Instrumental Variables Selection: a Comparison between Regularization and Post-Regularization Methods

by

Chiara Di Gravio

B.Sc. Statistics, Finance and Insurance, Sapienza University of Rome, 2009
M.Sc in Actuarial and Financial Science, Sapienza University of Rome, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Statistics)

The University of British Columbia
(Vancouver)

August 2015

© Chiara Di Gravio, 2015
Abstract

Instrumental variables are commonly used in statistics, econometrics, and epidemiology to obtain consistent parameter estimates in regression models when some of the predictors are correlated with the error term. However, the properties of these estimators are sensitive to the choice of valid instruments. Since in many applications, valid instruments come in a bigger set that includes also weak and possibly irrelevant instruments, the researcher needs to select a smaller subset of variables that are relevant and strongly correlated with the predictors in the model.

This thesis reviews part of the instrumental variables literature, examines the problems caused by having many potential instruments, and uses different variables selection methods in order to identify the relevant instruments. Specifically, the performance of different techniques is compared by looking at the number of relevant variables correctly detected, and at the root mean square error of the regression coefficients’ estimate. Simulation studies are conducted to evaluate the performance of the described methods.
Preface

This work was prepared under the supervision of Professor Gabriela Cohen Freue. The research question was earlier identified by Professor Gabriela Cohen Freue. Based on that, I compared already existing techniques with new approaches, and implemented the simulation studies reported in this thesis.
# Table of Contents

Abstract .......................................................... ii  
Preface ......................................................... iii  
Table of Contents ........................................ iv  
List of Tables ............................................... vi  
List of Figures ............................................. viii  
Acknowledgments ........................................... xi  
1 Introduction .............................................. 1  
2 Instrumental Variables: Two Stage Least Squares and Instruments  
   Selection .................................................. 5  
   2.1 Introduction .......................................... 5  
   2.2 Endogeneity and Instrumental Variables .......... 6  
   2.3 Two Stage Least Squares (2SLS) ................. 9  
   2.4 Instrumental Variables Problems ............... 12
List of Tables

Table 4.1  Coefficient RMSE for the simulated examples and for different instruments strength. $Cor(\varepsilon, v) = 0.3$. For the regularization and the post-regularization methods, the penalty parameter $\lambda$ has been chosen such that the cross-validation error is within one standard error from the minimum. 37

Table 4.2  Coefficient RMSE for the simulated examples and for different instruments strength. $Cor(\varepsilon, v) = 0.6$. For the regularization and the post-regularization methods, the penalty parameter $\lambda$ has been chosen such that the cross-validation error is within one standard error from the minimum. 38

Table 4.3  Average number of variables selected by each method. $Cor(\varepsilon, v) = 0.3$. For the regularization and the post-regularization methods, the penalty parameter $\lambda$ has been chosen such that the cross-validation error is within one standard error from the minimum. 39
Table 4.4 Average number of variables selected by each method. \( \text{Cor}(\varepsilon, v) = 0.6 \). For the regularization and the post-regularization methods, the penalty parameter \( \lambda \) has been chosen such that the cross-validation error is within one standard error from the minimum. 40

Table 4.5 Coefficient RMSE for the simulated examples and for different instruments strength. \( \text{Cor}(\varepsilon, v) = 0.3 \). For the regularization and the post-regularization methods, the penalty parameter \( \lambda \) has been chosen such that the cross-validation error is within one standard error from the minimum. 44

Table 4.6 Coefficient RMSE for the simulated examples and for different instruments strength. \( \text{Cor}(\varepsilon, v) = 0.6 \). For the regularization and the post-regularization methods, the penalty parameter \( \lambda \) has been chosen such that the cross-validation error is within one standard error from the minimum. 45

Table 4.7 Average number of variables selected by each method. \( \text{Cor}(\varepsilon, v) = 0.3 \). For the regularization and the post-regularization methods, the penalty parameter \( \lambda \) has been chosen such that the cross-validation error is within one standard error from the minimum. 46

Table 4.8 Average number of variables selected by each method. \( \text{Cor}(\varepsilon, v) = 0.6 \). For the regularization and the post-regularization methods, the penalty parameter \( \lambda \) has been chosen such that the cross-validation error is within one standard error from the minimum. 47
List of Figures

Figure 2.1 Mean estimates of $\hat{\beta}$ (dot) for different levels of $r^2_{xz}$. The dotted line represents the true value of $\beta$. The bars show the standard error of the estimates. ................................. 12

Figure 3.1 Regularization methods coefficients path for a real data example. The ratio of the sum of the absolute current estimate over the sum of the absolute OLS estimates is on the x-axis: as we move towards zero, more coefficients are excluded from the model. The standardized coefficients computed by multiplying the original coefficients by the norm of the predictors are on the y-axis. ................................................................. 24
Figure 4.1 Trend of the second stage coefficient’s RMSE for different values of $\lambda$. The data is generated based on Example 4, $\text{Cor}(\varepsilon, v) = 0.6$, and the first stage F-statistic is set to 40 (medium strength instruments). A grid of $\lambda$s is generated, and for each value in the grid $N = 500$ datasets are simulated. For each $\lambda$, the RMSE plotted on the y-axis is computed as the mean of the 500 replication. The black vertical line indicates the value of $\log(\lambda)$ such that the cross-validation error in the first stage is within one standard deviation of its minimum.

Figure 4.2 Trend of the second stage coefficient’s RMSE for different values of $\lambda$. The data is generated based on Example 8, $\text{Cor}(\varepsilon, v) = 0.6$, and the first stage F-statistic is set to 40 (medium strength instruments). A grid of $\lambda$s is generated, and for each value in the grid $N = 500$ datasets are simulated. For each $\lambda$, the RMSE plotted on the y-axis is computed as the mean of the 500 replication. The black vertical line indicates the value of $\log(\lambda)$ such that the cross-validation error in the first stage is within one standard deviation of its minimum.
Figure 4.3  Trend of the average number of variables selected for different values of $\lambda$. The data is generated based on Example 4, $Cor(\varepsilon, \nu) = 0.6$, the first stage F-statistic is set to 40 (medium strength instruments), and the Lasso is used to select the relevant instruments. A grid of $\lambda$s is generated, and for each value in the grid $N = 500$ datasets are simulated. For each $\lambda$, the number of variable selected, the true positive and the false positive plotted on the y-axis are computed as the mean of the 500 replication. The black vertical line indicates the value of $\log(\lambda)$ such that the cross-validation error in the first stage is within one standard deviation of its minimum.

Figure 4.4  Trend of the average number of variables selected for different values of $\lambda$. The data is generated based on Example 8, $Cor(\varepsilon, \nu) = 0.6$, the first stage F-statistic is set to 40 (medium strength instruments), and the Lasso is used to select the relevant instruments. A grid of $\lambda$s is generated, and for each value in the grid $N = 500$ datasets are simulated. For each $\lambda$, the number of variable selected, the true positive and the false positive plotted on the y-axis are computed as the mean of the 500 replication. The black vertical line indicates the value of $\log(\lambda)$ such that the cross-validation error in the first stage is within one standard deviation of its minimum.
Acknowledgments

First, I would like to thank my supervisor Professor Gabriela Cohen Freue for her support and guidance. Her constant encouragement, valuable comments and precious advice helped me not only to improve this thesis, but also to become a better researcher. Thanks Gaby, for being extremely patient and for dedicating so much time teaching me how to write my computer codes properly.

My time at UBC has been wonderful thanks to the people I met. In these past two years I have encountered incredible people in the Department of Statistics, I would like to thank them for every conversation, and every laughter we had in and outside the Department. You definitely make these past two years an unforgettable journey.

Most of all, I would like to thank my parents. They are the most amazing people I know, they have been providing me with so many opportunities, and they have been always encouraging me to follow my dreams. They are the ones that made all of this work possible.
Chapter 1

Introduction

The recent development of genomic technology has allowed researchers to carry out various studies in order to understand the molecular complexity of any disease process, and to develop tools for helping treatment, detection and prevention of a specific disease. Molecular biomarkers, such as genes or proteins, are measurable molecular indicators of the presence or progress of a particular disease, or the effect of a given treatment. Due to the ability of molecular biomarkers to provide an early indication of the disease, the number of biomarkers discovery studies has been increasing over the years. However, since most of these studies are characterized by many variables among which only a smaller set is relevant, new statistical methods “tailored for the analysis of -omics data” have been growing in the last decade.

The study of biomarkers can be carried out using either univariate or multivariate statistical methods [14]. Univariate methods can be used to identify significant biomarkers by considering each biomarker one at a time (e.g. Student’s t-test and non-parametric tests). Multivariate methods identify relevant biomarkers by considering the relationships existing between the candidate molecules. Among
univariate methods, ordinary least squares (OLS) might be used to discover which biomarkers are related to a particular disease. Specifically, assuming a linear relation between the potential biomarkers and the presence of the disease studied, a feasible model is represented by:

\[ y_i = x_i \beta + \epsilon_i, \]

\( i = 1, \ldots, n \)

where \( y \) is a continuous variable indicating the state of the disease of interest and \( x \) is a measurement of the considered biomarker. For instance, in studying Type 2 diabetes, it is possible to build a model where insulin resistance is the response variable, and potential molecular biomarkers are considered as predictors [11].

However, in this setting, the use of the OLS regression presents different problems that need to be addressed. First, some of the potential biomarkers that are included as predictors could have been measured with error; for instance, errors related to the machinery used, or to the collection and storage of the samples. Second, some of the potential biomarkers might not be observed, so it is possible to incorrectly exclude from the model predictors correlated with the ones considered (omitted variables). In both the cases of measurement error and omitted variables, the OLS estimates are biased and inconsistent. Instrumental variables estimators are commonly used in statistics, econometrics and epidemiology to address this problem.

Instrumental variables estimators consistently estimate the regression coefficients of the predictors by using new variables, not included in the model, called instruments. To be considered an instrument, a variable \( Z \) needs to be such that
changes in $Z$ are associated with changes in the predictor $X$, but do not lead directly to changes in the response $y$; in other words, $Z$ is an instrument if it is associated with the predictor $X$, but not with the error term $u$. This leads to:

$$Z \rightarrow X \rightarrow y \uparrow \downarrow u$$

Permutt and Hebel in [13] measured the effect of maternal smoking on their child’s birth weight, using an encouragement to stop smoking as the instrumental variable. Moreover, mendelian randomization studies can be seen as an application of instrumental variables; in fact, mendelian randomization uses genotype as an instrument for the exposure of interest. As pointed out by Didilez and Sheehan in [5], the genotype can be considered an instrument since it only affects the disease status indirectly, it is assigned randomly (given the parents genes), and it does not depend on any possible confounding. Furthermore, in proteomics biomarkers discovery studies, proteins are usually measured with error due to the technology use to quantify them. Since protein levels are regulated in part by gene expression, genes might also be used as instrumental variables when looking for molecular biomarkers.

