# ESSAYS ON INSTRUMENTAL VARIABLES

By

Enrique Pinzón García

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Economics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2012

Date of final oral examination: 04/25/12

The dissertation is approved by the following members of the Final Oral Committee:

Jack Porter, Professor, Economics
Bruce Hansen, Professor, Economics
Andrés Aradillas-López, Assistant Professor, Economics
Xiaoxia Shi, Assistant Professor, Economics
Chunming Zhang, Professor, Statistics

## 0.1. Introduction

The most important assumption made in the ordinary least squares regression model (OLS) is the orthogonality between the random disturbance term and the regressors. This assumption is not fulfilled in many relevant economic scenarios in which a set of explanatory variables, referred to as endogenous regressors, are correlated with the random disturbance. The earliest example, as documented by Stock and Trebbi (2003), arose with the estimation of demand equations and price elasticities. Researchers that explored this problem in the early twentieth century found that the result of regressing quantities demanded on prices was an upward sloping demand curve. The reason for this occurrence is that the equilibrium prices and quantities are affected by movements in the supply curve, something that yields a violation of the assumption that the error term and the regressors are independent. This problem is also present in the regression that relates individual wages to schooling choices and other individual characteristics. In this case there exist components of the random disturbance, for instance unobserved ability, that are correlated with schooling choices.

The instrumental variable procedure addresses this problem by introducing a set of regressors that are uncorrelated to the random disturbance and are related to the explanatory variable of interest only through the endogenous regressors. In the case of the demand equation, a valid instrument would be a variable that shifts the supply affecting the quantity demanded only through its effect on the equilibrium prices. In the case of the wage equation, a valid instrument would be one that affects the years of schooling of an individual but has no direct effect on the wages.

In this dissertation, I provide solutions to problems sometimes encountered by researchers when employing an instrumental variable methodology. I explore the instrumental variable problem in a nonparametric framework and in a situation where there exists a large set of instruments that are weakly correlated with the endogenous variable of interest. The latter case is referred to as the many weak instrument problem and is characterized by the fact that it yields inconsistent estimators with nonstandard asymptotic distributions. I also explore the model selection advantages of the least absolute shrinkage and selection operator (LASSO) as an instrument selection procedure in this context.

In the first chapter of this dissertation, I analyze a many weak instrument setting that extends the Chao and Swanson (2005a) framework to consider a potentially large number of irrelevant instruments. In this setting I propose a new 2SLS estimator that addresses two concerns: first, the selection of the relevant instruments; and second, inconsistent estimates that arise in a 2SLS context with many weak and irrelevant instruments. The methodology put forth addresses the first concern by disregarding with high probability those instruments that should not be included in the model using an adaptive absolute shrinkage and selection operator (LASSO). The second concern comes from the fact that

in the environment described, the traditional 2SLS estimator is not consistent. I prove that the proposed estimator is simultaneously consistent and asymptotically normal in the presence of many weak and irrelevant instruments.

The first stage can also be constructed using a mean independent instrument assumption to provide the possibility of a nonparametric version of the adaptive LASSO. The methodology allows for heteroskedasticity, which has been shown to affect the consistency of the structural parameter in a setting with many weak instruments (Chao et al. (2010); Hausman et al. (2010)). However, using the adaptive LASSO yields first stage estimates with considerable bias. To address this concern and exploit the instrument selection properties of the adaptive LASSO, I run an OLS regression with the selected instruments in the first stage as suggested by Belloni and Chernozhukov (2010).

The second chapter is an empirical application of the findings and method of the first chapter to the data set used in Angrist and Krueger (1991). This paper is the most commonly cited study when referring to the many weak instrument problem and provides a clear illustration of how the estimator introduced in chapter 1 can help select a set of instruments and alleviate the weak instrument problem. Moreover, the results found in this chapter can be readily compared to others presented in the literature. In particular I will contrast my estimation results with those presented in Belloni et al. (2010a).

In the third chapter, I explore a solution to the nonparametric instrumental variable problem within the context of reproducing kernel Hilbert Spaces (RKHs) by recognizing that the object of interest is the solution to a Fredholm integral equation of the first kind. RKHs are characterized by the fact that linear functionals in the space are bounded. Therefore, the results of the previous literature, which assume the function of interest lies in a bounded Hilbert space, can be mapped into a RKH.

Solutions to Fredholm integral equations of the first kind are called regularized solutions. The methodology proposed in this chapter is, as was typified by Nychka et al. (1984), a cross-validated spline solution. Within this framework the solution can be thought of as a penalized least squares estimate. The penalty over the roughness of the function, characteristic of these setups, is controlled by a regularization parameter that is chosen by Generalized Cross Validation (GCV). Except for Gagliardini and Scaillet (2006), the previous papers have no explicit mechanism for choosing the regularization parameter and some, like Newey and Powell (2003), recognize their estimator is very sensitive to the choice of parameters. One advantage of GCV over the methodology of Gagliardini and Scaillet (2006) is that its optimality has been established by Wahba (1977) within the context of integral equations which are the object of interest in the literature of nonparametric endogeneity.

# Contents

# Chapter 1

# Instrumental Variable Estimation and Selection with Many Weak and Irrelevant Instruments

## 1.1. Introduction

In this paper, I extend the many weak instruments framework of Chao and Swanson (2005) to allow for a potentially large number of irrelevant instruments. In a setting with many weak, possibly irrelevant instruments, traditional 2SLS leads to inconsistent estimates when all potential instruments are used, but it is unclear how to select the right instruments to include. I introduce a new 2SLS estimator that selects the correct instruments with high probability and is simultaneously consistent and asymptotically normal in the presence of heteroskedasticity.

Bound et al. (1995) show that even when instruments are plausible the resulting estimator can be inconsistent and have a significant finite sample bias if their correlation with the endogenous variable is small. They make the striking argument that results derived under weak instruments can be equivalent to those obtained with a set of randomly generated instruments. The work of Staiger and Stock (1997) proves that under weak instruments the traditional simultaneous equation estimators have a non-standard limiting distribution. When many instruments are available the conclusions are not as dire. Chao and Swanson (2003b, 2005b, 2003a) prove that when this is the case the traditional simultaneous equation estimators can be consistent and asymptotically normal even if the instrument set is weak. Their work suggests that using all available instruments can alleviate the weak instrument problem.

In the presence of many weak instruments, heteroskedasticity presents an additional challenge. Chao et al. (2010), Hausman et al. (2010), and Chao and Swanson (2003a), prove that simultaneous equations estimators that would be consistent in the presence of many weak instruments are rendered inconsistent. To solve this problem Chao et al. (2010) and Hausman et al. (2010) propose jackknife versions of the traditional simultaneous equations estimators.

Two asymptotic frameworks can be used to analyze the 2SLS estimator I propose in this paper. In both cases the inconsistency of the conventional 2SLS estimator can be characterized by the convergence of the estimator to an expression that depends on the ratio of the concentration parameter to the number of instruments. The concentration parameter is the canonical measure of instrument weakness and can be understood as capturing the signal to noise ratio of the reduced form equation in a linear instrumental variable estimation.

The first asymptotic framework, proposed by Staiger and Stock (1997), studies a sequence of models where the number of instruments is fixed and the concentration parameter remains constant in expectation. Under these conditions, if the concentration parameter is small, 2SLS is inconsistent. In the Chao and Swanson (2003b, 2005b) framework, a sequence of models that allows the concentration parameter and the number of instruments to grow with the sample size is considered. Here, the conventional 2SLS estimator is consistent if the rate at which the number of instruments grows is slower than the growth rate of the concentration parameter. Intuitively, this means that the signal to noise ratio per instrument is growing which implies additional instruments provide information asymptotically.

I contribute to the weak instrument literature by providing an estimator that is consistent and asymptotically normal in the presence of many weak and irrelevant instruments and heteroskedasticity. To do so, I employ the asymptotic framework of Chao and Swanson (2003b, 2005b) because it incorporates the many instruments framework of Bekker (1994) and Morimune (1983) and the fixed number of instruments framework of Staiger and Stock (1997). As was shown by Chao and Swanson (2003b), the selected asymptotic framework also characterizes the conditions under which a wide class of estimators is consistent in the presence of many weak instruments.

In contrast with Chao and Swanson (2003b, 2005b), I consider a framework were some of the instruments are irrelevant. These irrelevant instruments, ideally, should not be introduced. They are included because researchers are uncertain which regressors should be in the model, but the solution to the endogeneity problem demands the inclusion of all possible sources of exogenous variation.[1] Additionally, including many instruments, as was shown by Chao and Swanson (2003b, 2005b, 2003a), can result in the traditional simultaneous equation estimators[2] being consistent and asymptotically normal, even if the instruments set is weak. I argue that when a large instrument set is generated some of the instruments available may be potentially irrelevant. The results I present suggest that having an instrument selection procedure to exclude irrelevant instruments yields a more reliable estimator of the structural parameter of interest.

[1] Traditional solutions to the endogeneity problem assume the endogenous variable is composed of an endogenous and an exogenous component. The solution to the problem comes from subtracting an estimate of the endogenous component from the endogenous variable. Therefore, including all possible sources of exogenous variation is fundamental to solve the endogeneity problem.

[2] Here reference is being made to two stage least squares, 2SLS, limited information maximum likelihood, LIML, and the modified LIML estimator proposed by Fuller (1977).

The instrument selection procedure I employ in this paper to identify the relevant instruments is the adaptive LASSO. The adaptive LASSO selects the set of first stage regressors with non-zero coefficients and excludes the irrelevant ones with probability approaching one. However, the adaptive LASSO comes at the price of post selection bias. The procedure I introduce in this paper runs an OLS regression using the selected instruments, reducing the bias inherent to the adaptive LASSO as is proposed by Belloni and Chernozhukov (2010), Belloni et al. (2010a), and Belloni et al. (2010b).

Introducing an instrument selection procedure is important because it exploits the relationship between the number of instruments and their signal to noise ratio. Chao and Swanson (2003b), define instrument weakness employing this relationship. The definition is directly related to the first stage $F$ statistic, the most common measure of instrument weakness. Specifically, the $F$ statistic is an estimator of the ratio of the signal to noise ratio and the number of instruments. The adaptive LASSO excludes the irrelevant instruments, increasing the signal to noise ratio and reducing the number of instruments. Therefore, the signal to noise ratio per instrument, and hence the associated $F$ statistic, will be higher than if irrelevant instruments are included. This feature is essential for obtaining consistency and asymptotic normality for the 2SLS estimator I propose, in a scenario where many of the instruments are potentially irrelevant. In contrast, the conventional 2SLS that does not use an instrument selection procedure to exclude the irrelevant instruments will never achieve consistency and normality in the presence of instrument weakness.

With regard to the instrument selection literature, my paper can be compared to the work of Belloni et al. (2010a), Belloni et al. (2010b), Caner and Fan (2010), and Kuersteiner and Okui (2010). Belloni et al. (2010a) propose the use of the variables selected by the LASSO to run an OLS regression instead of using the LASSO estimates and prove an improvement in the bias properties of the estimators for this procedure. They refer to this methodology as a post-$l_1$-penalized estimator. Similarly, I use the adaptive LASSO as an instrument selection procedure and not as a first stage estimator directly. Caner and Fan (2010) use the adaptive LASSO to provide instrument selection and use the selected instruments to compute a GMM estimator. They conjecture that the estimator thus constructed will behave like a GMM estimator computed with the true set of instruments. Kuersteiner and Okui (2010) provide a model averaging version of the traditional simultaneous equations estimators. Their methodology is a combination of the instrument selection procedure proposed by Donald and Newey (2001) and the least squares model averaging introduced by Hansen (2007). Finally, Belloni et al. (2010b) extends the work of Belloni et al. (2010a) to include heteroskedasticity.

My work differs from these papers in two important ways: first, the post-selection bias inherent to the adaptive LASSO and the model averaging estimators proposed in

Kuersteiner and Okui (2010) is tackled using a post-$l_1$-penalized estimator; and second, I use a different data generating process and adopt a different asymptotic theory. In the work of Belloni et al. (2010a), Belloni et al. (2010b), and Caner and Fan (2010) the model consists of a set of strong instruments and a set of irrelevant instruments. In both cases the solution of the problem is to select enough strong instruments in the first stage. In contrast, in this paper the data generating process yields instruments that are weak or irrelevant and the solution to the problem is to exclude the irrelevant instruments and select enough of the weak instruments in the first stage. Therefore, the asymptotic behavior of the estimator I propose has a different limiting distribution than the one proposed by Belloni et al. (2010a) and Belloni et al. (2010b) that attain $\sqrt{n}$ consistency. As in Chao and Swanson (2003a) my estimator has a slower convergence rate that is associated to the degree of instrument weakness. Also, the Chao and Swanson (2005b) asymptotic framework allows the analysis of a fixed or a growing number of instruments and provides an insight to the effects of instrument selection using a LASSO type estimate when many of the instruments are weak or irrelevant. These effects can be understood using conventional weak instrument asymptotics.

I also allow for mean independence between the instruments and the structural random disturbance. The idea is for the researcher to be able to use the instruments in a flexible way. The traditional assumption of random disturbances being uncorrelated with the instruments only allows the instruments to enter linearly into the specification. The form of the first stage considered is one in which the conditional mean function consists of a large number of additive components. These additive components are approximated using truncated series expansions with B-splines. The problem is that many of the additive components are zero and some of the terms in the expansion are weakly related to the endogenous variable. The adaptive LASSO will act to identify which coefficients to include in the expansion. The Chao and Swanson (2003b, 2005b) definition of instrument weakness and their necessary and sufficient condition to establish consistency is extended to account for the possibility of terms in the expansion being weakly correlated to the endogenous variable.

The estimator in this study is therefore unique in the sense that it combines into one theoretical framework the model selection advantages of the adaptive LASSO, in its linear and nonparametric version, and the many instruments and adaptive LASSO asymptotics. At the same time it reduces the post-selection bias inherent to the adaptive LASSO using a post-$l_1$-penalized estimator. Moreover, it avoids the problems inherent to the estimators proposed by Chao et al. (2010), Hausman et al. (2010), and Chao and Swanson (2003b) in the presence of weak instruments and heteroskedasticity which do not possess high order moments and thus may lead to unreliable conclusions.[3] With regard to the instrument selection literature, the simulation studies I present below, under different degrees of instrument weakness and endogeneity, show that the estimators

[3] This is the case of LIML, which is amply documented in the literature and, in the simulations presented below, of the jackknife instrumental variable estimator in Chao et al. (2010) and Hausman et al. (2010).

put forth by Kuersteiner and Okui (2010) generate estimates with non-trivial bias. The estimator I propose reduces this bias significantly. Simulation results indicate that my procedure has better performance than the existent estimators for high levels of endogeneity and instrument weakness. However, the simulation results I present do not yield an estimator that dominates all other in terms of mean square error and median absolute deviation. For instance, under homoskedastic designs the Fuller estimator tends to be the most reliable alternative in terms of median absolute deviation. When endogeneity and instrument weakness are low 2SLS is a viable alternative. Finally, the Kuersteiner and Okui (2010) estimators reduce variance significantly, sometimes yielding a smaller mean square error than all other estimators.

The paper is structured as follows. Section 1.2 presents the model and its fundamental assumptions. Section 1.3 establishes the conditions under which the 2SLS estimator attains consistency. Section 1.4 presents the asymptotic normality results and suggests an inference procedure. Section 1.5 compares the performance of the estimator proposed in this study via simulation to those in Chao and Swanson (2003b), Chao et al. (2010), Caner and Fan (2010), and Kuersteiner and Okui (2010). Section 1.6 presents the concluding remarks and suggests possible extensions while all proofs are gathered in Section 1.7.

## 1.2. Framework

The framework is a simultaneous equations model where some of the regressors are irrelevant. The presentation of the model is divided into two subsections. The first subsection discusses instruments that are uncorrelated with the random disturbances, which is the traditional instrumental variable assumption. The second subsection introduces the possibility of having mean independence between the instruments and the random disturbance. The assumptions, regularity conditions, and a definition of instrument weakness are also presented and discussed.

### 1.2.1 Instruments Uncorrelated with the Random Disturbances

The model of interest is given by:

$$y_{1n} = Y_{2n}\beta + u_n \tag{1.1}$$

$$Y_{2n} = Z_n\Pi_n + V_n \tag{1.2}$$

$$\Pi_n = \begin{pmatrix} \Pi_{n1} & \Pi_{n2} \end{pmatrix} = \begin{pmatrix} \Pi_{n1} & \mathbf{0} \end{pmatrix} \tag{1.3}$$

$$E\left(u_i Y_{2ni}\right) \neq \mathbf{0}, \quad E\left(v_i' Z_{ni}\right) = 0 \quad \text{and} \quad E\left(u_i Z_{ni}\right) = \mathbf{0} \tag{1.4}$$

In (1.1) and (1.2) $y_{1n}$ has dimensions $n \times 1$, $Y_{2n}$ is $n \times 1$, and $Z_n$ is $n \times p_n$. $\begin{pmatrix} \Pi_{n1} & \Pi_{n2} \end{pmatrix}$ in (1.3) with $\Pi_{n1} \neq \mathbf{0}$, of dimensions $p_{n1} \times 1$, and $\Pi_{n2} = \mathbf{0}$, of dimensions $p_{n2} \times 1$, denotes the fact that some of the potential instruments have no explanatory power. In (1.4) $Z_{ni}$ is the $i^{th}$ column of $Z_n$ and $Y_{2ni}$, $u_i$ and $v_i$ are the $i^{th}$ elements of $Y_{2n}$, $u_n$ and $V_n$.[4] Without loss of generality all components of $Y_{2n}$ are considered to be endogenous and the data is assumed to be centered so intercepts are not included in the regression function.

The estimator I propose for the model described in (1.1) and (1.2) uses the adaptive LASSO to select a set of instruments, denoted by $\tilde{Z}_{n1}$. The selected instruments are then used in the first stage to run an OLS regression as is suggested by Belloni et al. (2010a) and Belloni et al. (2010b).[5] This yields a first stage estimator given by:

$$\hat{Y}_{2n} = \tilde{Z}_{n1} \left( \tilde{Z}'_{n1} \tilde{Z}_{n1} \right)^{-1} \tilde{Z}'_{n1} Y_{2n}$$

$\hat{Y}_{2n}$ is then used to obtain an estimator of the structural parameter of interest $\beta$ given by:

$$\hat{\beta} = \left( \hat{Y}'_{2n} \hat{Y}_{2n} \right)^{-1} \left( \hat{Y}'_{2n} y_{1n} \right) \tag{1.5}$$

The first stage regression (1.2) starts with an adaptive LASSO which is a modified ver-

[4] In this paper, I will focus on the case of one endogenous variable. I do this for ease of notation. Notice that the analysis could be generalized by allowing $Y_{2n}$ to be of dimensions $n \times K$ for $K > 1$. As is proposed by Caner and Fan (2010) to handle this case $Y_{2n}$ can be vectorized and the analysis follows in similar fashion to the case where $K = 1$.

[5] It is important to emphasize that the adaptive LASSO selects instruments and produces a first stage estimator. In this paper the adaptive LASSO is only used as an instrument selection procedure and the first stage estimates are not utilized.

sion of the LASSO suggested by Tibshirani (1996). The adaptive LASSO was introduced by Zou (2006), and its estimates are defined by:

$$\hat{\Pi}_n = \arg\min_{\Pi_n} \left( \left\| Y_{2n} - \sum_{j=1}^{p} Z_{nj}\Pi_{nj} \right\|_2^2 + \lambda_n \sum_{j=1}^{p} w_{nj} \left|\Pi_{nj}\right| \right) \tag{1.6}$$

The expression above has two components. The first one $\left\| Y_{2n} - \sum_{j=1}^{p} Z_{nj}\Pi_{nj} \right\|_2^2$ is the loss function of a least squares problem, where for a vector $a$ its norm is defined by $\|a\|_2 \equiv \left( \sum_{j=1} \left|a_j\right|^2 \right)^{1/2}$. The second expression is a penalty that arises from the restriction $\sum_{j=1}^{p} w_{nj} \left|\Pi_{nj}\right| \leq c$, where $c$ is a constant. In this context $\lambda_n$ is the Lagrange multiplier associated with the previously mentioned constraint that regulates the amount of shrinkage that is occurring. Notice that if $\lambda_n = 0$ the problem would simplify to a least squares estimation.

The main difference between the standard LASSO and the adaptive LASSO in (1.6) are the weights, $w_{nj}$, which are not present in the original Tibshirani (1996) formulation. Zou (2006) proposes that these weights be a function of the ordinary least squares estimate of $\Pi_n$, which I will denote by $\widehat{\Pi}_{n(OLS)}$. In his framework the weights are given by

$w_{nj} = \frac{1}{|\widehat{\Pi}_{nj(OLS)}|^{\gamma}}$, where $\gamma > 0$. The addition of the weights addresses the fact that the LASSO generates biased estimates of the larger coefficients and, under certain situations, is inconsistent for variable selection. By adding these weights, the adaptive LASSO solves these two difficulties and attains what the model selection literature refers to as the oracle property. Specifically, if we define $\mathcal{A} = \{j : \Pi_{nj} \neq 0\}$:

**Definition 1** (Oracle Property)**.** An estimation procedure has the **oracle property** if it asymptotically:

- Identifies the correct subset model $P\left(\{j : \hat{\Pi}_{nj} \neq 0\} = \mathcal{A}\right) \rightarrow 1$
- Has the optimal estimation rate, $\sqrt{n}\left(\hat{\Pi}_{nj_{\mathcal{A}}} - \Pi_{nj_{\mathcal{A}}}\right) \rightarrow_d N\left(0, V\right)$

Proposition 3 guarantees that the first component of the oracle property is attained by the adaptive LASSO. [6] This result is important because it allows researchers to argue for the validity of the selected instruments and eliminates with high probability instruments that provide no signal.

The adaptive LASSO introduced by Zou (2006) obtains the oracle property for a fixed number of regressors $p$. However, I exploit the properties of the adaptive LASSO when the number of regressors, $p_n$, grows with the sample size. Huang et al. (2007) has established the oracle properties under this scenario.[7]

[6] The second condition of the oracle property is not relevant here since the main interest is in the coefficients of the second stage.

[7] Following Huang et al. (2007), this study suggests using $w_{nj} = \frac{1}{|\hat{\Pi}_{nj(OLS)}|^{\gamma}}$ with $\gamma = 1$ when $p_n \leq n$. When $p_n \geq n$ a different set of weights is suggested. I will define these weights in section 1.2.2.

Using the adaptive LASSO first stage estimates, however, leaves unattended its consequent post-model-selection bias. This will be addressed using the estimator proposed by Belloni and Chernozhukov (2010). They suggest an $l_1$-penalized method, like the adaptive LASSO, to select the instruments and then run OLS with the selected regressors. They prove this yields a smaller bias than the LASSO. The simulations presented below, under different degrees of endogeneity and instrument weakness, confirm this fact. It is important to highlight that the method suggested by Belloni and Chernozhukov (2010) is used as a bias reducing mechanism only. Specifically, the assumptions that the errors are homoskedastic and Gaussian, as presented in Belloni and Chernozhukov (2010) and Belloni et al. (2010a), are not present here. The results in this paper follow in the presence of heteroskedasticity and may allow for errors with tail behavior that ranges from subgaussian to exponential and therefore is in line with Belloni et al. (2010b).

### 1.2.2 Mean Independence Between the Instruments and the Random Disturbances

The mean independence assumption modifies equation (1.4) which is now given by:

$$E\left(u_i|Y_{2i}\right) \neq \mathbf{0}, \quad E\left(v_i'|Z_{ni}\right) = 0 \quad \text{and} \quad E\left(u_i|Z_{ni}\right) = 0 \tag{1.7}$$

Equation (1.7) is stronger than the uncorrelated instruments and random disturbances in (1.4). The main reason for this new assumption is that the framework in this section will allow for arbitrary functions of $Z_n$ and not only a linear combination of its elements.

Under (1.7) and assuming an additive form for the first stage[8]:

$$Y_{2ni} = \sum_{j=1}^{p_n} f_j\left(Z_{nij}\right) + v_i \tag{1.8}$$

The $f_j$ functions in (1.8) are approximated using truncated series expansions with B-splines. I assume that many of the additive components are zero and some of the terms in the expansion are weakly related to the endogenous variable. As in section 1.2.1 a set of coefficients is going to be selected. These coefficients are going to be used to approximate each nonparametric component.

For completeness, I present the spline framework as introduced in Huang et al. (2010). Let $Z_{nij}$ take values in $[a,b]$ where $a < b$ are finite numbers. Let $[a,b]$ be partitioned into $T$ subintervals, $I_{Tk} = [\xi_k, \xi_{k+1}), k = 0, \ldots, T-1$, and $I_{TT} = [\xi_T, \xi_{T+1}]$, for $T \equiv T_n = n^{\nu}$ where $0 < \nu < 0.5$ satisfies $\max_{1 \leq k \leq T+1} |\xi_k - \xi_{k-1}| = O(n^{\nu})$. Defining $\mathcal{S}_n$ as the space of

[8] In Huang et al. (2010) the model is given by:

$$Y_{2i} = \mu + \sum_{j=1}^{p_n} f_j\left(Z_{nj}\right) + v_i$$

For notational simplicity $\mu = 0$. In an earlier version of the paper results are derived with $\mu \neq 0$.

polynomial splines of degree $l \geq 1$ consisting of functions $s$ that satisfy: (i) the restriction of $s$ to $I_{Tk}$ is a polynomial of degree $l$ for $1 \leq k \leq T$; (ii) for $l \geq 2$ and $0 \leq l' \leq l-2$, $s$ is $l'$ times continuously differentiable on $[a, b]$.

Under smoothness conditions stated in section 1.2.3, the $f_j$'s can be arbitrarily well approximated by a function $f_{nj} \in \mathcal{S}_n$. The $f_{nj} \in \mathcal{S}_n$ are defined by the existence of a normalized B-spline basis $\{\psi_t, 1 \leq \ t \leq s_n\}$ for $\mathcal{S}_n$, with $s_n \ \equiv \ T_n + l$ which yields the expression:

$$f_{nj}\left(z\right) = \sum_{t=1}^{s_n} \Pi_{jt} \psi_t \left(z\right) \tag{1.9}$$

Although $\Pi_{jt}$ has different dimensions than $\Pi_{nj}$ in section 1.2.1, I will refer to them both as $\Pi_{nj}$ throughout the document. It will be clear from the context if reference is made to random disturbances that are uncorrelated or mean independent of the instruments. The principal reason for this notational simplification is that the assumptions and properties regarding $\Pi_n$ will be analogous in both cases.

Furthermore, as in Huang et al. (2010) it is assumed that $Ef_j\left(Z_{nij}\right) = 0$. This gives rise to the sample analog:

$$\sum_{j=1}^{n}\sum_{t=1}^{s_n}\Pi_{jt}\psi_t\left(Z_{nij}\right)=0, \quad 1\leq j\leq p_n \tag{1.10}$$

To simplify the notation let $\psi_t(z) = \psi_{jt}(z)$ and $X_{nij} \equiv \left(\psi_1\left(Z_{nij}\right), \ldots, \psi_{s_n}\left(Z_{nij}\right)\right)'$. Also define $X_{nj} \equiv \left(X_{n1j}, \ldots, X_{nnj}\right)'$, the $n \times s_n$ design matrix associated with the $j$th covariate. Finally, let the design matrix be given by $X_n$.

Under this framework, denoting the basis functions selected by the nonparametric adaptive LASSO by $\tilde{X}_{n1}$, the estimator of the parameter of interest, $\beta$, is described by:

$$\begin{aligned}\hat{\beta} &= \left(\hat{Y}_{2n}'\hat{Y}_{2n}\right)^{-1}\left(\hat{Y}_{2n}y_{1n}\right)\\ \hat{Y}_{2n} &= \tilde{X}_{n1}\left(\tilde{X}_{n1}'\tilde{X}_{n1}\right)^{-1}\tilde{X}_{n1}'Y_{2n}\end{aligned}$$

Once again, the adaptive LASSO is only used for instrument selection. Its estimates are not used to construct $\hat{Y}_{2n}$.

The framework described above provides a version of (1.6) that incorporates the mean independence assumption and the additive form in (1.8) that is given by:

$$\hat{\Pi}_n = \arg\min_{\Pi_n}\left(\left\|Y_{2n}-\sum_{j=1}^{p_n}X_{nj}\Pi_{nj}\right\|_2^2+\lambda_{nl}\sum_{j=1}^{p_n}w_{nj}\left\|\Pi_{nj}\right\|_2\right) \tag{1.11}$$

The form of (1.11), which comes from the additivity structure assumed for (1.8), is important because the concepts of instrument weakness and instrument selection present themselves in a form that resembles the case of uncorrelated instruments and random disturbances of section 1.2.1. In particular, we can think of $\Pi_n$ as being zero or nearly zero as in section 1.2.1.

The main distinctions between equations (1.11) and (1.6) are that $Z_{nj}$ has been replaced by a set of basis functions $X_{nj}$, another regularization parameter, $\lambda_{nl}$, is used and the weights, defined in (1.12) below, will have a different form. As in section 1.2.1, the adaptive LASSO is selecting instruments. Through (1.11), a group of coefficients is being selected to approximate each nonparametric component $f_j$. This is clear by inspection of $X_{nj} \equiv \left(\psi_1\left(Z_{nj}\right), \ldots, \psi_{s_n}\left(Z_{nj}\right)\right)$ given that when $\Pi_{nj}$ is zero so is $Z_{nj}$. It follows that weak instruments, in the model described by (1.4) and the model described by (1.7), will refer to the $Z_{nj}$'s that do not provide a strong signal. The addition of another regularization parameter, $\lambda_{nl}$, implies that the amount of shrinkage is going to be different when the random disturbances are uncorrelated with the instruments.

Instead of obtaining the weights in (1.11) from a least squares optimization problem, I find them using LASSO to select a group of coefficients to approximate each additive component $f_j$. This incorporates the possibility that $p_n \geq n$ that I allow in the model with the mean independent random disturbances.[9] Defining the LASSO estimates as $\tilde{\Pi}_{nj}$ the

[9] In section 1.2, $p_n \leq n$. Here I relax that assumption to illustrate the possibility that, when $p_n \geq n$, a

weights are given by:

$$w_{nj} = \begin{cases} \|\tilde{\Pi}_{nj}\|_2^{-1} & \text{if } \|\tilde{\Pi}_{nj}\|_2 > 0, \\ \\ \infty & \text{if } \|\tilde{\Pi}_{nj}\|_2 = 0. \end{cases} \tag{1.12}$$

### 1.2.3 Assumptions, Definitions, and Regularity Conditions

This subsection lists and interprets the main assumptions, definitions, and regularity conditions used in the paper and discusses their relevance.

**Assumption 1.** $\Pi_n = \frac{C_n}{b_n}$ for some sequence of positive real numbers $\{b_n\}$, nondecreasing in $n$. In the parametric model of section 1.2.1 $\{C_n\}$ is a sequence of nonrandom, $p_n \times 1$ parameter vectors. In the nonparametric model of section 1.2.2 $\{C_n\}$ is a sequence of nonrandom, $(p_n s_n) \times 1$ parameter vectors.

**Assumption 2.** Let $\{Z_n\}$ be a triangular array of $\mathcal{R}^{p_n}$-valued random variables with finite first and second moments, $\tau$ denote a generic eigenvalue, $\underline{\lim}$ the limit inferior, and $\overline{\lim}$ the limit superior. Also, let $\{X_n\}$ be a set of $\mathcal{R}^{p_n \times s_n}$-valued basis functions depending on $Z_n$, and:

(a) For the model of section 1.2.1, $p_n \to \infty$ as $n \to \infty$ such that $\frac{p_n}{n} \to \alpha$ where $\alpha \in [0, 1)$

(b) There exists a sequence of positive real numbers $\{m_n\}$, nondecreasing in $n$, and con-

different set of weight functions and a different amount of regularization are needed.

stants $D_1$ and $D_2$, with $0 < D_1 \leq D_2 < \infty$, such that

$$D_1 \leq \underline{\lim}_{n\to\infty} \tau_{\min}\left(\frac{Z_n' Z_n}{m_{n1}}\right) \qquad \text{a.s}$$

and

$$\overline{\lim}_{n\to\infty} \tau_{\max}\left(\frac{Z_n' Z_n}{m_{n2}}\right) \leq D_2 \qquad \text{a.s}$$

(c) There exists a sequence of positive real numbers $\{m_{nx}\}$, nondecreasing in $n$, and constants $D_{1x}$ and $D_{2x}$, with $0 < D_{1x} \leq D_{2x} < \infty$, such that

$$D_{1x} \leq \underline{\lim}_{n\to\infty} \tau_{\min}\left(\frac{X_n' X_n}{m_{n1x}}\right) \qquad \text{a.s}$$

and

$$\overline{\lim}_{n\to\infty} \tau_{\max}\left(\frac{X_n' X_n}{m_{n2x}}\right) \leq D_{2x} \qquad \text{a.s}$$

(d) There exists a sequence of positive real numbers $\{h_n\}$, nondecreasing in $n$ and constants $D_3$ and $D_4$, with $0 < D_3 \leq D_4 < \infty$, such that

$$D_3 \leq \underline{\lim}_{n\to\infty} \tau_{\min}\left(\frac{C_n' C_n}{h_n}\right)$$

and

$$\overline{\lim}_{n\to\infty}\tau_{\max}\left(\frac{C_n'C_n}{h_n}\right)\leq D_4$$

**Assumption 3.** Let $z_i$ and $x_i$ be the $i^{th}$ rows of $Z_n$ and $X_n$ respectively. Then for constants $0 < D_7 < \infty$ and $0 < D_{7x} < \infty$:

$$\begin{aligned}\max_{1\leq i\leq n}\left|z_i'\left(Z_n'Z_n\right)z_i\right| &\leq D_7\frac{p_n}{m_{n1}}\\ \max_{1\leq i\leq n}\left|x_i'\left(X_n'X_n\right)x_i\right| &\leq D_{7x}\frac{p_n s_n}{m_{n1x}}\end{aligned}$$

**Assumption 4.** Let us define $S = w'sign\left(\Pi_n\right)$ where $w$ is the vector of weights defined in equation (1.6). It is assumed that there exist constants $0 < D_8 \leq D_9 < \infty$ and a sequence of positive real numbers $\{l_n\}$ such that:

$$\overline{\lim}_{n\to\infty}\tau_{\max}\left(\frac{S'S}{l_n}\right)\leq D_9 \qquad \text{a.s}$$

and

$$D_8\leq\underline{\lim}_{n\to\infty}\tau_{\min}\left(\frac{S'S}{l_n}\right) \qquad \text{a.s}$$

**Assumption 5.** Define $v_{ik}$ as an element of the matrix $V_n$ and $\eta_i \equiv \left(u_i, v_i'\right)'$ where $\eta$ are identically distributed, $v_i$ is a row $V_n$, and:

(a) $\eta_i \perp Z_n$, $E\left(\eta_i\right) = 0$, and $\mathrm{Var}\left(\eta_i\right) = \Sigma_i$, where $\Sigma_i$ is positive definite and defined to be

$$\begin{pmatrix} \sigma_{uu} & \sigma'_{Vu} \\ \sigma_{Vu} & \Sigma_{VV} \end{pmatrix}.$$

(b) $Eu_i^4 < \infty$ and $Ev_{ik}^4 < \infty$

(c) $E\left(u_i^3\right) = E\left(v_{ik}^3\right) = E\left(u_i^2 v_{ik}\right) = E\left(v_{ik}^2 u_i\right) = 0$

(d) The tail probabilities of the individual components of the random vector $V_n$, $v_i$, satisfy for certain constants $D_5 > 0$, $D_6$, and $1 \leq d \leq 2$, $P\left(|v_{ik}| > t\right) \leq D_6 exp\left(-D_5 t^d\right)$ for all $t \geq 0$ with $i = 1, 2, \ldots n$.

**Assumption 6.** For a constant $\vartheta$ such that $0 \leq \vartheta < \infty$:

(a) Define the ratio $r_n = \frac{m_{n2} h_n}{b_n^2}$. As $n \to \infty$, $\frac{r_n}{n} \to \vartheta$ and $\frac{\Pi'_n Z'_n Z_n \Pi_n}{r_n} \to \Phi$ almost surely for a nonrandom positive constant $\Phi$.

(b) Define the ratio $r_{nx} = \frac{m_{nx2} h_n}{b_n^2}$. As $n \to \infty$ $\frac{\Pi'_n X'_n X_n \Pi_n}{r_{nx}} \to \Phi_x$ almost surely for a positive constant $\Phi_x$.

**Assumption 7.** Let $\pi_n \equiv min\left\{\left|\Pi_{nj}\right| : j \in \mathcal{A}\right\}$. Then the variables $\{p_{n1}, p_{n2}, \lambda_n, \lambda_{nl}, l_n, m_{n2}, m_{nx2}, s_n\}$ satisfy the following conditions:

(a)

$$\frac{\lambda_n \sqrt{l_n}}{\sqrt{m_{n2}}}, \quad log\left(n\right)^{I\{d=1\}}\left(\frac{log(p_{n1})^{1/d}}{\sqrt{m_{n2}}\pi_n}\right), \quad \text{and} \quad log\left(n\right)^{I\{d=1\}}\left(\frac{log(p_{n2})^{1/d}}{\lambda_n}\right) \to 0$$

(b)

$$log\,(n)^{I\{d=1\}}\left(\frac{log(p_{n2})^{1/d}\sqrt{r_n}}{\lambda_n}\right)\to 0$$

(c)

$$\frac{qnlog(p_n s_n)}{m_{nx2}^2},\quad \frac{nq}{\sqrt{m_{nx2}}}\left[s_n^{-\partial}+\sqrt{\frac{s_n}{n}}\right]^2,\quad \frac{\lambda_{nl}^2 q\sqrt{r_{nx}}}{m_{nx2}^2}\quad \text{and}\quad \frac{\lambda_{nl}^2 q}{m_{nx2}^2}\to 0$$

**Assumption 8.** Let $\kappa$ be a nonnegative integer, and $\varrho \in (0,1]$ in such a way that $\partial = \kappa+\varrho > 0.5$. Also, let $\mathcal{F}$ be a class of functions on $[0,1]$ whose $\varrho$th derivative $f^{(\kappa)}$ exists and satisfies for a constant $C \geq 0$:

$$\left|f^{(\kappa)}(s)-f^{(\kappa)}(t)\right|\leq C\,|s-t|^{\varrho}\quad \text{for } s,t\in[a,b]$$

**Assumption 9.** Let $f_j(z)\neq 0$, $1\leq j\leq q$, but $f_j(z)\equiv 0$, $q+1\leq j\leq p_n$. Also define $\|f\|_2=\left[\int_a^b f^2(z)\,dz\right]^{1/2}$ for any function when the integral exists. Then the number of nonzero components $q$ is fixed and there exists a constant $c_f$ such that $\min_{1\leq j\leq q}\|f_j\|_2\geq c_f$.

