of TSLS estimator by adaptive lasso selection with the one by Post-Lasso estimation of IVs (using Algorithm 2.2 of Belloni et al. (2010)). To estimate the optimal IV set in their method, lasso procedure selects the relevant instruments first. We use Model 3, Model 4 and Model 4' as the DGP. Using Model 3 as an example, post-lasso is a method that estimates instrument set (and more generally, conditional expectation functions), where  $z = (z_1, z_2)$ , such that  $D(z) = E(x|z)$ . The post-lasso procedure doesn't rely on model selection of relevant  $z$ 's directly, but the selected  $z$  plays important role of predict  $D(z)$ . As showed in Caner and Fan (2011), adaptive lasso is consistent in model selection while lasso is not. From Table 4.15 we see that when the sample size is small and variance is relatively large, adaptive lasso selected TSLS method has smaller bias. When sample size is large, the two methods are similar and both are consistent. We also show the simulation results using BIC tuning and optimal inequality tuning parameter  $\lambda$ .

In Table 4.21 we show the correct model selection percentage of shrinkage methods (post lasso and adaptive lasso) when we have just identified model, and the only instrument is weak. We see the results are similar but adaptive lasso has slightly better overall performance.

Table 4.15: Summary statistics of TSLS for Model 3

	Post-Lasso					Adaptive Lasso				
	Bias	MSE	s.e.	MBias	MSP	Bias	MSE	s.e.	MBias	MSP
$n = 60, \sigma = 2, \rho = .5$	.092	.062	.232	.092	.794	.026	.077	.276	.011	.904
$n = 60, \sigma = 2, \rho = .99$	.107	.058	.215	.178	.812	.048	.076	.271	.034	.900
$n = 120, \sigma = 3, \rho = .5$	.091	.067	.243	.086	.836	.059	.074	.266	.046	.876
$n = 120, \sigma = 3, \rho = .99$	.097	.064	.234	.167	.834	.073	.075	.263	.092	.872
$n = 300, \sigma = 3, \rho = .5$	.004	.034	.184	-.013	.980	.004	.034	.183	-.010	.958
$n = 300, \sigma = 3, \rho = .99$	.006	.036	.189	-.024	.978	.007	.035	.188	-.021	.968
$n = 1000, \sigma = 3, \rho = .5$	-.005	.010	.097	-.009	.982	-.005	.010	.097	-.009	.984
$n = 1000, \sigma = 3, \rho = .99$	-.001	.009	.097	-.007	.986	-.001	.009	.097	-.007	.988

Table 4.16: Summary statistics of TSLS for Model 4 ( $\gamma_2 = .1/\sqrt{n}$ )

	Post-Lasso					Adaptive Lasso				
	Bias	MSE	s.e.	MBias	MSP	Bias	MSE	s.e.	MBias	MSP
$n = 60, \sigma = 2, \rho = .5$	.083	.062	.234	.088	.830	.032	.076	.274	.011	.898
$n = 60, \sigma = 2, \rho = .99$	.102	.059	.221	.164	.822	.054	.076	.270	.038	.878
$n = 120, \sigma = 3, \rho = .5$	.077	.066	.245	.073	.832	.060	.073	.263	.042	.844
$n = 120, \sigma = 3, \rho = .99$	.106	.065	.231	.157	.854	.075	.070	.255	.093	.874
$n = 300, \sigma = 3, \rho = .5$	.005	.034	.184	-.025	.990	.005	.034	.183	-.024	.974
$n = 300, \sigma = 3, \rho = .99$	-.007	.038	.194	-.038	.984	-.006	.038	.195	-.039	.974
$n = 1000, \sigma = 3, \rho = .5$	-.005	.010	.098	-.006	.992	-.005	.010	.098	-.006	.996
$n = 1000, \sigma = 3, \rho = .99$	.004	.009	.097	-.006	.982	.004	.009	.097	-.006	.980

Post-Lasso is the model which TSLS uses Post-Lasso estimated (first stage) instruments., Adaptive Lasso is the model which TSLS uses adaptive lasso selected (first stage) instruments.

Table 4.17: Summary statistics of TSLS for Model 4'

	Post-Lasso					Adaptive Lasso				
	Bias	MSE	s.e.	MBias	MSP	Bias	MSE	s.e.	MBias	MSP
$n = 60, \sigma = 2, \rho = .5$	.105	.063	.229	.098	.784	.040	.077	.276	.012	.870
$n = 60, \sigma = 2, \rho = .99$	.112	.058	.214	.175	.808	.079	.073	.259	.056	.844
$n = 120, \sigma = 3, \rho = .5$	.100	.067	.240	.099	.852	.082	.072	.256	.072	.862
$n = 120, \sigma = 3, \rho = .99$	.111	.065	.228	.173	.824	.076	.074	.261	.092	.858
$n = 300, \sigma = 3, \rho = .5$	.013	.034	.184	-.007	.972	.014	.034	.183	-.006	.956
$n = 300, \sigma = 3, \rho = .99$	.017	.033	.182	-.009	.968	.018	.033	.181	-.007	.948
$n = 1000, \sigma = 3, \rho = .5$	.003	.010	.098	-.010	.914	-.003	.010	.098	-.011	.926
$n = 1000, \sigma = 3, \rho = .99$	.005	.009	.097	-.007	.920	.005	.009	.097	-.007	.928

Post-Lasso is the model which TSLS uses Post-Lasso estimated (first stage) instruments., Adaptive Lasso is the model which TSLS uses adaptive lasso selected (first stage) instruments.

Table 4.18: Summary statistics of TSLS for Model 3

	Post-Lasso					Adaptive Lasso ( $\lambda$ )				
	Bias	MSE	s.e.	MBias	MSP	Bias	MSE	s.e.	MBias	MSP
$n = 60, \sigma = 2, \rho = .5$	.093	.062	.232	.094	.794	.090	.062	.231	.087	.826
$n = 60, \sigma = 2, \rho = .99$	.122	.060	.212	.188	.804	.117	.058	.210	.182	.832
$n = 120, \sigma = 3, \rho = .5$	.101	.068	.240	.095	.836	.182	.073	.200	.169	.714
$n = 120, \sigma = 3, \rho = .99$	.117	.064	.225	.183	.830	.174	.061	.174	.320	.726
$n = 300, \sigma = 3, \rho = .5$	.015	.035	.186	-.008	.980	.015	.034	.183	-.003	.990
$n = 300, \sigma = 3, \rho = .99$	.015	.035	.187	-.020	.978	.012	.034	.185	-.014	.992
$n = 1000, \sigma = 3, \rho = .5$	-.002	.009	.097	-.001	.990	-.003	.009	.097	-.001	1.000
$n = 1000, \sigma = 3, \rho = .99$	.004	.009	.096	-.003	.976	.004	.009	.097	-.004	1.000

Table 4.19: Summary statistics of TSLS for Model 4 ( $\gamma_2 = .1/\sqrt{n}$ )

	Post-Lasso					Adaptive Lasso ( $\lambda$ )				
	Bias	MSE	s.e.	MBias	MSP	Bias	MSE	s.e.	MBias	MSP
$n = 60, \sigma = 2, \rho = .5$	.072	.061	.236	.070	.834	.076	.059	.231	.079	.842
$n = 60, \sigma = 2, \rho = .99$	.109	.061	.222	.172	.796	.110	.058	.214	.188	.806
$n = 120, \sigma = 3, \rho = .5$	.081	.062	.236	.093	.808	.148	.062	.201	.155	.708
$n = 120, \sigma = 3, \rho = .99$	.111	.063	.225	.187	.818	.183	.062	.171	.336	.700
$n = 300, \sigma = 3, \rho = .5$	.008	.033	.182	-.011	.986	.008	.033	.180	-.006	.994
$n = 300, \sigma = 3, \rho = .99$	-.004	.036	.190	-.030	.976	-.009	.036	.190	-.029	.996
$n = 1000, \sigma = 3, \rho = .5$	.001	.009	.097	-.004	.982	-.001	.009	.097	-.004	1.000
$n = 1000, \sigma = 3, \rho = .99$	.003	.009	.096	-.003	.990	.002	.009	.097	-.004	.998

Post-Lasso is the model which TSLS uses Post-Lasso estimated (first stage) instruments., Adaptive Lasso is the model which TSLS uses adaptive lasso (optimal inequality  $\lambda$ ) selected (first stage) instruments.

Table 4.20: Summary statistics of TSLS for Model 4'

	Post-Lasso					Adaptive Lasso ( $\lambda$ )				
	Bias	MSE	s.e.	MBias	MSP	Bias	MSE	s.e.	MBias	MSP
$n = 60, \sigma = 2, \rho = .5$	.087	.061	.232	.080	.822	.096	.061	.227	.092	.824
$n = 60, \sigma = 2, \rho = .99$	.102	.058	.219	.173	.802	.115	.054	.202	.211	.784
$n = 120, \sigma = 3, \rho = .5$	.085	.068	.247	.080	.832	.160	.065	.198	.172	.700
$n = 120, \sigma = 3, \rho = .99$	.108	.065	.231	.166	.834	.176	.062	.175	.317	.726
$n = 300, \sigma = 3, \rho = .5$	-.013	.034	.184	-.015	.956	-.015	.034	.183	-.014	.992
$n = 300, \sigma = 3, \rho = .99$	.013	.034	.185	-.014	.954	.008	.033	.182	-.005	.982
$n = 1000, \sigma = 3, \rho = .5$	.002	.009	.096	-.002	.908	-.001	.009	.097	-.004	.992
$n = 1000, \sigma = 3, \rho = .99$	.002	.010	.098	-.009	.912	-.001	.010	.098	-.012	.994

Post-Lasso is the model which TSLS uses Post-Lasso estimated (first stage) instruments., Adaptive Lasso is the model which TSLS uses adaptive lasso (optimal inequality  $\lambda$ ) selected (first stage) instruments.

Table 4.21: Correct Model Selection Percentage for Just Identified Model (One weak  $.1/\sqrt{n}$ )

	Post Lasso	Adaptive Lasso (BIC)	Adaptive Lasso ( $\lambda$ )
$n = 60, \sigma = 2, \rho = .5$	.988	.944	.992
$n = 60, \sigma = 2, \rho = .99$	.988	.960	.994
$n = 60, \sigma = 3, \rho = .5$	.980	.956	.986
$n = 60, \sigma = 3, \rho = .99$	.990	.958	.996
$n = 120, \sigma = 3, \rho = .5$	.994	.986	.998
$n = 120, \sigma = 3, \rho = .99$	.974	.948	.996
$n = 300, \sigma = 3, \rho = .5$	.992	.984	1.000
$n = 300, \sigma = 3, \rho = .99$	.986	.988	.998
$n = 1000, \sigma = 3, \rho = .5$	.976	.984	1.000
$n = 1000, \sigma = 3, \rho = .99$	.982	.992	1.000

Adaptive Lasso (BIC), Adaptive Lasso ( $\lambda$ ) are adaptive lasso using BIC and optimal inequality  $\lambda$  tuning methods respectively.

## 4.6 Adaptive Lasso V.S. AIC and BIC

In this section we compare TSLS performance of adaptive lasso with that of AIC and BIC model selection. The model we use is

$$y_i = \beta x_i + \epsilon_i \quad (4.3)$$

$$x_i = z_i \gamma + \nu_i \quad (4.4)$$

where  $z_i = [z_{1i} \ z_{2i} \ \dots \ z_{9i}]$ ,  $\gamma = (\gamma_1 \ \gamma_2 \ \dots \ \gamma_9)^T$ . The model setup is the same as (4.1) and (4.2) except that the dimension of  $\gamma$  is 9 and the distribution of  $z_i$  is now  $N(0, \Sigma_z)$ . Let

$$Q = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{9}} & \frac{1}{\sqrt{9}} & \dots & \frac{1}{\sqrt{9}} \end{bmatrix} \text{ a triangular matrix. } Q_{ij} = \frac{1}{\sqrt{i}}, \text{ for } i \geq j \text{ and } Q_{ij} = 0, \text{ for } i < j.$$

$\Sigma_z = Q^T Q$ . The true value of  $\gamma$  is:

**Model 7** : four nonzero (strong) and five exact zero coefficients  $\gamma = (1, 0, .8, 0, .7, 0, 0, 0, .9)'$

**Model 8** : four nonzero (strong) and five exact zero coefficients  $\gamma = (1, \frac{t}{\sqrt{n}}, .8, \frac{t}{\sqrt{n}}, .7, \frac{t}{\sqrt{n}}, \frac{t}{\sqrt{n}}, \frac{t}{\sqrt{n}}, .9)'$ , where  $t$  is a constant real number

**Model 8'** : four nonzero (strong) and five exact zero coefficients  $\gamma = (1, .1, .8, .1, .7, .1, .1, .1, .9)'$ , where  $t$  is a constant real number

Adaptive lasso beats AIC in correct model selection. When sample size is small and the

noise is relatively big, BIC performs badly. Adaptive lasso is the best all-around considering both bias and model selection. We can also see that BIC and adaptive lasso is consistent in model selection and AIC is not.

Table 4.22: Summary statistics for Model 7: AIC / BIC v.s. adaptive lasso

	AIC			BIC			Adaptive Lasso		
	Bias	MSP	MSE	Bias	MSP	MSE	Bias	MSP	MSE
$n = 60, \sigma = 1.5, \rho = .5$	.024	.10	.004	.025	.07	.004	.017	.14	.004
$n = 60, \sigma = 1.5, \rho = .99$	.029	.11	.005	.023	.06	.005	.023	.16	.005
$n = 120, \sigma = 1.5, \rho = .5$	.017	.29	.002	.016	.32	.002	.015	.36	.002
$n = 120, \sigma = 1.5, \rho = .99$	.007	.29	.002	.005	.28	.002	.004	.39	.002
$n = 300, \sigma = 3, \rho = .5$	.023	.17	.005	.023	.01	.005	.021	.24	.005
$n = 300, \sigma = 3, \rho = .99$	.036	.16	.003	.030	.03	.003	.029	.22	.003
$n = 300, \sigma = 1.5, \rho = .5$	.006	.48	.001	.006	.77	.001	.006	.72	.001
$n = 300, \sigma = 1.5, \rho = .99$	.010	.44	.001	.010	.75	.001	.010	.67	.001
$n = 500, \sigma = 3, \rho = .5$	.013	.29	.002	.012	.25	.002	.011	.36	.002
$n = 500, \sigma = 3, \rho = .99$	.012	.29	.002	.012	.18	.002	.009	.35	.002
$n = 500, \sigma = 1.5, \rho = .5$	.006	.52	.001	.005	.91	.001	.005	.86	.001
$n = 500, \sigma = 1.5, \rho = .99$	.001	.43	.001	.001	.89	.001	.001	.82	.001
$n = 1000, \sigma = 1.5, \rho = .5$	.001	.46	.001	.001	.96	.001	.001	.90	.002
$n = 1000, \sigma = 1.5, \rho = .99$	.001	.49	.001	-.001	.97	.001	-.001	.93	.001

MSP is the rate of correct model selection

Table 4.23: Summary statistics for Model 8 ( $t = .1$ ): AIC / BIC v.s. adaptive lasso

