Selecting Strong and Exogenous IVs via Structural Error Criteria

Hippolyte Boucher*

November 2, 2022

Abstract

Instrumental variables (IVs) allow consistent estimation of the causal effect of endogenous variables on outcomes. However if IVs are not exogenous and jointly strong, estimators are inconsistent and t-test based Gaussian confidence intervals are invalid. Thus in this paper I design a procedure to select a subset of strong and exogenous IVs among a larger set of potentially weak and / or endogenous IVs in a linear setting. To do so I formally build losses, risks and risk estimators which are based on the structural errors being implicitly minimized when performing IV estimation. I shed light into the empirical and theoretical properties of the risks and find that IV subset selection via risk estimators minimization consistently select strong and exogenous subsets of IVs for the two-stage least squares (2SLS) estimator. More specifically, efficiency and consistency results are established by considering standard asymptotics, weak IV asymptotics and locally invalid IV asymptotics, while maintaining the total number of IVs fixed. I confirm the performances of my IV selection procedures against competing ones' using Monte Carlo simulations and lastly I estimate the causal effect of pre-trial detention on offenders guilt by selecting judge dummy IVs in the first stage.

*Toulouse School of Economics, Hippolyte.Boucher@outlook.com

Latest version available at https://github.com/HippolyteBoucher/Selecting-Strong-and-Exogenous-Instruments-via-Structural-Error-Criteria.git

Acknowledgements: The author thanks Pascal Lavergne, Frank Windmeijer, Eric Gautier and Nour Meddahi for helpful suggestions.

Keywords: Instrument Selection, Valid Instrument, Weak Instrument, Model Selection
JEL Codes: C52, C14

1 Introduction

IVs are used to estimate and infer on the causal effect of endogenous variables on outcomes. Yet applied researchers still struggle in their choice of IVs and specification because of a complicated trade-off between the quality and number of IVs, and the bias, the efficiency, and the asymptotic distribution of their IV estimator. Thus when econometricians have multiple IVs at their disposal one conventional solution is to use all of them. But for this solution to work all IVs must be exogenous and jointly strong, which may not be the case in practice. Another common solution is to use a single IV, typically the IV for which exogeneity can be best justified. However this choice is not data-driven and can actually be quite arbitrary. Instead in this paper I assume that there exists a strong and exogenous subset of IVs and propose data-driven methods to find this subset based on out-of-sample validation. To do so I formally define losses, risks and propose risk estimators in order to consistently choose strong and exogenous IV subsets even in the presence of endogenous and weak IVs.

Prior Work It is well-known that in the presence of weak or endogenous and possibly many IVs the traditional 2 stage least squares estimator (2SLS) is inconsistent and confidence intervals (CI) built from its Gaussian asymptotics have low coverage. See Stock, Wright, and Yogo (2002), Hahn and Hausman (2003), Hahn, Hausman, and Kuersteiner (2004), Kiviet and Kripfganz (2021) for general reviews on weak, many weak, many, and endogenous IVs problems respectively. The literature has treated each problem separately even tough in practice they are likely to occur at the same time.

Regarding weak IVs the literature has developed in the last 25 years with the maintained assumption that IVs are exogenous. It has mainly focused on detecting weak IVs, see Stock and Yogo (2005), Kleibergen (2007), and Olea and Pflueger (2013), and on inference procedures which are robust to weak identification, see Anderson and Rubin (1949), Kleibergen (2002), Moreira (2003), with their subvector counterparts, see Guggenberger, Kleibergen, Mavroeidis, and Chen (2012), and their nonlinear first

stage counterparts, see Antoine and Lavergne (2022) and Boucher (2022). The current consensus being that if weak IVs are detected such inference procedures should be used.

The literature on many (weak) IVs has studied the behavior of different k-class estimators under a many exogenous IVs or a many weak exogenous IVs assumption. Compared to 2SLS, other k-class estimators reduce finite sample bias, can be consistent under specific assumptions on the types of asymptotics, see Hahn et al. (2004), and can allow for valid inference, see Mikusheva and Sun (2021) and Andrews, Marmer, and Yu (2019) for reviews on inference under many weak IVs, and many IVs asymptotics respectively. Part of this literature has focused on regularizing these estimators, selecting IVs based on first stage fit, see Donald and Newey (2001), Bai and Ng (2010), Carrasco (2012), Belloni, Chen, Chernozhukov, and Hansen (2012), Chen, Chen, and Lewis (2021).

Lastly, to deal with (many) endogenous but strong IVs, the literature has focused on detecting endogeneity at the vector or subvector level using overidentifying restrictions types of tests with more recent papers allowing for heteroskedasticity and possibly many IVs, see Sargan (1958), K. Newey (1985), Hahn and Hausman (2002), Carrasco and Doukali (2021), and also focused on devising procedures to select the exogenous IVs directly or indirectly via regularization, see Andrews (1999), Hall and Peixe (2003), Caner (2009), Kang, Zhang, Cai, and Small (2016), Windmeijer, Farbmacher, Davies, and Smith (2018), Gautier and Rose (2021).

Contribution In a context with a finite number of IVs which is most common in practice, a strong argument can be made in favor of constructing criteria in order to select a strong and exogenous subset of IVs. Such criteria do not yet exist and would allow to remove the irrelevant and misleading information contained in the full set of IVs. Indeed in case all the IVs are exogenous or exogenous and strong, IV selection can only improve inference and lower finite sample bias by reducing the number of IVs and improving their overall strength. On the other hand when IVs can be endogenous selection is a necessity otherwise the true causal effect cannot be estimated consistently and valid inference cannot be performed. But thus far current popular methods for IV selection are either unable to pick exogenous IVs if endogenous IVs are present as in Donald and Newey (2001), Bai and Ng (2010), Belloni et al. (2012), or Carrasco and Tchuente (2016), either unable to pick up strong exogenous IVs if weak exogenous IVs are present as in Andrews (1999) or Kang et al. (2016), either require an a priori

consistent estimator as in Donald and Newey (2001) or Windmeijer et al. (2018).

Accordingly I consider a linear IV model where the total number of IVs remain fixed, where IVs can be correlated but also individually weak or strong, and where some IVs enter the structural equation. This setting with a finite number of IVs and an unknown subset of exogenous IVs is the most common in practice. Note that the endogenous IVs may directly affect the outcome or indirectly through some unobserved regressor which they are correlated to. The goal is then to find a subset of strong and exogenous IVs among the full set of IVs. This is a model selection problem (see Arlot and Celisse (2010) and Bates, Hastie, and Tibshirani (2021) for recent surveys on model selection) thus I design selection criteria for IV subsets based on the structural error being implicitly minimized in an IV setting. Hence I define three prediction losses for IV subsets: A loss based on the exogeneity condition; The mean squared error of prediction where the endogenous variable has been projected on the IVs; And the mean squared error of prediction. Then I define the corresponding risks, which are average prediction losses, and their corresponding cross-validation estimators for IV subsets. Finally the IV subset which minimizes the risk is selected and used as IVs whereas the rest of the IVs are considered endogenous and therefore used as control variables. In terms of theory I provide a decomposition of the risks and show that the IV subset selection procedures are efficient and consistent. If in the full set of IVs there exists a subset which is strong and exogenous it will be selected with probability one at the limit. These results are established by allowing for weak IVs (in the sense of Staiger and Stock (1997) and Andrews and Cheng (2012)), and by allowing for locally invalid IVs (as in the literature on local misspecificaiton, see Maasoumi and Phillips (1982), and on sensitivity analysis, see Andrews, Gentzkow, and Shapiro (2017)). I confirm these findings by looking at the performances of the 2SLS estimator and at the performances of the weak-identification robust inference procedure from Moreira (2003) in an extensive simulation exercise. I also apply my methods and select judge dummy variables in order to estimate the effect of pre-trial detention on the likelihood of being found guilty.

Outline The outline of the paper is as follows. In the second section of this paper I present the linear IV model when considering subset of IVs and interpret the estimated parameter in terms of losses. Then in the third section I define the losses, risks and risk estimators which are relevant for IV subset selection and present the selection proce-

dure. In the fourth section I present theoretical guaranties that strong and exogenous sets of IVs are systematically selected compared to weak and endogenous subsets. In the fifth section I show through simulations that the methods consistently select strong and exogenous IV subsets and therefore yield estimators and inference procedures with great performances. In the sixth section I apply my methods and estimate the effect of pre-trial detention on offenders' probability of being found guilty. I conclude in the seventh and final section.

2 Model, Estimator, and Loss Interpretation

2.1 Linear IV Model with Endogenous IVs

Consider the following linear IV model with outcome y_i, a single endogenous variable x_i, K_z IVs z_i among which z_{iE} is an exogenous subset and its complement $z_{i\overline{E}}$ is endogenous and enter the structural equation linearly

$$y_i = x_i\beta + z'_{i\overline{E}}\alpha + u_i, \qquad \mathbb{E}(u_i|z_i) = 0, \qquad \mathbb{E}(u_i^2|z_i) = \sigma_u^2, \qquad \mathbb{E}(u_i v_i|z_i) = \rho \quad (2.1)$$

$$x_i = z'_i\pi + v_i, \qquad \mathbb{E}(v_i z_i) = 0, \qquad \mathbb{E}(v_i^2|z_i) = \sigma_v^2 \quad (2.2)$$

for $i = 1, \ldots, n$ where $z'_i\pi \equiv z'_{iE}\pi_E + z'_{i\overline{E}}\pi_{\overline{E}}$ and $\pi_E \neq 0$. To simplify exposition and calculation I also impose that the data is centered. The model characterized by the structural equation or second stage (2.2) and the reduced form equation or first stage (2.1) is very common in applied work. A single causal effect β is of interest and a few IVs are used to try to consistently estimate it, see the example below. The set of IVs may be the result of some interactions between IVs and exogenous controls, or the result of modeling non-linearly the relation between the endogenous variable and IVs. The only departure from the usual linear IV model is that only the subset of IVs z_{iE} is exogenous. Intuitively, the subset of IVs $z_{i\overline{E}}$ does not satisfy the exclusion restriction and affects y_i either directly or indirectly through some unobservable, see figure 1 in appendix A.1 for some visualization. If $\alpha = 0$ then the model reduces to the usual IV model under exogeneity.

Note that the selection procedures and the formal results are later established under an independent and identically distributed and conditional homoskedasticity assumption which can be relaxed to allow for conditional heteroskedasticity without modifying

the selection criteria. Extending the results to the case where x_i is a vector is also possible but requires a more complex modelization of IV weakness at the vector level and additional assumptions to obtain the consistency of the selection procedures. As for exogenous controls they can be projected out a la Firsch-Waugh with little consequences and are therefore omitted. A setting with many IVs or a fully non-linear modelization of the first stage are outside the scope of this paper.

Clearly β is identified when using z_{iE} as IVs and $z_{i\overline{E}}$ as control variables but in practice it cannot be estimated consistently because E is unknown. For this reason it is key to reformulate the model and consider the 2SLS estimator for a specific subset of IVs but before that consider the following simple example of the model characterized by (2.1) and (2.2).

Example: Weather IVs Consider estimating the following demand curve for fish at the Fulton fish market as in Graddy (2006)

$$Q_i = P_i\beta + X_i'\delta + u_i$$

where i denotes the day, Q_i the total amount of fish sold during day i, P_i the average daily price of the fish, and X_i various control variables. Clearly P_i is endogenous because it is determined simultaneously with Q_i. Consequently the demand curve must augmented by the following reduced form first stage equation

$$P_i = \pi_1 cold_i + \pi_2 wind_i + \pi_3 rain_i + \pi_4 stormy_i + \pi_5 mixed_i + X_i'\gamma + v_i$$

where $cold_i$, $wind_i$, $rain_i$, $stormy_i$, and $mixed_i$ are available weather IVs. Then to identify β it must be that the weather variables are cost shifters which affect demand only through price. But some IVs such as $wind_i$ may affect supply significantly, and some IVs such as $cold_i$ may not have a direct effect on demand. Thus it is unclear which weather IV is truly exogenous and strong. Therefore instead of arguing (with difficulty) for the validity of specific weather IVs it seems much more natural to select a strong and exogenous subset of IVs in a data driven way.

2.2 Subset Model and Subset IV 2SLS

Before describing the IV subset selection method, the model and the 2SLS estimator for a given subset of IVs must be defined and additional notations must be introduced.

In the rest of the paper let $\mathcal{S}$ denote the collection of all non-empty subsets of

$$\{z_{i1}; z_{i2}; \ldots; z_{iK_z}\}$$

The cardinality of $\mathcal{S}$ is therefore $2^{K_z} - 1$. The complement of S is denoted as $\overline{S}$ in the sense that $S \cup \overline{S} = \{z_{i1}; z_{i2}; \ldots; z_{iK_z}\}$ and $S \cap \overline{S} = \emptyset$. The IVs associated to $S \in \mathcal{S}$ are denoted as z_{iS} which is a random vector of dimension $s = |S|_0$ where $|\cdot|_0$ denotes the counting norm, and π_S is the subvector of π of dimension s associated to S. Let $\Sigma \equiv \mathbb{E}(z_i z_i')$, and for any $S \in \mathcal{S}$ let $\Sigma_S = \mathbb{E}(z_{iS} z_{iS}')$. In addition to simplify notations I denote $w_i \equiv (y_i, x_i, z_i')'$ the observed variables for individual i. Furthermore let $y \equiv (y_1, y_2, \ldots, y_n)'$ be the $n \times 1$ vector of stacked outcomes over the sample, let $x \equiv (x_1, x_2, \ldots, x_n)'$ be the $n \times 1$ vector of stacked endogenous variables, let $z \equiv (z_1 \ z_2 \ \ldots \ z_n)'$ be the $n \times K_z$ matrix of stacked IVs, and let $w \equiv (w_1 \ w_2 \ \ldots \ w_n)'$ be the $n \times (K_z + 2)$ matrix of stacked observed variables. Similarly $u = (u_1, u_2, \ldots, u_n)$, $v = (v_1, v_2, \ldots, v_n)$, and $z_S = (z_{1S} \ z_{2S} \ \ldots \ z_{nS})'$ for any $S \in \mathcal{S}$.

As discussed the subset of exogenous IVs E is unknown a priori thus without selection it is not possible to estimate β consistently. As a consequence some candidate S has to be considered for instrumentation and its complement $\overline{S}$ has to enter as a vector of control variables. When S is picked for instrumentation and $S \subset E$ the model is called valid, whereas when $S = E$ the model is called the oracle model. In both these cases β can be estimated consistently, see figure 2 in appendix A.1. If the full vector of IVs z_i is used for instrumentation then no instrument should be included as a control variable. Consequently if subset S is considered for instrumentation the model can be rewritten as

$$y_i = x_i\beta + z_{i\overline{S}}'\alpha_{\overline{S}} + u_{i\overline{S}}, \qquad x_i = z_{iS}'\pi_S + z_{i\overline{S}}'\pi_{\overline{S}} + v_i, \qquad \mathbb{E}(z_i v_i) = 0 \tag{2.3}$$

for some $(\alpha_{\overline{S}}, u_{i\overline{S}})$. Since $z_{i\overline{S}}$ is considered a vector of control variables, it can be projected out a la Frisch-Waugh. Thus except in the rest of the paper except the proofs, I denote $(y_i, x_i, z_{iS}) \equiv (y_i, x_i, z_{iS}) - BLP((y_i, x_i, z_{iS})|z_{i\overline{S}})$ where $BLP(\cdot|z_{i\overline{S}})$ is the best linear projection on $z_{i\overline{S}}$. Consequently the model instrumented by S (2.3) can be rewritten as

$$y_i = x_i\beta + u_{iS}, \qquad \mathbb{E}(u_{iS} z_{i\overline{S}}) = \mathbb{E}(x_i z_{i\overline{S}}) = 0 \tag{2.4}$$

$$x_i = z_{iS}'\pi_S + v_i, \qquad \mathbb{E}(v_i z_i) = 0 \tag{2.5}$$

where $u_{iS} = z'_{i\overline{E}}\alpha + u_i$, $Var(u_{iS}|z_i) = \sigma_u^2$, $Var(v_i|z_i) = \sigma_v^2$ and $Cov(u_{iS}, v_i|z_i) = \rho$. Hence if subset S is considered for instrumentation the fact that β can be estimated consistently depends on whether or not $\mathbb{E}(z_{iS}u_{iS})$ is close to zero, which is the case when α is close to zero and when $\overline{S}$ is close to $\overline{E}$, and on whether or not π_S is equal to zero.

From the model instrumented by S characterized by (2.4) and (2.5) I define the 2SLS subset IV estimator

$$\hat{\beta}_S = \frac{x'P_{z_S}y}{x'P_{z_S}x} = \beta + \frac{x'P_{z_S}u_S}{x'P_{z_S}x}$$

where $P_{z_S} = z_S(z'_S z_S)^{-1}z_S$ is the orthogonal projection on z_S. For exposition the paper focuses on 2SLS but in appendix D.4 I define the k-class of IV subset estimators, which includes 2SLS and other common IV estimators such as Fuller, limited information maximum likelihood, debiased 2SLS, etc... k-class estimators enjoy better finite sample properties compared to 2SLS and can have better large sample properties, see Hausman, Newey, Woutersen, Chao, and Swanson (2012) for a review. Appendix D.4 also includes conditions under which the theoretical results derived in the paper hold for k-class IV estimators.

2.3 Interpreting the Causal Effect in terms of Losses

Having rewritten the model and estimator when subset S is considered for instrumentation, it is necessary to understand what are the losses of interest in the linear IV model. This will guide the choice of criteria for IV subset selection.

Traditionally the causal effect β is defined as the parameter for which the exogeneity condition is satisfied. Consequently when instrumenting with S the structural parameter β is the minimizer of a weighted sum of the squared correlations between the subset of IVs S and the error

$$\beta(S, W) = \underset{\tilde{\beta}}{Argmin}\ \mathbb{E}((y_i - x_i\tilde{\beta})z'_{iS})W\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta}))$$

for some symmetric full ranked weighting matrix W. $\beta(S, W)$ is the set of parameters that can be estimated given weights W and IVs S, it is the set of pseudo true values or target set induced by the minimization of the loss based on the exogeneity condition. Taking some empirical counterpart of the population moments in the objective and

assuming validity of the IVs will yield the different k-class and GMM estimators. A natural candidate for W is $W = \Sigma_S^{-1}$, ie the correlation structure of the IVs is controlled for. Therefore given IV set S let β_S be the target parameter set in IV estimation which minimizes the following exogeneity based loss

$$\beta_S = \underset{\tilde{\beta}}{Argmin}\ \mathbb{E}((y_i - x_i\tilde{\beta})z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta})) \tag{2.6}$$

If S is irrelevant as in $\pi_S = 0$ then $\beta_S = \mathbb{R}$, if S is relevant but possibly endogenous as in $\pi_S \neq 0$ then $\beta_S = \beta + (\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha$, if S is relevant and exogenous as in $\pi_S \neq 0$ and $\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = 0$ then and only then is the parameter to be estimated equal to the causal effect $\beta_S = \beta$. Hence, as the mean square error is the loss of interest in linear model, a loss of specific interest in linear IV models is based on the exogeneity condition.

More historically, the goal behind IV estimation is trying to perform ordinary least squares minimization while circumventing the endogenous nature of the regressor. So instead of using x_i as a regressor only its (potentially) exogenous part $BLP(x_i|z_{iS}) = z'_{iS}\pi_S$ the best linear projection of x_i on z_{iS} is used. In that sense β_S can be rewritten as

$$\beta_S = \underset{\tilde{\beta}}{Argmin}\ \mathbb{E}((y_i - z'_{iS}\pi_S\tilde{\beta})^2)$$

As before $\beta_S = \mathbb{R}$ if $\pi_S = 0$, $\beta_S = \beta + (\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha$ if $\pi_S \neq 0$, and $\beta_S = \beta$ if $\pi_S \neq 0$ and $\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = 0$. Hence a second natural loss to minimize in a linear IV model given IV set S is the mean square error with the endogenous variable projected on the IVs. See appendix D.1 and propositions 4.1 and 4.2 for formal results.

Having defined the IV subset estimator and the losses of interest in linear IV models for any candidate for instrumentation S, I formally design the criteria for the selection of IV subsets in the next section.

3 Risks for IV Sets

In statistics and machine learning model selection is now understood within a common framework, see Arlot and Celisse (2010), and IV subset selection can also be understood within this framework. Thus in this section I introduce prediction losses, risks, and

risk estimators for IV subsets. In practice these risk estimators are computed for each IV subset and the subset which minimizes them will be chosen for instrumentation. Lastly I describe the desirable properties of IV subset selection procedures.

3.1 Model Selection for IV Subsets

Prediction Losses In model selection the performances of an estimator (or of an IV subset in this case) are measured with prediction losses. Indeed if the right model is picked then it should perform very well with new data. Prediction losses are usually defined as average prediction errors or average out-of-sample discrepancies with respect to a new observation w^* conditional on the original sample $(w_i)_{i=1}^n$ where w^* has the same DGP but is independent of $(w_i)_{i=1}^n$. Let $\mathbb{E}_n(\cdot) \equiv \mathbb{E}(\cdot|(w_i)_{i=1}^n)$ denote the expectation conditional on the original sample, then define the following losses for any IV subset S

$$
\begin{aligned}
L_{EXO}(w^*; \tilde{\beta}, S) &= \mathbb{E}_n\left((y^* - x^*\tilde{\beta})z_S^{*'}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(y^* - x^*\tilde{\beta})\right) \\
L_{PMSE}(w^*; \tilde{\beta}, S) &= \mathbb{E}_n\left((y^* - z_S^{*'}\pi_S\tilde{\beta})^2\right) \\
L_{MSE}(w^*; \tilde{\beta}) &= \mathbb{E}_n\left((y^* - x^*\tilde{\beta})^2\right)
\end{aligned}
$$

L_{EXO} corresponds to an out-of-sample counterpart of the exogeneity based loss defined in the previous section, if a new observation w^* is at disposal then the correlation between the error and the IVs should be small and therefore L_{EXO} should be small. L_{PMSE} corresponds to an out-of-sample counterpart of the mean square error of prediction after projecting the endogenous variable on the IVs, again for a new observation it should naturally be small if the right IV subset was picked. L_{MSE} is the mean square error of prediction of the structural equation. While L_{MSE} is not of direct interest in the context of linear-IV models, it is already extensively used for nuisance parameter selection (bandiwdth, Lasso penalty, basis size, etc...) of various IV estimators and procedures, see Chernozhukov, Hansen, and Spindler (2015) or Kang et al. (2016), thus its IV subset selection properties are also studied. As will be shown in the next section the risk based on the mean square error of prediction can actually select strong and exogenous IV subsets in certain conditions.