**Assumption 10.** $Ef_j(Z_{nj})=0$ and $f_j\in\mathcal{F}, j=1,\dots q$.

**Assumption 11.** For the model described in (1.7) and (1.8) the covariate vectors $Z_{nj}$ have continuous densities and there exist constants $\underline{C}$ and $\bar{C}$ such that the density function $g_j$ of $Z_{nj}$ satisfies $0<\underline{C}\leq g_j\leq\bar{C}<\infty$ on $[a,b]$ for every $1\leq j\leq p_n$.

**Proposition 1.** Under assumptions 1 to 11 for $\lambda_{nl} \geq C\sqrt{nlog\left(p_n s_n\right)}$ and a constant $C > 0$. The $\tilde{\Pi}_{nj}$ of the nonparametric model described in (1.11) satisfy:

i) All $\Pi_{nj} \neq 0$ are selected with probability converging to one.

ii)

$$\sum_{j=1}^{p_n} \|\tilde{\Pi}_{nj} - \Pi_{nj}\|_2^2 = O_p\left(\frac{nqlog\left(p_n s_n\right)}{m_{nx2}^2}\right) + O\left(\frac{nq}{m_{nx2}}\left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^2\right) + O\left(\frac{\lambda_{nl}^2 q}{m_{nx2}^2}\right)$$

Proposition 1 refers to the behavior of the weights used by the adaptive LASSO in the nonparametric case. The initial weights should, in the linear and nonparametric cases, be consistent estimators for the first stage parameter. In the case of the linear model the OLS estimator clearly attains this goal. In the case of the nonparametric model Proposition 1 shows this.

Proposition 1 is grouped with the assumptions because it constitutes a condition that precedes the estimation stage and should be satisfied for the proposed methodology to be selection consistent, for the first stage estimator, and consistent and asymptotically normal, in the case of the second stage. A proof of Propositions 1 is provided in section 1.7.

### Discussion of Assumptions and Regularity Conditions

Chao and Swanson (2003b, 2005b) show that the concentration parameter, given assumptions 1 to 6 above, is of the order $O(r_n)$ almost surely whereas in Staiger and Stock (1997) it is fixed. It is via the characterization of the behavior of $r_n$ that Chao and Swanson (2003b, 2005b) bring together the local to zero Staiger and Stock (1997) framework with the many instrument ideas of Bekker (1994) and Morimune (1983). Allowing the concentration parameter to grow with the sample size guarantees consistency of LIML and Fuller by modeling the additional instruments as a source of increasing information asymptotically.

A useful way to interpret $r_n$, as discussed in Chao and Swanson (2003b), is to think of it as divided into two components. The first one is $\frac{h_n}{b_n^2}$. This component explicitly defines the local to zero part via $b_n^2$, but accounts for the possibility of the instruments having a signal asymptotically, $C_n' C_n = O(h_n)$. The second component, $m_{n2}$, is the rate at which the information in the instruments accumulates.

For instance, in the Staiger and Stock (1997) framework $b_n = \sqrt{n}$, $m_{n2} = n$, and the number of instruments is fixed for all $n$, making $C_n' C_n = O\left(1\right)$ and $h_n = 1$. This implies that the concentration parameter is constant: $r_n = 1$. Under this setup, no conventional simultaneous equations estimator is consistent if the instruments set is weak. If, however, the instruments provide some information asymptotically ($h_n \rightarrow \infty$ as $n$ grows and $r_n = h_n$), the concentration parameter growth can yield consistency even under instrument weakness as proved by Chao and Swanson (2003b).

In this paper these ideas are extended to the nonparametric case via $r_{nx}$ which has the same interpretation as $r_n$. In the case of the nonparametric model the concentration parameter is given by $\Sigma_{VV}^{-1/2}\Pi_n' X_n' X_n \Pi_n \Sigma_{VV}^{-1/2}$. Here it is important to highlight that the concentration parameter is defined with respect to the spline functions, $f_{nj}$, and not to $f_j$. Also, by definition $X_n$ is a set of basis functions that depend on $Z_n$ and, therefore, a weak signal is inherently related to $Z_n$.

Assumption 2 allows for different growth rates of the sequences $\Pi_n$, $Z_n$, $X_n$. In the case of $Z_n$, and analogously for $X_n$, assumption 2(b) implies that $\tau_{max}\left(Z_n' Z_n\right) = O(m_{n2})$ almost surely and $\tau_{min}\left(Z_n' Z_n\right) = O(m_{n1})$ almost surely and is in line with Portnoy (1984), a seminal work for linear regression when the regressors grow with the sample size, and Koenker and Machado (1999). Portnoy (1984) argues that for a fixed number of regressors $m_{n1} = m_{n2} = n$ but when the regressors grow with the sample size this is not necessarily true. He constructs results for the case where $m_{n1} = m_{n2} = n$ and then discusses under which circumstances the condition is satisfied. For example he shows that the setup with $m_{n1} = m_{n2} = n$ addresses situations where the regressors are normal or a scale mixture of normals. On the other hand, Koenker and Machado (1999) make explicit use of rates of growth of the eigenvalues of the information matrix that are different than $n$. They allow $\tau_{min}$ and $\tau_{max}$ to have different growth rates. In the case of the nonparametric model a similar analysis follows for $m_{n1x}$ and $m_{n2x}$ in the place of $m_{n1}$ and $m_{n2}$. The conditions in assumption 2 can be understood as equivalent to strong laws of large numbers.

In the context of this paper allowing for different rates of growth for $\tau_{max}\left(Z_n'Z_n\right)$ and $\tau_{min}\left(Z_n'Z_n\right)$ is going to be important because it denotes the fact that there are some instruments that provide different information than others. This is important because I am considering instruments that are weak along side with instruments that are useless. Chao and Swanson (2003b, 2005b) impose the condition that $\tau_{max}\left(Z_n'Z_n\right) = O(m_n)$ and $\tau_{min}\left(Z_n'Z_n\right) = O(m_n)$ almost surely which is tantamount to all the instruments considered being informationally equivalent.

Assumption 3 is assumption $X2$ of Koenker and Machado (1999). In their paper it is important to establish their inference results. In the case of this paper assumption 3 is going to be important to characterize the asymptotic behavior of the first stage estimators once they are plugged into the second stage.

Assumption 5(a) explicitly allows for heteroskedasticity of the variance-covariance matrix. The only requirement for the proposed estimator to be robust to heteroskedasticity is that the elements of the variance-covariance matrix are bounded. This requirement is given by Assumption 5(b).

Assumption 5(c) comes from Chao and Swanson (2003a). It imposes symmetry on the distribution of the disturbances and allows the results to hold for the family of elliptical distributions. It is similar to assumption $U1$ of Koenker and Machado (1999). The assumption is important to establish the asymptotic normality of the estimator.

Assumption 5(d) imposes restrictions on the tails of the distribution of the random disturbance $V_n$. For instance if $d = 2$ in the expression $P\left(|V_{ni}| > t\right) \leq D_6 exp\left(-D_5 t^d\right)$ the tail behavior is subgaussian and if $d = 1$ the tail behavior is exponential. This tail behavior assumption defines exponential Orlicz norms that allow the use of the maximal inequalities defined by van der Vaart and Wellner (1996). This restriction is also used in Huang et al. (2007) to obtain the variable selection consistency of the adaptive LASSO. Moreover, it provides a sense of the number of variables that will be selected by the adaptive LASSO. If the tails are exponential the number of variables selected will be fewer than if the tails are subgaussian. This distinction is not present in Belloni et al. (2010a) where the error terms are required to be Gaussian or in Huang et al. (2010) where the tail behavior of the errors is assumed to be subgaussian. I modify the proofs of Huang et al. (2010) to allow for the possibility that $d \in [1, 2]$.

With regards to assumption 6, the statement that $\frac{\Pi_n' Z_n' Z_n \Pi_n}{r_n} \to \Phi$ is not unique to this paper. Chao and Swanson (2005b) require this condition. It is also invoked for certain proofs in Chao and Swanson (2003a). However, the condition is not restrictive if one notes that by assumption 2, $\frac{\Pi_n' Z_n' Z_n \Pi_n}{r_n} = O(1)$ almost surely. The same assumption is extended to the nonparametric case.

Assumption 7(a) restricts the asymptotic behavior of $\lambda_n$ and the number of zero and nonzero coefficients. It is a restatement of assumption $A4$ of Huang et al. (2007) that incorporates the behavior of the information matrix of the instruments accounted for in assumption 2 instead of the assumption $\frac{1}{n}\sum_{i=1}^{n} Z_{nij}^2 = 1$ for $j = 1 \ldots p_n$.

Assumption 7(b) is needed for the proof of consistency and asymptotic normality of the proposed estimator. In the literature it is common to obtain results depending on the amount of regularization and the growth rate of $p_{n1}$ and $p_{n2}$.[10] Additionally, the growth rate of the concentration parameter and of the information matrix enter the expressions in assumption 7(b). This is a consequence of the conjunction of the two asymptotic frameworks in this paper. Assumption 7(c) is a restatement of (b) for the nonparametric case.

The terms in assumption 7(d) come from the assumptions of Huang et al. (2010) incorporating the notation of assumption 2. The elements of the assumption are necessary to guarantee consistency of the nonparametric estimator once the results of Proposition 2 are incorporated to the analysis. One important consequence of assumption 7(d) is that the approximating spline functions are of order $O\left(\frac{nq}{m_{nx2}}\left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^2\right)$. To provide some intuition, assumption 7(d) would follow if $q$ were taken to be constant and

[10] An excellent example of a wide class of solutions that arise in the LASSO literature, embedding a wide class of loss functions and regularization schemes, is Zhang et al. (2010).

$s_n = O\left(n^{1/(2\partial+1)}\right)$, when $\frac{n^{1/(2\partial+1)}}{\sqrt{m_{nx2}}} \to 0$. This condition would be satisfied immediately if $m_{nx2} = n$ and is likely to be satisfied if $f_j$ is sufficiently smooth.

Assumptions 8, 10, and 11 come from Huang et al. (2010). These assumption are used in the literature of nonparametric spline models to estimate the nonzero $f_j$ components. Assumption 9, that artificially bounds the $f_j$'s away from zero, is not present in the traditional nonparametric spline model literature. This is related to the fact that Huang et al. (2010) are considering zero and nonzero $f_j$ components. I will be able to establish this bound for the approximating spline functions $f_{nj}$.

Proposition 1 is satisfied if, as discussed in Huang et al. (2010), for an increasing sequence $e_n \to \infty$, that is defined by the convergence rates in part ii) of Proposition 1, the initial estimators $\tilde{\Pi}_{nj}$ satisfy two conditions:

$$e_n \max_{j \notin \mathcal{A}} \left\|\tilde{\Pi}_{nj}\right\|_2 = O_p\left(1\right)$$

And for a constant $c > 0$

$$P\left(\min_{j \in \mathcal{A}} \left\|\tilde{\Pi}_{nj}\right\|_2 \geq c\pi_n\right) \to 1$$

These two relationships imply that the zero an non-zero components are estimated consistently by the weights.

### Definition of Instrument Weakness

The notion of instrument weakness from Chao and Swanson (2003b) that will be used here is given by:

**Definition 2** (Instrument Weakness)**.** The following classification will be used to denote instrument weakness:

- The set of available instruments is **not weak** if $\frac{p_n}{r_n} \to 0$ as $n \to \infty$
- The set of available instruments is **weak** if $\frac{p_n}{r_n} \nrightarrow 0$

The definition highlights the importance of the growth rate of the concentration parameter relative to the number of instruments. For instance, if $r_n$ grows at a faster rate than $p_n$ it implies that the signal is growing for every additional instrument. In this case, the instruments are **not weak**. In a similar fashion if $\frac{p_n}{r_n} \to \delta_1$ bor $0 \leq \delta_1 < \infty$ the signal for every additional instrument is vanishing asymptotically and instruments are weak. Definition 2 can be extended to the nonparametric case by replacing $r_n$ with $r_{nx}$ and $p_n$ with $p_n s_n$. These two differences come from the fact that each instrument is associated with a number $s_n$ of basis function and the fact that the asymptotic behavior of the concentration parameter in the nonparametric framework is given by $r_{nx}$.

### 1.2.4 Smallest Nonzero Coefficient and Oracle Property

The objective in what follows is to understand how the proposed estimator behaves under the taxonomy of definition 2. The first component needed for the analysis is the threshold between the coefficients that are chosen to be zero and those that are not. The second component is the first part of the oracle property in definition 1, which establishes that the adaptive LASSO selects the correct model asymptotically.

Various studies derive theoretical results by imposing such a threshold. In this paper the approach of Zhang and Huang (2008) is used to determine it endogenously. Their proofs are modified to arrive at the conclusion for error terms that have subgaussian tails and for a behavior of the regularization parameter consistent with assumption 7(a). The idea is to obtain a conservative lower bound away from zero for the coefficients of the selected instruments. This is important because the conclusions of this paper depend on the fact that after instrument selection the term $\frac{\hat{p}_{n1}}{r_n} \to 0$, where $\hat{p}_{n1}$ is the number of selected instruments, while $\frac{p_n}{r_n} \to \infty$. The conditions $\frac{p_n}{r_n} \to \infty$ and $\frac{\hat{p}_{n1}}{r_n} \to 0$ imply that before introducing the adaptive LASSO instruments are weak but that after selection a condition that is equivalent to having strong instruments might be attained. Including a reduced number of instruments will make this conclusion artificially strong. Allowing subgaussian tails the maximum possible number of instruments and the most conservative lower bound possible for the selected coefficients is allowed. If $\frac{\hat{p}_{n1}}{r_n} \to 0$ is valid for subgaussian tails it will necessarily be true for heavier tails.

Zhang and Huang (2008) do not attempt for the LASSO to be selection consistent but try to analyze LASSO under weaker conditions. The idea in their work is that the coefficients outside an ideal model are assumed to be small, but not necessarily zero. By relaxing the assumption that the excluded coefficients are exactly zero, their approach is ideal for constructing a conservative lower bound.

To see how this works, define:

$$\zeta_{\alpha_1} \equiv \left( \sum_{j \in \mathcal{A}} \left| \Pi_{nj} \right|^{\alpha_1} I \left\{ \hat{\Pi}_{nj} = 0 \right\} \right)^{1/\alpha_1} \quad \alpha_1 \in [0, \infty] \tag{1.13}$$

$$B\left(\lambda\right) \equiv \left\| Z_n \Pi_n - Z_n \hat{\Pi}_n \right\|_2 \tag{1.14}$$

$$\eta_2 \equiv \max_{\mathcal{A}' \subset \mathcal{A}^c} \left\| \sum_{j \in \mathcal{A}'} Z_{nj} \Pi_{nj} \right\|_2 \tag{1.15}$$

Above (1.13) is a measure of the number of coefficients excluded by the model that should have been included, $B\left(\lambda\right)$ in (1.14) is a measure of the model's bias, and (1.15) is an upper bound on the contribution of a given set of the excluded regressors. Under this framework Zhang and Huang (2008) prove that the LASSO selects a model of the correct order of dimensionality $p_{n1}$, controls the bias of the selected model at a level determined by the contributions of small regression coefficients, and selects all coefficients of greater order than the bias of the selected model. This is equivalent to:

$$P\left\{\mathcal{A} \subset \hat{\mathcal{A}}, \quad B\left(\lambda\right) \leq \eta_2 \quad \text{and} \quad \zeta_{\alpha_1} = 0\right\} \rightarrow 1 \tag{1.16}$$

The arguments of Zhang and Huang (2008) imply that the threshold between the selected and excluded coefficients is related to the fact that the adaptive LASSO may include some variables that are not relevant and exclude some coefficients that are relevant. Specifically, under what they refer to as a sparse Riesz condition, the LASSO selects all coefficients of greater order than the bias of the selected model, $B\left(\lambda\right)$. The size of this error, under the assumptions of this paper, is $O\left(\frac{\lambda_n\sqrt{p_{n1}}}{\sqrt{m_{n2}}}\right)$. This also constitutes the threshold between the included and excluded instruments as stated in Proposition 2.

Belloni and Chernozhukov (2010) also acknowledge the existence of this bias and try to quantify it for Gaussian errors within their framework. In this paper the post-$l_1$-penalized methodology they propose is used to reduce this bias. Simulation results, however, show that bias reduction is also effective when the errors are not Gaussian and suggests that the post-$l_1$-penalized methodology is ideally endowed to reduce the bias of the estimators produced by the adaptive LASSO.

To allow for subgaussian tails the following regularity condition needs to be imposed[11]:

**Regularity Condition 1.** The constant $c_0$ which will determine the amount of shrinkage in (1.6) is such that $c_0 \geq 0$ and $c_0 \geq \frac{1}{D_5} - 1$.

The result that defines the threshold between the included and excluded coefficients is given by:

**Proposition 2** (Minimum $\Pi_{nj}$)**.** Under assumptions 2 to 7, and regularity condition 1 the set $\mathcal{A}^c \subset \{1, \ldots, p_n\}$ satisfies:

$$\#\{j \leq p_n : j \notin \mathcal{A}^c\} = p_{n1}, \qquad \sum_{j \in \mathcal{A}^c} |\Pi_{nj}| \leq \eta_1 \tag{1.17}$$

Under these conditions the smallest of the coefficients in $\Pi_{n1}$ is of order greater than $\frac{\lambda_n \sqrt{p_{n1}}}{\sqrt{m_{n2}}}$.

A proof of Proposition 2 is available in the appendix (Section 7.1.2). Proposition 2 gives us an approximation of the coefficients that are deemed to be non-zero by the

[11] In regularity condition 1 the constant $D_5$, which is presented in assumption 5, comes from the Orlicz norm used in Huang et al. (2007). The Orlicz norm as is introduced in van der Vaart and Wellner (1996) is defined by $\|X\|_\psi = inf \left\{ D_5 > 0 : E\psi\left(\frac{|X|}{D_5}\right) \leq 1 \right\}$. On the other hand, $c_0$ is a constant defined in Zhang and Huang (2008) that affects the level of the minimum $\lambda_n$. A bigger $c_0$ increases the possibility that the adaptive LASSO selects variables to be zero which are not.

adaptive LASSO. It is an asymptotic approximation of the order of $\Pi_n = \frac{C_n}{b_n}$ for the coefficients that are included. In particular, a consequence of Proposition 2 is that the coefficients of the disregarded instruments are of the order $O\left(\frac{\lambda_n\sqrt{p_{n1}}}{\sqrt{m_{n2}}}\right)$.

In Proposition 2 expression (1.17) is the sparsity condition of Zhang and Huang (2008). It tells us that there are a maximum of $p_{n1}$ large coefficients and that the sum of the $l_1$ norms of the small coefficients is no greater than $\eta_1$. Another way of thinking about it is that the excluded instruments are either zero or very small. Assumption 2 gives us the Riesz condition of Zhang and Huang (2008). It is necessary for Proposition 2 and states that the smallest eigenvalue of $\left(\frac{Z_n'Z_n}{m_{n2}}\right)$ is bounded away from zero and the largest one is finite. These two conditions imply the fulfillment of the sparse Riesz condition in Zhang and Huang (2008) and allow the use of their theoretical framework in the context of this paper.

**Corollary 1.** If the conditions of Proposition 2 are satisfied, for the nonparametric model the smallest of the coefficients in $\Pi_{n1}$ is of order greater than $\frac{\lambda_n\sqrt{p_{n1}s_n}}{m_{nx2}}$.

The distinctions between the instruments selection in the parametric and nonparametric case, given that the same assumptions are made about the $\Pi_n$ coefficients, are the dimensions of $\Pi_n$ and that the weights are based on the LASSO. Thus, the $p_{n1}s_n$ term instead of $p_{n1}$ and the additional $m_{nx2}$ in the statement of the Corollary. The last distinction is a consequence of the fact that the estimator of the weights, LASSO, is already excluding some of the coefficients. In particular Corollary 1 can be interpreted as a restatement of Corollary 2.1 of Wei and Huang (2008).

Proposition 2 and Corollary 1 establish the mechanics of how the adaptive LASSO selects instruments. The following propositions determine the accuracy of the procedure to exclude the irrelevant instruments:

**Proposition 3.** Under assumptions 1 to 7 the adaptive LASSO identifies the correct subset model and is sign consistent. In other words:

$$P\left(\widehat{\Pi}_{nj} =_s \Pi_{nj}\right) \to 1$$

**Proposition 4.** Under assumptions 1 to 11 and Proposition 1 the nonparametric adaptive LASSO identifies the correct subset model and is sign consistent. In other words:

$$P\left(\widehat{\Pi}_{nj} =_s \Pi_{nj}\right) \to 1$$

Proposition 3 states that, asymptotically, the adaptive LASSO will exclude the coefficients that are exactly zero and will provide the correct sign for the included coefficients. This feature is beneficial when the set of instruments at hand is numerous and many are potentially irrelevant, which is the framework studied in this paper. As I will show in section 1.3, being able to exclude the irrelevant coefficients is important to achieve the properties of the estimator I propose.

The proof of Proposition 3 follows directly from Huang et al. (2007) once it is shown that the assumptions in this paper are equivalent to the assumptions in their paper, as was discussed in section 1.2.3. Similarly, for the nonparametric estimator the result follows by virtue of Proposition 1 and noting that the assumptions in Huang et al. (2010) are satisfied in the context of the assumptions of this paper. The modifications to their proof are essentially the same as those presented in the proof of Proposition 1.

## 1.3. Consistency of 2SLS Using Adaptive LASSO

In this section, the conditions under which the procedure proposed is consistent are determined. The taxonomy of Chao and Swanson (2003b) is used to express the result for different degrees of instrument weakness.

Chao and Swanson (2003b) define consistency in terms of the growth of the concentration parameter. Specifically, consistency can be achieved in their framework if:

$$\frac{\widehat{Y}_{2n}' u_n}{r_n} \to 0 \tag{1.18}$$

This definition of consistency is a consequence of the fact that under weak instruments $E\left(\widehat{Y}_{2n}'u_n\right) \neq 0$. This implies that a stronger condition than $\frac{\widehat{Y}_{2n}'u_n}{n} \to 0$ is required to achieve consistency, i.e, $\frac{\widehat{Y}_{2n}'u_n}{r_n} \to 0$. This is a stronger statement because $r_n$ grows at a slower rate than $n$ under weak instruments.

In Chao and Swanson (2003b) 2SLS does not satisfy (1.18) and is inconsistent because $\frac{p_n}{r_n} \to \infty$. However, even if $\frac{p_n}{r_n} \to \infty$, the 2SLS estimator proposed in this paper can achieve consistency. Intuitively, by disregarding the useless instruments the remaining ones can achieve the condition $\frac{\hat{p}_{n1}}{r_n} = \frac{p_{n1}}{r_n} + o_p(1) \to 0$. This generates the possibility that even when the estimator is inconsistent using all of the instruments available, it will be consistent when using the reduced instrument set.

Eliminating the irrelevant instruments is important to achieve a greater signal per instrument. More importantly, it conveys the message that even though the solution to the endogeneity problem suggests including all sources of exogenous variation and the work of Chao and Swanson (2003b, 2005b) calls for the introduction of all available instruments, having irrelevant instruments affects the precision of the estimator of the structural parameter of interest.

Using Propositions 2 and 3 the conditions under which relationship (1.18) is satisfied can be found. The proof of the result is presented in the appendix. It states that:

**Theorem 1.** If assumptions 1 to 7 and regularity condition 1 are satisfied then:

$$\frac{\widehat{Y}_{2n}' u_n}{r_n} \to 0.$$

The same is true for consistency in the case of the nonparametric model described in section 1.2.2. In particular:

**Theorem 2.** Let $f_0 \equiv \left(\sum_{j=1}^{p_n} f_j\left(Z_{1j}\right), \ldots, \sum_{j=1}^{p_n} f_j\left(Z_{nj}\right)\right)$ and define the collection of estimators of each $f_j\left(Z_{ij}\right)$, i.e $\sum_{t=1}^{s_n} \hat{\Pi}_{jt}\psi_t(Z_{ij})$, by $\widehat{f_n}$. If assumptions 1 to 11 and regularity condition 1 are satisfied, for $\lambda_{nl} \geq C\sqrt{nlog\left(p_n s_n\right)}$ with $C > 0$:

$$\frac{\widehat{f}_n' u_n}{r_{nx}} \to 0.$$

## 1.4. Asymptotic Distribution

One of the advantages of the proposed estimator is that under the assumptions in the text it is asymptotically normal. This allows for reliable inference under many weak instruments.

Before the results are presented I define the following two matrices. For $\epsilon \equiv V_n\beta + \beta V_n' + u_n$:

$$\begin{aligned}\Lambda &= \Phi^{-1\prime}\mathrm{Var}\left(\frac{\Pi_{n1}'Z_{n1}'\epsilon}{\sqrt{r_n}}\right)\Phi^{-1}\\ \Lambda_x &= \Phi_x^{-1\prime}\mathrm{Var}\left(\frac{\Pi_{n1}'X_{n1}'\epsilon}{\sqrt{r_n}}\right)\Phi_x^{-1}\end{aligned}$$

The following theorem provides the asymptotic result:

**Theorem 3.** If assumptions 1 to 7 and regularity condition 1 are satisfied, then:

$$\sqrt{r_n}\left(\hat{\beta}-\beta\right)\rightarrow_d N\left(0,\Lambda\right)$$

It is important to mention that $r_n$ could be asymptotically equivalent to $n$ or slower depending on the growth rate of the concentration parameter. However, $r_n$ is not available to the researcher. The following result addresses this fact.

**Theorem 4.** Under the conditions of Theorem 3:

$$\hat{\Lambda}^{-1/2}\left(\hat{\beta}-\beta\right)\rightarrow_d N\left(0,1\right)$$

In the appendix I establish that $r_n\hat{\Lambda}\rightarrow_p \Lambda$.

From the previous discussion it follows that:

$$\hat{\Lambda}^{-1/2}\left(\hat{\beta}-\beta\right) \rightarrow \Lambda^{-1/2}\sqrt{r_n}\left(\hat{\beta}-\beta\right) \rightarrow_d N\left(0,1\right)$$

In the case of the nonparametric estimator:

**Theorem 5.** If assumptions 1 to 11 and regularity condition 1 are satisfied, for $\lambda \geq C\sqrt{nlog\left(p_n s_n\right)}$ with $C > 0$:

$$\sqrt{r_{nx}}\left(\hat{\beta}-\beta\right) \rightarrow_d N\left(0,\Lambda_x\right)$$

**Theorem 6.** Under the conditions of Theorem 5:

$$\hat{\Lambda}_x^{-1/2}\left(\hat{\beta}-\beta\right) \rightarrow_d N\left(0,1\right)$$

In the appendix I establish that $r_{nx}\hat{\Lambda}_x \rightarrow_p \Lambda$.

## 1.5. Simulation Analysis

The simulation analysis below compares eight different estimators and the post-$l_1$-penalized procedure I introduce. Three of these estimators are more traditional simultaneous equation estimators in the presence of endogeneity: 2SLS, limited information maximum likelihood (LIML), and a mean unbiased Fuller estimator. I selected the three estimators because they are the most commonly used and discussed in the literature. The fourth estimator is the adaptive LASSO without the post selection component. This estimator serves to highlight the importance of introducing a post-selection correction to the adaptive LASSO. The next three estimators are the model averaging versions of the 2SLS, LIML, and mean unbiased Fuller estimators as presented in Kuersteiner and Okui (2010) with unrestricted weight matrices.[12] These estimators can be thought of as including Donald and Newey (2001) as a subset when all the weights are the same. The last estimator is the jackknife instrumental variable estimator (JIVE) proposed by Chao et al. (2010). The authors of Chao et al. (2010) have also suggested other JIVE estimators based on LIML and Fuller. I select the 2SLS JIVE to see the performance of the post-$l_1$-penalized procedure when compared to an estimator that is robust to heteroskedasticity. Also, Chao et al. (2010) claim that under heteroskedasticity the 2SLS JIVE performs similarly to the other jackknife estimators.

The comparison of the estimators is based on three data generating processes. The first data generating process uses random disturbances and instruments that come from a normal distribution. In this setup I allow for the presence of heteroskedasticity. The second data generating process is analogous except for the fact that the design is ho-

[12] Kuersteiner and Okui (2010) suggest the possibility of using different weight matrices for the construction of their model averaging estimators. An unrestricted weight matrix is used because it allows for negative weights, something that Kuersteiner and Okui (2010) suggest is advisable for a situation in which many of the instruments are zero.

moskedastic. In the third data generating process I want to capture the behavior of the estimator when the tails of the random disturbance are thick. I use a homoskedastic design were the instruments and random disturbances come from a t-distribution with 5 degrees of freedom. All the designs incorporate the performance of the estimators under different degrees of endogeneity and instrument weakness. At the same time I test a scenario with a growing instrument set. The specifications yield six different scenarios for each data generating process and sample size. These scenarios are the result of the two levels of endogeneity and three measures of instrument weakness.

I use six criteria to compare the estimators. The first three criteria are the traditional mean square error and the associated bias and variance. This deviates from what is reported in Donald and Newey (2001) and Kuersteiner and Okui (2010) in which the concern for the existence of the moments of the estimators restricts the analysis to measures of central tendency and dispersion. However, considering the mean square error is important for researchers to analyze the potential benefits and caveats of employing the different estimators analyzed under the scenarios presented. The remaining three criteria, the median bias, median absolute deviation, with respect to the true $\beta$, and the interquartile range, are consistent with the work of Donald and Newey (2001) and Kuersteiner and Okui (2010).

### 1.5.1 Simulation Design

The simulation results presented below rely on the construction of a variance-covariance matrix given by:

$$\Sigma_i \equiv \text{Var}\left(u_n, V_n, Z_n\right) = \begin{pmatrix} 1 & \sigma_{uv} & 0 \ldots & 0 \\ \sigma_{uv} & 1 & 0 \ldots & 0 \\ 0 & 0 & & \\ \vdots & \vdots & \mathbf{I}_{p_n \times p_n} & \\ 0 & 0 & & \end{pmatrix} \tag{1.19}$$

In the matrix above, $\sigma_{uv}$ measures the level of endogeneity. In the simulations, two levels of endogeneity will be tested, 0.15 (low endogeneity) and 0.95 (high endogeneity), for sample sizes $n = 100$ and 500. The number of instruments associated with each sample size is 20 and 50 respectively. The results presented are for 1000 replications.

$\Pi_n$ in equation (1.2) will be constructed in two ways. The first one is a modification of Kuersteiner and Okui (2010) and is given by:

$$\begin{aligned} \pi_{nj} &= c\left(p_n\right)\left(1 - \frac{p_n/2 - j}{p_n/2 + 1}\right)^4 \quad \text{for} \quad j \leq p_n/2 \qquad (1.20) \\ \pi_{nj} &= 0 \quad \text{for} \quad j > p_n/2 \end{aligned}$$

In Kuersteiner and Okui (2010), the first $j \leq p_n/2$ are zero and not the last $j > p_n/2$ as presented above. The reason for this change is that the signal of the instruments will be unnecessarily small otherwise. In the expression above the coefficients are constructed for two different pseudo R-squared values, defined by $R_f^2 = \frac{\pi_{nj}'\pi_{nj}}{\pi_{nj}'\pi_{nj}+1} = 0.1,$ and 0.01. The pseudo R-squared is the way in which Kuersteiner and Okui (2010) introduce instrument weakness into their simulation exercise. The term $c\left(p_n\right)$ guarantees that $R_f^2$ keeps its predetermined values.

The second way in which the matrix $\Pi_n$ in equation (1.2) will be constructed follows the criteria of definition 2 for a set of moderately weak instruments. The vector of parameters $\Pi_n$ in this case is given by:

$$\Pi_n = \left(\mathbf{0}_{pn-8}, \frac{1}{\sqrt{n}}, \frac{-1}{\sqrt{n}}, \frac{-1}{\log\left(n\right)}, \frac{1}{\log\left(n\right)}, \frac{1}{\sqrt{\log(n)}}, \frac{-1}{\sqrt{\log(n)}}, \frac{1}{n^{1/3}}, \frac{-1}{n^{1/3}}\right) \tag{1.21}$$

Three data generating process are analyzed. The first one is devised so that both the random components $V_n$ and $u_n$ and the instruments come from a normal distribution with heteroskedastic variance. The second design is analogous except for the fact that the variance is homoskedastic. The final design is one in which the variance is homoskedastic but the draws come from a distribution with a thick tail, a t-distribution with 5 degrees of freedom. Formally, the data generating processes are given by:

**Data Generating Process 1** (DGP 1: $\Sigma_i$ Normal)**.**

$$\begin{aligned}
\tilde{u}_n &= u_n + Z'_{n1} Z_{n1} \\
Y_{2n} &= Z_n \Pi_n + V_n \\
y_{1n} &= 0.5 Y_{2n} + 2 Z_{n1} + \tilde{u}_n
\end{aligned}$$

**Data Generating Process 2** (DGP 2: $\Sigma_i$ Normal)**.**

$$\begin{aligned}
Y_{2n} &= Z_n \Pi_n + V_n \\
y_{1n} &= 0.5 Y_{2n} + 2 Z_{n1} + u_n
\end{aligned}$$

**Data Generating Process 3** (DGP 3: $\Sigma_i$ t-distribution with 5 d.f.)**.**

$$\begin{aligned}
Y_{2n} &= Z_n \Pi_n + V_n \\
y_{1n} &= 0.5 Y_{2n} + 2 Z_{n1} + u_n
\end{aligned}$$

### 1.5.2 DGP 1: Normal with Heteroskedastic Variance Covariance Matrix

In this section I analyze the performance of the post-$l_1$-penalized estimator I construct with respect to the traditional estimators, 2SLS, LIML, and Fuller, the model averaging versions of these estimators, and with respect to the JIVE estimator of Chao et al. (2010) for DGP 1. The full results of the simulations can be found at the end of the document in tables 1.10, 1.11, and 1.12. Tables 1.1, 1.2, and 1.3 summarize the results in a way that facilitates the understanding of how the estimators are behaving.

In tables 1.1, 1.2, and 1.3, if the post-$l_1$-penalized estimator is ranked as 1 it is the best along the dimension analyzed, for instance bias. If the value that appears on the table is 2a it means its performance is the second best after 2SLS, 2b denotes it comes after LIML and 2c, after Fuller, 2d after the adaptive LASSO, and 2j after JIVE. The other numbers follow a similar logic. The magnitudes that generate these rankings can be found in tables 1.10 to 1.12.

Table 1.1 shows the performance of the post-$l_1$-penalized estimator with respect to the traditional estimators. The results suggest that in general the post-$l_1$-penalized estimator has a lower variance than LIML and Fuller which under most the scenarios leads to a smaller means squared error (MSE). This is also the case for the robust measure of dispersion, the interquartile range (IQR), which is lower than that of LIML and Fuller. However, LIML and Fuller tend to have a smaller bias with the notable exception of the scenario in which the instruments are the weakest. When endogeneity is low the post-$l_1$-penalized estimator tends to be better centered around the true value according to the mean absolute deviation (MAD) criterion. When the endogeneity is high the LIML and Fuller are better centered around the true value in terms of MAD except for the case when instrument weakness is high. Therefore, when the instrument weakness is high, regardless of the level of endogeneity, the post-$l_1$-penalized estimator has a smaller MAD. With regard to 2SLS, the proposed estimator tends to exhibit a lower MSE and always has a smaller bias, median bias, and MAD. Nonetheless, the variance and IQR are higher than that of 2SLS.

Table 1.2 compares the post-$l_1$-penalized estimator to the model averaging estimators proposed by Kuersteiner and Okui (2010). The results suggest that the proposed estimator tends to have a smaller bias, median bias, MAD, and MSE, than the model averaging estimators. However, these estimators have a smaller variance than the post-$l_1$-penalized estimator.

The comparison with the JIVE estimator and the adaptive LASSO appears in Table 1.3. The conclusion is that when instrument weakness is the greatest the post-$l_1$-penalized estimator is superior in every respect. Also, the proposed estimator always has a smaller variance and IQR. When instruments are not as weak the JIVE estimator tends to exhibit a smaller bias, and median bias. When the instruments have low endogeneity the post-$l_1$-penalized estimator tends to have a smaller MAD, something that does not occur when endogeneity is high.

From the previous discussion one can conclude that the post-$l_1$-penalized estimator is more efficient than LIML, Fuller and JIVE. In the case of LIML and JIVE, as is presented in tables 1.10 to 1.12, the estimators do not seem to have a second moment. In terms of the bias these three estimators tend to perform better than the post-$l_1$-penalized estimator. With regard to the MAD the post-$l_1$-penalized estimator is a good option when instrument weakness is severe. When this is not the case LIML, Fuller, and JIVE tend to do better. Furthermore, the post-$l_1$-penalized estimator is better centered around the true value than 2SLS but has a greater variability. With respect to the model averaging estimators the post-$l_1$-penalized estimator has a lower bias and is better centered around the true value but has greater variability. Finally, the post-$l_1$-penalized estimator performs unambiguously better than the adaptive LASSO.

### 1.5.3 DGP 2: Normal with Homoskedastic Variance Covariance Matrix

In this section I analyze the performance of the post-$l_1$-penalized estimator I construct with respect to the traditional estimators, 2SLS, LIML, and Fuller, the model averaging versions of these estimators, and with respect to the JIVE estimator of Chao et al. (2010) for DGP 2. The full results of the simulations can be found at the end of the document in tables 1.13, 1.14, and 1.15. Tables 1.4, 1.5, and 1.6 summarize the results in a way that facilitates the understanding of how the estimators are behaving.

In tables 1.4, 1.5, and 1.6, if the post-$l_1$-penalized estimator is ranked as 1 it is the best along the dimension analyzed, for instance bias. If the value that appears on the table is 2a it means its performance is the second best after 2SLS, 2b denotes it comes after LIML and 2c, after Fuller, 2d after the adaptive LASSO, and 2j after JIVE. The other numbers follow a similar logic. The magnitudes that generate these rankings can be found in tables 1.13 to 1.15.

Table 1.4 suggests that the LIML and Fuller estimators tend to be less biased and better centered than the post-$l_1$-penalized estimator. This conclusion does not hold when instruments are the weakest and endogeneity is low, in which case only LIML is superior. Once again, LIML and Fuller tend to have a higher variance and IQR than the post-$l_1$-penalized estimator. This is not true, however, when endogeneity is high and instruments are completely weak, when the Fuller estimator tends to have a smaller variance yet still has a higher IQR. With regard to 2SLS the post-$l_1$-penalized estimator has a smaller bias, median bias, and MAD and has an inferior performance with respect to its variance and IQR.