	AIC			BIC			Adaptive Lasso		
	Bias	MSP	MSE	Bias	MSP	MSE	Bias	MSP	MSE
$n = 60, \sigma = 1.5, \rho = .5$	.025	.11	.005	.027	.04	.005	.023	.16	.005
$n = 60, \sigma = 1.5, \rho = .99$	.019	.06	.003	.015	.09	.003	.015	.16	.003
$n = 120, \sigma = 1.5, \rho = .5$	.005	.33	.002	.007	.28	.002	.002	.39	.002
$n = 120, \sigma = 1.5, \rho = .99$	.023	.22	.003	.021	.24	.003	.020	.31	.003
$n = 300, \sigma = 3, \rho = .5$	.023	.17	.005	.022	.01	.005	.021	.24	.005
$n = 300, \sigma = 3, \rho = .99$	.033	.16	.003	.030	.05	.003	.029	.22	.003
$n = 300, \sigma = 1.5, \rho = .5$	.002	.41	.001	.001	.75	.001	.001	.68	.001
$n = 300, \sigma = 1.5, \rho = .99$	.002	.40	.001	.002	.74	.001	.002	.65	.001
$n = 500, \sigma = 3, \rho = .5$	.009	.29	.002	.007	.21	.002	.007	.33	.002
$n = 500, \sigma = 3, \rho = .99$	.025	.28	.002	.024	.29	.002	.022	.34	.002
$n = 500, \sigma = 1.5, \rho = .5$	.003	.45	.001	.003	.87	.001	.003	.77	.001
$n = 500, \sigma = 1.5, \rho = .99$	.003	.40	.001	.002	.88	.001	.001	.76	.001
$n = 1000, \sigma = 1.5, \rho = .5$	.003	.47	.001	.003	.98	.001	.003	.94	.001
$n = 1000, \sigma = 1.5, \rho = .99$	.001	.40	.001	-.001	.97	.001	-.001	.94	.001

MSP is the rate of correct model selection

## 4.7 Adaptive Lasso and Best Subset Selection

The advantage of adaptive lasso compared to other traditional model selection method such as best subset selection using AIC/BIC is of threefold. First, when the number of variables is large, the computation of subset selection becomes infeasible. Second, subset selection is instable because of its discreteness (Breiman 1995). Small changes in the data can result in very different model selection. Third, step wise selection ignores stochastic errors in the variable selection process. We illustrate this problem and compare the model performance of adaptive lasso and best subset selection using a simple example. In the reduced form equation (4.4),  $z_i \sim N(0, \Sigma_z)$ .  $Q_{ij} = \frac{1}{\sqrt{i}}$ , for  $i \geq j$  and  $Q_{ij} = 0$ , for  $i < j$ .  $\Sigma_z = Q^T Q$ , we have true value of

**Model 9** :  $\gamma = (1, 0, .7, 0, .9, 0, 0, 0)$

**Model 10** :  $\gamma = (1, \frac{t}{\sqrt{n}}, .7, \frac{t}{\sqrt{n}}, .9, \frac{t}{\sqrt{n}}, \frac{t}{\sqrt{n}}, \frac{t}{\sqrt{n}})$

Table 4.24 and Table 4.25 show the result. In small samples adaptive lasso has significant advantage over BIC. If we increase the sample size to 1000, BIC is not dominated by adaptive lasso, since both are consistent method.

Table 4.24: Success rate of model selection: Model 9

	BIC	Adaptive Lasso
$n = 60, \sigma = 1.5$	.15	.30
$n = 120, \sigma = 1.5$	.41	.52
$n = 300, \sigma = 1.5$	.84	.85
$n = 1000, \sigma = 1.5$	.98	.98

Table 4.25: Success rate of model selection: Model 10

	BIC	Adaptive Lasso
$n = 60, \sigma = 1.5$	.08	.20
$n = 120, \sigma = 1.5$	.76	.64
$n = 300, \sigma = 1.5$	.93	.86
$n = 1000, \sigma = 1.5$	.95	.93

In Chapter 3 we show adaptive lasso can estimate parameters of weak instruments as exact

zero even in finite samples. We use this property to select strong instruments. Notice in the weak instruments case, the correct variable selection implies the weak instruments will be estimated as zero.

## Chapter 5

# Returns to Education Revisited

We attempt to give a theoretical guideline for instruments selection by adaptive lasso in previous chapters. Our Monte Carlo study shows that adaptive lasso method performs better compared to other existing methods in the perspective of reducing finite sample bias and correct model selection. It is very important to point out that our method is computationally efficient so that empirical researcher would find that adaptive lasso first stage selection is easy to implement. It is thus necessary to study the instruments selection problem in a real data set as a demonstration of how our method works. In this chapter we replicate and compare the adaptive lasso method with the celebrated empirical study of returns to education by Angrist and Krueger (1991) (Hereafter referred to as AK 1991). We make important contribution to the empirical results by improving the bias term.

Let us briefly introduce the research question raised by AK 1991. One important question in labor economics is that, will one extra year of compulsory education increase the wage of a worker? School education is one of the crucial factors for human capital accumulation, which will in turn improve the labor's performance at work. There are many other forms of human capital accumulation which are heterogeneous, such as professional training (non-degree) and learning by doing. But some of these activities are not recorded in the survey questions such as the U.S. census data. Therefore one common problem in the study of returns to education is that the variable 'years of education' could be endogenous due to the correlation of education level and missing variable 'ability' in the error term. In AK 1991, quarter of birth is employed as pivotal instrumental variable which is manipulated to create more instruments. Also in AK 1991, year of birth dummy variables along with the interaction terms are used in the TSLS regression. Our 'instruments pool' is the full instruments set used in AK 1991. So we do not create any more instruments than the original paper.

But as we discussed in Chapter 1, there is no theoretical framework to justify the inclusion of some of the instruments in the model (such as the large number of dummy variables and

interaction terms). In empirical studies, it is a common theme that there is no clear boundary of instruments set other than using the F-test for ‘all or none’ decision (as we show in Chapter 4, it does not work). We apply adaptive lasso method to select the instruments for years of education in the first stage. In the second stage, we include only the selected instruments. We compare the TSLS results using all instruments and the selected instruments as reported in Table 5.1.

First, we briefly describe the model and data of AK 1991. To control for the endogeneity and time trend, they use the following TSLS model:

$$\ln W_i = X_i\beta + \Sigma_c Y_{ic}\xi_c + \rho E_i + \mu_i \quad (5.1)$$

$$E_i = X_i\pi + \Sigma_c Y_{ic}\delta_c + \Sigma_c \Sigma_j Y_{ic}Q_{ij}\theta_{jc} + \epsilon_i \quad (5.2)$$

In the structural equation, the dependent variable  $\ln W$  is the log of weekly wage.  $X$  is the vector of covariates which includes race, marital status, region of residence (SMSA), age, square of age, etc.  $Y$  is the year of birth dummy variable. The variable of interests is years of education  $E$ . Quarter of birth ( $Q$ ) and year of birth ( $Y$ ) dummy variables (and interaction terms) are used as instruments<sup>6</sup>. The sample was drawn from 1980 U.S. Census data. In this replicate we use the cohort of 329,509 men whose birth year is between 1930 and 1939. We use the same variables and data as in AK 1991 Table V. And we will do adaptive lasso model selection in the first stage. We report the TSLS estimate with ‘all instruments’ as in AK 1991 Table V and the adaptive lasso selected instruments in Table 5.1. We interpret the coefficient of years of education  $\hat{\beta}$  as following. If the representative individual increases the years of education by one more year, she would expect to have an increase (decrease if negative  $\hat{\beta}$ ) of wage by  $100 \times \hat{\beta}$  %. For example in column 6 of Table 5.1, if the representative worker has one more year of education, her expected wage will increase by 6.65%.

Table 5.1 contains the TSLS estimate of years of education and other included exogenous variables. For each structural model setup (different included exogenous variables as shown in the very left column), we report the original TSLS estimate of AK 1991 (in odd columns) and the adaptive lasso selected TSLS (in even columns adjacent to the AK 1991 estimate). From the last row of Table 5.1 we see that adaptive lasso method always selects a subset of the instruments set. This result implies that each full model (with all instruments) has unwanted instruments. In our first stage adaptive lasso estimation, year of birth dummy variables are in most of the selected sets. But some year and quarter of birth interaction terms are not selected (shrunk to zero by adaptive lasso). These instruments that we throw out may satisfy the orthogonality condition but they are not correlated with the endogenous variable hence they are very weak. By doing adaptive lasso shrinkage, we are able to get a desired parsimonious

---

<sup>6</sup>Included exogenous variables are also used as instruments in TSLS regression.

model with less ‘noise’ from many junk instruments. This is important for finite sample bias term especially when the instruments set is large and sparse such as the one used in AK 1991. Therefore our method can give us the guideline in empirical studies that which instrument is strong and which one is weak. In table 5.1, the TOLS estimates are similar except for columns 5 and 6. The adaptive lasso TOLS estimate in column 6 is much smaller than the original estimate and less significant. This result gives us another angle to look at the effect of years of education. It is also clearly shown in the first row of the table, the adaptive lasso TOLS estimator is different from the AK 1991. The difference is due to finite sample properties and different instrument set.

Another concern of this study is that quarter of birth is a weak instrument (Bound et al, 1995). The conventional asymptotic properties of TOLS will not hold under this scenario. Confidence intervals will have distorted coverage of the true coefficient. We compare adaptive lasso and conditional likelihood ratio (CLR) confidence interval reported in Table 3 of Cruz and Moreira (2005). In that paper they emphasize on the weak IV problem. It is likely that quarter of birth itself is exogenous, but it is very weakly correlated to years of education. Therefore all the instruments that are created using this instrument might also be weak. Cruz and Moreira (2005) suggests using first stage F-test as an indicator of the strength of instruments. If the instruments were very weak, more powerful test such as conditional likelihood ratio or Wald should be reported to give the correct confidence interval. We also think weak IV is a realistic problem in this case, so it might be a good idea to look for other strong instruments.

To compare the CLR confidence interval by Cruz and Moreira (2005), we report here the confidence interval of the adaptive lasso selected TOLS estimator. The adaptive lasso confidence interval comparable to model II in Table 5.2 is [.0421, .0909]. The CLR confidence interval is [.047, .122]. CLR confidence interval is larger compared to that of adaptive lasso albeit the latter is lightly skewed to the left. So without the variables age and squared age, the coefficient of years of education is statistically significant and the effect is clearly positive. With the concerns of weak instruments, the researcher should consider using a more powerful test such as CLR. But we can also see that sometimes the confidence interval could be meaningless such as in the cohort of men born 1920 – 1929, the CLR confidence interval is  $(-\infty, \infty)$  (Table 2 of Cruz and Moreira, 2005). The adaptive lasso confidence interval comparable to model III in Table 5.2 is [−.0100, .1324]. The CLR confidence interval is [−.212, .267]. So with the addition of variables age and squared age, the coefficient on years of education is not statistically significant from 0. CLR confidence interval is wider (on both ends) than that of adaptive lasso. In our TOLS regression, if we add age and squared age, the data matrix is prone to computational singularity. In conclusion if it were true that quarter of birth is very weak, then the CLR confidence interval can reject the false null hypothesis with higher probability (Moreira, 2003).

Table 5.1: TOLS And Adaptive Lasso TOLS Estimate of Return to Education

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Covariates	TOLS	AD TOLS	TOLS	AD TOLS	TOLS	AD TOLS	TOLS	AD TOLS
Years of Education	.0891 (.0161)	.0869 (.0199)	.0760 (.0290)	.0702 (.0420)	.0806 (.0164)	.0665 (.0244)	.0600 (.0299)	.0612 (.0712)
Race	–	–	–	–	-.2302 (.0261)	-.2524 (.0387)	-.2626 (.0458)	-.2607 (.1121)
SMSA	–	–	–	–	-.1581 (.0174)	-.1729 (.0258)	-.1797 (.0305)	-.1784 (.0747)
Married	–	–	–	–	.2440 (.0049)	.2471 (.0063)	.2486 (.0073)	.2483 (.0163)
9 Year of birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region of residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
Age	–	–	-.0801 (.0645)	-.0011 (.0049)	–	–	-.0741 (.0626)	-.0894 (.1556)
Age squared	–	–	.0008 (.0007)	-.0001 (.0001)	–	–	.0007 (.0007)	.0009 (.0018)
Number of instruments	39	17	41	17	50	24	52	26

TOLS is the estimate directly from AK 1991. AD TOLS is the estimate of adaptive lasso selected model. Standard errors are in parentheses. Number of instruments is the the number of selected instruments which are used in TOLS regression (for AK 1991, all instruments are used).

Table 5.2: Cruz and Moreira (2005) replication (Table 3 of the paper)

	I	II	III	IV
OLS	.0632	.0632	.0632	.0628
Standard error	.0003	.0003	.0003	.0003
TOLS	.0990	.0806	.0600	.0811
Standard error	.0207	.0164	.0290	.0109
LIML	.0999	.0838	.0574	.0982
Standard error	.0210	.0179	.0385	.0153
Anderson-Rubin	[.052,.153]	[-.002,.179]	[-.441,.493]	[-.015,.240]
Wald	[.059,.140]	[.052,.110]	[.012,.112]	[.060,.102]
Conditional Wald	[.059,.138]	[.052,.118]	[-.231,.353]	[.059,.142]
Score	[.059,.144]	[.048,.122]	[-.079,.192]	[.057,.142]
Conditional Likelihood Ratio	[.059,.144]	[.047,.122]	[-.212,.267]	[.056,.141]
F (first stage)	30.53	4.75	1.61	1.87
F (overidentification)	1.16	.78	.72	.92
Partial $R^2$ (excluded instruments $\times$ 100)	.028	.043	.014	.101
Age, Age <sup>2</sup>	no	no	yes	yes
State of birth	no	no	no	yes
Quarter of birth	yes	yes	yes	yes
Quarter of birth, year of birth	no	yes	yes	yes
Quarter of birth, state of birth	no	no	no	yes
Number of instruments	3	30	28	178

All specifications include a constant, race, metropolitan area, married dummies, eight regional dummies and nine year-of-birth dummies as controls.

## Chapter 6

# Conclusion

In this dissertation we have proposed the adaptive lasso for instrumental variables selection. We have shown that adaptive lasso enjoys the oracle properties when the instruments are irrelevant (parameter equals 0). And for the small but not exactly zero parameters (weak instruments), adaptive lasso treats those as zero asymptotically. In simulations we show that, TSLS estimator bias will improve as weak instruments eliminated. We also extend the adaptive lasso method in a TSLS model that performs model selection in both stages. Further investigation is needed to simultaneously select the correct model in the structural equation with the strong instruments in the reduced form. It is also worthwhile to study the finite sample property of TSLS estimator so that we can understand the bias reduction mechanism better.

## REFERENCES

- [1] Anderson, P. K. and Gill, R. D. (1982), “Coxs Regression Model for Counting Processes: A Large Sample Study”, *The Annals of Statistics*, 10, 1100–1120.
- [2] Andrews, Donald W. K. (1999), “Consistent Moment Selection Procedures for Generalized Method of Moments”, *Econometrica*, 67, 543–564.
- [3] Andrews, D.W.K and B. Lu (2001), “Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models”, *Journal of Econometrics*, 101, 123–165.
- [4] Andrews, Donald W. K., Moreira, Marcelo J. and Stock, James H. (2006), “Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression”, *Econometrica*, 74, 715–52.
- [5] Angrist, J. D., Imbens, G. W. and Krueger, A. B. (1999), “Jackknife Instrumental Variables Estimation,” *Journal of Applied Econometrics*, 14, 57–67.
- [6] Angrist, J. D. and Krueger, A. B. (1991), “Does Compulsory School Attendance Affect Schooling and Earnings,” *The Quarterly Journal of Economics*, 106, 979–1014.
- [7] Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C. “Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain”, working paper.
- [8] Breiman, L. (1995), “Better Subset Regression Using the Nonnegative Garotte”, *Technometrics*, 37, 373–384.
- [9] Breiman, L. (1996), “Heuristics of Instability and Stabilization in Model Selection”, *The Annals of Statistics*, 24, 2350–2383.
- [10] Bound, J., Jaeger, D. A. and Baker, R. M. (1995), “Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak”, *Journal of the American Statistical Association*, 90, 443–450.
- [11] Caner, M. (2009), “Lasso-type GMM estimator”, *Econometric Theory*, 25, 270–290.
- [12] Caner, M. (2007), “Boundedly Pivotal Structural Change Tests in Continuous Updating GMM with Strong, Weak Identification and Completely Unidentified Cases”, *Journal of Econometrics*, 137, 28–67.
- [13] Caner, M., “Pivotal Structural Change Tests in Linear Simultaneous Equations Models with Weak Identification”, *Econometric Theory*, forthcoming.
- [14] Caner, M. (2010), “Testing, Estimation in GMM and CUE with Nearly-Weak Identification”, *Econometric Reviews*, 29, 330–363.
- [15] Caner, M., “Exponential Tilting with Weak Instruments: Estimation and Testing”, *Oxford Bulletin of Economics and Statistics*, 72, 307–326.