Other types of prediction losses can be discussed but they are less appealing. Excess losses are versions of losses which are "centered" with respect to the true β, but in the

linear-IV context "centering" is difficult because β is identified by IV subset E which is unknown a priori. Losses which do not depend on S are in line with the literature on predictors selection in linear models but too different from the losses being implicitly minimized during IV estimation. Alternatively losses built from the log-likelihood of (y^*, x^*) given z_S^* could be used. However they require the first stage (2.5) to be causal which is very hard to argue for. Finally integrated losses are much more suited to a non-linear setting with fewer variables. See Arlot and Celisse (2010) for formal definitions.

Risks To properly assess the performances of $\hat{\beta}_S$, losses have to be evaluated at $\hat{\beta}_S$ and averaged because they are random. These average prediction losses are called risks. The risks for any IV subset S to consider are therefore

$$\begin{aligned}
R_{EXO}(S) &= \mathbb{E}\left(\mathbb{E}_n\left((y^* - x^*\hat{\beta}_S)z_S^{*\prime}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(y^* - x^*\hat{\beta}_S)\right)\right) \\
R_{PMSE}(S) &= \mathbb{E}\left(\mathbb{E}_n\left((y^* - z_S^{*\prime}\pi_S\hat{\beta}_S)^2\right)\right) \\
R_{MSE}(S) &= \mathbb{E}\left(\mathbb{E}_n\left((y^* - x^*\hat{\beta}_S)^2\right)\right)
\end{aligned}$$

for some new observation w^*. The risks are thoroughly decomposed and interpreted in section 4.1. Two other risks have been formalized in the literature, Donald and Newey (2001) assume that all IVs are exogenous and strong and directly consider the conditional mean squared error of $\hat{\beta}$

$$\forall S \in \mathcal{S} \qquad R_{DN}(S) = \mathbb{E}\left[(\hat{\beta}_S - \beta)^2 | (z_i)_{i=1}^n\right]$$

which the authors approximate using an a priori consistent estimator of β. More precisely the authors use Nagar (1959) expansions to approximate the bias of different IV estimators and make use of an a priori consistent estimator of β. These expansions are known to be unstable even in the best case scenario and do not hold if the IVs being considered are weak, very weak or endogenous, see Chaudhuri and Zivot (2011). On the other hand Andrews (1999) assumes that all IVs are strong and coins different criteria based on Sargan-Hansen J statistics to pick the largest set of exogenous IVs. Hence the risk the author actually estimate is

$$\forall S \in \mathcal{S} \qquad R_A(S) = \mathbb{E}\left[(y_i - x_i\hat{\beta}_S)z_{iS}'\right] Var\left[z_{iS}(y_i - x_i\hat{\beta}_S)\right]^{-1} \mathbb{E}\left[z_{iS}(y_i - x_i\hat{\beta}_S)\right]$$

up to some normalization to account for $s = |S|_0$. Note that $R_A(S)$ is a normalized in-sample version of $R_{EXO}(S)$.

Risk Estimators The risks R_{EXO}, R_{PMSE} and R_{MSE} are unknown and need to be estimated. I consider their cross-validation average estimators[1] denoted as $\hat{R}_{EXO}$, $\hat{R}_{PMSE}$, and $\hat{R}_{MSE}$. To obtain them in practice the following steps can be followed:

Cross-Validation Average Risk Estimator

1. Split the original sample into a validation sample of size n_c and a training sample of size $n - n_c$
2. Compute $\hat{\beta}_S$ using the data from the training sample only
3. Use the validation sample to estimate R_{EXO}, R_{PMSE} and R_{MSE} but plug-in the estimator $\hat{\beta}_S$ created using the training sample
4. Repeat the process B times and average

To be more specific let B be the number of times the original sample is split, n_c be the validation sets sample size, and $n - n_c$ be the training sets sample size. Then for any $b = 1, \ldots, B$ let $(w_i)_{i \in I_b}$ be the validation sample for split b of size n_c and let $(w_i)_{i \in \overline{I}_b}$ be the training sample for split b of size $n - n_c$. Finally let $\hat{\beta}_{S,b}$ be the 2SLS estimator associated to split b which uses the training sample $\overline{I}_b$ only. Formally for any $S \in \mathcal{S}$ the risk estimators are

$$\hat{R}_{EXO}(S) = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{n_c(s+1)\hat{\sigma}_b^2} \sum_{i \in I_b} \left((y_i - x_i\hat{\beta}_{S,b})z_{iS}'\right) \hat{\Sigma}_S^{-1} \frac{1}{n_c} \sum_{i \in I_b} \left(z_{iS}(y_i - x_i\hat{\beta}_{S,b})\right)$$

$$\hat{R}_{PMSE}(S) = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{n_c\hat{\sigma}_b^2} \sum_{i \in I_b} \left((y_i - z_{iS}'\hat{\pi}_S\hat{\beta}_{S,b})^2\right)$$

$$\hat{R}_{MSE}(S) = \frac{1}{B}\sum_{b=1}^{B} \frac{1}{n_c\hat{\sigma}_b^2} \sum_{i \in I_b} \left((y_i - x_i\hat{\beta}_{S,b})^2\right)$$

where $\hat{\sigma}_b^2$ is a normalization which controls for differences in variations across splits. This normalization is useful in finite sample, for instance $\hat{\sigma}_b^2 = \frac{1}{n_c}\sum_{i \in I_b}^{n_c}(y_i - \overline{y})^2$ or $\hat{\sigma}_b^2 = \frac{1}{n_c}\sum_{i \in I_b}^{n_c}(x_i - \overline{x})^2$, in large samples one can set $\hat{\sigma}_b^2 = 1$. Other types of normalizations are possible for instance adding a degenerate bonus term in s as in Andrews (1999) would allow to select larger subsets of IVs. The practical choice of B, n_c and $n - n_c$ is

[1] In practice other methods such as out-of-bag bootstrap validation or k-fold cross-validation can be used, moreover instead of the average risk the median risk or most voted risk can also be used.

up to the researcher. A standard choice in machine learning is forty splits with a third of the data used for validation and two thirds of the data used for training, ie $B = 20$, $n_c = \frac{n}{3}$, and $n - n_c = \frac{2n}{3}$. To establish asymptotic results a requirement is that B, n_c, and $n - n_c$ increase with n.

Selection Procedure For a certain risk $k \in \{EXO; PMSE; MSE\}$ the selected subset of IVs is simply the minimizer of the risk estimator

$$\hat{S}_{\hat{R}_k} = \underset{S \in \mathcal{S}}{Argmin}\ \hat{R}_k(S)$$

In that sense if K_z is large it becomes very time consuming to compute $\hat{R}_k(S)$ for all $2^{K_z} - 1$ subsets in $\mathcal{S}$. However even if $K_z \geqslant 10$ it is still possible to simplify the problem by minimizing the risks over only part of $\mathcal{S}$ or if there are groups of uncorrelated IVs by minimizing the risks in each group as in Windmeijer, Liang, Hartwig, and Bowden (2021).

3.2 Ideal Properties of Risk Estimators

Before deriving the theoretical performances of the selection procedures I characterize their ideal properties.

Efficiency A selection method $\hat{S}_{\hat{R}_k}$ is deemed efficient if its minimum converges to the minimum of the risk it is trying to estimate

$$\frac{\underset{S \in \mathcal{S}}{min}\ \hat{R}_k(S)}{\underset{S \in \mathcal{S}}{min}\ R_k(S)} \xrightarrow{\mathbb{P}} 1 \tag{3.7}$$

A procedure being efficient does not directly imply that the selection procedure will select a "good" model however. This is especially the case in linear IV models.

Consistency A consistent model selection procedure is a procedure which selects the true model with probability 1 at the limit. But defining a true or good model in the linear IV context is difficult. A candidate set of good models of interest may be set of all IV subsets which allow to identify β

$$\mathcal{S}_{id} = \{S \in \mathcal{S} : \alpha = 0, \pi_S \neq 0\}$$

If $\mathcal{S}_{id}$ is non-empty then there exists at least one valid subset of IVs and therefore β is identified. But identification of β does not guarantee its consistent estimation, in fact a local lack of identification does not prevent consistent estimation and.

Thus IV subsets of much more interest are the subsets for which $\hat{\beta}_S$ is a consistent estimator of β and the subsets for which the t-statistic $t_S = \frac{\hat{\beta}_S - \beta}{\hat{Var}(\hat{\beta}_S)}$ is asymptotically standard normal. To characterize these sets I let a_S and b_S represent respectively the strength of IV z_{iS}, ie $\pi_S \propto n^{-a_S}$, and the level of endogeneity of IV z_{iS}, ie $\mathbb{E}(z_{iS}u_{iS}) \propto n^{-b_S}$. Allowing π_S and $\mathbb{E}(z_{iS}u_{iS})$ to depend on sample size is a way to model IV weakness and IV local endogeneity. This generalization resembles that of Andrews and Cheng (2012) and is a theoretical way to approximate the behavior of IV estimators and inference procedures under unfavorable conditions in practice. Hence define the three following categories of IV subsets

$$\mathcal{S}_c = \{S \in \mathcal{S} : b_S - a_S > 0, a_S < 1/2\}$$
$$\mathcal{S}_{an} = \{S \in \mathcal{S} : a_S < 1/2, b_S > 1/2\}$$

It can be shown that $\mathcal{S}_c$ represents all the subsets of IVs such that $plim\ \hat{\beta}_S = \beta$ whereas $\mathcal{S}_{an}$ represents all the IV subsets such that $\hat{\beta}_S$ is consistent and asymptotically normal in the sense that under the null $H_0 : \beta = 0$ the ration $\hat{\beta}_S$ and an estimator of its standard deviation is asymptotically standard normal and therefore the usual Gaussian confidence intervals are valid. Another category of IV subsets of specific interest is

$$\mathcal{S}_r = \{S \in \mathcal{S} : b_S > 1/2\}$$

which characterizes the subsets which yield valid weak identification robust confidence sets for β via test inversion as in Anderson and Rubin (1949). See appendix D.2 and proposition 4.3 for formal proofs.

Going forward $\hat{S}_{\hat{R}_k}$ is c-consistent where $c \in \{c; an; r\}$ if

$$\mathbb{P}(\hat{S}_{\hat{R}_k} \in \mathcal{S}_c) \to 1 \tag{3.8}$$

Efficiency and consistency of IV selection via $\hat{R}_{EXO}$, $\hat{R}_{PMSE}$ and $\hat{R}_{MSE}$ are proven in the next section.

Valid Post Model Selection Inference Lastly, let $CI_{\alpha,S}(\beta)$ be a confidence interval with nominal coverage α using subset of IVs S based on either weak identification robust methods, either t-tests Gaussian asymptotics. Then if there exists some

$S \in S_r$ or $S \in S_{an}$ valid inference is possible, ie $lim\ \mathbb{P}(\beta \in CI_{\alpha,S}(\beta)) \geqslant \alpha$. Ideally post-selection inference should also be valid however $\hat{S}_{\hat{R}_k}$ is correlated with the data therefore

$$\mathbb{P}(\beta \in CI_{\alpha,\hat{S}_{\hat{R}_k}}(\beta) | \hat{S}_{\hat{R}_k} = S) \neq \mathbb{P}(\beta \in CI_{\alpha,S}(\beta))$$

Valid inference on β post selection via $\hat{R}_{EXO}$, $\hat{R}_{PMSE}$ and $\hat{R}_{MSE}$ is not formally proven in this paper. But from the extensive simulation exercise in section 5 this contamination does not seem to be a concern as confidence intervals have nominal coverage.

Still this issue can be completely bypassed and exact inference can be recovered by using sample-splitting[2] with one sample used for finding $\hat{S}_{\hat{R}_k}$ and the other used for estimation and inference. Other common methods for valid-post-selection inference in econometrics and statistics systematically involve immunization of either the inference procedure or the estimation procedure to the choice of nuisance parameter or to the models. For instance if the estimator is modified by Neyman orthogonalization of the score of the criterion it is built from, as in Chernozhukov et al. (2015) or Singh and Sun (2021) in the IV context, it will require all the IVs to be exogenous. A partial identification approach immune to the choice of IVs and therefore endogeneity itself may be used however the confidence interval will be very large unless strong additional assumptions are made. Thus these approaches are not very appealing, even more so because the IV estimator is taken as given. In fact ideally the estimator and inference procedure should not be immune to the choice of IV subset otherwise it would be impossible to assert the performances of each IV subset.

On a final note $\mathcal{S}_{an}$ being non-empty implies that there exists at least some IVs which are only locally endogenous, if instead $\mathcal{S}_{an}$ is empty valid post-selection inference may be possible for a pseudo-true-value instead.

4 Theoretical Properties

To understand the properties of IV selection methods via minimization of $\hat{R}_{EXO}$, $\hat{R}_{PMSE}$ and $\hat{R}_{MSE}$ defined in section 3.1, the risks are first decomposed then their theoretical asymptotic properties are derived.

[2]Note that sample splitting will effectively reduce sample size and therefore could aggravate weak IVs problems but at the same time it could improve the level of exogeneity of the IVs.

4.1 Risks Decomposition

In linear models the mean square error of prediction decomposes into the squared bias and variance of estimators and the same decomposition exercise can be performed with R_{EXO}, R_{PMSE} and R_{MSE}, formal proofs are in appendix D.3. For any $S \in \mathcal{S}$ the risks can be rewritten as

$$R_{EXO}(S) = \mathbb{E}\left(||\Sigma_S^{-1/2}\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha - \Sigma_S^{1/2}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$

$$R_{PMSE}(S) = \mathbb{E}\left((u^* + v^*\beta)^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$

$$R_{MSE}(S) = \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$

Clearly the risks do not decompose into squared bias and variance. Instead they mainly depend on some average distance between $z_{\overline{E}}^{*'}\alpha$ and $z_S^{*'}\pi_S(\hat{\beta}_S - \beta)$ and it is possible to go further.

Strong and Endogenous IVs Assume that IV subset S is strong but endogenous, ie assume that $\pi_S \neq 0$ and is fixed and that $\mathbb{E}(z_{iS}u_{iS}) = \mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha \neq 0$ and is fixed. Then it can be shown that

$$\hat{\beta} - \beta = \frac{\pi'_S z'_S z_E \alpha}{\pi'_S z'_S z_S \pi_S} + o_P(1) = \frac{\pi'_S \mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha}{\pi'_S \mathbb{E}(z_{iS}z'_{iS})\pi_S} + o_P(1)$$

where $o_P(1)$ is the small o in probability notation[3]. As a consequence the risks can be rewritten as

$$R_{EXO}(S) = \alpha'\mathbb{E}(z_{\overline{E}}^* z_S^{*'})M_1\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha + o_P(1)$$

$$R_{PMSE}(S) = \mathbb{E}\left((u^* - v^*\beta)^2\right) + \alpha' M_2 \alpha + o_P(1)$$

$$R_{MSE}(S) = \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \alpha' M_2 \alpha + o_P(1)$$

where M_1 and M_2 are positive semi-definite matrices defined as

$$M_1 = \Sigma_S^{-1} - \pi_S(\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S, \qquad M_2 = \Sigma_E - \mathbb{E}(z_{\overline{E}}^* z_S^{*'})\pi_S(\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S\mathbb{E}(z_S^* z_{\overline{E}}^{*'})$$

From this decomposition it is clear that the risks are mainly quadratic functions of $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$ because $\hat{\beta}_S - \beta$ is also a function of $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$. Thus the larger the amount

[3]Formally if random sequence $X_n = o_P(1)$ then $\forall \varepsilon > 0\ \mathbb{P}(|X_n| > \varepsilon) \to 0$. If random sequence $X_n = O_P(1)$ then $\forall \varepsilon > 0\ \exists M > 0, \exists N > 0 : \forall n > N\ \mathbb{P}(|X_n| > M) < \varepsilon$.

of endogeneity or equivalently the larger $\mathbb{E}(z_S^* u_S^*) = \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$ is, the larger are the risks. Note the first term in R_{MSE} also depends on $\hat{\beta}_S - \beta$ thus this risk may misclassify some IV subsets.

Exogenous IVs This time assume that the right IVs were chosen $S = E$, or that all the IVs are exogenous $\alpha = 0$, then the risks reduce to quadratic terms of $\hat{\beta}_S - \beta$

$$R_{EXO}(S) = \mathbb{E}\left(||\Sigma_S^{1/2}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$
$$R_{PMSE}(S) = \mathbb{E}\left((u^* - v^*\beta)^2\right) + \mathbb{E}\left(||z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$
$$R_{MSE}(S) = \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$

Thus if the subset S is strong then $\hat{\beta}_S$ is a consistent estimator of β and the three risks further reduce asymptotically, whereas if the subset S is weak then $\hat{\beta}_S$ does not converge to β and the three risks are large.

To summarize when ranking IV subsets the risks are not weighting subsets to balance bias and variance, instead they are larger if IVs are endogenous and slightly larger if the IVs are weak. In that sense endogeneity is first order whereas weakness is second order, this is a desirable property because valid inference on β can still be performed regardless of the level of strength as long as the IVs being picked are exogenous.

4.2 Asymptotic Results

In order to establish the efficiency of the selection procedures three sets of assumptions are made. Assumption A characterizes the model and the data generating process (DGP). Assumption B places restrictions on the k-class of IV subset estimators. Assumption C characterizes the risk estimator to make sure it convergences towards the risk.

Assumption A
(i) The sample $(y_i, x_i, z_i)_{i=1}^n$ is iid such that (2.1) and (2.2) hold at β
(ii) z_i, x_i and y_i possess finite moments of order 4, z_i is not perfectly colinear
(iii) Without loss of generality for any $S \in \mathcal{S}$, $\pi_S \equiv n^{-a_S}\kappa_S$ for some fixed $\kappa_S \in \mathbb{R}_*^s$ and some $a_S \in \mathbb{R}^+ \cup \{+\infty\}$

(iv) Without loss of generality for any $S \in \mathcal{S}$, $\mathbb{E}(z_{iS}u_{iS}) = \mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = n^{-b_S}\delta_S$ for some fixed $\delta_S \in \mathbb{R}^s_*$ and some $b_S \in \mathbb{R}^+ \cup \{+\infty\}$

Assumption A(i) determines the model whereas A(ii) is a common moments condition. Alternatively conditional heteroskedasticity could be assumed. A(iii) and A(iv) formally allow the level of weakness and the level of endogeneity of any subset of IVs to vary with n, thus the asymptotic behavior of the IV estimator and risk estimators is characterized by the values of (a_S, b_S). For instance when $a_S \geqslant 1/2$, IV subset S is weak and therefore the estimator $\hat{\beta}_S$ is random at the limit. See lemma in appendix C for details.

Assumption B
For any $S \in \mathcal{S}$ there exists some $e > 0$ such that

- *If $a_S \geqslant 1/2$*
$$\mathbb{P}(x'P_{z_S}x < e) = 0$$
- *If $a_S < 1/2$*
$$\mathbb{P}(n^{2a_S-1}x'P_{z_S}x < e) = 0$$

Assumption B is a condition which ensures the existence of moments of the 2SLS estimator (more precisely they ensure the uniform integrability of the risk estimators). B is almost always satisfied in practice, for instance if x_i is continuous, or if it is discrete but doesn't have zero in its support. An unnatural counterexample would be the case where x is binary with a very large probability to be equal to zero. For k-class estimators a different assumption is made, see **??** in appendix D.4.

Assumption C
For $k \in \{EXO; PMSE; MSE\}$ let $\hat{R}_{k,b}(S)$ be the risk estimator computed for split b. Then n_c and B are such that
(i) $n_c \underset{n\to+\infty}{\to} +\infty$, $n - n_c \underset{n\to+\infty}{\to} +\infty$, and $B \underset{n\to+\infty}{\to} +\infty$
(ii) There exists some $c \in (0; 1)$ such that for $k \in \{EXO; PMSE; MSE\}$, for any $S \in \mathcal{S}$, for any $b = 1, \ldots, B$

$$\sum_{b'=1}^{B} Cov(\hat{R}_{k,b}(S), \hat{R}_{k,b'}(S)) \leqslant \sum_{n_t=0}^{n_c} Var(\hat{R}_{k,b}(S))c^{n_c-n_t}$$

Assumption C characterizes the sampling process used to obtain the risk estimators. This assumption is specific to the linear IV context with potentially endogenous and potentially weak IVs. As such from C(i) the training sample size and validation sample size need to increase with n, this implies that leave-one-out cross-validation cannot be used for risk estimation. This is necessary in order to estimate the out-of-sample correlation between the IVs and the error. C(ii) is a sufficient condition to ensure convergence of the average risk and can be ignored when all IVs are strong. Intuitively, it forces the correlation between the estimated risk across splits to be proportional to the number of common observations across splits. Consequently even when 2SLS is random at the limit because the IVs are weak, the estimators across splits are not too correlated. A simple way to satisfy C(ii) which is quite common in the machine learning literature is to randomly split the data into B folds of size n/B and split those B folds again into training and validation samples. This way the risk estimators across folds are effectively independent. Note that C(ii) is always satisfied in practice from simulation evidence (the constant c can be close to 1).

Theorem 4.1 states that under the above assumptions IV subset selection via cross-validation is efficient, its proof is in appendix C.

Theorem 4.1
Under assumptions A, B, and C, for $k \in \{EXO, PMSE, MSE\}$

$$\frac{\min\limits_{S \in \mathcal{S}} \hat{R}_k(S)}{\min\limits_{S \in \mathcal{S}} R_k(S)} \xrightarrow{\mathbb{P}} 1$$

An almost sure version of this result cannot be obtained unless one assumes that all IVs are strong ($a_S < 1/2$). This is due to the fact the 2SLS estimator is random at the limit when IVs are weak ($a_S \geqslant 1/2$). As mentioned efficiency in this context does not guarantee the selection of good IV subsets.

Consistency can be established with another assumption.

Assumption D
(i) For any $S \in \mathcal{S}$ *at least one of the following conditions hold*

- $k \in \{EXO, PMSE\}$

- $sign(\rho) = sign(\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha)$
- $a_S \neq b_S$
- $\frac{\sigma_v^2}{2\rho}\kappa'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha > \kappa'_S\Sigma_S\kappa_S$

(ii) Let K_w denote the dimension of the largest subset $S \in \mathcal{S}$ such that $a_S \geqslant 1/2$ then at least one of the following conditions hold

- $K_w \leqslant 2$
- $\sigma_u^2\sigma_v^2 > \rho^2 \ max\{\frac{K_w}{2}; \frac{K_w(K_w-1)}{2}\}$

Assumption D(i) ensures that R_{MSE} will correctly rank subset with varying levels of endogeneity. As mentioned in section 4.1, R_{MSE} decomposes into a first term $\mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right)$ and a term which is quadratic in $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$. Under some very specific conditions the dependence of the first term on $\hat{\beta}_S - \beta$ may lead R_{MSE} to miss-classify some IV subsets. Note that in practice when D(i) is not satisfied, it seems that the selection procedure of Andrews (1999) also has its performances halved. D(ii) ensures that when 2SLS is used the three risks will correctly rank strong subsets below weak subsets. Indeed 2SLS is close to OLS in terms of behavior when IVs are weak and numerous, in fact it can be shown that on average 2SLS behaves like OLS in such conditions therefore on average it can be very efficient. Consequently D(ii) is unnecessary when using other k-class estimators due to their lower bias.