While instrumental variables estimators provide consistent parameter estimates, their properties are sensitive to the choice of valid instruments. Since in many applications, valid instruments come in a bigger set that includes also weak and possibly irrelevant instruments, the researcher needs to select a smaller subset of variables that are relevant and strongly correlated with the predictors in the model.
This thesis examines the problems caused by having many potential instruments, and uses different selection methods to identify the relevant ones. In particular, it compares the performance of the different techniques used in terms of number of instruments selected and model’s coefficients’ estimate.

The thesis is organized as follows. In Chapter 2, we present background of instrumental variables and we define a classical instrumental variables estimator: the two stage least squares (2SLS). We highlight the problem of having to choose valid instruments among a bigger set, and we consider instrumental variables selection as a feasible solution. In Chapter 3, we discuss a wide variety of selection methods; specifically, we review the Lasso, the elastic net, the adaptive Lasso, the post-Lasso and the supervised principal component analysis (SPCA). We also propose a methodology called post-EN/L2. Finally, in Chapter 4 we report a series of simulation studies designed to compare different selection method in terms of selecting the right subset of instruments and reducing the model’s coefficients’ bias.
Chapter 2

Instrumental Variables: Two Stage Least Squares and Instruments Selection

2.1 Introduction

Chapter 1 briefly introduces the consequences of fitting an OLS regression when the predictors in the model are measured with error or when potential confounding variables are not considered. In both cases, the OLS estimates are biased and inconsistent. Instrumental variables estimators can solve the inconsistency problem by using variables not included in the model, called instruments, to estimate the coefficients in the OLS regression. This chapter presents in detail a classical instrumental variables estimator, and illustrates some of the issues related to having numerous instruments from which the researcher needs to select a smaller subset.
Specifically, Section 2.2 introduces the concept of endogeneity and instrumental variables. Section 2.3 describes the two stage least squares (2SLS) as a possible method to consistently estimate the coefficient in the model. Section 2.4 explains the problems of having weak and irrelevant instruments. Finally, Section 2.5 suggests instrumental variables selection as a method to improve the model’s coefficients’ estimate when the set of available variables includes relevant, weak and possibly irrelevant instruments.

### 2.2 Endogeneity and Instrumental Variables

Consider the model:

\[ y_i = x_i^\ast \beta + u_i, \quad (2.1) \]

\[ i = 1, \ldots, n, \]

\[ \mathbb{E}[u] = 0 \quad \text{and} \quad \text{Var}(u) = \sigma_u. \]

where \( y_i \) is the response variable and \( x_i^\ast \) is a single predictor. Under certain regularity conditions, the classical OLS regression produces unbiased and consistent estimates. One important assumption is that the predictors in the model are uncorrelated with the error term: \( \mathbb{E}[x_i^\ast u_i] = 0 \). When the aforementioned condition is not met, then the OLS estimate is biased and inconsistent, and the predictor correlated with the error term is called endogenous.

Endogeneity originates in several contexts: when the predictors are measured with error (errors in variables), when they are correlated with unobserved predictors (omitted variables), or when they simultaneously affect and are affected by
the response variable (simultaneity). As the consequences of endogeneity are the same regardless of the possible causes, this section only focuses on the endogeneity caused by measurement error.

Suppose that, while (2.1) represents the true model, the predictor $x_i^\ast$ cannot be observed. Specifically, the best the researcher can do is to observe a noisy measure of $x_i^\ast$:

$$x_i = x_i^\ast + v_i,$$  \hspace{1cm} (2.2)

where $v$ is the measurement error with variance $\sigma_v^2$. Then, the model in (2.1) becomes:

$$y_i = x_i^\ast \beta + u_i = (x_i - v_i) \beta + u_i = x_i \beta + (u_i - v_i \beta).$$ \hspace{1cm} (2.3)

If $v_i$ and $u_i$ are assumed to be independent, then the predictor $x_i$ is endogenous as it is correlated with the error term $\varepsilon_i$:

$$\mathbb{E}[x_i \varepsilon_i] = \mathbb{E}[(x_i^\ast + v_i)(u_i - v_i \beta)] = -\beta \sigma_v^2.$$ \hspace{1cm} (2.4)

It can be easily proved that the OLS estimator $\hat{\beta}_{OLS} = \sum i \frac{x_i y_i}{x_i^2}$ resulting from regressing $y_i$ on $x_i$ is biased:

$$bias \left( \hat{\beta}_{OLS} \right) = \mathbb{E} \left[ \hat{\beta}_{OLS} \right] - \beta = \mathbb{E} \left[ \frac{\sum_i x_i \varepsilon_i}{\sum_i x_i^2} \right] \approx \frac{Cov(X, \varepsilon)}{\sigma_x^2},$$ \hspace{1cm} (2.5)
and inconsistent:

\[ \text{plim} \hat{\beta}_{OLS} = \frac{\text{plim} \frac{1}{n} \left( \sum_i x_i y_i \right)}{\text{plim} \frac{1}{n} \left( \sum_i x_i^2 \right)} = \frac{\beta}{1 + (\sigma_e^2 / \sigma_x^2)} \neq \beta. \]  

(2.6)

To be able to consistently estimate the regression coefficient of the endogenous predictor, it is possible to use new variables, not included in the model, called instruments. A set of variables can be considered as potential instruments if they satisfy two important conditions: first, they need to be correlated with the endogenous predictor \( X \); second, they have to be uncorrelated with the error term \( \varepsilon \). The first condition requires that there is some sort of association between the instruments and the endogenous predictor. The second condition excludes the possibility that the instruments are predictors for \( y \); in fact, if both \( X \) and \( Z \) are predictors for \( y \), and \( y \) is only regressed on \( X \), then \( Z \) would have been absorbed into the error term \( \varepsilon \), and \( \mathbb{E}[Z^T \varepsilon] \neq 0 \). If an instrument fails to satisfy the first condition, then the instrument is said to be irrelevant; however, if an instrument satisfies the first condition, but it is only weakly correlated with the endogenous variable, then the instrument is defined as weak. If an instrument satisfies both the conditions, then the coefficient of the endogenous predictor can be consistently estimated using a classical instrumental variables estimator called two stage least squares (2SLS). In what follow, we consider one endogenous predictor \( X \) and \( p \) instruments; however, the results obtained in the next sections can be easily extended to the case where multiple endogenous predictors are included in the model and to the case where both endogenous and non endogenous predictors are considered. To ease the notation, we omit the subscript \( i \) and we write all the formulas using matrix notation.
2.3 Two Stage Least Squares (2SLS)

Given the model in (2.3) and a set of valid instrument $Z$, it is possible to estimate the predictor’s coefficient $\beta$ by fitting a procedure called two stage least squares (2SLS). Formally, assuming that the predictor $X$ and the instruments $Z$ are linearly related, we can extend the model (2.3) to:

$$y = X\beta + \epsilon$$  \hspace{1cm} (2.7)

$$X = Z\Pi + w.$$  \hspace{1cm} (2.8)

where $y$ ($n \times 1$) is the response variable, $X$ ($n \times 1$) is the endogenous predictor, $Z$ ($n \times p$) is the matrix of $p$ instruments, $\Pi$ is the vector of coefficients associated with each instruments, $E[X^T \epsilon] \neq 0$, $E[Z^T w] = E[Z^T \epsilon] = 0$, $Var(\epsilon) = \sigma_\epsilon^2$, $Var(w) = \sigma_w^2$ and $Cov(\epsilon, w) = \sigma_{w\epsilon}$ measures the endogeneity level. Without loss of generality it is possible to assume that the endogenous variable and the instruments, and the response variable are centered.

The idea behind 2SLS is to first use the model in (2.8) to separate the endogenous predictor $X$ in two components: the non problematic component $Z\Pi$ ($Z$ is chosen such that it is independent of $\epsilon$), and the problematic one $w$ ($w$ and $\epsilon$ are correlated). Then, 2SLS aims to estimate $\beta$ in (2.7) by using $Z\Pi$ only. However, since $\Pi$ is not known a priori, the non problematic component cannot be computed exactly and needs to be estimated using an OLS regression. Hence, this method as been referred to as 2SLS. In other words, to estimate $\beta$, first, $X$ is regressed on $Z$ and the predicted $\hat{X}$ is computed as $\hat{X} = Z\hat{\Pi}$; then, $y$ is regressed on $\hat{X}$ and the
coefficient $\beta$ is estimated as:

$$\hat{\beta}_{2SLS} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T \hat{Y} = (X^T P_Z X)^{-1} X^T P_Z Y$$  \hspace{1cm} (2.9)$$

where $P_Z = Z (Z^T Z)^{-1} Z^T$. Using the coefficient’s estimate in (2.9) and replacing $X$ with the equation in (2.8), the approximate bias of the 2SLS estimates is [9]:

$$\mathbb{E} [\hat{\beta}_{2SLS}] - \beta \approx \frac{\sigma_{we} p}{\Pi^T Z^T Z \Pi + \rho \sigma_e^2} \left( \frac{\Pi^T Z^T Z \Pi}{\rho \sigma_e^2} + 1 \right)^{-1}. \hspace{1cm} (2.10)$$

Thus, $\hat{\beta}_{2SLS}$ is still biased. However, it can be proved that $\hat{\beta}_{2SLS}$ is consistent and asymptotically normal. Since $\hat{\Sigma}_{xz} = \frac{1}{n} X^T Z \rightarrow \Sigma_{xz}$ (by the law of large numbers),

$$\text{plim} \hat{\beta}_{2SLS} = \text{plim} \left( (X^T P_z X)^{-1} X^T P_Z Y \right) = \text{plim} \left( (X^T P_z X)^{-1} X^T P_Z (X \beta + \epsilon) \right)$$

$$= \beta + \text{plim} \left( \left( \frac{X^T Z}{n} \right) \left( \frac{Z^T Z}{n} \right)^{-1} \left( \frac{Z^T X}{n} \right)^{-1} \left( \frac{Z^T \epsilon}{n} \right) \right)$$

$$= \beta + \text{plim} \left( \Sigma_{xz}^{-1} \Sigma_{xx}^{-1} \Sigma_{cx} \right)$$

$$= \beta + \Sigma_{xz}^{-1} \Sigma_{cc}^{-1} \Sigma_{cx}$$

$$= \beta. \hspace{1cm} (2.11)$$

Moreover,

$$\hat{\beta}_{2SLS} \xrightarrow{d} N \left( \beta, \frac{\omega^2}{n} \right), \hspace{1cm} (2.12)$$

where the asymptotic variance is given by $\omega^2 = \sigma_e^2 \left[ \Sigma_{xz} \Sigma_{cc}^{-1} \Sigma_{xz} \right]^{-1}$.