- [16] Cragg, J. G. and Donald S. G. (1997) “Testing Identifiability and Specification in Instrumental Variable Models”, *Econometric Theory*, 9, 222–240.
- [17] Cruz, L. M. and Donald M. J. (2005) “On the Validity of Econometric Techniques with Weak Instruments Inference on Returns to Education Using Compulsory School Attendance Laws”, *The Journal Of Human Resources*, 40, 393–410.
- [18] Donald, S. G. and Newey, W. (2001), “Choosing the Number of Instruments”, *Econometrica*, 69, 1161–1191.
- [19] Donoho, D. and Johnstone, I. (1994), “Ideal Spatial Adaptation via Wavelet Shrinkages,” *Biometrika*, 81, 425–455.
- [20] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), “Least Angle Regression”, *The Annals of Statistics*, 32, 407–499.
- [21] Fan, J. and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties”, *Journal of the American Statistical Association*, 96, 1348–1360.
- [22] Fan, J. and Peng, H. (2004), “On Nonconcave Penalized Likelihood with Diverging Number of Parameters”, *The Annals of Statistics*, 32, 928–961.
- [23] Friedman, J., Hastie, T. and Tibshirani, R. (2007). “Regularization Paths for Generalized Linear Models via Coordinate Descent”, *The Annals of Applied Statistics*, 1, 302–332.
- [24] Fu, W. (1998). “Penalized Regression: the Bridge Versus the Lasso”, *Journal of computational and graphical statistics*, 7, 397–416.
- [25] Fuller, W. A. (1977). “Some Properties of a Modification of the Limited Information Estimator”, *Econometrica*, 45, 939–954.
- [26] Hall, A. and Peixe, F. (2003), “A Consistent Method for the Selection of Relevant Instruments”, *Econometric Reviews*, 22, 269–287.
- [27] Hahn J. and Hausman, J. (2002), “A New Specification Test for The Validity of Instrumental Variables”, *Econometrica*, 70, 163–189.
- [28] Hansen, B. E. (2007). “Least Squares Model Averaging”, *Econometrica*, 75, 1175–1189.
- [29] Hansen, C., Hausman, J. and Newey, W. K. (2004), “Many Instruments, Weak Instruments, and Microeconomic Practice”, Working Paper, M.I.T.
- [30] Kleibergen, F. (2002), “Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression,” *Econometrica*, 70, 1781–1803.
- [31] Knight, K. and Fu, W. (2000), “Asymptotics for Lasso Type Estimators”, *The Annals of Statistics*, 28, 1356–1378.
- [32] Kuersteiner, G. and Okui, R. (2010), “Constructing Optimal Instruments by First-stage Prediction Averaging”, *Econometrica*, 78, 697–718.

- [33] Moreira, M. (2003), “A Conditional Likelihood Ratio Test for Structural Models”, *Econometrica*, 71, 1027–1048.
- [34] Nagar, A. L. (1959), ‘The Bias and Moment Matrix of The General K–class Estimators of The Parameters in Simultaneous Equations’, *Econometrica*, 27, 575–595.
- [35] Nelson, C. R. and Startz, R. (1990), “The Distribution of the Instrumental Variable Estimator and Its t–Ratio When the Instrument Is a Poor One”, *Journal of Business*, 63, S125–S140.
- [36] Okui, R. (2009), “The Optimal Choice of Moments in Dynamic Panel Data Methods”, *Journal of Econometrics*, 151, 1–16.
- [37] Pollard, D. (1991), “Asymptotics for Least Absolute Deviation Regression Estimators”, *Econometric Theory*, 7, 186–199.
- [38] Staiger, D. and Stock, J. (1997), “Instrumental Variables Regression with Weak Instruments”, *Econometrica*, 65, 557–586.
- [39] Stock, J. and Wright, J. (2000), “GMM with Weak Identification”, *Econometrica*, 68, 1055–1096.
- [40] Stock, J., Wright, J. and Yogo, M. (2002), “A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments”, *Journal of the American Statistical Association*, 20, 518–529.
- [41] Tibshirani, R. (1996), “Regression Shrinkage and Selection via The Lasso”, *Journal of the Royal Statistical Society*, 58, 267–288.
- [42] Wang, H. and Leng C. (2007) “Unified LASSO estimation via least squares approximation”, *Journal of the American Statistical Association*, 102, 1039–1048.
- [43] Wang, H., Li, R. and Tsai, C. (2007), “Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method”, *Biometrika*, 94, 553–568.
- [44] Wright, J. (2003), “Detecting Lack of Identification in GMM”, *Econometric Theory*, 19, 322–330.
- [45] Zhang, H. and Lu, W. (2007), “Adaptive–Lasso for Coxs Proportional Hazards Model”, *Biometrika*, 94, 691–703.
- [46] Zou, H. (2006), “The Adaptive lasso and its oracle properties”, *Journal of the American Statistical Association*, 101, 1418–1429.
- [47] Zou, H. and Zhang, H. (2009), “On the Adaptive Elastic–net with a Diverging Number of Parameters”, *The Annals of Statistics*, 37, 1733–1751.

## APPENDICES

# Appendix A

## Proofs

### A.1 Proof of Theorem 1

First we show consistency of adaptive lasso estimator. The objective function we minimize is

$$(X_V - Z_K \gamma_V)'(X_V - Z_K \gamma_V) + \lambda_n \sum_{k=1}^p \sum_{j=1}^q \hat{w}_{jk} |\gamma_{jk}| \quad (\text{A.1})$$

Define the function

$$L_n(\gamma) = \frac{1}{n} (X_V - Z_K \gamma_V)'(X_V - Z_K \gamma_V) + \frac{\lambda_n}{n} \sum_{k=1}^p \sum_{j=1}^q \hat{w}_{jk} |\gamma_{jk}| \quad (\text{A.2})$$

and  $L_n(\gamma)$  is minimized at  $\gamma = \hat{\gamma}_n$  by the definition of  $\hat{\gamma}_n$ .

Now consider each term of (A.2).

$$\begin{aligned} & \frac{1}{n} (X_V - Z_K \gamma_V)'(X_V - Z_K \gamma_V) \\ &= \frac{1}{n} [\nu_V - Z_K(\gamma_V - \text{vec}(\gamma^*))]' [\nu_V - Z_K(\gamma_V - \text{vec}(\gamma^*))] \\ &= \nu_V' \nu_V / n - \frac{1}{n} 2\nu_V' Z_K(\gamma_V - \text{vec}(\gamma^*)) + (\gamma_V - \text{vec}(\gamma^*))' \frac{Z_K' Z_K}{n} (\gamma_V - \text{vec}(\gamma^*)) \\ &\rightarrow p\sigma_V^2 + (\gamma_V - \text{vec}(\gamma^*))' C (\gamma_V - \text{vec}(\gamma^*)) \end{aligned}$$

by Assumption 2.1 and Assumption 2.2.

Then the penalty term,

Since  $\hat{w}_{jk} = \frac{1}{|\tilde{\gamma}_{jk}|^\tau}$  and  $\tilde{\gamma}_{jk} = O_p(\frac{1}{\sqrt{n}})$ , we have  $\hat{w}_{jk} = O_p(n^{\frac{\tau}{2}})$ .

If we have  $\frac{\lambda_n \hat{w}_{jk}}{n} = \frac{\lambda_n O_p(n^{\frac{\tau}{2}})}{n} \rightarrow_p 0$ , which boils down to  $\lambda_n n^{\frac{\tau}{2}-1} \rightarrow 0$ . And this is true

by Assumption 2.4.

In Zou (2001) p. 1420, it is possible that  $\tilde{\gamma}$  converges to  $\gamma^*$  in a much slower rate than  $\sqrt{n}$ , then the estimation consistency proof here is trivial, it basically follows Theorem 1 of Knight and Fu (2000).

$\hat{\gamma}_n$  is consistent if  $\hat{\gamma}_n \rightarrow_p \arg \min(L)$  where

$$L(\gamma) = [\gamma_V - \text{vec}(\gamma^*)]' C [\gamma_V - \text{vec}(\gamma^*)]$$

We need to show that

$$\sup_{\gamma \in K} |L_n(\gamma) - L(\gamma) - p\sigma_V^2| \rightarrow_p 0 \quad (\text{A.3})$$

for any compact set  $K$  and that

$$\hat{\gamma}_n = O_p(1) \quad (\text{A.4})$$

$L_n$  is convex, thus (A.3) and (A.4) follow from the point-wise convergence in probability of  $L_n(\gamma)$  to  $L(\gamma) + p\sigma_V^2$  by applying standard results of [Anderson and Gill (1982); Pollard (1991)].

Then we show asymptotic normality. Following the proof of Zou (2006), let

$$\gamma_{jk} = \gamma_{jk}^* + \frac{u_{jk}}{\sqrt{n}} \quad j = 1, 2, \dots, q; \quad k = 1, 2, \dots, p$$

And

$$\Psi_n(u) = [X_V - Z_K(\gamma_V + \frac{u_V}{\sqrt{n}})]' [X_V - Z_K(\gamma_V + \frac{u_V}{\sqrt{n}})] + \lambda_n \sum_{k=1}^p \sum_{j=1}^q \hat{w}_{jk} |\gamma_{jk}^* + \frac{u_{jk}}{\sqrt{n}}| \quad (\text{A.5})$$

where  $u_V = \text{vec}(u)$ , notice we now minimize the objective function with respect to  $u$  instead of  $\gamma$ .

Let  $\hat{u}_n = \arg \min_u \Psi_n(u)$ , then  $\text{vec}(\hat{\gamma}_n) = \text{vec}(\gamma^*) + \frac{\text{vec}(\hat{u}_n)}{\sqrt{n}}$  or  $\text{vec}(\hat{u}_n) = \sqrt{n}[\text{vec}(\hat{\gamma}_n) - \text{vec}(\gamma^*)]$ , which is the asymptotic form we need.

We set  $\Psi_n(u) - \Psi_n(0) = V_n(u)$ , where

$$V_n(u) = u_V' (\frac{1}{n} Z_K' Z_K) u_V - 2 \frac{\nu_V' Z_K}{\sqrt{n}} u_V + \frac{\lambda_n}{\sqrt{n}} \sum_{k=1}^p \sum_{j=1}^q \hat{w}_{jk} \sqrt{n} (|\gamma_{jk}^* + \frac{u_{jk}}{\sqrt{n}}| - |\gamma_{jk}^*|) \quad (\text{A.6})$$

The original problem is equivalent to minimize  $V_n(u)$ .

Now consider each term in (A.6). By Assumption 2.2,  $\frac{1}{n} Z_K' Z_K \rightarrow C$ . By Assumption 2.1,  $\frac{\nu_V' Z_K}{\sqrt{n}} \rightarrow_d W \equiv N(0, \sigma^2 C)$ .

Now consider the limiting behavior of the penalty term of  $V_n(u)$ .

If  $\gamma_{jk}^* \neq 0$ , then  $\hat{w}_{jk} \rightarrow_p \frac{1}{|\gamma_{jk}^*|^\tau}$ .

And

$$\sqrt{n}(|\gamma_{jk}^* + \frac{u_{jk}}{\sqrt{n}}| - |\gamma_{jk}^*|) \rightarrow u_{jk} \operatorname{sgn}(\gamma_{jk}^*)$$

By Assumption 2.4,  $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$ .

By Slutsky's theorem,  $\frac{\lambda_n}{\sqrt{n}} \hat{w}_{jk} \sqrt{n}(|\gamma_{jk}^* + \frac{u_{jk}}{\sqrt{n}}| - |\gamma_{jk}^*|) \rightarrow_p 0$ .

If  $\gamma_{jk}^* = 0$ , then  $\sqrt{n}(|\gamma_{jk}^* + \frac{u_{jk}}{\sqrt{n}}| - |\gamma_{jk}^*|) = |u_{jk}|$  and  $\frac{\lambda_n}{\sqrt{n}} \hat{w}_{jk} = \frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau}{2}} (\sqrt{n} \tilde{\gamma}_{jk})^{-\tau}$ , where  $\sqrt{n} \tilde{\gamma}_{jk} = O_p(1)$ . By Assumption 2.4,  $\frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau}{2}} \rightarrow \infty$ . Therefore, by Slutsky's theorem,

$$\frac{\lambda_n}{\sqrt{n}} \hat{w}_{jk} \sqrt{n}(|\gamma_{jk}^* + \frac{u_{jk}}{\sqrt{n}}| - |\gamma_{jk}^*|) \rightarrow_p \infty.$$

Now combine above results for (A.6), by Slutsky's theorem,  $V_n(u) \rightarrow_d V(u)$  for every  $u$ , where

$$V(u) = \begin{cases} u'_{\mathcal{V}\mathcal{A}} C_{11} u_{\mathcal{V}\mathcal{A}} - 2u'_{\mathcal{V}\mathcal{A}} W_{\mathcal{A}} & \text{if } u_j = 0 \forall \gamma_j^* \notin \mathcal{A}, \\ \infty & \text{otherwise.} \end{cases} \quad (\text{A.7})$$

And  $u_{\mathcal{V}\mathcal{A}} = \operatorname{vec}(u_j)$ , which is  $q_0 p$ -vector represents the nonzero  $\gamma_j^* \in \mathcal{A}$ .  $W_{\mathcal{A}} = \operatorname{plim} \frac{\nu'_{\mathcal{V}} Z_K}{\sqrt{n}}$ , which is normally distributed,  $q_0 p$ -vector representing the  $\gamma_j^* \in \mathcal{A}$ .  $V_n(u)$  is a convex function, and the unique minimum of  $V_n(u)$  is  $(C_{11}^{-1} W_{\mathcal{A}}, 0)'$ . Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), let  $\hat{u}_{\mathcal{V}\mathcal{A}} = \operatorname{vec}(\hat{u}_j)$ ,  $\hat{\gamma}_j \in \mathcal{A}_n$  and  $\hat{u}_{\mathcal{V}\mathcal{A}^c} = \operatorname{vec}(\hat{u}_j)$ ,  $\hat{\gamma}_j \notin \mathcal{A}_n$ , we have:

$$\hat{u}_{\mathcal{V}\mathcal{A}} \rightarrow_d C_{11}^{-1} W_{\mathcal{A}} \quad (\text{A.8})$$

$$\hat{u}_{\mathcal{V}\mathcal{A}^c} \rightarrow_d 0 \quad (\text{A.9})$$

Since  $W_{\mathcal{A}} = N(0, \sigma^2 C_{11})$ ,  $\hat{u}_{\mathcal{V}\mathcal{A}} \rightarrow_d N(0, \sigma_{\mathcal{V}}^2 C_{11}^{-1})$ , we proved the asymptotic normality.