Theorem 4.2 establishes consistency under the above assumptions, its proof is in appendix C.

Theorem 4.2
Under assumptions A, B, C for $k \in \{EXO, PMSE, MSE\}$

- *If $\mathcal{S}_c \neq \emptyset$ and assumption D hold then*

$$\mathbb{P}(\underset{S\in\mathcal{S}}{Argmin}\ \hat{R}_k(S) \in \mathcal{S}_c) \to 1$$

- *If $\mathcal{S}_{an} \neq \emptyset$ and assumption D hold then*

$$\mathbb{P}(\underset{S\in\mathcal{S}}{Argmin}\ \hat{R}_k(S) \in \mathcal{S}_{an}) \to 1$$

- *If* $\mathcal{S}_r \neq \emptyset$ *and assumption D(i) holds then*

$$\mathbb{P}(\underset{S \in \mathcal{S}}{Argmin}\ \hat{R}_k(S) \in \mathcal{S}_r) \to 1$$

Intuitively if there exists some IV subset such that the estimator is consistent or asymptotically normal then one such subset will be picked with probability one at the limit. Similarly if there exists some IV subset such that valid weak identification robust inference can be performed then one of such subset will be picked with probability at the limit. This result has some caveats however. If the mean square error of prediction R_{MSE} is used to picked the IV subset then it may not detect endogeneity in certain situations. If the 2SLS estimator is used it may lead the risks to not detect weak IVs if they are too numerous.

It is very difficult to establish more precise results regarding the exact identity of the IV subsets being picked. Such results could be established by either assuming normality of the data and that IVs are non-random, which are very unrealistic assumptions, or assuming that all IVs are strong and use Nagar expansions as in Nagar (1959), which are known to be unreliable. In the next section I assess the empirical performances of the IV subset selection procedures.

5 Simulations

In this section I perform an extensive simulation exercise in order to assess the behavior of different IV selection procedures in different settings: The general case where IVs can be endogenous and / or weak; The case where IVs are strong but can be endogenous IVs are exogenous but some may be weak; The case where IVs are exogenous but can be weak.

Performance Measures I evaluate the performances of the selection methods through the performances of the 2SLS estimator post selection and a weak identification robust confidence interval based on the conditional likelihood ratio test with normal approximation of Mikusheva (2010) post selection. Thus I consider the following metrics over $20,000$ simulations: The interquartile range of the estimators across simulations; The median absolute bias; The median squared bias; The empirical coverage of β using

normal asymptotics for inference with nominal coverage 95%; The median length of the confidence interval (CI) using normal asymptotics; The average number of IVs being picked; The empirical coverage of β using the weak identification robust confidence interval (RCI); The median length of the the weak identification RCI; And the percentage of times the weak identification robust confidence interval exists and is finite.

Selection Methods The risk estimators from section 3.1 are computed using the cross-validation risk estimators with $B = 40$ resamples and $n_c = n/2$. Furthermore to make comparisons, I compute the same post selection metrics using the following selection methods: The "mean square error" criterion of Donald and Newey (2001) using the jacknife and the 2SLS using all the IVs as a first stage; The GMM-BIC procedure of Andrews (1999); The post-lasso of Kang et al. (2016) with the penalty obtained by cross-validation; The post-adaptive-lasso of Windmeijer et al. (2018) with the penalty obtained by cross-validation and using the median estimator as a first stage; The oracle which only uses strong and exogenous IVs. To extend on the introduction, Donald and Newey (2001) should fail in case some IVs are endogenous, Andrews (1999) should fail in case some IVs are weak, Kang et al. (2016) should fail unless the majority of IVs are exogenous and all IVs are strong, and Windmeijer et al. (2018) should fail unless the "largest group" of IVs is exogenous and all IVs are strong. Note that the Lasso and post-lasso are much better suited to a setting with a larger numbers of IVs unlike the methods developed in this paper.

Data Generating Process The data generating process I consider through the simulations is of the following form

$$\begin{aligned} y_i &= 2x_i + \alpha_2 z_{i2} + \alpha_4 z_{i4} + \alpha_6 z_{i6} + u_i \\ x_i &= 0.5\left(z_{i1} + z_{i2} + \frac{c_1}{\sqrt{n}} z_{i3} + \frac{c_1}{\sqrt{n}} z_{i4} + \frac{c_2}{n} z_{i5} + \frac{c_2}{n} z_{i6}\right) + v_i \end{aligned} \tag{5.9}$$

where $(u_i, v_i, z_i')_{i=1}^n$ is iid, normally distributed with mean 0, individual variance 1, correlation between IVs is equal to 0.1, and correlation between u_i and v_i is equal to 0.5. Thus there are six IVs, with three which potentially do not satisfy the exclusion restriction. The parameters which determine the different settings are $(c_1, c_2) \in \mathbb{R}^2$ and $(\alpha_2, \alpha_4, \alpha_6) \in \mathbb{R}^3$. I also allow the sample size n to be equal to either 400 or 4000.

Note that using this specification some IV subsets end up at the exact cutoff levels in terms of endogeneity and strength which determine the asymptotics of the risks, and some IVs have exactly the same level of strength, therefore it constitutes a worst-case scenario for the selection procedures. Moreover this specification also resembles the simulation designs in Belloni et al. (2012). In addition this specification is of specific interest because it allows to control for the range of IVs strength and for the level of bias in terms of pseudo-true-value for all the possible sets of IVs. In each setting I mention the bias of the OLS estimator, the range of the concentration parameters divided by the number of IVs and the range of the (pseudo-true-value) bias of the 2SLS estimator over all the possible subsets S in $\mathcal{S}$. The bias of the OLS estimator can be written as

$$Bias(OLS) = \left| \frac{\rho + \pi' \mathbb{E}(z_i z'_{i\overline{E}})\alpha}{\mathbb{E}(x_i^2)} \right|$$

The pseudo-true-value bias due to using subset S as IVs is

$$Bias(Pseudo_S) = \left| \frac{\pi'_S \mathbb{E}(z_{iS} z'_{i\overline{E}})\alpha}{\pi'_S \mathbb{E}(z_{iS} z'_{iS})\pi_S} \right|$$

The concentration parameter of set S is a measure of strength of the IVs S which is defined as

$$\mu_S^2 = \frac{n\pi'_S \mathbb{E}(z_{iS} z'_{iS})\pi_S}{s\sigma_v^2}$$

It is typically considered low and IVs are typically considered jointly weak when it is inferior to 20, see Stock et al. (2002).

5.1 General Case

The general setting I consider is such that $c_1 = 1$, $c_2 = 1$, and $\alpha_2 = \alpha_4 = \alpha_6 = 1$. Thus model 5.9 can be rewritten as

$$y_i = 2x_i + z_{i2} + z_{i4} + z_{i6} + u_i, \qquad x_i = 0.5\left(z_{i1} + z_{i2} + \frac{1}{\sqrt{n}} z_{i3} + \frac{1}{\sqrt{n}} z_{i4} + \frac{1}{n} z_{i5} + \frac{1}{n} z_{i6}\right) + v_i$$

In addition $Bias(OLS) = 0.83$, for any $S \in \mathcal{S}$ $Bias(Pseudo_S) \in [0; 800]$ and $\mu_S^2 \in [0; 100]$ for $n = 400$ whereas $Bias(Pseudo_S) \in [0; 8000]$ and $\mu_S^2 \in [0; 1000]$ for $n = 4000$. Consequently there are some very weak and endogenous IV subsets and some strong and exogenous IV subsets. The IV subsets which will yield an asymptotically normal estimator are therefore composed of z_{i1} and possibly of z_{i3} and z_{i5}

$$\mathcal{S}_{an} = \{z_{i1}; \{z_{i1}; z_{i3}\}; \{z_{i1}; z_{i5}\}; \{z_{i1}; z_{i3}; z_{i5}\}\}$$

The oracle only uses z_{i1} for instrumentation.

Simulations results for 2SLS in this general setting are in table 1 of appendix A.2. The criterion of Donald and Newey (2001) balances bias and variance assuming exogeneity of all IVs and therefore picks a subset of strong IVs regardless of their level of endogeneity. Consequently the estimator has high bias, very low coverage, but is efficient. Andrews (1999) fails to pick any decent subset of IVs because it requires all of them to be strong, and for there to be no corner solutions. Its estimates post-selection are extremely biased and are very different across simulations, this indicates that weak IVs were picked. Lastly the post-lasso of Kang et al. (2016) and the post-adaptive-lasso of Windmeijer et al. (2018) have high dispersion, which means that for each simulation very different IV subsets are picked. As a consequence the performance of the post-lasso and post-adaptive-lasso are not great, there is bias and coverage is low. This is due to the fact that IVs are allowed to be weak and that the exogeneity conditions required for these procedures to work do not hold. The three risks R_{EXO}, R_{PMSE}, and R_{MSE} are performing as well as the oracle in large sample in terms of bias, coverage and length of Gaussian confidence interval, and dispersion. Note that R_{PMSE} picks more IVs than the other two methods. In addition in smaller samples, some coverage seems to be lost especially when R_{EXO} is used for selection.

Simulations diagnostics for the RCI using the CLR test are in the first three columns of table 5 in appendix A.2. As with the performances of the post selection 2SLS estimator, the RCI using R_{EXO}, R_{PMSE}, and R_{MSE} for selection performs at oracle level in terms coverage, interval lengths, and percentage of finite confidence interval. In small sample this is less the case, using R_{MSE} the RCI undercovers and using R_{EXO} the RCI can be of infinite length. Other selection methods perform worse in terms of coverage and interval length.

5.2 Strong IV Case

I consider two strong IV specifications. In the first strong IV setting I let $c_1 = \sqrt{n}$, $c_2 = n$, $\alpha_2 = \alpha_4 = \alpha_6 = 1$ and model 5.9 can be rewritten as

$$y_i = 2x_i + z_{i2} + z_{i4} + z_{i6} + u_i, \qquad x_i = 0.5 \sum_{k=1}^{6} z_{ik} + v_i$$

Thus $Bias(OLS) = 0.84$, for any $S \in \mathcal{S}$ $Bias(Pseudo_S) \in [0; 2]$ and for $n = 400$ $\mu_S^2 \in [100; 150]$ whereas for $n = 4000$ $\mu_S^2 \in [1000; 1500]$. Hence all IV subsets are

strong, some are very endogenous however, and this time $\mathcal{S}_{an}$ is composed of any combination of z_{i1}, z_{i3} and z_{i5}. The oracle uses (z_{i1}, z_{i3}, z_{i5}).

Simulations results for 2SLS in this strong IV setting are in table 2 of appendix A.2. Again Donald and Newey (2001) selects strong IVs thus it picks all six IVs because all of them are strong leading to an efficient but very biased estimator and to a confidence interval with coverage 0. Andrews (1999) J statistic criterion fails to pick the exogenous IVs, thus the 2SLS estimator post-selection is still highly biased, and coverage of the Gaussian confidence interval is low. This is most likely due to the fact that in this setting the J test statistic is low even when IVs are endogenous, indeed in many settings the J test has low power under the alternative, see Kiviet and Kripfganz (2021) for a recent review and for formal conditions under which the J test has no power. As before the post-lasso and post-adaptive Lasso perform badly, this is due to the fact that the IVs do not satisfy the right exogeneity conditions. Thus only R_{EXO}, R_{PMSE} and R_{MSE} are selecting strong IV subsets and have performances comparable to the oracle with a large sample, and slightly lower performances in small sample.

The second strong IV setting which I call favorable setting is one where a strict majority of IVs is exogenous and where the levels of endogeneity of the endogenous IVs are different from each other and large. In such setting the Sargan Hansen J statistic will be large if endogenous subsets are picked, the median IV estimator which is used to set-up the weights in the adaptive Lasso procedure is consistent, and there can be no confusion between weak and exogenous IV subsets and weak and endogenous IV subsets in the Lasso procedure. Let $c_1 = \sqrt{n}$, $c_2 = n$, $\alpha_2 = 0$, $\alpha_4 = 1$, and $\alpha_6 = 3$ and 5.9 can be rewritten as

$$y_i = 2x_i + z_{i4} + 4z_{i6} + u_i, \qquad x_i = 0.5 \sum_{k=1}^{6} z_{ik} + v_i$$

Then $Bias(OLS) = 1.07$, for any $S \in \mathcal{S}$ $Bias(Pseudo_S) \in [0; 6]$, for $n = 400$ $\mu_S^2 \in [100; 150]$ and for $n = 4000$ $\mu_S^2 \in [1000; 1500]$. The oracle uses $(z_{i1}, z_{i2}, z_{i3}, z_{i5})$.

Simulations results for 2SLS in this strong and favorable setting are in table 3 of appendix A.2. All the selection procedures are performing well except Donald and Newey (2001) which does not recognize endogenous IVs. Note that the 2SLS estimators obtained after selection via Lasso and adaptive Lasso are slightly more dispersed across simulations, have higher bias and their Gaussian confidence intervals have lower coverage compared to estimators after selection using Andrews (1999), R_{EXO}, R_{PMSE}

or R_{MSE}. This is not surprising, Lasso and adaptive Lasso can be useful when the total number of IVs is large and when sample size is large. When the total number of IVs is low there is few reason to use Lasso, just like there is no reason to use Lasso for control variable selection in linear models when there aren't many control variables in the first place.

In these two strong IV settings, simulations diagnostics for the RCI using the CLR test are in the fourth to ninth columns of table 5 in appendix A.2. Once again because Donald and Newey (2001) picks endogenous IVs, using it to select IVs yields an RCI with very low coverage in both strong IV settings. Andrews (1999) doesn't pick exogenous IVs in the first setting whereas it does in the second, this is reflected in the performances of its RCI which is at oracle level in the second setting but not in the first. The RCI after Lasso or adaptive-lasso selection perform very badly in the first setting for the same reasons post-lasso and post-adaptive-lasso perform badly, in the setting they perform very well but not at oracle level. Lastly the RCI using R_{EXO}, R_{PMSE}, and R_{MSE} for selection have oracle level performances in both settings.

5.3 Exogenous IV Case

Lastly I consider an exogenous IVs setting such that $c_1 = c_2 = 1$ and $\alpha_2 = \alpha_4 = \alpha_6 = 0$

$$y_i = 2x_i + u_i, \qquad x_i = 0.5\left(z_{i1} + z_{i2} + \frac{1}{\sqrt{n}}z_{i3} + \frac{1}{\sqrt{n}}z_{i4} + \frac{1}{n}z_{i5} + \frac{1}{n}z_{i6}\right) + v_i$$

Then $Bias(OLS) = 0.32$, whereas $Bias(Pseudo_S) = 0$ for any $S \in \mathcal{S}$, $\mu_S^2 \in [0; 100]$ for $n = 400$ and $\mu_S^2 \in [0; 1000]$ for $n = 4000$. Therefore there are very weak sets of IVs and strong sets of IVs and as long as a set of strong IVs is picked the estimator should estimate the true causal parameter of interest. The oracle use (z_{i1}, z_{i2}) for instrumentation.

Simulation results for 2SLS in this exogenous setting are in table 4 of appendix A.2. In this case, there are very little differences in terms of estimator performances between each selection method, it seems all of them are picking strong IVs and some of them also pick weaker IVs. Post-lasso seem to select too few IVs which is why it seems less efficient. On another note, in principle Andrews (1999) picks the largest IV subset which is exogenous, this explains why it picks all six IVs. As for simulations diagnostics for the RCI using the CLR test they are in the tenth to twelfth columns of table 5

in appendix A.2. All methods yields a post selection RCI with oracle performances except, as argued previously, Lasso which select too few IVs and thus has an RCI with a slightly greater length.

In the next section I estimate the effect of pre-trial detention on guilt and I select judge dummies which act as instrumental variables with the methods designed in this paper.

6 Application

Since Kling (2006) a large literature in Economics and Law which utilizes the random assignment of judges to cases and differences in the degree of severity of judges has developed in order to estimate the causal effect of prison on offenders' outcomes. It is now well-established that, controlling for the offender's characteristics, for the case's characteristics, for other time and place variables, judges differ significantly in their propensity to send offenders to prison, including their propensity to send offenders to pre-trial detention. Thus in practice judge dummies generate supposedly exogenous variation in detention / pre-trial detention which allow to identify and estimate causal effects, most often the JIVE estimator is used instead of 2SLS leading to the famous judge leniency IV or jacknifed judge IV. This identification strategy can fail for multiple reasons however.

First, judges may not differ significantly in their leniency, which can generate a weak IVs' problem. Second, the identity of the judge assigned to a case (and his level of leniency) is known by the offender and his lawyer before the trial is held, this can lead to voluntary postponement of the trial in order to get a more lenient judge, bribing of the judge, differing defenses during the trial depending on the judges' leniency, plea deals before the actual trial, etc... Furthermore when evaluating the effect of pre-trial detention on detention, the identity of the judge present during the pre-trial hearing is also known by judge present for the actual trial. Consequently the judge present during the trial has the possibility of doubling-down on the signal sent by the judge during the pre-trial hearing if they deem them trustworthy, or on the contrary compensate for the pre-trial judgement if they deem the judge present during the pre-trial hearing untrustworthy. Finally judges are never completely randomly assigned to a case, at best a judge among a subset of available judges is randomly assigned to a case. As

a consequence the identity of the judge can directly affect the offender outcomes or indirectly through unobserved cofounders such as offender income, level of education, lawyer quality, defense strategy, psychological state.

For the aforementioned reasons selecting the judges which differ most in their leniency and which satisfy best the exclusion restriction is a priority. In this specific application I use data on 331,971 court cases in Philadelphia and Miami from September 2006 until February 2013 and study the effect of pre-trial detention on the likelihood of being found guilty using 8 judge dummy variables as IVs. Among other control variables the data includes the case characteristics, the offender criminal history and some of their characteristics such as their race, the date and time of the day when the pre-trial hearing is held. Most information about the actual trial are unknown, including the identity of the judge present during the trial. To be more specific, after their arrest an offender is assigned a judge who will preside over a pre-trial hearing, in Philadelphia and Miami pre-trial hearings happen at most a few days after initial arrest. During this hearing the judge decides whether they offer the offender a plea deal, whether they offer the offender a bail deal, or whether they directly send the offender to prison before their trial. Note that because the data is from the US failure of the exclusion restriction due to bribing is unlikely. Heterogenous effects are also unlikely because pre-trial detention sends the same signal to the judge which presides over the actual trial. Finally from table 7 in appendix A.3 the observable covariates seem relatively balanced across judges. See Stevenson (2018) for more details on the data and the set-up.

To be more specific the oracle model to estimate is

$$guilt_i = predet_i\beta + X_i'\delta + \sum_{j\in E} \alpha_j 1\{judge_i = j\} + u_i$$

$$predet_i = \sum_{j=1}^{8} \pi_j 1\{judge_i = j\} + X_i'\gamma + v_i$$

where $guilt_i$ is a dummy variable which equals 1 if the offender was found guilty and equals 0 otherwise, $predet_i$ is a dummy variable which equals 1 if the offender was kept in prison before his trial and equals 0 otherwise, for $j = 1, \ldots, 8$ $1\{judge_i = j\}$ is a dummy variable which equals 1 if judge j is the judge who oversees the pre-trial hearing of individual i and equals 0 otherwise, X_i is a vector of control variables which include an intercept, and there exists some j such that $\alpha_j = 0$. Of course a priori it is

unknown which judge dummy enters the structural equation, ie $\overline{E}$ is unknown, which is why the judge dummy IVs must be selected.

In table 6 in appendix A.3 are pre-trial judges' descriptive statistics including what percentage of cases they oversaw and what percentage of offenders were sent to prison before their trial unconditionally and conditionally on different control dummy variables. All 8 judges have supervised a high number of cases but some more than others, across all judges offenders are sent to prison before their trial between 39% and 44% of the time. When conditioning on offender characteristics and criminal record such as race, gender or the number of prior offense, the differences in judge propensities to send offenders to prison before the trial are maintained. On the other hand when conditioning on the case characteristics such as the time when the offender was arrested or the reason for their arrest, differences in judge propensities for pre-trial detention become very different. This clearly indicates that judges greatly differ in their leniency, that offenders demographics enter linearly in the first stage as fixed effects but that case characteristics do not.

In table 8 in appendix A.3 are the first stage estimates, heteroskedasticity robust standard errors, and relevance p-values, of the judge dummy variables on the endogenous dummy variable pre-trial detention for three different specifications: (1) only includes time and date fixed effects, (2) also includes the case characteristics, and (3) also includes the offender characteristics and their criminal history. First stage F-statistics are also reported, and so are Sargan-Hansen J statistics after using the 2SLS estimator with the full set of IVs. Note that because an intercept is included a judge dummy variable must be excluded to prevent multicolinearity, I excluded judge 2. As a consequence the judge fixed effects on pre-trial detention are all relative to judge 2. This choice is made mainly because judge 2 oversees the lowest amount of cases which limits potential endogeneity bias, and because judge 2 has the highest probability to send anyone to pre-trial detention which increases the amount of strong IV subsets. A thorough explanation and analysis of the choice of excluded judge and its impact along with a guess on which IV is likely to be endogenous are in appendix B. After excluding judge 2, for any specification the effect of judge 6 is insignificant, ie judge 6 is as harsh as judge 2. On the other hand judge 1 and judge 5 seem to have the same moderate effect on the likelihood of the offender going to pre-trial detention. Judge 4 and judge 8 are significantly are the most lenient judges. Taken jointly the IVs are not

weak but are not very strong either given that the first stage F statistic is close to 40 across all specifications. Consequently small size and power distortions of tests are to be expected. Regarding the Sargan-Hansen J statistics, they cannot tell us whether or not the IVs are exogenous. As mentioned, the Sargan-Hansen test is known to have poor power properties, see Kiviet and Kripfganz (2021), and indeed specification (1) with the fewest controls is the least likely to satisfy the exclusion restriction and yet it has the lowest J statistic compared to the J statistics in specification (2) and (3) which both reject exogeneity at level 10%.

Next in table 9 of appendix A.3 I report the OLS estimator, the 2SLS estimator using all the IVs, the 2SLS estimator after selection via Donald and Newey (2001) (DN), via Andrews (1999) (AN), via Kang et al. (2016) (Post-lasso), via Windmeijer et al. (2018) (Post-adalasso), and my methods along with their heteroskedasticity-robust standard errors. I also report the set of judges selected for instrumentation IVs by each method. Thus the judges which are not included as IVs were used as control variables except for judge 2 which was excluded. In addition in table 10 are the first stage F statistics and the J statistics computed post selection for each method.