The asymptotic variance of $\hat{\beta}_{2SLS}$ plays a primary role in understanding the precision of the 2SLS estimate. For simplicity, considering only one instrument,
\( \omega^2 \) can be rewritten as:

\[
\omega^2 = \text{Var} \left( \hat{\beta}_{2SLS} \right) = \sigma_e^2 \left[ \Sigma_{xx}^{-1} \Sigma_{xz} \right]^{-1} = \frac{\sigma_e^2}{\sigma_{\hat{z}}^2 \rho_{xz}^2}, \tag{2.13}
\]

where \( \rho_{xz}^2 \) is the correlation between the endogenous variable \( X \) and the instrument \( Z \), it is clear how the precision of \( \hat{\beta}_{2SLS} \) increases when the instrument included in the model is strongly correlated with the endogenous variable. Similarly, if only one instrument is considered, the approximate bias of the 2SLS estimate in (2.10) can be rewritten as:

\[
E[\hat{\beta}_{2SLS}] - \beta \approx \frac{\sigma_{we}}{\sigma_w^2} \left( \frac{\pi^2 \sigma_x^2}{\sigma_w^2} + 1 \right)^{-1} = \frac{\sigma_{we}}{\sigma_w^2} \left[ \left( \frac{\rho_{xz}^2}{\sigma_x^2} \right) + 1 \right]^{-1}. \tag{2.14}
\]

Hence, the estimate’s bias decreases when the instruments are strongly associated with the predictor. Therefore, in order to have a more precise and a less biased estimate, the researcher needs to carefully individuate which of the available instruments is highly correlated with the endogenous predictor. Figure 2.1 shows how the variability and the bias of the 2SLS estimate decrease when the correlation between the endogenous predictor \( X \) and the instrumental variable \( Z \) increases. In this case, we considered one endogenous predictor and one instrument \( Z \) and we generated a sample size of \( n = 100 \); then, for each level of association, we simulated \( N = 200 \) datasets and we estimated \( \beta_{2SLS} \) for each datasets.
2.4 Instrumental Variables Problems

Section 2.3 briefly introduces the problem of having to choose the optimal instrument (or set of instruments). First, it is essential to choose a variable \( Z \) that is not endogenous (i.e., \( Z \) has to be uncorrelated with the error term \( \varepsilon \)); second, the variable \( Z \) needs to be correlated with the endogenous predictor \( X \); finally, the correlation between \( X \) and \( Z \) needs to be strong. Moreover, it is important to keep in mind that the use of the instrumental variables estimator decreases the precision of the coefficients’ estimate when compared to the OLS.

In order to facilitate the choice of whether it is worth using the instrumen-
tal variables estimator, and which instruments to include in the model, the instru-
mental variables literature provides different tests that are useful for understanding
whether a specific variable is endogenous, and if potential instruments are only
weakly correlated with the endogenous predictor. This section discusses one of the
most common tests used to check for endogeneity, provides a formal definition of
weak instruments, and it outlines some of the issues that originate from including
weak instruments in the model.

2.4.1 The Hausman Test for Endogeneity

Even though the instrumental variables estimator removes the inconsistency of the
OLS, it causes a loss of efficiency that needs to be taken into account if the re-
searcher suspects that the predictor in the model is not endogenous. The Hausman
test for endogeneity, introduced by J.A. Hausman in [10], is essentially a test of
whether this loss in efficiency is worth removing the inconsistency of the OLS es-
timator. Formally, given the model discussed in Section 2.3, the test is based on
the following hypotheses:

\[ H_0 : \text{Cov}(X, \varepsilon) = 0 \]

\[ H_a : \text{Cov}(X, \varepsilon) \neq 0. \]

and it involves fitting the model using both the instrumental variables and the OLS
regression, and comparing a weighted square of the difference between the two
estimators. Under the null hypothesis, both the OLS and the instrumental variables
estimators are consistent; hence, if the sample size is large, the difference between
them converges to zero. That is:

\[
d = \frac{\left(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS}\right)^2}{\text{Var}(\hat{\beta}_{2SLS}) - \text{Var}(\hat{\beta}_{OLS})} \to 0,
\]

(2.15)

where it can be proved that \(d \sim_{\text{approx.}} \chi^2\). Therefore, if \(d\) is significantly different from zero, then the use of instrumental variables is justified; otherwise, if \(d\) is not significantly different from zero, then the predictor \(X\) might not be endogenous and it is not worth losing estimate’s precision by using 2SLS. While this test is extremely simple to carry out, it is based on the assumption that the available sample size is large; therefore, it might work poorly in small sample problems.

### 2.4.2 Weak Instruments: Definition and Finite Sample Bias

Given the model discussed in Section 2.3, once the researcher proves the presence of endogeneity, he/she needs to identify one or more instruments to include. Selecting a set of relevant instruments that satisfies the two conditions listed in Section 2.2, it is not a simple task. In practice, since testing for the absence of correlation between \(Z\) and \(e\) presents numerous difficulties, most of the instruments are included solely based on what the researcher believes to be appropriate [15]; however, for every instrument considered in the model it is possible to easily quantify its strength.

A practical approach used to measure the strength of a set of instruments looks at the first stage F-statistic testing the joint significance of the coefficients of the instruments. Specifically, given the model in (2.8) and regardless of the available sample size, if the first stage F-statistic is less than 10, then the instruments are classified as weak [15]. The definition of weakness provided in [15] needs to
be viewed as a simple rule of thumb, and it is just one among the many available. While many authors have argued against the use of the first stage F-statistic as a measure of instruments weakness, this rule is still widely used in many applications; consequently, we will adopt it in this thesis as a benchmark for identifying weak instruments.

It is essential to understand the consequences of including weak instruments in a 2SLS model. Section 2.3 shows how using instruments that are only weakly correlated to the endogenous variable decreases the precision of the estimate; however, additional problems might affect the results. In [4], Bound et al. identify two main issues caused by the use of weak instruments: the inconsistency of \( \hat{\beta}_{2SLS} \), and the so-called finite sample bias. To understand the inconsistency problem, it is fundamental to remember that in all cases the lack of correlation between the instrument in the model and the error term \( e \) cannot be formally tested. If we consider an instrumental variables estimator, intuition might suggest that, when the correlation between \( e \) and the instruments is much lower than the correlation between the \( e \) and the endogenous variable \( X \), the bias of the 2SLS estimator should be lower than the bias of OLS; however, this intuition can be seriously wrong when the instruments in the model are only weakly correlated with the endogenous predictor. More formally, comparing the probability limit:

\[
\text{plim} \hat{\beta}_{2SLS} = \beta + \frac{\text{plim} \frac{\hat{X}^T \epsilon}{n}}{\text{plim} \frac{\hat{X}^T X}{n}} , \quad (2.16)
\]

with:
\[
\text{plim} \hat{\beta}_{OLS} = \beta + \frac{\text{plim} X^T \epsilon}{\text{plim} X^T X_n}, \tag{2.17}
\]

clearly shows how a high covariance between \( X \) and \( \epsilon \) could translate into modest inconsistency for \( \hat{\beta}_{OLS} \) if the variance of \( X \) is sufficiently large, while a small covariance between \( \hat{X} \) and \( \epsilon \) could translate into a large inconsistency for \( \hat{\beta}_{2SLS} \) if the covariance between \( \hat{X} \) and \( X \) is very small (i.e., if the instruments are weak).

The finite sample bias problem can be easily understood by rewriting the bias in (2.10) as:

\[
\text{bias} \left( \hat{\beta}_{2SLS} \right) = \mathbb{E}[\hat{\beta}_{2SLS}] - \beta \approx \frac{\sigma_{we}}{\sigma_w^2} \left( \frac{1}{F + 1} \right), \tag{2.18}
\]

where \( F = \frac{\hat{F}^T \hat{Z} \hat{Z}^T \hat{F}}{p \sigma_w^2} \) is the value of first stage F-statistic. In this case, in fact, it can be noted that if \( \sigma_x^2 = \sigma_w^2 \), and the F-statistic is equal to zero (i.e. the instrument used does not provide any information about the endogenous predictor) then:

\[
\text{bias} \left( \hat{\beta}_{2SLS} \right) = \mathbb{E}[\hat{\beta}_{2SLS}] - \beta \approx \frac{\sigma_{we}}{\sigma_w^2} = \frac{\sigma_{we}}{\sigma_x^2} \approx \mathbb{E}[\hat{\beta}_{OLS}] - \beta = \text{bias} \left( \hat{\beta}_{OLS} \right). \tag{2.19}
\]

Thus, when the F-statistics approaches to 0 (i.e. the instrument included provides little information about the endogenous predictor), then the 2SLS coefficient’s bias increases and it approaches to the OLS coefficient’s bias.

The use of instruments that are only weakly correlated with the endogenous predictor is not the only source of finite sample bias. The bias can also arise when there are “too many” instruments in the model compared to the sample size \( n \). For
instance, given the model summarized in (2.7) and (2.8), if $n$ valid instruments are included, then, in the first stage regression, the number of available parameters will be equal to the available data points. Consequently, a perfect fit will be obtained and the predicted values $\hat{X}$ will be equal to the actual values of the endogenous variable $X$. In this scenario, the second stage regression coincides with the standard OLS regression; hence, the estimated coefficient of the endogenous variable is biased and inconsistent. In other words, as the number of instruments approaches the sample size, the 2SLS estimator tends to be equal to the OLS estimator.

A feasible solution that can be used to estimate the coefficient $\beta$, and to reduce finite sample bias is to select a smaller set of instruments that are relevant and strongly correlated with the endogenous predictor. As different selection methods can be used, it is essential to understand which is the most appropriate based on the estimated bias of the endogenous predictor’s coefficient’s, and on the ability of the method to select the “right” set of instruments.