Now we show the variable selection consistency part:  $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$ .

$\forall \gamma_j^* \in \mathcal{A}$ , the asymptotic normality indicates that  $\hat{\gamma}_j \rightarrow_p \gamma_j^*$ , thus  $P(\hat{\gamma}_j \in \mathcal{A}_n) \rightarrow 1$ . Therefore to show variable selection consistency, it is sufficient to show that

$$\forall \gamma_{j'}^* \notin \mathcal{A}, P(\hat{\gamma}_{j'} \in \mathcal{A}_n) \rightarrow 0$$

We show the impossibility of interior solution of multivariate adaptive lasso. Consider the event  $\hat{\gamma}_{j'} \in \mathcal{A}_n$ . By the Karush-Kuhn-Tucker optimality conditions, we know that for each  $\hat{\gamma}_{j'k}$ ,  $k = 1, 2, \dots, p$ , we have

$$2Z'_{j'}(X_k - Z\hat{\gamma}_k) = \lambda_n \hat{w}_{j'k} \quad (\text{A.10})$$

where  $Z_{j'}$  ( $n \times 1$ ) and  $X_k$  ( $n \times 1$ ) are the  $j'^{th}$  and  $k^{th}$  column of  $Z$ ,  $X$  matrix respectively,  $\hat{\gamma}_k$  ( $q \times 1$ ) is the  $k^{th}$  column of the  $\hat{\gamma}$  matrix.

Divide (A.10) by  $\sqrt{n}$ , the RHS

$$\frac{\lambda_n \hat{w}_{j'k}}{\sqrt{n}} = \frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau}{2}} \frac{1}{|\sqrt{n} \tilde{\gamma}_{j'k}|^\tau} \rightarrow_p \infty$$

the LHS

$$\frac{2Z'_{j'}(X_k - Z\hat{\gamma}_k)}{\sqrt{n}} = \frac{2Z'_{j'}Z\sqrt{n}(\gamma_k^* - \hat{\gamma}_k)}{n} + 2\frac{Z'_{j'}\nu_k}{\sqrt{n}}$$

By (A.8), (A.9) and Slutsky's theorem, we know that

$$\frac{2Z'_{j'}Z\sqrt{n}(\gamma_k^* - \hat{\gamma}_k)}{n} \rightarrow_d \text{some normal distribution}$$

and

$$2\frac{Z'_{j'}\nu_k}{\sqrt{n}} \rightarrow_d N(0, 4\sigma_\nu^2 \|Z_{j'}\|^2)$$

Thus,  $P(\hat{\gamma}_{j'} \in \mathcal{A}_n) \leq P(2Z'_{j'}(X_k - Z\hat{\gamma}_k) = \lambda_n \hat{w}_{j'k}) \rightarrow 0$ . Q.E.D.

## A.2 Proof of Theorem 2

The proof of Theorem 2 is essentially the same as those of asymptotic normality part of Theorem 1.

Let

$$\gamma_{jk} = \gamma_{njk} + \frac{u_{jk}}{n}$$

Consider the limiting behavior of the third term of (3.3).

If  $\gamma_j^* \neq 0$ , then  $\hat{w}_{jk} \rightarrow_p \frac{1}{|\gamma_{jk}^*|^\tau}$

and

$$\begin{aligned} & \sqrt{n}(|\gamma_{njk} + \frac{u_{jk}}{\sqrt{n}}| - |\gamma_{njk}|) \\ &= \sqrt{n}(|\gamma_{jk}^* + \frac{u_{jk} + t_{jk}}{\sqrt{n}}| - |\gamma_{jk}^* + \frac{t_{jk}}{\sqrt{n}}|) \\ &\rightarrow u_{jk} \operatorname{sgn}(\gamma_{jk}^*) \end{aligned}$$

By Slutsky's theorem, we have

$$\frac{\lambda_n}{\sqrt{n}} \sum_{k=1}^p \sum_{j=1}^q \hat{w}_{jk} \sqrt{n} (|\gamma_{nj k} + \frac{u_{jk}}{\sqrt{n}}| - |\gamma_{nj k}|) \rightarrow_p 0$$

If  $\gamma_j^* = 0$ , then

$$\sqrt{n} (|\gamma_{nj k} + \frac{u_{jk}}{\sqrt{n}}| - |\gamma_{nj k}|) = \sqrt{n} (|\frac{u_{jk} + t_{jk}}{\sqrt{n}}| - |\frac{t_{jk}}{\sqrt{n}}|) = |u_{jk} + t_{jk}| - |t_{jk}|$$

and  $\frac{\lambda_n}{\sqrt{n}} \hat{w}_{jk} \rightarrow \infty$ .

Thus  $\arg \min_u(V) = -t_j$ ,  $\sqrt{n}(\hat{\gamma}_n - \gamma_n) \rightarrow_d -t_j$ . Q.E.D.

## Appendix B

# Adaptive Lasso type TSLS estimator

In the previous work we showed oracle properties of adaptive lasso in the linear reduced form equation. In this appendix, we will extend the adaptive lasso method to the second stage structural model. Assuming that we have the correct variable selection in the first stage (the model is identifiable with valid instruments), our study focuses on the linear TSLS model which is sparse in the structural equation. Using the instruments selected in first stage, we introduce an adaptive lasso type TSLS estimator which can perform both variable selection and coefficients estimation simultaneously. In Caner (2009), a bridge-type GMM estimator is shown to have oracle properties. We show the same oracle properties hold for adaptive lasso.

### B.1 The Basic Model

Recall the structural equation of interest is

$$y = X\beta^* + \epsilon \tag{B.1}$$

where  $y$  is  $n \times 1$  dependent variable,  $X$  is  $n \times p$  matrix of endogenous variables,  $\epsilon$  is  $n \times 1$  random disturbances,  $\beta^*$  is  $p \times 1$  coefficient vector. Let  $\mathcal{A} = \text{support}(\beta^*) = \{j \in \{1, 2, \dots, p\} : |\beta_j^*| > 0\}$ . The dimension of  $\mathcal{A}$ ,  $\|\mathcal{A}\|_0 = \sum_{j=1}^p 1\{\beta_j^*\} = p_0 < p$ . For simplicity we assume there is no included exogenous variables.

Let  $Z$  be  $n \times q$  matrix of instruments, assume that  $Z$  consists of strong instruments only, and  $q \geq p$ . In Chapter 2 and Chapter 3 we showed that adaptive lasso can select the nonzero instruments with probability 1. In this chapter we develop theoretical adaptive lasso for TSLS based on the asymptotic property that we can select the strong instruments. In simulations, we will test how the adaptive lasso selection in both reduced form and structural equations works against non-selection in structural equation like Donald and Newey (2001).

The adaptive lasso type TSLS estimator is defined as:

$$\hat{\beta}_n = \arg \min_{\beta} \left\{ (y - X\beta)' Z(Z'Z)^{-1} Z'(y - X\beta) + \lambda_n \sum_{j=1}^p \hat{w}_j |\beta_j| \right\} \quad (\text{B.2})$$

Define each element of coefficient vector  $\hat{\beta}_n$  as  $\hat{\beta}_k$ ,  $k = 1, 2, \dots, p$ .  $\mathcal{A}_n$  is the support of adaptive lasso estimates  $\hat{\beta}_n$ . Let  $P_Z = Z(Z'Z)^{-1}Z'$ , the projection matrix of  $Z$ . Therefore we can write (B.2) as

$$\hat{\beta}_n = \arg \min_{\beta} \left\{ (\tilde{y} - \tilde{X}\beta)' (\tilde{y} - \tilde{X}\beta) + \lambda_n \sum_{j=1}^p \hat{w}_j |\beta_j| \right\} \quad (\text{B.3})$$

where  $\tilde{y} = P_Z y$ ,  $\tilde{X} = P_Z X$  and  $\tilde{\epsilon} = P_Z \epsilon$ . We need this transformation so that proofs can be handled easily. Also this format will extend to efficient linear GMM easily as explained in the subsection. Suppose  $\tilde{\beta}$  is a  $\sqrt{n}$  consistent unpenalized estimator to  $\beta^*$ , e.g., the regular 2SLS from second stage after the adaptive lasso variable selection. The adaptive weight  $\hat{w}_j = 1/|\tilde{\beta}_j|^\tau$ , where  $0 < \tau \leq 1$ ,  $j = 1, 2, \dots, p$ . Let  $\mathcal{A}_n = \{\text{the nonzero elements in } \hat{\beta}_n\}$ .

## B.2 Assumptions

We then present the assumptions that are useful in providing the oracle properties for adaptive lasso type TSLS estimator.

**Assumption 1** : the unobserved errors  $\epsilon$  are independent identically distributed (i.i.d.) with mean 0 and variance  $\sigma_\epsilon^2$ , and  $E(\epsilon | Z) = 0$ .

**Assumption 2** :  $C_n = \frac{1}{n} X' P_Z X = \left(\frac{X'Z}{n}\right) \left(\frac{Z'Z}{n}\right)^{-1} \left(\frac{Z'X}{n}\right) \rightarrow \Sigma_{XZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX} = C$ , where  $\frac{X'Z}{n} \rightarrow_p \Sigma_{XZ}$ ,  $\frac{Z'Z}{n} \rightarrow_p \Sigma_{ZZ}$ ,  $\frac{Z'X}{n} \rightarrow_p \Sigma_{ZX}$  and  $C$  is a positive definite matrix and finite. Without loss of generality, we assume the first  $p_0$  variables are nonzero, so that  $\mathcal{A} = \{\beta_1^*, \beta_2^*, \dots, \beta_{p_0}^*\}$ .

Let  $C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$ , where  $C_{11}$  is  $p_0 \times p_0$ ,  $C_{22}$  is  $(p - p_0) \times (p - p_0)$  and  $C_{12} = C'_{21}$ .

**Assumption 3** :  $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$  and  $\lambda_n n^{\frac{\tau-1}{2}} \rightarrow \infty$

The first two assumptions are adopted from the setup of Knight and Fu (2000) for the analysis of large sample theory.

The two components of Assumption 3 are stronger than the assumptions in Knight and Fu (2000). Under Assumption 3 we will have stronger results, namely shrinkage to 0 with probability 1 of irrelevant variables asymptotically.

### B.3 The Limit Theory for Adaptive Lasso IV estimator

In this section we present the asymptotic properties of adaptive lasso estimator. Adaptive lasso has ideal asymptotic properties of estimation and selection consistency. Adaptive lasso can choose the correct model as if it were known, which is also known as the oracle properties.

**Theorem 3** (The consistency of adaptive lasso IV estimator)

1. If  $\lambda_n n^{\frac{\tau}{2}-1} = O_p(1)$ , then  $\hat{\beta}_n \rightarrow_p \operatorname{argmin}(\tilde{L})$  where

$$\tilde{L} = (\beta - \beta^*)' C (\beta - \beta^*) + \sum_{j=1}^p \lambda_{0w_j} |\beta_j|$$

and  $\lambda_{0w_j} \geq 0$

2. If  $\lambda_n n^{\frac{\tau}{2}-1} = o_p(1)$ , then  $\hat{\beta}_n - \beta^* \rightarrow_p 0$

In the proofs, we will need

$$\lambda_n n^{\frac{\tau}{2}-1} = \frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau-1}{2}} \rightarrow 0$$

and this is implied by Assumption 3 since  $0 < \tau \leq 1$ .

This result is necessary for the estimation consistency which is not explicitly shown in Zou (2006).

**Theorem 4** (The oracle properties of adaptive lasso IV estimator) Under Assumption 1-3,

1. Asymptotic normality:  $\sqrt{n}(\hat{\beta}_{\mathcal{A}} - \beta_{\mathcal{A}}) \rightarrow_d N(0, \sigma_{\epsilon}^2 C_{11}^{-1})$
2. Consistency in variable selection:  $\lim_{n \rightarrow \infty} P(\mathcal{A}_n = \mathcal{A}) = 1$

### B.4 Standard Errors of Adaptive Lasso TSLS Estimator

As in Caner (2009), We do not analyze the post-selection inference in this paper. In this section we briefly discuss the standard errors of adaptive lasso TSLS estimator. We follow the standard adaptive lasso standard error sandwich formula by Zou (2006). Efron et al. (2004) Tibshirani (1996), Fan and Li (2001) show that covariance of nonzero penalized estimates can be approximated by iteratively computing the ridge solution. Zou's (2006) standard error formula follows local quadratic approximation (LQA) approach which can provide a consistent sandwich formula for computing the covariance of the estimates of nonzero parameters (Fan and Peng, 2004).

## B.5 Proof of Theorems

### B.5.1 Proof of Estimation Consistency

First we show estimation consistency of adaptive lasso type TSLS estimator.

Set

$$L_n(\beta) = \frac{1}{n}(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta) + \frac{\lambda_n}{n} \sum_{j=1}^p \hat{w}_j |\beta_j| \quad (\text{B.4})$$

Transform  $y$ ,  $X$ ,  $\epsilon$  in the following way

$$\tilde{y} = P_Z y = P_Z(X\beta^* + \epsilon) = \tilde{X}\beta^* + \tilde{\epsilon} \quad (\text{B.5})$$

Use this in the following

$$\frac{1}{n}(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta) = \frac{1}{n}[\tilde{\epsilon} - \tilde{X}(\beta - \beta^*)]'[\tilde{\epsilon} - \tilde{X}(\beta - \beta^*)] \quad (\text{B.6})$$

Now in (B.6) we can write the RHS as following

$$\frac{\tilde{\epsilon}'\tilde{\epsilon}}{n} - \frac{2\tilde{\epsilon}'\tilde{X}(\beta - \beta^*)}{n} + \frac{(\beta - \beta^*)'\tilde{X}'\tilde{X}(\beta - \beta^*)}{n} \quad (\text{B.7})$$

Now simplify each term in (B.7) :

$$\frac{\tilde{\epsilon}'\tilde{\epsilon}}{n} = \frac{\epsilon'P_Z\epsilon}{n} = \left(\frac{\epsilon'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'\epsilon}{n}\right) \rightarrow_p [E(Z'\epsilon)]'(\Sigma_{ZZ})^{-1}[E(Z'\epsilon)] = 0 \quad (\text{B.8})$$

since  $\frac{Z'\epsilon}{n} \rightarrow_p E(Z'\epsilon) = 0$  by Assumption 1 and  $\frac{Z'Z}{n} \rightarrow_p \Sigma_{ZZ}$  by Assumption 2.

Next in (B.7) we have

$$\begin{aligned} \frac{\tilde{\epsilon}'\tilde{X}(\beta - \beta^*)}{n} &= \frac{\epsilon'P_Z X(\beta - \beta^*)}{n} = \left(\frac{\epsilon'Z}{n}\right)\left(\frac{Z'Z}{n}\right)^{-1}\left(\frac{Z'X}{n}\right)(\beta - \beta^*) \\ &\rightarrow_p [E(\epsilon'Z)](\Sigma_{ZZ})^{-1}(\Sigma_{ZX})(\beta - \beta^*) = 0 \end{aligned} \quad (\text{B.9})$$

again  $E(\epsilon'Z) = 0$  by Assumption 1.

Next in (B.7),

$$\frac{(\beta - \beta^*)'\tilde{X}'\tilde{X}(\beta - \beta^*)}{n} = \frac{(\beta - \beta^*)'X'P_Z X(\beta - \beta^*)}{n} \rightarrow_p (\beta - \beta^*)'C(\beta - \beta^*) \quad (\text{B.10})$$

by Assumption 2.