From table 9 for the three specifications the OLS estimate of the effect of pre-trial detention on guilt is close to zero. This is in accordance with the literature, it is believed that OLS underestimates the effect effect of pre-trial detention on the likelihood of being found guilty due to omitted variables. Indeed variables such as the offender level of education or the offender income are unobserved and negatively correlated with pre-trial detention. On the other hand the 2SLS estimates using all the judge dummies as IVs lie between 15% and 18.5% in the three specifications which is slightly lower than what is expected in the literature, see Kling (2006) or Dobbie, Goldin, and Yang (2018). Regarding the selection procedures, quite interestingly the AN and DN use all the judges as IVs in the three specifications. A possible explanation is that all IVs are exogenous and strong but based on the J test statistics in table 10 the full set of IVs is unlikely to be exogenous. Thus the most plausible explanation is that the conditions for these two procedures to work are not met. Next, note that the Lasso uses judge 4 as the sole IV and the 2SLS estimator is approximately equal to 0.25 in all specifications. Judge 4 is both the most lenient judge and is the judge with the highest number of cases so this IV could be endogenous. This cannot be confirmed by the implied J statistic in table 10 because they are equal to zero by construction. Conceptually the Lasso will work if a majority of the IVs are exogenous.

Thus it is not clear why it picks the same judge dummy in (1) and (3) even tough the judge dummies in (1) are most likely endogenous. The selection via adaptive Lasso is difficult to interpret, it uses judge 4 and 6 in specification (1), all judges except judge 1 as IVs in specification (2), and judges 6 in specification (3). The post adaptive Lasso 2SLS estimates vary between 0.20 and 0.52 and are significant except in setting (3), and the J statistics are small. In addition the selection procedure is quite sensitive to the choice to the penalty choice, different non-nested IV sets can be picked depending on the penalty value. A plausible explanation is that there is no "largest group" of exogenous IVs. In addition the adaptive Lasso can fail when some IVs are weak, which is the case for judge 6 which is selected by the adaptive Lasso in (2) and (3).

Regarding the procedures developed in this paper R_{EXO} selects judges 6 and 7 in specifications (1) and (2), and judges 4 and 8 in specification (3). This choice in specifications (1) and (2) seem to be linked to the fact that there is not enough control variables and therefore the true effect of pre-trial detention on guilt cannot be properly captured. This would also explain why the J statistics are so low, the procedure may have focused on picking the most exogenous IV subsets possible. In (3) it selects a completely different set of judges, the effect is significant and close to 0.26, and the J statistic is very small. R_{PMSE} selects judge 4 in (1) and (3) and selects judges 4 and 8 in specification (2). Similar IV subsets are selected across specifications and are very close to that of Lasso. The estimates post-selection with R_{PMSE} are close to 0.25 and are significant in all specifications. Finally R_{MSE} selects judges 1, 3, 7 and 8 in specification (1) and (2), and judges 1 and 6 in specification (3). The corresponding estimators are not statistically significant, indeed the corresponding first F statistics are not large (< 30). In addition the estimator is negative in specification (3). This is most likely due to the fact that the sign conditions for R_{MSE} to rank properly IV subsets are not satisfied. Indeed the OLS estimator is smaller than the 2SLS estimator thus the OLS bias is negative, ie $\rho < 0$. Additionally as mentioned some judges which supervise pre-trials may be positive signal for the judge during the actual trial. Hence judge dummies which enter the structural equation may have a positive effect on the likelihood of being found guilty, ie $\mathbb{E}(z_{iS}z_{i\bar{E}})\alpha > 0$. This would violate assumption D(i).

The estimates post-selection with R_{EXO} and R_{PMSE} in specification (3) appear quite trustworthy compared to other procedures, especially in light of the simulation exercise. This means that the effect of pre-trial detention on the likelihood of being

found guilty is actually equal to 25%. This is more in line with the literature compared to a 2SLS estimator equal to 18% found using all the judge dummies.

7 Concluding Remarks

In this paper I formally define and study losses, risks, and risk estimators for the selection of strong and exogenous subsets of IVs in the linear IV model for the 2SLS estimator. To do so, I utilize the losses implicitly minimized during IV estimation and obtain three risk: one based on the exogeneity condition, one based on the mean square error of prediction after projecting the endogenous variable on the IVs, and the mean square error of prediction. These risks do not balance squared bias and variance, instead they trade in priority endogenous IVs for exogenous ones. I show that choosing the IV subsets which minimize these risks is a consistent procedure to obtain an estimator which converges towards the true structural parameter of interest and is asymptotically normal. This implies that in practice, applied researchers can use this IV selection method to easily strengthen the credibility of their results. If they have multiple IVs at their disposal they have theoretical guarantees that they will select a strong and exogenous subset. These results are corroborated by the simulation exercise and the application.

From a broader perspective, this paper resembles earlier works on the selection of regressors via risk minimization when the number of regressors is finite. It is no surprise then that, given a fixed and small total number of IVs, the risks developed in this paper have better chances to select exogenous IV subsets compared to the Lasso and adaptive Lasso which are best suited to a setting with many IVs. A next natural step in the investigation of risks in IV models is understanding when the procedures developed in this paper might fail. Introducing heteroskedasticity, non-linearity, or heterogeneity in the structural equation may affect the selection methods in which case they would require correction. In fact it may be possible to derive new predictions losses and risks for non-linear IV models. These risks could then be used to select the tuning parameters of regularized two-step IV estimators. For instance the penalty terms of the Lasso and adaptive Lasso could be chosen using risks similar to the ones coined in this paper.

Bibliography

Anderson, T. W. and H. Rubin (1949): "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *The Annals of Mathematical Statistics*, 20, 46–63.

Andrews, D. W. K. (1999): "Consistent Moment Selection Procedures for Generalized Method of Moments Estimation," *Econometrica*, 67, 543–564.

Andrews, D. W. K. and X. Cheng (2012): "Estimation and Inference With Weak, Semi-Strong, and Strong Identification," *Econometrica*, 80, 2153–2211.

Andrews, D. W. K., V. Marmer, and Z. Yu (2019): "On optimal inference in the linear IV model," *Quantitative Economics*, 10, 457–485.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quaterly Journal of Economics*, 132, 1553–1992.

Antoine, B. and P. Lavergne (2022): "Identification-robust nonparametric inference in a linear IV model," *Journal of Econometrics*.

Arlot, S. and A. Celisse (2010): "A survey of cross-validation procedures for model selection," *Statistics Surveys*, 4.

Bai, J. and S. Ng (2010): "Instrument Variable Estimation in a Data Rich Environment," *Econometric Theory*, 26, 1577–1606.

Bates, S., T. Hastie, and R. Tibshirani (2021): "Cross-validation: what does it estimate and how well does it do it?" .

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments With an Application to Eminent Domain," *Econometrica*, 80, 2369–2429.

Caner, M. (2009): "Testing, Estimation in GMM and CUE with Nearly-Weak Identification," *Econometric Reviews*, 29, 330–363.

Carrasco, M. (2012): "A regularization approach to the many instruments problem," *Journal of Econometrics*, 170, 383–398.

Carrasco, M. and M. Doukali (2021): "Testing overidentifying restrictions with many instruments and heteroscedasticity using regularised jackknife IV," *The Econometrics Journal*, 25, 71–97.

Carrasco, M. and G. Tchuente (2016): "Efficient Estimation with Many Weak Instruments Using Regularization Techniques," *Econometric Reviews*, 35, 1609–1637.

Chaudhuri, S. and E. Zivot (2011): "A new method of projection-based inference in GMM with weakly identified nuisance parameters," *Journal of Econometrics*, 164, 239–251.

Chen, J., D. L. Chen, and G. Lewis (2021): "Mostly Harmless Machine Learning: Learning Optimal Instruments in Linear IV Models," .

Chernozhukov, V., C. Hansen, and M. Spindler (2015): "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments," *American Economic Review*, 105, 486–490.

Dobbie, W., J. Goldin, and C. Yang (2018): "The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges," *American Economic Review*, 108, 201–40.

Donald, S. G. and W. K. Newey (2001): "Choosing the Number of Instruments," *Econometrica*, 69, 1161–1191.

Gautier, E. and C. Rose (2021): "High-dimensional instrumental variables regression and confidence sets," .

Graddy, K. (2006): "Markets: The Fulton Fish Market," *Journal of Economic Perspectives*, 20, 207–220.

Guggenberger, P., F. Kleibergen, S. Mavroeidis, and L. Chen (2012): "On the Asymptotic Sizes of Subset Anderson-Rubin and Lagrange Multiplier Tests in Linear Instrumental Variables Regression," *Econometrica*, 80, 2649–2666.

Hahn, J. and J. Hausman (2002): "A New Specification Test for the Validity of Instrumental Variables," *Econometrica*, 70, 163–189.

——— (2003): "Weak Instruments: Diagnosis and Cures in Empirical Econometrics," *American Economic Review*, 93, 118–125.

Hahn, J., J. Hausman, and G. Kuersteiner (2004): "Estimation with weak instruments: Accuracy of higher-order bias and MSE approximations," *The Econometrics Journal*, 7, 272–306.

Hall, A. R. and F. P. M. Peixe (2003): "A Consistent Method for the Selection of Relevant Instruments," *Econometric Reviews*, 22, 269–287.

Hausman, J. A., W. K. Newey, T. Woutersen, J. C. Chao, and N. R. Swanson (2012): "Instrumental variable estimation with heteroskedasticity and many instruments: Instrumental variable estimation," *Quantitative Economics*, 3, 211–255.

K. Newey, W. (1985): "Generalized method of moments specification testing," *Journal of Econometrics*, 29, 229–256.

Kang, H., A. Zhang, T. T. Cai, and D. S. Small (2016): "Instrumental Variables Estimation With Some Invalid Instruments and its Application to Mendelian Randomization," *Journal of the American Statistical Association*, 111, 132–144.

Kiviet, J. F. and S. Kripfganz (2021): "Instrument approval by the Sargan test and its consequences for coefficient estimation," *Economics Letters*, 205.

Kleibergen, F. (2002): "Pivotal Statistics for Testing Structural Parameters in Instrumental Variables Regression," *Econometrica*, 70, 1781–1803.

——— (2007): "Generalizing weak instrument robust IV statistics towards multiple parameters, unrestricted covariance matrices and identification statistics," *Journal of Econometrics*, 139, 181–216.

Kling, J. R. (2006): "Incarceration Length, Employment, and Earnings," *American Economic Review*, 96, 863–876.

Maasoumi, E. and P. C. Phillips (1982): "On the behavior of inconsistent instrumental variable estimators," *Journal of Econometrics*, 19, 183–201.

Mikusheva, A. (2010): "Robust confidence sets in the presence of weak instruments," *Journal of Econometrics*, 157, 236–247.

Mikusheva, A. and L. Sun (2021): "Inference with Many Weak Instruments," *The Review of Economic Studies*, 89, 2663–2686.

Moreira, M. J. (2003): "A Conditional Likelihood Ratio Test for Structural Models," *Econometrica*, 71, 1027–1048.

Nagar, A. L. (1959): "The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations," *Econometrica*, 27, 575–595.

Olea, J. L. M. and C. Pflueger (2013): "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 31, 358–369.

Sargan, J. D. (1958): "The Estimation of Economic Relationships using Instrumental Variables," *Econometrica*, 26, 393.

Singh, R. and L. Sun (2021): "Automatic Kappa Weighting for Instrumental Variable Models of Complier Treatment Effects," .

Staiger, D. and J. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stevenson, M. T. (2018): "Distortion of Justice: How the Inability to Pay Bail Affects Case Outcomes," *The Journal of Law, Economics, and Organization*, 34, 511–542.

Stock, J. and M. Yogo (2005): *Testing for Weak Instruments in Linear IV Regression*, New York: Cambridge University Press, 80–108.

Stock, J. H., J. H. Wright, and M. Yogo (2002): "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–529.

Windmeijer, F., H. Farbmacher, N. Davies, and G. D. Smith (2018): "On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments," *Journal of the American Statistical Association*, 114, 1339–1350.

Windmeijer, F., X. Liang, F. P. Hartwig, and J. Bowden (2021): "The confidence interval method for selecting valid instrumental variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 752–776.

A Figures and Tables

A.1 Directed Acyclic Graphs

Figure 1: Linear IV model with endogenous IVs DAG, direct effect (top); Linear IV model with endogenous IVs DAG, indirect effect (bottom)

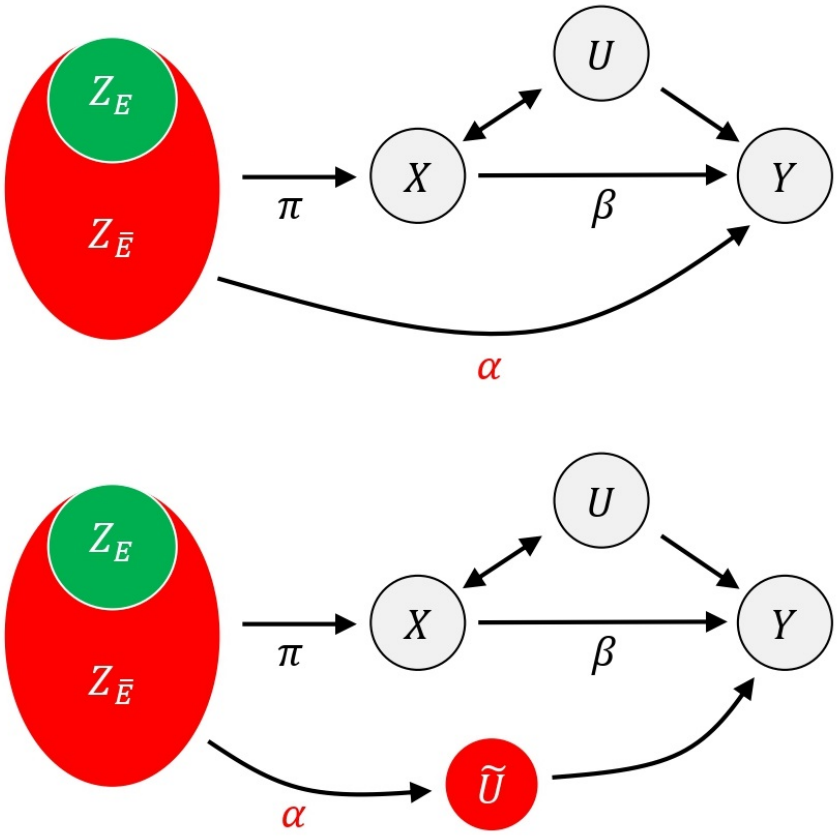
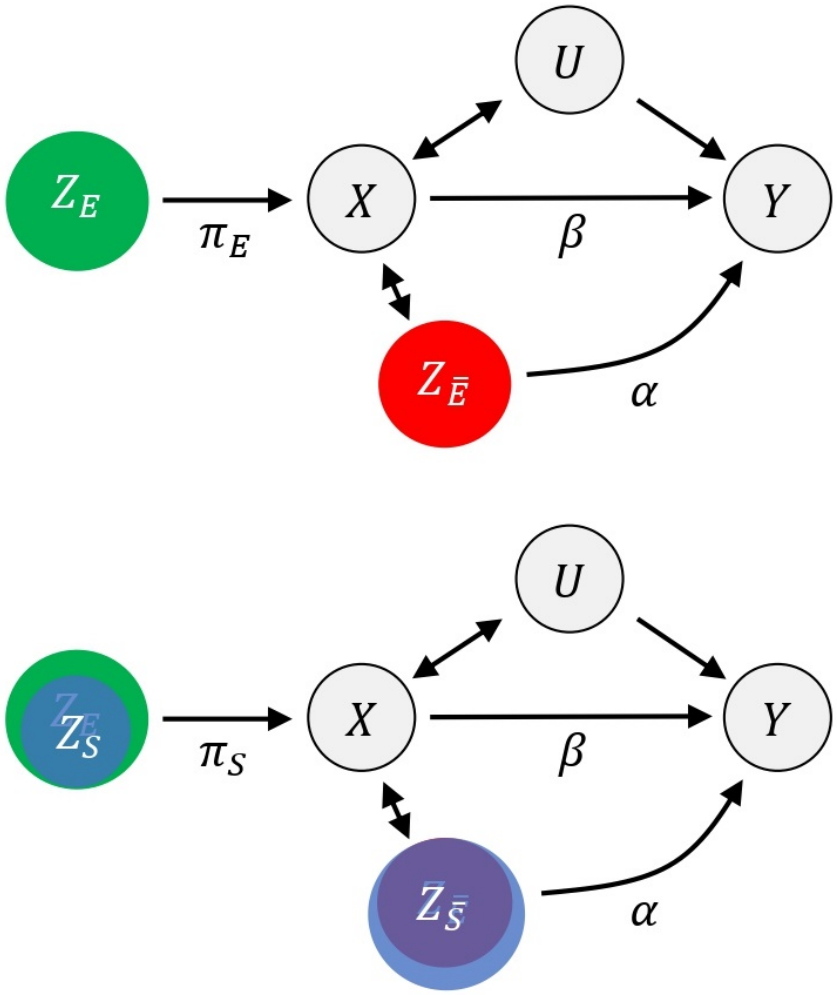


Figure 2: Linear IV model with endogenous IVs DAG, oracle model (top); Linear IV model with endogenous IVs DAG, valid model (bottom)



A.2 Simulation Results

Table 1: 2SLS diagnostics by selection method, general setting

		Diagnostics					
		Intrqr	Med Abs Bias	Med Sq Bias	Emp Cov CI An	Med Len CI An	Avg # of IVs
n=400	DN	0.299	1.087	1.181	0.000	0.392	3.421
	AN	4.434	2.143	4.592	0.337	0.889	3.268
	Post-lasso	1.687	0.730	0.532	0.657	0.996	1.610
	Post-adalasso	1.819	0.726	0.527	0.562	0.705	2.576
	R_{EXO}	0.155	0.079	0.006	0.935	0.400	2.064
	R_{PMSE}	0.144	0.072	0.005	0.928	0.389	2.995
	R_{MSE}	0.172	0.087	0.008	0.816	0.373	2.566
	Oracle	0.144	0.072	0.005	0.954	0.399	1.000
n=4000	DN	0.201	1.112	1.237	0.000	0.133	3.422
	AN	0.130	2.013	4.053	0.278	0.227	2.720
	Post-lasso	1.925	0.929	0.863	0.602	0.241	2.284
	Post-adalasso	1.752	1.203	1.448	0.562	0.246	2.440
	R_{EXO}	0.046	0.023	0.001	0.948	0.126	2.058
	R_{PMSE}	0.043	0.022	0.000	0.951	0.125	2.998
	R_{MSE}	0.044	0.022	0.000	0.947	0.125	2.407
	Oracle	0.044	0.022	0.000	0.948	0.126	1.000

Table 2: 2SLS diagnostics by selection method, strong IVs setting

		Diagnostics					
		Intrqr	Med Abs Bias	Med Sq Bias	Emp Cov CI An	Med Len CI An	Avg # of IVs
n=400	DN	0.076	0.995	0.991	0.000	0.199	6.000
	AN	1.992	1.801	3.243	0.489	0.250	2.997
	Post-lasso	1.654	0.986	0.972	0.348	0.369	2.096
	Post-adalasso	0.133	0.992	0.984	0.172	0.209	4.303
	R_{EXO}	0.082	0.042	0.002	0.908	0.217	2.766
	R_{PMSE}	0.078	0.040	0.002	0.892	0.212	2.990
	R_{MSE}	0.083	0.042	0.002	0.843	0.208	3.127
	Oracle	0.074	0.037	0.001	0.940	0.211	3.000
n=4000	DN	0.025	1.001	1.002	0.000	0.063	6.000
	AN	1.997	1.944	3.778	0.458	0.109	2.999
	Post-lasso	0.041	1.001	1.003	0.151	0.064	4.163
	Post-adalasso	0.025	1.001	1.002	0.000	0.063	6.000
	R_{EXO}	0.026	0.013	0.000	0.919	0.068	2.868
	R_{PMSE}	0.023	0.012	0.000	0.934	0.068	2.998
	R_{MSE}	0.023	0.011	0.000	0.945	0.068	3.000
	Oracle	0.023	0.011	0.000	0.945	0.068	3.000

Table 3: 2SLS diagnostics by selection method, strong IVs favorable setting

		Diagnostics					
		Intrqr	Med Abs Bias	Med Sq Bias	Emp Cov CI An	Med Len CI An	Avg # of IVs
n=400	DN	0.146	1.328	1.764	0.000	0.368	6.000
	AN	0.064	0.032	0.001	0.939	0.176	3.946
	Post-lasso	0.079	0.040	0.002	0.865	0.185	3.539
	Post-adalasso	0.066	0.033	0.001	0.920	0.177	3.898
	R_{EXO}	0.069	0.035	0.001	0.911	0.178	3.771
	R_{PMSE}	0.064	0.032	0.001	0.935	0.176	3.986
	R_{MSE}	0.064	0.032	0.001	0.927	0.175	4.014
	Oracle	0.063	0.032	0.001	0.939	0.176	4.000
n=4000	DN	0.046	1.335	1.783	0.000	0.117	6.000
	AN	0.020	0.010	0.000	0.948	0.056	3.986
	Post-lasso	0.024	0.013	0.000	0.880	0.058	3.463
	Post-adalasso	0.026	0.015	0.000	0.862	0.066	3.078
	R_{EXO}	0.021	0.011	0.000	0.932	0.056	3.909
	R_{PMSE}	0.020	0.010	0.000	0.947	0.056	3.998
	R_{MSE}	0.020	0.010	0.000	0.949	0.056	4.000
	Oracle	0.020	0.010	0.000	0.949	0.056	4.000

Table 4: 2SLS diagnostics by selection method, exogenous IVs setting

		Diagnostics					
		Intrqr	Med Abs Bias	Med Sq Bias	Emp Cov CI An	Med Len CI An	Avg # of IVs
n=400	DN	0.091	0.046	0.002	0.944	0.263	3.417
	AN	0.091	0.046	0.002	0.933	0.258	5.923
	Post-lasso	0.116	0.063	0.004	0.947	0.384	1.230
	Post-adalasso	0.097	0.048	0.002	0.930	0.259	5.760
	R_{EXO}	0.118	0.058	0.003	0.927	0.275	3.099
	R_{PMSE}	0.091	0.045	0.002	0.940	0.257	5.982
	R_{MSE}	0.091	0.045	0.002	0.935	0.257	5.104
	Oracle	0.091	0.046	0.002	0.953	0.271	2.000
n=4000	DN	0.029	0.015	0.000	0.942	0.085	3.421
	AN	0.029	0.015	0.000	0.941	0.083	5.971
	Post-lasso	0.038	0.020	0.000	0.945	0.124	1.577
	Post-adalasso	0.032	0.016	0.000	0.929	0.084	5.460
	R_{EXO}	0.033	0.016	0.000	0.925	0.086	3.269
	R_{PMSE}	0.029	0.015	0.000	0.946	0.083	5.993
	R_{MSE}	0.029	0.015	0.000	0.946	0.083	5.806
	Oracle	0.030	0.015	0.000	0.942	0.086	2.000