### 2.5 Choosing the Relevant Instruments: Instrumental Variables Selection

Section 2.4.2 briefly suggests instrumental variables selection as a possible method to solve the problems related to having to choose the right variables among a large set of instruments that might be relevant, weak or possibly irrelevant. This section aims to further explore the reasons behind instrumental variables selection, to describe briefly how the selection can be done, and to explain how selecting variables might reduce the finite sample bias.

While introducing the 2SLS estimator, one of the assumptions we made was the availability of a set of relevant instruments $Z$; however, we would like to extend the
estimator discussed in Section 2.3 by taking into account both relevant and possibly irrelevant instruments. In order to reach our objective we still adopt the framework presented in (2.7) and (2.8), and we define the relevance of each instrument based on the vector of coefficients $\Pi$; specifically:

- The instrument $i$ is relevant if $\pi_i \neq 0$ (i.e., there is a linear association between the instrument $i$ and the endogenous predictor $X$).
- The instrument $i$ is irrelevant if $\pi_i = 0$ (i.e., there is not a linear association between the instrument $i$ and the endogenous predictor $X$).

Since the elements of the vector $\Pi$ are not known in advance, the 2SLS procedure estimates them by regressing the endogenous predictor $X$ on the set of instruments $Z$; however, as in this new framework we would like some of the $\hat{\pi}_i$ to be exactly zero, we need to resort to a variables selection method to be able to keep in the model only the relevant instruments and to discard the rest.

As introduced by Belloni et al. in [3], it is possible to use different selection methods in order to choose a set of relevant instruments and to optimally predict $\hat{X}$. In this thesis, we look at selection methods that work well in both the case where many instruments are relevant and the case where most of the instruments are irrelevant (i.e., sparse case); in particular, we focus on three main categories of selection methods: regularization (Lasso, elastic net, adaptive Lasso), post-regularization (post-Lasso, post-EN/L2), and principal components (supervised principal component analysis).

Before introducing each method in detail, it is important to highlight that while most of the regularization methods shrink the estimated coefficients so that the prediction error of the out-of-sample data is minimized, the aim for our problem is
to select only the relevant instruments such that the estimated endogenous predictor $\hat{X}$ is as free as possible from noise and the coefficient $\beta$ of the endogenous variable can be estimated consistently.
Chapter 3

First Stage: Instrumental Variables Selection

3.1 Introduction

As shown in Chapter 2, the properties of the 2SLS are sensitive to the choice of valid instruments; consequently, when weak instruments are considered, or when too many instruments are included in the model, the traditional 2SLS might lead to inconsistent and highly biased estimates. Moreover, as briefly introduced in the previous chapter, there could be situations where the number of potential instruments $p$ is greater than the sample size $n$; in these settings, 2SLS breaks down and new methods need to be employed in order to estimate the model’s parameters.

For instance, suppose we are carrying out a proteomics biomarker study and we are considering genes as instruments. For each protein, we have a great number of potential instruments, among which we need to search for the most relevant ones. In addition, the large number of instruments is often accompanied by a small
sample (i.e., $n < p$). In this setting, a possible solution consists in selecting a smaller set of instruments in the first stage, and using only the instruments selected to find an estimate of the endogenous predictor $\hat{X}$ that will be passed into the second stage.

This chapter introduces different methods that can be used to select a set of instruments. Section 3.2 briefly reviews both the Lasso and the elastic net methods. Sections 3.3 and 3.4 describe some post-regularization methods, the adaptive Lasso, and introduce a possible modification of already existing techniques. Finally, Section 3.5 discusses the supervised principal component analysis as an additional method to select the relevant instruments and predict the endogenous variable in the first stage.

### 3.2 Brief Review of Regularization Methods: Lasso and Elastic Net

Given a data set with $n$ observations and $p$ predictors, the regression of the endogenous predictor $X$ on a set of instruments $Z$ can be written as:

$$X = Z\Pi + w,$$

$$w \sim N(0, \sigma).$$  \hspace{1cm} (3.1)

Without loss of generality, it is possible to assume that the covariates $Z_j$ ($j = 1, \ldots, p$) are standardized (i.e. all the covariates have mean 0 and variance one) and that the response variable $X$ has mean 0. Then, according to [16], for any fixed
non-negative penalty $\lambda$, the Lasso criterion is defined as:

$$L(\lambda, \Pi) = |X - Z\Pi|^2 + \lambda |\Pi|_1,$$

where $|\cdot|_1$ indicates the $l_1$-norm $|a|_1 = \sum_i |a_i|$, and $|\cdot|^2$ refers to the $l_2$-norm $|a|^2 = \sum_i a_i^2$. Given (3.2), the Lasso coefficients are estimated as:

$$\hat{\Pi}_{\text{lasso}} = \text{argmin}_{\Pi} \{ L(\lambda, \Pi) \} = \text{argmin}_{\Pi} \left\{ \sum_{i=1}^{n} (x_i - \sum_{j=1}^{p} z_{ij}\Pi_j)^2 + \lambda \sum_{j=1}^{p} |\Pi_j| \right\},$$

where the amount of shrinkage is controlled by the penalty parameter $\lambda$. Thus, the larger the value of $\lambda$, the greater the amount of shrinkage. When $\lambda = 0$, no shrinkage is applied and Lasso simply becomes an OLS estimator. The idea behind Lasso is to select a subset of relevant variables by imposing an $l_1$ penalty on the regression coefficients. Specifically, Lasso simultaneously selects a subset of variables and shrinks the chosen coefficients toward zero (compared with the OLS solution). Since, the minimization problem in (3.3) corresponds to a convex minimization procedure, any algorithm that finds the minimum it is guaranteed to detect the global, overall minimum. Consequently, the Lasso estimates can be computed using fast algorithms (e.g. LARS, cyclical coordinate descent) that work well even when $p$ is greater than $n$.

Even though the Lasso procedure is able to select a model easy to interpret and accurate in terms of prediction error, it has few drawbacks. First, when $p > n$, Lasso selects at most $n$ variables; second, when there is a group of highly correlated variables, it tends to randomly select only one variable for that group and to ignore
the rest. The above limitations are addressed by Zou and Hastie who introduce the elastic net procedure [18]. Specifically, for any fixed non-negative $\lambda_1$ and $\lambda_2$, the elastic net criterion is defined as:

$$L(\lambda_1, \lambda_2, \Pi) = |X - Z\Pi|^2 + \lambda_1|\Pi|_1 + \lambda_2|\Pi|_2^2,$$  \hspace{1cm} (3.4)

Given (3.4), the elastic net coefficients are estimated as:

$$\hat{\Pi}_{EN} = \arg\min_{\Pi} \{L(\lambda_1, \lambda_2, \Pi)\}$$ \hspace{1cm} (3.5)

$$= \arg\min_{\Pi} \left\{ \sum_{i=1}^{n} (x_i - \sum_{j=1}^{p} z_{ij}\pi_j)^2 + \lambda_1 \sum_{j=1}^{p} |\pi_j| + \lambda_2 \sum_{j=1}^{p} |\pi_j|^2 \right\}$$

which is still a convex minimization problem (i.e. as in the Lasso case, the minimum is guaranteed to be a global one). As in the Lasso case, the elastic net simultaneously does automatic variables selection and continuous shrinkage; however, unlike the Lasso, both an $l_1$ penalty and an $l_2$ penalty are considered. In particular, the $l_1$ part in (3.4) does automatic variables selection, whereas the $l_2$ part facilitates grouped selection. Consequently, the elastic net can select groups of correlated features when those groups are not known in advance.

In [18], Zou and Hastie describe the elastic net procedure as a regularized least squares problem; specifically, given $\alpha = \lambda_1/(\lambda_1 + \lambda_2)$, finding $\hat{\Pi}_{EN}$ in (3.5) is equivalent to finding the coefficients that minimize the following:

$$\hat{\Pi}_{EN} = \arg\min_{\Pi} \left\{ \sum_{i=1}^{n} (x_i - \sum_{j=1}^{p} z_{ij}\pi_j)^2 \right\}$$ \hspace{1cm} (3.6)

$$((1 - \alpha)|\Pi|_1 + \alpha|\Pi|^2) \leq t,$$
where $t$ is a constant that controls the amount of shrinkage and $\alpha$ is the elastic net parameter: when $\alpha = 0$ elastic net reduces to a Lasso problem, when $\alpha = 1$ elastic net is equivalent to ridge regression. As in [18], only the cases where $\alpha \in [0,1)$ are considered. Figure 3.1 shows, for the prostate cancer data discussed by Tibshirani in [16], the Lasso and the elastic net estimates as a function of the proportion of shrinkage. Specifically, since each colored line represents the path of a single coefficient shrinking towards zero, it is possible to see how each coefficient is selected in both the considered methods. The comparison between the two panels in Figure 3.1 clearly shows how elastic net is sensitive to having a group of correlated variables, and how it tends to stabilize the solution path.

![Figure 3.1](image)

**Figure 3.1**: Regularization methods coefficients path for a real data example. The ratio of the sum of the absolute current estimate over the sum of the absolute OLS estimates is on the x-axis: as we move towards zero, more coefficients are excluded from the model. The standardized coefficients computed by multiplying the original coefficients by the norm of the predictors are on the y-axis.
3.3 Post-Regularization Methods

Although regularization methods allow the researcher to select only the relevant variables and to reduce the estimates’ variance, the coefficients estimated in these models tend to be biased. The shrinkage in the regularization methods, in fact, causes the estimates of non zero-coefficients to be biased towards zero (i.e., the positive coefficients will be underestimated, while the negative coefficients will be overestimated), and non consistent [8].

To solve these issues different modifications of regularization methods have been suggested. One of these, introduced in 2004, is a hybrid LARS/OLS procedure where the bias of the model’s coefficients is reduced by fitting a least squares to the model selected by LARS [6]. Based on the LARS/OLS procedure, Belloni et al. proposed the post-Lasso estimator as a feasible method to select only the relevant variables and to reduce the non-zero coefficients’ bias [2].

In the next sections we discuss in detail the post-Lasso procedure, and we propose a modification of post-Lasso that might be helpful in all the situations where the number of relevant predictors is greater than the available sample size.