The results of Table 1.5 show that the post-$l_1$-penalized estimator has a lower bias and median bias, than the modeling average estimators proposed by Kuersteiner and Okui (2010) when endogeneity is low. This is the same when endogeneity is high except in one of the six cases analyzed. Also, the post-$l_1$-penalized estimator tends to have a lower MAD and MSE. With respect to the variance and IQR the modeling average estimators always outperforms the post-$l_1$-penalized estimator.

The comparison with the JIVE estimator and the adaptive LASSO appears in Table 1.6. When the instruments are the weakest, the post-$l_1$-penalized estimator has a smaller bias, median bias, MAD, and variance. However, when instruments are not as weak JIVE tends to be have smaller bias, median bias, and MAD but a larger variance and IQR.

From the previous discussion one can conclude that the post-$l_1$-penalized estimator is more efficient than LIML, Fuller and JIVE. The LIML and JIVE estimators, as is presented in tables 1.13 to 1.15, do not seem to have a second moment. In terms of the bias these three estimators tend to perform better than the post-$l_1$-penalized estimator and, LIML and Fuller outperform JIVE. With regard to the MAD the post-$l_1$-penalized estimator is a good option when instrument weakness is severe. When this is not the case LIML, Fuller, and JIVE tend to do better. With respect to the model averaging estimators the post-$l_1$-penalized estimator has a lower bias and is better centered around the true value but has greater variability. Furthermore, the post-$l_1$-penalized estimator is better centered around the true value than 2SLS but has a greater variability. Finally, the post-$l_1$-penalized estimator is unambiguously better than the adaptive LASSO.

### 1.5.4 DGP 3: t-distribution with Homoskedastic Variance Covariance Matrix

In this section I analyze the performance of the post-$l_1$-penalized estimator I construct with respect to the traditional estimators, 2SLS, LIML, and Fuller, the model averaging versions of these estimators, and with respect to the JIVE estimator of Chao et al. (2010) for DGP 3. The full results of the simulations can be found at the end of the document in tables 1.16, 1.17, and 1.18. Tables 1.7, 1.8, and 1.9 summarize the results in a way that permits the understanding of how the estimators are behaving.

In tables 1.7, 1.8, and 1.9, if the post-$l_1$-penalized estimator is ranked as 1 it is the best along the dimension analyzed, for instance bias. If the value that appears on the table is 2a it means its performance is the second best after 2SLS, 2b denotes it comes after LIML and 2c, after Fuller, 2d after the adaptive LASSO, and 2j after JIVE. The other numbers follow a similar logic. The magnitudes that generate these rankings can be found in tables 1.16 to 1.18.

Table 1.7 suggests that 2SLS tends to have a higher bias than LIML, Fuller, and the post-$l_1$-penalized estimator when endogeneity is high but has a smaller MAD and MSE when instruments have low endogeneity. Once again, 2SLS has the smallest variance and IQR with respect to the traditional estimators and the post-$l_1$-penalized estimator. Table 1.7 also suggests that the LIML and Fuller estimators tend to be less biased and better centered than the post-$l_1$-penalized estimator according to MAD and both measures of bias. As in the previous DGPs LIML and Fuller have a higher variance and IQR than the post-$l_1$-penalized estimator.

The results of Table 1.7 show that the post-$l_1$-penalized estimator has a lower bias and median bias, unambiguously, than the modeling average estimators proposed by Kuersteiner and Okui (2010) when endogeneity is low. This is the same when endogeneity is high except in one of the six cases analyzed. Also, the post-$l_1$-penalized estimator tends to have a lower MAD except when instruments are moderately weak. The MSE that in the previous two DGPs had been lower for the post-$l_1$-penalized estimator is now unambiguously inferior when instruments have low endogeneity. When endogeneity is high MSE still tends to be lower for the proposed estimator. With respect to the variance and IQR the modeling average estimators always outperforms the proposed estimator.

The comparison with the JIVE estimator and the adaptive LASSO appears in Table 1.9. When the instruments are the weakest the post-$l_1$-penalized estimator has a smaller bias, median bias, MAD, and variance. However, when instruments are not as weak JIVE tends to be have smaller bias, median bias, and MAD but a larger variance and IQR.

From the previous discussion the conclusion is that the post-$l_1$-penalized estimator is more efficient than LIML, Fuller and JIVE. In the case of LIML and JIVE, tables 1.16 to 1.18

suggest the estimators do not have a second moment. In terms of the bias these three estimators tend to perform better than the post-$l_1$-penalized estimator and, LIML and Fuller outperform JIVE. With regards to the MAD this also holds except when endogeneity of the instruments is low in which case 2SLS is better.With respect to the model averaging estimators the post-$l_1$-penalized estimator has a lower bias and is better centered around the true value but has greater variability. Finally, the post-$l_1$-penalized estimator is unambiguously better than the adaptive LASSO.

## 1.6. Conclusion

In this paper, I consider a setting with many weak, possibly irrelevant instruments, where traditional 2SLS leads to inconsistent estimates when all potential instruments are used, but it is unclear how to select the right instruments to include. I introduce a new 2SLS estimator that selects the correct instruments with high probability, and is simultaneously consistent and asymptotically normal.

I exploit the idea that signal per instrument is what determines the quality of the instrument set and provide a methodology to attain a higher signal per instrument. I achieve this objective using the adaptive LASSO, a selection procedure that excludes the irrelevant instruments with probability approaching one. The result is the possibility that even though the initial signal per instrument stays constant as the sample size increases, after instrument selection the signal per instrument diverges asymptotically. This implies that estimators that were inconsistent before instrument selection, because the signal per instrument remained constant asymptotically, become consistent and asymptotically normal. I derive these results bringing together the many weak instrument asymptotic framework of Chao and Swanson (2003b, 2005b) and the adaptive LASSO asymptotics of Huang et al. (2007).

The results I obtain are robust to heteroskedasticity which, as Chao et al. (2010) and Hausman et al. (2010) prove, is an important consideration because it affects not only the properties of confidence intervals but the consistency of the estimator of the structural parameter. In addition, by allowing for the assumption that the instruments are mean independent of the structural random disturbance, I allow the possibility of a nonparametric adaptive LASSO in the first stage.

Finally, simulation results show that the proposed estimator performs better than the estimators discussed in section 1.5 under heteroskedasticity, when the instruments are the weakest and endogeneity is high. This suggests that the estimator is fulfilling the description of the theoretical results and that it is an important alternative for researchers that suspect high levels of endogeneity, a set of considerably weak instruments, and the possibility of the presence of irrelevant instruments in their specification.

# 1.7. Appendix

All proofs of the results in the text are presented in this section. Below $C$ will be used generically to refer to a constant.

## 1.7.1 Proof of Propositions

In this subsection I present a proof of Propositions 1 and 2. Both results are important to establish the existence of the asymptotic order of the minimum non-zero coefficient selected by the adaptive LASSO.

**Proposition 2**

*Proof.* (Proposition 2) First notice that from the Kuhn-Tucker conditions, $\widehat{\Pi}_n = \left(\widehat{\Pi}_{n1} \ldots \widehat{\Pi}_{nn}\right)'$ is the unique solution to the adaptive LASSO if:

$$\begin{aligned} z_j'\left(Y_2 - Z_n\widehat{\Pi}_n\right) &= \lambda_n w_{nj} sgn\left(\widehat{\Pi}_{nj}\right), \qquad \widehat{\Pi}_{nj} \neq 0 \\ \left|z_j'\left(Y_2 - Z_n\widehat{\Pi}_n\right)\right| &< \lambda_n w_{nj}, \qquad \widehat{\Pi}_{nj} = 0 \end{aligned} \tag{1.22}$$

The proof below follows from verifying that the assumptions in Theorem 1 and Theorem 2 of Zhang and Huang (2008) are satisfied using $\lambda* \equiv \lambda_n w_{nj}$. The reason for this is that they use the LASSO, i.e $w_{nj} = 1\, \forall j$, instead of the adaptive LASSO. Also, they assume $V_n$ is normally distributed. Here instead of having normal residuals, assumption 5 is used to attain the results for subgaussian tails. As was explained in the text this allows us to obtain a conservative lower bound for the minimum of the excluded coefficients.

Let us define:

$$\zeta_{\alpha_1} \equiv \left( \sum_{j \in \mathcal{A}} \left| \Pi_{nj} \right|^{\alpha_1} I \left\{ \hat{\Pi}_{nj} = 0 \right\} \right)^{1/\alpha_1} \quad \alpha_1 \in [0, \infty] \tag{1.23}$$

$$B(\lambda) \equiv \left\| Z_n \Pi_n - Z_n \hat{\Pi}_n \right\| \tag{1.24}$$

$$\eta_2 \equiv \max_{\mathcal{A}' \subset \mathcal{A}^c} \left\| \sum_{j \in \mathcal{A}'} Z_{nj} \Pi_{nj} \right\| \tag{1.25}$$

Also, for $m \leq p_n$, $I$ denoting an identity matrix, $Z_{nY}$ denoting a matrix of instruments of dimensions $n \times Y$, and $P_Y \equiv Z_{nY} \left( Z'_{nY} Z_{nY} \right)^{-1} Z'_{nY}$:

$$\chi^*_m \equiv \max_{|A|=m} \max_{s \in \{\pm 1\}^m} \left| V'_n \frac{Z_{nA} \left( Z'_{nA} Z_{nA} \right)^{-1} s\lambda_* - (I - P_A) Z_n \Pi_n}{\left\| Z_{nA} \left( Z'_{nA} Z_{nA} \right)^{-1} s\lambda_* - (I - P_A) Z_n \Pi_n \right\|} \right| \tag{1.26}$$

Finally for $c_0 \geq 0$ and $a_n \geq 0$ satisfying $p_n / \left( p_n \vee a_n \right)^{1+c_0} \to 0$:

$$\Omega_0 \equiv \left\{ (Z_n, V_n) : \chi_m^* \leq \sqrt{2\,(1 + c_0)\,(m \vee 1)\,log\,(p_n \vee a_n)} \right\} \tag{1.27}$$

Above (1.23) is a measure of the number of coefficients excluded by the model that should have been included, $B(\lambda)$ in (1.24) is a measure of the model's bias, and (1.25) is an upper bound on the contribution of a given set of the excluded regressors.

If $(Z_n, V_n) \in \Omega_0$ and the sparse Riesz condition of Zhang and Huang (2008) is fulfilled the result of Proposition 2 follows. Expression (1.17) is the definition of sparsity in Zhang and Huang (2008). It tells us that there are a maximum of $p_{n1}$ large coefficients and that the sum of the $l_1$ norms of the small coefficients is no greater than $\eta_1$. The estimator of the weights guarantees that this condition is satisfied. Also, assumption 2 gives us the Riesz condition of Zhang and Huang (2008). It states that the smallest eigenvalue of $\left(\frac{Z_n' Z_n}{m_{n1}}\right)$ is bounded away from zero and the largest one is finite. These two conditions imply the fulfillment of the sparse Riesz condition in Zhang and Huang (2008).

What remains to be shown is that $(Z_n, V_n) \in \Omega_0$. This in turn implies, denoting the set of selected instruments by $\hat{\mathcal{A}}$, that:

$$P\left\{\mathcal{A} \subset \hat{\mathcal{A}}, \quad B(\lambda) \leq \eta_2 \quad \text{and} \quad \zeta_{\alpha_1} = 0\right\} \to 1$$

For $V_n$ in assumption 5 with $d = 2$, $(Z_n, V_n) \in \Omega_0$ is given by:

$$\begin{aligned}
1 - P\left\{(Z_n, V_n) \in \Omega_0\right\} &\leq \sum_{m=0}^{\infty} 2^{m \vee 1} \binom{p_n}{m} D_6 exp\left\{-D_5 (m \vee 1)(1 + c_0) log (p \vee a_n)\right\} \\
&\leq \frac{2D_6}{(p_n \vee a_n)^{(1+c_0)D_5}} + D_6 exp\left(\frac{2p_n}{(p_n \vee a_n)^{(1+c_0)D_5}}\right) - D_6 \\
&\rightarrow 0
\end{aligned}$$

Regularity condition 1 guarantees that the final expression goes to zero.

By Theorem 1 and 2 of Zhang and Huang (2008) it follows that $(Z_n, V_n) \in \Omega_0$. This implies that the excluded instruments are of the order $O\left(\frac{\lambda * \sqrt{p_{n1}}}{m_{n2}}\right)$. Also, given that $w_{nj} = O\left(\sqrt{m_{n2}}\right)$, the excluded instruments are of the order $O\left(\frac{\lambda * \sqrt{p_{n1}}}{m_{n2}}\right)$ translates to the excludes instruments being of the order:

$$O\left(\frac{\lambda_n \sqrt{p_{n1}}}{\sqrt{m_{n2}}}\right) \tag{1.28}$$

□

**Proposition 1**

*Proof.* Proposition 1 is a restatement of Theorem 1 of Huang et al. (2010) with two differences. The first difference is that the tail behavior of the errors in the first stage are not assumed to be subgaussian. The second one is that the eigenvalues of the information matrix of the selected basis functions $X'_{n1}X_{n1}$ is determined by $m_{nx2}$ in assumption 2. This fact is equivalent to Lemma 3 in their paper being fulfilled. The proof below will first address part *ii*) of Proposition 1 and then part *i*).

Below the Lemmas used to prove Theorem 1 in Huang et al. (2010) will be modified to incorporate these differences. As was mentioned previously, Lemma 3 in their paper follows from assumption 2. Also, Lemma 1m which comes directly from the theory of regression splines, is not modified.

The first component is to modify Lemma 2 of Huang et al. (2010) to allow for the tail behavior of $V_n$ to range from exponential to subgaussian. Let us define $T_{jk} \equiv \frac{\sqrt{s_n}}{\sqrt{n}} \sum_{i=1}^{n} \psi_k \left(Z_{ij}\right) v_i$ for $1 \leq j \leq p_n, 1 \leq k \leq s_n$ and $T_n \equiv \max_{1\leq j\leq p_n, 1\leq k\leq s_n} \left|T_{jk}\right|$. Also, define $t_{njk}^2 \equiv \sum_{i=1}^{n} \psi_k^2 \left(Z_{ij}\right)$ and $t_n^2 \equiv \max_{1\leq j\leq p_n, 1\leq k\leq s_n} t_{njk}^2$. The first detail to notice is that conditional on the $Z_{ij}$'s the $T_{jk}$'s behave according to the exponential Orlicz norm given assumption 5. This implies that by van der Vaart and Wellner (1996):

$$E\left(\max_{1\leq j\leq p_n, 1\leq k\leq s_n} \left|T_{jk}\right| \mid \left\{Z_{ij}, 1 \leq i \leq n, 1 \leq j \leq p_n\right\}\right) \leq C\frac{\sqrt{s_n}}{\sqrt{n}} t_n \left(log\left(p_n s_n\right)\right)^{1/d} \quad \text{for} \quad d \in [1,2]$$

By the law of iterated expectations the expression above becomes:

$$E\left(\max_{1\leq j\leq p_n,1\leq k\leq s_n}\left|T_{jk}\right|\right)\leq C\frac{\sqrt{s_n}}{\sqrt{n}}\left(log\left(p_ns_n\right)\right)^{1/d}E\left(t_n\right)\quad\text{for}\quad d\in[1,2]$$

The remaining arguments in Huang et al. (2010) remain unmodified. The conclusion of Lemma 2 in their text then becomes:

$$E\left(T_n\right)\leq C\frac{\sqrt{s_n}}{\sqrt{n}}\left(log\left(p_ns_n\right)\right)^{1/d}\left(\sqrt{2C\frac{n}{s_n}log(p_ns_n)}+4log(2p_ns_n)+C\frac{n}{s_n}\right)^{1/2}\quad\text{for}\quad d\in[1,2]$$

It remains now to restate the elements of the proof of Theorem 1 in Huang et al. (2010) that have been modified by the changes introduced to the Lemmas in their text.

Lets first define $\varsigma\equiv Y_{2n}-\sum_{j=1}^{p_n}X_{nj}\Pi_{nj}$ and $\varsigma*\equiv X_{n1}\left(X'_{n1}X_{n1}\right)^{-1}X'_{n1}\varsigma$. The expression for $\varsigma$ can be rewritten to incorporate the fact that $\sum_{j=1}^{p_n}X_{nj}\Pi_{nj}$ is an approximation to $\sum_{j=1}^{p_n}f_j\left(Z_{nj}\right)$ by noting that:

$$\begin{aligned}\varsigma_i &= Y_{2i}-\sum_{j=1}^{p_n}X_{nj}\Pi_{nj}-\sum_{j=1}^{p_n}f_j\left(Z_{nj}\right)+\sum_{j=1}^{p_n}f_j\left(Z_{nj}\right)\\ &= v_i+\sum_{j=1}^{p_n}f_j\left(Z_{nj}\right)-\sum_{j=1}^{p_n}X_{nj}\Pi_{nj}\end{aligned}$$

By the fact that $\left\|\sum_{j=1}^{p_n}f_j\left(Z_{nj}\right)-\sum_{j=1}^{p_n}X_{nj}\Pi_{nj}\right\|_\infty=O\left(s_n^{-\partial}+\sqrt{\frac{s_n}{n}}\right)$, which comes from results of Lemma 1 of Huang et al. (2010), where $\partial$ is defined in assumption 8, it follows that:

$$\|\varsigma*\|_2^2 \leq 2\,\|v_i*\|_2^2 + O\left(nq\left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^2\right)$$

Where $v_i* \equiv X_{n1}\left(X_{n1}'X_{n1}\right)^{-1}X_{n1}'v_i$. Furthermore, by assumption 2 it follows that $nc_n*$, which is the equivalent to the growth rate of the information matrix in Huang et al. (2010), can be replaced by $m_{nx2}$, and:

$$\begin{aligned}
\|v_i*\|_2^2 &= \left\|X_{n1}\left(X_{n1}'X_{n1}\right)^{-1}X_{n1}'v_i\right\|_2^2 \\
&= \frac{C}{m_{nx2}}\left\|X_{n1}'v_i\right\|_2^2
\end{aligned}$$

It follows from the arguments above and those of Theorem 1 in Huang et al. (2010) that:

$$\|v_i*\|_2^2 = O_p(1)\frac{O_p(q)nlog\left(p_n s_n\right)}{m_{nx2}}$$

Again using assumption 2 paired with Theorem 1 of Huang et al. (2010):

$$\left\|\tilde{\Pi}_{n1} - \Pi_{n1}\right\|_2^2 \leq \frac{C\left\|\varsigma*\right\|_2^2}{m_{nx}} + \frac{C\lambda_{nl}^2 O_p(q)}{m_{nx2}^2}$$

It follows from the arguments above that:

$$\left\|\tilde{\Pi}_{n1} - \Pi_{n1}\right\|_2^2 \leq O_p\left(\frac{qnlog\left(p_n s_n\right)}{m_{nx}^2}\right) + O\left(\frac{nq}{m_{nx2}}\left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^2\right) + O\left(\frac{\lambda_{nl}^2 q}{m_{nx2}^2}\right)$$

By virtue of assumption 7:

$$\frac{qnlog(p_n s_n)}{m_{nx2}^2} \to 0, \quad \frac{nq}{m_{nx2}}\left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^2 \to 0 \quad \text{and} \quad \frac{\lambda_{nl}^2 q}{m_{nx2}^2} \to 0$$

The above relationships and the arguments found in Huang et al. (2010) guarantee that part (i) of Proposition 1 is satisfied. This completes the proof of the Proposition.

□

### 1.7.2 Theorem 1

#### Lemmas

The first lemma presented below allows the use of the true set of instruments in the proofs, as opposed to those selected by the adaptive LASSO. This is going to simplify the notation and it also facilitates the analysis of the implications of incorporating the adaptive LASSO into a two stage procedure.

**Lemma 1.** Defining $\hat{p}_{n1}$ as the number of selected instrument, if the conditions of Proposition 2 are satisfied by assumption 7(b) and Proposition 3:

$$\frac{\hat{p}_{n1}}{r_n} \to_p 0$$

*Proof.* By Proposition 3, $\frac{\hat{p}_{n1}}{r_n} = \frac{p_{n1}}{r_n} + o_p(1)$. Replacing $r_n$ in the expression $\frac{p_{n1}}{r_n}$ by its definition and using Proposition 2:

$$\frac{p_{n1}}{r_n} = \frac{p_{n1}b_n^2}{m_{n2}h_n} \leq C\frac{p_{n1}m_{n2}^2}{m_{n2}^2\lambda_n^2 p_{n1}} = C\frac{1}{\lambda_n^2} \to 0 \tag{1.29}$$

□

It is important to highlight that division by $\frac{b_n^2}{h_n}$, the asymptotic order of the first stage coefficient vector, is bounded above by Proposition 2. In other words dividing by $\frac{b_n^2}{h_n}$ necessarily yields a smaller number than dividing by the asymptotic order of the smallest of the non-zero coefficients. This yields the first inequality and is the key argument to prove lemma 1. The conclusion follows by Assumption 7(b).

The first lemma presented below allows the use of the true set of instruments in the proofs, as opposed to those selected by the adaptive LASSO. This is going to simplify the notation and it also facilitates the analysis of the implications of incorporating the adaptive LASSO into a two stage procedure.

Let us define $\hat{\beta}$ as the estimator proposed in this paper using the instruments selected by the adaptive LASSO, $\tilde{Z}_{n1}$, and $\hat{\beta}_{\text{oracle}}$ as the unfeasible estimator that could be computed if the true set of relevant instruments, $Z_{n1}$, were known. It follows that:

$$\begin{aligned}
\sqrt{r_n}\left(\hat{\beta}-\beta\right) &= \sqrt{r_n}\left(\hat{\beta}-\beta\right) I\left\{Z_{n1}=\tilde{Z}_{n1}\right\}+\sqrt{r_n}\left(\hat{\beta}-\beta\right) I\left\{Z_{n1}\neq\tilde{Z}_{n1}\right\} \\
&= \sqrt{r_n}\left(\hat{\beta}_{\text{oracle}}-\beta\right)+\sqrt{r_n}\left(\hat{\beta}-\beta\right) I\left\{Z_{n1}\neq\tilde{Z}_{n1}\right\}
\end{aligned}$$

From Theorem 1 of Huang et al. (2007) and assumptions 5 and 7, for $\mathcal{A} = \left\{j : \Pi_{nj} \neq 0\right\}$, the sign consistency of the adaptive LASSO implies that:

$$\begin{aligned}
1-P\left(\left\{j:\hat{\Pi}_{nj}\neq 0\right\}=\mathcal{A}\right) &= O_p\left(log\left(n\right)^{I\{d=1\}}\left(\frac{log(p_{n2})^{1/d}}{\lambda_n}\right)\right) \\
&= o_p(1)
\end{aligned} \tag{1.30}$$

The second equality is a consequence of assumption 7. In the expression above $d$ represents the tail behavior of the random disturbances as described in assumption 5.

Assumption 7(b) and equation (1.30) guarantee that:

$$\begin{aligned}\sqrt{r_n} I\left\{Z_{n1} \neq \tilde{Z}_{n1}\right\} &= O_p\left(log\left(n\right)^{I\{d=1\}}\left(\frac{log(p_{n2})^{1/d}\sqrt{r_n}}{\lambda_n}\right)\right) \\ &= o_p\left(1\right)\end{aligned}$$

If I can show that for any random set of instruments $\left(\hat{\beta}-\beta\right) = O_p\left(1\right)$ then it will follow that:

$$\sqrt{r_n}\left(\hat{\beta}-\beta\right) = \sqrt{r_n}\left(\hat{\beta}_{oracle}-\beta\right) + o_p\left(1\right)$$

This is what is proved in Lemma 2.

**Lemma 2.** For any random set of instruments $Z_{nk}$:

$$\left(\hat{\beta}-\beta\right) = O_p\left(1\right)$$

*Proof.* Let us define $P_z = Z_n\left(Z_n' Z_n\right)^{-1} Z_n$ as the projection matrix using all the instruments, $P_{z1} = Z_{n1}\left(Z_{n1}' Z_{n1}\right)^{-1} Z_{n1}$ as the projection matrix using the correct set of instruments, and $P_{zk} = Z_{nk}\left(Z_{nk}' Z_{nk}\right)^{-1} Z_{nk}'$ as the projection matrix associated with a random set of $k$ instruments.

The analysis will focus on the estimator that would arise for any random set of instruments:

$$\hat{\beta} - \beta = \left(Y_{2n}' P_{zk} Y_{2n}\right)^{-1} \left(Y_{2n}' P_{zk} u_n\right)$$

I will first focus on the denominator. The first step is to rewrite the denominator as:

$$\begin{aligned}
Y_{2n}' P_{zk} Y_{2n} &= \left(Z_n \Pi_n + V_n\right)' P_{zk} \left(Z_n \Pi_n + V_n\right) \\
&= \Pi_n' Z_n' P_{zk} Z_n \Pi_n + V_n' P_{zk} Z_n \Pi_n + \Pi_n' Z_n' P_{zk} V_n + V_n' P_{zk} V_n
\end{aligned}$$

To analyze the first term above, let $z_{ki}$ be the $i^{th}$ row of $Z_{nk}$ and $z_i$ be the $i^{th}$ rows of $Z_n$. Also, I define $w_i \equiv \left(Z_{nk}' Z_{nk}\right)^{-1/2} z_{ki}$:

$$\begin{aligned}
& E \left\| \Pi_n' Z_n' P_{zk} Z_n \Pi_n \right\|^2 \\
= \; & E \left( \sum_{i=1}^{n} \sum_{i=1}^{n} \Pi_n' z_i z_{ki}' \left(Z_{nk}' Z_{nk}\right)^{-1} z_{kj} z_j' \Pi_n \right)^2 \\
= \; & E \left( \sum_{i=1}^{n} \sum_{i=1}^{n} w_i' w_j z_j' \Pi_n \Pi_n' z_i \right)^2 \\
= \; & E \left( \sum_{i=1}^{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \sum_{l=1}^{n} w_i' w_j w_k' w_l z_j' \Pi_n \Pi_n' z_i z_l' \Pi_n \Pi_n' z_k \right) \\
= \; & E \left( \sum_{i=1}^{n} \left(w_i' w_i\right)^2 \left(z_i' \Pi_n \Pi_n' z_i\right)^2 + 4 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(w_i' w_j\right)^2 \left(z_i' \Pi_n \Pi_n' z_j\right)^2 \right) \\
+ \; & E \left( 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(z_i' \Pi_n \Pi_n' z_i\right) \left(z_j' \Pi_n \Pi_n' z_j\right) \left[ \left(w_i' w_i\right) \left(w_j' w_j\right) + 2 \left(w_i' w_j\right)^2 \right] \right) \\
+ \; & E \left( 4 \sum_{i=2}^{n} \sum_{j=1}^{i-1} \left(w_i' w_i\right) \left(w_i' w_j\right) \left(z_i' \Pi_n \Pi_n' z_i\right) \left(z_i' \Pi_n \Pi_n' z_j\right) \right)
\end{aligned}$$

$$
\begin{aligned}
\leq\;& E\left(\sum_{i=1}^{n}\left(w_i'w_i\right)^2\left(z_i'\Pi_n\Pi_n'z_i\right)^2+4\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(w_i'w_i\right)\left(w_j'w_j\right)\left(z_i'\Pi_n\Pi_n'z_j\right)^2\right)\\
+\;& E\left(2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(z_i'\Pi_n\Pi_n'z_i\right)\left(z_j'\Pi_n\Pi_n'z_j\right)\left[\left(w_i'w_i\right)\left(w_j'w_j\right)+2\left(w_i'w_i\right)\left(w_j'w_j\right)\right]\right)\\
+\;& E\left(4\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(w_i'w_i\right)\sqrt{\left(w_i'w_i\right)\left(w_j'w_j\right)}\left(z_i'\Pi_n\Pi_n'z_i\right)\left(z_i'\Pi_n\Pi_n'z_j\right)\right)\\
\leq\;& E\left(\left(\max_{1\leq i\leq n}\left|z_{ki}'\left(Z_{nk}'Z_{nk}\right)^{-1}z_{ki}\right|\right)^2\left[\sum_{i=1}^{n}\left(z_i'\Pi_n\Pi_n'z_i\right)^2+2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(z_i'\Pi_n\Pi_n'z_j\right)^2\right]\right)\\
+\;& E\left(3\left(\max_{1\leq i\leq n}\left|z_{ki}'\left(Z_{nk}'Z_{nk}\right)^{-1}z_{ki}\right|\right)^2\left[2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(z_i'\Pi_n\Pi_n'z_i\right)\left(z_j'\Pi_n\Pi_n'z_j\right)\right]\right)\\
+\;& E\left(\left(\max_{1\leq i\leq n}\left|z_{ki}'\left(Z_{nk}'Z_{nk}\right)^{-1}z_{ki}\right|\right)^2\left[2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(z_i'\Pi_n\Pi_n'z_i\right)\left(z_i'\Pi_n\Pi_n'z_j\right)\right]\right)\\
\leq\;& E\left(C\frac{p_{nk}^2}{m_{n1}^2}\left[\sum_{i=1}^{n}\left(z_i'\Pi_n\Pi_n'z_i\right)^2+4\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(z_i'\Pi_n\Pi_n'z_j\right)^2\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
+\;& E\left(C\frac{p_{nk}^2}{m_{n1}^2}\left[2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(z_i'\Pi_n\Pi_n'z_i\right)\left(z_j'\Pi_n\Pi_n'z_j\right)\right]\right)\\
+\;& E\left(C\frac{p_{nk}^2}{m_{n1}^2}\left[2\sum_{i=2}^{n}\sum_{j=1}^{i-1}\left(z_i'\Pi_n\Pi_n'z_i\right)\left(z_i'\Pi_n\Pi_n'z_j\right)\right]\right)\\
\leq\;& C\frac{E\left(p_{nk}^2\right)r_n^2}{m_{n1}^2}\\
\leq\;& C\frac{E\left(p_{nk}^2\right)r_n^2}{m_{n1}^2}
\end{aligned}
$$

In the expression above the first inequality uses the Cauchy-Schwarz inequality. The second and third inequalities exploit assumption 3. The third inequality follows from assumptions 1, 2, and 6.

For the second term:

$$
\begin{aligned}
& E\left\|V_n' P_{zk} Z_n \Pi_n\right\|^2 \\
= \;& E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} v_i z_{ki}'\left(Z_{nk}'Z_{nk}\right)^{-1} z_{kj} z_i' \Pi_n\right)^2 \\
= \;& E\left(\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} v_i v_k z_j' \Pi_n \Pi_n' z_l w_i' w_j w_k' w_l\right) \\
= \;& E\left(\sum_{i=1}^{n} v_i^2 z_i' \Pi_n \Pi_n' z_i \left(w_i' w_i\right)^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} v_i^2 z_i' \Pi_n \Pi_n' z_i \left(w_j' w_i\right)^2\right) \\
+ \;& E\left(2\sum_{i=2}^{n}\sum_{j=1}^{i-1} v_i^2 z_i' \Pi_n \Pi_n' z_j \left(w_i' w_i\right)\left(w_i' w_j\right) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} v_j^2 z_i' \Pi_n \Pi_n' z_j \left(w_i' w_i\right)\left(w_i' w_j\right)\right) \\
\leq \;& E\left(\sum_{i=1}^{n} v_i^2 z_i' \Pi_n \Pi_n' z_i \left(w_i' w_i\right)^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} v_i^2 z_i' \Pi_n \Pi_n' z_i \left(w_i' w_i\right)\left(w_j' w_j\right)\right) \\
+ \;& E\left(2\sum_{i=2}^{n}\sum_{j=1}^{i-1} v_i^2 z_i' \Pi_n \Pi_n' z_j \left(w_i' w_i\right)\sqrt{\left(w_i' w_i\right)\left(w_j' w_j\right)} + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} v_j^2 z_i' \Pi_n \Pi_n' z_j \left(w_i' w_i\right)\sqrt{\left(w_i' w_i\right)\left(w_j' w_j\right)}\right) \\
\leq \;& E\left(\max_i E\left(v_i^2|z\right)\max_{1\leq i\leq n}\left|z_{ki}'\left(Z_{nk}'Z_{nk}\right)^{-1} z_{ki}\right|^2\left[\sum_{i=1}^{n} z_i' \Pi_n \Pi_n' z_i + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} z_i' \Pi_n \Pi_n' z_i\right]\right)
\end{aligned}
$$

$$
\begin{aligned}
+ \;& E\left(\max_i E\left(v_i^2|z\right)\max_{1\leq i\leq n}\left|z_{ki}'\left(Z_{nk}'Z_{nk}\right)^{-1} z_{ki}\right|^2\left[4\sum_{i=2}^{n}\sum_{j=1}^{i-1} z_i' \Pi_n \Pi_n' z_j\right]\right) \\
\leq \;& C\frac{r_n E\left(p_{nk}\right)^2}{m_{n1}^2}
\end{aligned}
$$

The first inequality above follows from the Cauchy-Schwarz inequality. The second and third inequalities use assumptions 5 and 3.

For the last term, let $p_{ij}$ denote the $(i,j)^{th}$ element of $P_{zk}$ and $v_i$ the $i^{th}$ element of $V_n$

and let $C_1$ be a positive constant:

$$\begin{aligned}
E\left\|V_n'P_{zk}V_n\right\|^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} E\left(p_{ij}p_{kl}v_iv_jv_kv_l\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{l=1}^{n} E\left(p_{ij}p_{kl}\right)E\left(v_iv_jv_kv_l\right) \\
&= \sum_{i=1}^{n} E\left(p_{ii}^2\right)E\left(v_i^4\right) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\left(v_i^2\right)E\left(v_j^2\right)E\left(2p_{ij}^2 + p_{ii}p_{jj}\right) \\
&\leq \max_i E\left(v_i^4\right)\sum_{i=1}^{n} E\left(p_{ii}^2\right) + \max_i E\left(v_i^2\right)E\left(v_j^2\right)2\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\left(2p_{ij}^2 + p_{ii}p_{jj}\right) \\
&\leq \max_i E\left(v_i^4\right)\sum_{i=1}^{n} E\left(p_{ii}\right) + \max_i E\left(v_i^2\right)E\left(v_j^2\right)2\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\left(2p_{ij}^2 + p_{ii}p_{jj}\right) \\
&= \max_i E\left(v_i^4\right)E\left(\text{trace}\left(P_{zk}\right)\right) \\
&+ \max_i E\left(v_i^2\right)E\left(v_j^2\right)\left[E\left(\text{trace}\left(P_{zk}\right)^2\right) + 2\text{trace}\left(E\left(P_{zk}'P_{zk}\right)\right) - 3\sum_{i=1}^{n} E\left(p_{ii}^2\right)\right] \\
&= \left(C_1 + o_p\left(1\right)\right)\left(E\left(p_{nk}\right)^2 + E\left(p_{nk}\right)\right)
\end{aligned}$$

The second equality comes from the independence of $V_n$ and $Z_n$. The third equality comes by virtue of two facts. The first is that the $v_i$s are independent and the second that only when $i = j = k = l$ or when two equalities of the elements in $\{i, j, k, l\}$ hold, do the terms in the summation become different than zero. The first inequality is a consequence of the boundedness of the second and fourth moments of $V_n$. The second inequality comes from the fact that $0 \leq p_{ii} \leq 1$. The first argument of the fourth equality is the definition of the trace. For the second term in the fourth equality two relationships are used. The first:

$$\begin{aligned} E\left(\text{trace}\left(P_{zk}\right)^2\right) &= E\left(\sum_{i=1}^{n} p_{ii}\right)^2 \\ &= E\left(\sum_{i=1}^{n} p_{ii}^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} p_{ii}p_{jj}\right) \end{aligned}$$

The second argument used to derive the fourth inequality of $E\left\|V_n' P_{zk} V_n\right\|^2$ is given by the fact that:

$$\text{trace}\left(E\left(P_{zk}' P_{zk}\right)\right) = \sum_{i=1}^{n} E\left(p_{ii}^2\right) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\left(p_{ij}^2\right)$$

Combining these two expression I obtain that:

$$\begin{aligned} & E\left(\text{trace}\left(P_{zk}\right)^2\right) + 2\text{trace}\left(E\left(P_{zk}' P_{zk}\right)\right) - 3\sum_{i=1}^{n} E\left(p_{ii}^2\right) \\ = \; & E\left(\sum_{i=1}^{n} p_{ii}^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} p_{ii}p_{jj}\right) + 2\left(\sum_{i=1}^{n} E\left(p_{ii}^2\right) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\left(p_{ij}^2\right)\right) - 3\sum_{i=1}^{n} E\left(p_{ii}^2\right) \\ = \; & 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} E\left(2p_{ij}^2 + p_{ii}p_{jj}\right) \end{aligned}$$

The conclusion from the analysis of the first three terms is that, for a positive constant $C_1$:

$$Y_{2n}' P_{zk} Y_{2n} = O_p(1)\frac{E\left(p_{nk}\right) r_n}{m_{n1}} + O_p(1)\frac{E\left(p_{nk}\right)\sqrt{r_n}}{m_{n1}} + \left(C_1 + o_p(1)\right)\left(E\left(p_{nk}\right) + E\left(\sqrt{p_{nk}}\right)\right)$$

From a similar analysis it follows that:

$$
\begin{aligned}
Y_{2n}' P_{zk} u_n &= \left(Z_n \Pi_n + V_n\right)' P_{zk} u_n \\
&= \Pi_n' Z_n' P_{zk} u_n + V_n' P_{zk} u_n \\
&= O_p\left(1\right) \frac{E\left(p_{nk}\right) \sqrt{r_n}}{m_{n1}} + \left(C_1 + o_p\left(1\right)\right) \left(E\left(p_{nk}\right) + \sqrt{E\left(p_{nk}\right)}\right)
\end{aligned}
$$

It follows that:

$$
\begin{aligned}
& \hat{\beta} - \beta \\
= & \left(O_p\left(1\right) \frac{E\left(p_{nk}\right) r_n}{m_{n1}} + O_p\left(1\right) \frac{E\left(p_{nk}\right) \sqrt{r_n}}{m_{n1}} + \left(C_1 + o_p\left(1\right)\right) \left(E\left(p_{nk}\right) + \sqrt{E\left(p_{nk}\right)}\right)\right)^{-1} \\
& \left(O_p\left(1\right) \frac{E\left(p_{nk}\right) \sqrt{r_n}}{m_{n1}} + \left(C_1 + o_p\left(1\right)\right) \left(E\left(p_{nk}\right) + \sqrt{E\left(p_{nk}\right)}\right)\right) \\
= & \left(O_p\left(1\right) \frac{r_n}{m_{n1}} + O_p\left(1\right) \frac{\sqrt{r_n}}{m_{n1}} + \left(C_1 + o_p\left(1\right)\right) \left(1 + \frac{1}{\sqrt{E\left(p_{nk}\right)}}\right)\right)^{-1} \\
& \left(O_p\left(1\right) \frac{\sqrt{r_n}}{m_{n1}} + \left(C_1 + o_p\left(1\right)\right) \left(1 + \frac{1}{\sqrt{E\left(p_{nk}\right)}}\right)\right) \\
= & \left(C_1 + o_p\left(1\right)\right)^{-1} \left(C_1 + o_p\left(1\right)\right) \\
= & O_p\left(1\right)
\end{aligned}
$$

In the expression above the first three equalities follow from algebraic manipulations and the definitions of $r_n$ and $m_{n1}$. The last equality results from the fact that $C_1$ is a positive constant.