Table 5: CLR diagnostics by selection method

		Diagnostics											
		General setting			Strong IV setting			Strong IV favorable setting			Exogenous IV setting		
		Emp Cov CI R	Med Len CI R	% finite CI	Emp Cov CI R	Med Len CI R	% finite CI	Emp Cov CI R	Med Len CI R	% finite CI	Emp Cov CI R	Med Len CI R	% finite CI
n=400	DN	0.00	1.51	1.00	0.00	0.24	1.00	0.00	1.04	1.00	0.95	0.27	1.00
	AN	0.46	0.71	0.56	0.47	0.29	1.00	0.95	0.18	1.00	0.94	0.27	1.00
	Post-lasso	0.65	0.65	0.62	0.34	0.38	1.00	0.90	0.19	1.00	0.96	0.39	0.96
	Post-adalasso	0.58	0.62	0.71	0.16	0.26	1.00	0.94	0.18	1.00	0.94	0.27	1.00
	R_{EXO}	0.94	0.41	0.94	0.93	0.22	1.00	0.93	0.18	1.00	0.93	0.27	0.95
	R_{PMSE}	0.94	0.41	1.00	0.93	0.22	1.00	0.95	0.18	1.00	0.95	0.27	1.00
	R_{MSE}	0.84	0.41	0.98	0.86	0.21	1.00	0.94	0.18	1.00	0.95	0.27	1.00
	Oracle	0.95	0.41	1.00	0.95	0.22	1.00	0.95	0.18	1.00	0.95	0.28	1.00
n=4000	DN	0.00	0.64	1.00	0.00	0.08	1.00	0.00	0.33	1.00	0.95	0.09	1.00
	AN	0.33	0.22	0.67	0.48	0.07	1.00	0.95	0.06	1.00	0.94	0.08	1.00
	Post-lasso	0.65	0.21	0.57	0.15	0.08	1.00	0.90	0.06	1.00	0.95	0.12	0.98
	Post-adalasso	0.61	0.22	0.53	0.00	0.08	1.00	0.88	0.07	1.00	0.93	0.08	1.00
	R_{EXO}	0.95	0.13	0.95	0.93	0.07	1.00	0.93	0.06	1.00	0.94	0.09	0.95
	R_{PMSE}	0.95	0.13	1.00	0.94	0.07	1.00	0.95	0.06	1.00	0.95	0.08	1.00
	R_{MSE}	0.95	0.13	0.99	0.95	0.07	1.00	0.95	0.06	1.00	0.95	0.08	1.00
	Oracle	0.95	0.13	1.00	0.95	0.07	1.00	0.95	0.06	1.00	0.95	0.09	1.00

A.3 Application Figures and Tables

Table 6: Judge descriptive statistics

Descriptive Statistic	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8
% cases allocated	0.065	0.039	0.163	0.170	0.101	0.166	0.125	0.170
% pre-trial detention	0.402	0.432	0.418	0.395	0.413	0.432	0.413	0.398
% pre-trial detention, cond. male	0.423	0.450	0.448	0.422	0.440	0.462	0.437	0.423
% pre-trial detention, cond. black	0.452	0.487	0.467	0.451	0.473	0.487	0.465	0.452
% pre-trial detention, cond. one prior	0.438	0.479	0.466	0.433	0.462	0.475	0.460	0.445
% pre-trial detention, cond. three priors	0.474	0.529	0.511	0.468	0.514	0.523	0.508	0.490
% pre-trial detention, cond. possess	0.213	0.255	0.182	0.160	0.167	0.186	0.197	0.219
% pre-trial detention, cond. agg assault	0.515	0.475	0.524	0.530	0.541	0.577	0.555	0.486
% pre-trial detention, cond. felony	0.568	0.566	0.589	0.583	0.604	0.609	0.588	0.545
% pre-trial detention, cond. misdemeanor	0.389	0.414	0.403	0.385	0.400	0.419	0.402	0.389

Table 7: Balance check: mean variable by judge and by pretrial detention status

	Judge 1	Judge 2	Judge 3	Judge 4	Judge 5	Judge 6	Judge 7	Judge 8	Pretrial detention	No pretrial detention
Prior # cases	4.820	4.660	4.890	4.880	4.990	4.830	4.920	4.910	3.900	6.290
Prior # felony charges	0.999	0.808	1.320	1.320	1.580	1.320	1.480	1.300	0.984	1.810
Prior # guilt	1.440	1.360	1.630	1.640	1.790	1.640	1.740	1.640	1.320	2.110
Felony	0.518	0.522	0.503	0.509	0.512	0.512	0.503	0.503	0.360	0.720
Misdemeanor	0.932	0.932	0.932	0.935	0.933	0.933	0.935	0.935	0.953	0.906
Summary offense	0.049	0.049	0.060	0.059	0.063	0.057	0.065	0.059	0.056	0.064
Felony type 1	0.142	0.149	0.141	0.144	0.137	0.142	0.137	0.138	0.056	0.262
Felony type 2	0.110	0.111	0.116	0.117	0.116	0.114	0.115	0.112	0.052	0.205
Felony type 3	0.223	0.218	0.202	0.204	0.194	0.203	0.195	0.201	0.126	0.312
Other felony	0.133	0.153	0.129	0.135	0.142	0.138	0.131	0.135	0.127	0.147
Misdemeanor type 1	0.365	0.357	0.364	0.364	0.355	0.360	0.360	0.356	0.260	0.504
Misdemeanor type 2	0.374	0.363	0.373	0.368	0.363	0.366	0.363	0.366	0.285	0.484
Misdemeanor type 3	0.078	0.079	0.081	0.078	0.076	0.079	0.085	0.078	0.075	0.085
Other misdemeanors	0.425	0.444	0.409	0.419	0.415	0.419	0.411	0.422	0.512	0.283
Robbery	0.078	0.076	0.070	0.077	0.068	0.076	0.072	0.070	0.021	0.148
Aggravated assault	0.085	0.079	0.092	0.090	0.096	0.091	0.091	0.090	0.072	0.117
Drug possession	0.156	0.133	0.133	0.137	0.118	0.137	0.134	0.136	0.186	0.062
Selling drugs	0.125	0.144	0.121	0.127	0.132	0.128	0.123	0.127	0.124	0.131
1st offense DUI	0.063	0.059	0.067	0.066	0.068	0.063	0.067	0.065	0.099	0.016
Guilt	0.450	0.452	0.491	0.489	0.523	0.495	0.505	0.492	0.492	0.493
Bail date	13,858	13,667	14,542	14,522	15,018	14,497	14,804	14,512	14,514	14,548
White	0.283	0.278	0.289	0.282	0.273	0.285	0.287	0.284	0.300	0.260
Black	0.604	0.586	0.569	0.576	0.568	0.576	0.575	0.577	0.524	0.651
Age	32.1	32.1	32.6	32.4	32.7	32.4	32.6	32.6	32.8	32.0
Male	0.830	0.831	0.829	0.836	0.828	0.837	0.829	0.831	0.795	0.885
One prior	0.759	0.753	0.761	0.761	0.768	0.759	0.764	0.762	0.704	0.844
Three priors	0.513	0.518	0.523	0.519	0.532	0.522	0.524	0.522	0.444	0.635
Waiting for another trial	0.638	0.638	0.636	0.643	0.646	0.638	0.640	0.638	0.567	0.744
Morning pretrial	0.326	0.309	0.355	0.343	0.363	0.350	0.364	0.351	0.359	0.337
Evening pretrial	0.344	0.352	0.321	0.331	0.336	0.328	0.332	0.333	0.335	0.326
Early morning pretrial	0.329	0.339	0.324	0.326	0.301	0.322	0.304	0.316	0.306	0.337
Weekend pretrial	0.029	0.039	0.038	0.034	0.029	0.031	0.033	0.035	0.034	0.033
Day of the year	168	142	199	200	221	197	217	199	198	199
Monday	0.577	1	0.234	0.238	0	0.258	0	0.243	0.242	0.237
Tuesday	0.423	0	0.169	0.174	0	0.177	0.219	0.176	0.173	0.170
Wednesday	0	0	0.161	0.168	0.252	0.151	0.213	0.160	0.165	0.152
Thursday	0	0	0.144	0.146	0.254	0.148	0.193	0.150	0.150	0.145
Friday	0	0	0.150	0.137	0.252	0.135	0.196	0.143	0.145	0.144
Saturday	0	0	0.142	0.137	0.242	0.131	0.178	0.127	0.126	0.152

Table 8: Estimates of judge fixed effects on pre-trial detention, robust standard errors

	(1)	(2)	(3)
Judge 1	-0.0322^{****}	-0.0306^{****}	-0.0305^{****}
	(0.0056)	(0.0052)	(0.0050)
Judge 3	-0.0232^{****}	-0.0211^{****}	-0.0189^{****}
	(0.0051)	(0.0047)	(0.0046)
Judge 4	-0.0455^{****}	-0.0452^{****}	-0.0440^{****}
	(0.0051)	(0.0047)	(0.0045)
Judge 5	-0.0326^{****}	-0.0311^{****}	-0.0296^{****}
	(0.0056)	(0.0052)	(0.0050)
Judge 6	-0.0086^{*}	-0.0080^{*}	-0.0071
	(0.0051)	(0.0047)	(0.0046)
Judge 7	-0.0298^{****}	-0.0272^{****}	-0.0250^{****}
	(0.0055)	(0.0050)	(0.0049)
Judge 8	-0.0419^{****}	-0.0373^{****}	-0.0358^{****}
	(0.0051)	(0.0047)	(0.0046)
Time effects	Yes	Yes	Yes
Case characteristics	-	Yes	Yes
Offender characteristics	-	-	Yes
F statistic	35.22^{****}	40.39^{****}	43.06^{****}
J statistic	8.43	10.88^{*}	12.16^{*}

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 9: Estimates of effect of pre-trial detention on guilt, robust standard errors

		(1)	(2)	(3)
OLS	Est	0.0002	0.0563****	0.0289****
	Sd	(0.0018)	(0.0019)	(0.0019)
2SLS	Est	0.1509	0.1879***	0.1818***
	Sd	0.06526	0.06457	0.06403
	Judge IV	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}
2SLS AN	Est	0.1509**	0.1879***	0.1818***
	Sd	0.06526	0.06457	0.06403
	Judge IV	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}
2SLS DN	Est	0.1509**	0.1879***	0.1818***
	Sd	0.06526	0.06457	0.06403
	Judge IV	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}	{1;3;4;5;6;7;8}
2SLS Post-lasso	Est	0.2549**	0.2552**	0.2594**
	Sd	0.1166	0.1108	0.1127
	Judge IV	{4}	{4}	{4}
2SLS Post-adalasso	Est	0.2117***	0.2018***	0.5206
	Sd	0.07552	0.06505	0.7508
	Judge IV	{4;6}	{3;4;5;6;7;8}	{6}
2SLS R_{EXO}	Est	0.3832**	0.5025***	0.2595**
	Sd	0.1504	0.1606	0.1126
	Judge IV	{6;7}	{6;7}	{4;8}
2SLS R_{PMSE}	Est	0.2549**	0.2556**	0.2594**
	Sd	0.1166	0.1108	0.1127
	Judge IV	{4}	{4;8}	{4}
2SLS R_{MSE}	Est	0.06212	0.1190	-0.01643
	Sd	0.1086	0.1171	0.1477
	Judge IV	{1;3;7;8}	{1;3;7;8}	{1;6}
Time effects		Yes	Yes	Yes
Case characteristics		-	Yes	Yes
Demographics and other		-	-	Yes

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 10: Diagnostics post selection when estimating the effect of pre-trial detention on guilt, cluster robust standard errors

Selection method		(1)	(2)	(3)
No selection	F statistic	35.2162****	40.3948****	43.0605****
	J statistic	8.4281	10.8814*	12.1602*
AN	F statistic	35.2162****	40.3948****	43.0605****
	J statistic	8.4281	10.8814*	12.1602*
DN	F statistic	35.2162****	40.3948****	43.0605****
	J statistic	8.4281	10.8814*	12.1602*
Lasso	F statistic	80.2850****	97.8957****	99.6765****
	J statistic	0.0000	0.0000	0.0000
Adalasso	F statistic	93.9719****	46.5946****	2.5798
	J statistic	0.2422	6.9497	0.0000
R_{EXO}	F statistic	25.9478****	26.3822****	49.9060****
	J statistic	0.0416	0.0050	0.0019
R_{PMSE}	F statistic	80.2850****	48.9627****	99.6765****
	J statistic	0.0000	0.0439	0.0000
R_{MSE}	F statistic	21.8693****	21.2626****	27.8508****
	J statistic	7.1293*	9.0086**	0.6231
Time effects		Yes	Yes	Yes
Case characteristics		-	Yes	Yes
Demographics and other		-	-	Yes

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 11: 2SLS estimators and p-values by excluded judge dummy and by judge used as IV, robust standard errors, specification (1)

		Excluded judge dummy variable							
		1	2	3	4	5	6	7	8
Judge IV	1	-	0.1205	-0.8127	0.5811	19.3040	-0.02174	-3.3776	0.4673
	2	0.1205	-	0.4841*	0.2549**	0.3633**	0.5105	0.4096**	0.2008
	3	-0.8127	0.4841*	-	0.01663	0.0668	0.4684**	0.1470	-0.1495
	4	0.5811	0.2549**	0.01663	-	-0.02023	0.1952**	-0.03786	0.8918
	5	19.3040	0.3633**	0.0668	-0.02023	-	0.3105**	-0.1168	-0.3692
	6	-0.02174	0.5105	0.4684**	0.1952**	0.3105**	-	0.3685**	0.1208
	7	-3.3776	0.4096**	0.1470	-0.03786	-0.1168	0.3685**	-	-0.3096
	8	0.4673	0.2008	-0.1495	0.8918	-0.3692	0.1208	-0.3096	-

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 12: 2SLS estimators and p-values by excluded judge dummy and by judge used as IV, robust standard errors, specification (2)

		Excluded judge dummy variable							
		1	2	3	4	5	6	7	8
Judge IV	1	-	0.05449	-0.7387	0.6742**	21.3701	-0.0888	-3.4442	1.0862
	2	0.05449	-	0.4113*	0.2552**	0.4229**	0.4563	0.4928**	0.2396*
	3	-0.7387	0.4113*	-	0.1188	0.4473	0.3835*	0.7754	0.01574
	4	0.6742**	0.2552**	0.1188	-	-0.1147	0.2117***	-0.1025	0.3281
	5	21.3701	0.4229**	0.4473	-0.1147	-	0.4112***	-0.05897	-0.6879
	6	-0.0888	0.4563	0.3835*	0.2117***	0.4112***	-	0.5081***	0.1800*
	7	-3.4442	0.4928**	0.7754	-0.1025	-0.05897	0.5081***	-	-0.4422
	8	1.0862	0.2396*	0.01574	0.3281	-0.6879	0.1800*	-0.4422	-

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

Table 13: 2SLS estimators and p-values by excluded judge dummy and by judge used as IV, robust standard errors, specification (3)

		Excluded judge dummy variable							
		1	2	3	4	5	6	7	8
Judge IV	1	-	0.05902	-0.5675	0.7104**	-12.5002	-0.08053	-2.2185	1.3849
	2	0.05902	-	0.4449	0.2594**	0.4376**	0.5206	0.5577**	0.2560*
	3	-0.5675	0.4449	-	0.1202	0.4248	0.3995	0.9041	0.04559
	4	0.7104**	0.2594**	0.1202	-	-0.1062	0.2093***	-0.1330	0.2740
	5	-12.5002	0.4376**	0.4248	-0.1062	-	0.4115***	-0.2176	-0.6094
	6	-0.08053	0.5206	0.3995	0.2093***	0.4115***	-	0.5724***	0.1908*
	7	-2.2185	0.5577**	0.9041	-0.1330	-0.2176	0.5724***	-	-0.4430
	8	1.3849	0.2560*	0.04559	0.2740	-0.6094	0.1908*	-0.4430	-

$^{****}p < 0.001$; $^{***}p < 0.01$; $^{**}p < 0.05$; $^{*}p < 0.1$

B Application Appendix

In this section I justify the choice of "excluding" judge 2 from the set of eight potential judge dummy IVs and guess which judge dummy is possibly endogenous. First notice that by projecting out the exogenous regressors except the intercept the model can be written as

$$guilt_i = \gamma_0 + predet_i\beta + \sum_{j\in\overline{E}} \alpha_j 1\{judge_i = j\} + u_i$$

$$predet_i = \delta_0 + \sum_{j=1}^{8} \pi_j 1\{judge_i = j\} + v_i$$

Then let for $j = 1, \ldots, 8$ $p_j = \mathbb{P}(judge_i = j)$, $\overline{predet} = \mathbb{E}(predet_i)$ and $\overline{guilt} = \mathbb{E}(guilt_i)$ so that the demeaned variables can be written for $j = 1, \ldots, 8$ as $D_{ij} = 1\{judge_i = j\} - p_j$, $\tilde{predet}_i = predet_i - \overline{predet}$ and $\tilde{guilt}_i = guilt_i - \overline{guilt}$. The intercept can then be projected out so that the model can be rewritten as

$$\tilde{guilt}_i = \tilde{predet}_i\beta + \sum_{j\in\overline{E}} \alpha_j D_{ij} + \tilde{u}_i$$

$$\tilde{predet}_i = \sum_{j=1}^{8} \pi_j D_{ij} + \tilde{v}_i$$

Thus because $\sum_{j=1}^{8} D_{ij} = 0$ there is a multicolinearity problem in the first stage and some judge variable must be removed.

Then what if, for instance, judge 8 is excluded and not used either as an IV or as a control variable? Can it still enter the structural equation? If judge 8 is excluded and enters the structural equation then it is actually $D_{i8} = 1\{judge_i = 8\} - p_8 = \sum_{j=1}^{7}(p_j - 1\{judge_i = j\}) = \sum_{j=1}^{7} D_{ij}$ which enters the structural equation. Hence affirming that excluded judge 8 enters the structural equation implies that there is at least one other judge dummy variable which is endogenous which is the usual case considered in the paper. On the other hand affirming that judge 8 doesn't enter the structural equation, and thus doesn't imply that all other judge dummy variables are exogenous hence it is also the usual case. Consequently endogeneity or exogeneity of the excluded variable is entirely tied to endogeneity or exogeneity of the rest of the judge dummy variables which can be dealt with IV selection methods.

Still in order to limit as much as possible the overall amount of endogeneity I exclude the judge which has the lowest amount of cases, ie judge 2 with $\hat{p}_2 = 0.039$, see 6 in appendix A.3. Indeed, if judge 2 is excluded but still enters the structural equation through the sum of the other dummies, endogeneity is limited compared to

when excluding other judges. When the model is instrumented by subset S with $\overline{S}$ projected out

$$\tilde{guilt}_i = \tilde{predet}_i\beta + \tilde{u}_{iS}, \qquad \tilde{predet}_i = \sum_{j\in S}\pi_j D_{ij} + v_i$$

where the error is $\tilde{u}_{iS} = \alpha_2 D_{i2} + \sum_{j\in\overline{E}}\alpha_j D_{ij}\tilde{u}_i$. Thus for any S part of its level of endogeneity is determined by $\mathbb{E}\left(\sum_{j\in S} D_{ij}D_{i2}\right) = -\mathbb{P}(judge_i = 2)\sum_{j\in S}\mathbb{P}(judge_i = j)$ which is small because $\mathbb{P}(judge_i = 2)$ is the smallest out of all judge propensity to take a case.

In addition, as mentioned before excluding judge j implies that all other judge effects are relative to judge j in the first stage. This means that the IVs which are individually insignificant in the first stage depend on the judge which is excluded. Thus the excluded judge determines which IV subsets are weak, this is quite problematic because IV selection methods are sensitive to weak IVs, including the ones developed in this paper to some extent. For this reason excluding judge 2 is a great choice because judge 2 has the largest propensity to send someone to pre-trial detention, thus when excluding judge 2 other judge dummies are more likely to be individually significant.

In tables 11, 12, and 13 are 2SLS estimators computed for specification (1), (2) and (3) respectively, where each column correspond to which judge was excluded, and where each row corresponds to which judge was used as the sole IV to construct 2SLS (other judge dummies act as controls). First notice that the tables are symmetric, this is due to the fact that if in the model six judge dummies and an intercept are controlled for then the two last judge dummy variables are equal to each other hence excluding one or the other does not change the estimator. Second note that, across specifications, when judge 2 and to a lesser extent judge 8 are excluded (or used as the IV) the 2SLS estimators are much more likely to be significant, this is because they are the harshest judges and relative to them other judge dummies have a significant effect on pre-trial detention. Third, notice that across specifications some entries are extremes, sometimes negative, especially when judge 5 or judge 7 are excluded (or used as the IV). This is either due to the fact that being imprisoned before a trial is taken as a signal that the offender is not guilty, which is implausible, or this is due to these judge dummy variables being endogenous and producing a negative bias. Indeed if the true model is D_{ij} is used as the sole IV but enters the structural equation as in

$$\tilde{guilt}_i = \tilde{predet}_i\beta + \alpha_j D_{ij} + \tilde{u}_i, \qquad \tilde{predet}_i = \pi_j D_{ij} + \tilde{v}_i$$

This can generate a negative bias when $sign(\alpha_j) \neq sign(\pi_j)$. If a pre-trial judge is too lenient with a negative π_j then the judge which supervises the trial may compensate and be harsher with a positive α_j. Fourth the differences between the individual judge IV estimators in tables 11, 12 and 13 could also be due (in part) to heterogeneity in β or to endogeneity of all judge IVs, although both are unlikely.

To conclude, there are very good reasons to exclude judge 2.

C Main Theorems Proofs

The proofs of the main asymptotic results are divided in four parts. First the limit in distribution of all the risks, risk estimators and the limit of their expectation are found. Second IV sets are ranked in accordance to limit of each risk. Third I prove that the feasible risk estimators and the risks have the same limit in probability. Fourth I combine these results to prove the efficiency and the consistency of selection via risk estimator minimization.

Before writing the proofs, some notations and conventions are introduced. Unless specified, all limits are taken with respect to n. Additionally, we denote by the expression $X = o_P(n^a)$ a random variable or statistic X which is asymptotically degenerate of order n^a, ie $X = o_P(n^a) \Leftrightarrow \forall e > 0 \ \mathbb{P}(|X|n^{-a} > e) \underset{n\to\infty}{\to} 0$, and denote by $X = O_p(n^a)$ a random variable which is (bounded in probability) of order n^a, ie $\forall e > 0 \ \exists M > 0, \exists N : \forall n \geqslant N \ \mathbb{P}(|X|n^{-a} > M) < e$. The usual properties of o_P and O_P random variables are used throughout these proofs. In adition, $plim\ X$ denotes the limit in probability of a random variable or statistic X whereas $dlim\ X$ denotes its limit in distribution.

Furthermore denote $\mathbb{E}(v_{iS}^2) = \sigma_{vS}^2$ and $\Sigma_S = \mathbb{E}(z_{iS}z_{iS}')$. To simplify notations $\hat{C}_S = \frac{x'(P_S - \alpha(P_S)I_n)u}{x'(P_S - \alpha(P_S)I_n)x}$ only in lemma 3.1 and lemma 3.2, for the rest of the proofs $\hat{C}_S = \frac{x'P_{z_S}u}{x'P_{z_S}x}$ if assumption **??**(ii) holds or $\hat{C}_S = \frac{x'P_{z_S}u - s\rho}{x'P_{z_S}x - s\sigma_{vS}^2}$ if assumption **??**(iii) holds and $a_S \geqslant 1/2$. This simplification is made in order to replace $\hat{C}_S$ by a simpler expression which has the same limit in probability, see the proofs of lemma 3.1 and lemma 3.2 for details.