Next combine (B.8) - (B.10) in (B.7) we have

$$\frac{1}{n}(\tilde{y} - \tilde{X}\beta)'(\tilde{y} - \tilde{X}\beta) \rightarrow_p (\beta - \beta^*)'C(\beta - \beta^*) \quad (\text{B.11})$$

So if we had that  $\lambda_n n^{\frac{\tau}{2}-1} = O_p(1)$ , then

$$\frac{\lambda_n \hat{w}_j}{n} \rightarrow_p \lambda_{0w_j}, \quad j = 1, 2, \dots, p$$

where  $\lambda_{0w_j} \geq 0$ , since  $\tilde{\beta}_j = O_p(\frac{1}{\sqrt{n}})$  and  $\hat{w}_j = \frac{1}{|\tilde{\beta}_j|^\tau}$ .

Then,

$$L_n(\beta) \rightarrow_p (\beta - \beta^*)'C(\beta - \beta^*) + \sum_{j=1}^p \lambda_{0w_j} |\beta_j|$$

Since  $L_n(\beta)$  is convex,

$$\sup_{\beta} \left\{ L_n(\beta) - (\beta - \beta^*)'C(\beta - \beta^*) - \sum_{j=1}^p \lambda_{0w_j} |\beta_j| \right\} \rightarrow_p 0$$

and also,

$$\hat{\beta}_n = O_p(1)$$

This follows from the point-wise convergence by applying [Anderson and Gill (1982); Pollard (1991)]. But if  $\lambda_{0w_j} = 0$ , then  $\hat{\beta}_n$  is consistent. So this means  $\lambda_n n^{\frac{\tau}{2}-1} = o_p(1)$  provides consistency.

To see this more clearly,

$$\lambda_n n^{\frac{\tau}{2}-1} = o_p(1)$$

then

$$\arg \min_{\beta} L_n(\beta) = \beta^*$$

Thus  $\hat{\beta}_n$  is consistent.

## B.5.2 Proof of Oracle Property

Now we show asymptotic normality. Following the proof of Zou (2006), let

$$\beta_j = \beta_j^* + \frac{u_j}{\sqrt{n}}, \quad j = 1, 2, \dots, p$$

And define

$$\Psi_n(u) = [\tilde{y} - \tilde{X}(\beta^* + \frac{u}{\sqrt{n}})]' [\tilde{y} - \tilde{X}(\beta^* + \frac{u}{\sqrt{n}})] + \lambda_n \sum_{j=1}^p \hat{w}_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| \quad (\text{B.12})$$

$$\Psi_n(0) = (\tilde{y} - \tilde{X}\beta^*)'(\tilde{y} - \tilde{X}\beta^*) + \lambda_n \sum_{j=1}^p \hat{w}_j |\beta_j^*| \quad (\text{B.13})$$

Notice we now minimize the objective function with respect to  $u$  instead of  $\beta$ .

Let  $\hat{u}_n = \arg \min_u \Psi_n(u)$  or  $\hat{u}_n = \sqrt{n}(\hat{\beta}_n - \beta^*)$ . Since  $\tilde{y} = \tilde{X}\beta^* + \tilde{\epsilon}$  and we can write (B.13) as

$$\Psi_n(0) = \tilde{\epsilon}'\tilde{\epsilon} + \lambda_n \sum_{j=1}^p \hat{w}_j |\beta_j^*| \quad (\text{B.14})$$

and (B.12) as

$$\begin{aligned} \Psi_n(u) &= \left[ \tilde{y} - \tilde{X}\beta^* - \tilde{X} \frac{u}{\sqrt{n}} \right]' \left[ \tilde{y} - \tilde{X}\beta^* - \tilde{X} \frac{u}{\sqrt{n}} \right] + \lambda_n \sum_{j=1}^p \hat{w}_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| \\ &= \left[ \tilde{\epsilon} - \tilde{X} \frac{u}{\sqrt{n}} \right]' \left[ \tilde{\epsilon} - \tilde{X} \frac{u}{\sqrt{n}} \right] + \lambda_n \sum_{j=1}^p \hat{w}_j \left| \beta_j^* + \frac{u_j}{\sqrt{n}} \right| \end{aligned} \quad (\text{B.15})$$

Note that  $\Psi_n(u) - \Psi_n(0) = V_n(u)$ , where

$$V_n(u) = u' \left( \frac{1}{n} \tilde{X}' \tilde{X} \right) u - 2 \frac{\tilde{\epsilon}' \tilde{X}}{\sqrt{n}} u + \frac{\lambda_n}{\sqrt{n}} \sum_{j=1}^p \hat{w}_j \sqrt{n} (|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) \quad (\text{B.16})$$

Then consider each term in (B.16).

$$\frac{1}{n} \tilde{X}' \tilde{X} = \frac{X' P_Z X}{n} = \left( \frac{X' Z}{n} \right) \left( \frac{Z' Z}{n} \right)^{-1} \left( \frac{Z' X}{n} \right) \rightarrow C \quad (\text{B.17})$$

by Assumption 2.

Next since  $\tilde{X} = P_Z X$  and  $\tilde{\epsilon} = P_Z \epsilon$ ,

$$\begin{aligned} \frac{\tilde{\epsilon}' \tilde{X}}{\sqrt{n}} &= \frac{\epsilon' P_Z X}{\sqrt{n}} = \left( \frac{\epsilon' Z}{\sqrt{n}} \right) \left( \frac{Z' Z}{n} \right)^{-1} \left( \frac{Z' X}{n} \right) \\ &\rightarrow_d N(0, \sigma_\epsilon^2 \Sigma_{XZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX}) = D \equiv N(0, \sigma_\epsilon^2 C) \end{aligned} \quad (\text{B.18})$$

Now consider the penalty term of (B.16).

There are 2 cases to analyze,  $\beta_j^* \neq 0$ , the other is  $\beta_j^* = 0$ .

Let's start with  $\beta_j^* \neq 0$ . These are the nonzero parameters in the structural equation.

$$\hat{w}_j \rightarrow_p \frac{1}{|\beta_j^*|^\tau} \quad (\text{B.19})$$

via Lasso GMM by Caner (2009).

Next we have

$$\sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) \rightarrow u_j \text{sgn}(\beta_j^*) \quad (\text{B.20})$$

Then use Slutsky's theorem, given (B.19) and (B.20) and  $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$ ,

$$\frac{\lambda_n}{\sqrt{n}} \hat{w}_j \sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) \rightarrow_p 0 \quad (\text{B.21})$$

Next we consider the zero parameters ( $\beta_j^* = 0$ ) and then we will come back to analysis of nonzero parameters and we will discuss (B.18) and (B.21).

Now we analyze when  $\beta_j^* = 0$ . These are the irrelevant variables in the structural equation. There are two possibilities in this case.

First,

$$\sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) = |u_j| \quad (\text{B.22})$$

Then by multiplying and dividing by  $n^{\frac{\tau}{2}}$  and use  $\hat{w}_j = 1/|\tilde{\beta}_j|^\tau$

$$\frac{\lambda_n}{\sqrt{n}} \hat{w}_j = \frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau}{2}} |\sqrt{n} \tilde{\beta}_j|^{-\tau} \rightarrow +\infty \quad (\text{B.23})$$

where  $\sqrt{n} \tilde{\beta}_j = O_p(1)$  from the first stage  $\sqrt{n}$  consistent estimator (this can be Lasso GMM of Caner (2009)) and by Assumption 3,  $\frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau}{2}} \rightarrow \infty$ .

The first possibility is that  $u_j = 0$ , where  $j \in \mathcal{A}^c$  (since  $\mathcal{A}$  represent the set of nonzero coefficients,  $j \in \mathcal{A}^c$  represents all the  $u_j$ 's that corresponds to the zero coefficients), we have by (B.16)

$$V_n(u) = \Psi_n(u) - \Psi_n(0) \rightarrow_d u'_A C_{11} u_A - 2u'_A D_A \quad (\text{B.24})$$

where  $u_A$  represents the  $u$ 's of nonzero coefficients and  $C_{11}$  represents the upper left block ( $p_0 \times p_0$ ) matrix which is nonsingular and positive definite. And  $D_A$  is the  $p_0 \times 1$  subset (corresponding to the first  $p_0$  elements) of  $D$  vector in (B.18).

But also see that when  $u_j \neq 0$  for  $j \in \mathcal{A}^c$  (the second possibility for  $u$ 's where  $j \in \mathcal{A}^c$ ), then through (B.22) and (B.23), the penalty diverges to  $\infty$  and dominates the distribution hence

$$\Psi_n(u) - \Psi_n(0) \rightarrow +\infty \quad (\text{B.25})$$

So, since  $\Psi_n(u) - \Psi_n(0)$  is convex and the unique minimizer of  $\Psi_n(u) - \Psi_n(0)$  is  $((C_{11}^{-1}D_{\mathcal{A}}, 0_{p-p_0}))'$ . Then follow the epi-convergence of Knight and Fu (2000),

$$\hat{u}_{\mathcal{A}^c} \rightarrow_d u_{\mathcal{A}^c} = 0 \quad (\text{B.26})$$

So we see that when  $u_j = 0$  where  $j \in \mathcal{A}^c$  we have the following simplification

$$u'Cu - 2u'D = u'_{\mathcal{A}}C_{11}u_{\mathcal{A}} - 2u'_{\mathcal{A}}D_{\mathcal{A}}$$

given (B.26).

Now we consider nonzero parameters again. Note that for  $\beta_j^* \neq 0$  we obtain (B.24) by (B.21), since  $\hat{u}_{\mathcal{A}}$  is the unique minimizer of (B.16), we get

$$\hat{u}_{\mathcal{A}} \rightarrow_d C_{11}^{-1}D_{\mathcal{A}} \quad (\text{B.27})$$

Then we show the variable selection consistency:

$$\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$$

$\forall \beta_j^* \in \mathcal{A}$ , the consistency indicates that  $\hat{\beta}_j \rightarrow_p \beta_j^*$ , thus  $P(\hat{\beta}_j \in \mathcal{A}_n) \rightarrow 1$ . Therefore to show variable selection consistency, it is sufficient to show that  $\forall \beta_{j'}^* \notin \mathcal{A}$ ,  $P(\hat{\beta}_{j'} \in \mathcal{A}_n) \rightarrow 0$ .

Consider the event  $\hat{\beta}_{j'} \in \mathcal{A}_n$ . By the Karush-Kuhn-Tucker optimality conditions, we know that for each  $\hat{\beta}_{j'}$ ,  $j' \notin \mathcal{A}$ , we have

$$2\tilde{x}'_{j'}(\tilde{y} - \tilde{X}\hat{\beta}_n) = \lambda_n \hat{w}_{j'}$$

where  $\tilde{x}'_{j'}$  is the  $j'^{th}$  column of  $\tilde{X}$ . By Assumption 3 and  $\tilde{\beta}_j$  being a  $\sqrt{n}$  consistent estimator

$$\frac{\lambda_n \hat{w}_{j'}}{\sqrt{n}} = \frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau}{2}} \frac{1}{|\sqrt{n}\tilde{\beta}_{j'}|^{\tau}} \rightarrow_p \infty$$

and by (B.5)

$$\frac{2\tilde{x}'_{j'}(\tilde{y} - \tilde{X}\hat{\beta}_n)}{\sqrt{n}} = \frac{2\tilde{x}'_{j'}\tilde{X}\sqrt{n}(\beta^* - \hat{\beta}_n)}{n} + 2\frac{\tilde{x}'_{j'}\epsilon}{\sqrt{n}}$$

By (B.17), (B.27) and (B.26) and Slutsky's theorem, given  $\hat{u}_n = \sqrt{n}(\hat{\beta}_n - \beta^*)$  we know that

$$\frac{2\tilde{x}'_{j'}\tilde{X}\sqrt{n}(\beta^* - \hat{\beta}_n)}{n} \rightarrow_d \text{some normal distribution}$$

and

$$2 \frac{\tilde{x}'_j \epsilon}{\sqrt{n}} \rightarrow_d N(0, 4\sigma^2 \|\tilde{x}_{j'}\|^2)$$

Thus,  $P(\hat{\beta}_{j'} \in \mathcal{A}_n) \leq P(2\tilde{x}'_j(\tilde{y} - \tilde{X}\hat{\beta}_j) = \lambda_n \hat{w}_j) \rightarrow 0$ . Q.E.D.

## B.6 Simulation Results

### B.6.1 Adaptive lasso V.S. Post-Lasso and Model Averaging

In this section we compare the results of TSLS Adaptive lasso Estimator with that of Post-Lasso (Belloni et al., 2010) and Model Averaging of KO (2010). Our DGP is as follows:

The structural and reduced form equations are:

$$y = X\beta + \epsilon$$

$$X = Z\gamma + \nu$$

where  $X$  is the  $n \times 4$  matrix of endogenous variables, in our design,  $\beta = (2, 0, 0, 0)$ . So we have a sparse structural equation model.  $Z$  is the  $n \times 5$  matrix of instruments. We assume there are four strong instruments and one irrelevant instrument.

$$\gamma = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$\epsilon$  and  $\nu$  have mean 0,  $\sigma_\epsilon^2 = 1$ ,  $\sigma_\nu^2 = I_4$ .  $\text{corr}(\epsilon_i, \nu_{ij}) = \rho$ ,  $i = 1, \dots, n$ ,  $j = 1, 2, 3, 4$ . In Design 1,  $\rho = .5$ . In Design 2,  $\rho = .1$

In the first stage, we use three methods (adaptive lasso, post-lasso and model averaging) to select (estimate) the instruments respectively. In the second stage, we use adaptive lasso TSLS estimator presented in Appendix B to estimate the model using the selected instruments. For post-lasso and model averaging, we use regular TSLS. We could use Andrews and Lu (2001)'s BIC to select the structural model for the post-lasso and model averaging methods, but the procedure is not very easy to implement and it is left for further studies.

We report the bias and MSE of structural equation coefficients in Table B.1 and Table B.2. Adaptive lasso TSLS estimator is shown to have variable selection consistency as well as estimation consistency. Post-lasso and model averaging do not select the correct model in the second stage.  $n=100$ .

Table B.1: Summary statistics of TSLS for Design 1

	Post Lasso		Model Averaging		Adaptive Lasso TSLS	
	Bias	MSE	Bias	MSE	Bias	MSE
$\beta_1$	.071	.745	.014	.344	-.070	.015
$\beta_2$	.011	.710	.018	.324	0	0
$\beta_3$	.075	.804	.041	.361	0	0
$\beta_4$	-.090	.698	-.017	.339	0	0

Post Lasso is the model which TSLS uses Post-Lasso estimated (first stage) instruments., Adaptive Lasso TSLS is the model which TSLS uses adaptive lasso selected (first stage) instruments and in the second stage TSLS adaptive lasso method is used, MA is KO (2010) method of model averaging.

Table B.2: Summary statistics of TSLS for Design 2

	Post Lasso		Model Averaging		Adaptive Lasso TSLS	
	Bias	MSE	Bias	MSE	Bias	MSE
$\beta_1$	-.013	.454	-.046	.339	-.056	.010
$\beta_2$	.045	.578	.014	.267	0	0
$\beta_3$	-.020	.464	.039	.323	0	0
$\beta_4$	-.086	.709	-.082	.348	0	0

Post Lasso is the model which TSLS uses Post-Lasso estimated (first stage) instruments., Adaptive Lasso TSLS is the model which TSLS uses adaptive lasso selected (first stage) instruments and in the second stage TSLS adaptive lasso method is used, MA is KO (2010) method of model averaging.