□

For the nonparametric case a similar analysis follows. In particular by assumption 7(c):

$$\begin{aligned}\sqrt{r_{nx}} I\left\{X_{n1} \neq \tilde{X}_{n1}\right\} &= O_p\left(\frac{\lambda_{nl}\sqrt{r_{nx}}}{m_{nx2}^2}\right) \\ &= o_p(1)\end{aligned}$$

It then remains to be established that:

**Lemma 3.** For any random set of instruments and basis functions defined by $X_{nk}$:

$$\left(\hat{\beta} - \beta\right) = O_p(1)$$

*Proof.* Defining $\theta \equiv f_0 - X_{n1}\Pi_{n1}$, the selected $X_n$s by $\tilde{X}_{n1}$, the true nonzero ones by $X_{n1}$, $P_{\tilde{X}} = \tilde{X}_{n1}\left(\tilde{X}_{n1}'\tilde{X}_{n1}\right)^{-1}\tilde{X}_{n1}$, $P_x = X_{n1}\left(X_{n1}'X_{n1}\right)^{-1}X_{n1}$, and the number of random $X_{nk}$s by $p_{nk}s_n$:

$$\left(\hat{\beta} - \beta\right) = \left(Y_{2n}'P_{\tilde{X}}Y_{2n}\right)^{-1}Y_{2n}'P_{\tilde{X}}u_n$$

As in Lemma 2 the analysis starts with:

$$\begin{aligned}Y_{2n}'P_{\tilde{X}}Y_{2n} &= \left(\theta' + \Pi_{n1}'X_{n1}' + V_n'\right)P_{\tilde{X}}\left(\theta + X_{n1}\Pi_{n1} + V_n\right) \\ &= \theta'P_{\tilde{X}}\theta + \Pi_{n1}'X_{n1}'P_{\tilde{X}}\theta + V_n'P_{\tilde{X}}\theta + \Pi_{n1}'X_{n1}'P_{\tilde{X}}X_{n1}\Pi_{n1}\end{aligned}$$

$$+ \quad \Pi_{n1}' X_{n1}' V_n + V_n' P_{\tilde{X}} V_n + AE \tag{1.31}$$

In the expression above $AE$ refers to terms that are asymptotically equivalent to those already accounted for.

From Lemma 2 it follows that for a positive constant $C_1$:

$$\begin{aligned} Y_{2n}' P_{\tilde{X}} Y_{2n} &= \theta' P_{\tilde{X}} \theta + \Pi_{n1}' X_{n1}' P_{\tilde{X}} \theta + V_n' P_{\tilde{X}} \theta + \frac{O_p(1) E(p_{nk} s_n) r_{nx}}{m_{n1x}} \\ &+ \frac{O_p(1) E(p_{nk} s_n) \sqrt{r_{nx}}}{m_{n1x}} + (C_1 + o_p(1)) \left( E(p_{nk} s_n) + \sqrt{E(p_{nk} s_n)} \right) + AE \end{aligned}$$

Also, by the theory of regression splines, assumption 5, assumption 6, and assumption 7:

$$\begin{aligned} E \left\| \theta' P_{\tilde{X}} \theta \right\|^2 &\leq E \left\| \theta' \theta \right\|^2 \left\| P_{\tilde{X}} \right\|^2 \\ &= o_p(1) E(p_{nk} s_n) \end{aligned}$$

$$\begin{aligned} E \left\| \Pi_{n1}' X_{n1}' P_{\tilde{X}} \theta \right\|^2 &\leq E \left\| \Pi_{n1}' X_{n1} \right\|^2 \left\| P_{\tilde{X}} \right\|^2 \left\| \theta \right\|^2 \\ &= E(p_{nk} s_n) \sqrt{r_{nx}} \frac{nq}{m_{nx2}} \left[ s_n^{-\partial} + \sqrt{\frac{s_n}{n}} \right]^2 \\ &= E(p_{nk} s_n) o_p(1) \end{aligned}$$

$$E \left\| V_n' P_{\tilde{X}} \theta \right\|^2 \leq E \left\| \theta \right\|^2 E \left\| P_{\tilde{X}} \right\|^2 E \left\| V_n' \right\|^2$$

$$= o_p(1)\, E(p_{nk}s_n)$$

Using similar arguments it can be shown that:

$$\begin{aligned} Y_{2n}' P_{\tilde{X}} u_n &= \theta' P_{\tilde{X}} u_n + V_n' P_{\tilde{X}} u_n + \Pi_{n1}' X_{n1}' P_{\tilde{X}} u_n \\ &= o_p(1)\, E(p_{nk}s_n) + \left(C_1 + o_p(1)\right) \left(E(p_{nk}s_n) + \sqrt{E(p_{nk}s_n)}\right) \end{aligned}$$

From the arguments used in the proof of Lemma 2 it is immediate that:

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

□

Below Lemma 2 and Lemma 3 are used. The aforementioned Lemmas imply that in the proofs below the true $p_{n1}$ and $Z_{n1}$ can be used instead of their estimated counterparts.

**Theorem 1**

*Proof.*

$$\widehat{\Pi}_{n1} = \left(Z_{n1}' Z_{n1}\right)^{-1} Z_{n1}' Y_{2n}$$

Thus, $\widehat{Y}_{2n}$ satisfies:

$$\widehat{Y}_{2n} = Z_{n1}\left(Z'_{n1}Z_{n1}\right)^{-1}Z'_{n1}Y_{2n} \tag{1.32}$$

I analyze the expression $\frac{\widehat{Y}'_{2n}u_n}{r_n}$:

$$\begin{aligned}
\frac{\widehat{Y}'_{2n}u_n}{r_n} &= \frac{Y'_{2n}Z_{n1}\left(Z'_{n1}Z_{n1}\right)^{-1}Z'_{n1}u_n}{r_n} \\
\frac{\widehat{Y}'_{2n}u_n}{r_n} &= \frac{\left(Z_{n1}\Pi_{n1}+V_n\right)'Z_{n1}\left(Z'_{n1}Z_{n1}\right)^{-1}Z'_{n1}u_n}{r_n} \\
\frac{\widehat{Y}'_{2n}u_n}{r_n} &= \frac{\Pi'_{n1}Z'_{n1}u_n}{r_n} + \frac{V'_nZ_{n1}\left(Z'_{n1}Z_{n1}\right)^{-1}Z'_{n1}u_n}{r_n} \qquad (1.33)
\end{aligned}$$

For the first term above:

$$\begin{aligned}
E\left\|\frac{\Pi'_{n1}Z'_{n1}u_n}{r_n}\right\|^2 &= \frac{E\left(\text{trace}\left(u'_nZ_{n1}\Pi_{n1}\Pi'_{n1}Z'_{n1}u_n\right)\right)}{r_n^2} \\
&\le \frac{CE\left(\text{trace}\left(Z_{n1}\Pi_{n1}\Pi'_{n1}Z'_{n1}\right)\right)}{r_n^2} \\
&\le \frac{C}{r_n} \\
&= o_p(1)
\end{aligned}$$

The first equality comes from the definition of the matrix norm. The first inequality uses assumption 5 and the law of iterated expectations. The second inequality and the conclusion are a consequence of assumption 6.

For the second term:

$$\begin{aligned} E\left\| \frac{V_n' Z_{n1} \left(Z_{n1}' Z_{n1}\right)^{-1} Z_{n1}' u_n}{r_n} \right\|^2 &= \frac{E\left(\text{trace}\left(u_n' Z_{n1} \left(Z_{n1}' Z_{n1}\right)^{-1} Z_{n1}' V_n V_n' Z_{n1} \left(Z_{n1}' Z_{n1}\right)^{-1} Z_{n1}' u_n\right)\right)}{r_n^2} \\ &\leq \frac{CE\left(\text{trace}\left(Z_{n1} \left(Z_{n1}' Z_{n1}\right)^{-1} Z_{n1}'\right)\right)}{r_n^2} \\ &= \frac{C p_{n1}}{r_n^2} \\ &= o_p\left(1\right) \end{aligned}$$

The first inequality uses the law of iterated expectations and assumption 5. The second inequality follows by the properties of the trace operator. The conclusion comes from Lemma 1.

□

### 1.7.3 Theorem 2

#### Lemmas for Theorem 2

**Lemma 4.** If the conditions of Proposition 2 are satisfied by assumption 7(c):

$$\frac{p_{n1} s_n}{r_{nx}} \to 0 \qquad \text{as} \quad n \to \infty$$

*Proof.* Replacing $r_{nx}$ in the expression $\frac{p_{n1}s_n}{r_{nx}}$ by its definition and using the corollary of Proposition 2:

$$\frac{p_{n1}s_n}{r_{nx}} = \frac{p_{n1}s_n b_n^2}{m_{nx}h_n} \leq C\frac{p_{n1}m_{nx}^2}{m_{nx}^2\lambda_{nl}^2 p_{n1}s_n} \to 0 \tag{1.34}$$

□

**Theorem 2**

*Proof.* As in Theorem 1 the starting point is:

$$\widehat{\Pi}_{n1} = \left(X_{n1}'X_{n1}\right)^{-1} X_{n1}'Y_{2n} \tag{1.35}$$

$$= \left(X_{n1}'X_{n1}\right)^{-1} X_{n1}'V_n + \left(X_{n1}'X_{n1}\right)^{-1} X_{n1}'\theta + \Pi_{n1} \tag{1.36}$$

From (1.35) and defining $P_x = X_{n1}\left(X_{n1}'X_{n1}\right)^{-1} X_{n1}'$ it follows that:

$$\hat{f}_n = P_x V_n + P_x\theta + X_{n1}\Pi_{n1} \tag{1.37}$$

It needs to be shown that:

$$\frac{\hat{f}_n' u_n}{r_{nx}} = o_p(1)$$

Relationship (1.37) implies that:

$$\frac{\hat{f}_n' u_n}{r_{nx}} = \frac{V_n' P_x u_n + \theta' P_x u_n + \Pi_{n1}' X_{n1}' u_n}{r_{nx}} \tag{1.38}$$

For the first term of (1.38):

$$\begin{aligned}
\frac{E\left\|V_n' P_x u_n\right\|^2}{r_{nx}^2} &= \frac{E\left(\text{trace}\left(u_n' P_x V_n V_n' P_x u_n\right)\right)}{r_{nx}^2} \\
&\leq \frac{CE\left(\text{trace}\left(P_x\right)\right)}{r_{nx}^2} \\
&\leq C\frac{p_{n1} s_n}{r_{nx}^2} \\
&\to 0
\end{aligned}$$

The first equality is the definition of the matrix norm. The second inequality is a consequence of the law of iterated expectations and assumption 5. The third inequality comes from the definition of $P_x$ and manipulations inside the trace operator. The conclusion follows from Lemma 4.

For the second term:

$$\begin{aligned}
\frac{E\left\|\theta' P_x u_n\right\|^2}{r_{nx}^2} &= \frac{E\left(\text{trace}\left(u_n' P_x \theta\theta' P_x u_n\right)\right)}{r_{nx}^2} \\
&\leq \frac{CE\left(\text{trace}\left(\theta'\theta\right)^2 \text{trace}\left(P_x\right)^2\right)}{r_{nx}^2} \\
&\leq \frac{Cn\left(qp_{n1}s_n\right)^2}{r_{nx}^2 m_{nx}^2}\left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^4 \\
&\rightarrow 0
\end{aligned}$$

The first inequality is a result of the law of iterated expectations and the fact that $\|AB\| \leq \|A\| \, \|B\|$. The second inequality is a consequence of Lemma 1 of Huang et al. (2010). The conclusion follows by assumption 7.

For the third term:

$$\begin{aligned}
\frac{E\left\|\Pi_{n1}' X_{n1}' u_n\right\|^2}{r_{nx}^2} &= \frac{E\left(\text{trace}\left(u_n' X_{n1}\Pi_{n1}\Pi_{n1}' X_{n1}' u_n\right)\right)}{r_{nx}^2} \\
&\leq \frac{CE\left(\text{trace}\left(X_{n1}\Pi_{n1}\Pi_{n1}' X_{n1}'\right)\right)}{r_{nx}^2} \\
&\leq \frac{C}{r_{nx}} \\
&\rightarrow 0
\end{aligned}$$

The first inequality is a consequence of the law of iterated expectations. The second inequality follows from assumption 6. The conclusion is a result of the fact that $r_{nx}$ grows to infinity as $n$ grows to infinity.

The above statements imply that:

$$\frac{\hat{f}_n' u_n}{r_{nx}} = o_p(1)$$

□

### 1.7.4 Theorem 3

*Proof.* Let us first define the proposed estimator of $\beta$, $\hat{\beta}$, and rewrite the latter in a way that is more malleable to construct the argument:

$$\begin{aligned}
\hat{\beta} &\equiv \left(\widehat{Y}_{2n}'\widehat{Y}_{2n}\right)^{-1}\widehat{Y}_{2n}'y_{1n} \\
&= \left(\widehat{\Pi}_{n1}'Z_{n1}'Z_{n1}\widehat{\Pi}_{n1}\right)^{-1}\widehat{\Pi}_{n1}'Z_{n1}'y_{1n} \\
&= \left(\widehat{\Pi}_{n1}'Z_{n1}'Z_{n1}\widehat{\Pi}_{n1}\right)^{-1}\widehat{\Pi}_{n1}'Z_{n1}'\left(Z_{n1}\Pi\beta + V_n\beta + u_n\right)
\end{aligned} \tag{1.39}$$

To construct the proof equation 1.39 above is multiplied on both sides by $\sqrt{r_n}$. This yields for the right hand side, $\left(\frac{\widehat{\Pi}_{n1}'Z_{n1}'Z_{n1}\widehat{\Pi}_{n1}}{r_n}\right)^{-1}\frac{\widehat{\Pi}_{n1}'Z_{n1}'}{\sqrt{r_n}}\left(Z_{n1}\Pi\beta + V_n\beta + u_n\right)$ The analysis starts with the term $\frac{\widehat{\Pi}_{n1}'Z_{n1}'Z_{n1}\widehat{\Pi}_{n1}}{r_n} \equiv B_1 + B_2$. The expression below comes from replacing $\widehat{\Pi}_n$.

$$\begin{aligned}
B_1 &\equiv \left(\frac{\Pi_{n1}'Z_{n1}'Z_{n1}\Pi_{n1}}{r_n}\right) + \frac{\Pi_{n1}'Z_{n1}'V_n}{r_n} \\
&= \Phi + \frac{\Pi_{n1}'Z_{n1}'V_n}{r_n} + o_p(1) \\
&= \Phi + O_p\left(\sqrt{\frac{1}{r_n}}\right) + o_p(1) \\
&= \Phi + o_p(1)
\end{aligned} \tag{1.40}$$

In equation 1.40 above the first equality uses the fact that $\left(\frac{\Pi_{n1}'Z_{n1}'Z_{n1}\Pi_{n1}}{r_n}\right) \to \Phi$. The second equality follows from similar arguments to those employed in the proof of Theorem 1. The last equality follows by assumption 7 and by the definition of $r_n$.

$$\begin{aligned}
B_2 &\equiv \frac{V_n'Z_{n1}\Pi_{n1}}{r_{1n}} + \frac{V_n'Z_{n1}\left(Z_{n1}'Z_{n1}\right)^{-1}Z_{n1}'V_n}{r_n} \\
&= O_p\left(\sqrt{\frac{1}{r_n}}\right) + O_p\left(\frac{\sqrt{p_{n1}}}{r_n}\right) \\
&= o_p(1)
\end{aligned}$$

The first equality follows from similar arguments to those employed in the proof of Theorem 1 and the arguments used for $B_1$. The last equality follows by assumption 7.

The summary of the three statements above is that:

$$\frac{\widehat{\Pi}'_{n1}Z'_{n1}Z_{n1}\widehat{\Pi}_{n1}}{r_n} = \Phi + o_p(1) \tag{1.41}$$

The second term under analysis is $\frac{\widehat{\Pi}'_{n1}Z'_{n1}(Z_{n1}\Pi\beta+V_n\beta+u_n)}{\sqrt{r_n}} \equiv B_3$. Defining $\epsilon \equiv 2V_n\beta + u_n$ and $\tilde{\epsilon} \equiv V_n\beta + u_n$, $B_3$ can be rewritten as:

$$\begin{aligned}
B_3 &= \frac{2\Pi'_{n1}Z'_{n1}V_n\beta + \Pi'_{n1}Z'_{n1}u_n + V_nZ_{n1}\left(Z'_{n1}Z_{n1}\right)^{-1}Z'_{n1}V_n\beta + V_nZ_{n1}\left(Z'_{n1}Z_{n1}\right)^{-1}Z'_{n1}u_n}{\sqrt{r_n}} \\
&+ \frac{\Pi'_{n1}Z'_{n1}Z'_{n1}\Pi_{n1}\beta}{\sqrt{r_n}} \\
&= \Phi\beta\sqrt{r_n} + \frac{\Pi'Z'_{n1}\epsilon}{\sqrt{r_n}} + \frac{V'_nZ_{n1}\left(Z'_{n1}Z_{n1}\right)^{-1}Z'_{n1}\tilde{\epsilon}}{\sqrt{r_n}} + o_p(1) \\
&= \Phi\beta\sqrt{r_n} + \frac{\Pi'Z'_{n1}\epsilon}{\sqrt{r_n}} + O_p\left(\sqrt{\frac{p_{n1}}{r_n}}\right) + o_p(1) \\
&= \Phi\beta\sqrt{r_n} + \frac{\Pi'Z'_{n1}\epsilon}{\sqrt{r_n}} + o_p(1)
\end{aligned}$$

The second equality is a consequence of the assumption that $\frac{\Pi'_{n1}Z'_{n1}Z'_{n1}\Pi_{n1}}{r_n} \to \Phi$ and the third equality follows by arguments analogous to the ones used in the proof of Theorem 1 . The fourth equality equality is the result of some algebraic manipulation of the last term in the third equality, the definition of $r_n$. The result follows by assumption 7.

Combining the previous arguments:

$$\sqrt{r_n}\left(\hat{\beta}-\beta\right)=\Phi^{-1}\left(\frac{\Pi'_{n1}Z'_{n1}\epsilon}{\sqrt{r_n}}\right)+o_p(1)$$

It now suffices to show that $\frac{\Pi'_{n1}Z'_{n1}\epsilon}{\sqrt{r_n}}$ satisfies Liupanov's condition to render the expression asymptotically normal. First, note that by the law of iterated expectations and assumption 5, $E\left(\frac{\Pi'_{n1}Z'_{n1}\epsilon}{\sqrt{r_n}}\right)=0$. Also:

$$\begin{aligned}\text{Var}\left(\frac{\Pi'_{n1}Z'_{n1}\epsilon}{\sqrt{r_n}}\right) &= \frac{E\left(Z'_{n1}\Pi'_{n1}\epsilon\epsilon' Z_{n1}\Pi_{n1}\right)}{r_n}\\ &\equiv \Omega\end{aligned}$$

By assumptions 5 and 6 this term is finite. What remains to be determined is that for a $\delta>0$, $\lim_{n\rightarrow\infty}E\left|\frac{\Pi'_{n1}Z'_{n1}\epsilon}{\sqrt{r_n}}\right|^{2+\delta}=0$. For $\delta=1$ and by the symmetry imposed by assumption 5(c):

$$\begin{aligned}E\left|\frac{\Pi'_{n1}Z'_{n1}\epsilon}{\sqrt{r_n}}\right|^{3} &= \frac{2E\left(Z'_{n1}\Pi'_{n1}\epsilon\epsilon' Z_{n1}\Pi_{n1}Z'_{n1}\Pi'_{n1}\epsilon\right)}{r_n^{3/2}}\\ &= \frac{2E\left(Z'_{n1}\Pi'_{n1}E\left(\epsilon\epsilon'|Z\right)Z_{n1}\Pi_{n1}Z'_{n1}\Pi'_{n1}E\left(\epsilon|Z\right)\right)}{r_n^{3/2}}\\ &= 0\end{aligned}$$

The conclusions above follow by assumption 5. In particular the symmetry assumption 5(c) is crucial.

The previous results imply that:

$$\sqrt{r_n}\left(\hat{\beta}-\beta\right)=N\left(0,\Lambda\right) \tag{1.42}$$

In the expression above $\Phi^{-1'}\Omega\Phi^{-1}\equiv\Lambda$.

□

### 1.7.5 Theorem 4

In this subsection a proof of Theorem 4 is provided. The first step is to prove the consistency of the variance-covariance estimator of $\Lambda$. To do so a set of Lemmas is established that are applications of the Law of Large Numbers.

**Proofs of Lemmas**

**Lemma 5.**

$$\hat{\sigma}_v^2\equiv\widehat{\text{Var}}\left(V_n\beta\right)\to\sigma_v^2\equiv\text{Var}\left(V_n\beta\right)$$

*Proof.* Let $Z_{n1}\left(Z'_{n1}Z_{n1}\right)^{-1}Z'_{n1}\equiv P_z$. Then from the proof of Theorem 3 and $\hat{V}_n=Y_{2n}-\hat{Y}_{2n}$

it follows that:

$$\begin{aligned}\hat{V}_n\hat{\beta} &= \left(V_n - P_z V_n\right)\left(\beta + o_p\left(1\right)\right)\\ &= V_n\beta - P_z V_n\beta + o_p\left(1\right)\end{aligned}$$

The previous development allows the estimated variance to be written as:

$$\begin{aligned}\hat{\sigma}_v^2 &= \frac{\sum_{i=1}^n \left(\hat{v}_{ni}\hat{\beta}\right)^2}{n - p_{n1}}\\ &= \frac{\sum_{i=1}^n \left(v_{ni}\beta\left(1 - P_{zi}\right)\right)^2}{n - p_{n1}}\\ &= \frac{\beta' V_n' \left(I_{n\times n} - P_z\right) V_n\beta}{n - p_{n1}}\\ &= C\end{aligned}$$

The analysis below will follow for $E\left(\left(\hat{\sigma}_v^2\right)\right)$

$$\begin{aligned}E\left(C\right) &= \frac{E\left(E\left(C|Z\right)\right)}{n - p_{n1}}\\ &= \frac{E\left(E\left(\text{trace}\left(\beta' V_n' \left(I_{n\times n} - P_z\right) V_n\beta\right)|Z\right)\right)}{n - p_{n1}}\\ &= \text{Var}\left(V_n\beta\right)\frac{n - p_{n1}}{n - p_{n1}}\end{aligned}$$

$$= \quad \mathrm{Var}\left(V_n\beta\right)$$

The conclusion follows by the law of iterated expectations and the properties of the trace operator.The conclusion of the Lemma follows directly by the Law of Large Numbers.

□

**Lemma 6.**

$$\hat{\sigma}_{v\beta u} \equiv \frac{\hat{\beta}'\hat{V}_n'\hat{u}_n}{n-p_{n1}} - \hat{\sigma}_v^2 \to \sigma_{v\beta u} \to \mathrm{Cov}\left(V_n\beta, u_n\right)$$

*Proof.* Noting that $y_{1n} - \hat{y}_{1n} = \hat{u}_n$:

$$\begin{aligned} \hat{u}_n &= Z_{n1}\Pi_{n1}\beta + V_n\beta + u_n - \left(Z_{n1}\Pi_{n1} + P_zV_n\right)\left(\beta + o_p\left(1\right)\right) \\ &= V_n\beta - P_zV_n\beta + u_n + o_p(1) \end{aligned}$$

It follows immediately from Lemma 5 and the expression above that:

$$\frac{\hat{\beta}'\hat{V}_n'\hat{u}_n}{n-p_{n1}} = \hat{\sigma}_v^2 + \frac{u_n'\left(I_{n\times n} - P_z\right)V_n\beta}{n-p_{n1}} + o_p(1)$$

The first term on the right hand side above was analyzed in the previous lemma. It remains to explore the term $\frac{u_n'(I_{n\times n}-P_z)V_n\beta}{n-p_{n1}}$.

$$
\begin{aligned}
E\left(\frac{u_n'\left(I_{n\times n}-P_z\right)V_n\beta}{n-p_{n1}}\right) &= E\left(\text{trace}\left(\frac{u_n'\left(I_{n\times n}-P_z\right)V_n\beta}{n-p_{n1}}\right)\right)\\
&= \text{Cov}\left(V_n\beta, u\right)\\
&= \sigma_{v\beta u}
\end{aligned}
$$

The conclusion follows by the law of iterated expectations and similar algebra to the one used in Lemma 5. The conclusion of the Lemma follows directly by the Law of Large Numbers.

□

**Lemma 7.**

$$
\hat{\sigma}_{uu}^2 \equiv \frac{\hat{u}_n'\hat{u}_n}{n-p_{n1}} - 2\hat{\sigma}_{v\beta u} - \hat{\sigma}_v^2 \to \sigma_{uu} \equiv \text{Var}\left(u_n\right)
$$

*Proof.* From Lemma 6 the following holds:

$$
\begin{aligned}
\hat{u}_n &= Z_{n1}\Pi_{n1}\beta + V_n\beta + u_n - \left(Z_{n1}\Pi_{n1} + P_zV_n\right)\left(\beta + o_p\left(1\right)\right)\\
&= Z_{n1}\Pi_{n1}\beta + V_n\beta + u_n + P_zu_n - P_zu_n - \left(Z_{n1}\Pi_{n1} + P_zV_n\right)\left(\beta + o_p\left(1\right)\right)\\
&= \left(I_{n\times n}-P_z\right)V_n\beta + \left(I_{n\times n}-P_z\right)u_n + P_zu_n
\end{aligned}
$$

Using the expression above:

$$\begin{aligned}\frac{\hat{u}_n'\hat{u}_n}{n-p_{n1}} &= \frac{\hat{\sigma}_v^2+\beta' V_n'\left(I_{n\times n}-P_z\right)u_n+u_n'\left(I_{n\times n}-P_z\right)V_n\beta+u_n'\left(I_{n\times n}-P_z\right)u_n+u_n'P_zu_n+R}{n-p_{n1}} \\ &= \frac{\hat{\sigma}_v^2+2\hat{\sigma}_{v\beta u}+u_n'\left(I_{n\times n}-P_z\right)u_n+u_n'P_zu_n+R}{n-p_{n1}}\end{aligned}$$

In the equation above $R$ contains terms that have a conditional mean zero or are directly zero because they have terms of the form $\left(I_{n\times n}-P_z\right)P_z$. The second equality above comes from Lemma 6. Also:

$$\begin{aligned}\frac{E\left(u_n'P_zu_n\right)}{n-p_{n1}} &= \frac{E\left(\text{trace}\left(u_n'P_zu_n\right)\right)}{n-p_{n1}} \\ &= \frac{\sigma_{uu}p_{n1}}{n-p_{n1}} \\ &\to 0\end{aligned}$$

In the expression above the fact that $\frac{p_{n1}}{n}\to 0$, assumption 5, and the law of iterated expectations yield the conclusion.

Finally by arguments similar to the ones presented in the lemmas above:

$$E\left(\frac{u_n'\left(I_{n\times n}-P_z\right)u_n}{n-p_{n1}}\right)=\sigma_{uu}$$

The lemma follows by the arguments above and the Law of Large Numbers.

□

The first important fact to realize about Lemmas 5 to 7 is that they imply the consistency of $\hat{\Sigma}_{\epsilon\epsilon}$. $\Sigma_{\epsilon\epsilon} = \text{Var}\left(2V_n\beta + u_n\right)$ is equivalent to $4\text{Var}\left(V_n\beta\right) + \text{Var}\left(u_n\right) + 4\text{Cov}\left(V_n\beta, u_n\right)$. Therefore, by Lemmas 5 to 7 $\hat{\Sigma}_{\epsilon\epsilon} = 4\hat{\sigma}_v^2 + \hat{\sigma}_{uu}^2 + 4\hat{\sigma}_{v\beta u}$ is a consistent estimator of $\Sigma_{\epsilon\epsilon}$. Moreover, this estimator simplifies to a terse expression:

$$\begin{aligned}
4\hat{\sigma}_v^2 + \hat{\sigma}_{uu}^2 + 4\hat{\sigma}_{v\beta u} &= 4\hat{\sigma}_v^2 + \frac{\hat{u_n} u_n}{n - p_{n1}} - 2\hat{\sigma}_{v\beta u} - \hat{\sigma}_v^2 + 4\hat{\sigma}_{v\beta u} \\
&= 3\hat{\sigma}_v^2 + \frac{\hat{u_n}'\hat{u_n}}{n - p_{n1}} + 2\hat{\sigma}_{v\beta u} \\
&= 3\hat{\sigma}_v^2 + \frac{\hat{u_n}'\hat{u_n}}{n - p_{n1}} + 2\left(\frac{\hat{\beta}'\hat{V}_n'\hat{u}_n}{n - p_{n1}} - \hat{\sigma}_v^2\right) \\
&= \frac{\hat{\beta}'\hat{V}_n'\hat{V}_n\hat{\beta}}{n - p_{n1}} + \frac{\hat{u_n}'\hat{u_n}}{n - p_{n1}} + 2\frac{\hat{\beta}'\hat{V}_n'\hat{u}_n}{n - p_{n1}}
\end{aligned}$$

**Lemma 8.**

$$r_n\hat{\Lambda} \to \Lambda + o_p(1)$$

*Proof.* By Lemmas 5 to 7, equation (1.41) and assumption 6:

$$\begin{aligned} r_n\hat{\Lambda} &= r_n\hat{\Phi}^{-1'}\hat{\Omega}\hat{\Phi}^{-1} \\ &= \left(\frac{\widehat{\Pi}'_{n1}Z'_{n1}Z_{n1}\widehat{\Pi}_{n1}}{r_n}\right)^{-1}\frac{\widehat{\Pi}'_{n1}Z'_{n1}}{\sqrt{r_n}}\hat{\Sigma}_{\epsilon\epsilon}\frac{Z_{n1}\widehat{\Pi}_{n1}}{\sqrt{r_n}}\left(\frac{Z_{n1}\widehat{\Pi}_{n1}\widehat{\Pi}'_{n1}Z'_{n1}}{r_n}\right)^{-1} \\ &= \Lambda + o_p(1) \end{aligned} \tag{1.43}$$

The third equality is a consequence of assumption 6 and Lemma 1 noting that:

$$\begin{aligned} \frac{\widehat{\Pi}'_{n1}Z'_{n1}}{\sqrt{r_n}} &= \frac{\Pi'_{n1}Z'_{n1}+V'_nP_z}{\sqrt{r_n}} \\ &= \Phi + O_p\left(\sqrt{\frac{p_{n1}}{r_n}}\right) \\ &= \Phi + o_p\left(1\right) \end{aligned}$$

□

**Proof of Theorem 4**

*Proof.* Lemma 8 yields that:

$$\begin{aligned} \hat{\Lambda}^{-1/2}\left(\hat{\beta}-\beta\right) &\to \Lambda^{-1/2}\sqrt{r_n}\left(\hat{\beta}-\beta\right) \\ &\to_d N\left(0,1\right) \end{aligned} \tag{1.44}$$

□

### 1.7.6 Theorem 5

*Proof.* Let us first define the proposed estimator of $\beta$, $\hat{\beta}$, and rewrite the latter in a way that is more malleable to construct the argument:

$$\begin{aligned}
\hat{\beta} &\equiv \left(\widehat{Y}_{2n}'\widehat{Y}_{2n}\right)^{-1}\widehat{Y}_{2n}'y_{1n} \\
&= \left(\widehat{\Pi}_{n1}'X_{n1}'X_{n1}\widehat{\Pi}_{n1}\right)^{-1}\widehat{\Pi}_{n1}'X_{n1}'y_{1n} \\
&= \left(\widehat{\Pi}_{n1}'X_{n1}'X_{n1}\widehat{\Pi}_{n1}\right)^{-1}\widehat{\Pi}_{n1}'X_{n1}'\left(Y_{2n}\beta+u_n\right) \\
&= \left(\widehat{\Pi}_{n1}'X_{n1}'X_{n1}\widehat{\Pi}_{n1}\right)^{-1}\widehat{\Pi}_{n1}'X_{n1}'\left(\left(Y_{2n}+f_0-f_0+X_{n1}\Pi_{n1}-X_{n1}\Pi_{n1}\right)\beta+u_n\right) \\
&= \left(\frac{\widehat{\Pi}_{n1}'X_{n1}'X_{n1}\widehat{\Pi}_{n1}}{r_{nx}}\right)^{-1}\frac{\widehat{\Pi}_{n1}'X_{n1}'}{r_{nx}}\left(V_n\beta+\theta\beta+X_{n1}\Pi_{n1}\beta+u_n\right)
\end{aligned}$$

First it is established that:

$$\left(\frac{\widehat{\Pi}_{n1}'X_{n1}'X_{n1}\widehat{\Pi}_{n1}}{r_{nx}}\right)\to\Phi_x \tag{1.45}$$

As in Theorem 3 the first order conditions give us:

$$\begin{aligned}
\widehat{\Pi}_{n1} &= \left(X_{n1}'X_{n1}\right)^{-1}X_{n1}'Y_{2n} \\
&= \left(X_{n1}'X_{n1}\right)^{-1}X_{n1}'\left(Y_{2n}+f_0-f_0+X_{n1}\Pi_{n1}-X_{n1}\Pi_{n1}\right) \\
&= \left(X_{n1}'X_{n1}\right)^{-1}X_{n1}'V_n+\left(X_{n1}'X_{n1}\right)^{-1}X_{n1}'\theta+\Pi_{n1}
\end{aligned}$$

In the expression above $\theta \equiv f_0 - X_{n1}\Pi_{n1}$. It is important to notice that the expression is similar to the one derived in Theorem 3 above with the exception of one term: $\left(X'_{n1}X_{n1}\right)^{-1}X'_{n1}\theta$. Therefore, the results of Theorem 3 follow by assumption 7(c) and focus can be centered on expressions that involve $\left(X'_{n1}X_{n1}\right)^{-1}X'_{n1}\theta$ to get a final result.

$$\begin{aligned}\frac{\widehat{\Pi}'_{n1}X'_{n1}X_{n1}\widehat{\Pi}_{n1}}{r_{nx}} &= \frac{\Pi'_{n1}X'_{n1}X_{n1}\Pi_{n1}}{r_{nx}} + L_1 + L_2 \\ &= \Phi_x + o_p(1)\end{aligned}$$

In the expression above $L_1$ is given by:

$$\begin{aligned}L_1 &= \left[\Pi_{n1} + \left(X'_{n1}X_{n1}\right)^{-1}X'_{n1}V_n\right]'\frac{X'_{n1}X_{n1}}{r_{nx}} \\ &\quad \left[\Pi_{n1} + \left(X'_{n1}X_{n1}\right)^{-1}X'_{n1}V_n\right]\end{aligned}$$

The terms in the equation above are analogous to those in Theorem 3 and are of order $o_p(1)$ by assumption 7(c) and the arguments of Theorem 3. On the other hand, defining $P_x = X_{n1}\left(X'_{n1}X_{n1}\right)^{-1}X'_{n1}$, $L_2$ is given by:

$$L_2 = \frac{\Pi'_{n1}X'_{n1}\theta + V'_nP_x\theta + \theta'P_x\theta}{r_{nx}} + AE$$

In the equation above $AE$ is a collection of terms that are asymptotically equivalent to the other terms in the equation.

For the first term:

$$\begin{aligned}
\frac{E\left\|\Pi_{n1}' X_{n1}' \theta\right\|^2}{r_{nx}^2} &= \frac{E\left(\text{trace}\left(\theta' X_{n1} \Pi_{n1} \Pi_{n1}' \theta\right)\right)}{r_{nx}^2} \\
&\leq \Phi_x \frac{E\left(\text{trace}\left(\theta'\theta\right)\right)}{r_{nx}} \\
&\leq C\left(\frac{nq}{r_{nx} m_{nx2}}\left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^2\right) \\
&\to 0
\end{aligned}$$

In the expression above the first equality comes from the matrix norm. The first inequality comes from assumption 6(b). The second inequality follows from Lemma 1 of Huang et al. (2010). The conclusion comes from assumptions 7(c) and 7(d).

For the second term:

$$\begin{aligned}
\frac{E\left\|V_n' P_x \theta\right\|^2}{r_{nx}^2} &= \frac{E\left(\text{trace}\left(\theta' P_x V_n V_n' P_x \theta\right)\right)}{r_{nx}^2} \\
&\leq C \frac{E\left(\text{trace}\left(\theta' P_x P_x \theta\right)\right)}{r_{nx}^2} \\
&\leq C \frac{E\left(\text{trace}\left(\theta'\theta\right)^2 \text{trace}\left(P_x\right)^2\right)}{r_{nx}^2} \\
&= C \frac{\left(p_{n1} s_n\right)^2 E\left(\text{trace}\left(\theta'\theta\right)^2\right)}{r_{nx}^2}
\end{aligned}$$

$$\begin{aligned} &= CE\left(\text{trace}\left(\theta'\theta\right)^2\right)\frac{\left(p_{n1}s_n\right)^2}{r_{nx}^2} \\ &\to 0 \end{aligned}$$

Above the first equality is the matrix norm. The first inequality comes from assumption 5 that bounds the moments of $V_n$. The second inequality comes from the fact that $\text{trace}\,(AB)^n \leq \text{trace}(A'A)^n\text{trace}(B'B)^n$ for $n \geq 0$. The second equality arises from the properties of the trace operator and the dimension of $P_x$. The conclusion follows by the properties of the spline approximation and by $\frac{p_{n1}s_n}{r_{nx}} \to 0$ which shall be proved in Lemma 4 below.

For the third term:

$$\begin{aligned} \frac{E\left\|\theta' P_x \theta\right\|^2}{r_{nx}^2} &= \frac{E\left(\text{trace}\left(\theta' P_x \theta\theta' P_x \theta\right)\right)}{r_{nx}^2} \\ &\leq \frac{E\left(\text{trace}\left(\theta'\theta\right)^4 \text{trace}\left(P_x\right)^2\right)}{r_{nx}^2} \\ &= \frac{p_{n1}s_n}{r_{nx}^2}E\left(\text{trace}\left(\theta'\theta\right)^4\right) \\ &\to 0 \end{aligned}$$

The first inequality is due to $\text{trace}\,(AB)^n \leq \text{trace}(A'A)^n\text{trace}(B'B)^n$ for $n \geq 0$. The conclusion follows by assumption 7, Lemma 1 of Huang et al. (2010), and Lemma 4.