3.3.1 The Post-Lasso Estimator

Similar to the LARS/OLS procedure, Belloni et al. introduce the post-Lasso which fits an OLS to the model selected using the $l_1$ penalty [2]. Formally, if $\hat{T} = \{ j \in \{1, 2, ..., p\} : |\hat{\pi}_j| > 0 \}$ indicates the model chosen by the Lasso procedure, then the post-Lasso coefficients $\overline{\Pi}_{post}$ can be defined as:

$$\overline{\Pi}_{post} = \arg\min_{\Pi} \left\{ \sum_{i=1}^{n} (x_i - \sum_j z_{ij}\hat{\pi}_j)^2 \right\}, j \in \hat{T}. \quad (3.7)$$
Belloni et al. prove that, when $p > n$, the post-Lasso estimator can perform at least as well as Lasso in terms of the rate of convergence, and it has a smaller coefficients’ bias. In addition, they show that Lasso and post-Lasso performs similarly when the absolute value of the non-zero coefficients is small (less than 0.5) and that post-Lasso has a better performance when the coefficients of the relevant predictors are greater (in absolute value) than 0.5.

3.3.2 The Post-EN/L2 Estimator

As already discussed in Section 3.1, there might be scenarios where given the set of all variables $p$, the number of relevant predictors $s$ is greater than the sample size $n$. In these cases, any selection method based on an $l_1$ penalty will be able to select at most $n$ variables and it will miss some relevant features. A possible solution that can be adopted when $s > n$ is to select the relevant predictors using an elastic net penalty, and to reduce the endogenous predictor’s coefficient bias using only the set of variables selected.

However, unlike post-Lasso, it might be impossible to apply an OLS to the chosen model if more than $n$ variables are selected. To be able to simultaneously attenuate the coefficients’ bias and select more than $n$ variables, we propose a new estimator called post-EN/L2. The idea of the post-EN/L2 is to fit ridge regression using only the variables selected by the elastic net. Formally, if $\hat{T}_{EN} = \{ j \in \{1, 2, ..., p\} : |\hat{\beta}_j| > 0 \}$ indicates the model chosen by the elastic net, then the post-EN/L2 coefficients $\hat{\Pi}_{\text{postEN/L2}}$ can be defined as:

$$\hat{\Pi}_{\text{postEN/L2}} = \arg\min\Pi \left\{ \frac{1}{n} \sum_{i=1}^{n} (x_i - \sum_j z_{ij} \pi_j)^2 + \lambda_2 \sum_j \pi_j^2 \right\}, j \in \hat{T}_{EN}. \quad (3.8)$$
Fitting ridge regression using the selected variables only allows us to consider more than \( n \) predictors, to take into account the presence of groups of correlated variables, and to separate the two effects of any regularization method (i.e., variable selection and coefficient shrinkage). In addition, by using a two-stage regularization procedure we now consider two penalty parameters: a larger one, on the first pass, used to select a small number of variables, and a small penalty, on the second pass, used to shrink the non-zero coefficients.

### 3.4 Adaptive Lasso

As already introduced in Section 3.3, Zou in [17] proves that the Lasso estimates are only consistent under specific circumstances, and he proposes a new estimator called the adaptive Lasso. The proposed method introduces weights to penalize differently the coefficients in the \( l_1 \) penalty and to guarantee the consistency of the Lasso estimates. Formally, given the model in (3.1), \( \Pi \) can be estimated as:

\[
\hat{\Pi}_{\text{ada.lasso}} = \arg\min_{\Pi} \left\{ \sum_{i=1}^{n} (y_i - \sum_{j=1}^{p} z_{ij} \pi_j)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\}
\]  

(3.9)

where \( w = (w_1, \ldots, w_p) \) is a vector of the data dependent weights. In particular, given the coefficient estimated by OLS, \( \hat{\Pi}_{\text{OLS}} \), the vector \( w \) can be estimated by:

\[
\hat{w} = 1/|\hat{\Pi}_{\text{OLS}}|^\gamma
\]

(3.10)

where \( \gamma > 0 \). Since the optimization problem in (3.9) is convex, once the minimum is found, it is guaranteed to be the global minimum.
Based on the weights in (3.10), Zou proves that, when \( n > p \), the adaptive Lasso provides consistent estimates under every circumstance, identifies the right subset of relevant variables, and achieves the optimal estimation rate. As in the Lasso case, both \( \gamma \) and the penalty parameter \( \lambda \) are chosen based on cross-validation; specifically, for a given \( \gamma \), Zou uses cross-validation paired with LARS to exclusively search for \( \lambda \). Moreover, while it is also possible to change \( \hat{\Pi}_{OLS} \) with any other consistent estimates, Zou in [17] suggests to use \( \hat{\Pi}_{OLS} \) unless the researcher is concerned of the effects of multicollinearity. In that case, the use of \( \hat{\Pi}_{ridge} \) (i.e., the estimate resulting from fitting ridge regression) is advised.

### 3.5 Supervised Principal Components Analysis

An alternative method that can be used to select the relevant covariates and to predict the response variable \( X \) from a set of \( p \) predictors \( (Z_1, \ldots, Z_p) \) measured on each of the \( n \) sample, is the supervised principal components (SPCA). SPCA is a four steps methodology, introduced by Bair et al. in [1], that allows the researcher to predict the outcome of interest and to identify relevant predictors.

Formally, given \( Z = (Z_1, \ldots, Z_p) \) an \( n \times p \) matrix of variables and \( X \) an \( n \times 1 \) vector of outcome measurements, and assuming that (without loss of generality) all the variables in the model are centered, SPCA first estimates the standard regression coefficient for each predictor as:

\[
s_i = \frac{Z_i^T X}{\sqrt{Z_i^T Z_i}}.
\]  

(3.11)

Afterwards, by using cross-validation of the likelihood ratio test it estimates a threshold \( \theta \), and it forms a reduced data matrix \( Z_\theta \) considering only the collection
of indices, $C_\theta$, such that $|s_i| > \theta$. Next, SPCA computes the first principal component (or the first few) of the reduced data matrix, $u_{\theta,1}$. Finally, it fits a univariate regression model with response $X$ and predictor $u_{\theta,1}$:

$$X^{\text{spec.}, \theta} = \gamma_0 + \gamma_1 u_{\theta,1} \quad (3.12)$$

Since $u_{\theta,1}$ is a left singular vector of $Z_\theta$, it has mean zero and norm one; hence, $\hat{\gamma}_1 = u_{\theta,1}^T X$, $\hat{\gamma}_0 = \bar{x}$ and:

$$\hat{X}^{\text{spec.}, \theta} = X + \hat{\gamma}_1 u_{\theta,1} \quad (3.13)$$

Consequently, the model in (3.12) is a restricted linear model built using all the selected predictors that can be used to make reliable predictions of $X$.

By performing principal components analysis on only a smaller set of variables, SPCA tries to consider only those predictors with the strongest estimated correlation with the outcome of interest; thus, it tends to eliminate the “excessive noise”. Moreover, the use of PCA allows the researcher to find a more compact representation of the data by reducing a set of $p$ variables (more or less correlated) into $k$ (where $k < p$) uncorrelated linear combinations of the original variables called principal components that contain most of the relevant information of the data [14].

Having derived the predictor $u_{\theta,1}$, Bair et al. also introduce the concept of importance score to assess the contribution of each individual covariate. Specifically, they define the importance score as the correlation between each variable and the
supervised principal component predictor:

\[ \text{imp}_i = \text{Cor}(Z_i, u_{\beta,1}) \]  

(3.14)

Based on (3.14), one can understand which are the most relevant predictors (i.e., the features with higher importance score), discard the irrelevant variables, and prove that the SPCA coefficients’ estimates are consistent. Although SPCA allows the researcher to select a smaller set of features, it is essential to highlight that the ability of cross-validation to select the correct threshold has not been established theoretically in [1].

3.6 Using Variables Selection in Instrumental Variables Regression

Every selection method introduced in this chapter allows the researcher to predict the outcome variable \( X \) based on a selected smaller subset of relevant predictors and to select a smaller subset of relevant predictors. In this thesis, we use these variables selection techniques in order to primarily choose a smaller set of valid instruments from a collection of variables that might also include weak and possibly irrelevant instruments. Given the framework in (2.7) and (2.8), we employ a two steps procedure. Specifically, first, we use a variable selection method (regularization, post-regularization and SPCA) to select a set of relevant variable and to calculate \( \hat{X} \) as free as possible from noise; then, we regress \( y \) on \( \hat{X} \) in order to estimate the coefficient \( \beta \) of the endogenous predictor. Ideally, we would like to
chose a method such that all the relevant instruments are detected correctly:

\[ P[(j : \hat{\pi}_j \neq 0) = (j : \pi_j \neq 0)] \to 1, \quad (3.15) \]

and the estimated coefficient bias of the endogenous predictor tends to zero.

In the next chapter we compare the performance of SPCA to the one of regularization and post-regularization methods using different simulation settings that take into account both the cases when \( n > p \) and when \( n < p \).
Chapter 4

Simulation Study

4.1 Introduction

We carried out multiple simulation studies in order to understand which selection method minimizes the estimate’s bias and selects the right set of instruments. In this chapter, we explain in detail the settings used, we report the results obtained, and we compare all the methods employed.

In our simulation study, we consider one endogenous variable, \( n \) observations, and \( p \) instruments (where \( p \) can be either less or greater than the sample size). Specifically, the simulation is based on the following model:

\[
y = X\beta + \varepsilon
\]

\[
X = Z\Pi + w.
\]

\[
\begin{pmatrix}
\varepsilon_i \\
w_i
\end{pmatrix}
\sim N
\begin{pmatrix}
0, \\
\begin{pmatrix}
1 & \sigma_{e_2} \\
\sigma_{e_1} & \sigma_w^2
\end{pmatrix}
\end{pmatrix}
\]
$\beta = 1$ is the parameter of interest, $Z$ is the $n \times p$ matrix of instruments generated differently in different examples, but such that in all considered cases $\mathbb{E}[Z^Tw] = \mathbb{E}[Z^T\varepsilon] = 0$. The endogenous predictor $X$ and the outcome variable $y$ are generated hierarchically according to (4.1). For the other parameters, a wide variety of different settings are used. Specifically, we consider two main scenarios (the case where $n > p$, and the one where $n < p$), two different levels of endogeneity (low and medium), and three different values of $\sigma^2_w$ which are chosen to benchmark three different strengths of instruments. The two endogeneity levels are fixed by setting $\text{Cor}(\varepsilon,w)$ to either 0.3 (low) or 0.6 (medium); whereas the instrument strength is introduced in the model by computing $\sigma^2_w$ as:

$$
\sigma^2_w = \frac{n\Pi' \Sigma_e \Pi}{F \Pi' \Pi},
$$

(4.2)

where $\Sigma_e$ is the variance and covariance matrix of $Z$, and $F$ is the first stage F-statistic set to be 10 (weak instruments), 40 (medium strength instruments) and 160 (strong instruments).