Define the following within-sample and out-of-sample risk estimators of $nR_1(S)$

$$\hat{R}_w(S) = \frac{1}{n}(y - x\hat{\beta}_S)'z_S\Sigma_S z_S'(y - x\hat{\beta}_S)$$
$$\hat{R}_o(S) = \frac{1}{n}(y - x\hat{\beta}_S^*)'z_S\Sigma_S z_S'(y - x\hat{\beta}_S^*)$$

where $\hat{\beta}_S^*$ is built from the sample $(w_i^*)_{i=1}^n$ which has the same DGP but is independent of $(w_i)_{i=1}^n$. And define the apparent risk estimators

$$\hat{R}_{EXO,app}(S) = \frac{1}{n^2}(y - x\hat{\beta}_S)'z_S\Sigma_S z_S'(y - x\hat{\beta}_S)$$
$$\hat{R}_{PMSE,app}(S) = \frac{1}{n}\sum_{i=1}^n (y_i - z_{iS}'\pi_S\hat{\beta}_S)^2$$
$$\hat{R}_{MSE,app}(S) = \frac{1}{n}\sum_{i=1}^n (y_i - x_i\hat{\beta}_S)^2$$

where $\hat{\beta}_S$ was computed from original sample $(w_i)_{i=1}^n$.

C.1 Technical lemmas: Expected Risk Estimators Asymptotics

Lemma 3.1

Under assumptions **??**, **??***(i), and* **??***(ii), for any* $S \in \mathcal{S}$

- *If* $a_S \in [0; 1/2)$ *then*

$$plim\ \hat{\beta}_S = \ plim\ \frac{x' P_{z_S} y}{x' P_{z_S} x}$$

Furthermore

$$plim\ \hat{C}_S = \ plim\ \hat{C}_S^2 = 0, \qquad n^{1/2-a_S}\hat{C}_S = \frac{\kappa_S' \Sigma_S^{1/2} \lambda_u}{\kappa_S' \Sigma_S \kappa_S} = O_P(1)$$

- *If* $a_S = 1/2$ *then*

$$dlim\ \hat{\beta}_S = \ dlim\ \frac{x' P_{z_S} y}{x' P_{z_S} x}$$

Furthermore

$$dlim\ \hat{C}_S \equiv C_w = \frac{\tilde{\lambda}_{vS}' \lambda_u}{||\tilde{\lambda}_{vS}||^2}$$

- *If* $a_S > 1/2$ *then*

$$dlim\ \hat{\beta}_S = \ dlim\ \frac{x' P_{z_S} y}{x' P_{z_S} x}$$

Furthermore

$$dlim\ \hat{C}_S \equiv C_{S,vw} = \frac{\lambda_{vS}' \lambda_u}{||\lambda_{vS}||^2}$$

where $(\lambda_u, \lambda_{vS}, \tilde{\lambda}_{vS})$ *is Gaussian.*

Proof. By assumption **??**(ii) if $a_S \in [0; 1/2)$ then

$$n^{a_S - 1/2} x'(P_S - \alpha(P_S) I_n) u = n^{a_S - 1/2} x' P_{z_S} u + o_P(1)$$
$$n^{2a_S - 1} x'(P_S - \alpha(P_S) I_n) x = n^{2a_S - 1} x' P_{z_S} x + o_P(1)$$

Thus using by the Continuous Mapping Theorem (CMT)

$$\begin{aligned} n^{1/2-a_S}(\hat{\beta}_S - \beta) = n^{1/2-a_S}\hat{C}_S &= n^{1/2-a_S}\frac{n^{2a_S-1}}{n^{2a_S-1}}\frac{x'(P_S - \alpha(P_S)I_n)u}{x'(P_S - \alpha(P_S)I_n)x} \\ &= \frac{n^{a_S-1/2}}{n^{2a_S-1}}\frac{x'(P_S - \alpha(P_S)I_n)u}{x'(P_S - \alpha(P_S)I_n)x} \\ &= \frac{n^{a_S-1/2}}{n^{2a_S-1}}\frac{x'P_{z_S}u}{x'P_{z_S}x} + o_P(1) \end{aligned}$$

which implies

$$\hat{C}_S = \frac{x'P_{z_S}u}{x'P_{z_S}x} + o_P(1), \quad plim\ \hat{\beta}_S = \ plim\ \frac{x'P_{z_S}y}{x'P_{z_S}x}$$

Similarly when $a_S \geqslant 1/2$ then

$$\begin{aligned} x'(P_S - \alpha(P_S)I_n)u &= x'P_{z_S}u + o_P(1) \\ x'(P_S - \alpha(P_S)I_n)x &= x'P_{z_S}x + o_P(1) \end{aligned}$$

Thus

$$\hat{C}_S = \frac{x'P_{z_S}u}{x'P_{z_S}x} + o_P(1), \quad dlim\ \hat{C}_S = dlim\ \frac{x'P_{z_S}u}{x'P_{z_S}x}, \quad dlim\ \hat{\beta}_S = \ dlim\ \frac{x'P_{z_S}y}{x'P_{z_S}x}$$

Next take note of the following decompositions of $\hat{C}_S$

$$\begin{aligned} \hat{C}_S &= \frac{v_S'P_{z_S}u + \pi_S'z_S'u}{\pi_S'z_S'z_S\pi_S + 2\pi_S'z_Sv_S + v_S'P_{z_S}v_S} = \frac{v_S'P_{z_S}u + n^{-a_S}\kappa_S'z_S'u}{n^{-2a_S}\kappa_S'z_S'z_S\kappa_S + 2n^{-a_S}\kappa_S'z_Sv_S + v_S'P_{z_S}v_S} \\ &= \frac{n^{2a_S-1}v_S'P_{z_S}u + n^{a_S-1}\kappa_S'z_S'u}{n^{-1}\kappa_S'z_S'z_S\kappa_S + 2n^{a_S-1}\kappa_S'z_Sv_S + n^{2a_S-1}v_S'P_{z_S}v_S} \end{aligned}$$

Then take note that

$$v_S'P_{z_S}u = \frac{1}{n}v_S'z_S(\frac{1}{n}z_S'z_S)^{-1}\frac{1}{\sqrt{n}}z_S'u$$

The following Law of Large Numbers (LLN) and Central Limit Theorem (CLT) then apply by assumption **??**

$$\begin{aligned} &n^{-1/2}z_S'u \xrightarrow{d} \mathcal{N}(0, \sigma_u^2\Sigma_S), \qquad n^{-1/2}z_S'v_S \xrightarrow{d} \mathcal{N}(0, \sigma_{vS}^2\Sigma_S) \\ &n^{-1}z_S'z_S \xrightarrow{\mathbb{P}} \Sigma_S \end{aligned}$$

Thus denote

$$\begin{aligned} \lambda_u &\equiv dlim(n^{-1/2}z_S'u\Sigma_S^{-1/2}) \sim \mathcal{N}(0, \sigma_u^2I_s) \\ \lambda_{vS} &\equiv dlim(n^{-1/2}z_S'v_S\Sigma_S^{-1/2}) \sim \mathcal{N}(0, \sigma_{vS}^2I_s) \end{aligned}$$

and note that for any $j = 1, \ldots, s \quad \mathbb{E}(\lambda_{u,j}\lambda_{vS,j}) = \rho$ but $\mathbb{E}(\lambda_{u,j}\lambda_{vS,j'}) = 0$ if $j' \neq j$. Thus by properties of Gaussian vectors

$$\lambda_u = \frac{\rho}{\sigma_{vS}^2}\lambda_{vS} + \varepsilon_S$$

where $\varepsilon_{S,j}$ is independent of $\lambda_{vS,j}$ and $\varepsilon_S \sim \mathcal{N}\left(0, (\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2})I_s\right)$. Consequently

- If $a_S \in [0; 1/2)$ then

$$\begin{aligned}\hat{C}_S &= \frac{n^{2a_S-1}v_S'P_{z_S}u + n^{a_S-1}\kappa_S'z_S'u}{n^{-1}\kappa_S'z_S'z_S\kappa_S + 2n^{a_S-1}\kappa_S'z_Sv_S + n^{2a_S-1}v_S'P_{z_S}v_S} = \frac{O_P(n^{2a_S-1}) + O_P(n^{a_S-1/2})}{O_P(1) + O_P(n^{a_S-1/2}) + O_P(n^{2a_S-1})} \\ &= o_P(1)\end{aligned}$$

  which implies that $plim\ \hat{\beta}_S = \ plim\ \frac{x'P_{z_S}y}{x'P_{z_S}x} = \beta$. Furthermore by the CLT and Slutsky's lemma (SL)

$$\begin{aligned}n^{1/2-a_S}\hat{C}_S &= \frac{n^{a_S-1/2}v_S'P_{z_S}u + n^{-1/2}\kappa_S'z_S'u}{n^{-1}\kappa_S'z_S'z_S\kappa_S + 2n^{a_S-1}\kappa_S'z_Sv_S + n^{2a_S-1}v_S'P_{z_S}v_S} \\ &= \frac{O_P(n^{a_S-1/2}) + O_P(1)}{O_P(1) + O_P(n^{a_S-1/2}) + O_P(n^{2a_S-1})} \\ &= O_P(1) \\ \Rightarrow\ dlim\ n^{1/2-a_S}\hat{C}_S &= \frac{\kappa_S'\Sigma_S^{1/2}\lambda_u}{\kappa_S'\Sigma_S\kappa_S}\end{aligned}$$

- If $a_S = 1/2$ then by the CLT and SL

$$\begin{aligned}\hat{C}_S &= \frac{v_S'P_{z_S}u + n^{-1/2}\kappa_S'z_S'u}{n^{-1}\kappa_S'z_S'z_S\kappa_S + 2n^{-1/2}\kappa_S'z_Sv_S + v_S'P_{z_S}v_S} = O_P(1) \\ \Rightarrow\ dlim\ \hat{C}_S &= \frac{\lambda_{vS}'\lambda_u + \kappa_S\Sigma_S^{1/2}\lambda_u}{\lambda_{vS}'\lambda_{vS} + 2\kappa_S'\Sigma^{1/2}\lambda_{vS} + \kappa_S'\Sigma_S\kappa_S} = \frac{(\lambda_{vS} + \Sigma_S^{1/2}\kappa_S)'\lambda_u}{||\lambda_{vS} + \Sigma_S\kappa_S||^2} \\ &\equiv C_{S,w} \equiv \frac{\tilde{\lambda}_{vS}'\lambda_u}{||\tilde{\lambda}_{vS}||^2}\end{aligned}$$

  where $\tilde{\lambda}_{vS} = \lambda_v + \Sigma_S^{1/2}\kappa_S$. Going further using properties of Gaussian vectors λ_u can be rewritten as

$$\lambda_u = \frac{\rho}{\sigma_{vS}^2}\tilde{\lambda}_{vS} - \frac{\rho}{\sigma_{vS}^2}\kappa\Sigma_S^{1/2} + \varepsilon_S$$

Then note that

$$C_{S,w} = \frac{\rho}{\sigma_{vS}^2} + \frac{\varepsilon'\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2} - \frac{\rho}{\sigma_{vS}^2}\kappa_S'\Sigma_S^{1/2}\frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2}$$

$$C_{S,w}^2 = \frac{\rho^2}{\sigma_{vS}^4} + \frac{\varepsilon' P_{\tilde{\lambda}_{vS}}\varepsilon}{||\tilde{\lambda}_{vS}||^2} + \frac{\rho^2}{\sigma_{vS}^4}\kappa_S'\Sigma_S^{1/2}\frac{P_{\tilde{\lambda}_{vS}}}{||\tilde{\lambda}_{vS}||^2}\Sigma_S^{1/2}\kappa_S +$$
$$+ 2\frac{\rho}{\sigma_{vS}^2}\frac{\varepsilon'\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2} - 2\frac{\rho}{\sigma_{vS}^2}\kappa_S'\Sigma_S^{1/2}\frac{P_{\tilde{\lambda}_{vS}}}{||\tilde{\lambda}_{vS}||^2}\varepsilon - 2\frac{\rho^2}{\sigma_{vS}^4}\kappa_S'\Sigma_S^{1/2}\frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2}$$

$$\Rightarrow \mathbb{E}(C_{S,w}) = \frac{\rho}{\sigma_{vS}^2}\left(1 - \kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2}\right)\right)$$

$$\Rightarrow \mathbb{E}(C_{S,w}^2) = \frac{\rho^2}{\sigma_{vS}^4}\left(1 - 2\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2}\right) + \kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}}{||\tilde{\lambda}_{vS}||^2}\right)\Sigma_S^{1/2}\kappa_S\right)$$
$$+ \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\mathbb{E}(||\tilde{\lambda}_{vS}||^{-2})$$
$$= \frac{\rho^2}{\sigma_{vS}^4}\mathbb{E}\left(||1 - \frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2}\Sigma_S^{1/2}\kappa_S||^2\right) + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)(\sigma_{vS}^2(s-2))^{-1}$$

- If $a_S > 1/2$ then

$$\hat{C}_S = \frac{v_S'P_{z_S}u + n^{-a_S}\kappa_S'z_S'u}{n^{-2a_S}\kappa_S'z_S'z_S\kappa_S + 2n^{-a_S}\kappa_S'z_Sv_S + v_S'P_{z_S}v_S}$$
$$= = \frac{v_S'P_{z_S}u}{v_S'P_{z_S}v_S} + o_P(1)$$
$$\Rightarrow \ dlim\ \hat{C}_S \equiv C_{S,vw} = \frac{\lambda_{vS}'\lambda_u}{||\lambda_{vS}||^2}$$

Using the decomposition mentioned before note that

$$C_{S,vw} = \frac{\rho}{\sigma_{vS}^2} + \frac{\varepsilon'\lambda_{vS}}{||\lambda_{vS}||^2} \Rightarrow \mathbb{E}(C_{S,vw}) = \frac{\rho}{\sigma_{vS}^2}$$

$$C_{S,vw}^2 = \frac{\rho^2}{\sigma_{vS}^4} + \frac{\varepsilon' P_{\lambda_{vS}}\varepsilon}{||\lambda_{vS}||^2} + 2\frac{\rho}{\sigma_{vS}^2}\frac{\varepsilon'\lambda_{vS}}{||\lambda_{vS}||^2} \Rightarrow \mathbb{E}(C_{S,vw}^2) = \frac{\rho^2}{\sigma_{vS}^4} + \left(\sigma_u^2 - \frac{\rho}{\sigma_{vS}^2}\right)\mathbb{E}(||\lambda_{vS}||^{-2})$$
$$\Rightarrow \mathbb{E}(C_{S,vw}^2) = \frac{\rho^2}{\sigma_{vS}^4} + \left(\sigma_u^2 - \frac{\rho}{\sigma_{vS}^2}\right)(\sigma_{vS}^2(s-2))^{-1}$$

□

Lemma 3.2

Under assumptions **??**, **??***(i), and* **??***(iii), for any* $S \in \mathcal{S}$

- *If* $a_S \in [0; 1/2)$ *then*

$$plim\ \hat{\beta}_S = \ plim\ \frac{x' P_{z_S} y}{x' P_{z_S} x}$$

Furthermore

$$plim\ \hat{C}_S = \ plim\ \hat{C}_S^2 = 0, \qquad n^{1/2-a_S} \hat{C}_S = \frac{\kappa_S' \Sigma_S^{1/2} \lambda_u}{\kappa_S' \Sigma_S \kappa_S} = O_P(1)$$

- *If* $a_S = 1/2$ *then*

$$dlim\ \hat{\beta}_S = \beta + \ dlim\ \frac{x' P_{z_S} u - s\rho}{x' P_{z_S} x - s\sigma_{vS}^2}$$

Furthermore

$$dlim\ \hat{C}_S \equiv C_w = \frac{\tilde{\lambda}_{vS}' \lambda_u - s\rho}{||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2}$$

- *If* $a_S > 1/2$ *then*

$$dlim\ \hat{\beta}_S = \beta + \ dlim\ \frac{x' P_{z_S} u - s\rho}{x' P_{z_S} x - s\sigma_{vS}^2}$$

Furthermore

$$dlim\ \hat{C}_S \equiv C_{S,vw} = \frac{\lambda_{vS}' \lambda_u - s\rho}{||\lambda_{vS}||^2 - s\sigma_{vS}^2}$$

where $(\lambda_u, \lambda_{vS}, \tilde{\lambda}_{vS})$ *is Gaussian.*

Proof. Case by case

- If $a_S \in [0; 1/2)$ then the result follows from the proof of lemma 3.1. If $a_S \geqslant$ then by assumption **??**(iii)

$$n^{a_S - 1/2} x'(P_S - \alpha(P_S) I_n) u = n^{a_S - 1/2} x' P_{z_S} u + o_P(1)$$
$$n^{2a_S - 1} x'(P_S - \alpha(P_S) I_n) x = n^{2a_S - 1/2} x' P_{z_S} x + o_P(1)$$

Thus by the CMT

$$\hat{C}_S = \frac{x'(P_S - \alpha(P_S) I_n) u}{x'(P_S - \alpha(P_S) I_n) x} = \frac{x' P_{z_S} u - s\rho}{x' P_{z_S} x - s\sigma_{vS}^2} + o_P(1)$$

and therefore

$$\hat{C}_S = \frac{x' P_{z_S} u - s\rho}{x' P_{z_S} x - s\sigma_{vS}^2} + o_P(1), \quad dlim\, \hat{C}_S = \ dlim\ \frac{x' P_{z_S} u - s\rho}{x' P_{z_S} x - s\sigma_{vS}^2}, \quad dlim\, \hat{\beta}_S = \beta + dlim\ \frac{x' P_{z_S} u - s\rho}{x' P_{z_S} x - s\sigma_{vS}^2}$$

- If $a_S = 1/2$ then based on the proof of lemma 3.1 and SL

$$dlim\ \hat{C}_S = C_{S,w} = \frac{\tilde{\lambda}'_{vS}\lambda_u - s\rho}{||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS}}$$

Therefore based on the decomposition of λ_u

$$\begin{aligned}
C_{S,w} &= \frac{||\tilde{\lambda}_{vS}||^2\frac{\rho}{\sigma^2_{vS}} - \tilde{\lambda}'_{vS}\Sigma_S^{1/2}\kappa_S\frac{\rho}{\sigma^2_{vS}} + \tilde{\lambda}'_{vS}\varepsilon - s\sigma^2_{vS}\frac{\rho}{\sigma^2_{vS}}}{||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS}} \\
&= \frac{\rho}{\sigma^2_{vS}}\left(1 - \frac{\tilde{\lambda}'_{vS}\Sigma_S^{1/2}\kappa_S}{||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS}}\right) + \frac{\tilde{\lambda}'_{vS}\varepsilon}{||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS}} \\
\Rightarrow \mathbb{E}(C_{S,w}) &= \frac{\rho}{\sigma^2_{vS}}\left(1 - \kappa'_S\Sigma_S^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}\Sigma_S^{1/2}\kappa_S}{||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS}}\right)\right) \\
C^2_{S,w} &= \frac{\rho^2}{\sigma^4_{vS}}\left(1 - \frac{\tilde{\lambda}'_{vS}\Sigma_S^{1/2}\kappa_S}{||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS}}\right)^2 + \varepsilon' P_{\tilde{\lambda}_{vS}}\varepsilon\frac{||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS})^2} \\
&\quad + 2\frac{\rho}{\sigma^2_{vS}}\left(1 - \frac{\tilde{\lambda}'_{vS}\Sigma_S^{1/2}\kappa_S}{||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS}}\right)\frac{\tilde{\lambda}'_{vS}\varepsilon}{||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS}} \\
\Rightarrow \mathbb{E}(C^2_{S,w}) &= \frac{\rho^2}{\sigma^2_{vS}}\left(1 + \left(\kappa'_S\Sigma_S^{1/2}P_{\tilde{\lambda}_{vS}}\Sigma_S^{1/2}\kappa_S\frac{||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS})^2}\right) - 2\kappa'_S\Sigma_{vS}^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS}}\right)\right) \\
&\quad + \left(\sigma^2_u - \frac{\rho^2}{\sigma^2_{vS}}\right)\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma^2_{vS})^2}\right)
\end{aligned}$$

- If $a_S > 1/2$ then based on the proof of lemma 3.1 and SL

$$dlim\ \hat{C}_S = C_{S,w} = \frac{\lambda'_{vS}\lambda_u - s\rho}{||\lambda_{vS}||^2 - s\sigma^2_{vS}}$$

Therefore based on the decomposition of λ_u

$$\begin{aligned}
C_{S,w} &= \frac{||\lambda_{vS}||^2\frac{\rho}{\sigma^2_{vS}} + \lambda'_{vS}\varepsilon - s\sigma^2_{vS}\frac{\rho}{\sigma^2_{vS}}}{||\lambda_{vS}||^2 - s\sigma^2_{vS}} \\
&= \frac{\rho}{\sigma^2_{vS}} + \frac{\lambda'_{vS}\varepsilon}{||\lambda_{vS}||^2 - s\sigma^2_{vS}} \\
\Rightarrow \mathbb{E}(C_{S,w}) &= \frac{\rho}{\sigma^2_{vS}} \\
C^2_{S,w} &= \frac{\rho^2}{\sigma^4_{vS}} + \varepsilon' P_{\lambda_{vS}}\varepsilon\frac{||\lambda_{vS}||^2}{(||\lambda_{vS}||^2 - s\sigma^2_{vS})^2} + 2\frac{\rho}{\sigma^2_{vS}}\frac{\lambda'_{vS}\varepsilon}{||\lambda_{vS}||^2 - s\sigma^2_{vS}} \\
\Rightarrow \mathbb{E}(C^2_{S,w}) &= \frac{\rho^2}{\sigma^2_{vS}} + \left(\sigma^2_u - \frac{\rho^2}{\sigma^2_{vS}}\right)\mathbb{E}\left(\frac{||\lambda_{vS}||^2}{(||\lambda_{vS}||^2 - s\sigma^2_{vS})^2}\right)
\end{aligned}$$

□

Lemma 3.3 *Under assumptions* **??**, **??***(i), and* **??***(ii) or* **??***(iii), for any* $S \in \mathcal{S}$

$$plim\ \hat{R}_{EXO,app}(S) = \ lim\ \mathbb{E}(\hat{R}_{EXO,app}(S)) = \ lim\ R_1(S) = 0$$

Proof. First decompose $\hat{R}_{EXO,app}(S) = \frac{1}{n}(y - x\hat{\beta}_S)' z_S \Sigma_S^{-1} \frac{1}{n} z_S'(y - x\hat{\beta}_S)$

$$\begin{aligned}
\hat{R}_{EXO,app} &= \frac{1}{n} u' z_S \Sigma_S^{-1} \frac{1}{n} z_S' u - \frac{2\hat{C}_S}{n} u' z_S \Sigma_S^{-1} \frac{1}{n} z_S' x + \frac{\hat{C}^2}{n} x' z_S \Sigma_S^{-1} \frac{1}{n} z_S' x \\
&= \frac{1}{n} u' z_S \Sigma_S^{-1} \frac{1}{n} z_S' u - \frac{2\hat{C}_S}{n} u' z_S \Sigma_S^{-1} \frac{1}{n} z_S' z_S \pi_S - \frac{2\hat{C}_S}{n} u' z_S \Sigma_S^{-1} \frac{1}{n} z_S' v_S \\
&\quad + \frac{\hat{C}_S^2}{n} \pi_S' z_S' z_S \Sigma_S^{-1} \frac{1}{n} z_S' z_S \pi_S + 2\frac{\hat{C}_S^2}{n} \pi_S' z_S' z_S \Sigma_S^{-1} \frac{1}{n} z_S' v_S + \frac{\hat{C}_S^2}{n} v_S' z_S \Sigma_S^{-1} \frac{1}{n} z_S' v_S
\end{aligned}$$

Then by the LLN and the CMT the first term converge to zero because $\mathbb{E}(u_i z_i) = 0$. Next I consider three cases

- If $a_S \in [0; 1/2)$ then, as established in lemma 3.1 and 3.2, $\hat{C}_S = o_P(1)$. Additionally by the LLN $\frac{1}{n} z_S' x = O_P(1)$ thus by the CMT $\hat{R}_{EXO,app}(S) = o_P(1)$.
- If $a_S = 1/2$ then $\hat{C}_S = O_P(1)$ however $\frac{1}{\sqrt{n}} \pi_S' z_S' z_S = O_P(1)$ thus with a slight abuse of notations by the CMT and SL

$$\hat{R}_{EXO,app}(S) = o_P(1) - \frac{2}{n} O_P(1) - \frac{2}{n} O_P(1) + \frac{1}{n} O_P(1) + \frac{2}{n} O_P(1) + \frac{1}{n} O_P(1) = o_P(1)$$

- If $a_S > 1/2$ then $\hat{C}_S = O_P(1)$ however $\frac{1}{\sqrt{n}} \pi_S' z_S' z_S = o_P(1)$ thus by the CMT and SL $\hat{R}_{EXO,app} = o_P(1)$.