To complete the proof the second part of (1.45) has to be shown to be asymptotically normal. Defining this term as $F$ and $P_x = X_{n1}\left(X_{n1}'X_{n1}\right)^{-1}X_{n1}'$:

$$
\begin{aligned}
F &= \frac{\widehat{\Pi}_{n1}'X_{n1}'}{\sqrt{r_{nx}}}\left(V_n\beta + \theta\beta + X_{n1}\Pi_{n1}\beta + u_n\right) \\
&= \left(\frac{V_n'P_x + \theta' P_x + \Pi_{n1}'X_{n1}'}{\sqrt{r_{nx}}}\right)\left(V_n\beta + \theta\beta + X_{n1}\Pi_{n1}\beta + u_n\right) \\
&= F_1 + F_2 + F_3 + AE
\end{aligned}
$$

In the expression above AE refers to terms that are asymptotically equivalent to $F_1$ to $F_4$. Starting the analysis with $F_1$

$$
\begin{aligned}
F_1 &\equiv \frac{V_n'P_xV_n\beta + V_nP_x\theta\beta + V_n'P_xu_n}{\sqrt{r_{nx}}} \\
&= O_p\left(\sqrt{\frac{p_{n1}s_n}{r_{nx}}}\right) + \frac{V_n'P_x\theta\beta}{\sqrt{r_{nx}}} \\
&= O_p\left(\sqrt{\frac{p_{n1}s_n}{r_{nx}}}\right) + O_p\left(\sqrt{\frac{p_{n1}s_n}{r_{nx}}}\sqrt{\frac{qn}{m_{nx2}}}\left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]\right) \\
&\to 0
\end{aligned}
$$

The first equality follows by assumption 5 and the law of iterated expectations using arguments similar to those employed in the proof of Theorem 2. The conclusion follows by assumption 7. The second equality is a consequence of:

$$\begin{aligned}
\frac{E\left\|V_n' P_x \theta\beta\right\|^2}{r_{nx}} &= \frac{E\left(\text{trace}\left(\beta'\theta' P_x V_n V_n' P_x \theta\beta\right)\right)}{r_{nx}} \\
&\leq \frac{C p_{n1} s_n n q}{r_{nx} m_{nx2}} \left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^2 \\
&= O_p\left(\sqrt{\frac{p_{n1} s_n}{r_{nx}}} \sqrt{\frac{qn}{m_{nx2}}} \left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]\right)
\end{aligned}$$

Above the first inequality is a consequence of assumption 5 and Lemma 1 of Huang et al. (2010). The second equality is a result of assumption 6 and Proposition 2.

For $F_2$:

$$\begin{aligned}
F_2 &= \frac{\theta' P_x \theta\beta + \theta' X_{n1} \Pi_{n1} \beta + AE}{\sqrt{r_{nx}}} \\
&= O_p\left(\sqrt{\frac{nq}{m_{nx}}} \left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]\right) + O_p\left(\sqrt{\frac{p_{n1} s_n}{r_{nx}}} \frac{nq}{m_{nx}} \left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^2\right) \\
&\rightarrow 0
\end{aligned}$$

For the first term in $F_2$:

$$\begin{aligned}
\frac{E\left\|\theta' P_x \theta\beta\right\|^2}{r_{nx}} &= \frac{E\left(\text{trace}\left(\beta'\theta' P_x \theta\theta' P_x \theta\beta\right)\right)}{r_{nx}} \\
&\leq \frac{CE\left(\text{trace}\left(\theta'\theta\right)^4 \text{trace}\left(P_x\right)^2\right)}{r_{nx}} \\
&\leq \frac{C p_{n1} s_n}{r_{nx}} \frac{p_{n1} s_n}{m_{nx2}^2} \frac{n^4 q^4}{m_{nx2}^2} \left[s_n^{-\partial} + \sqrt{\frac{s_n}{n}}\right]^8
\end{aligned}$$

$$\rightarrow 0$$

The second inequality comes from Lemma 1 of Huang et al. (2010). The conclusion follows from assumption 7.

For the second term in $F_2$:

$$\begin{aligned}
\frac{E\left\|\theta' X_{n1}\Pi_{n1}\beta\right\|^2}{r_{nx}} &= \frac{E\left(\text{trace}\left(\beta'\Pi_{n1}' X_{n1}'\theta\theta' X_{n1}\Pi_{n1}\beta\right)\right)}{r_{nx}} \\
&\leq CE\left(\text{trace}\left(\theta'\theta\right)\right) \\
&\rightarrow 0
\end{aligned}$$

Let $\epsilon \equiv 2V_n\beta + u_n$. Then $F_3$ is given by:

$$\begin{aligned}
F_3 &= \frac{\Pi_{n1}' X_{n1}'\epsilon}{\sqrt{r_{nx}}} + \frac{\Pi_{n1}' X_{n1}' X_{n1}\Pi_{n1}\beta}{\sqrt{r_{nx}}} \\
&= \frac{\Pi_{n1}' X_{n1}'\epsilon}{\sqrt{r_{nx}}} + \sqrt{r_{nx}}\Phi_x\beta + o_p(1)
\end{aligned}$$

Combining the previous arguments:

$$\sqrt{r_{nx}}\left(\hat{\beta} - \beta\right) = \Phi_x^{-1}\left(\frac{\Pi_{n1}' X_{n1}'\epsilon}{\sqrt{r_{nx}}}\right) + o_p(1)$$

It now suffices to show that $\frac{\Pi_{n1}' X_{n1}' \epsilon}{\sqrt{r_{nx}}}$ satisfies Liupanov's condition to render the expression asymptotically normal. First, note that by the law of iterated expectations and assumption 5, $E\left(\frac{\Pi_{n1}' X_{n1}' \epsilon}{\sqrt{r_{nx}}}\right) = 0$. By virtue of this argument:

$$\begin{aligned} \text{Var}\left(\frac{\Pi_{n1}' X_{n1}' \epsilon}{\sqrt{r_n}}\right) &= \frac{E\left(X_{n1}' \Pi_{n1}' \epsilon \epsilon' X_{n1} \Pi_{n1}\right)}{r_{nx}} \\ &\equiv \Omega \end{aligned}$$

From assumptions 5 and 6 it follows that the term above is constant. What remains to be determined is that for a $\delta > 0$, $\lim_{n \to \infty} E\left|\frac{\Pi_{n1}' X_{n1}' \epsilon}{\sqrt{r_{nx}}}\right|^{2+\delta} = 0$. For $\delta = 1$ and by the symmetry imposed by assumption 5(c):

$$\begin{aligned} E\left|\frac{\Pi_{n1}' X_{n1}' \epsilon}{\sqrt{r_{nx}}}\right|^3 &= \frac{2E\left(X_{n1}' \Pi_{n1}' \epsilon \epsilon' X_{n1} \Pi_{n1} X_{n1}' \Pi_{n1}' \epsilon\right)}{r_{nx}^{3/2}} \\ &= \frac{2E\left(X_{n1}' \Pi_{n1}' E\left(\epsilon \epsilon' | X\right) X_{n1} \Pi_{n1} X_{n1}' \Pi_{n1}' E\left(\epsilon | X\right)\right)}{r_{nx}^{3/2}} \\ &= 0 \end{aligned}$$

The conclusions above follow by assumption 5. In particular the symmetry assumption 5(c) is crucial.

The previous results imply that:

$$\sqrt{r_{nx}}\left(\hat{\beta}-\beta\right) = N\left(0, \Lambda_x\right) \tag{1.46}$$

In the expression above $\Phi_x^{-1'}\Omega\Phi_x^{-1} \equiv \Lambda_x$.

□

**Lemma 9.**

$$r_n\hat{\Lambda}_x \to \Lambda_x + o_p(1)$$

*Proof.* By Lemmas 5 to 7 and arguments similar to those used in Theorem 3:

$$\begin{aligned} r_{nx}\hat{\Lambda}_x &= r_{nx}\hat{\Phi}_{x1}^{-1'}\hat{\Omega}\hat{\Phi}_{x1}^{-1} \\ &= \left(\frac{\widehat{\Pi}'_{n1}X'_{n1}X_{n1}\widehat{\Pi}_{n1}}{r_{nx}}\right)^{-1'}\left(\frac{\widehat{\Pi}'_{n1}X'_{n1}}{\sqrt{r_{nx}}}\right)\hat{\Sigma}_{\epsilon\epsilon}\left(\frac{X_{n1}\widehat{\Pi}_{n1}}{\sqrt{r_{nx}}}\right)\left(\frac{X_{n1}\widehat{\Pi}_{n1}\widehat{\Pi}'_{n1}X'_{n1}}{r_{nx}}\right)^{-1} \\ &= \Lambda_x + o_p(1) \end{aligned} \tag{1.47}$$

□

**Proof of Theorem 6**

*Proof.* Lemma 9 yields that:

$$\begin{aligned} \hat{\Lambda}_x^{-1/2}\left(\hat{\beta}-\beta\right) &\rightarrow \Lambda_x^{-1/2}\sqrt{r_{nx}}\left(\hat{\beta}-\beta\right) \\ &\rightarrow_d N\left(0,1\right) \end{aligned} \tag{1.48}$$

□

## 1.8. Tables

Table 1.1: Ranking of Post-$l_1$-Penalized against Traditional Estimators for DGP 1

| Instrument Weakness | Sample Size | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|---|
| **Low Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 1 | 3bc | 2a | 3bc | 1 | 2a |
| | 500 | 1 | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R^2_f = .1$) | 100 | 2a | 2c | 2a | 3bc | 1 | 2a |
| | 500 | 1 | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R^2_f = .01$) | 100 | 1 | 1 | 2a | 1 | 1 | 2a |
| | 500 | 2a | 2c | 2a | 3bc | 1 | 2a |
| **High Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 2c | 3bc | 2a | 3bc | 3bc | 2a |
| | 500 | 3bc | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R^2_f = .1$) | 100 | 1 | 3bc | 2a | 3bc | 2b | 2a |
| | 500 | 2c | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R^2_f = .01$) | 100 | 1 | 1 | 2a | 1 | 1 | 2a |
| | 500 | 1 | 1 | 2a | 3bc | 1 | 2a |

**Note 1**: 1 means that the Post-$l_1$-Penalized estimator had the lowest value with regards to the measure of the respective column. 2a means it ranks second to 2SLS, 2b to LIML, and 2c to the Fuller estimator. The same logic follows for the other rankings.

**Note 2**: DGP 1 is defined by by a normal distribution with the variance covariance matrix (1.19), with heteroskedasticity introduced via $\tilde{u}_n = u_n + Z'_{n1}Z_{n1}$. $\sigma_{vu} = .95$ defines a high level of endogeneity, $\sigma_{vu} = .15$ defines a low level of endogeneity, and $R^2_f$ is the pseudo R-squared of Kuersteiner and Okui (2010).

**Note 3:** The terms Moderately Weak and Completely Weak follow definition 2. For both values of $R^2_f$, with regards to the criteria of definition 2 the instruments are completely weak.

**Note 4**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.2: Ranking of Post-$l_1$-Penalized against Model Averaging Estimators for DGP 1

| **Low Endogeneity** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Instrument Weakness | Sample Size | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
| Moderately Weak | 100 | 1 | 1 | 1 | 1 | 1 | 3bc |
| | 500 | 1 | 1 | 1 | 1 | 1 | 4 |
| Completely Weak ($R_f^2 = .1$) | 100 | 3bc | 1 | 4 | 1 | 2a | 4 |
| | 500 | 1 | 1 | 4 | 1 | 2a | 4 |
| Completely Weak ($R_f^2 = .01$) | 100 | 3bc | 1 | 4 | 1 | 1 | 4 |
| | 500 | 3bc | 1 | 4 | 1 | 1 | 4 |
| **High Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 1 | 2a | 1 | 1 | 1 | 4 |
| | 500 | 1 | 1 | 1 | 1 | 1 | 4 |
| Completely Weak ($R_f^2 = .1$) | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 1 | 1 | 4 | 1 | 1 | 4 |
| Completely Weak ($R_f^2 = .01$) | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 1 | 1 | 4 | 1 | 1 | 4 |

**Note 1**: 1 means that the Post-$l_1$-Penalized estimator had the lowest value with regards to the measure of the respective column. 2a means it ranks second to the model averaging 2SLS, 2b to the model averaging LIML, 2c to the the model averaging Fuller estimator. The same logic follows for the other rankings.

**Note 2**:DGP 1 is defined by by a normal distribution with the variance covariance matrix (1.19), with heteroskedasticity introduced via $\tilde{u}_n = u_n + Z'_{n1}Z_{n1}$. $\sigma_{vu} = .95$ defines a high level of endogeneity, $\sigma_{vu} = .15$ defines a low level of endogeneity, and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010).

**Note 3:** The terms Moderately Weak and Completely Weak follow definition 2. For both values of $R_f^2$, with regards to the criteria of definition 2 the instruments are completely weak.

**Note 4**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.3: Ranking of Post-$l_1$-Penalized against JIVE and Adaptive LASSO for DGP 1

**Low Endogeneity**

| Instrument Weakness | Sample Size | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|---|
| Moderately Weak | 100 | 1 | 2j | 1 | 2j | 1 | 1 |
| | 500 | 1 | 2j | 1 | 2j | 2j | 1 |
| Completely Weak ($R^2_f = .1$) | 100 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 500 | 1 | 2j | 1 | 2j | 1 | 1 |
| Completely Weak ($R^2_f = .01$) | 100 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 500 | 1 | 1 | 1 | 1 | 1 | 1 |
| **High Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 1 | 2j | 1 | 2j | 2j | 1 |
| | 500 | 1 | 2j | 1 | 2j | 2j | 1 |
| Completely Weak ($R^2_f = .1$) | 100 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 500 | 1 | 2j | 1 | 2j | 2j | 1 |
| Completely Weak ($R^2_f = .01$) | 100 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 500 | 1 | 1 | 1 | 1 | 1 | 1 |

**Note 1**: 1 means that the Post-$l_1$-Penalized estimator had the lowest value with regards to the measure of the respective column. 2d means it ranks second to the adaptive LASSO and 2j to the JIVE.

**Note 2**:DGP 1 is defined by by a normal distribution with the variance covariance matrix (1.19), with heteroskedasticity introduced via $\tilde{u}_n = u_n + Z'_{n1}Z_{n1}$. $\sigma_{vu} = .95$ defines a high level of endogeneity, $\sigma_{vu} = .15$ defines a low level of endogeneity, and $R^2_f$ is the pseudo R-squared of Kuersteiner and Okui (2010).

**Note 3:** The terms Moderately Weak and Completely Weak follow definition 2. For both values of $R^2_f$, with regards to the criteria of definition 2 the instruments are completely weak.

**Note 4**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.4: Ranking of Post-$l_1$-Penalized against Traditional Estimators for DGP 2

| **Low Endogeneity** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Instrument Weakness | Sample Size | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
| Moderately Weak | 100 | 3bc | 3bc | 2a | 3bc | 3bc | 2a |
| | 500 | 3bc | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R_f^2 = .1$) | 100 | 1 | 3bc | 2a | 3bc | 3bc | 2a |
| | 500 | 3bc | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R_f^2 = .01$) | 100 | 1 | 2b | 2a | 1 | 1 | 3ac |
| | 500 | 1 | 2b | 2a | 3bc | 2b | 2a |
| **High Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 3bc | 3bc | 2a | 3bc | 3bc | 2a |
| | 500 | 3bc | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R_f^2 = .1$) | 100 | 2b | 3bc | 3ab | 3bc | 3bc | 2a |
| | 500 | 3bc | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R_f^2 = .01$) | 100 | 1 | 2b | 3ac | 3bc | 3bc | 4 |
| | 500 | 2b | 3bc | 3ac | 3bc | 3bc | 2a |

**Note 1**: 1 means that the Post-$l_1$-Penalized estimator had the lowest value with regards to the measure of the respective column. 2a means it ranks second to 2SLS, 2b to LIML, 2c to the Fuller estimator. The same logic follows for the other rankings.

**Note 2**: DGP 2 is defined by by a normal distribution with the variance covariance matrix (1.19) and mean zero. $\sigma_{vu} = .15$ defines a high level of endogeneity, $\sigma_{vu} = .95$ defines a high level of endogeneity, and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010).

**Note 3:** The terms Moderately Weak and Completely Weak follow definition 2. For both values of $R_f^2$, with regards to the criteria of definition 2 the instruments are completely weak.

**Note 4**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.5: Ranking of Post-$l_1$-Penalized against Model Averaging Estimators for DGP 2

**Low Endogeneity**

| Instrument Weakness | Sample Size | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|---|
| Moderately Weak | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 1 | 1 | 4 | 1 | 1 | 4 |
| Completely Weak ($R_f^2 = .1$) | 100 | 2a | 1 | 4 | 1 | 2a | 4 |
| | 500 | 1 | 1 | 4 | 1 | 2a | 4 |
| Completely Weak ($R_f^2 = .01$) | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 1 | 1 | 4 | 1 | 1 | 4 |
| **High Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 1 | 1 | 4 | 1 | 1 | 4 |
| Completely Weak ($R_f^2 = .1$) | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 2a | 1 | 4 | 2a | 2a | 4 |
| Completely Weak ($R_f^2 = .01$) | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 1 | 1 | 4 | 1 | 1 | 4 |

**Note 1**: 1 means that the Post-$l_1$-Penalized estimator had the lowest value with regards to the measure of the respective column. 2a means it ranks second to the model averaging 2SLS, 2b to the model averaging LIML, 2c to the the model averaging Fuller estimator. The same logic follows for the other rankings.

**Note 2**: DGP 2 is defined by by a normal distribution with the variance covariance matrix (1.19) and mean zero. $\sigma_{vu} = .15$ defines a low level of endogeneity, $\sigma_{vu} = .95$ defines a high level of endogeneity, and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010).

**Note 3:** The terms Moderately Weak and Completely Weak follow definition 2. For both values of $R_f^2$, with regards to the criteria of definition 2 the instruments are completely weak.

**Note 4**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.6: Ranking of Post-$l_1$-Penalized against JIVE and Adaptive LASSO for DGP 2

**Low Endogeneity**

| Instrument Weakness | Sample Size | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|---|
| Moderately Weak | 100 | 1 | 2j | 1 | 2j | 2j | 1 |
| | 500 | 2j | 2j | 1 | 2j | 2j | 1 |
| Completely Weak ($R^2_f = .1$) | 100 | 1 | 1 | 1 | 2j | 1 | 1 |
| | 500 | 1 | 2j | 1 | 2j | 2j | 1 |
| Completely Weak ($R^2_f = .01$) | 100 | 1 | 1 | 1 | 1 | 1 | 2j |
| | 500 | 1 | 1 | 1 | 1 | 1 | 1 |
| **High Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 1 | 2j | 1 | 2j | 2j | 1 |
| | 500 | 1 | 2j | 1 | 2j | 2j | 1 |
| Completely Weak ($R^2_f = .1$) | 100 | 1 | 2j | 1 | 2j | 1 | 1 |
| | 500 | 1 | 2j | 1 | 2j | 2j | 1 |
| Completely Weak ($R^2_f = .01$) | 100 | 1 | 1 | 1 | 1 | 1 | 2j |
| | 500 | 1 | 1 | 1 | 1 | 1 | 1 |

**Note 1**: 1 means that the Post-$l_1$-Penalized estimator had the lowest value with regards to the measure of the respective column. 2d means it ranks second to the adaptive LASSO and 2j to the JIVE.

**Note 2**: DGP 2 is defined by by a normal distribution with the variance covariance matrix (1.19) and mean zero. $\sigma_{vu} = .15$ defines a low level of endogeneity, $\sigma_{vu} = .95$ defines a high level of endogeneity, and $R^2_f$ is the pseudo R-squared of Kuersteiner and Okui (2010).

**Note 3:** The terms Moderately Weak and Completely Weak follow definition 2. For both values of $R^2_f$, with regards to the criteria of definition 2 the instruments are completely weak.

**Note 4**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.7: Ranking of Post-$l_1$-Penalized against Traditional Estimators for DGP 3

| Instrument Weakness | Sample Size | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|---|
| **Low Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 2a | 2c | 2a | 3bc | 2a | 2a |
| | 500 | 2a | 3bc | 2a | 3bc | 1 | 2a |
| Completely Weak ($R_f^2 = .1$) | 100 | 2a | 1 | 2a | 1 | 2a | 2a |
| | 500 | 2a | 3bc | 2a | 3bc | 2a | 2a |
| Completely Weak ($R_f^2 = .01$) | 100 | 2a | 1 | 2a | 1 | 2a | 2a |
| | 500 | 2a | 2b | 2a | 3bc | 2a | 2a |
| **High Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 2c | 3bc | 2a | 3bc | 3bc | 2a |
| | 500 | 3bc | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R_f^2 = .1$) | 100 | 2c | 3bc | 2a | 3bc | 3bc | 2a |
| | 500 | 3bc | 3bc | 2a | 3bc | 3bc | 2a |
| Completely Weak ($R_f^2 = .01$) | 100 | 1 | 1 | 3ac | 2b | 1 | 2a |
| | 500 | 1 | 3bc | 2a | 3bc | 3bc | 2a |

**Note 1**: 1 means that the Post-$l_1$-Penalized estimator had the lowest value with regards to the measure of the respective column. 2a means it ranks second to 2SLS, 2b to LIML, 2c to the Fuller estimator and 2d to the Adaptive LASSO. The same logic follows for the other rankings.

**Note 2**: DGP 3 is defined by a zero mean t-distribution with 5 degrees of freedom using the variance covariance matrix (1.19). $\sigma_{vu} = .95$ defines a high level of endogeneity while $\sigma_{vu} = .15$ defines a low level of endogeneity and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010).

**Note 3:** The terms Moderately Weak and Completely Weak follow definition 2. For both values of $R_f^2$, with regards to the criteria of definition 2 the instruments are completely weak.

**Note 4**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.8: Ranking of Post-$l_1$-Penalized against Model Averaging Estimators for DGP 3

| **Low Endogeneity** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Instrument Weakness | Sample Size | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
| Moderately Weak | 100 | 4 | 1 | 4 | 1 | 4 | 4 |
| | 500 | 2a | 1 | 4 | 1 | 2a | 4 |
| Completely Weak ($R_f^2 = .1$) | 100 | 4 | 1 | 4 | 1 | 4 | 4 |
| | 500 | 4 | 1 | 4 | 1 | 4 | 4 |
| Completely Weak ($R_f^2 = .01$) | 100 | 4 | 1 | 4 | 1 | 4 | 4 |
| | 500 | 4 | 1 | 4 | 1 | 4 | 4 |
| **High Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 1 | 1 | 4 | 1 | 1 | 4 |
| Completely Weak ($R_f^2 = .1$) | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 2a | 1 | 4 | 2a | 2a | 4 |
| Completely Weak ($R_f^2 = .01$) | 100 | 1 | 1 | 4 | 1 | 1 | 4 |
| | 500 | 1 | 1 | 4 | 1 | 1 | 4 |

**Note 1**: 1 means that the Post-$l_1$-Penalized estimator had the lowest value with regards to the measure of the respective column. 2a means it ranks second to the model averaging 2SLS, 2b to the model averaging LIML, 2c to the the model averaging Fuller estimator. The same logic follows for the other rankings.

**Note 2**: DGP 3 is defined by a zero mean t-distribution with 5 degrees of freedom using the variance covariance matrix (1.19). $\sigma_{vu} = .95$ defines a high level of endogeneity while $\sigma_{vu} = .15$ defines a low level of endogeneity and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010).

**Note 3:** The terms Moderately Weak and Completely Weak follow definition 2. For both values of $R_f^2$, with regards to the criteria of definition 2 the instruments are completely weak.

**Note 4**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.9: Ranking of Post-$l_1$-Penalized against JIVE and Adaptive LASSO for DGP 3

| **Low Endogeneity** | | | | | | | |
|---|---|---|---|---|---|---|---|
| Instrument Weakness | Sample Size | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
| Moderately Weak | 100 | 1 | 1 | 1 | 2j | 1 | 1 |
| | 500 | 1 | 2j | 1 | 2j | 1 | 1 |
| Completely Weak ($R_f^2 = .1$) | 100 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 500 | 1 | 1 | 1 | 2j | 1 | 1 |
| Completely Weak ($R_f^2 = .01$) | 100 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 500 | 1 | 1 | 1 | 2j | 1 | 1 |
| **High Endogeneity** | | | | | | | |
| Moderately Weak | 100 | 1 | 2j | 1 | 3 | 3 | 1 |
| | 500 | 1 | 2j | 1 | 2j | 3 | 1 |
| Completely Weak ($R_f^2 = .1$) | 100 | 1 | 1 | 1 | 1 | 2j | 1 |
| | 500 | 1 | 2j | 1 | 2j | 2j | 1 |
| Completely Weak ($R_f^2 = .01$) | 100 | 1 | 2j | 1 | 1 | 1 | 2j |
| | 500 | 1 | 1 | 1 | 1 | 1 | 1 |

**Note 1**: 1 means that the Post-$l_1$-Penalized estimator had the lowest value with regards to the measure of the respective column. 2d means it ranks second to the adaptive LASSO and 2j to the JIVE.

**Note 2**: DGP 3 is defined by a zero mean t-distribution with 5 degrees of freedom using the variance covariance matrix (1.19). $\sigma_{vu} = .95$ defines a high level of endogeneity while $\sigma_{vu} = .15$ defines a low level of endogeneity and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010).

**Note 3:** The terms Moderately Weak and Completely Weak follow definition 2. For both values of $R_f^2$, with regards to the criteria of definition 2 the instruments are completely weak.

**Note 4**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.10: DGP 1: Low Endogeneity Level Using $\Pi_n$ of equation (1.20)

| $R_f^2 = 0.1$; n=100 | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|
| 2SLS | 0.92 | 1.36 | 0.44 | 1.11 | 1.02 | 0.68 |
| Liml | 1053.75 | 2.46 | 1579.73 | 0.71 | 1.77 | 2.62 |
| Fuller | 3.77 | 0.67 | 5.45 | 0.77 | 1.48 | 2.16 |
| Lasso | 3.71 | 1.98 | 3.62 | 1.28 | 1.37 | 1.48 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 0.92 | 1.30 | 0.53 | 1.21 | 1.20 | 0.34 |
| MaFuller | 0.91 | 1.30 | 0.53 | 1.21 | 1.19 | 0.34 |
| Ma2SLS | 1.22 | 1.57 | 0.59 | 1.07 | 0.97 | 0.85 |
| JIVE | 707.76 | 3.15 | 1058.09 | 1.09 | 2.20 | 3.02 |
| $R_f^2 = 0.1$; n=500 | | | | | | |
| 2SLS | 1.03 | 1.07 | 0.74 | 1.03 | 1.06 | 0.90 |
| Liml | 3.66 | 0.21 | 12.37 | 0.60 | 0.90 | 2.14 |
| Fuller | 2.04 | 0.11 | 6.91 | 0.62 | 0.86 | 2.06 |
| Lasso | 2.43 | 1.61 | 2.08 | 1.23 | 1.56 | 1.34 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.99 | 1.60 | 0.59 | 1.27 | 1.65 | 0.34 |
| MaFuller | 1.99 | 1.60 | 0.59 | 1.27 | 1.65 | 0.34 |
| Ma2SLS | 1.00 | 1.08 | 0.61 | 1.00 | 0.99 | 0.99 |
| JIVE | 11.71 | 0.51 | 39.28 | 0.55 | 1.01 | 2.45 |
| $R_f^2 = 0.01$; n=100 | | | | | | |
| 2SLS | 1.01 | 2.20 | 0.33 | 1.35 | 1.33 | 0.38 |
| Liml | 124701.06 | 16.33 | 146837.88 | 1.30 | 3.41 | 2.30 |
| Fuller | 4.68 | 2.11 | 4.72 | 1.30 | 2.31 | 1.60 |
| Lasso | 3.82 | 2.53 | 3.36 | 1.39 | 1.49 | 1.46 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 0.79 | 1.90 | 0.29 | 1.34 | 1.33 | 0.17 |
| MaFuller | 0.80 | 1.91 | 0.29 | 1.35 | 1.33 | 0.16 |
| Ma2SLS | 1.24 | 2.54 | 0.31 | 1.32 | 1.32 | 0.45 |
| JIVE | 235.50 | 2.10 | 276.62 | 1.38 | 2.66 | 1.78 |
| $R_f^2 = 0.01$; n=500 | | | | | | |
| 2SLS | 0.88 | 1.19 | 0.25 | 1.05 | 1.10 | 0.63 |
| Liml | 1505.77 | 1.34 | 3298.76 | 0.85 | 2.40 | 5.17 |
| Fuller | 6.63 | 0.82 | 13.74 | 0.86 | 1.91 | 4.20 |
| Lasso | 5.52 | 2.40 | 5.24 | 1.57 | 2.06 | 1.82 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 0.86 | 1.22 | 0.12 | 1.08 | 1.15 | 0.19 |
| MaFuller | 0.86 | 1.22 | 0.12 | 1.08 | 1.15 | 0.20 |
| Ma2SLS | 1.00 | 1.31 | 0.12 | 1.05 | 1.09 | 0.74 |
| JIVE | 52476.96 | 14.94 | 114772.56 | 1.12 | 2.36 | 4.39 |

**Note 1**: The values in the table are computed as $\frac{\text{Estimator Measure}}{\text{Post Lasso Measure}}$ to capture the estimators' relative performance with respect to the Post-$l_1$-Penalized estimator.

**Note 2**: DGP 1 is defined by by a normal distribution with the variance covariance matrix (1.19), with heteroskedasticity introduced via $\tilde{u}_n = u_n + Z'_{n1}Z_{n1}$. $\sigma_{vu} = .15$ defines a low level of endogeneity and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010). The $\Pi_n$ was constructed following $\pi_{nj} = c\left(p_n\right)\left(1 - \frac{j - p_n/2}{p_n/2+1}\right)^4$ for $j \leq p_n/2$ and $\pi_{nj} = 0$ for $j > p_n/2$.

**Note 3**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.11: DGP 1: High Endogeneity Level Using $\Pi_n$ of equation (1.20)

| $R_f^2 = 0.1$; n=100 | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|
| 2SLS | 1.13 | 1.41 | 0.21 | 1.10 | 1.18 | 0.57 |
| Liml | 422.02 | 0.64 | 864.89 | 0.73 | 1.22 | 2.64 |
| Fuller | 1.90 | 0.99 | 2.87 | 0.81 | 0.91 | 1.93 |
| Lasso | 3.45 | 1.87 | 3.39 | 1.33 | 1.56 | 1.64 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.28 | 1.45 | 0.42 | 1.18 | 1.30 | 0.24 |
| MaFuller | 1.27 | 1.44 | 0.42 | 1.18 | 1.30 | 0.25 |
| Ma2SLS | 1.12 | 1.33 | 0.45 | 1.08 | 1.14 | 0.67 |
| JIVE | 2885.49 | 2.97 | 5907.18 | 1.06 | 1.72 | 3.01 |
| $R_f^2 = 0.1$; n=500 | | | | | | |
| 2SLS | 1.07 | 1.07 | 0.45 | 1.03 | 1.05 | 0.83 |
| Liml | 136.38 | 0.32 | 1147.24 | 0.45 | 0.53 | 2.77 |
| Fuller | 0.96 | 0.08 | 8.01 | 0.48 | 0.50 | 2.54 |
| Lasso | 2.98 | 1.68 | 4.09 | 1.32 | 1.56 | 1.71 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.86 | 1.43 | 0.52 | 1.23 | 1.41 | 0.28 |
| MaFuller | 1.86 | 1.43 | 0.52 | 1.23 | 1.41 | 0.27 |
| Ma2SLS | 1.00 | 1.03 | 0.54 | 1.00 | 1.00 | 0.91 |
| JIVE | 183.33 | 0.42 | 1541.85 | 0.43 | 0.78 | 4.29 |
| $R_f^2 = 0.01$; n=100 | | | | | | |
| 2SLS | 1.23 | 1.84 | 0.14 | 1.18 | 1.28 | 0.23 |
| Liml | 652.61 | 1.42 | 981.63 | 1.12 | 1.83 | 1.36 |
| Fuller | 2.50 | 1.87 | 2.00 | 1.14 | 1.34 | 0.89 |
| Lasso | 3.57 | 2.11 | 3.13 | 1.39 | 1.63 | 1.53 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.17 | 1.73 | 0.25 | 1.19 | 1.30 | 0.09 |
| MaFuller | 1.17 | 1.73 | 0.25 | 1.18 | 1.30 | 0.10 |
| Ma2SLS | 1.27 | 1.81 | 0.26 | 1.18 | 1.29 | 0.27 |
| JIVE | 214848.36 | 24.27 | 323205.09 | 1.14 | 1.66 | 1.04 |
| $R_f^2 = 0.01$; n=500 | | | | | | |
| 2SLS | 1.06 | 1.25 | 0.10 | 1.03 | 1.05 | 0.57 |
| Liml | 1802.37 | 1.32 | 5257.55 | 0.83 | 1.45 | 5.83 |
| Fuller | 2.69 | 1.01 | 5.90 | 0.86 | 1.11 | 4.45 |
| Lasso | 4.86 | 2.23 | 4.68 | 1.58 | 1.90 | 2.45 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.10 | 1.28 | 0.09 | 1.06 | 1.09 | 0.17 |
| MaFuller | 1.10 | 1.27 | 0.09 | 1.06 | 1.09 | 0.16 |
| Ma2SLS | 1.08 | 1.26 | 0.09 | 1.03 | 1.05 | 0.68 |
| JIVE | 8844.86 | 6.32 | 25740.39 | 1.09 | 1.46 | 4.60 |

**Note 1**: The values in the table are computed as $\frac{\text{Estimator Measure}}{\text{Post Lasso Measure}}$ to capture the estimators' relative performance with respect to the Post-$l_1$-Penalized estimator.

**Note 2**: DGP 1 is defined by by a normal distribution with the variance covariance matrix (1.19), with heteroskedasticity introduced via $\tilde{u}_n = u_n + Z_{n1}' Z_{n1}$. $\sigma_{vu} = .95$ defines a high level of endogeneity and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010). The $\Pi_n$ was constructed following $\pi_{nj} = c\left(p_n\right)\left(1 - \frac{j - p_n/2}{p_n/2+1}\right)^4$ for $j \leq p_n/2$ and $\pi_{nj} = 0$ for $j > p_n/2$.

**Note 3**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.12: DGP 1: High and Low Endogeneity Levels Using $\Pi_n$ of equation (1.21)

| $\sigma_{uv} = 0.15$; n=100 | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|
| 2SLS | 1.03 | 1.19 | 0.87 | 1.03 | 1.06 | 0.98 |
| Liml | 1.70 | 0.09 | 2.37 | 0.75 | 1.09 | 1.45 |
| Fuller | 1.47 | 0.02 | 2.06 | 0.78 | 1.05 | 1.40 |
| Lasso | 1.41 | 1.22 | 1.37 | 1.03 | 1.13 | 1.11 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 2.36 | 1.40 | 2.52 | 1.38 | 2.08 | 0.63 |
| MaFuller | 2.35 | 1.39 | 2.52 | 1.38 | 2.09 | 0.62 |
| Ma2SLS | 1.08 | 1.60 | 2.53 | 1.02 | 1.06 | 1.01 |
| JIVE | 6.19 | 0.54 | 8.54 | 0.71 | 1.18 | 1.64 |
| $\sigma_{uv} = 0.95$; n=100 | | | | | | |
| 2SLS | 1.16 | 1.19 | 0.70 | 1.07 | 1.14 | 0.91 |
| Liml | 5.07 | 0.09 | 13.79 | 0.61 | 0.71 | 1.65 |
| Fuller | 0.97 | 0.01 | 2.65 | 0.64 | 0.68 | 1.55 |
| Lasso | 1.66 | 1.29 | 1.63 | 1.07 | 1.17 | 1.19 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 3.37 | 1.99 | 2.37 | 1.47 | 2.11 | 0.51 |
| MaFuller | 3.36 | 1.98 | 2.37 | 1.47 | 2.11 | 0.51 |
| Ma2SLS | 1.13 | 0.61 | 2.44 | 1.04 | 1.08 | 0.95 |
| JIVE | 121.25 | 0.71 | 329.24 | 0.54 | 0.87 | 2.13 |
| $\sigma_{uv} = 0.15$; n=500 | | | | | | |
| 2SLS | 1.34 | 1.44 | 0.82 | 1.09 | 1.26 | 0.93 |
| Liml | 1.05 | 0.03 | 1.80 | 0.81 | 0.86 | 1.29 |
| Fuller | 1.03 | 0.01 | 1.76 | 0.81 | 0.84 | 1.28 |
| Lasso | 1.28 | 1.13 | 1.27 | 1.02 | 1.10 | 1.12 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 7.01 | 3.89 | 1.21 | 1.63 | 3.59 | 0.45 |
| MaFuller | 6.99 | 3.89 | 1.21 | 1.63 | 3.57 | 0.46 |
| Ma2SLS | 1.18 | 1.04 | 1.25 | 1.05 | 1.16 | 0.99 |
| JIVE | 1.24 | 0.16 | 2.10 | 0.78 | 0.93 | 1.48 |
| $\sigma_{uv} = 0.95$; n=500 | | | | | | |
| 2SLS | 1.53 | 1.34 | 0.63 | 1.12 | 1.32 | 0.77 |
| Liml | 0.51 | 0.06 | 2.21 | 0.63 | 0.48 | 1.34 |
| Fuller | 0.48 | 0.01 | 2.08 | 0.65 | 0.47 | 1.34 |
| Lasso | 1.30 | 1.13 | 1.34 | 1.04 | 1.12 | 1.12 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 6.71 | 2.90 | 1.08 | 1.70 | 2.88 | 0.32 |
| MaFuller | 6.67 | 2.89 | 1.08 | 1.70 | 2.87 | 0.33 |
| Ma2SLS | 1.12 | 1.05 | 1.16 | 1.03 | 1.10 | 0.87 |
| JIVE | 2.16 | 0.27 | 9.14 | 0.60 | 0.61 | 1.70 |

**Note 1**: The values in the table are computed as $\frac{\text{Estimator Measure}}{\text{Post Lasso Measure}}$ to capture the estimators' relative performance with respect to the Post-$l_1$-Penalized estimator.

**Note 2**: DGP 1 is defined by by a normal distribution with the variance covariance matrix (1.19), with heteroskedasticity introduced via $\tilde{u}_n = u_n + Z'_{n1}Z_{n1}$. $\sigma_{vu} = .95$ defines a high level of endogeneity, $\sigma_{vu} = .15$ defines a low level of endogeneity, and $R^2_f$ is the pseudo R-squared of Kuersteiner and Okui (2010). The $\Pi_n$ was constructed following $\Pi_n = \left(\mathbf{0}_{pn-8}, \frac{1}{\sqrt{n}}, \frac{-1}{\sqrt{n}}, \frac{-1}{\log(n)}, \frac{1}{\log(n)}, \frac{1}{\sqrt{\log(n)}}, \frac{-1}{\sqrt{\log(n)}}, \frac{1}{n^{1/3}}, \frac{-1}{n^{1/3}}\right)$.