We report the bias and MSE of structural equation coefficients in Table B.3 and Table B.4. Adaptive lasso TSLS estimator is shown to have model selection consistency as well as estimation consistency. Post-lasso and one-step adaptive lasso do not select the correct model in the second stage.  $n=300$ .

### B.6.2 Adaptive Lasso V.S. bridge

In previous subsection we select variables in both stages. Now we will focus on the structural equation with the presumption that enough valid instruments are available. We compare the results of TSLS Adaptive lasso Estimator with the bridge estimator of Caner (2009). Our DGP is as follows:

The structural and reduced form equations are:

$$y = X\beta + \epsilon$$

Table B.3: Summary statistics of TSLS for Design 1

	Post Lasso		Adaptive Lasso		ADTS	
	Bias	MSE	Bias	MSE	Bias	MSE
$\beta_1$	-.038	.706	-.030	.641	-.049	.023
$\beta_2$	-.001	.728	.038	.664	0	0
$\beta_3$	.089	.679	.075	.626	0	0
$\beta_4$	-.025	.565	-.018	.531	0	0

Post-Lasso is the model which TSLS uses Post-Lasso estimated (first stage) instruments, ADTS (Adaptive Lasso) is the model which TSLS uses adaptive lasso selected (first stage) instruments and in the second stage TSLS adaptive lasso method is used, Adaptive Lasso is first stage using adaptive lasso and second stage no model selection TSLS.

Table B.4: Summary statistics of TSLS for Design 2

	Post Lasso		Adaptive Lasso		ADTS	
	Bias	MSE	Bias	MSE	Bias	MSE
$\beta_1$	-.017	.542	-.029	.554	-.045	.015
$\beta_2$	-.010	.491	.038	.529	0	0
$\beta_3$	.102	.429	.129	.488	0	0
$\beta_4$	-.091	.529	-.112	.589	0	0

Post-Lasso is the model which TSLS uses Post-Lasso estimated (first stage) instruments, ADTS (Adaptive Lasso) is the model which TSLS uses adaptive lasso selected (first stage) instruments and in the second stage TSLS adaptive lasso method is used, Adaptive Lasso is first stage using adaptive lasso and second stage no model selection TSLS.

$$X = Z\gamma + \nu$$

where  $X$  is the  $n \times 4$  matrix of endogenous variables, in our design,  $\beta = (2, 0, 0, 0)$ . So we have a sparse structural equation model.  $Z$  is the  $n \times 5$  matrix of instruments. We assume all five are strong instruments.

$$\gamma = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$\epsilon$  and  $\nu$  have mean 0,  $\sigma_\epsilon^2 = 1$ ,  $\sigma_\nu^2 = I_4$ .  $\text{corr}(\epsilon_i, \nu_{ij}) = \rho$ ,  $i = 1, \dots, n$ ,  $j = 1, 2, 3, 4$ . In Design 1,  $\rho = .5$ . In Design 2,  $\rho = .1$

In the second stage, we use adaptive lasso TSLS estimator presented in Section 2 and bridge

in Caner (2009) to estimate the model. We could use Andrews and Lu (2001)'s BIC to select the structural model, but the procedure is not very easy to implement and it is left for further studies.

We report the median bias and MSE of structural equation coefficients of the two model designs in Table B.5 and Table B.6, for  $n = 100$ ,  $n = 300$  respectively. Adaptive lasso TSLS estimator is shown to have variable selection consistency as well as estimation consistency. Bridge estimator is also consistent. Adaptive lasso has smaller variance and MSE than bridge.

Table B.5: Summary statistics for Design 1

	$n = 100$				$n = 300$			
	AD Lasso		bridge		AD Lasso		bridge	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\beta_1$	-.048	.006	-.100	.011	-.034	.007	-.101	.010
$\beta_2$	0	0	0	0	0	0	0	0
$\beta_3$	0	0	0	0	0	0	0	0
$\beta_4$	0	0	0	0	0	0	0	0

Table B.6: Summary statistics for Design 2

	$n = 100$				$n = 300$			
	AD Lasso		bridge		AD Lasso		bridge	
	Bias	MSE	Bias	MSE	Bias	MSE	Bias	MSE
$\beta_1$	-.049	.006	-.107	.012	-.034	.004	-.103	.011
$\beta_2$	0	0	0	0	0	0	0	0
$\beta_3$	0	0	0	0	0	0	0	0
$\beta_4$	0	0	0	0	0	0	0	0

## Appendix C

# One Step Adaptive Lasso with No Structural Equation Variable Selection

In previous chapters we showed the oracle properties of Adaptive Lasso estimator that can select the correct instruments with probability approaching to 1. In Appendix B, we showed the Adaptive Lasso type TSLS estimator where we assume that first stage selection is correct and the model is identifiable. In this appendix, we show consistency and asymptotic normality of one step adaptive lasso. In first stage, we use adaptive lasso as described in Chapter 2. We use the adaptive lasso predicted value of endogenous variable  $\hat{X}$  in the structural equation. LS is then applied in the estimation.

### C.1 The DGP

The IV model to be estimated is the same as in Chapter 2

$$\begin{aligned}y &= X\beta^* + \epsilon \\X &= Z\gamma^* + \nu\end{aligned}$$

where  $y$  is  $n \times 1$  dependent variable,  $X$  is  $n \times p$  matrix of endogenous variables,  $Z$  is  $n \times q$  matrix of instruments,  $\epsilon$  and  $\nu$  are  $n \times 1$ ,  $n \times p$  random disturbances respectively,  $\beta^*$  and  $\gamma^*$  are  $p \times 1$ ,  $q \times p$  coefficients respectively. Again, we assume there is no included exogenous variables. We assume the need of variable selection in the first stage (true model has zero coefficients) equation but not in the second stage. Without loss of generality, we assume the first  $q_0$  row of  $\gamma^*$  is nonzero, and also  $q_0 \geq p$ . Define  $Z_0$ ,  $Z_1$  as the conformable block matrices of partitioned

matrix  $Z = [Z_0 \ Z_1]$  respectively corresponding to the nonzero and zero coefficient matrix  $\gamma^*$ .

The LS estimator is

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$$

where the predicted value of regressors,  $\hat{X} = Z\hat{\gamma}$ , and  $\hat{\gamma}$  is adaptive lasso estimator in the first stage as defined in Chapter 2.

## C.2 Assumptions

We then present the assumptions that are useful in providing the oracle properties for one step adaptive lasso type TSLS estimator.

**Assumption 1** :  $\epsilon_i$  is i.i.d. with mean 0 and variance  $\sigma_\epsilon^2$ .  $\nu_i$  i.i.d. with mean 0 and variance  $\sigma_\nu^2$ . And  $Cov(\epsilon_i, \nu_{ik}) = \sigma_{\nu\epsilon} \neq 0$ ,  $i = 1, 2, \dots, n$ ;  $k = 1, 2, \dots, p$ .  $Cov(\nu_i, \epsilon_{i'k}) = 0$ ,  $i, i' = 1, 2, \dots, n$ ,  $i \neq i'$ .

**Assumption 2** :  $\frac{Z'Z}{n} \rightarrow \Sigma_{ZZ}$ .  $D_n = (\frac{X'Z_0}{n})(\frac{Z_0'Z_0}{n})^{-1}(\frac{Z_0'X}{n}) \rightarrow \Sigma'_{Z_0X}\Sigma_{Z_0Z_0}^{-1}\Sigma_{Z_0X} = D$ , where  $\frac{Z_0'Z_0}{n} \rightarrow_p \Sigma_{Z_0Z_0}$ ,  $\frac{Z_0'X}{n} \rightarrow_p \Sigma_{Z_0X}$  and  $D$  is a positive definite matrix and finite.

**Assumption 3** :  $E(\epsilon | Z) = 0$ ,  $E(\nu | Z) = 0$ .

**Assumption 4** :  $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$  and  $\lambda_n n^{\frac{\tau-1}{2}} \rightarrow \infty$

## C.3 The Limit Theory for One Step Adaptive Lasso Estimator

In this section we present the asymptotic properties of adaptive lasso estimator. Adaptive lasso has ideal asymptotic properties of estimation and selection consistency. Adaptive lasso can choose the correct model as if it were known, which is also known as the oracle properties.

**Theorem 5** (The consistency and asymptotic normality of one step adaptive lasso IV estimator)

Under Assumption 1-4,

1. The consistency of one step adaptive lasso IV estimator

$$\hat{\beta} \rightarrow_p \beta^*$$

2. The asymptotic normality of one step adaptive lasso IV estimator

$$\sqrt{n}(\hat{\beta} - \beta^*) \rightarrow_d N(0, \sigma_\epsilon^2 D^{-1})$$

## C.4 Proof of Theorems

Consistency of  $\hat{\beta}$ :

$$\begin{aligned}
\hat{\beta} &= (\hat{X}'\hat{X})^{-1}\hat{X}'\mathbf{y} \\
&= (\hat{\gamma}'Z'Z\hat{\gamma})^{-1}(\hat{\gamma}'Z'\mathbf{y}) \\
&= \left(\frac{\hat{\gamma}'Z'Z\hat{\gamma}}{n}\right)^{-1}\left(\frac{\hat{\gamma}'Z'X}{n}\beta^* + \frac{\hat{\gamma}'Z'\epsilon}{n}\right) \\
&= \left(\frac{\hat{\gamma}'Z'Z\hat{\gamma}}{n}\right)^{-1}\left(\frac{\hat{\gamma}'Z'X}{n}\right)\beta^* + \left(\frac{\hat{\gamma}'Z'Z\hat{\gamma}}{n}\right)^{-1}\left(\frac{\hat{\gamma}'Z'\epsilon}{n}\right)
\end{aligned} \tag{C.1}$$

Then via

$$\hat{\gamma} \rightarrow_p \gamma^*$$

And by Assumption 2,  $\frac{Z'Z}{n} \rightarrow \Sigma_{ZZ}$

$$\hat{\gamma}'\left(\frac{Z'Z}{n}\right)\hat{\gamma} \rightarrow_p \gamma^{*\prime}\Sigma_{ZZ}\gamma^* \tag{C.2}$$

$$\hat{\gamma}'\left(\frac{Z'\epsilon}{n}\right) \rightarrow_p 0 \tag{C.3}$$

Since  $E(Z_i\epsilon_i) = 0$ .

$$\hat{\gamma}'\left(\frac{Z'X}{n}\right) = \hat{\gamma}'\left(\frac{Z'Z}{n}\right)\gamma^* + \hat{\gamma}'\left(\frac{Z'\nu}{n}\right)$$

Again, by  $\hat{\gamma} \rightarrow_p \gamma^*$ ,

$$\hat{\gamma}'\left(\frac{Z'Z}{n}\right)\gamma^* \rightarrow_p \gamma^{*\prime}\Sigma_{ZZ}\gamma^*$$

And

$$\hat{\gamma}'\left(\frac{Z'\nu}{n}\right) \rightarrow_p 0$$

Since  $E(Z_i\nu_i) = 0$ .

So

$$\hat{\gamma}'\left(\frac{Z'X}{n}\right) \rightarrow_p \gamma^{*\prime}\Sigma_{ZZ}\gamma^* \tag{C.4}$$

So from (C.2) and (C.4) we have

$$\left(\frac{\hat{\gamma}'Z'Z\hat{\gamma}}{n}\right)^{-1}\left(\frac{\hat{\gamma}'Z'X}{n}\right)\beta^* - \beta^* \rightarrow_p 0 \tag{C.5}$$

Through (C.2) and (C.3) we have

$$\left(\frac{\hat{\gamma}'Z'Z\hat{\gamma}}{n}\right)^{-1}\left(\frac{\hat{\gamma}'Z'\epsilon}{n}\right) \rightarrow_p 0 \tag{C.6}$$

So (C.5) and (C.6) provide

$$\hat{\beta} \rightarrow_p \beta^* \quad (\text{C.7})$$

Proof of Asymptotic normality:

Assume only one endogenous variable in  $X$  to simplify the notation without loss of generality.

So  $p = 1$ .

$$\hat{\beta} = (\hat{X}'\hat{X})^{-1}(\hat{X}'y)$$

$$\sqrt{n}\hat{\beta} = \left(\frac{\hat{X}'\hat{X}}{n}\right)^{-1}\left(\frac{\hat{X}'y}{\sqrt{n}}\right) \quad (\text{C.8})$$

$$\sqrt{n}\beta^* = \left(\frac{\hat{X}'\hat{X}}{n}\right)^{-1}\left(\frac{\hat{X}'\hat{X}\beta^*}{\sqrt{n}}\right) \quad (\text{C.9})$$

Next subtract (C.9) from (C.8)

$$\sqrt{n}(\hat{\beta} - \beta^*) = \left(\frac{\hat{X}'\hat{X}}{n}\right)^{-1}\left(\frac{\hat{X}'y - \hat{X}'\hat{X}\beta^*}{\sqrt{n}}\right) \quad (\text{C.10})$$

Analyze RHS of (C.10).

By Theorem 1 of Chapter 2,  $\hat{\gamma}$  enjoys oracle property. Namely,  $\hat{\gamma} \rightarrow_p \hat{\gamma}_{\text{OLS}}$  if  $\gamma_j^* \neq 0$ ; and  $\hat{\gamma} \rightarrow_p 0$  if  $\gamma_j^* = 0$ ,  $j = 1, 2, \dots, q$ .

$$\hat{\gamma} = \begin{bmatrix} \hat{\gamma}_{\mathcal{A}_n} \\ \hat{\gamma}_{\mathcal{A}_n^c} \end{bmatrix} = \begin{bmatrix} \left(\frac{Z_0'Z_0}{n}\right)^{-1}\left(\frac{Z_0'X}{n}\right) \\ \mathbf{0} \end{bmatrix} + o_p(1) \quad (\text{C.11})$$

So, via

$$\hat{\gamma} = \begin{bmatrix} \left(\frac{Z_0'Z_0}{n}\right)^{-1}\left(\frac{Z_0'X}{n}\right) \\ \mathbf{0} \end{bmatrix} + o_p(1) \rightarrow_p \begin{bmatrix} \Sigma_{z_0z_0}^{-1}\Sigma_{z_0x} \\ \mathbf{0} \end{bmatrix} \quad (\text{C.12})$$

and

$$\frac{Z'Z}{n} = \begin{bmatrix} \frac{Z_0'Z_0}{n} & \frac{Z_0'Z_1}{n} \\ \frac{Z_1'Z_0}{n} & \frac{Z_1'Z_1}{n} \end{bmatrix} \rightarrow \begin{bmatrix} \Sigma_{z_0z_0} & \Sigma_{z_0z_1} \\ \Sigma_{z_1z_0} & \Sigma_{z_1z_1} \end{bmatrix} \quad (\text{C.13})$$

$$\hat{\gamma}'\left(\frac{Z'Z}{n}\right)\hat{\gamma} \rightarrow_p \begin{bmatrix} \Sigma_{z_0x}'\Sigma_{z_0z_0}^{-1} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \Sigma_{z_0z_0} & \Sigma_{z_0z_1} \\ \Sigma_{z_1z_0} & \Sigma_{z_1z_1} \end{bmatrix} \begin{bmatrix} \Sigma_{z_0z_0}^{-1}\Sigma_{z_0x} \\ \mathbf{0} \end{bmatrix} = \Sigma_{z_0x}'\Sigma_{z_0z_0}^{-1}\Sigma_{z_0x} \quad (\text{C.14})$$

Therefore

$$\left(\frac{\hat{X}'\hat{X}}{n}\right)^{-1} = \left(\hat{\gamma}'\left(\frac{Z'Z}{n}\right)\hat{\gamma}\right)^{-1} \rightarrow_p \left(\Sigma_{z_0x}'\Sigma_{z_0z_0}^{-1}\Sigma_{z_0x}\right)^{-1} \quad (\text{C.15})$$