Then via Vitali's convergence theorem, convergence in distribution / in probability and uniform integrability of $\hat{R}_{EXO,app}(S)$, see lemma **??**, imply convergence of the mean.

Finally recall that

$$\begin{aligned}
R_1(S) &= \mathbb{E}(||\mathbb{E}_n((y^* - x^* \hat{\beta}_S) z_S^*) \Sigma_S - 1/2||^2) \\
&= \mathbb{E}(u^* z_E') \Sigma_S^{-1} \mathbb{E}(z_E u^*) + \mathbb{E}(\hat{C}_S^2) \mathbb{E}(x^* z_E') \Sigma_S^{-1} \mathbb{E}(z_E x^*) - 2\mathbb{E}(\hat{C}_S) \mathbb{E}(x^* z_E') \Sigma_S^{-1} \mathbb{E}(z_E u^*) \\
&= \mathbb{E}(\hat{C}_S^2) \pi_S' \Sigma_S \pi_S = n^{-2a_S} \mathbb{E}(\hat{C}_S^2) \kappa_S' \Sigma_S \kappa_S
\end{aligned}$$

Thus $lim\ R_1(S)$ for any S because $\hat{C}_S = o_P(1)$ if $a_S \in [0; 1/2)$ and $\hat{C}_S = O_P(1)$ if $a_S \geqslant 1/2$. □

Lemma 3.4

Under assumptions **??**, *under* **??***(i), and* **??***(ii) or* **??***(iii), for any* $S \in \mathcal{S}$

$$\begin{aligned} plim\ \hat{R}_{PMSE,app}(S) =&\ lim\ \mathbb{E}(\hat{R}_{PMSE,app}(S)) =\ lim\ R_2(S) \\ &= \sigma_u^2 + \sigma_v^2\beta^2 + 2\beta\rho + \pi' A_{\overline{S}}' \left(\Sigma_{\overline{S}} - \mathbb{E}(z_{i\overline{S}} z_{iS}')\Sigma_S^{-1}\mathbb{E}(z_{iS} z_{i\overline{S}}')\right) A_{\overline{S}}\pi \end{aligned}$$

Furthermore this limit is minimized when $S = (z_{i1}, z_{i2}, \ldots, z_{K_z})$ *such that*

$$plim\ \hat{R}_{PMSE,app}(S) =\ lim\ R_2(S) = \sigma_u^2 + \sigma_v^2\beta^2 + 2\beta\rho$$

Proof. First decompose $\hat{R}_{PMSE,app}(S) = \frac{1}{n}\sum_i (y_i - z_{iS}\hat{\pi}_S\hat{\beta}_S)^2$

$$\begin{aligned} \hat{R}_{PMSE,app}(S) &= \frac{1}{n}\sum_i (-z_{iS}'\pi_S\hat{C}_S + v_{iS}\beta + u_i + z_{iS}'(\pi_S - \hat{\pi}_S)\hat{\beta}_S)^2 \\ &= \frac{1}{n}\sum_i (u_i + v_{iS}\beta)^2 + \frac{\hat{C}_S^2}{n}\sum_i \pi_S' z_{iS} z_{iS}'\pi_S - \frac{2\hat{C}_S}{n}\sum_i (u_i + v_{iS}\beta) z_{iS}'\pi_S \\ &\quad + (\pi_S - \hat{\pi}_S)' \left(\frac{\hat{\beta}_S^2}{n}\sum_i z_{iS} z_{iS}'(\pi_S - \hat{\pi}_S) + \frac{2\hat{\beta}_S}{n}\sum_i z_{iS}(u_i + v_{iS}\beta) - \frac{2\hat{\beta}_S\hat{C}_S}{n}\sum_i z_{iS} z_{iS}'\pi_S \right) \end{aligned}$$

The terms on the 2nd line are degenerate because they are bounded in probability and because $plim\ \hat{\pi}_S = \pi_S$. Indeed, for any $a_S \in \mathbb{R}^+$

$$\hat{\beta}_S = O_P(1), \quad \hat{C}_S = O_P(1), \quad \frac{1}{n}\sum_i z_{iS} z_{iS}' = O_P(1), \quad \frac{1}{n}\sum_i z_{iS}(u_i + v_{iS}\beta) = O_P(1), \quad \pi_S - \hat{\pi}_S = o_P(1)$$

Consequently by the LLN and the CMT

$$plim\ \hat{R}_{PMSE,app}(S) = \mathbb{E}((u_i + v_{iS}\beta)^2) + plim\ \left(\frac{\hat{C}_S^2}{n}\sum_i \pi_S' z_{iS} z_{iS}'\pi_S - \frac{2\hat{C}_S}{n}\sum_i (u_i + v_{iS}\beta) z_{iS}'\pi_S \right)$$

Then consider 2 cases: If $a_S \in [0; 1/2)$ then as seen previously $plim\ \hat{C}_S = 0$ thus the second and third terms are degenerate. If $a_S \geqslant 1/2$ then $\hat{C}_S$ converges weakly either towards $C_{S,w}$ either towards $C_{S,vw}$ but because $\pi_S = \frac{\kappa_S}{n^{a_S}}$ then $\frac{1}{n}\sum_i \pi_S' z_{iS} z_{iS}'\pi_S = o_P(1)$ and $\frac{1}{n}\sum_i (u_i + v_{iS}\beta) z_{iS}'\pi_S = o_P(1)$ so the second and third terms are again degenerate. Convergence of the mean is implied by uniform integrability of $\hat{R}_{PMSE,app}$ via Vitali's convergence theorem.

Then based on the definition of v_{iS} it follows that

$$\mathbb{E}(u_i + v_{iS}\beta)^2) = \sigma_u^2 + \sigma_v^2\beta^2 + 2\rho\beta + \pi' A_{\overline{S}}' \left(\Sigma_{\overline{S}} - \mathbb{E}(z_{i\overline{S}} z_{iS}')\Sigma_S^{-1}\mathbb{E}(z_{iS} z_{i\overline{S}}')\right) A_{\overline{S}}\pi$$

The matrix $\Sigma_{\overline{S}} - \mathbb{E}(z_{i\overline{S}} z_{iS}')\Sigma_S^{-1}\mathbb{E}(z_{iS} z_{i\overline{S}}')$ is positive semi-definite because it is symmetric and letting $BLP(z_{iS}|z_{i\overline{S}}) = z_{i\overline{S}}\Sigma_{\overline{S}}^{-1}\mathbb{E}(z_{i\overline{S}} z_{iS}')$ notice that

$$\Sigma_{\overline{S}} - \mathbb{E}(z_{i\overline{S}} z_{iS}')\Sigma_S^{-1}\mathbb{E}(z_{iS} z_{i\overline{S}}') = \mathbb{E}((z_{i\overline{S}} - BLP(z_{i\overline{S}}|z_{iS}))(z_{i\overline{S}} - BLP(z_{i\overline{S}}|z_{iS}))') \geqslant 0$$

The set S which minimizes $\mathbb{E}(u_i + v_{iS}\beta)^2)$ is the whole set of IVs as in that case $\overline{S} = \emptyset$ and $z_{i\overline{S}} = 0$.

Finally, consider a decomposition of $R_2(S)$

$$\begin{aligned} R_2(S) &= \mathbb{E}\left(\mathbb{E}_n\left((y_i^* - z_{iS}^{*\prime}\pi_S\hat{\beta}_S)^2\right)\right) = \mathbb{E}((u_i + v_{iS}\beta)^2) + \pi_S'\Sigma_S\pi_S\mathbb{E}(\hat{C}_S^2) - 2\mathbb{E}(\hat{C}_S)\mathbb{E}((v_{iS}\beta + u_i)z_{iS}')\pi_S \\ &= \mathbb{E}((u_i + v_{iS}\beta)^2) + \pi_S'\Sigma_S\pi_S\mathbb{E}(\hat{C}_S^2) + 0 \\ &\to \mathbb{E}((u_i + v_{iS}\beta)^2) = \ plim\ \hat{R}_{PMSE,app} \end{aligned}$$

Based on previous arguments $\pi_S'\Sigma_S\pi_S\mathbb{E}(\hat{C}_S^2) \to 0$ therefore $lim\ R_2(S) = \ plim\ \hat{R}_{PMSE,app}(S)$ for any S. □

Lemma 3.5
Under assumptions **??** *and* **??***(i), for any* $S \in \mathcal{S}$

- *If* $a_S \in [0; 1/2)$ *then*

$$plim\ \hat{R}_{4,app}(S) = \sigma_u^2$$

which implies

$$lim\ \mathbb{E}(\hat{R}_{4,app}(S)) = \sigma_u^2$$

- *If* $a_S = 1/2$ *then*

$$dlim\ \hat{R}_{4,app}(S) = \sigma_u^2 s + C_{S,w}^2\sigma_{vS}^2 - 2C_{S,w}\rho$$

which implies if **??***(ii) holds*

$$lim\ \mathbb{E}(\hat{R}_{4,app}(S)) = \sigma_u^2 + \frac{\rho^2}{\sigma_{vS}^2}\left(\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}}{||\tilde{\lambda}_{vS}||^2}\right)\Sigma_S^{1/2}\kappa_S - 1\right) + (\sigma_u^2\sigma_{vS}^2 - \rho^2)\mathbb{E}(||\tilde{\lambda}_{vS}||^{-2})$$

or if **??***(iii) holds*

$$lim\ \mathbb{E}(\hat{R}_{4,app}(S)) = \sigma_u^2 + \frac{\rho^2}{\sigma_{vS}^2}\left(\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right)\Sigma_S^{1/2}\kappa_S - 1\right) + (\sigma_u^2\sigma_{vS}^2 - \rho^2)\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right)$$

- *If* $a_S > 1/2$ *then*

$$dlim\ \hat{R}_{4,app}(S) = \sigma_u^2 s + C_{S,vw}^2\sigma_{vS}^2 - 2C_{S,vw}\rho$$

which implies if **??***(ii) holds*

$$lim\ \mathbb{E}(\hat{R}_{4,app}(S)) = \sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2} + (\sigma_u^2\sigma_{vS}^2 - \rho^2)\mathbb{E}(||\lambda_{vS}||^{-2})$$

or if **??***(iii) holds*

$$lim\ \mathbb{E}(\hat{R}_{4,app}(S)) = \sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2} + (\sigma_u^2\sigma_{vS}^2 - \rho^2)\mathbb{E}\left(\frac{||\lambda_{vS}||^2}{(||\lambda_{vS}||^2 - s\sigma_{vS}^2)^2}\right)$$

where $(C_{S,w}, C_{S,vw})$ *is random. Furthermore*

$$plim\ \hat{R}_{4,app}(S) = \ lim\ R_4(S)$$

Proof. First decompose $\hat{R}_{4,app}(S) = \frac{1}{n}\sum_i (y_i - x_i\hat{\beta}_S)^2$

$$\hat{R}_{4,app}(S) = \frac{1}{n}\sum_i u_i^2 + \frac{\hat{C}_S^2}{n}\sum_i x_i^2 - \frac{2\hat{C}_S}{n}\sum_i u_i x_i$$

Then I prove the 3 statements in order

- If $a_S \in [0; 1/2)$ then, as seen previously, $\hat{C}_S = o_P(1)$ so that by the LLN and the CMT $\hat{R}_{4,app} = o_P(1)$. Convergence in mean is a direct implication of lemma **??** and Vitali's convergence theorem.

- If $a_S = 1/2$ then by previous arguments

$$plim\ \frac{1}{n}\sum_i x_i^2 = \sigma_{vS}^2, \qquad plim\ \frac{1}{n}\sum_i u_i x_i = \rho, \qquad dlim\ \hat{C}_S = C_{S,w} = \frac{\tilde{\lambda}_{vS}'\lambda_u}{\tilde{\lambda}_{vS}'\tilde{\lambda}_{vS}}$$

Thus by SL

$$dlim\ \hat{R}_4(S) = \sigma_u^2 + C_{S,w}^2\sigma_{vS}^2 - 2C_{S,w}\rho$$

which implies, reusing Vitali's convergence theorem and the formula of $\mathbb{E}(C_{S,w})$ and $\mathbb{E}(C_{S,w}^2)$, if **??**(ii) holds

$$lim\ \mathbb{E}(\hat{R}_4(S)) = \sigma_u^2 + \frac{\rho^2}{\sigma_{vS}^2}\left(\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}}{||\tilde{\lambda}_{vS}||^2}\right)\Sigma_S^{1/2}\kappa_S - 1\right) + (\sigma_u^2\sigma_v^2 - \rho^2)\mathbb{E}(||\tilde{\lambda}_{vS}||^{-2})$$

and if **??**(iii) holds

$$lim\ \mathbb{E}(\hat{R}_4(S)) = \sigma_u^2 + \frac{\rho^2}{\sigma_{vS}^2}\left(\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right)\Sigma_S^{1/2}\kappa_S - 1\right) + (\sigma_u^2\sigma_v^2 - \rho^2)\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right)$$

- If $a_S > 1/2$ then

$$plim\ \frac{1}{n}\sum_i x_i^2 = \sigma_{vS}^2, \qquad plim\ \frac{1}{n}\sum_i u_i x_i = \rho, \qquad dlim\ \hat{C}_S = C_{S,w} = \frac{\lambda_{vS}'\lambda_u}{\lambda_{vS}'\lambda_{vS}}$$

Thus by SL

$$dlim\ \hat{R}_4(S) = \sigma_u^2 + C_{S,vw}^2\sigma_{vS}^2 - 2C_{S,vw}\rho$$

which implies, reusing lemma **??**, Vitali's convergence theorem and the formula of $\mathbb{E}(C_{S,vw})$ and $\mathbb{E}(C_{S,vw}^2)$, if **??**(ii) holds

$$lim\ \mathbb{E}(\hat{R}_4(S)) = \sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2} + (\sigma_u^2\sigma_{vS}^2 - \rho^2)\mathbb{E}(||\lambda_{vS}||^{-2})$$

and if **??**(iii) holds

$$lim\ \mathbb{E}(\hat{R}_4(S)) = \sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2} + (\sigma_u^2\sigma_{vS}^2 - \rho^2)\mathbb{E}\left(\frac{||\lambda_{vS}||^2}{(||\lambda_{vS}||^2 - s\sigma_{vS}^2)^2}\right)$$

Finally notice that

$$R_4(S) = \mathbb{E}_n\left(\mathbb{E}\left((y^* - x^*\hat{\beta}_S)^2\right)\right) = \sigma_u^2 + \mathbb{E}(\hat{C}_S^2)\mathbb{E}(x^{*2}) - 2\mathbb{E}(\hat{C}_S)\mathbb{E}(x^*u^*)$$

$$\to \begin{cases} \sigma_u^2 s & \text{if } a_S \in [0; 1/2) \\ \sigma_u^2 + \mathbb{E}(C_{S,w}^2)\sigma_{vS}^2 - 2\mathbb{E}(C_{S,w})\rho & \text{if } a_S = 1/2 \\ \sigma_u^2 + \mathbb{E}(C_{S,vw}^2)\sigma_{vS}^2 - 2\mathbb{E}(C_{S,vw})\rho & \text{if } a_S > 1/2 \end{cases}$$

which is exactly the same as the limit of $\mathbb{E}(\hat{R}_{4,app}(S))$. □

Lemma 3.6 *Under assumptions* **??** *and* **??***(i), for any* $S \in \mathcal{S}$

- *If* $a_S \in [0; 1/2)$ *then*

$$dlim\ \hat{R}_w(S) = \lambda_u' M_{\Sigma_S^{1/2}\kappa_S}\lambda_u, \qquad dlim\ \hat{R}_o(S) = \lambda_u'(I_s + P_{\Sigma_S^{1/2}\kappa_S})\lambda_u$$

which implies

$$lim\ \mathbb{E}\left(\frac{1}{s-1}\hat{R}_w(S)\right) = \ lim\ \mathbb{E}\left(\frac{1}{s+1}\hat{R}_o(S)\right) = \sigma_u^2$$

- *If* $a_S = 1/2$ *then*

$$dlim\ \hat{R}_w(S) = \lambda_u' M_{\tilde{\lambda}_{vS}}\lambda_u, \qquad dlim\ \hat{R}_o(S) = \lambda_u'\lambda_u + \lambda_u^{*\prime} P_{\tilde{\lambda}_{vS}^*}\lambda_u^* \frac{||\tilde{\lambda}_{vS}||^2}{||\tilde{\lambda}_{vS}^*||^2} - 2\frac{\tilde{\lambda}_{vS}^{*\prime}\tilde{\lambda}_u^*}{||\tilde{\lambda}_{vS}^*||^2}\tilde{\lambda}_{vS}'\lambda_u$$

which implies if **??***(ii) holds*

$$\begin{aligned}
lim\ \mathbb{E}(\hat{R}_w(S)) &= \sigma_u^2 s - \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right) - \frac{\rho^2}{\sigma_{vS}^4}\mathbb{E}(||\tilde{\lambda}_{vS} - P_{\tilde{\lambda}_{vS}}\Sigma_S^{1/2}\kappa_S||^2) \\
lim\ \mathbb{E}(\hat{R}_o(S)) &= \sigma_u^2 s + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\mathbb{E}(||\tilde{\lambda}_{vS}||^2)\mathbb{E}(||\tilde{\lambda}_{vS}||^{-2}) + \frac{\rho^2}{\sigma_{vS}^4}(\kappa_S'\Sigma_S\kappa_S - s\sigma_{vS}^2) \\
&\quad + \frac{\rho^2}{\sigma_{vS}^4}\left(\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}\tilde{\lambda}_{vS}'}{||\tilde{\lambda}_{vS}||^4}\right)\Sigma_S^{1/2}\kappa_S\mathbb{E}(||\tilde{\lambda}_{vS}||^2) - 2\kappa_S'\Sigma_S\kappa_S\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2}\right)\right)
\end{aligned}$$

and if **??***(iii) holds*

$$\begin{aligned}
lim\ \mathbb{E}(\hat{R}_w(S)) &= \sigma_u^2 s + \frac{\rho^2}{\sigma_{vS}^4}(\kappa_S'\Sigma_S\kappa_S - s\sigma_{vS}^2) \\
&\quad + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\left(\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^4}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right) - 2\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^2}{||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2}\right)\right) \\
&\quad + \frac{\rho^2}{\sigma_{vS}^4}\kappa_S'\Sigma_S^{1/2}\left(\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}||\tilde{\lambda}_{vS}||^4}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right) - 2\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}||\tilde{\lambda}_{vS}||^2}{||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2}\right)\right)\Sigma_S^{1/2}\kappa_S \\
lim\ \mathbb{E}(\hat{R}_o(S)) &= \sigma_u^2 s + \frac{\rho^2}{\sigma_{vS}^4}(\kappa_S'\Sigma_S\kappa_S - s\sigma_{vS}^2) \\
&\quad + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right)\mathbb{E}(||\tilde{\lambda}_{vS}||^2) \\
&\quad + \frac{\rho^2}{\sigma_{vS}^4}\kappa_S'\Sigma_S^{1/2}\mathbb{E}(||\tilde{\lambda}_{vS}||^2)\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right)\Sigma_S^{1/2}\kappa_S \\
&\quad - 2\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2}\right)\mathbb{E}(\tilde{\lambda}_{vS})'\Sigma_S^{1/2}\kappa_S
\end{aligned}$$

- *If* $a_S > 1/2$ *then*

  $$dlim\ \hat{R}_w(S) = \lambda_u' M_{\lambda_{vS}} \lambda_u, \qquad dlim\ \hat{R}_o(S) = \lambda_u'\lambda_u + \lambda_u^{*\prime} P_{\lambda_{vS}^*} \lambda_u^* \frac{||\lambda_{vS}||^2}{||\lambda_{vS}^*||^2} - 2\frac{\lambda_{vS}^{*\prime}\lambda_u^*}{||\lambda_{vS}^*||^2}\lambda_{vS}'\lambda_u$$

  which implies if **??***(ii) holds*

  $$lim\ \mathbb{E}(\hat{R}_w(S)) = \sigma_u^2 s - \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right) - \frac{s\rho^2}{\sigma_{vS}^2}$$
  $$lim\ \mathbb{E}(\hat{R}_o(S)) = \sigma_u^2 s - \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\mathbb{E}(||\lambda_{vS}||^2)\mathbb{E}(||\lambda_{vS}||^{-2}) - \frac{s\rho^2}{\sigma_{vS}^2}$$

  and if **??***(iii) holds*

  $$lim\ \mathbb{E}(\hat{R}_w(S)) = \sigma_u^2 s - \frac{s\rho^2}{\sigma_{vS}^2} + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\left(\mathbb{E}\left(\frac{||\lambda_{vS}||^4}{(||\lambda_{vS}||^2 - s\sigma_{vS}^2)^2}\right) - 2\mathbb{E}\left(\frac{||\lambda_{vS}||^2}{||\lambda_{vS}||^2 - s\sigma_{vS}^2}\right)\right)$$
  $$lim\ \mathbb{E}(\hat{R}_o(S)) = \sigma_u^2 s - \frac{s\rho^2}{\sigma_{vS}^2} + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right) s\sigma_{vS}^2 \mathbb{E}\left(\frac{||\lambda_{vS}||^2}{(||\lambda_{vS}||^2 - s\sigma_{vS}^2)^2}\right)$$