**Note 3**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.13: DGP 2: Low Endogeneity Level Using $\Pi_n$ of equation (1.20)

| $R_f^2 = 0.1$; n=100 | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|
| 2SLS | 1.00 | 1.33 | 0.20 | 1.06 | 1.02 | 0.67 |
| Liml | 260.13 | 0.75 | 532.74 | 0.67 | 0.97 | 2.84 |
| Fuller | 1.38 | 0.62 | 2.42 | 0.73 | 0.75 | 2.17 |
| Lasso | 4.08 | 1.97 | 4.30 | 1.29 | 1.47 | 1.75 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.28 | 1.57 | 0.06 | 1.16 | 1.21 | 0.35 |
| MaFuller | 1.28 | 1.56 | 0.06 | 1.15 | 1.21 | 0.34 |
| Ma2SLS | 0.99 | 1.36 | 0.07 | 1.03 | 0.96 | 0.78 |
| JIVE | 446.81 | 2.02 | 911.78 | 0.95 | 1.39 | 3.51 |
| $R_f^2 = 0.1$; n=500 | | | | | | |
| 2SLS | 1.13 | 1.08 | 0.66 | 1.03 | 1.07 | 0.86 |
| Liml | 0.55 | 0.15 | 5.96 | 0.58 | 0.42 | 2.22 |
| Fuller | 0.45 | 0.07 | 4.97 | 0.60 | 0.41 | 2.09 |
| Lasso | 2.62 | 1.61 | 2.93 | 1.22 | 1.53 | 1.56 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 2.64 | 1.70 | 0.12 | 1.27 | 1.66 | 0.38 |
| MaFuller | 2.64 | 1.70 | 0.12 | 1.27 | 1.66 | 0.37 |
| Ma2SLS | 1.00 | 1.04 | 0.16 | 1.00 | 0.99 | 0.96 |
| JIVE | 7.67 | 0.45 | 84.06 | 0.53 | 0.56 | 3.16 |
| $R_f^2 = 0.01$; n=100 | | | | | | |
| 2SLS | 1.11 | 2.19 | 0.09 | 1.12 | 1.23 | 0.18 |
| Liml | 3431.54 | 0.63 | 4379.85 | 1.03 | 1.83 | 1.15 |
| Fuller | 2.00 | 2.05 | 1.39 | 1.06 | 1.18 | 0.75 |
| Lasso | 3.76 | 2.48 | 3.09 | 1.39 | 1.76 | 1.51 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.08 | 2.22 | 0.02 | 1.13 | 1.25 | 0.09 |
| MaFuller | 1.08 | 2.22 | 0.02 | 1.13 | 1.25 | 0.09 |
| Ma2SLS | 1.26 | 2.39 | 0.02 | 1.12 | 1.24 | 0.21 |
| JIVE | 244.82 | 2.89 | 310.17 | 1.15 | 1.63 | 0.79 |
| $R_f^2 = 0.01$; n=500 | | | | | | |
| 2SLS | 1.00 | 1.21 | 0.08 | 1.01 | 1.02 | 0.66 |
| Liml | 178.02 | 1.05 | 539.08 | 0.80 | 1.17 | 5.58 |
| Fuller | 2.20 | 0.91 | 4.99 | 0.83 | 0.92 | 4.49 |
| Lasso | 5.59 | 2.41 | 5.20 | 1.57 | 2.03 | 2.83 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.09 | 1.27 | 0.01 | 1.04 | 1.08 | 0.22 |
| MaFuller | 1.09 | 1.27 | 0.01 | 1.04 | 1.08 | 0.22 |
| Ma2SLS | 1.01 | 1.23 | 0.01 | 1.01 | 1.02 | 0.76 |
| JIVE | 2894.45 | 3.66 | 8774.63 | 1.00 | 1.34 | 5.06 |

**Note 1**: The values in the table are computed as $\frac{\text{Estimator Measure}}{\text{Post Lasso Measure}}$ to capture the estimators' relative performance with respect to the Post-$l_1$-Penalized estimator.

**Note 2**: DGP 2 is defined by by a normal distribution with the variance covariance matrix (1.19) and mean zero. $\sigma_{vu} = .15$ defines a high level of endogeneity and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010). The $\Pi_n$ was constructed following $\pi_{nj} = c\left(p_n\right)\left(1 - \frac{j-p_n/2}{p_n/2+1}\right)^4$ for $j \leq p_n/2$ and $\pi_{nj} = 0$ for $j > p_n/2$.

**Note 3**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.14: DGP 2: High Endogeneity Level Using $\Pi_n$ of equation (1.20)

| $R_f^2 = 0.1$; n=100 | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|
| 2SLS | 1.17 | 1.40 | 0.02 | 1.04 | 1.06 | 0.42 |
| Liml | 201.12 | 0.29 | 487.11 | 0.45 | 0.43 | 3.21 |
| Fuller | 0.31 | 0.54 | 0.34 | 0.61 | 0.37 | 1.14 |
| Lasso | 3.42 | 1.85 | 3.40 | 1.31 | 1.51 | 2.87 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.48 | 1.59 | 0.00 | 1.12 | 1.20 | 0.14 |
| MaFuller | 1.48 | 1.59 | 0.00 | 1.12 | 1.20 | 0.14 |
| Ma2SLS | 1.08 | 1.35 | 0.01 | 1.01 | 1.01 | 0.58 |
| JIVE | 717.56 | 0.99 | 1736.92 | 0.92 | 1.12 | 4.94 |
| $R_f^2 = 0.1$; n=500 | | | | | | |
| 2SLS | 1.10 | 1.08 | 0.12 | 1.03 | 1.05 | 0.67 |
| Liml | 0.18 | 0.10 | 2.94 | 0.42 | 0.22 | 2.79 |
| Fuller | 0.11 | 0.01 | 1.88 | 0.46 | 0.20 | 2.43 |
| Lasso | 3.02 | 1.68 | 6.09 | 1.31 | 1.55 | 3.06 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 2.09 | 1.49 | 0.01 | 1.25 | 1.45 | 0.15 |
| MaFuller | 2.09 | 1.49 | 0.01 | 1.25 | 1.45 | 0.15 |
| Ma2SLS | 0.96 | 1.01 | 0.03 | 0.99 | 0.98 | 0.95 |
| JIVE | 201.42 | 0.61 | 3480.60 | 0.41 | 0.50 | 7.48 |
| $R_f^2 = 0.01$; n=100 | | | | | | |
| 2SLS | 1.30 | 1.84 | 0.00 | 1.02 | 1.03 | 0.03 |
| Liml | 724.68 | 0.02 | 1173.33 | 0.82 | 0.82 | 0.71 |
| Fuller | 1.16 | 1.70 | 0.10 | 0.91 | 0.86 | 0.23 |
| Lasso | 3.57 | 2.10 | 3.05 | 1.33 | 1.50 | 1.60 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.33 | 1.86 | 0.00 | 1.03 | 1.05 | 0.01 |
| MaFuller | 1.33 | 1.87 | 0.00 | 1.03 | 1.05 | 0.01 |
| Ma2SLS | 1.28 | 1.83 | 0.00 | 1.02 | 1.02 | 0.04 |
| JIVE | 91.28 | 2.42 | 144.17 | 1.03 | 1.06 | 0.15 |
| $R_f^2 = 0.01$; n=500 | | | | | | |
| 2SLS | 1.11 | 1.24 | 0.00 | 1.01 | 1.01 | 0.52 |
| Liml | 296.98 | 0.86 | 1055.97 | 0.42 | 0.45 | 11.95 |
| Fuller | 0.43 | 0.66 | 0.44 | 0.63 | 0.43 | 2.99 |
| Lasso | 4.98 | 2.23 | 4.97 | 1.55 | 1.84 | 10.52 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.19 | 1.28 | 0.00 | 1.03 | 1.05 | 0.07 |
| MaFuller | 1.19 | 1.29 | 0.00 | 1.03 | 1.05 | 0.08 |
| Ma2SLS | 1.09 | 1.23 | 0.00 | 1.00 | 1.00 | 0.69 |
| JIVE | 32.00 | 1.27 | 109.86 | 1.01 | 1.07 | 7.25 |

**Note 1**: The values in the table are computed as $\frac{\text{Estimator Measure}}{\text{Post Lasso Measure}}$ to capture the estimators' relative performance with respect to the Post-$l_1$-Penalized estimator.

**Note 2**: DGP 2 is defined by by a normal distribution with the variance covariance matrix (1.19). $\sigma_{vu} = .95$ defines a high level of endogeneity and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010). The $\Pi_n$ was constructed following $\pi_{nj} = c\left(p_n\right)\left(1 - \frac{j - p_n/2}{p_n/2+1}\right)^4$ for $j \leq p_n/2$ and $\pi_{nj} = 0$ for $j > p_n/2$.

**Note 3**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.15: DGP 2: High and Low Endogeneity Levels Using $\Pi_n$ of equation (1.21)

| $\sigma_{uv} = 0.15$; n=100 | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|
| 2SLS | 1.19 | 1.20 | 0.80 | 1.04 | 1.16 | 0.93 |
| Liml | 0.86 | 0.13 | 2.14 | 0.75 | 0.65 | 1.38 |
| Fuller | 0.75 | 0.02 | 1.89 | 0.78 | 0.64 | 1.32 |
| Lasso | 1.59 | 1.23 | 1.69 | 1.03 | 1.11 | 1.30 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 4.81 | 2.75 | 0.60 | 1.39 | 2.56 | 0.64 |
| MaFuller | 4.79 | 2.75 | 0.60 | 1.39 | 2.55 | 0.64 |
| Ma2SLS | 1.12 | 1.19 | 0.70 | 1.02 | 1.09 | 0.98 |
| JIVE | 14.26 | 0.57 | 35.41 | 0.74 | 0.76 | 1.65 |
| $\sigma_{uv} = 0.95$; n=100 | | | | | | |
| 2SLS | 1.24 | 1.18 | 0.49 | 1.06 | 1.14 | 0.78 |
| Liml | 0.40 | 0.14 | 2.46 | 0.58 | 0.36 | 1.64 |
| Fuller | 0.28 | 0.00 | 1.81 | 0.63 | 0.36 | 1.50 |
| Lasso | 1.76 | 1.30 | 2.21 | 1.08 | 1.19 | 1.46 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 4.45 | 2.29 | 0.25 | 1.48 | 2.18 | 0.43 |
| MaFuller | 4.44 | 2.28 | 0.24 | 1.48 | 2.18 | 0.42 |
| Ma2SLS | 1.16 | 1.14 | 0.47 | 1.03 | 1.08 | 0.84 |
| JIVE | 120.79 | 0.61 | 775.89 | 0.54 | 0.57 | 2.98 |
| $\sigma_{uv} = 0.15$; n=500 | | | | | | |
| 2SLS | 1.69 | 1.44 | 0.68 | 1.09 | 1.42 | 0.88 |
| Liml | 0.46 | 0.03 | 1.66 | 0.79 | 0.51 | 1.30 |
| Fuller | 0.45 | 0.01 | 1.60 | 0.80 | 0.51 | 1.28 |
| Lasso | 1.29 | 1.13 | 1.30 | 1.02 | 1.11 | 1.14 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 12.10 | 4.08 | 0.23 | 1.63 | 3.99 | 0.42 |
| MaFuller | 12.08 | 4.08 | 0.23 | 1.62 | 3.98 | 0.43 |
| Ma2SLS | 1.23 | 1.26 | 0.31 | 1.03 | 1.16 | 0.96 |
| JIVE | 0.68 | 0.15 | 2.39 | 0.78 | 0.59 | 1.50 |
| $\sigma_{uv} = 0.95$; n=500 | | | | | | |
| 2SLS | 1.64 | 1.34 | 0.29 | 1.12 | 1.32 | 0.57 |
| Liml | 0.16 | 0.04 | 1.57 | 0.63 | 0.26 | 1.28 |
| Fuller | 0.15 | 0.01 | 1.43 | 0.65 | 0.26 | 1.23 |
| Lasso | 1.28 | 1.13 | 1.24 | 1.04 | 1.11 | 1.13 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 7.81 | 2.95 | 0.05 | 1.70 | 2.86 | 0.19 |
| MaFuller | 7.78 | 2.94 | 0.05 | 1.70 | 2.86 | 0.20 |
| Ma2SLS | 1.14 | 1.12 | 0.19 | 1.03 | 1.09 | 0.80 |
| JIVE | 2.36 | 0.27 | 22.04 | 0.61 | 0.38 | 2.08 |

**Note 1**: The values in the table are computed as $\frac{\text{Estimator Measure}}{\text{Post Lasso Measure}}$ to capture the estimators' relative performance with respect to the Post-$l_1$-Penalized estimator.
**Note 2**: DGP 2 is defined by by a normal distribution with the variance covariance matrix (1.19). $\sigma_{vu} = .95$ defines a high level of endogeneity while $\sigma_{vu} = .15$ defines a low level of endogeneity and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010). The $\Pi_n$ was constructed following $\Pi_n = \left(\mathbf{0}_{pn-8}, \frac{1}{\sqrt{n}}, \frac{-1}{\sqrt{n}}, \frac{-1}{\log(n)}, \frac{1}{\log(n)}, \frac{1}{\sqrt{\log(n)}}, \frac{-1}{\sqrt{\log(n)}}, \frac{1}{n^{1/3}}, \frac{-1}{n^{1/3}}\right)$.
**Note 3**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.16: DGP 3: Low Endogeneity Level Using $\Pi_n$ of equation (1.20)

| $R_f^2 = 0.1$; n=100 | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|
| 2SLS | 0.57 | 8.68 | 0.48 | 1.14 | 0.74 | 0.69 |
| Liml | 1025.09 | 57.02 | 1022.44 | 1.00 | 2.29 | 2.33 |
| Fuller | 6.43 | 5.10 | 6.41 | 1.02 | 1.98 | 2.02 |
| Lasso | 3.31 | 13.62 | 3.09 | 1.13 | 1.54 | 1.39 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 0.32 | 12.79 | 0.13 | 1.16 | 0.57 | 0.43 |
| MaFuller | 0.32 | 12.77 | 0.13 | 1.17 | 0.58 | 0.43 |
| Ma2SLS | 0.71 | 21.57 | 0.15 | 1.15 | 0.79 | 0.75 |
| JIVE | 3373.93 | 52.58 | 3374.65 | 1.14 | 2.61 | 2.56 |
| $R_f^2 = 0.1$; n=500 | | | | | | |
| 2SLS | 0.68 | 1.19 | 0.57 | 1.01 | 0.89 | 0.80 |
| Liml | 53.87 | 0.56 | 61.86 | 0.83 | 2.08 | 2.35 |
| Fuller | 11.47 | 0.01 | 13.18 | 0.83 | 2.04 | 2.30 |
| Lasso | 3.83 | 2.96 | 3.09 | 1.20 | 1.72 | 1.37 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 0.48 | 1.70 | 0.12 | 1.06 | 0.93 | 0.43 |
| MaFuller | 0.48 | 1.70 | 0.12 | 1.06 | 0.93 | 0.43 |
| Ma2SLS | 0.77 | 2.25 | 0.13 | 1.00 | 0.91 | 0.87 |
| JIVE | 2388.55 | 2.57 | 2743.80 | 0.82 | 2.20 | 2.45 |
| $R_f^2 = 0.01$; n=100 | | | | | | |
| 2SLS | 0.54 | 1.92 | 0.45 | 1.66 | 0.50 | 0.53 |
| Liml | 44832.31 | 14.37 | 45983.47 | 1.63 | 2.16 | 2.61 |
| Fuller | 6.30 | 1.87 | 6.38 | 1.66 | 1.70 | 1.99 |
| Lasso | 3.82 | 1.77 | 3.83 | 1.26 | 1.14 | 1.35 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 0.26 | 2.65 | 0.09 | 1.69 | 0.38 | 0.29 |
| MaFuller | 0.27 | 2.70 | 0.09 | 1.69 | 0.37 | 0.29 |
| Ma2SLS | 0.64 | 4.69 | 0.09 | 1.66 | 0.52 | 0.59 |
| JIVE | 6079.38 | 2.91 | 6235.98 | 1.79 | 1.79 | 2.08 |
| $R_f^2 = 0.01$; n=500 | | | | | | |
| 2SLS | 0.46 | 2.59 | 0.31 | 1.06 | 0.69 | 0.58 |
| Liml | 1674.74 | 0.59 | 1714.69 | 0.87 | 3.73 | 4.25 |
| Fuller | 17.55 | 1.54 | 17.91 | 0.89 | 3.45 | 3.84 |
| Lasso | 14.91 | 8.39 | 13.59 | 1.43 | 2.20 | 1.83 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 0.25 | 2.98 | 0.04 | 1.08 | 0.65 | 0.26 |
| MaFuller | 0.24 | 2.96 | 0.04 | 1.08 | 0.65 | 0.25 |
| Ma2SLS | 0.51 | 4.49 | 0.04 | 1.07 | 0.75 | 0.63 |
| JIVE | 2652.64 | 5.92 | 2715.08 | 0.96 | 3.33 | 3.88 |

**Note 1**: The values in the table are computed as $\frac{\text{Estimator Measure}}{\text{Post Lasso Measure}}$ to capture the estimators' relative performance with respect to the Post-$l_1$-Penalized estimator.

**Note 2**: DGP 3 is defined by a zero mean t-distribution with 5 degrees of freedom using the variance covariance matrix (1.19). $\sigma_{vu} = .15$ defines a low level of endogeneity and $R_f^2$ and is the pseudo R-squared of Kuersteiner and Okui (2010). The $\Pi_n$ was constructed following $\pi_{nj} = c\,(p_n)\left(1 - \frac{j-p_n/2}{p_n/2+1}\right)^4$ for $j \leq p_n/2$ and $\pi_{nj} = 0$ for $j > p_n/2$.

**Note 3**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.17: DGP 3: High Endogeneity Level Using $\Pi_n$ of equation (1.20)

| $R_f^2 = 0.1$; n=100 | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|
| 2SLS | 1.15 | 1.46 | 0.06 | 1.05 | 1.09 | 0.44 |
| Liml | 481.76 | 0.20 | 1020.12 | 0.51 | 0.56 | 2.37 |
| Fuller | 0.75 | 0.48 | 1.32 | 0.60 | 0.46 | 1.53 |
| Lasso | 3.83 | 1.71 | 4.82 | 1.16 | 1.27 | 2.46 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.41 | 1.63 | 0.01 | 1.13 | 1.22 | 0.25 |
| MaFuller | 1.41 | 1.63 | 0.01 | 1.13 | 1.22 | 0.24 |
| Ma2SLS | 1.07 | 1.42 | 0.02 | 1.02 | 1.04 | 0.52 |
| JIVE | 4448.35 | 5.72 | 9383.19 | 0.93 | 1.17 | 3.56 |
| $R_f^2 = 0.1$; n=500 | | | | | | |
| 2SLS | 1.07 | 1.05 | 0.32 | 1.03 | 1.05 | 0.73 |
| Liml | 0.52 | 0.14 | 8.83 | 0.46 | 0.28 | 2.46 |
| Fuller | 0.33 | 0.08 | 5.61 | 0.48 | 0.27 | 2.32 |
| Lasso | 2.39 | 1.44 | 7.67 | 1.18 | 1.34 | 2.38 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.81 | 1.38 | 0.04 | 1.20 | 1.38 | 0.29 |
| MaFuller | 1.81 | 1.39 | 0.04 | 1.20 | 1.38 | 0.29 |
| Ma2SLS | 0.96 | 1.00 | 0.06 | 0.99 | 0.99 | 0.87 |
| JIVE | 2117.46 | 0.18 | 36896.15 | 0.44 | 0.49 | 4.84 |
| $R_f^2 = 0.01$; n=100 | | | | | | |
| 2SLS | 1.29 | 1.98 | 0.02 | 1.06 | 1.09 | 0.08 |
| Liml | 334.77 | 2.08 | 492.90 | 0.99 | 1.11 | 0.59 |
| Fuller | 1.47 | 1.89 | 0.46 | 1.00 | 1.01 | 0.42 |
| Lasso | 3.45 | 1.98 | 3.21 | 1.19 | 1.30 | 1.46 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.44 | 2.10 | 0.02 | 1.07 | 1.10 | 0.05 |
| MaFuller | 1.32 | 2.02 | 0.01 | 1.07 | 1.10 | 0.05 |
| Ma2SLS | 1.28 | 1.99 | 0.01 | 1.06 | 1.09 | 0.10 |
| JIVE | 211.92 | 0.82 | 313.01 | 1.07 | 1.15 | 0.35 |
| $R_f^2 = 0.01$; n=500 | | | | | | |
| 2SLS | 1.09 | 1.23 | 0.02 | 1.02 | 1.03 | 0.59 |
| Liml | 235.60 | 0.32 | 829.45 | 0.70 | 0.78 | 7.26 |
| Fuller | 1.18 | 0.87 | 2.23 | 0.75 | 0.63 | 5.17 |
| Lasso | 70.17 | 2.28 | 234.02 | 1.42 | 1.66 | 5.25 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.16 | 1.27 | 0.00 | 1.04 | 1.06 | 0.24 |
| MaFuller | 1.16 | 1.27 | 0.00 | 1.04 | 1.06 | 0.24 |
| Ma2SLS | 1.08 | 1.23 | 0.00 | 1.01 | 1.02 | 0.64 |
| JIVE | 1753.49 | 2.63 | 6157.66 | 1.01 | 1.09 | 4.50 |

**Note 1**: The values in the table are computed as $\frac{\text{Estimator Measure}}{\text{Post Lasso Measure}}$ to capture the estimators' relative performance with respect to the Post-$l_1$-Penalized estimator.

**Note 2**: DGP 3 is defined by a zero mean t-distribution with 5 degrees of freedom using the variance covariance matrix (1.19). $\sigma_{vu} = .95$ defines a high level of endogeneity and $R_f^2$ and is the pseudo R-squared of Kuersteiner and Okui (2010). The $\Pi_n$ was constructed following $\pi_{nj} = c\left(p_n\right)\left(1 - \frac{j - p_n/2}{p_n/2+1}\right)^4$ for $j \leq p_n/2$ and $\pi_{nj} = 0$ for $j > p_n/2$.

**Note 3**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

Table 1.18: DGP 3: High and Low Endogeneity Levels Using $\Pi_n$ of equation (1.21)

| $\sigma_{uv} = 0.15$; n=100 | MSE | BIAS | VAR | M.BIAS | MAD | IQR |
|---|---|---|---|---|---|---|
| 2SLS | 0.73 | 1.26 | 0.70 | 1.01 | 0.95 | 0.87 |
| Liml | 3237.05 | 9.17 | 3343.69 | 0.90 | 1.39 | 1.38 |
| Fuller | 2.57 | 0.23 | 2.66 | 0.90 | 1.38 | 1.35 |
| Lasso | 2.48 | 2.11 | 2.41 | 1.02 | 1.31 | 1.30 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 0.70 | 2.65 | 0.48 | 1.08 | 0.94 | 0.71 |
| MaFuller | 0.69 | 2.60 | 0.48 | 1.09 | 0.95 | 0.70 |
| Ma2SLS | 0.77 | 2.99 | 0.49 | 1.01 | 0.94 | 0.88 |
| JIVE | 271.04 | 3.93 | 279.69 | 0.92 | 1.45 | 1.42 |
| $\sigma_{uv} = 0.95$; n=100 | | | | | | |
| 2SLS | 1.16 | 1.19 | 0.55 | 1.04 | 1.11 | 0.78 |
| Liml | 1.00 | 0.17 | 3.27 | 0.64 | 0.45 | 1.24 |
| Fuller | 0.57 | 0.08 | 1.90 | 0.66 | 0.44 | 1.18 |
| Lasso | 9.46 | 1.23 | 28.14 | 0.98 | 0.97 | 1.35 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 3.56 | 2.22 | 0.31 | 1.39 | 2.09 | 0.56 |
| MaFuller | 3.54 | 2.22 | 0.31 | 1.39 | 2.09 | 0.58 |
| Ma2SLS | 1.13 | 1.20 | 0.39 | 1.03 | 1.07 | 0.82 |
| JIVE | 510.93 | 0.69 | 1710.91 | 0.61 | 0.64 | 2.12 |
| $\sigma_{uv} = 0.15$; n=500 | | | | | | |
| 2SLS | 0.84 | 1.37 | 0.74 | 1.03 | 1.01 | 0.91 |
| Liml | 2.55 | 0.10 | 2.80 | 0.93 | 1.34 | 1.50 |
| Fuller | 2.47 | 0.07 | 2.71 | 0.93 | 1.33 | 1.49 |
| Lasso | 1.76 | 1.72 | 1.65 | 1.05 | 1.28 | 1.23 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 1.11 | 3.06 | 0.31 | 1.13 | 1.41 | 0.61 |
| MaFuller | 1.10 | 3.04 | 0.31 | 1.13 | 1.39 | 0.61 |
| Ma2SLS | 0.83 | 2.48 | 0.32 | 1.02 | 0.98 | 0.95 |
| JIVE | 4.63 | 0.33 | 5.06 | 0.92 | 1.34 | 1.47 |
| $\sigma_{uv} = 0.95$; n=500 | | | | | | |
| 2SLS | 1.64 | 1.37 | 0.45 | 1.12 | 1.34 | 0.72 |
| Liml | 0.25 | 0.07 | 1.44 | 0.67 | 0.32 | 1.22 |
| Fuller | 0.23 | 0.04 | 1.39 | 0.68 | 0.32 | 1.20 |
| Lasso | 1.09 | 1.01 | 1.37 | 1.00 | 0.99 | 1.24 |
| **Post Lasso** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| MaLiml | 6.69 | 2.83 | 0.11 | 1.61 | 2.79 | 0.36 |
| MaFuller | 6.68 | 2.83 | 0.11 | 1.61 | 2.79 | 0.36 |
| Ma2SLS | 1.32 | 1.25 | 0.19 | 1.06 | 1.18 | 0.81 |
| JIVE | 8.20 | 0.35 | 48.22 | 0.64 | 0.42 | 1.69 |

**Note 1**: The values in the table are computed as $\frac{\text{Estimator Measure}}{\text{Post Lasso Measure}}$ to capture the estimators' relative performance with respect to the Post-$l_1$-Penalized estimator.

**Note 2**: DGP 3 is defined by a zero mean t-distribution with 5 degrees of freedom using the variance covariance matrix (1.19). $\sigma_{vu} = .95$ defines a high level of endogeneity while $\sigma_{vu} = .15$ defines a low level of endogeneity and $R_f^2$ is the pseudo R-squared of Kuersteiner and Okui (2010). The $\Pi_n$ was constructed following $\Pi_n = \left(\mathbf{0}_{pn-8}, \frac{1}{\sqrt{n}}, \frac{-1}{\sqrt{n}}, \frac{-1}{\log(n)}, \frac{1}{\log(n)}, \frac{1}{\sqrt{\log(n)}}, \frac{-1}{\sqrt{\log(n)}}, \frac{1}{n^{1/3}}, \frac{-1}{n^{1/3}}\right)$.

**Note 3**: MSE(Mean Square Error); VAR(Variance); M.BIAS(Median Bias); MAD(Median Absolute Deviation); IQR(Interquartile Range).

# Chapter 2

# Angrist and Krueger (1991) Using a Post Adaptive LASSO

## 2.1. Introduction

The most commonly cited example of an instrumental variables specification in the presence of many weak instruments is Angrist and Krueger (1991). It is therefore important to be able to determine the performance of any procedure that addresses the many weak instruments problem with respect to the Angrist and Krueger (1991) paper. In this chapter of my dissertation I show the performance of the estimator I introduced in the first chapter and compare the results of my estimation with those encountered by Belloni et al. (2010a), which is the paper that most closely resembles mine. I also include the results that would arise if the traditional simultaneous equations estimators were used.

The problem faced by Angrist and Krueger (1991) is to determine the impact of schooling on wages. As is usually the case when considering a wage equation, the endogeneity of schooling is a motive for concern. Specifically, unobserved factors may influence both educational attainment and earnings, resulting in a biased coefficient in an Ordinary Least Squares regression of schooling on wages. Angrist and Krueger (1991) argue that an individual's quarter of birth is a valid instrument in this scenario. Their claim is that the quarter of birth affects the years of schooling of an individual due to specific state regulations. For instance, if state legislation determines that anybody that is four years of age by June 30 of a particular year can enroll in school, those born in June 31 will have one less year of compulsory schooling than those born on June 30. The fundamental intuition behind the quarter of birth as an instrument is that it affects the years of schooling of an individual but should not be a direct determinant of wage. Thus, it satisfies the exogeneity condition that is demanded from a valid instrument.

## 2.2. Background

### 2.2.1 Returns to Education

The economic literature on the returns to schooling arises from the work of Becker (1964) and Mincer (1958). The building block of the original theory is that individuals maximize lifetime earnings by selecting how much education to attain. The early empirical literature, however, starts with the work of Mincer (1974) who devise a production function representation of the relationship between accumulated skills, viewed as an output, and education, experience and abilities, viewed as inputs. This idea is summarized in what is known in the literature as the Mincer wage regression, proposed by Mincer (1974):

$$log\,(W_t) = \phi_0 + \phi_1(S_t) + \phi_2(Exp_t) + \varepsilon_t \tag{2.1}$$

$\phi_1(S_t)$ represents the effects of schooling, $\phi_2(Exp_t)$ represents the effects of post-schooling decisions, usually approximated by experience, and $\varepsilon_t$ can be understood as an idiosyncratic productivity shock.

The estimation of the returns to schooling using the Mincerian wage regression (2.1) became one of the most widely analyzed topics in applied econometrics. Griliches (1977) discusses numerous econometric problems in estimating the returns to schooling. In particular, problems arise related to the measurement of both schooling and ability. More importantly, he suggests that the endogeneity of schooling decisions is a serious problem that constitutes an obstacle in establishing the causal effect of education on earnings. The main concern is the correlation between unobserved individual characteristics (predominantly the ability of individuals) and schooling choices.

The work of Angrist and Krueger (1991) can then be viewed as instrumental variable approach to estimating the Mincer equation that tries to address this endogeneity problem. The instruments in this framework originate from natural experiments in the compulsory education and enrollment legislation in each state. Also, Angrist and Krueger (1991) can be understood as a static model, in the spirit of Becker (1964), in which once education decisions are taken individuals start working and there are no dynamics that allow for individuals to transition between the labor force and educational institutions. The previous consideration is one of the reasons for an additional literature on the structural modeling of schooling decisions pioneered by Keane and Wolpin (1997).

There thus exist a structural dynamic approach and a static reduced form approach that yields two different sets of estimates of the returns to schooling. As Belzil (2006) mentions, instrumental variable estimators of the returns to education fall in a range of a 10 to 15 percent increase the in weekly wage for an additional year of schooling. On the other hand, structural dynamic models, as discussed by Belzil (2006), find a range of 4 to 7 percent increases in the weekly wage resulting from an additional year of schooling. In this chapter, I present results of an instrumental variable estimation but do not take a stance on which is the preferred way of modeling the problem as the literature has not settled this debate.

### 2.2.2 Angrist and Krueger (1991)

The data for the computations in the Angrist and Krueger (1991) paper comes from the 1970 and 1980 census. For the 1970 census a 1 percent sample of white and black men born between 1920 and 1929 is used. For the 1980 census a 5 percent sample of white and black men born between 1930 and 1939 is used. In both cases only data for individuals that have positive earnings are used. Furthermore, observations that were imputed by the Census Bureau were dropped.[13] For my estimations I use the data as is provided by Angrist on his website.[14]

The model estimated by Angrist and Krueger (1991) is a cross sectional simultaneous equations model given by:

$$\begin{aligned} wage_i &= \beta_0 + \beta_1 education_i + \beta_2 X_i + \varepsilon_i \\ education_i &= \Pi_0 + \Pi_1 quarterofbirth_i + \Pi_2 X_i + \Pi_3 Z_i + \eta_i \end{aligned} \tag{2.2}$$

In the expression above the subscript $i$ refers to the individual $i$, $wage_i$ is the natural logarithm of weekly wage, $education_i$ are the years of education, $X_i$ are other determinants of wage, $Z_i$ are the excluded instruments, and $quarterofbirth_i$ is the quarter of birth instrument proposed by Angrist and Krueger (1991).

The authors obtain this instrument set by interacting the quarter of birth instrument with the state of birth and year of birth of the individual. Also, they use the state of

[13] A more thorough description of the data set used can be found at the end of Angrist and Krueger (1991)

[14] http://economics.mit.edu/faculty/angrist/data1/data/angkru1991

birth, the region of birth, age, age squared, and a dummy variable for marital status as part of the included instrument set. This gives rise to 180 excluded instruments plus the 62 included instruments coming from state and region of birth dummy variables. As is argued by Bound et al. (1995), and more widely by the many weak instrument literature, these instrument are weak. In the first chapter of my dissertation I demonstrate that in this scenario and under the possibility that some of the instruments are irrelevant, disregarding the irrelevant instruments is fundamental to obtain a consistent estimator in a 2SLS framework. Therefore, my purpose is to disregard the irrelevant instruments and to illustrate the consequences of using a large set of weak and irrelevant instruments.

In table 2.1 I present the result of the traditional simultaneous equations models, two stage least squares (2SLS), limited information maximum likelihood (LIML), and Fuller, and, of ordinary least squares (OLS) for the model in (2.2). Under endogeneity we expect largest biases in estimation to arise under OLS. In the presence of many weak instruments, the weaker the instruments are the closer they are to the OLS estimates. As the results of table 2.1 illustrate, 2SLS, LIML, and Fuller are all close to OLS when the entire instrument set is used. Another important consideration that arises from table 2.1 is that the first stage F-statistic is low, below the rule of thumb value of 10 suggested by Stock et al. (2002) an often cited in applied work as an indication of a weak instrument problem. As I mentioned in the first chapter, the F-statistic is an estimator of the concentration parameter and as such is a measure of the instrument weakness. The results of table 2.1, therefore, suggest the presence of weak instruments and estimates of the return to education that are inconsistent.

Another important reflection comes from the solution of the many weak instrument problem that is proposed by Chao and Swanson (2005a). They recommend that when there is evidence of weak instruments a plausible solution is to include more instruments in the specification. Their results suggest that, even though the set of instruments is weak, the information provided by the additional instruments can help the LIML and Fuller estimators to achieve consistency. The basic intuition that arises from their asymptotic theory is that additional instruments provide a signal asymptotically, that even if small it helps to attain consistency. The practitioner that encounters instrument weakness might then be advised to include extra instruments to address the difficulty at hand. In the first chapter I argue that this is not the case if additional instruments are irrelevant.

Table 2.2 shows a situation a researcher that uses the Angrist and Krueger (1991) data might face. In the case illustrated the researcher uses 153 excluded instruments,coming from quarter of birth interactions with state and quarter of birth, and 62 included instruments, the state dummies and other determinants of wage, and finds that the first stage F-statistic is below 10. The suggestion of Chao and Swanson (2005a) would be to include a larger set of instruments. Table 2.1 illustrates what occurs in the case you include 27 additional excluded instruments that come from interacting quarter of birth with year of birth. The F-statistic decreases from 5.63 to 2.36 and the coefficients move closer to the OLS estimate. The conclusion from this simple example is that by including additional instruments the weak instrument problem can become worse. In the language of chapter 1, including instruments that are irrelevant or are providing more noise than information compounds the difficulties of the researcher. As I demonstrated in chapter 1, the key is the ability to be able to exclude the irrelevant instruments and to include instruments that provide enough information in order to guarantee the possibility that 2SLS, after instrument selection, will be consistent.

## 2.3. Results

In this section I present the results that arise from using the post-adaptive LASSO method proposed in Chapter 1. I also try to ascertain the robustness of the results by revisiting the criticism of Bound et al. (1995). Finally, I describe the computation of the estimator.

Table 2.3 shows that when using the adaptive LASSO on the first stage to select the instruments and then running a 2SLS estimate with the selected regressors as I proposed in Chapter 1. It is important to highlight that there is a reduction from 242 to 166 instruments and that the F-statistic increases to 17.88, which is outside the danger zone for weak instruments stipulated in the literature. As is to be expected, the quarter of birth instruments are included. Angrist and Krueger (1991) do include specifications with only the quarter of birth instruments; it is only when they depart from this specification that the weak instrument problem arises. This suggests that the set of instruments selected by any procedure should be somewhere in between 3 (with only quarter of birth instruments) and 242 (with the entire instrument set discussed above). The uncertainty inherent in deciding exactly which instruments actually belong in the estimation, paired with the weak instrument problem, that makes the estimator in Chapter 1 appealing. The estimate of the return to education in table 2.3 which has a value of 14 percent falls within the range found in similar studies summarized by Belzil (2006).

Belloni et al. (2010a) also study instrument selection in Angrist and Krueger (1991). In Table 5 of Belloni et al. (2010a), the authors show two instrument sets selected by the procedure they propose. The first one, advocated by them as the preferred method, only selects 1 instrument. The second method, a then fold cross-validation procedure which is not presented in the paper, selects 12 instruments including the quarter of birth instruments. Their method to select the regularization parameter appears to impose heavy shrinkage. This also suggests a deeper discussion about the selection of the regularization parameter is needed.

In graph 1 I show the cross-validation function that I use to select the regularization parameter. In this context, I find that cross-validation performed better in simulations than generalized cross-validation which was the proposed method in Fu (1998).

### 2.3.1 Revisiting Bound et al. (1995)

Bound et al. (1995) are skeptical about the results presented in Angrist and Krueger (1991). Their main concern is the fact that education is weakly correlated with quarter of birth. However, they also highlight the possibility that quarter of birth and education might be correlated for reasons other than compulsory education rules and that there might be some correlation between quarter of birth and wages. Therefore, they are not only preoccupied by instrument weakness but by the validity, i.e. exogeneity, of the instrument itself.

Angrist and Krueger (1991) present many specifications, some of which are clearly weaker than others. Bound et al. (1995) focus on the case in which instruments are the weakest in order to illustrate the potential problems that arise in this situation. They generate a simulated quarter of birth variable and estimate the specifications presented in Angrist and Krueger (1991). These instruments, by construction, are uncorrelated with education and are therefore ultimately weak, i.e. irrelevant. Bound et al. (1995) show that under these conditions their results are the same as those in Angrist and Krueger (1991) in the specification in which instrument weakness is highest.