Next, consider the rest of the RHS term of (C.10).

$$\begin{aligned}
& \frac{\hat{X}'y - \hat{X}'\hat{X}\beta^*}{\sqrt{n}} \\
&= \frac{\hat{X}'(X\beta^* + \epsilon) - \hat{X}'\hat{X}\beta^*}{\sqrt{n}} \\
&= \frac{\hat{X}'\epsilon}{\sqrt{n}} + \frac{(\hat{X}'X - \hat{X}'\hat{X})\beta^*}{\sqrt{n}}
\end{aligned} \tag{C.16}$$

The first term of (C.16). Define  $\frac{Z'\epsilon}{\sqrt{n}} = \begin{bmatrix} \frac{Z'_0\epsilon}{\sqrt{n}} \\ \frac{Z'_1\epsilon}{\sqrt{n}} \end{bmatrix}$  and use (C.11), by Slutsky's theorem

$$\frac{\hat{X}'\epsilon}{\sqrt{n}} = \hat{\gamma}'\left(\frac{Z'\epsilon}{\sqrt{n}}\right) \rightarrow_d N(0, \sigma_\epsilon^2 \Sigma'_{z_0x} \Sigma_{z_0z_0}^{-1} \Sigma_{z_0x}) \tag{C.17}$$

Second term of (C.16),

$$\frac{(\hat{X}'X - \hat{X}'\hat{X})}{\sqrt{n}} = \frac{\hat{\gamma}'Z'X - \hat{\gamma}'Z'Z\hat{\gamma}}{\sqrt{n}} \tag{C.18}$$

Then analyze the numerator in (C.18),

$$\hat{\gamma}'Z'X = \hat{\gamma}'Z'Z\gamma^* + \hat{\gamma}'Z'\nu \tag{C.19}$$

So in (C.18), using (C.19)

$$\begin{aligned}
& \frac{\hat{\gamma}'Z'X - \hat{\gamma}'Z'Z\hat{\gamma}}{\sqrt{n}} \\
&= \frac{\hat{\gamma}'Z'Z\gamma^*}{\sqrt{n}} + \frac{\hat{\gamma}'Z'\nu}{\sqrt{n}} - \frac{\hat{\gamma}'Z'Z\hat{\gamma}}{\sqrt{n}} \\
&= \hat{\gamma}'\left(\frac{Z'Z}{n}\right)\sqrt{n}(\gamma^* - \hat{\gamma}) + \hat{\gamma}'\frac{Z'\nu}{\sqrt{n}}
\end{aligned} \tag{C.20}$$

For the second term of (C.20), use (C.11)

$$\hat{\gamma} = \begin{bmatrix} \left(\frac{Z'_0Z_0}{n}\right)^{-1}\left(\frac{Z'_0X}{n}\right) \\ \mathbf{0} \end{bmatrix} + o_p(1)$$

$$\begin{aligned}
\hat{\gamma}' \frac{Z' \nu}{\sqrt{n}} &= \left[ \left( \frac{Z'_0 X}{n} \right)' \left( \frac{Z'_0 Z_0}{n} \right)^{-1} \mathbf{0} \right] \begin{bmatrix} \frac{Z'_0 \nu}{\sqrt{n}} \\ \frac{Z'_1 \nu}{\sqrt{n}} \end{bmatrix} + o_p(1) \\
&= \left( \frac{Z'_0 X}{n} \right)' \left( \frac{Z'_0 Z_0}{n} \right)^{-1} \left( \frac{Z'_0 \nu}{\sqrt{n}} \right) + o_p(1)
\end{aligned} \tag{C.21}$$

Now the first term of (C.20),

$$\sqrt{n}(\hat{\gamma} - \gamma^*) = \begin{bmatrix} \left( \frac{Z'_0 Z_0}{n} \right)^{-1} \left( \frac{Z'_0 \nu}{\sqrt{n}} \right) \\ \mathbf{0} \end{bmatrix} + o_p(1)$$

Through (A.8) and (A.9) in Appendix A.

So

$$\sqrt{n}(\gamma^* - \hat{\gamma}) = \begin{bmatrix} -\left( \frac{Z'_0 Z_0}{n} \right)^{-1} \left( \frac{Z'_0 \nu}{\sqrt{n}} \right) \\ \mathbf{0} \end{bmatrix} + o_p(1) \tag{C.22}$$

Apply (C.22), (C.11) and (C.13) to the first term of (C.20),

$$\begin{aligned}
&\hat{\gamma}' \left( \frac{Z' Z}{n} \right) \sqrt{n}(\gamma^* - \hat{\gamma}) \\
&= \left\{ \left[ \left( \frac{Z'_0 X}{n} \right)' \left( \frac{Z'_0 Z_0}{n} \right)^{-1} \mathbf{0} \right] + o_p(1) \right\} \begin{bmatrix} \frac{Z'_0 Z_0}{n} & \frac{Z'_0 Z_1}{n} \\ \frac{Z'_1 Z_0}{n} & \frac{Z'_1 Z_1}{n} \end{bmatrix} \left\{ \begin{bmatrix} -\left( \frac{Z'_0 Z_0}{n} \right)^{-1} \left( \frac{Z'_0 \nu}{\sqrt{n}} \right) \\ \mathbf{0} \end{bmatrix} + o_p(1) \right\} \\
&= -\left( \frac{Z'_0 X}{n} \right)' \left( \frac{Z'_0 Z_0}{n} \right)^{-1} \left( \frac{Z'_0 \nu}{\sqrt{n}} \right) + o_p(1)
\end{aligned} \tag{C.23}$$

Therefore, we come back to (C.18), which is also (C.20), using (C.21) and (C.23)

$$\frac{\hat{\gamma}' Z' X - \hat{\gamma}' Z' Z \hat{\gamma}}{\sqrt{n}} = \hat{\gamma}' \left( \frac{Z' Z}{n} \right) \sqrt{n}(\gamma^* - \hat{\gamma}) + \hat{\gamma}' \frac{Z' \nu}{\sqrt{n}} \rightarrow_p 0 \tag{C.24}$$

Through (C.15) and (C.17) we have

$$\sqrt{n}(\hat{\beta} - \beta^*) \rightarrow_d N\left(0, \sigma_\epsilon^2 (\Sigma'_{z_0 x} \Sigma_{z_0 z_0}^{-1} \Sigma_{z_0 x})^{-1}\right)$$

which is asymptotic efficiency under conditional homoskedasticity and i.i.d. errors.

In heteroskedasticity case the limiting theory is different.

## C.5 Simulation Results

In this section we compare the results of One Step Adaptive Lasso Estimator with that of Post-Lasso (Belloni et al., 2010) and our own Adaptive Lasso TSLS estimator from Appendix

B without structural equation selection. Our DGP is as follows:

The structural and reduced form equations are:

$$y = X\beta + \epsilon$$

$$X = Z\gamma + \nu$$

where  $X$  is the  $n \times 4$  matrix of endogenous variables, in our design,  $\beta = (2, 0, 0, 0)$ . So we have a sparse structural equation model.  $Z$  is the  $n \times 5$  matrix of instruments. We assume there are four strong instruments and one irrelevant instrument.

$$\gamma = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$\epsilon$  and  $\nu$  have mean 0,  $\sigma_\epsilon^2 = 1$ ,  $\sigma_\nu^2 = I_4$ .  $\text{corr}(\epsilon_i, \nu_{ij}) = \rho$ ,  $i = 1, \dots, n$ ,  $j = 1, 2, 3, 4$ . In Design 1,  $\rho = .5$ . In Design 2,  $\rho = .1$

In the first stage, we use three methods (adaptive lasso, post-lasso and model averaging) to select (estimate) the instruments respectively. In the second stage, we use adaptive lasso TSLS estimator presented in Appendix B to estimate the model using the selected instruments. For post-lasso and model averaging, we use regular TSLS. We could use Andrews and Lu (2001)'s BIC to select the structural model for the post-lasso and model averaging methods, but the procedure is not very easy to implement and it is left for further studies.

We report the bias and MSE of structural equation coefficients in Table C.1. Adaptive lasso TSLS estimator is shown to have variable selection consistency as well as estimation consistency. Post-lasso and model averaging do not select the correct model in the second stage.  $n=100$ .

Table C.1: Summary statistics for Model 7

	Post Lasso		ADTS		OSAD	
	Bias	MSE	Bias	MSE	Bias	MSE
$n = 60, \sigma = 1.5, \rho = .5$	.017	.043	.008	.046	.042	.139
$n = 60, \sigma = 1.5, \rho = .9$	.023	.043	.022	.046	.057	.170
$n = 120, \sigma = 1.5, \rho = .5$	-.009	.021	-.006	.021	.020	.064
$n = 120, \sigma = 1.5, \rho = .9$	.014	.021	.017	.021	.041	.080
$n = 300, \sigma = 3, \rho = .5$	-.001	.033	.001	.033	.029	.102
$n = 300, \sigma = 3, \rho = .9$	.011	.038	.016	.037	.050	.138
$n = 1000, \sigma = 3, \rho = .5$	.012	.009	.012	.009	.029	.029
$n = 1000, \sigma = 3, \rho = .9$	-.006	.010	-.006	.010	.013	.037

Post Lasso is the model which TSLS uses Post-Lasso estimated (first stage) instruments., ADTS (Adaptive Lasso) is the model which TSLS uses adaptive lasso selected (first stage) instruments and in the second stage TSLS adaptive lasso method is used. OSAD is the one step adaptive lasso.

## Appendix D

# One Step Adaptive Lasso with Structural Equation Variable Selection

In previous section we showed the consistency and asymptotic normality of one step adaptive lasso (first stage adaptive lasso, second stage OLS). We use the adaptive lasso predicted value of endogenous variable  $\hat{X}$  in the structural equation. In this section we show the oracle property of adaptive lasso in both stages and with predicted endogenous variables  $\hat{X}$  as the regressor. The IV model to be estimated is the same as presented in Appendix C

$$\begin{aligned}y &= X\beta^* + \epsilon \\X &= Z\gamma^* + \nu\end{aligned}$$

where the variables  $y, X, Z, \epsilon, \nu$  are defined in Appendix C. We assume the need of variable selection in both stage (sparsity in coefficients  $\gamma^*$  and  $\beta^*$ ). Again, we assume there is no included exogenous variables. Let  $\mathcal{A} = \text{support}(\beta^*) = \{j \in \{1, 2, \dots, p\} : |\beta_j^*| \neq 0\}$ . The dimension of  $\mathcal{A}$ ,  $\|\mathcal{A}\|_0 = \sum_{j=1}^p 1\{\beta_j^*\} < p$ . Use the Assumption 1 to 4 in Appendix C.

We use the predicted  $\hat{X} = Z\hat{\gamma}$  in the structural equation, where  $\hat{\gamma}$  is the adaptive lasso estimator in the first stage. The adaptive lasso estimator with  $\hat{X}$  as the regressor is

$$\hat{\beta}_n = \arg \min_{\beta} \left\{ (y - \hat{X}\beta)'(y - \hat{X}\beta) + \lambda_n \sum_{k=1}^p \hat{w}_k |\beta_k| \right\}$$

where the tuning parameter  $\lambda_n$ , adaptive weight  $\hat{w}_k$  and  $\mathcal{A}_n$  are defined in Appendix B. In Zou (2006), estimation consistency of adaptive lasso is not explicitly shown in the paper. Now we show the consistency of  $\hat{\beta}_n$ :

Define the function

$$L_n(\beta) = \frac{1}{n}(y - \hat{X}\beta)'(y - \hat{X}\beta) + \frac{\lambda_n}{n} \sum_{k=1}^p \hat{w}_k |\beta_k| \quad (\text{D.1})$$

Transform  $y$  in the following way

$$\begin{aligned} y &= X\beta^* + \epsilon + \hat{X}\beta^* - \hat{X}\beta^* \\ &= \hat{X}\beta^* + \epsilon + (X - \hat{X})\beta^* \\ &= \hat{X}\beta^* + \epsilon + (Z\gamma^* + \nu - Z\hat{\gamma})\beta^* \\ &= \hat{X}\beta^* + \epsilon + [Z(\gamma^* - \hat{\gamma}) + \nu]\beta^* \\ &= \hat{X}\beta^* + \hat{\epsilon} \end{aligned} \quad (\text{D.2})$$

where  $\hat{\epsilon} \equiv \epsilon + [Z(\gamma^* - \hat{\gamma}) + \nu]\beta^*$

Use (D.2) in the non-penalized part in (D.1),

$$\begin{aligned} \frac{1}{n}(y - \hat{X}\beta)'(y - \hat{X}\beta) &= \frac{1}{n}[\hat{\epsilon} - \hat{X}(\beta - \beta^*)]'[\hat{\epsilon} - \hat{X}(\beta - \beta^*)] \\ &= \frac{\hat{\epsilon}'\hat{\epsilon}}{n} - \frac{2\hat{\epsilon}'\hat{X}(\beta - \beta^*)}{n} + \frac{(\beta - \beta^*)'\hat{X}'\hat{X}(\beta - \beta^*)}{n} \end{aligned} \quad (\text{D.3})$$

Now we look at each components of (D.3), first

$$\begin{aligned} \frac{\hat{\epsilon}'\hat{\epsilon}}{n} &= \frac{[\epsilon + Z(\gamma^* - \hat{\gamma})\beta^* + \nu\beta^*]'[\epsilon + Z(\gamma^* - \hat{\gamma})\beta^* + \nu\beta^*]}{n} \\ &= \frac{\epsilon'\epsilon}{n} + 2\frac{\epsilon'Z}{n}(\gamma^* - \hat{\gamma})\beta^* + 2\frac{\epsilon'\nu}{n}\beta^* + \beta^{*'}(\gamma^* - \hat{\gamma})'\left(\frac{Z'Z}{n}\right)(\gamma^* - \hat{\gamma})\beta^* + 2\beta^{*'}\frac{\nu'Z}{n}(\gamma^* - \hat{\gamma})\beta^* + \beta^{*'}\frac{\nu'\nu}{n}\beta^* \\ &\rightarrow_p \sigma_\epsilon^2 + 2\sigma_{\epsilon\nu}\mathbf{1}'\beta^* + \beta^{*'}\beta^*\sigma_\nu^2 \end{aligned} \quad (\text{D.4})$$

where  $\mathbf{1}$  is a conformable vector of 1's.

Since  $\frac{\epsilon'Z}{n} \rightarrow_p 0$ ,  $\frac{\nu'Z}{n} \rightarrow_p 0$ ,  $\hat{\gamma} - \gamma^* \rightarrow_p 0$  and

$$(\gamma^* - \hat{\gamma})'\left(\frac{Z'Z}{n}\right)(\gamma^* - \hat{\gamma}) \rightarrow_p (\gamma^* - \gamma^*)'\Sigma_{ZZ}(\gamma^* - \gamma^*) = 0$$

Since  $\hat{\gamma} - \gamma^* \rightarrow_p 0$ .

Next, in (D.3)

$$\begin{aligned} \frac{2\epsilon' \hat{X}(\beta - \beta^*)}{n} &= 2 \frac{\{\epsilon + [Z(\gamma^* - \hat{\gamma}) + \nu]\beta^*\}' Z \hat{\gamma}}{n} (\beta - \beta^*) \\ &= 2 \frac{\epsilon' Z}{n} \hat{\gamma}(\beta - \beta^*) + 2\beta^{*'} (\gamma^* - \hat{\gamma})' \frac{Z' Z}{n} \hat{\gamma}(\beta - \beta^*) + 2\beta^{*'} \frac{\nu' Z}{n} \hat{\gamma}(\beta - \beta^*) \\ &\rightarrow_p 0 \end{aligned} \quad (\text{D.5})$$

Since  $\frac{\epsilon' Z}{n} \rightarrow_p 0$ ,  $\hat{\gamma} - \gamma^* \rightarrow_p 0$ ,  $\frac{Z' Z}{n} \rightarrow_p \Sigma_{ZZ}$  and  $\frac{\nu' Z}{n} \rightarrow_p 0$ .