For each combination of the simulation parameter values, we simulate $N = 500$ datasets. For each dataset, we estimate the coefficient $\beta$ of the endogenous variable using all the methods listed in Chapter 3, and we compare their root mean square error (RMSE):

$$
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left[ \hat{\beta}_i - \beta_i \right]^2},
$$

(4.3)

as well as the average number of variables selected in the first stage, and the average number of relevant instruments correctly identified by the chosen method (i.e. true
positive). All the reported results have been rounded to have two significant digits.

4.2  The Case of $n > p$

In the scenario where $n > p$, we consider four examples: the first three examples are taken from the original elastic net paper [18] and they are partially changed in order to add endogeneity, whereas the last one is taken from the Belloni et. al’s simulation study in [3]. Each example is characterized by a different $n$, $p$, $\Pi$ and $Z$. The first two examples represent cases in which Lasso and elastic net are supposed to perform similarly, the third example creates a grouped variables situation, and the last example depicts a scenario where most of the potential instruments are irrelevant. The details of the four examples are reported below:

**Example 1.** The data are simulated from the true model (4.1). The sample size $n = 20$, while the number of instruments $p = 8$. Among the 8 instruments only $s = 3$ are relevant:

$$\Pi = (3, 1.5, 0, 0, 2, 0, 0, 0).$$

The matrix of instruments $Z$ is generated as $N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$.

**Example 2.** It is the same as example 1 except that all the instruments considered are relevant but each instrument does not have a strong influence on the endogenous predictor $X$: $\Pi_i = 0.85$ ($i = 1, 2, \ldots, 8$).

**Example 3.** The data are simulated from the model (4.1) where the number of observations $n = 50$ and the number of instruments $p = 40$. Among the 40
instruments only $s = 15$ are relevant:

$$
\Pi = (3, \ldots, 3, 0, \ldots, 0).$$

Example 3 considers three equally important groups of five instruments each, and other 25 irrelevant variables. The matrix of potential instruments is generated as:

- $z_i = W_1 + v_i$ with $W_1 \sim N(0, 1), i = 1, \ldots, 5$
- $z_i = W_2 + v_i$ with $W_2 \sim N(0, 1), i = 6, \ldots, 10$
- $z_i = W_3 + v_i$ with $W_3 \sim N(0, 1), i = 11, \ldots, 15$
- $z_i \sim N(0, 1), i = 16, \ldots, 40$

where $v_i (i = 1, \ldots, 15)$ are independent identically distributed $N(0, 0.01)$.

Example 4. The data are simulated from the model (4.1) where the number of observations $n = 500$ and the number of potential instruments $p = 100$. Among the 100 instruments, only $s = 5$ are relevant:

$$
\Pi = (1, \ldots, 1, 0, \ldots, 0).$$

The matrix of instruments $Z$ is generated as $N(0, \Sigma)$ with $\Sigma_{ij} = 0.5|i-j|$.

4.2.1 Simulation’s Results

The comparison between selection methods is done based on the endogenous predictor’s coefficient’s RMSE and the average number of instruments selected by each method. Tables 4.1, 4.2, 4.3 and 4.4 report the obtained results.

Tables 4.1 and 4.2 show that, regardless of the endogeneity level and the sim-
ulation design, post-Lasso minimizes the coefficient’s RMSE. On the other hand, Lasso and elastic net lead to the highest coefficient’s RMSE; surprisingly, both the methods have a worse performance than the 2SLS. As previously observed by Belloni et. al results show that the selection of a subset of instruments comes at a price of an excessive shrinkage of the coefficients in the first stage [3]. Fitting a ridge regression (post-EN/L2) or a least squares (post-Lasso) only on the selected variables helps to reduce the coefficient’s RMSE. While every regularization method is outperformed by post-Lasso, it is essential to highlight that SPCA leads to a coefficient’s RMSE close to the one estimated using post-Lasso.

Tables 4.3 and 4.4 show the number of variables selected by each method. As some of the methods used are based on the same selection, the comparison reported only inspects the differences between the Lasso, the elastic net, the adaptive Lasso and the SPCA. Before drawing any conclusion, it is essential to remark that, in every example, when the number of selected variables is higher than the number of relevant instruments, $s$, the model selects all the right coefficients (hence, the number of true positives is equal to $s$). When the number of selected variables is lower than the number of relevant instruments, the model does not select any of the irrelevant coefficients (thus, the number of false positives is equal to zero). The Lasso, the elastic net and the adaptive Lasso tend to choose the same number of variables, whereas the SPCA selects less variables and tends to miss some of the true instruments.
Table 4.1: Coefficient RMSE for the simulated examples and for different instruments strength. \(Cor(\varepsilon, v) = 0.3\). For the regularization and the post-regularization methods, the penalty parameter \(\lambda\) has been chosen such that the cross-validation error is within one standard error from the minimum.
Table 4.2: Coefficient RMSE for the simulated examples and for different instruments strength. $\text{Cor}(\epsilon, v) = 0.6$. For the regularization and the post-regularization methods, the penalty parameter $\lambda$ has been chosen such that the cross-validation error is within one standard error from the minimum.
### Weak Instruments

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>20</td>
<td>8</td>
<td>3</td>
<td>3.75</td>
<td>3.83</td>
<td>3.96</td>
<td>3.11</td>
</tr>
<tr>
<td>Example 2</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>4.20</td>
<td>4.30</td>
<td>5.11</td>
<td>5.17</td>
</tr>
<tr>
<td>Example 3</td>
<td>50</td>
<td>40</td>
<td>15</td>
<td>21.60</td>
<td>20.92</td>
<td>19.06</td>
<td>10.43</td>
</tr>
<tr>
<td>Example 4</td>
<td>500</td>
<td>100</td>
<td>5</td>
<td>1.66</td>
<td>1.73</td>
<td>2.60</td>
<td>16.71</td>
</tr>
</tbody>
</table>

### Medium Strength Instruments

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>20</td>
<td>8</td>
<td>3</td>
<td>3.89</td>
<td>4</td>
<td>3.62</td>
<td>3.14</td>
</tr>
<tr>
<td>Example 2</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>6.89</td>
<td>7</td>
<td>7.15</td>
<td>5.96</td>
</tr>
<tr>
<td>Example 3</td>
<td>50</td>
<td>40</td>
<td>15</td>
<td>21.41</td>
<td>21.15</td>
<td>16.88</td>
<td>10.00</td>
</tr>
<tr>
<td>Example 4</td>
<td>500</td>
<td>100</td>
<td>5</td>
<td>4.80</td>
<td>4.81</td>
<td>5.1</td>
<td>8.97</td>
</tr>
</tbody>
</table>

### Strong Instruments

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>20</td>
<td>8</td>
<td>3</td>
<td>3.84</td>
<td>4.01</td>
<td>3.19</td>
<td>3.15</td>
</tr>
<tr>
<td>Example 2</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>7.86</td>
<td>7.93</td>
<td>7.94</td>
<td>6.16</td>
</tr>
<tr>
<td>Example 3</td>
<td>50</td>
<td>40</td>
<td>15</td>
<td>20.92</td>
<td>21.13</td>
<td>15.18</td>
<td>10.42</td>
</tr>
<tr>
<td>Example 4</td>
<td>500</td>
<td>100</td>
<td>5</td>
<td>5.04</td>
<td>5.07</td>
<td>5.52</td>
<td>5.29</td>
</tr>
</tbody>
</table>

*Table 4.3:* Average number of variables selected by each method. $\text{Cor}(\varepsilon,v) = 0.3$. For the regularization and the post-regularization methods, the penalty parameter $\lambda$ has been chosen such that the cross-validation error is within one standard error from the minimum.
<table>
<thead>
<tr>
<th>Weak Instruments</th>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>20</td>
<td>8</td>
<td>3</td>
<td>3.73</td>
<td>3.80</td>
<td>3.91</td>
<td>3.10</td>
</tr>
<tr>
<td>Example 2</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>4.16</td>
<td>4.29</td>
<td>5.13</td>
<td>5.21</td>
</tr>
<tr>
<td>Example 3</td>
<td>50</td>
<td>40</td>
<td>15</td>
<td>21.50</td>
<td>20.92</td>
<td>19.02</td>
<td>10.36</td>
</tr>
<tr>
<td>Example 4</td>
<td>500</td>
<td>100</td>
<td>5</td>
<td>1.68</td>
<td>3.08</td>
<td>2.6</td>
<td>16.41</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Medium Strength Instruments</th>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>20</td>
<td>8</td>
<td>3</td>
<td>3.88</td>
<td>3.96</td>
<td>3.62</td>
<td>3.12</td>
</tr>
<tr>
<td>Example 2</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>6.91</td>
<td>6.98</td>
<td>7.16</td>
<td>5.96</td>
</tr>
<tr>
<td>Example 3</td>
<td>50</td>
<td>40</td>
<td>15</td>
<td>21.34</td>
<td>21.10</td>
<td>16.72</td>
<td>10.48</td>
</tr>
<tr>
<td>Example 4</td>
<td>500</td>
<td>100</td>
<td>5</td>
<td>4.80</td>
<td>4.81</td>
<td>4.98</td>
<td>8.98</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strong Instruments</th>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>20</td>
<td>8</td>
<td>3</td>
<td>3.84</td>
<td>4.01</td>
<td>3.17</td>
<td>3.15</td>
</tr>
<tr>
<td>Example 2</td>
<td>20</td>
<td>8</td>
<td>8</td>
<td>7.85</td>
<td>7.94</td>
<td>7.95</td>
<td>6.18</td>
</tr>
<tr>
<td>Example 3</td>
<td>50</td>
<td>40</td>
<td>15</td>
<td>20.86</td>
<td>21.13</td>
<td>15.18</td>
<td>10.56</td>
</tr>
<tr>
<td>Example 4</td>
<td>500</td>
<td>100</td>
<td>5</td>
<td>5.04</td>
<td>5.07</td>
<td>5.70</td>
<td>5.28</td>
</tr>
</tbody>
</table>

Table 4.4: Average number of variables selected by each method. $\text{Cor}(\varepsilon, v) = 0.6$. For the regularization and the post-regularization methods, the penalty parameter $\lambda$ has been chosen such that the cross-validation error is within one standard error from the minimum.
4.3 The Case of \( n < p \)

In biomarker discovery studies where the genes can be considered as potential instruments, the number of potential instrumental variable \( p \) is usually higher than the sample size \( n \). In these settings, we would like to identify all the relevant instruments in order to have a complete picture of the problem investigated; consequently, it becomes essential to employ methodologies that are able select more than \( n \) variables.