In table 2.4, I repeat the Bound et al. (1995) exercise. I also include the LIML and Fuller estimators. As in Bound et al. (1995) I find that the 2SLS result is extremely close to the OLS result which suggests an extreme level of instrument weakness, as established by

Stock et al. (2002). The same occurs with LIML. In the three cases mentioned previously the coefficient associated with education is positive and statistically significant yet given the weakness of the instruments coefficient is unreliable. For the Fuller estimator, the standard errors are large, suggesting that the instruments, as expected, do not have identifying power.

As with the Fuller estimator, my estimator exhibits a large standard error, see table 2.4. Furthermore, my estimator only selects 10 instruments, in contrast to the case with actual instruments where my estimator selects 166. That is, my estimator method treats the cases with many instruments and with many simulated irrelevant instruments quite differently, suggesting that there is information content in the original instruments, that was not captured by the standard estimators considered in Bound et al. (1995). Yet, at the same time, the F-statistic is smaller than that of the other three estimators, suggesting that the model is ultimately weak. From these results I conclude that in the case where instruments are by construction uncorrelated with the endogenous regressor, my estimator eliminates most of the instruments but, given the small value of the F-statistic, even after instrument selection the instruments do not have much identifying power.

### 2.3.2 Computation of the Adaptive LASSO

In this section, I present the algorithm I use to compute the adaptive LASSO. The adaptive LASSO is computed using the procedure proposed by Fu (1998). This methodology is selected for its ease of implementation.[15] The optimization procedure and results will be illustrated for a single endogenous variable, given that the main concern is to demonstrate the behavior of the estimator in the presence of many instruments. Further, this aids discussion given that many of the finite sample distribution results in the literature are derived under this framework.

Let us define $Z_n \equiv (Z_i, X_i, quarterofbirth)$ where $Z_i$ and $X_i$ and quarter of birth are the first stage instruments presented in (2.2) and $\Pi_n \equiv (\Pi_0, \Pi_1, \Pi_2)$. The first order conditions of the problem defined in equation (1.6) yield, for the $j^{th}$ column of $Z_n$, $Z_n^{(j)}$:

$$2Z_n^{(j)'}Z_n^{(j)}\Pi_{nj} + 2\sum_{i \neq j} Z_n^{(j)'}Z_n^{(i)}\Pi_{ni} - 2Z_n^{(j)'}Y_{2n} = -\lambda_n w_{nj} sign\left(\Pi_{nj}\right)$$

To simplify the exposition the right hand side (RHS), left hand side (LHS), and the regularization parameter in the expression above are redefined to be:

$$\begin{aligned} LHS &= S_j\left(\Pi_{nj}, \Pi_n^{-j}, Z_n, Y_{2n}\right) \\ RHS &= d(\Pi_{nj}, \lambda_n *) \\ \lambda_n * &= \lambda_n w_{nj} \end{aligned}$$

[15] Much has been written about the computation of the LASSO and its variants. A good source for computational aspects concerning the LASSO is http://www-stat.stanford.edu/t̃ibs/lasso.html and the references cited there.

where $\Pi_n^{-j}$ refers to the coefficients that are different from $\Pi_{nj}$.

The algorithm used to compute the proposed adaptive LASSO estimator is:

(i) Start with $\Pi_{n0} = \Pi_{OLS} = \left(\hat{\Pi}_1, \ldots, \hat{\Pi}_{p_n}\right)'$

(ii) At step $m$, for each $j = 1, \ldots, p$, let $S_0 = S_j\left(0, \Pi_n^{-j}, Z_n, Y_{2n}\right)$ and set

$$\hat{\Pi}_{nj} = \begin{cases} \frac{\lambda_n * - S_0}{2Z_n^{(j)'} Z_n^{(j)}} & \text{if } S_0 > \lambda_n * \\ \frac{-\lambda_n * - S_0}{2Z_n^{(j)'} Z_n^{(j)}} & \text{if } S_0 < -\lambda_n * \\ 0 & \text{if } |S_0| \leq \lambda_n * \end{cases}$$

Form a new estimator $\hat{\Pi}_m = \left(\hat{\Pi}_1, \ldots, \hat{\Pi}_{p_n}\right)'$ after updating $\hat{\Pi}_{nj}$

(iii) Repeat $(ii)$ until $\hat{\Pi}_m$ converges.

This procedure selects instruments and at the same time generates an estimator. Here, the selected instruments are used as an input to perform a 2SLS regression as proposed by Belloni and Chernozhukov (2010).

## 2.4. Conclusion

The results in this section highlight the benefits of using the estimator proposed in Chapter 1. As is explicit in table 2.3 the F-statistic increases significantly with respect to the result where all instruments are included, suggesting that some of the instruments provide more noise than signal or are altogether insignificant. In this regard both Belloni et al. (2010a) and this study agree. The estimator thus derived is more reliable as it addresses the many weak instrument problem. However, with respect to the instrument selection procedure my results and those of Belloni et al. (2010a) are inconclusive. This is due to the fact that there exist a considerable difference with the results presented in Belloni et al. (2010a) and there is no definitive optimality criteria to select the regularization parameter. In my case this was determined using cross-validation given my experience with the simulation experiments. Finally, as was mentioned in Belzil (2006) the value of the coefficients in my study falls within the range of values found by similar studies in the literature.

## 2.5. Tables and Graphs

Table 2.1: Return to Education for Men Born 1930-1939: 1980 Census

| | OLS | 2SLS | LIML | Fuller |
|---|---|---|---|---|
| Education | 0.06 | 0.07 | 0.08 | 0.08 |
| Standard Error | (0.00) | (0.01) | (0.02) | (0.02) |
| First Stage F-Statistic | — | 2.36 | 2.36 | 2.36 |
| Number of Instruments | — | 242 | 242 | 242 |
| Number of Observations | 329,509 | 329,509 | 329,509 | 329,509 |

The dependent variable is the log of weekly earnings. The excluded regressors are age, age squared, dummies for race, marital status, state of birth, and quarter of birth, and interactions of quarter of birth with year of birth and state of birth.

Table 2.2: Return to Education for Men Born 1930-1939: 1980 Census. Without Year of Birth time Quarter of Birth Interactions

| | OLS | 2SLS | LIML | Fuller |
|---|---|---|---|---|
| Education | 0.06 | 0.08 | 0.09 | 0.09 |
| Standard Error | (0.00) | (0.02) | (0.02) | (0.02) |
| First Stage F-Statistic | — | 5.63 | 5.63 | 5.63 |
| Number of Instruments | — | 215 | 215 | 215 |
| Number of Observations | 329,509 | 329,509 | 329,509 | 329,509 |

The dependent variable is the log of weekly earnings. The excluded regressors are age, age squared, dummies for race, marital status, state of birth, and quarter of birth, and interactions of quarter of birth with state of birth.

Graph 1:

Cross-Validation Function Values for 100 grid points

Table 2.3: Return to Education for Men Born 1930-1939: 1980 Census. Using a Post-Adaptive LASSO

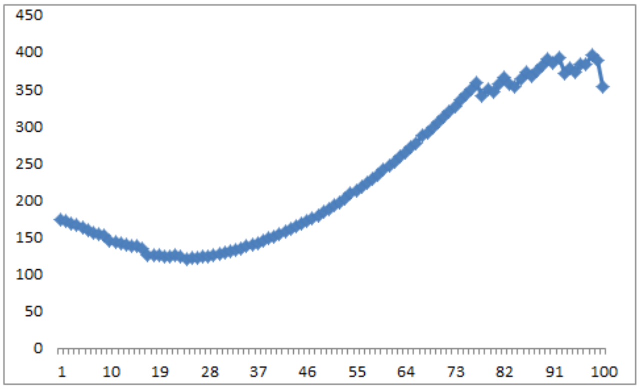
| | OLS | 2SLS | LIML | Fuller | **Post-A LASSO** |
|---|---|---|---|---|---|
| Education | 0.06 | 0.08 | 0.08 | 0.08 | 0.14 |
| Standard Error | (0.00) | (0.01) | (0.02) | (0.02) | (0.04) |
| First Stage F-Statistic | — | 2.36 | 2.36 | 2.36 | 17.88 |
| Instruments | — | 242 | 242 | 242 | 166 |
| Observations | 329,509 | 329,509 | 329,509 | 329,509 | 329,509 |

The dependent variable is the log of weekly earnings. The excluded regressors are age, age squared, dummies for race, marital status, state of birth, and quarter of birth, and interactions of quarter of birth with state of birth.

Table 2.4: Return to Education for Men Born 1930-1939: 1980 Census. Revisiting Bound et al. (1995)

| | OLS | 2SLS | LIML | Fuller | **Post-A LASSO** |
|---|---|---|---|---|---|
| Education | 0.062 | 0.064 | 0.056 | 0.416 | 0.055 |
| Standard Error | (0.000) | (0.015) | (0.005) | (4.879) | (0.135) |
| First Stage F-Statistic | — | 1.001 | 1.008 | 1.008 | 0.210 |
| Instruments | — | 242 | 242 | 242 | 10 |
| Observations | 329,509 | 329,509 | 329,509 | 329,509 | 329,509 |

The dependent variable is the log of weekly earnings. The excluded regressors are age, age squared, dummies for race, marital status, state of birth, and quarter of birth, and interactions of quarter of birth with state of birth.


450
400
350
300
250
200
150
100
50
0
1
10
19
28
37
46
55
64
73
82
91
100

# Chapter 3

# A Cross-Validated Spline Method for Nonparametric Instrumental Variable Estimation

## 3.1. Introduction

Empirical studies in economics often deal with the challenges posed by potential endogeneity in their estimations. A common solution to this difficulty is based on an instrumental variable procedure. The instrumental variable methodology was extended to the nonparametric framework by the work of Brown and Matzkin (1998), Newey et al. (1999), Altonji and Matzkin (2005), Darolles et al. (2003), Ai and Chen (2003), Newey and Powell (2003), Hall and Horowitz (2005), and Gagliardini and Scaillet (2006).

This paper solves the nonparametric instrumental variable problem within the context of reproducing kernel Hilbert Spaces (RKHs) recognizing that the object of interest is the solution to a Fredholm integral equation of the first kind. RKHs are characterized by the fact that linear functionals in the space are bounded. Therefore, the results of the previous literature, which assume the function of interest lies in a bounded Hilbert space, can be mapped into a RKH.

Solutions to Fredholm integral equations of the first kind are called regularized solutions. The methodology proposed in this paper is, as was typified by Nychka et al. (1984), a cross-validated spline solution. Within this framework the solution can be thought of as a penalized least squares estimate. The penalty over the roughness of the function, characteristic of these setups, is controlled by a regularization parameter that is chosen by Generalized Cross Validation (GCV). Except for Gagliardini and Scaillet (2006), the previous papers have no explicit mechanism to choose the regularization parameter and some, like Newey and Powell (2003), recognize their estimator is very sensitive to the choice of parameters. One advantage of GCV over the methodology of Gagliardini and Scaillet (2006) is that its optimality has been established by Wahba (1977) within the context of integral equations which are the object of interest in the literature of nonparametric endogeneity.

The solution proposed follows the strategy of Wahba (1969, 1973) and Kress (1989) to solve ill-posed inverse problems. However, many of the objects that are observed in the context of integral equations are unknown in the economic framework modeled in this study. These unknown objects will be replaced by nonparametric estimates. In this sense the solution in this paper is a modified version of the traditional cross-validated spline solutions to integral equations.

## 3.2. Background

The model of interest, which is a variation of the one presented by Newey and Powell (2003), is of the form

$$\begin{aligned} y &= g_0(x) + \kappa \\ E[\kappa|z] &= 0 \end{aligned} \tag{3.1}$$

In the expression above $y$ is an observable scalar random variable, $g_0$ represents the true structural function, $x$ is an explanatory variable vector of dimension $d_x \times 1$, $z$ is a vector of instruments of dimension $d_z \times 1$, and $\kappa$ is a disturbance.

Taking the conditional expectation of equation (3.1) one obtains the expression:

$$E\left[y|z\right] = E\left[g_0|z\right] = \int g_0\left(x\right) f\left(x|z\right) d_x \tag{3.2}$$

The relationship described in (3.2) is a Fredholm integral equation of the first kind, which in this case leads to an ill-posed inverse problem.[16] Fredholm integral equations of the first kind are usually written as:

$$w\left(z\right) = \int_x T\left(x, z\right) \varphi\left(x\right) d_x$$

For the expression above, in the integral equation literature, all the components are known except for $\varphi\left(x\right)$. It follows that (3.2) fits into the theory of integral equations with the function $T\left(.\right)$ being $f(x|z)$, the function that only depends on the constant term $z$ on the left hand side being $E(y|z)$ and the unknown function being $g_0\left(x\right)$.

The ill-posed inverse problem, as is stated in Kress (1989), is a consequence of $g_0\left(x\right)$ being an element of a space of functions, an infinite dimensional space. Specifically Theorem 15.4 of Kress (1989)ascertains:

[16] Chapter 15 of Kress (1989)provides a discussion of the ill-posed inverse problems and gives some examples. Chapter 8 of Wahba (1990) briefly discusses the problem in a framework that is closer to how I present it here.

*Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces and let $A : \mathcal{X} \to \mathcal{Y}$ be a compact linear operator. Then for the unknown function $\varphi$ the equation of the first kind $A\varphi = \Psi$ is improperly posed if $\mathcal{X}$ is not of finite dimension.*

The theorem is trying to convey that, given the infinite dimension of $\mathcal{X}$, the operator $A$ does not have a bounded inverse. This is exactly what happens in equation (3.2) where $g_0\left(x\right)$ is an element of a space of functions. The solution to this difficulty in theory consists in finding a bounded approximation to the unbounded inverse operator.

The ill-posed inverse problem also has a computational manifestation. To recover $g_0\left(x\right)$ from equation (3.2) the components of the solution must be computed discretely. As is pointed out by Wahba (1990), these discrete approximations are often numerically unstable. The situation deteriorates as the degree of discretization increases.

In summary, any solution to the ill-posed inverse problem needs to impose some bounds to the inverse integral operator that recovers $g_0$. As was discussed above the solution is sensitive to the bounds imposed. Therefore, a careful choice of the parameters of the problem is fundamental.

In what follows the methodology proposed to solve the problem is described. Then the convergence rates of the solution are derived.

## 3.3. Framework

Wahba (1969, 1973) propose a solution to Fredholm integral equations of the first kind within the context of RKHs. A fundamental difference arises, however, between the nonparametric instrumental variable problem and hers. In Wahba's approach a noisy version of $E(y|z)$, $E(y|z_i)^*$, is observed for some values of $z$ and $f(x|z)$ is known. Her methodology cannot be applied directly to equation (3.2) because $f(x|z)$ and $E(y|z_i)^*$ are unknown. Instead the solution she proposes is taken as a starting point and $f(x|z)$ and $E(y|z_i)^*$ are replaced by nonparametric estimates.

The underlying model in Wahba (1969, 1973), for $i = 1, \ldots, n$, is:

$$\begin{aligned} E(y|z_i)^* &= E(y|z_i) + \varepsilon_i \\ E(\varepsilon_i) &= 0 \end{aligned}$$

Within her framework the solution to the problem becomes to find a function in a RKHS, $\mathcal{H}_R$, which satisfies the following expression:

$$\min_{g \in \mathcal{H}_R} \sum_{i=1}^{n} \left( \int g_0(x) f(x|z_i) \, d_x - E(y|z_i)^* \right)^2 + \alpha \, \|g\|^2_{\mathcal{H}_R} \tag{3.3}$$

In the expression above $\int_X g(x) f(x|z) d_x$ is what is referred to as an approximate solution for (3.2) which becomes exactly $E(y|z_i)$ at $g_0\,(x)$.

Before commenting on the solution of equation (3.3) some basic concepts of RKHs that are related to the solution of (3.3) using Wahba's methodology are introduced.

**Definition 1** (Reproducing Kernel Hilbert Space)**.** A RKHS is a Hilbert space of real-valued functions on an index set $\mathcal{T}$, for instance $\mathcal{T} = [0, 1]$, with the property that for $t \in \mathcal{T}$, the evaluation functional $L_t$, which associates $g$ with $g(t)$, is a bounded linear functional in the sense that, $\exists M$ such that:

$L_t h = |g(t)| \leq M \, \|g\|$ for all $h$ in the RKH,

where $\|.\|$ is the norm in the Hilbert space.

It is a well known result in the literature of RKHs[17] that to every RKH there corresponds a unique positive definite function on $\mathcal{T} x \mathcal{T}$ that is referred to as its reproducing kernel. By definition the reproducing kernel is an object which has the property that its inner product with any function in the RKH space yields the function. It will be denoted by $R(t, t')$ and assumed that:

**Assumption 1.** The reproducing Kernel $R(t, t')$ is continuous and $\int \int_{\mathcal{T}} R^2(t, t') d_t d_t' < \infty$

[17] A more detailed description can be found in Aronszajn (1950).

For every space of functions that satisfies definition 1 and assumption 1 the following properties of RKHs can be defined.

**Property of RKHS 1.** By the theorems of Hilbert, Schmidt, and Mercer [18] the reproducing kernel can be written as:

$$R(t,t') = \sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu \left(t\right) \phi_\nu \left(t'\right)$$

where $\{\phi_\nu\}_{\nu=1}^{\infty}$ is a complete orthonormal system of eigenfunctions on the space with corresponding eigenvalues $\{\lambda_\nu\}_{\nu=1}^{\infty}$, $\lambda_\nu > 0$, $\sum_{\nu=1}^{\infty} \lambda_\nu < \infty$, and $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$.

**Property of RKHS 2.** Definition 1 of the RKHS $\mathcal{H}_R$ above can be stated explicitly in terms of the inner product of the RKHS, $\langle . \rangle_{\mathcal{H}_R}$, by the following relationships:

$$\begin{aligned} \mathcal{H}_R &= \left\{ g : g \in \mathcal{L}_2\left(\mathcal{T}\right), \quad \sum_{\nu=1}^{\infty} \frac{\left(g, \phi_\nu\right)^2}{\lambda_\nu} < \infty \right\} \\ \langle g, g' \rangle_{\mathcal{H}_R} &= \sum_{\nu=1}^{\infty} \frac{\left(g, \phi_\nu\right)\left(g', \phi_\nu\right)}{\lambda_\nu} \end{aligned}$$

In the expression above $(.)$ is the inner product in the $\mathcal{L}_2\left(\mathcal{T}\right)$ space.

**Property of RKHS 3.** As a consequence of properties 1 and 2 we have that:

a. For a fixed value $t$, $R(t,t') = R_t(t') \in \mathcal{H}_R$

b. $\langle R_t, g \rangle_{\mathcal{H}_R}, \quad t \in \mathcal{T}$

[18] The theorems can be found in Riesz (1955) in pages 242-246

**Property of RKHS 4.** For any bounded linear functional one can find its representer $\eta_t$ which is defined by:

$L_t = \langle \eta_t, g \rangle_{\mathcal{H}_R}$ and $\eta_t(t') = \langle \eta_t, R_{t'} \rangle_{\mathcal{H}_R}$

The intuition behind properties 3 and 4 is that by knowing the reproducing kernel and the representer of a space one can characterize any function and linear functional in the RKH. Furthermore, as will be seen when we study the convergence properties of the solution to equation (3.3), the characterization of the reproducing kernel as $\sum_{\nu=1}^{\infty} \lambda_\nu \phi_\nu(t) \phi_\nu(t')$ and of the inner product in property 2 allows an analysis of convergence rate of elements in the space by imposing conditions on the decreasing sequence $\{\lambda_\nu\}_{\nu=1}^{\infty}$ of property 1. More importantly, these properties allow the main problem to be written as a finite dimensional problem using the methodology proposed by Kimeldorf and Wahba (1971).

In assumption 2 below conditions on the function $g_0$ are imposed to explicitly obtain its reproducing kernel. A different assumption on the function $g_0$ will yield a different reproducing kernel and different representers. The properties imposed on $g_0$ in assumption 2 are restrictive. However, they are in accordance with the proposed solutions in the literature to nonparametric instrumental variables.

**Assumption 2.** The function $g_0$ defined in equation (3.2) belongs to the Hilbert space $W^m$ defined by:

$$W^m\left[X\right]=\left\{g:g^{(m-1)}\quad\text{is absolutely continuous and }g^{(m)}\in\mathcal{L}_2\left[X\right]\right\}$$

In the expression above $g^{(m)}$ is the $m^{th}$ derivative of $g$. By virtue of assumption 2 and following Wahba (1990), the reproducing kernel of the space is defined to be[19]:

$$R(x,x')=\sum_{\nu=1}^{m}\frac{x^{\nu-1}x'^{\nu-1}}{(\nu-1)!\,(\nu-1)!}+\int_X\frac{(x-u)_+^{m-1}\,(x'-u)_+^{m-1}}{(m-1)!\,(m-1)!}d_u \tag{3.4}$$

where $(x)_+=x$ for $x\geq 0$ and otherwise $(x)_+=0$.

An explicit expression for the representers of property 4 can be found. To find it note that $L_t$ of property 4 is the integral defined in equation (3.3) and that $R_t$ is defined by equation (3.4). The representers are defined by:

$$\eta_z(x)=\int_X f\left(x'|z\right)R\left(x,x'\right)dx'\quad x,x'\in X \tag{3.5}$$

Now that the properties of RKHS have been stated and that exact expressions for the representers and the reproducing kernel have been rendered, I can describe the solution

[19] Wahba (1990) has a detailed discussion of the construction of this reproducing kernel which is based on the Taylor approximation to an $m$ times continuously differentiable function.

to the problem in (3.3) using RKHs can be described.

A starting point for the solution of (3.3) in Wahba (1969, 1973) is the knowledge of $E(y|z_i)^*$ for certain values of $z \in \{z_1, z_2 \ldots, z_n\}$ where $z_1 < z_2 < \ldots < z_n$. Also, by property 4 of RKHS the approximate solution to $E(y|z_i)$ can be written in terms of the representers. Specifically,

$$\int_X g(x) f\left(x|z_i\right) d_x = \langle n_{z_i}, g \rangle_{\mathcal{H}_R} \tag{3.6}$$

Using (3.6) expression (3.3) can be rewritten as:

$$\min_{g \in \mathcal{H}_R} \sum_{i=1}^{n} \left( E\left(y|z_i\right)^* - \langle n_{z_i}, g \rangle_{\mathcal{H}_R} \right)^2 + \alpha \left\| g \right\|^2_{\mathcal{H}_R} \tag{3.7}$$

The problem in (3.7) consists of two parts. The first component $\sum_{i=1}^{n} \left( E\left(y|z_i\right)^* - \langle n_{z_i}, g \rangle_{\mathcal{H}_R} \right)^2$ can be thought of as the fidelity of the data, while the second component is controlling the amount of smoothing. The parameter alpha controls this trade-off and imposes a bound on the variability of the function of interest. This is another reason why constructing a data driven mechanism to choose $\alpha$ is important.

Within this framework Kimeldorf and Wahba (1971) have established the following result:

**Proposition 1.** The solution to problem (3.3) using (3.7) is given by:

$$g^{*}\left(x\right)=\left(\eta_{z_1}\left(x\right),\ldots,\eta_{z_n}\left(x\right)\right)\left(Q_n+\alpha I\right)^{-1}\left(E\left(y|z_1\right)^{*},\ldots,E\left(y|z_n\right)^{*}\right) \tag{3.8}$$

In equation (3.8) $Q_n$ is an $n \times n$ matrix whose $ij^{th}$ element is $\left\langle \eta_{z_i},\eta_{z_j}\right\rangle_{\mathcal{H}_R}$.

The solution $g^{*}\left(x\right)$ in (3.8) can be thought of as lying in a space spanned by the representers. Therefore, the representers can be viewed as a basis of the space where the solution exists. This is the intuition that underlies the latter computation of the solution and is another way of stating the main conclusion of the representer theorem of Kimeldorf and Wahba (1971) . This way of understanding and writing the solution relies critically on the use of the RKHS machinery.

Another important consideration is that the regularized solution given in (3.8) is not the solution to the problem in (3.3) but an approximation that tries to control for the ill-posedness of the problem via $\alpha$. The true solution to the problem, which is infeasible by the ill-posedness embedded in the matrix $Q_n$, occurs when $\alpha = 0$. The convergence rates of the solution then depend on the rate at which a sequence of regularization parameters tends to zero as n increases. This will be discussed later with respect to the convergence rates.

The estimator in (3.8) is, however, infeasible. The first reason for this is that $f(x|z)$ is unknown. Also the zero mean perturbed version of $E(y|zi)$ is not observed. What will be observed is a nonparametric estimate of $E(y|zi)$ that will be used to approximate $E(y|z)^*$. This biased estimator is defined as $\Upsilon\left(z_i\right)$ and will be computed using a series estimator to be described below. The estimate of $f(x|z)$, on the other hand, will be done using the methodology of Fan et al. (1996). The estimator under this framework becomes:

$$\tilde{g}\left(x\right) = \left(\hat{\eta}_{z_1}\left(x\right), \ldots, \hat{\eta}_{z_n}\left(x\right)\right)\left(\hat{Q}_n + \alpha I\right)^{-1}\left(\Upsilon\left(z_1\right), \ldots, \Upsilon\left(z_n\right)\right) \tag{3.9}$$

In equation (3.9) $\hat{Q}_n$ is an $n \times n$ matrix whose $ij^{th}$ element is $\left\langle \hat{\eta}_{z_i}, \hat{\eta}_{z_j} \right\rangle_{\mathcal{H}_R}$ and:

$$\hat{\eta}_z\left(x\right) = \int_X \hat{f}\left(x'|z\right) R\left(x, x'\right) dx' \quad x, x' \in X$$

## 3.4. Estimation of $E\left(y|z\right)$, $f\left(x|z\right)$, and selection of $\alpha$

For the estimation of $f\left(x|z\right)$ we will adopt the estimator of Fan et al. (1996) that can be written as:

$$\hat{f}\left(x|z\right) = \frac{1}{h_1 h_2} \sum_{i=1}^{n} W_0^n \left(\frac{z_i - z}{h_1}\right) N \left(\frac{x_i - x}{h_2}\right) \tag{3.10}$$

In the above equation $N(.)$ is a kernel weight, $h_1$ and $h_2$ are bandwidths, and $W_0^n(.)$, which can be thought of as a local quadratic approximation to obtain the conditional density, is determined by the following expressions:

$$W_0^n = \tau_0^T S_n^{-1} \left(1, h_1 t, h_1 t^2\right)^T \times W\left(t\right) \tag{3.11}$$

$$\begin{aligned} S_n &= \begin{pmatrix} s_{n,0} & s_{n,1} & s_{n,2} \\ s_{n,1} & s_{n,2} & s_{n,3} \\ s_{n,2} & s_{n,3} & s_{n,4} \end{pmatrix} \\ s_{n,j} &= \frac{1}{h_1} \sum_{i=1}^{n} \left(Z_i - z\right)^j W\left(\frac{Z_i - z}{h_1}\right) \end{aligned} \tag{3.12}$$

In equation (3.12) $W(.)$ is another kernel weight and $\tau_0$ is the unit vector with the first element equal to one. For the estimation an Epanechnikov kernel is used. The choice of kernel obeys to the fact that it satisfies the assumptions of Newey (1994), which are important for the convergence rate results, and because the integral that defines the representers has a closed form solution[20].

To determine how to select the regularization parameter in (3.9) and the bandwidths $h_1$ and $h_2$ Generalized Cross Validation as proposed by Wahba (1977) and Golub et al. (1979) will be used. For the purpose of writing the cross-validation criterion function (3.9) can be expressed in terms of the parameters as:

$$\begin{aligned} \tilde{g}\left(x,\alpha,h_1,h_2\right) &= A\left(x,\alpha,h_1,h_2\right) \mathrm{Y}\left(z\right) && (3.13) \\ A\left(x,\alpha,h_1,h_2\right) &= \left(\eta_{z_1}\left(x,h_1,h_2\right),\ldots,\eta_{z_n}\left(x,h_1,h_2\right)\right)\left(\hat{Q}_n\left(h_1,h_2\right)+\alpha I\right)^{-1} && (3.14) \\ \mathrm{Y}\left(z\right) &= \left(\mathrm{Y}\left(z_1\right),\ldots,\mathrm{Y}\left(z_n\right)\right) && (3.15) \end{aligned}$$

One can obtain the optimal regularization parameter, $\alpha$, and the bandwidths that minimize the generalized cross-validation function given by:

$$V\left(\alpha,h_1,h_2\right) = \sum_{i=1}^{n} \frac{\left(y_i - \tilde{g}\left(x,\alpha,h_1,h_2\right)\right)^2}{\left[\frac{1}{n}\mathrm{Trace}\left(I - A\left(x,\alpha,h_1,h_2\right)\right)\right]^2} \tag{3.16}$$

This choice of the regularization parameter as well as the solution to the ill-posed problem was suggested by Wahba in the context of RKHS. This study adds the choice of bandwidths to this procedure thus providing a data driven mechanism to choose the

[20] The Gaussian kernel which is the other conventional choice does not satisfy these two conditions

parameters of the estimation. The fact that the choice of the regularization parameter and the bandwidths comes from the minimization of the criterion function $V(\alpha, h1, h2)$ and is not to be arbitrarily chosen by the researcher is a contribution of this paper.

The computation of $E(y|z)$ will use a power series estimator. Let us define a vector of $L$ approximating functions $p^L(.)$ of dimension $L \times 1$ by:

$$p^L(z) = (p_1(z), \ldots, p_L(z))' \tag{3.17}$$

Also, let us define $P$ to be an $n \times L$ matrix whose $i^{th}$ row is given by $p^L(Z_i)'$. Under this framework the estimate of $E(y|z)$ is given by:

$$\Upsilon(z) = p^L(z)' \left(P'P\right)^{-} P'y \tag{3.18}$$

In (3.18) the term $(.)^{-}$ denotes the generalized inverse of $(.)$.

The number of series terms of this estimator is once more chosen using Generalized Cross Validation. In this specific case the generalized cross-validation function to be minimized has the form:

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \tilde{g}\left(x,\alpha,h_1,h_2\right)\right)^2\left(1-\frac{L}{n}\right)^{-2}$$

## 3.5. Properties of the Proposed Solution

The results established in this section come from three sources. The first is the paper of Lukas (1988) , who studies the properties of the convergence rates for regularized solutions like the one in (3.8) for a wide class of RKHs. The second source is Newey (1994) , who studies the convergence rates of kernel estimators under a Sobolev supremum norm. The third is the paper of Newey (1997) , which studies the properties of power series.

The convergence rates of the estimator proposed in (3.9) are determined in two stages. In the first stage the convergence properties of $g^*$ as an estimate of $g_0$ are established. This result comes directly from the literature on ill-posed inverse problems as is presented in Lukas (1988). In the second stage $\tilde{g}$ as an approximation of $g^*$ is studied. This is achieved using the conclusions in Newey (1994, 1997) and Lemma 1 of the Appendix. As noted in the Appendix, Lemma 1 provides a link between the work of Lukas (1988) and the results of Newey (1994, 1997). The inputs from the two stages will be united to determine the convergence rates that arise from the relationship $\|g_0-\tilde{g}\|_{W_u} = \|g_0-\tilde{g}+g*-g*\|_{W_u} \leq \|g_0-g^*\|_{W_u}+\|g^*-\tilde{g}\|_{W_u}$. The norm $\|.\|_{W_u}$

comes from Lukas (1988) and will be defined below.

Some notation is introduced to simplify the exposition. For an arbitrary function $g$, define the operators: $\hat{\mathcal{K}}g = \int \hat{f}\,(x|z)\,h(.)dx$ and $\mathcal{K}g = \int f\,(x|z)\,h(.)dx$. Also by property 4 of the RKHs and equation (3.5):

$$\hat{\mathcal{K}}\tilde{g}\,(x) = \left(\hat{Q}\,(z_1, z)\,, \ldots, \hat{Q}\,(z_n, z)\right)\left(\hat{Q}_n + \alpha I\right)^{-1}\left(\Upsilon\,(z_1)\,, \ldots, \Upsilon\,(z_n)\right)$$

Following Kimeldorf and Wahba (1971) and Lukas (1988) it can be established that $\hat{Q}$ is continuous on $Z \times Z$ and, for an arbitrary function $g$, the following operator $\mathcal{Q} : \mathcal{L}^2[Z] \to \mathcal{L}2[Z]$ can be defined by:

$$\mathcal{Q}g\,(z) = \int_Z \hat{Q}\,(z, z')\,g\,(z')\,d'_z \tag{3.19}$$

Lukas (1988) shows that $\mathcal{Q}$ is bounded and positive definite. Assumption 1 and Definition 1 are satisfied and so $\hat{Q}(z, z')$ is a reproducing kernel for the RKH. I shall refer to it as $\mathcal{H}_{RQ}$. Let us define its non-increasing sequence of eigenvalues by $\lambda_{1Q} \geq \lambda_{2Q} \geq \ldots \geq 0$ and the corresponding orthonormal eigenfunctions by $\{\phi_{vQ}\}_{v=1}^{\infty}$. Therefore, $\mathcal{H}_{RQ}$ is defined in a similar way as $\mathcal{H}_R$ in property 2 with $\lambda_{iQ}$ in place of $\lambda_i$ and $\phi_{vQ}$ in place of $\phi_v$. The following assumption on the $\lambda_{iQ}$ is imposed.

**Assumption 3.** $0 < a_1 i^{-2b} \leq \lambda_{iQ} \leq a_2 i^{-2b}$, $i = 1, 2, \ldots$ and $b > 1/2$ for non negative constants $a_1$ and $a_2$.

Lukas (1988) studies a class of RKHS defined by:

$$H_u = \left\{ g \in \mathcal{L}^2 \left[ Z \right] : \sum_{i=1}^{\infty} \frac{\left( g, \phi_{iQ} \right)^2}{\lambda_{iQ}^u} < \infty \right\}$$

and inner product given by,

$$\left( g, g' \right)_{\mathcal{H}_u} = \sum_{i=1}^{\infty} \frac{\left( g, \phi_{iQ} \right) \left( g', \phi_{iQ} \right)}{\lambda_{iQ}^u}$$

The expressions above are analogous to those put forth in property 2. The difference arises in the $\lambda_{iQ}^u$ term that imposes faster convergence rates on the Fourier coefficients for higher $u$ [21]. An important fact is that for u$=1$ we obtain $\mathcal{H}_R$ defined in property 2 of RKHs and that for $v < u$ we have that $\mathcal{H}_u \subset \mathcal{H}_u$ [22]. By allowing different values of $u$ Lukas (1988) studies the convergence rate for a wide class of RKHs.

Lukas (1988) defines,

[21] In each of these cases the reproducing kernel changes for each $u$. Specifically, for each $u$ the reproducing kernel will be written, using property 1 of RKHS, by $\sum_{i=1}^{\infty} \lambda_{iQ}^u \phi_{iQ}(t) \phi_{iQ}(t')$.

[22] The previous result comes from the fact that $\|g\|_{\mathcal{H}_v}^2 = \sum_{i=1}^{\infty} \frac{\left( g, \phi_{iQ} \right)^2}{\lambda_{iQ}^v} = \sum_{i=1}^{\infty} \frac{\left( g, \phi_{iQ} \right)^2 \lambda_{iQ}^{u-v}}{\lambda_{iQ}^u} \leq max \left\| \lambda_{iQ}^{u-v} \right\| \|g\|_{\mathcal{H}_u}^2$.

$$W_u \equiv \{g \in \mathcal{H}_{RQ} : \mathcal{K}g \in \mathcal{H}_u\}$$

and shows the existence of an isometric isomorphism between $W_u$ and $\mathcal{H}_u$ defined by the relationship:

$$(g, g')_{W_u} = (\mathcal{K}g, \mathcal{K}g')_{\mathcal{H}_u} \tag{3.20}$$

These elements of Lukas (1988) are used to define the distance between the true function, $g_0$, and the regularized solution, $g^*$. Under regularity conditions presented in the Appendix, it can be established that for $E(y|z)^*$ that belongs to $H_s$, where $s \geq \max u, v$:

for $u \leq s \leq u + 2$,

$$\|g_0 - g^*\|_{W_u} = O_p\left(\left(\alpha^{s-u}\left\|E\left(y|z\right)^*\right\|_s^2 + \frac{\sigma_u^2}{n}\alpha^{-u-1/2b}\right)^{1/2}\right)$$

and for $s \geq u + 2$,

$$\|g_0 - g^*\|_{W_u} = O_p\left(\left(\alpha^{2}\left\|E\left(y|z\right)^*\right\|_s^2 + \frac{\sigma_u^2}{n}\alpha^{-u-1/2b}\right)^{1/2}\right)$$

The above statement gives the first component of the convergence result of this paper. The outcome depends on the rate of decay of the Fourier coefficients. This occurs via $b$ that establishes bounds on the eigenvalues, the constant $s$ that governs the rate of decay of the eigenvalues on $\mathcal{H}_s$, and $u$ which, similarly, controls the rate of decay of the eigenvalues in $\mathcal{H}_u$.

The second component of the convergence rates is established in Theorem 1, stated precisely and proved in the Appendix. Theorem 1 concludes that:

$$\|g^* - \tilde{g}\|_{W_u} = O_p\left(\frac{L^{3/2}}{\sqrt{n}} + L^{1-\tau} + ln(n)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

In the expression above, $L$ refers to the number of series terms; $\tau$ is rate at which the approximation error of the power series estimator to $E(y|z)$ shrinks under the Sobolev Supremum norm, which is exactly $O\left(L^{-\tau}\right)$; $\bar{\omega}$ is the order of the kernel $W_0^n(.)$; $h$ is a bandwidth that is asymptotically equivalent to $h_1$ and $h_2$; $m$ is the order of the derivatives of the Sobolev space defined in Assumption 2; and $k$ denotes the existence of an extension of $f\left(x|z\right)$ to all of $\mathcal{R}^k$ that is continuously differentiable to order $\zeta$ on $\mathcal{R}^k$. $\zeta$ is a non-negative integer.

Using the results given for $\|g - g*\|_{W_u}$ and $\|g^* - \tilde{g}\|_{W_u}$ and the triangle inequality the conclusion of Theorem 2, presented rigorously in the Appendix, can be ascertained.