Next, in (D.3)

$$\frac{(\beta - \beta^*)' \hat{X}' \hat{X}(\beta - \beta^*)}{n} = (\beta - \beta^*)' \hat{\gamma}' \left( \frac{Z' Z}{n} \right) \hat{\gamma}(\beta - \beta^*) \rightarrow_p (\beta - \beta^*)' \gamma^{*'} \Sigma_{ZZ} \gamma^*(\beta - \beta^*) \quad (\text{D.6})$$

Next combine (D.4) - (D.6) in (D.3) we have

$$\frac{1}{n} (y - \hat{X}\beta)' (y - \hat{X}\beta) \rightarrow_p \sigma_\epsilon^2 + 2\sigma_{\epsilon\nu} \mathbf{1}' \beta^* + \beta^{*'} \beta^* \sigma_\nu^2 + (\beta - \beta^*)' \gamma^{*'} \Sigma_{ZZ} \gamma^*(\beta - \beta^*) \quad (\text{D.7})$$

Then the penalty term in (D.1),

$\frac{\lambda_n \hat{w}_k}{n} \rightarrow_p 0$  is shown in Appendix B, so the penalty term in (D.1) converge to 0.

Therefore from (D.1), (D.7) and since  $\frac{\lambda_n \hat{w}_k}{n} \rightarrow_p 0$ ,  $\hat{\beta}_n$  is consistent if  $\hat{\beta}_n \rightarrow_p \arg \min(L)$  where

$$L(\beta) = [(\beta - \beta^*)' \gamma^{*'}] \Sigma_{ZZ} [\gamma^*(\beta - \beta^*)] + \sigma_\epsilon^2 + 2\sigma_{\epsilon\nu} \mathbf{1}' \beta^* + \beta^{*'} \beta^* \sigma_\nu^2$$

We need to show that

$$\sup_{\beta \in K} |L_n(\beta) - L(\beta)| \rightarrow_p 0 \quad (\text{D.8})$$

for any compact set K and that

$$\hat{\beta}_n = O_p(1) \quad (\text{D.9})$$

$L_n(\beta)$  is convex, thus (D.8) and (D.9) follow from the point-wise convergence in probability of  $L_n(\beta)$  to  $L(\beta)$  by applying standard results of [Anderson and Gill (1982); Pollard (1991)].

Then we show asymptotic normality.

$\hat{\beta}_n$  is consistent, let

$$\beta_k = \beta_k^* + \frac{u_k}{\sqrt{n}}, \quad k = 1, 2, \dots, p$$

And define

$$\Psi_n(u) = \left[ y - \hat{X} \left( \beta^* + \frac{u}{\sqrt{n}} \right) \right]' \left[ y - \hat{X} \left( \beta^* + \frac{u}{\sqrt{n}} \right) \right] + \lambda_n \sum_{k=1}^p \hat{w}_k \left| \beta_k^* + \frac{u_k}{\sqrt{n}} \right| \quad (\text{D.10})$$

$$\Psi_n(0) = (y - \hat{X}\beta^*)'(y - \hat{X}\beta^*) + \lambda_n \sum_{k=1}^p \hat{w}_k |\beta_k^*| \quad (\text{D.11})$$

Notice we now minimize the objective function with respect to  $u$  instead of  $\beta$ .

Let  $\hat{u}_n = \arg \min_u \Psi_n(u)$  or  $\hat{u}_n = \sqrt{n}(\hat{\beta}_n - \beta^*)$ .

Let

$$\begin{aligned} V_n(u) &= \Psi_n(u) - \Psi_n(0) \\ &= \left[ y - \hat{X} \left( \beta^* + \frac{u}{\sqrt{n}} \right) \right]' \left[ y - \hat{X} \left( \beta^* + \frac{u}{\sqrt{n}} \right) \right] + \lambda_n \sum_{k=1}^p \hat{w}_k \left| \beta_k^* + \frac{u_k}{\sqrt{n}} \right| \\ &\quad - (y - \hat{X}\beta^*)'(y - \hat{X}\beta^*) - \lambda_n \sum_{k=1}^p \hat{w}_k |\beta_k^*| \\ &= -2 \frac{(y - \hat{X}\beta^*)' \hat{X} u}{\sqrt{n}} + \frac{u' \hat{X}' \hat{X} u}{n} + \lambda_n \sum_{k=1}^p \hat{w}_k \left( \left| \beta_k^* + \frac{u_k}{\sqrt{n}} \right| - |\beta_k^*| \right) \\ &= -2 \frac{\hat{\epsilon}' \hat{X}}{\sqrt{n}} u + u' \left( \frac{\hat{X}' \hat{X}}{n} \right) u + \lambda_n \sum_{k=1}^p \hat{w}_k \left( \left| \beta_k^* + \frac{u_k}{\sqrt{n}} \right| - |\beta_k^*| \right) \end{aligned} \quad (\text{D.12})$$

The solution to (D.10) is equivalent to that of  $V_n(u)$ .

Since  $y - \hat{X}\beta^* = \hat{\epsilon} = \epsilon + [Z(\gamma^* - \hat{\gamma}) + \nu]\beta^*$ .

First consider the non-penalized terms in (D.12).

By (C.14)

$$\frac{\hat{X}' \hat{X}}{n} = \hat{\gamma}' \frac{Z' Z}{n} \hat{\gamma} \rightarrow_p \Sigma'_{z_0 x} \Sigma_{z_0 z_0}^{-1} \Sigma_{z_0 x} \equiv D \quad (\text{D.13})$$

Using  $\hat{X} = Z\hat{\gamma}$  and since  $\frac{Z' Z}{n} \rightarrow \begin{bmatrix} \Sigma_{z_0 z_0} & \Sigma_{z_0 z_1} \\ \Sigma_{z_1 z_0} & \Sigma_{z_1 z_1} \end{bmatrix}$  and  $\hat{\gamma} \rightarrow_p \begin{bmatrix} \Sigma_{z_0 z_0}^{-1} \Sigma_{z_0 x} \\ 0 \end{bmatrix}$  by (C.13) and (C.12).

Let  $D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}$ , where  $D_{11}$  is  $p_0 \times p_0$  matrix,  $D_{22}$  is  $(p - p_0) \times (p - p_0)$  matrix and  $D_{12} = D'_{21}$ .

Next in (D.12)

$$\begin{aligned} \frac{\hat{\epsilon}' \hat{X}}{\sqrt{n}} &= \frac{[\epsilon' + \beta^{*'}(\gamma^* - \hat{\gamma})' Z' + \beta^{*'} \nu'] Z \hat{\gamma}}{\sqrt{n}} \\ &= \frac{\epsilon' Z \hat{\gamma}}{\sqrt{n}} + \beta^{*'} \left[ \sqrt{n}(\gamma^* - \hat{\gamma})' \left( \frac{Z' Z}{n} \right) \hat{\gamma} + \frac{\nu' Z \hat{\gamma}}{\sqrt{n}} \right] \end{aligned} \quad (\text{D.14})$$

Analyze each term in (D.14). By (C.17),

$$\frac{\epsilon' Z \hat{\gamma}}{\sqrt{n}} \rightarrow_d N(0, \sigma_\epsilon^2 D) \equiv W \quad (\text{D.15})$$

By (C.24),

$$\sqrt{n}(\gamma^* - \hat{\gamma})' \left( \frac{Z'Z}{n} \right) \hat{\gamma} + \frac{\nu' Z \hat{\gamma}}{\sqrt{n}} \rightarrow_p 0$$

Now consider the limiting behavior of the penalty term of (D.12).

There are 2 cases to analyze, one is  $\beta_k^* \neq 0$ , the other is  $\beta_k^* = 0$ .

If  $\beta_k^* \neq 0$ , then  $\hat{w}_k \rightarrow_p \frac{1}{|\beta_k^*|^\tau}$ .

And

$$\sqrt{n}(|\beta_k^* + \frac{u_k}{\sqrt{n}}| - |\beta_k^*|) \rightarrow u_k \operatorname{sgn}(\beta_k^*)$$

By Assumption 2.4,  $\frac{\lambda_n}{\sqrt{n}} \rightarrow 0$ .

By Slutsky's theorem,  $\frac{\lambda_n}{\sqrt{n}} \hat{w}_k \sqrt{n}(|\beta_k^* + \frac{u_k}{\sqrt{n}}| - |\beta_k^*|) \rightarrow_p 0$ .

If  $\beta_k^* = 0$ , then  $\sqrt{n}(|\beta_k^* + \frac{u_k}{\sqrt{n}}| - |\beta_k^*|) = |u_k|$ .

And  $\frac{\lambda_n}{\sqrt{n}} \hat{w}_k = \frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau}{2}} (\sqrt{n} \tilde{\beta}_k)^{-\tau}$ , where  $\sqrt{n} \tilde{\beta}_k = O_p(1)$ .

By Assumption 2.4,  $\frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau}{2}} \rightarrow \infty$ .

Therefore, by Slutsky's theorem,

$$\frac{\lambda_n}{\sqrt{n}} \hat{w}_k \sqrt{n}(|\beta_k^* + \frac{u_k}{\sqrt{n}}| - |\beta_k^*|) \rightarrow_p \infty$$

Now combine above results for (D.12), by Slutsky's theorem,  $V_n(u) \rightarrow_d V(u)$  for every  $u$ , where

$$V(u) = \begin{cases} u'_A D_{11} u_A - 2u'_A W_A & \text{if } u_k = 0 \forall \beta_k^* \notin \mathcal{A}, \\ \infty & \text{otherwise.} \end{cases} \quad (\text{D.16})$$

where  $u_A$  represents the  $u$ 's of nonzero coefficients. And  $W_A$  is the  $p_0 \times 1$  subset (corresponding to the first  $p_0$  elements) of  $W$  vector in (D.15), which is distributed as  $N(0, \sigma_\epsilon^2 D_{11})$ .  $V_n(u)$  is a convex function, and the unique minimum of  $V_n(u)$  is  $(D_{11}^{-1} W_A, \mathbf{0})'$ .

Following the epi-convergence results of Geyer (1994) and Knight and Fu (2000), let  $\hat{u}_A$ ,  $\hat{u}_{A^c}$  represents the nonzero and zero  $\hat{\beta}_k$  in  $\mathcal{A}_n$  respectively, we have:

$$\hat{u}_A \rightarrow_d D_{11}^{-1} W_A \quad (\text{D.17})$$

$$\hat{u}_{A^c} \rightarrow_d \mathbf{0} \quad (\text{D.18})$$

Since  $W_A = N(0, \sigma_\epsilon^2 D_{11})$ ,  $\hat{u}_A \rightarrow_d N(0, \sigma_\epsilon^2 D_{11}^{-1})$ .

Now we show the variable selection consistency:  $\lim_n P(\mathcal{A}_n = \mathcal{A}) = 1$ .

$\forall \beta_k^* \in \mathcal{A}$ , the asymptotic normality indicates that  $\hat{\beta}_k \rightarrow_p \beta_k^*$ , thus  $P(\hat{\beta}_k \in \mathcal{A}_n) \rightarrow 1$ .

Therefore to show variable selection consistency, it is sufficient to show that

$$\forall \beta_{k'}^* \notin \mathcal{A}, P(\hat{\beta}_{k'} \in \mathcal{A}_n) \rightarrow 0$$

We show the impossibility of interior solution of adaptive lasso. Consider the event  $\hat{\beta}_{k'} \in \mathcal{A}_n$ . By the Karush-Kuhn-Tucker optimality conditions, we know that for each  $\hat{\beta}_{k'}$  corresponding to  $\beta_{k'}^* \notin \mathcal{A}$ , we have

$$2\hat{x}'_{k'}(y - \hat{X}\hat{\beta}_n) = \lambda_n \hat{w}_{k'} \quad (\text{D.19})$$

where  $\hat{x}'_{k'}$  is the  $k'^{th}$  column of  $\hat{X}$ ,  $\hat{\beta}_n$  is the adaptive lasso estimator.

By Assumption 3 and  $\tilde{\beta}_k$  being a  $\sqrt{n}$  consistent estimator of  $\beta_k^*$  (from the definition of adaptive weight),  $\sqrt{n}\tilde{\beta}_k = O_p(1)$ , for the RHS of (D.19)

$$\frac{\lambda_n \hat{w}_{k'}}{\sqrt{n}} = \frac{\lambda_n}{\sqrt{n}} n^{\frac{\tau}{2}} \frac{1}{|\sqrt{n}\tilde{\beta}_{k'}|^\tau} \rightarrow_p \infty$$

And now the LHS of (D.19), by (D.2),  $y = \hat{X}\beta^* + \hat{\epsilon}$

$$\frac{2\hat{x}'_{k'}(y - \hat{X}\hat{\beta}_n)}{\sqrt{n}} = \frac{2\hat{x}'_{k'}\hat{X}\sqrt{n}(\beta^* - \hat{\beta}_n)}{n} + 2\frac{\hat{x}'_{k'}\hat{\epsilon}}{\sqrt{n}}$$

By (D.13), (D.17) and (D.18) and Slutsky's theorem, recall that  $\hat{u}_n = \sqrt{n}(\hat{\beta}_n - \beta^*)$

$$\frac{2\hat{x}'_{k'}\hat{X}\sqrt{n}(\beta^* - \hat{\beta}_n)}{n} \rightarrow_d \text{some normal distribution}$$

Then,

$$\begin{aligned} 2\frac{\hat{x}'_{k'}\hat{\epsilon}}{\sqrt{n}} &= 2\frac{\hat{x}'_{k'}\{\epsilon + [Z(\gamma^* - \hat{\gamma}) + \nu]\beta^*\}}{\sqrt{n}} \\ &= 2\frac{\hat{x}'_{k'}\epsilon}{\sqrt{n}} + 2\frac{\hat{x}'_{k'}\nu\beta^*}{\sqrt{n}} + 2\frac{\hat{x}'_{k'}Z(\gamma^* - \hat{\gamma})\beta^*}{\sqrt{n}} \end{aligned} \quad (\text{D.20})$$

Consider each term in (D.20)

By (D.13) and Assumption 2.1,

$$\frac{\hat{x}'_{k'}\epsilon}{\sqrt{n}} \rightarrow_d \text{some normal distribution}$$

Similarly,

$$\frac{\hat{x}'_{k'}\nu\beta^*}{\sqrt{n}} \rightarrow_d \text{some normal distribution}$$

By (D.13) and, (C.22)

$$\frac{\hat{x}'_{k'} Z(\gamma^* - \hat{\gamma})\beta^*}{\sqrt{n}} = \frac{\hat{x}'_{k'} Z}{n} \sqrt{n}(\gamma^* - \hat{\gamma})\beta^* \rightarrow_d \text{some normal distribution}$$

Combine all three terms in (D.20)

$$2 \frac{\hat{x}'_{k'} \hat{\epsilon}}{\sqrt{n}} \rightarrow_d \text{some normal distribution}$$

Thus,  $P(\hat{\beta}_{k'} \in \mathcal{A}_n) \leq P(2\hat{x}'_j(\hat{y} - \hat{X}\hat{\beta}_j) = \lambda_n \hat{w}_j) \rightarrow 0.$  Q.E.D.