In this part of the simulation study, we look at the performance of the instrumental variables selection methods when \( p > n \). Specifically, given a set of \( p \) potential instruments, four examples are considered: in the first one, the number of relevant instruments \( s \) is greater than the total sample size \( n \); in the second example, the number of relevant instruments is less than the sample size \( n \); in the third one all the considered instruments are relevant but they do not have a strong influence on the endogenous predictor; finally, the last example depicts a sparse scenario where most of the potential instruments are actually irrelevant. The details of the four examples are reported below:

**Example 5.** The data are simulated from the true model (4.1) where the number of observations \( n = 40 \), the number of predictors \( p = 80 \) and the number of relevant variables \( s = 50 \):

\[
\Pi = (2, \ldots, 2, 0, \ldots, 0, 2, \ldots, 2, 0, \ldots, 0)_{25 \times 20}^{25 \times 10}.
\]

The matrix of instruments is generated as: \( Z \sim N(0, \Sigma) \) with \( \Sigma_{ij} = 0.5^{|i-j|} \).

**Example 6.** It is the same as example 5 except that the number of relevant
variable $s = 20$ is smaller than the sample size $n = 40$.

\[ \Pi = (2, \ldots, 2, 0, \ldots, 0, 2, \ldots, 2, 0, \ldots, 0). \]

**Example 7.** It is the same as example 5 except that all the instruments included are relevant but each instrument does not have a strong influence on the endogenous predictor $X$: $\Pi_j = 0.85$ ($j = 1, 2, \ldots, 80$).

**Example 8.** The data are simulated from the model (4.1) where the number of observations $n = 50$ and the number of instruments $p = 1000$. Among the 1000 instruments, $s = 100$ are relevant:

\[ \Pi = (1, \ldots, 1, 0, \ldots, 0). \]

The matrix of instruments $Z$ is generated as $N(0, \Sigma)$ with $\Sigma_{ij} = 0.5^{|i-j|}$.

### 4.3.1 Simulation’s Results

Tables 4.5 and 4.6 report the coefficient’s RMSE for the examples considered. Both tables show results similar to the ones obtained when $n > p$. For instance, we can see that the post-Lasso outperforms all the other regularization methods, and that post-EN/L2 reduces the coefficient’s RMSE of the endogenous predictor’s coefficient by diminishing the first stage coefficient’s bias. Unlike the case where $n > p$, the SPCA has the same performance of the post-Lasso in all the considered settings. The behavior of the SPCA agrees with what was observed in the previous section, and it can be considered as an indicator that SPCA might have a better performance when both $n$ and $p$ increase.
Tables 4.7 and 4.8 show the number of variables selected by each method. Like in the case where \( n > p \), the comparison reported examines only the methods that are based on different selection mechanisms. Both tables clearly show the Lasso and the adaptive Lasso’s inability to select more instruments than the available sample size. On the other hand, the elastic net tends to select more instruments with a higher number of true positive detected.

As a genomic dataset resembles a sparse situation, it is interesting to observe the results obtained in Example 8. In this scenario, the regularization methods fail to select a complete subset of relevant instruments: the maximum number of variables selected is 70.5 out of 100 non-zero coefficients among which only 58 are correctly detected. The performance of SPCA is not as good either; in fact, while SPCA selects all the relevant instruments, it adds, on average, other 94 false positives. Regardless of the strength of the instruments included in the model, the level of endogeneity considered, and the number of relevant instruments selected, both post-Lasso and SPCA lead to the same value for the RMSE of the endogenous predictor’s coefficient. It is interesting to observe that the first principal component retains most of the relevant information from the 194 variables; as a result, the estimated endogenous predictor \( \hat{X} \) is close to that obtained from fitting an OLS regression with only the five variables selected by Lasso.
<table>
<thead>
<tr>
<th></th>
<th>post-Lasso</th>
<th>Lasso</th>
<th>EN</th>
<th>Post EN/L2</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5</td>
<td>0.01</td>
<td>0.27</td>
<td>0.10</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 6</td>
<td>0.01</td>
<td>0.16</td>
<td>0.34</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 7</td>
<td>0.01</td>
<td>0.48</td>
<td>0.10</td>
<td>0.06</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 8</td>
<td>0.01</td>
<td>0.75</td>
<td>1.08</td>
<td>0.10</td>
<td>0.75</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>post-Lasso</th>
<th>Lasso</th>
<th>EN</th>
<th>Post EN/L2</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5</td>
<td>0.01</td>
<td>0.22</td>
<td>0.10</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 6</td>
<td>0.01</td>
<td>0.08</td>
<td>0.27</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 7</td>
<td>0.01</td>
<td>0.41</td>
<td>0.05</td>
<td>0.10</td>
<td>0.10</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 8</td>
<td>0.01</td>
<td>1.11</td>
<td>1.51</td>
<td>0.11</td>
<td>0.64</td>
<td>0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>post-Lasso</th>
<th>Lasso</th>
<th>EN</th>
<th>Post EN/L2</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5</td>
<td>0.01</td>
<td>0.20</td>
<td>0.09</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 6</td>
<td>0.01</td>
<td>0.05</td>
<td>0.27</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 7</td>
<td>0.01</td>
<td>0.41</td>
<td>0.04</td>
<td>0.11</td>
<td>0.09</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 8</td>
<td>0.01</td>
<td>0.75</td>
<td>1.03</td>
<td>0.13</td>
<td>0.61</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 4.5: Coefficient RMSE for the simulated examples and for different instruments strength. Cor(\(\varepsilon, v\)) = 0.3. For the regularization and the post-regularization methods, the penalty parameter \(\lambda\) has been chosen such that the cross-validation error is within one standard error from the minimum.
### Weak Instruments

<table>
<thead>
<tr>
<th></th>
<th>post-Lasso</th>
<th>Lasso</th>
<th>EN</th>
<th>Post EN/L2</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5</td>
<td>0.01</td>
<td>0.26</td>
<td>0.11</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 6</td>
<td>0.01</td>
<td>0.17</td>
<td>0.34</td>
<td>0.04</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 7</td>
<td>0.01</td>
<td>0.49</td>
<td>0.11</td>
<td>0.06</td>
<td>0.11</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 8</td>
<td>0.01</td>
<td>0.92</td>
<td>1.05</td>
<td>0.09</td>
<td>0.69</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### Medium Strength Instruments

<table>
<thead>
<tr>
<th></th>
<th>post-Lasso</th>
<th>Lasso</th>
<th>EN</th>
<th>Post EN/L2</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5</td>
<td>0.01</td>
<td>0.24</td>
<td>0.10</td>
<td>0.04</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 6</td>
<td>0.01</td>
<td>0.08</td>
<td>0.27</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 7</td>
<td>0.01</td>
<td>0.41</td>
<td>0.06</td>
<td>0.09</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 8</td>
<td>0.01</td>
<td>1.23</td>
<td>1.15</td>
<td>0.11</td>
<td>0.64</td>
<td>0.01</td>
</tr>
</tbody>
</table>

### Strong Instruments

<table>
<thead>
<tr>
<th></th>
<th>post-Lasso</th>
<th>Lasso</th>
<th>EN</th>
<th>Post EN/L2</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5</td>
<td>0.01</td>
<td>0.21</td>
<td>0.09</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 6</td>
<td>0.01</td>
<td>0.05</td>
<td>0.27</td>
<td>0.04</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 7</td>
<td>0.01</td>
<td>0.39</td>
<td>0.04</td>
<td>0.11</td>
<td>0.08</td>
<td>0.01</td>
</tr>
<tr>
<td>Example 8</td>
<td>0.01</td>
<td>0.76</td>
<td>1.18</td>
<td>0.14</td>
<td>0.58</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Table 4.6:** Coefficient RMSE for the simulated examples and for different instruments strength. $\text{Cor}(\varepsilon, v) = 0.6$. For the regularization and the post-regularization methods, the penalty parameter $\lambda$ has been chosen such that the cross-validation error is within one standard error from the minimum.
### Weak Instruments

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5</td>
<td>40</td>
<td>80</td>
<td>50</td>
<td>53.59</td>
<td>26.69</td>
<td>25.66</td>
<td>27.28</td>
</tr>
<tr>
<td>Example 6</td>
<td>40</td>
<td>80</td>
<td>20</td>
<td>24.96</td>
<td>23.20</td>
<td>20.44</td>
<td>19.92</td>
</tr>
<tr>
<td>Example 7</td>
<td>40</td>
<td>80</td>
<td>80</td>
<td>63.90</td>
<td>22.66</td>
<td>22.68</td>
<td>28.41</td>
</tr>
<tr>
<td>Example 8</td>
<td>50</td>
<td>1000</td>
<td>100</td>
<td>70.50</td>
<td>4.80</td>
<td>8.24</td>
<td>191.77</td>
</tr>
</tbody>
</table>

### Medium Strength Instruments

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5</td>
<td>40</td>
<td>80</td>
<td>50</td>
<td>55.79</td>
<td>27.62</td>
<td>25.06</td>
<td>27.42</td>
</tr>
<tr>
<td>Example 6</td>
<td>40</td>
<td>80</td>
<td>20</td>
<td>26.56</td>
<td>26.74</td>
<td>21.48</td>
<td>20.23</td>
</tr>
<tr>
<td>Example 7</td>
<td>40</td>
<td>80</td>
<td>80</td>
<td>74.96</td>
<td>25.11</td>
<td>23.22</td>
<td>29.23</td>
</tr>
<tr>
<td>Example 8</td>
<td>50</td>
<td>1000</td>
<td>100</td>
<td>56.50</td>
<td>4.78</td>
<td>9.64</td>
<td>196.99</td>
</tr>
</tbody>
</table>

### Strong Instruments

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 5</td>
<td>40</td>
<td>80</td>
<td>50</td>
<td>56.12</td>
<td>28.05</td>
<td>25.16</td>
<td>27.46</td>
</tr>
<tr>
<td>Example 6</td>
<td>40</td>
<td>80</td>
<td>20</td>
<td>25.92</td>
<td>28.14</td>
<td>20.28</td>
<td>10.02</td>
</tr>
<tr>
<td>Example 7</td>
<td>40</td>
<td>80</td>
<td>80</td>
<td>76.56</td>
<td>25.60</td>
<td>21.82</td>
<td>29.73</td>
</tr>
<tr>
<td>Example 8</td>
<td>50</td>
<td>1000</td>
<td>100</td>
<td>70.50</td>
<td>4.80</td>
<td>9.82</td>
<td>195.52</td>
</tr>
</tbody>
</table>