Specifically:

for $u \leq s \leq u+2$,

$$\|g_0 - \tilde{g}\|_{W_u} = O_p\left(\left(\alpha^{s-u} + \frac{\alpha^{-u-1/2b}}{n}\right)^{1/2} + \frac{L^{3/2}}{\sqrt{n}} + L^{1-\tau} + ln(n)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

for $s > u+2$ :

$$\|g_0 - \tilde{g}\|_{W_u} = O_p\left(\left(\alpha^{2} + \frac{\alpha^{-u-1/2b}}{n}\right)^{1/2} + \frac{L^{3/2}}{\sqrt{n}} + L^{1-\tau} + ln(n)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

From Theorem 2 it is clear that the convergence rates of the solution are neatly separated between the effect of the series estimate of $E\left(y|z\right)$, the regularized solution, and the kernel estimate of the conditional density. The convergence rate of $\tilde{g}$ to $g_0$ is then dominated by the convergence of the slowest of these terms. From this analysis and as a consequence of Theorem 2 the following result holds:

**Corollary 1.** Let $\delta_n \equiv (ln\,(n))^{\frac{4}{2(\bar{\omega}+m)+k}}$. Then the optimal $\alpha$, $h$, and $L$ are given by:

$$\begin{aligned} h^* &= O_p\left(\delta_n n^{\frac{-1}{2(\bar{\omega}+m)+k}}\right) \\ L^* &= O_p\left(n^{\frac{-1}{n^{2\tau+1}}}\right) \end{aligned}$$

for $u \leq s \leq u+2$,

$$\alpha^* = O_p\left(n^{\frac{-b}{2bs+1}}\right)$$

and for $s > u+2$:

$$\alpha^* = O_p\left(n^{\frac{-b}{4b+2bu+1}}\right)$$

Moreover,

for $u \leq s \leq u+2$,

$$\|g_o - \tilde{g}\|_{W_u} = O_p\left(\min\left\{\left(n^{\frac{-b(s-u)}{2bs+1}} + n^{\frac{-2b(2s-u)-1}{2(2bs+1)}}\right)^{1/2}; n^{\frac{-\tau-2}{2\tau-1}}; \delta_n^{\frac{2(\bar{\omega}+m)+k-1}{2}} n^{\frac{\bar{\omega}}{2(\bar{\omega}+m)+k}}\right\}\right)$$

for $s > u+2$:

$$\|g_o - \tilde{g}\|_{W_u} = O_p\left(\min\left\{\left(n^{\frac{-2b}{4b+2bu+1}} + n^{\frac{-2b(4b+u)-1}{4b+2bu+1}}\right)^{1/2}; n^{\frac{-\tau-2}{2\tau-1}}; \delta_n^{\frac{2(\bar{\omega}+m)+k-1}{2}} n^{\frac{\bar{\omega}}{2(\bar{\omega}+m)+k}}\right\}\right)$$

## 3.6. Results

This section illustrates how the methodology proposed works. It is tested on the function:

$$\begin{aligned} y &= g(x) + \kappa = -sin(8x) + 0.025\kappa && (3.21) \\ x &= z + v && (3.22) \\ \begin{pmatrix} \kappa \\ v \\ x \end{pmatrix} &\sim N\begin{pmatrix} & 1 & 0.5 & 0 \\ 0, & 0.5 & 1 & 0 \\ & 0 & 0 & 1 \end{pmatrix} && (3.23) \end{aligned}$$

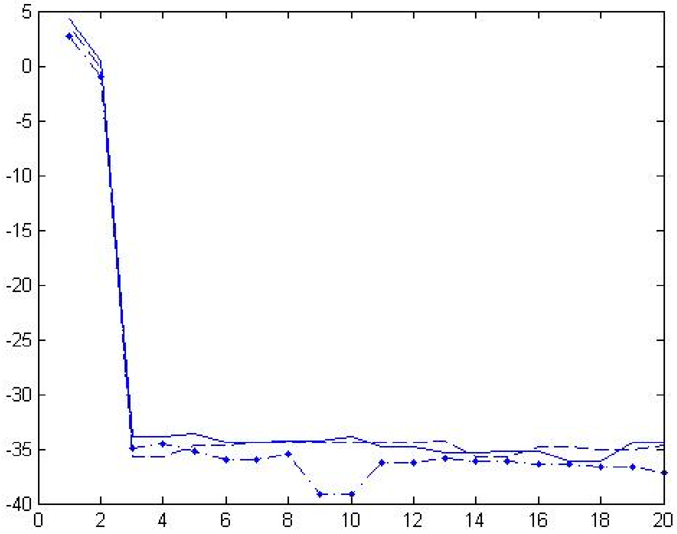
The $x$ and $z$ are normalized to lie between zero and one and given this modification $\kappa$ is rescaled by $c_1$. The estimations are performed for 50 and 100 grid points of the vector $z$. This is equivalent to the number of representers used to span the space of functions.

As the number of grid points augments, the ill-posedness of the problem increases, so having too many grid points is not advisable. One can test the degree of ill-posedness by plotting the log of the eigenvalues of the matrix $\hat{Q}_n$. The plot should be decreasing until at some point it starts to show an oscillating behavior. If the plot for a number of grid points starts to fluctuate before another, the matrix is more ill-conditioned. A higher level of oscillation also indicates a more ill-conditioned matrix. If the plots are similar, the one with the least number of grid points should be selected. Although not a formal theory of how to determine the number of grid points, this method can serve as a guideline.

Graph 1 below shows the plot for 20, 50, and 100 grid points. The solid line represents the eigenvalues of $\hat{Q}_n$ for 100 grid points, the dashed line for 50, and the dash-dot line is for 20 grid points. According to the suggested criteria a choice of 50 grid points seems to be the most adequate. However, results are presented both for 50 and 100 grid points.

Graph 1:

Eigenvalues of the matrix $\hat{Q}_n$ for three different grids for equation (3.21).



The dashed line corresponds to the grid with 50 points, the dotted line is for 20, and the solid for 100

Table 3.1: Values of the Parameters Using Generalized Cross-Validation

| | |
|---|---|
| Regularization Parameter $\alpha$ | 1.235 |
| Bandwidth for Z's | 3.335 |
| Bandwidth for the X's | 0.009 |

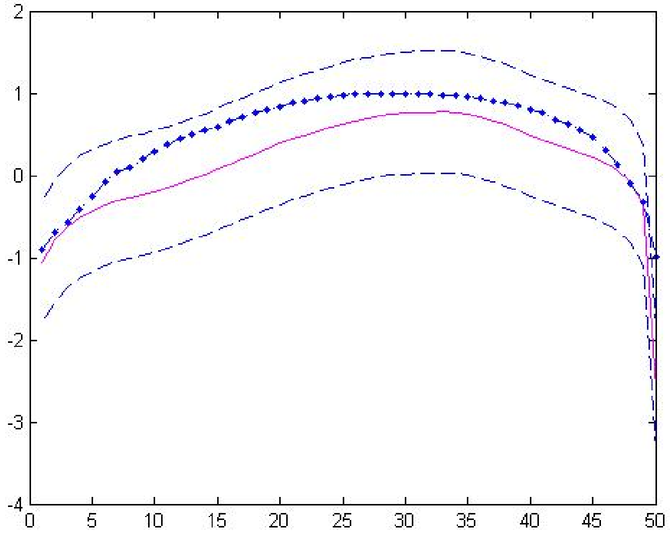
Results correspond to grid size of 50 for equation (3.21).

In Table 1 below the values of the parameters are presented for the function in equation (3.21), for the two sets of grid sizes selected. As was explained above they were chosen by minimizing the function in equation (3.16). The minimization was done by a grid search. The optimization was implemented in this way because the generalized cross validation function is very flat in some regions and built-in minimization routines in software can be very sensitive to starting values.

In Graphs 2 and Graph 3 the estimation results are presented for equation (3.21) for the two grid sizes selected. The graphs correspond to the mean after 500 realizations of the estimation. The confidence intervals are constructed $\pm 1.96\sqrt{\text{mean standard error}}$ to this estimate. In the graphs the confidence intervals are the dashed lines, the estimator is solid line, and the true function is the dotted line.

Graph 2:

Estimation Results for $\tilde{g}$ in (3.21).



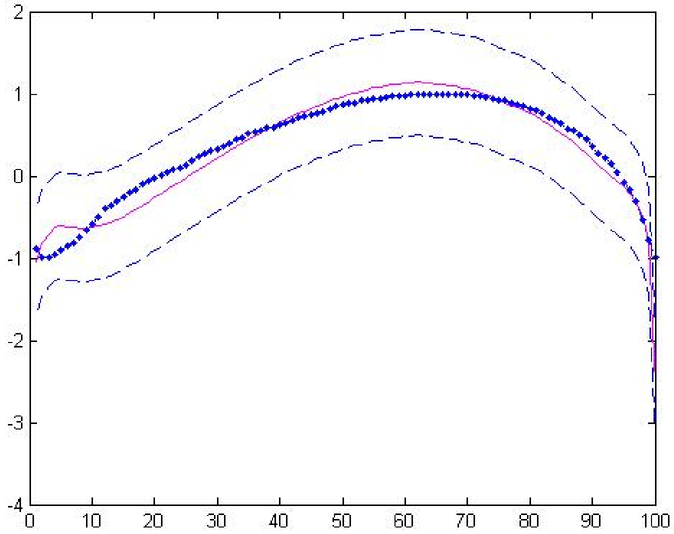
Estimates with 50 grid points and parameters of Table 1

Dashed line for confidence intervals, dotted for true $g$, and solid for $\tilde{g}$

The results indicate that the methodology proposed does a good job in recovering the shape of the function of interest for the grid points selected. Also, it performs well in the regions of the curve that are less flat. Graphs 1a and 1b, moreover, show that for the experiment presented the matrices are very ill-conditioned. Careful choice of the parameters is therefore crucial and in this sense this paper makes an important contribution.

Graph 3:

Estimation Results for $\tilde{g}$ in (3.21).



Estimates with 50 grid points and parameters of Table 1

Dashed line for confidence intervals, dotted for true $g$, and solid for $\tilde{g}$

## 3.7. Appendix

Here, the results of the section where the properties of the solution were described is presented in detail. The assumptions used to derive the conclusions are stated explicitly and all proofs are provided. First, results of $g^*$ as an approximation of $g_0$ are given, followed by those of $\tilde{g}$ as an estimator of $g^*$.

Using the result in (3.20), and following Lukas (1988), the mean squared error for $g^*$ as an estimate of $g_0$ can be determined by the following relationship:

$$E\left\|g_o - g^*\right\|_{W_u}^2 = E\left\|\mathcal{K}g_o - \mathcal{K}g^*\right\|_{H_u}^2 \tag{3.24}$$

Assumptions 4 and 5, needed to derive the convergence rates of the relation expressed in equation (3.24), are stated as follows:

**Assumption 4.** There exists $v$, $0 < v < 1 - 1/4b$, and a sequence $c_n \to 0$ such that for all $g, g' \in \mathcal{H}$:

$$\left|\int_Z gg' - \frac{1}{n}\sum_{i=1}^{n} g\left(z_i\right) g'\left(z_i\right)\right| \leq c_n \left\|g\right\|_{H_v} \left\|g'\right\|_{H_v}$$

**Assumption 5.** The random disturbances $\varepsilon_i$ in equation (3.3) satisfies, $E\varepsilon_i = 0$, $\sup_i Var\left(\varepsilon_i\right) = \sigma_u^2$ and the $\varepsilon_i's$ are independent.

Assumption 5 is a slight modification of the assumptions of Lukas (1988) . Specifically, Lukas (1988) assumes a homoskedastic variance. Here it is assumed instead that the variance is bounded above. The results of Lukas (1988) are not modified by this and proceed in the same way with $\sigma_u^2$ in place of the homoskedastic variance.

A restatement of Lukas (1988) Theorem 2.1 using the notation of this study is:

*Under assumptions 3 to 5, let $E\left(y|z\right)^*$ belongs to $H_s$, where $s \geq \max u, v$ and $u < 2 - v - 1/2b$. Suppose that $\alpha = \alpha(n) \to 0$ as $n \to \infty$ in such a way that $c_n \alpha^{-v-1/4b} \to 0$ and, if $u > v$*

*and* $s > v + 2$, $c_n \alpha^{\frac{-v}{2} - \frac{-u}{2} - \frac{1}{4b}} \to 0$. *Then for* $u \leq s \leq u + 2$,

$$E \left\| g_0 - g^* \right\|_{W_u}^2 = E \left\| \mathcal{K} g_0 - E(y|z)^* \right\|_{H_u}^2 = O\left( \alpha^{s-u} \left\| E(y|z)^* \right\|_s^2 + \frac{\sigma_u^2}{n} \alpha^{-u-1/2b} \right)$$

*and for* $s \geq u + 2$,

$$E \left\| g_0 - g^* \right\|_{W_u}^2 = E \left\| \mathcal{K} g_0 - E(y|z)^* \right\|_{H_u}^2 = O\left( \alpha^{2} \left\| E(y|z)^* \right\|_s^2 + \frac{\sigma_u^2}{n} \alpha^{-u-1/2b} \right)$$

Assumptions 6-a2ch8 below come from Newey (1997) and Assumptions 9-11 from Newey (1994) . They are used by him to derive the convergence rate of series estimators and of kernel estimators respectively. They will be used here to construct the second element necessary to arrive at the convergence rate of the solution in (3.9), that is:

$$\left\| g^* - \tilde{g} \right\|_{W_u} \tag{3.25}$$

**Assumption 6.** $(y_1, z_1), \ldots, (y_n, z_n)$ are *i.i.d* and $Var\,(y|z)$ is bounded.

**Assumption 7.** For every $L$, as defined in (3.17), there is a nonsingular matrix $B$ such that for $P^L(z) = B_p^L(z)$; (i) the smallest eigenvalue of $E\left[P^L(z_i)\, P^L(z_i)'\right]$ is bounded away from zero uniformly in $L$ and; (ii) there is a sequence of constants $\xi_0(L)$ satisfying $\sup_{z \in Z} \left\| P^L(z) \right\| \leq \xi_0(L)$ and $L = L(n)$ such that $\xi_0(L)$ follows $\xi_0(L)^2 L/n \to 0$ as

$n \to \infty$.

**Assumption 8.** For an integer $m \geq 0$ there are $\tau$ and $\beta_L$ such that $E\left\|E\left(y|z\right)-P^{L'}\beta_L\right\|_m = O\left(L^{-\tau}\right)$ as $L \to \infty$.

In assumption 8 for an arbitrary function $h$, $\|h\|_m = \max_{|\theta|\leq m}\sup_{z\in Z}\left\|\partial^{|\theta|}h\left(z\right)\right\|$ where $\theta = \left(\theta_1,\ldots,\theta_r\right)'$ is a vector of nonnegative integers with the same dimension as $z$, $|\theta| = \sum_{j=1}^{r}\theta_j$ and $\partial^{|\theta|}h\left(z\right)$ is a vector of partial derivatives. This is a Sobolev supremum norm.

**Lemma 1.** Using Assumption 1, if $g \in W^m$ converges in the Sobolev supremum norm $\|g\|_m = \max_{|\theta|\leq m}\sup_{z\in Z}\left\|\partial^{|\theta|}g\left(z\right)\right\|$ it converges in the norm for $H_u$, where $W^m$ is as defined in Assumption 2.

*Proof.* Assume that a sequence of functions $g_n(x)$ in $W^m$ converges in the Sobolev supremum norm to $g$. Define the reproducing kernel by $R_x$. Then the following relationship holds:

$$\begin{aligned}
|g_n(x)-g(x)| &= \left|\left(g_n, g, R_x\right)_{H_u}\right| \\
&\leq \|g_n - g\|_{H_u}\|R_x\| \\
&= \sqrt{\sum_{v=0}^{m-1}\left[\partial^m\left(g_n-g\right)\left(0\right)\right]^2+\int_0^1\left[\partial^m\left(g_n-g\right)\left(u\right)\right]^2 d_u\,\|R_x\|} \\
&= \sqrt{m\max_{|\theta|\leq m}\left[\partial^m\left(g_n-g\right)\left(0\right)\right]^2+\max_{|\theta|\leq m}\sup_{x\in X}\left[\partial^m\left(g_n-g\right)\left(u\right)\right]^2 d_u}\,\|R_x\| \\
&\leq (m+1)\|g_n-g\|_m\|R_x\|
\end{aligned}$$

The first equality comes from property 3 of RKHs. The first three relationships use the fact that norm convergence in a RKHS implies pointwise convergence (Wahba (1990), pg.2). The second equality is the definition of the norm of the space in assumption 2. Assumption 1 bounds $\|R_x\|$ and the fact that $\|h\|_m$ converges in the norm guarantees the convergence of $\|h_n - h\|_{H_u}$.

□

Lemma 1 allows the use of the results of Newey (1994) and Newey (1997) in conjunction with the results of Lukas (1988) . Specifically, Lemma 1 facilitates the analysis of the convergence rate of equation (3.24) under the norm of $H_u$ or $W_u$ regardless of the fact that the convergence of some of its elements was studied under different norms.

Before stating assumptions 9-11 a connection between Newey (1994) and the object of study of this paper will be established. To simplify notation let $(\hat{\eta}_{z_1}(x), \ldots, \hat{\eta}_{z_n}(x)) \equiv \hat{\eta}$ and $\left(\hat{Q}_n + \alpha I\right)^{-1} \equiv \hat{Q}_\alpha$, similarly $(\eta_{z_1}(x), \ldots, \eta_{z_n}(x)) \equiv \eta$ and $(Q_n + \alpha I)^{-1} \equiv Q_\alpha$.

Expression (3.25) using the previous simplifications can be rewritten as:

$$\begin{aligned}
\|g^* - \tilde{g}\|_{W_u} &= \left\|\eta Q_\alpha E(y|z) * - \hat{\eta}\hat{Q}_\alpha Y(z)\right\|_{W_u} \\
&= \left\|\eta Q_\alpha E(y|z) * - \hat{\eta}\hat{Q}_\alpha Y(z) + \eta Q_\alpha E(y|z) - \eta Q_\alpha E(y|z) + \hat{\eta}\hat{Q}_\alpha E(y|z) - \hat{\eta}\hat{Q}_\alpha E(y|z)\right\|_{W_u} \\
&= \left\|\eta Q_\alpha \varepsilon_i + \hat{\eta}\hat{Q}_\alpha \left(E(y|z) - Y(z)\right) + E(y|z)\left(\hat{\eta}\hat{Q}_\alpha - \alpha Q_\alpha\right)\right\|_{W_u} \qquad (3.26)
\end{aligned}$$

The distance between $E(y|z) - \Upsilon(z)$ is studied by Newey (1997) and assumptions 6-8 can be used to ascertain its properties. On the other hand, Newey (1994) can help in the analysis $\hat{\eta}\hat{Q}_{\alpha} - \eta Q_{\alpha}$. In particular given that the matrices $\hat{\eta}\hat{Q}_{\alpha}$ and $\eta Q_{\alpha}$ are functions of $\hat{\eta}$ and $\eta$ and are invertible and bounded, to examine $\hat{\eta}\hat{Q}_{\alpha} - \eta Q_{\alpha}$ it suffices to look at $\hat{\eta}$ and $\eta$. Specifically, the following relationship will be used:

$$\begin{aligned}
\|e\hat{t}a - \eta\|_{W_u} &= \left\| \int_X \left[ \hat{f}\left(x'|z\right) - f\left(x'|z\right) \right] R\left(x, x'\right) d_x' \right\|_{W_u} \\
&\leq \left\| \hat{f} - f \right\|_m \left\| \int_X R\left(x, x'\right) d_x' \right\|_{W_u} \\
&= C \left\| \hat{f} - f \right\|_m \qquad (3.27)
\end{aligned}$$

The inequality in (3.27) applies the Sobolev supremum norm. This norm will be noted as $\|.\|_m$, where m denotes the number of derivatives assumed. It is important to notice that the reproducing kernel as an element of the RKH is bounded and so can be replaced by the constant $C$. From now on $C$ will be used to denote an arbitrary constant in different contexts. Therefore, the object to analyze becomes $\left\| \hat{f} - f \right\|_m$. This is where the work of Newey (1994) becomes relevant for the purpose of this study, as he derives convergence rates under the Sobolev supremum norm for derivatives of kernel estimators.

Newey (1994) studies the nonparametric estimation of functions of the form $\Lambda_0(z) = E\left(\rho|z\right) * \psi_0\left(z\right)$ where the approximation is represented by:

$$\hat{\Lambda}_0\left(z\right) = \frac{1}{nh}\sum_{i=1}^{n}\rho_i K\left(\frac{z-z_i}{h}\right) \tag{3.28}$$

In the expression above K is a kernel weight. The object of interest as is presented by Fan et al. (1996) originates in a similar framework. Fan et al. (1996) construct the conditional density using the following idea:

$$f\left(x|r\right) \approx E\left(N\left(\frac{X_i - x}{h_2}\right)|Z = r\right)$$

A Taylor expansion around $z \in Z$ gives:

$$f\left(x|r\right) \approx f\left(x|z\right) + \frac{\partial f\left(x|z\right)}{\partial z}\left(r-z\right) + \frac{\partial^2 f\left(x|z\right)}{\partial^2 z^2}\left(r-z\right)^2$$

Combining the previous two statements results in:

$$\begin{aligned} f\left(x|z\right) &\approx E\left(N\left(\frac{X_i - x}{h_2}\right)|Z = r\right) - \frac{\partial f\left(x|z\right)}{\partial z}\left(r-z\right) - \frac{\partial^2 f\left(x|z\right)}{\partial^2 z^2}\left(r-z\right)^2 \\ f\left(x|z\right) &\approx E\left(N\left(\frac{X_i - x}{h_2}\right)|Z = r\right)\Psi\left(z\right) \end{aligned} \tag{3.29}$$

In the expression above:

$$\Psi\left(z\right) \equiv \left[1 - \left(E\left(N\left(\frac{X_i - x}{h_2}\right)|Z = r\right)\right)^{-1}\left(\frac{\partial f\left(x|z\right)}{\partial z}\left(r - z\right) + \frac{\partial^2 f\left(x|z\right)}{\partial^2 z^2}\left(r - z\right)^2\right)\right]$$

Within this framework Fan et al. (1996) build their estimator. In analogy to Newey (1994) and using the Appendix of Fan et al. (1996), the following estimator can be constructed:

$$\begin{aligned}\hat{m}\left(x,z\right) &= \frac{1}{nh_1h_2}\sum_{i=1}^{n} W_0^n\left(\frac{Z_i - z}{h_1}\right)N\left(\frac{X_i - x}{h_2}\right) \\ &= = \frac{1}{nh_1h_2}\sum_{i=1}^{n} W_0^n\left(\frac{Z_i - z}{h_1}\right)\pi_i \qquad (3.30)\end{aligned}$$

Above $N\left(\frac{X_i - x}{h_2}\right) \equiv \pi_i$.

From the previous arguments a correspondence between the conditional density estimation in this paper and the work of Newey (1994) can be established. Explicitly, $W_0^n(.)$ is equivalent to $K(.)$ and $\pi_i$ to $\rho_i$. Likewise, equation (3.29) is analogous to $\Lambda_0\left(z\right) = E\left(\rho|z\right)\psi_0\left(z\right)$ where $E\left(N\left(\frac{X_i - x}{h_2}\right)|Z = r\right) \sim E(\rho|z)$, $\psi_0\left(z\right) \sim \Psi\left(z\right)$, and $\Lambda_0\left(z\right) \sim f\left(x|z\right)$.

The assumptions and arguments of Newey (1994) can be rephrased in terms of equation (3.29) and equation (3.30). As will be seen below two sets of assumptions are being made. The first set restricts the behavior of $W_0^n$. The second set imposes conditions on the moments of $N(.)$.

Using the remark after Theorem 1 in Fan et al. (1996) it can be noted that $h_1$ and $h_2$ are asymptotically equivalent. Therefore, without loss of generality for the convergence rate results we can write equation (3.30) as:

$$\hat{m}\left(x,z\right) = \frac{1}{nh^2}\sum_{i=1}^{n} W_0^n\left(\frac{Z_i - z}{h}\right)\pi_i \tag{3.31}$$

Assumptions 9-11, which come from Newey (1994), in this framework become:

**Assumption 9.** There are positive integers $\Delta$ and $\bar{\omega}$ such that $W_0^n\left(u\right)$ is differentiable of order $\Delta$, the derivatives of order $\Delta$ are Lipschitz, $W_0^n\left(u\right)$ is zero outside a bounded set, for all $m < \bar{\omega}$, $\int W_0^n\left(u\right)\left[\otimes_{l=1}^{m}\right] d_u = 0$ and $\int W_0^n\left(u\right) d_u = 1$.

**Assumption 10.** There is a non-negative integer $\zeta$ and an extension of $f\left(x|z\right)$ to all of $\mathcal{R}^k$ that is continuously differentiable to order $\zeta$ on $\mathcal{R}^k$.

**Assumption 11.** For $\mu \geq 4$, $E\left[\|\pi_i\|^{\mu}\right] < \infty$ and $E\left[\|\pi_i\|^{\mu}|z\right]\Psi\left(z\right)$ is bounded.

**Assumption 12.** The perturbation $varepsilon_i$ is independent of $E\left(y|z\right) - \Upsilon\left(z\right)$.

Assumption 12 is trying to impose that for $E\left(y|z_i\right)^* = E\left(y|z_i\right) + \varepsilon_i$ the perturbations are independent of $E(y|z_i)$ and via this channel independent from the error in the estimation of $E(y|z_i)$. In other words, the source of randomness of the observed $E(y|z_i)^*$ is unrelated to the error of the nonparametric estimation of $E(y|z)$.

Using the notation of this paper Lemma B.1 of Newey (1994) can be written as:

Suppose $E\left[\|\pi\|^{\mu}\right] < \infty$ for $\mu > 2$, $E\left[\|\pi\|^{\mu} |z\Psi(z)\right]$ is bounded, $Z$ is compact, assumption 9 is satisfied for $m \leq \Delta$, and $h = h(n)$ such that $h(n)$ is bounded and $\frac{n^{1-\frac{2}{\mu}}}{ln(n)} \to \infty$. Then:

$$\left\|\hat{f} - E(f)\right\|_m = O_p\left(ln(n)^2\left(nh^{k+2m}\right)^{-1/2}\right)$$

Likewise Lemma B.2 of **?** can be reworded as:

If Assumptions 9-11 are satisfied for $\zeta \geq m + \bar{\omega}$ then:

$$\left\|E\left(\hat{f}\right) - f\right\|_m = O\left(h^{\bar{\omega}}\right)$$

**Proposition 2.** If the conditions of Lemma B.1 and B.2 in Newey (1994) are satisfied and assumption 10 is fulfilled for $\zeta \geq m + \bar{\omega}$ then:

$$\left\|\hat{f} - f\right\|_m = O_p\left(ln\left(n\right)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

*Proof.* The proof follows from Lemma B.1 and Lemma B.2 in Newey (1994) using the triangle inequality □

**Proposition 3.**

$$\|\hat{\eta}\hat{Q}_\alpha - \eta Q_\alpha\|_m = O_p\left(ln\left(n\right)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

*Proof.* The outline of the proof was given in the discussion after equation (3.27). It follows by applying Proposition 2 to equation (3.28), by the fact that $\hat{Q}_\alpha$ and $Q_\alpha$ are invertible and bounded, and by the result that convergence in the Sobolev supremum norm implies convergence in the Sobolev norm (Lemma 1). □

The elements constructed so far can now be used to analyze equation (3.27). In particular the following result is true:

**Theorem 1.** If Assumptions 5-12 are satisfied, the conditions of Proposition 3 are fulfilled, and $\frac{L^3}{n} \to 0$ then:

$$\|g^* - \tilde{g}\|_{W_u} = O_p\left(\frac{L^{3/2}}{\sqrt{n}} + L^{1-\tau} + ln\left(n\right)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

*Proof.* Let us assume that Assumptions 5-12 are satisfied, the conditions of Proposition 3 are fulfilled, and $\frac{L^3}{n} \to 0$. Then:

$$\left\|\eta Q_\alpha \varepsilon_i + \hat{\eta}\hat{Q}_\alpha \left(E\left(y|z\right) - \Upsilon\left(z\right)\right) + E\left(y|z\right)\left(\hat{\eta}\hat{Q}_\alpha - \eta Q_\alpha\right)\right\|_{W_u}$$

$$\leq \left\|\eta Q_\alpha \varepsilon_i + \eta Q_\alpha \left(E\left(y|z\right) - \Upsilon\left(z\right)\right)\right\|_m + O_p\left(ln\left(n\right)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

In this first inequality the Sobolev supremum norm and the result of Proposition 3 are imposed. Since $\eta$ and $Q_\alpha$ are elements of a RKH they are bounded. This allows the inequality above to be written as:

$$\left\|C\varepsilon_i + C\left(E\left(y|z\right) - \Upsilon\left(z\right)\right)\right\|_m + O_p\left(ln\left(n\right)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

$$\leq \left\|C\varepsilon_i\right\|_m + E\left\|C\left(E\left(y|z\right) - \Upsilon\left(z\right)\right)\right\|_m + O_p\left(ln\left(n\right)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

Here Assumption 12 and the triangle inequality are used. Now merging the results of Newey (1997) for series estimators when $\frac{L^3}{n} \to 0$ and Assumption 5 it follows that:

$$\begin{aligned}
&\left\|C\varepsilon_i\right\|_m + E\left\|C\left(E\left(y|z\right) - \Upsilon\left(z\right)\right)\right\|_m + O_p\left(ln\left(n\right)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right) \\
\leq\ & O_p\left(\frac{L^{3/2}}{\sqrt{n}} + L^{1-\tau}\right) + O_p\left(ln\left(n\right)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right) + C^2\sigma_u^2 \\
=\ & O_p\left(\frac{L^{3/2}}{\sqrt{n}} + L^{1-\tau} + ln\left(n\right)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)
\end{aligned}$$

□

With the above results the distance from the true function the estimator proposed in this paper can be ascertained. Before that, it is important to notice that from the restatement of Lukas (1988) Theorem 2.1 using the notation of this study it follows that:

for $u \leq s \leq u+2$,

$$\|g_0 - g^*\|_{W_u} = O_p\left(\left(\alpha^{s-u}\left\|E\left(y|z\right)^*\right\|_s^2 + \tfrac{\sigma_u^2}{n}\alpha^{-u-1/2b}\right)^{1/2}\right)$$

and for $s \geq u+2$,

$$\|g_0 - g^*\|_{W_u} = O_p\left(\left(\alpha^{2}\left\|E\left(y|z\right)^*\right\|_s^2 + \tfrac{\sigma_u^2}{n}\alpha^{-u-1/2b}\right)^{1/2}\right)$$

Also, by Assumption 5, which bounds the conditional variance $E(y|z)$, and Assumption 6, that does the same for $E(y|z_i)^* - E(y|z_i)$, $E(y|z)^*$ $\sigma_u^2$ above can be taken to be constants.

Combining these results the convergence rates of the proposed estimator are obtained.

**Theorem 2.** If the conditions of Theorem 1 are fulfilled and $u \leq s \leq u+2$,

$$\|g_0 - \tilde{g}\|_{W_u} = O_p\left(\left(\alpha^{s-u} + \frac{\alpha^{-u-1/2b}}{n}\right)^{1/2} + \frac{L^{3/2}}{\sqrt{n}} + L^{1-\tau} + ln(n)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

for $s > u + 2$ :

$$\|g_0 - \tilde{g}\|_{W_u} = O_p\left(\left(\alpha^2 + \frac{\alpha^{-u-1/2b}}{n}\right)^{1/2} + \frac{L^{3/2}}{\sqrt{n}} + L^{1-\tau} + ln(n)^2\left(nh^{k+2m}\right)^{-1/2} + h^{\bar{\omega}}\right)$$

# Bibliography

Ai, C. and X. Chen (2003): “Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions,” *Econometrica*, 71, 1795–1843.

Altonji, J. and R. Matzkin (2005): “Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors,” *Econometrica*, 73, 1053–1102.

Angrist, J. D. and A. B. Krueger (1991): “Does Compulsory School Attendance Affect Schooling and Earnings?” *The Quarterly Journal of Economics*, 106, 979–1014.

Aronszajn, N. (1950): “Theory of Reproducing Kernels,” *Transactions of the American Mathematical Society*, 68, 337–404.

Becker, G. (1964): *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education*, New York, NY.: National Bureau of Economic Research.

BEKKER, P. A. (1994): “Alternative Approximations to the Distributions of Instrumental Variable Estimators,” *Econometrica*, 62, 657–681.

BELLONI, A. AND V. CHERNOZHUKOV (2010): “Post-*l1*-Penalized Estimators in High-Dimensional Linear Regression Models,” Working Paper, MIT.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2010a): “LASSO Methods for Gaussian Instrumental Variable Models,” Working Paper, MIT.

BELLONI, A., D.CHEN, V. CHERNOZHUKOV, AND C. HANSEN (2010b): “Sparse Models and Methods for Optimal Instruments,” Working Paper, MIT.

BELZIL, C. (2006): “Returns to Schooling in Structural Dynamic Models: A Survey,” Working Paper.

BOUND, J., D. A. JAEGER, AND R. M. BAKER (1995): “Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variables is Weak,” *Journal of the American Statistical Association*, 90, 443–450.

BROWN, D. AND R. MATZKIN (1998): “Estimation of Nonparametric Functions in Simultaneous Equations Models, with application to Consumer Demand,” Cowles Foundation Discussion Paper, Yale University.

CANER, M. AND Q. FAN (2010): “The Adaptive Lasso Method for Instrumental Variable Selection,” Working Paper, North Carolina State University.

CHAO, J. C., J. A. HAUSMAN, W. K. NEWEY, T. WOUTERSEN, AND N. R. SWANSON (2010): "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," Forthcoming *Econometric Theory*.

CHAO, J. C. AND N. R. SWANSON (2003a): "Asymptotic Normality of Single-Equation Estimators for the Case with a Large Number of Weak Instruments," Working Paper, Rutgers University.

——— (2003b): "Consistent Estimation with a Large Number of Weak Instruments," Working Paper, Yale University.

——— (2005a): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73, 1673–1692.

——— (2005b): "Consistent Estimation with a Large Number of Weak Instruments," *Econometrica*, 73, 1673–1692.

DAROLLES, S., J. FLORENS, AND E. RENAULT (2003): "Nonparametric Instrumental Regression," Working Paper.

DONALD, S. G. AND W. K. NEWEY (2001): "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191.

FAN, J., Q. YAO, AND H. TONG (1996): "Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems," *Biometrika*, 83, 189–206.

FU, W. J. (1998): "Penalized Regressions: The Bridge versus the Lasso," *Journal of Computational and Graphical Statistics*, 7, 397–416.

FULLER, W. E. (1977): "Some Properties of a Modification of the Limited Information Estimator," *Econometrica*, 45, 939–953.

GAGLIARDINI, P. AND O. SCAILLET (2006): "Tikhonov Regularization for Nonparametric Instrumental Variable Estimators," Working Paper.

GOLUB, G. H., M. HEATH, AND G. WAHBA (1979): "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215–223.

GRILICHES, Z. (1977): "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1–22.

HALL, P. AND J. HOROWITZ (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33, 2904–2929.

HANSEN, B. E. (2007): "Least Squares Model Averaging," *Econometrica*, 75, 1175–1189.

HAUSMAN, J. A., J. C. CHAO, , W. K. NEWEY, T. WOUTERSEN, AND N. R. SWANSON (2010): "Instrumental Variable Estimation with Heteroskedasticity and Many Instruments," Working Pape, MIT.

HUANG, J., J. L. HOROWITZ, AND F. WEI (2010): "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313.

HUANG, J., S. MA, AND C.-H. ZHANG (2007): "Adaptive Lasso for Sparse High-Dimensional Regression Models," Technical Report, University of Iowa.

KEANE, M. P. AND K. WOLPIN (1997): "The Career Decisions of Young Men," *Journal of Political Economy*, 45, 1–22.

KIMELDORF, G. AND G. WAHBA (1971): "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, 33, 82–95.

KOENKER, R. AND J. A. MACHADO (1999): "GMM inference when the number of moment conditions is large," *Journal of Econometrics*, 551–566.

KRESS, R. (1989): *Linear integral equations*, New York: Springer-Verlag.

KUERSTEINER, G. AND R. OKUI (2010): "Constructing Optimal Instruments by First-Stage Prediction Average," *Econometrica*, 78, 697–718.

LUKAS, M. A. (1988): "Convergence Rates for Regularized Solutions," *Mathematics of Computation*, 51, 107–131.

MINCER, J. (1958): "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy*, 66, 281–302.

——— (1974): *Schooling, Experience, and Earnings*, New York, NY.: National Bureau of Economic Research.

MORIMUNE, K. (1983): "Approximate Distributions of k-Class Estimators when the Degree of Overidentifiability is Large Compared with the Sample Size," *Econometrica*, 51, 821–841.

NEWEY, W. K. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233–253.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79, 147–168.

NEWEY, W. K. AND J. L. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71, 1565–1578.

NEWEY, W. K., J. L. POWELL, AND F. VELLA (1999): "Nonparametric Estimation of Triangular Simultaneous Equations Models," *Econometrica*, 67, 565–603.

NYCHKA, D., G. WAHBA, S. GOLDFARB, AND T. PUGH (1984): "Cross-Validated Spline Methods for the Estimation of Three-Dimensional Tumor Size Distributions From Observations on Two-Dimensional Cross Sections," *Journal of the American Statistical Association*, 79, 738–782.

PORTNOY, S. (1984): "Asymptotic Behavior of M-Estimators of $p$ Regression Parameters when $p^2/n$ is Large I. Consistency," *The Annals of Statistics*, 12, 551–566.

RIESZ, F. (1955): *Functional analysis*, New York,: Ungar, translation of Lecons d'analyse fonctionnelle. OCLC: (OCoLC)00527978.

STAIGER, D. AND J. H. STOCK (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

STOCK, J., J. WRIGHT, AND M. YOGO (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business and Economic*, 20, 1–22.

STOCK, J. H. AND F. TREBBI (2003): "Who Invented Instrumental Variable Regression," *Journal of Economic Perspectives*, 17, 2282–2313.

TIBSHIRANI, R. (1996): "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society.Series B (Methodological)*, 58, 267–288.

VAN DER VAART, W. AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*, New York, NY.: Springer Series in Statistics.

WAHBA, G. (1969): "On the Numerical Solution of Fredholm Integral Equations of the First Kind," Technical Report 990, Mathematics Research Center, University of Wisconsin, Madison.

——— (1973): "A class of approximate solutions to linear operator equations," *Journal of Approximation Theory*, 9, 61–77.

——— (1977): "Practical Approximate Solutions to Linear Operator Equations when the Data are Noisy," *SIAM Journal on Numerical Analysis*, 14, 651–667.

——— (1990): *Spline models for observational data*, Philadelphia, Pa.: Society for Industrial and Applied Mathematics.

WEI, F. AND J. HUANG (2008): "Consistent group selection in high-dimensional linear regression," Technical Report 387, Department of Statistics and Actuarial Science, University of Iowa.

ZHANG, C., Y. JIANG, AND Y. CHAI (2010): "Penalized Bregman Divergence for Large Dimensional Regression and Classification," *Biometrika*, 97, 551–566.

ZHANG, C.-H. AND J. HUANG (2008): "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594.

ZOU, H. (2006): "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.