**Table 4.7:** Average number of variables selected by each method. $\text{Cor}(\epsilon, v) = 0.3$. For the regularization and the post-regularization methods, the penalty parameter $\lambda$ has been chosen such that the cross-validation error is within one standard error from the minimum.
Weak Instruments

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>80</td>
<td>50</td>
<td>53.68</td>
<td>26.82</td>
<td>25.90</td>
<td>27.45</td>
</tr>
<tr>
<td>40</td>
<td>80</td>
<td>20</td>
<td>24.95</td>
<td>23.25</td>
<td>20.50</td>
<td>19.92</td>
</tr>
<tr>
<td>40</td>
<td>80</td>
<td>80</td>
<td>63.84</td>
<td>22.59</td>
<td>22.84</td>
<td>28.38</td>
</tr>
<tr>
<td>50</td>
<td>1000</td>
<td>100</td>
<td>70.50</td>
<td>4.94</td>
<td>8.04</td>
<td>192.21</td>
</tr>
</tbody>
</table>

Medium Strength Instruments

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>80</td>
<td>50</td>
<td>55.81</td>
<td>27.71</td>
<td>24.32</td>
<td>27.45</td>
</tr>
<tr>
<td>40</td>
<td>80</td>
<td>20</td>
<td>26.64</td>
<td>26.74</td>
<td>21.52</td>
<td>19.97</td>
</tr>
<tr>
<td>40</td>
<td>80</td>
<td>80</td>
<td>74.97</td>
<td>25.20</td>
<td>22.90</td>
<td>29.58</td>
</tr>
<tr>
<td>50</td>
<td>1000</td>
<td>100</td>
<td>56.60</td>
<td>4.69</td>
<td>8.56</td>
<td>198.09</td>
</tr>
</tbody>
</table>

Strong Instruments

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>s</th>
<th>EN</th>
<th>Lasso</th>
<th>Adaptive Lasso</th>
<th>SPCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>40</td>
<td>80</td>
<td>50</td>
<td>56.27</td>
<td>28.08</td>
<td>25.02</td>
<td>28.46</td>
</tr>
<tr>
<td>40</td>
<td>80</td>
<td>20</td>
<td>25.99</td>
<td>28.24</td>
<td>20.32</td>
<td>20.07</td>
</tr>
<tr>
<td>40</td>
<td>80</td>
<td>80</td>
<td>76.56</td>
<td>25.60</td>
<td>21.92</td>
<td>29.77</td>
</tr>
<tr>
<td>50</td>
<td>1000</td>
<td>100</td>
<td>70.50</td>
<td>4.94</td>
<td>9.94</td>
<td>191.98</td>
</tr>
</tbody>
</table>

Table 4.8: Average number of variables selected by each method. \( Cor(\varepsilon, v) = 0.6 \). For the regularization and the post-regularization methods, the penalty parameter \( \lambda \) has been chosen such that the cross-validation error is within one standard error from the minimum.
4.4 The Choice of the Penalty Parameter $\lambda$

As already introduced in Section 3.3, the penalty parameter $\lambda$ used in regularization and post-regularization methods can be chosen based on different criteria. In this section, we consider a possible grid of $\lambda$s, and we look at the performance of both regularization and post-regularization methods in order to illustrate the importance of choosing the appropriate penalty.

The most commonly used penalties are the ones introduced by Hastie et al. in [8]. To compute the optimal $\lambda$, they fix a grid of $\lambda$, then for each $\lambda$ in the grid, they carry out a $k$ folds cross validation; specifically, they divide the available data in $k$ subsamples, train the model in $k - 1$ subsamples and test it in the $k^{th}$ subsample. They repeat the cross-validation procedure $k$ times and the $k$ results from the different subsamples are combined to produce a single estimation for the model’s prediction error. Using the aforementioned procedure, they distinguish between two different penalties:

- $\lambda_{\text{min}}$: the value of $\lambda$ that gives the minimum mean cross-validated error;
- $\lambda_{1se}$: the value of $\lambda$ that gives the most regularized model such that the cross-validation error is within one standard deviation of its minimum.

Since the error at each value of $\lambda$ is the average error over the $k$ folds, this estimate is characterized by uncertainty. Consequently, they advise to take a more conservative approach and to use $\lambda_{1se}$ as penalty parameter that gives a slightly simpler model than the best model, and takes into account the uncertainty in the $k$ folds cross-validation estimate.

Another possible choice of $\lambda$ is introduced by Belloni et al. in [2]. In particular,
the penalty parameter adopted by Belloni et. al is not computed based on the cross-validation error of the first stage, but it estimated looking only the data available without considering the out-of-sample prediction error. Moreover, they justify the choice of a data-driven penalty by proving that, if the penalty is not chosen using the data, then it tends to be quite conservative (i.e. too large) when the predictors in the model are highly correlated [2]. As there exists multiple choices for the penalty parameter, we set a grid of possible values to look at the trend of the endogenous predictor coefficient’s RMSE and of the average number of variables selected over the grid.

Given the model in (4.1), since the variables selection is carried out in the first stage, the choice of $\lambda$ refers to the regression of $X$ on $Z$; however, as the main focus of the study is the estimate of $\beta$, we could check whether the $\lambda$ chosen in the first stage is the one that minimizes the coefficient’s RMSE in the second stage. Figures 4.1 and 4.2 show the trend of the coefficient’s RMSE for Examples 4 and 8, given a grid of possible $\lambda$s. We use a vertical black line to indicate the value of $\log(\lambda_{1se})$ that gives the most regularized model such that the cross-validation error in the first stage is within one standard deviation of its minimum. When $n > p$, we only plot the trend of the coefficient’s RMSE for the Lasso as it is quite similar to the one generated fitting the elastic net. In both examples, it can be observed how the classical regularization methods lead to a higher coefficient’s RMSE as the shrinkage increases. Moreover, it is essential to highlight that the coefficient’s RMSE in the second stage it is not minimized at the value of $\log(\lambda_{1se})$ computed based on the cross-validation error. However, the difference between the absolute minimum and the coefficient’s RMSE estimated using $\log(\lambda_{1se})$ is minimal, and in all the cases considered the coefficient’s RMSE is minimized when post-Lasso
is fitted. Consequently, we can think of $\log(\lambda_{1se})$ as a good choice for a penalty parameter even though it was not tuned based on the second stage.

Figures 4.3 and 4.4 show the trend of the average number of variables selected, true positives and false positives for Examples 4 and 8, given a grid of possible $\lambda$s. In both examples it can be observed that, for smaller values of the penalty parameter, the majority of the variables included in the model are false positives. As the value of the penalty increases (i.e., as less variables are selected), the number of false positives drops to zero while the true positives are retained into the model. At $\log(\lambda_{1se})$, it can be immediately noted that, when $n > p$ the model chooses on average all the relevant variables (5 out of 5) and does not include any false positives. On the other hand, when $n < p$, at $\log(\lambda_{1se})$ the model selects on average five variables out of the 100 that are set to be relevant; in particular, among the five variables chosen only two have been selected correctly. However, regardless of the value of the penalty parameter $\lambda$, the model never performs well in term of instruments selection. Nevertheless, since $\log(\lambda_{1se})$ is closer to the value of the penalty that minimize the coefficient’s RMSE, it still represents a good choice for a penalty.
Figure 4.1: Trend of the second stage coefficient’s RMSE for different values of $\lambda$. The data is generated based on Example 4, $\text{Cor}(\varepsilon, y) = 0.6$, and the first stage F-statistic is set to 40 (medium strength instruments). A grid of $\lambda$s is generated, and for each value in the grid $N = 500$ datasets are simulated. For each $\lambda$, the RMSE plotted on the y-axis is computed as the mean of the 500 replication. The black vertical line indicates the value of $\log(\lambda)$ such that the cross-validation error in the first stage is within one standard deviation of its minimum.
Figure 4.2: Trend of the second stage coefficient’s RMSE for different values of $\lambda$. The data is generated based on Example 8, $\text{Cor}(\varepsilon, \nu) = 0.6$, and the first stage F-statistic is set to 40 (medium strength instruments). A grid of $\lambda$s is generated, and for each value in the grid $N = 500$ datasets are simulated. For each $\lambda$, the RMSE plotted on the y-axis is computed as the mean of the 500 replication. The black vertical line indicates the value of $\log(\lambda)$ such that the cross-validation error in the first stage is within one standard deviation of its minimum.
Figure 4.3: Trend of the average number of variables selected for different values of $\lambda$. The data is generated based on Example 4, $Cor(\epsilon, \nu) = 0.6$, the first stage F-statistic is set to 40 (medium strength instruments), and the Lasso is used to select the relevant instruments. A grid of $\lambda$s is generated, and for each value in the grid $N = 500$ datasets are simulated. For each $\lambda$, the number of variable selected, the true positive and the false positive plotted on the y-axis are computed as the mean of the 500 replication. The black vertical line indicates the value of $log(\lambda)$ such that the cross-validation error in the first stage is within one standard deviation of its minimum.
Figure 4.4: Trend of the average number of variables selected for different values of $\lambda$. The data is generated based on Example 8, $\text{Cor}(\varepsilon, \nu) = 0.6$, the first stage F-statistic is set to 40 (medium strength instruments), and the Lasso is used to select the relevant instruments. A grid of $\lambda$s is generated, and for each value in the grid $N = 500$ datasets are simulated. For each $\lambda$, the number of variable selected, the true positive and the false positive plotted on the y-axis are computed as the mean of the 500 replication. The black vertical line indicates the value of $\log(\lambda)$ such that the cross-validation error in the first stage is within one standard deviation of its minimum.
Chapter 5

Conclusions

In this thesis we employed different variables selection methods in the context of instrumental variables selection. Looking at scenarios where the relevant instruments are part of a bigger set that contains also weak and possibly irrelevant instrumental variables, we employed each method with the objective of selecting only the relevant instruments. We reviewed some existing methods and we introduced new approaches that might help the researcher in all the settings where the number of available instruments is greater than the sample size.

Based on the results obtained from different simulation studies, we concluded that, regardless of the instruments strength, the available sample size, and the number of instruments, post-regularization methods and supervised principal components outperform regularization methods by reducing the first stage coefficients’ bias. When the sample size and the number of instruments increase, post-Lasso and supervised principal components analysis lead to the same coefficient’s RMSE even though they select a different number of instruments.
Bibliography


56