Proof. First I decompose $\hat{R}_w(S)$

$$\begin{aligned}\hat{R}_w(S) &= \frac{1}{n}(y - x\hat{\beta}_S)' z_S \Sigma_S^{-1} z_S'(y - x\hat{\beta}_S) \\ &= \frac{1}{n} u' z_S \Sigma_S^{-1} z_S' u + \frac{\hat{C}_S^2}{n} x' z_S \Sigma_S^{-1} z_S' x - 2\frac{\hat{C}_S}{n} x' z_S \Sigma_S^{-1} z_S' u\end{aligned}$$

Then notice by the CLT and SL that the first component has limit

$$dlim\ \frac{1}{n} u' z_S \Sigma_S^{-1} z_S' u = \lambda_u'\lambda_u$$

for any a_S. Then I decompose the second and third components

$$\frac{\hat{C}_S^2}{n} x' z_S \Sigma_S^{-1} z_S x = \frac{\hat{C}_S^2}{n} v_S' z_S \Sigma_S^{-1} z_S v_S + n^{1-2a_S}\frac{\hat{C}_S^2}{n^2}\kappa_S' z_S' z_S \Sigma_S^{-1} z_S' z_S \kappa_S + 2n^{1/2-a_S}\frac{\hat{C}_S^2}{n^{3/2}} v_S' z_S \Sigma_S^{-1} z_S' z_S \kappa_S$$

$$-2\frac{\hat{C}_S}{n} x' z_S \Sigma_S^{-1} z_S' u = -2\frac{\hat{C}_S}{n} v_S' z_S \Sigma_S^{-1} z_S' u - 2n^{1/2-a_S}\frac{\hat{C}_S}{n^{3/2}}\kappa_S' z_S' z_S \Sigma_S^{-1} z_S' u$$

Then case by case

- If $a_S \in [0; 1/2)$ then

$$
\begin{aligned}
&plim\ \frac{\hat{C}_S^2}{n} v_S' z_S \Sigma_S^{-1} z_S v_S = 0 \\
&dlim\ n^{1-2a_S} \frac{\hat{C}_S^2}{n^2} \kappa_S' z_S' z_S \Sigma_S^{-1} z_S' z_S \kappa_S = \lambda_u' P_{\Sigma_S^{1/2}\kappa_S} \lambda_u \\
&plim\ 2n^{1/2-a_S} \frac{\hat{C}_S^2}{n^{3/2}} v_S' z_S \Sigma_S^{-1} z_S' z_S \kappa_S = 0 \\
&plim\ -2\frac{\hat{C}_S}{n} v_S' z_S \Sigma_S^{-1} z_S' u = 0 \\
&dlim\ 2n^{1/2-a_S} \frac{\hat{C}_S}{n^{3/2}} \kappa_S' z_S' z_S \Sigma_S^{-1} z_S' u = -2\frac{\lambda_u' \Sigma_S^{1/2} \kappa_S}{\kappa_S' \Sigma_S \kappa_S} \lambda_u' \Sigma_S^{1/2} \kappa_S
\end{aligned}
$$

because $\hat{C}_S = O_P(n^{1/2-a_S})$ and $dlim\ n^{1/2-a_S} \hat{C}_S = \frac{\lambda_u' \Sigma_S^{1/2} \kappa_S}{\kappa_S' \Sigma_S \kappa_S}$. Therefore by lemma **??** and Vitali's convergence theorem

$$
\begin{aligned}
&dlim\ \hat{R}_w(S) = \lambda_u' M_{\Sigma_S^{1/2}\kappa_S} \lambda_u \ \Rightarrow\ lim\ \mathbb{E}(\hat{R}_w(S)) = (s-1)\sigma_u^2 \\
&dlim\ \hat{R}_o(S) = \lambda_u' (I_s + P_{\Sigma_S^{1/2}\kappa_S}) \lambda_u + o_P(1) \ \Rightarrow\ lim\ \mathbb{E}(\hat{R}_o(S)) = (s+1)\sigma_u^2
\end{aligned}
$$

- If $a_S = 1/2$ then

$$
\begin{aligned}
&dlim\ \frac{\hat{C}_S^2}{n} v_S' z_S \Sigma_S^{-1} z_S v_S = C_{S,w}^2 \lambda_{vS}' \lambda_{vS} \\
&dlim\ n^{1-2a_S} \frac{\hat{C}_S^2}{n^2} \kappa_S' z_S' z_S \Sigma_S^{-1} z_S' z_S \kappa_S = C_{S,w}^2 \kappa_S' \Sigma_S \kappa_S \\
&dlim\ 2n^{1/2-a_S} \frac{\hat{C}_S^2}{n^{3/2}} v_S' z_S \Sigma_S^{-1} z_S' z_S \kappa_S = 2C_{S,w}^2 \lambda_{vS}' \Sigma_S^{1/2} \kappa_S \\
&dlim\ -2\frac{\hat{C}_S}{n} v_S' z_S \Sigma_S^{-1} z_S' u = -2C_{S,w} \lambda_{vS}' \lambda_u \\
&dlim\ 2n^{1/2-a_S} \frac{\hat{C}_S}{n^{3/2}} \kappa_S' z_S' z_S \Sigma_S^{-1} z_S' u = -2C_{S,w} \kappa_S' \Sigma_S^{1/2} \lambda_u
\end{aligned}
$$

As a consequence by SL

$$
\begin{aligned}
dlim\ \hat{R}_w(S) &= \lambda_u' \lambda_u + C_{S,w}^2 \tilde{\lambda}_{vS}' \tilde{\lambda}_{vS} - 2C_{S,w} \tilde{\lambda}_{vS}' \lambda_u = \lambda_u' M_{\tilde{\lambda}_{vS}} \lambda_u \\
dlim\ \hat{R}_o(S) &= \lambda_u' \lambda_u + C^{*2}_{S,w} \tilde{\lambda}_{vS}' \tilde{\lambda}_{vS} - 2C^*_{S,w} \tilde{\lambda}_{vS}' \lambda_u
\end{aligned}
$$

Thus by lemma **??** and Vitali's convergence theorem if **??**(ii) holds then by lemma

3.1

$$lim\ \mathbb{E}(\hat{R}_w(S)) = \sigma_u^2 s - \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right) - \frac{\rho^2}{\sigma_{vS}^2}\mathbb{E}(||\tilde{\lambda}_{vS} - P_{\tilde{\lambda}_{vS}}\Sigma_S^{1/2}\kappa_S||^2)$$

$$lim\ \mathbb{E}(\hat{R}_o(S)) = \sigma_u^2 s + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\mathbb{E}(||\tilde{\lambda}_{vS}||^2)\mathbb{E}(||\tilde{\lambda}_{vS}||^{-2}) + \frac{\rho^2}{\sigma_{vS}^4}(\kappa_S'\Sigma_S\kappa_S - s\sigma_{vS}^2)$$
$$+ \frac{\rho^2}{\sigma_{vS}^4}\left(\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}\tilde{\lambda}_{vS}'}{||\tilde{\lambda}_{vS}||^4}\right)\Sigma_S^{1/2}\kappa_S\mathbb{E}(||\tilde{\lambda}_{vS}||^2) - 2\kappa_S'\Sigma_S\kappa_S\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2}\right)\right)$$

and if **??**(iii) holds then based on lemma 3.2

$$lim\ \mathbb{E}(\hat{R}_w(S)) = \sigma_u^2 s + \frac{\rho^2}{\sigma_{vS}^4}(\kappa_S'\Sigma_S\kappa_S - s\sigma_{vS}^2)$$
$$+ \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\left(\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^4}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right) - 2\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^2}{||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2}\right)\right)$$
$$+ \frac{\rho^2}{\sigma_{vS}^4}\kappa_S'\Sigma_S^{1/2}\left(\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}||\tilde{\lambda}_{vS}||^4}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right) - 2\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^2}{||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2}\right)\right)\Sigma_S^{1/2}\kappa_S$$

$$lim\ \mathbb{E}(\hat{R}_o(S)) = \sigma_u^2 s + \frac{\rho^2}{\sigma_{vS}^4}(\kappa_S'\Sigma_S\kappa_S - s\sigma_{vS}^2)$$
$$+ \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right)\mathbb{E}(||\tilde{\lambda}_{vS}||^2)$$
$$+ \frac{\rho^2}{\sigma_{vS}^4}\kappa_S'\Sigma_S^{1/2}\mathbb{E}(||\tilde{\lambda}_{vS}||^2)\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}||\tilde{\lambda}_{vS}||^2}{(||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2)^2}\right)\Sigma_S^{1/2}\kappa_S$$
$$- 2\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{\tilde{\lambda}_{vS}}{||\tilde{\lambda}_{vS}||^2 - s\sigma_{vS}^2}\right)\mathbb{E}(\tilde{\lambda}_{vS})'\Sigma_S^{1/2}\kappa_S$$

- If $a_S > 1/2$ then

$$dlim\ \frac{\hat{C}_S^2}{n}v_S'z_S\Sigma_S^{-1}z_Sv_S = C_{S,vw}^2\lambda_{vS}'\lambda_{vS}$$

$$plim\ n^{1-2a_S}\frac{\hat{C}_S^2}{n^2}\kappa_S'z_S'z_S\Sigma_S^{-1}z_S'z_S\kappa_S = 0$$

$$plim\ 2n^{1/2-a_S}\frac{\hat{C}_S^2}{n^{3/2}}v_S'z_S\Sigma_S^{-1}z_S'z_S\kappa_S = 0$$

$$dlim\ -2\frac{\hat{C}_S}{n}v_S'z_S\Sigma_S^{-1}z_S'u = -2C_{S,vw}\lambda_{vS}'\lambda_u$$

$$plim\ 2n^{1/2-a_S}\frac{\hat{C}_S}{n^{3/2}}\kappa_S'z_S'z_S\Sigma_S^{-1}z_S'u = 0$$

As a consequence by SL

$$dlim\ \hat{R}_w(S) = \lambda_u'\lambda_u + C_{S,vw}^2\lambda_{vS}'\lambda_{vS} - 2C_{S,vw}\lambda_{vS}'\lambda_u = \lambda_u' M_{\lambda_{vS}}\lambda_u$$
$$dlim\ \hat{R}_o(S) = \lambda_u'\lambda_u + C^{*2}_{S,vw}\lambda_{vS}'\lambda_{vS} - 2C^*_{S,vw}\lambda_{vS}'\lambda_u$$

Thus by lemma **??** and Vitali's convergence theorem if **??**(ii) holds then by lemma 3.1

$$lim\ \mathbb{E}(\hat{R}_w(S)) = \sigma_u^2 s - \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right) - \frac{s\rho^2}{\sigma_{vS}^2}$$
$$lim\ \mathbb{E}(\hat{R}_o(S)) = \sigma_u^2 s + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\mathbb{E}(||\lambda_{vS}||^2)\mathbb{E}(||\lambda_{vS}||^{-2}) - \frac{s\rho^2}{\sigma_{vS}^2}$$

and if **??**(iii) holds then by lemma 3.2

$$lim\ \mathbb{E}(\hat{R}_w(S)) = \sigma_u^2 s - \frac{s\rho^2}{\sigma_{vS}^2} + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\left(\mathbb{E}\left(\frac{||\lambda_{vS}||^4}{(||\lambda_{vS}||^2 - s\sigma_{vS}^2)^2}\right) - 2\mathbb{E}\left(\frac{||\lambda_{vS}||^2}{||\lambda_{vS}||^2 - s\sigma_{vS}^2}\right)\right)$$
$$lim\ \mathbb{E}(\hat{R}_w(S)) = \sigma_u^2 s - \frac{s\rho^2}{\sigma_{vS}^2} + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right) s\sigma_{vS}^2\mathbb{E}\left(\frac{||\lambda_{vS}||^2}{(||\lambda_{vS}||^2 - s\sigma_{vS}^2)^2}\right)$$

□

C.2 Technical Lemmas: IV Set Rankings via Risk

Lemma 3.7

Under assumptions **??** *and* **??***(i)-(ii) then for any* (S, S') *such that* $a_S \in [0; 1/2)$, $a_{S'} = 1/2$ *and* $\frac{\sigma_u^2\sigma_{vS'}^2}{\rho^2} > s' - 1$

$$lim\ R_4(S) < lim\ R_4(S')$$

and for any (S, S') *such that* $a_S = 1/2$, $a_{S'} > 1/2$ *and* $\frac{s-2+\frac{\kappa_S'\Sigma_S\kappa_S}{\sigma_{vS}^2}}{s'-2} > 1$ *then*

$$lim\ R_4(S) < lim\ R_4(S')$$

Under assumptions **??**, **??***(i), and* **??***(iii) then for any* (S, S', S'') *such that* $a_S \in [0; 1/2)$, $a_{S'} = 1/2$ *and* $a_{S''} > 1/2$

$$lim\ R_4(S) < lim\ R_4(S') < lim\ R_4(S'')$$

Proof. To prove the 1st statement I just need to find the conditions under which

$$\frac{\rho^2}{\sigma_{vS}^2}\left(\kappa_S'\Sigma_S^{1/2}\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS}}}{||\tilde{\lambda}_{vS}||^2}\right)\Sigma_S^{1/2}\kappa_S - 1\right) + \left(\sigma_u^2 - \frac{\rho^2}{\sigma_{vS}^2}\right)\sigma_{vS}^2\mathbb{E}(||\tilde{\lambda}_{vS}||^{-2}) > 0$$

based on the limit of $R_4(S)$ when $a_S = 1/2$, see lemma 3.5. Let $a_S < 1/2$, $a_{S'} = 1/2$ then

$$\begin{aligned} lim\ R_4(S) - \ lim\ R_4(S') = \\ &\geqslant \frac{\rho^2}{\sigma_{vS'}^2}\left(\frac{s'-2}{\mu_{S'}^2+s'-2} + (\mu_{S'}^2+s'-2)^{-1}\right) - \sigma_u^2(\mu_{S'}^2+s'-2)^{-1} \\ &\frac{\rho^2}{\sigma_{vS'}^2}\left(\frac{s'-2}{\mu_{S'}^2+s'-2} + (\mu_{S'}^2+s'-2)^{-1}\right) - \sigma_u^2(\mu_{S'}^2+s'-2)^{-1} > 0 \\ &\Leftrightarrow \frac{\rho^2}{\sigma_{vS'}^2}(s'-1) - \sigma_u^2 > 0 \Leftrightarrow \frac{\sigma_u^2\sigma_{vS'}^2}{\rho^2} < s'-1 \end{aligned}$$

□

Lemma 3.8

Under assumptions **??** *and* **??***(i)-(ii) then for any* (S, S') *such that* $a_S \in [0; 1/2)$, $a_{S'} = 1/2$, *and* $s \geqslant s'$ *or* $s < s'$ *and* $\frac{\sigma_u^2\sigma_{vS'}}{\rho^2} > \frac{s(s'-1)}{s'-s}$

$$lim\ \frac{n}{s}\mathbb{E}(R_o(S)) < \ lim\ \frac{n}{s'}\mathbb{E}(R_o(S'))$$

and for any (S, S') *such that* $a_S = 1/2$, $a_{S'} > 1/2$ *and* $s \geqslant s'$ *then*

$$lim\ \frac{n}{s}\mathbb{E}(R_o(S)) < \ lim\ \frac{n}{s'}\mathbb{E}(R_o(S'))$$

Under assumptions **??**, **??***(i), and* **??***(iii) then for any* (S, S', S'') *such that* $a_S \in [0; 1/2)$, $a_{S'} = 1/2$ *and* $a_{S''} > 1/2$

$$lim\ \frac{n}{s}\mathbb{E}(R_o(S)) < \ lim\ \frac{n}{s'}\mathbb{E}(R_o(S')) < \ lim\ \frac{n}{s''}\mathbb{E}(R_o(S''))$$

Proof. Risk by risk

- If $a_S \in [0; 1/2)$, $a_{S'} = 1/2$ then if **??**(ii) holds

$$lim\ \frac{1}{s'}R_3(S') - \ lim\ \frac{1}{s'}R_3(S) = (\sigma_u^2\sigma_{vS'}^2 - \rho^2)\mathbb{E}(||\tilde{\lambda}_{vS'}||^{-2}) + \frac{\rho^2}{\sigma_{vS'}^2}\left(\kappa'_{S'}\Sigma_{S'}^{1/2}\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS'}}}{||\tilde{\lambda}_{vS'}||^2}\right)\Sigma_{S'}^{1/2}\kappa_{S'} - 1\right)$$
$$\leqslant (\sigma_u^2\sigma_{vS'}^2 - \rho^2)\mathbb{E}(||\tilde{\lambda}_{vS'}||^2)^{-1} + \frac{\rho^2}{\sigma_{vS'}^2}\left(\kappa'_{S'}\Sigma_{S'}\kappa_{S'}\mathbb{E}(||\tilde{\lambda}_{vS'}||^2)^{-1} - 1\right)$$
$$= \frac{\sigma_u^2\sigma_{vS'}^2 - \rho^2}{\kappa'_{S'}\Sigma_{S'}\kappa_{S'} + s'\sigma_{vS'}^2} - \frac{\rho^2 s'}{\kappa'_{S'}\Sigma_{S'}\kappa_{S'} + s'\sigma_{vS'}^2}$$
$$= \frac{\sigma_u^2\sigma_{vS'}^2 - \rho^2(s'+1)}{\kappa'_{S'}\Sigma_{S'}\kappa_{S'} + s'\sigma_{vS'}^2}$$

because $P_{\tilde{\lambda}_{vS'}} \leqslant I_{s'}$, using Jensen inequality and because $\mathbb{E}(||\tilde{\lambda}_{vS'}||^2) = \kappa'_{S'}\Sigma_{S'}\kappa_{S'} + s'\sigma_{vS'}^2$. Therefore

$$\frac{\sigma_u^2\sigma_{vS'}^2}{\rho^2} < s' + 1 \Rightarrow\ lim\ \frac{1}{s'}R_3(S') <\ lim\ \frac{1}{s}R_3(S)$$

If **??**(iii) holds

$$lim\ \frac{1}{s'}R_3(S') - \ lim\ \frac{1}{s}R_3(S) = \frac{\rho^2}{\sigma_{vS'}^2}\left(\kappa'_{S'}\Sigma_{S'}^{1/2}\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS'}}||\tilde{\lambda}_{vS'}||^2}{(||\tilde{\lambda}_{vS'}||^2 - s'\sigma_{vS'}^2)^2}\right)\Sigma_{S'}^{1/2}\kappa_{S'} - 1\right)$$
$$+ (\sigma_u^2\sigma_{vS'}^2 - \rho^2)\mathbb{E}\left(\frac{||\tilde{\lambda}_{vS'}||^2}{(||\tilde{\lambda}_{vS'}||^2 - s'\sigma_{vS'}^2)^2}\right) > 0$$

Additionally if $a_{S''} > 1/2$ then if **??**(ii) holds then

$$lim\ \frac{1}{s''}R_3(S'') - \ lim\ \frac{1}{s'}R_3(S') = (\sigma_u^2\sigma_{vS'}^2 - \rho^2)\left(\mathbb{E}(||\lambda_{vS''}||^{-2}) - \mathbb{E}(||\tilde{\lambda}_{vS'}||^{-2})\right)$$
$$- \frac{\rho^2}{\sigma_{vS'}^2}\left(\kappa'_{S'}\Sigma_{S'}^{1/2}\mathbb{E}\left(\frac{P_{\tilde{\lambda}_{vS'}}}{||\tilde{\lambda}_{vS'}||^2}\right)\Sigma_{S'}^{1/2}\kappa_{S'}\right) + \rho^2(\sigma_{vS'}^{-2} - \sigma_{vS''}^{-2})$$

□

C.3 Technical Lemma: Convergence of Risk Estimators

Lemma 3.9

Consider an iid sample of random variables $(X_i)_{i=1}^n$ with finite second moments and a bootstrap sample $(X_{i,b})_{i=1}^{n_c}$, then

$$plim\ \frac{1}{n_c}\sum_{i=1} X_{i,b} = \mathbb{E}(X_i)$$

Proof. Because $(X_{i,b})_{i=1}^{n_c}$ is a bootstrapped sample, conditional on $(X_i)_{i=1}^n$, $X_{i,b}$ has support $\{X_1; X_2; \ldots; X_n\}$ and for any i for any j $\mathbb{P}(X_{i,b} = X_j|(X_i)_{i=1}^n) = \frac{1}{n}$. Next note by the Law of Total Probability (LTP) that for any function f such that $\mathbb{E}(f(X_i))$ exists

$$\mathbb{E}(f(X_{i,b})) = \mathbb{E}(\mathbb{E}(f(X_{i,b})|(X_i)_{i=1}^n)) = \mathbb{E}\left(\sum_{j=1}^n \frac{1}{n} f(X_j)\right) = \mathbb{E}(f(X_i))$$

Therefore $\mathbb{E}(X_{i,b}) = \mathbb{E}(X_i)$, $\mathbb{E}(X_{i,b}^2) = \mathbb{E}(X_i^2)$, $Var(X_{i,b}) = Var(X_i)$. Furthermore, without loss of generality there exists $D_{i,b}$ whose support is $\{1; 2; \ldots; n\}$ and $\mathbb{P}(D_{i,b} = j) = \frac{1}{n}$, moreover $D_{i,b}$ is independent of $D_{j,b}$ and $(X_m)_{m=1}^n$. Thus $X_{i,b}$ can be represented as

$$X_{i,b} = \sum_{j=1}^n 1\{D_{i,b} = j\} X_j$$

where $1\{\cdot\}$ denotes the indicator function. Consequently for $i \neq j$

$$\begin{aligned}
Cov(X_{i,b}, X_{j,b}) &= \mathbb{E}(X_{i,b}X_{j,b}) - \mathbb{E}(X_{i,b})\mathbb{E}(X_{j,b}) \\
&= \mathbb{E}\left(\mathbb{E}\left(\sum_{l,k} 1\{D_{i,b} = l\}1\{D_{j,b} = k\}X_l X_k \middle| (X_m)_{m=1}^n\right)\right) - \mathbb{E}(X_i)^2 \\
&= \mathbb{E}\left(\sum_l \mathbb{E}\left(1\{D_{i,b} = l\}1\{D_{j,b} = l\}\middle|(X_m)_{m=1}^n\right) X_l^2\right) \\
&\quad + \mathbb{E}\left(\sum_{l,k\neq l} \mathbb{E}\left(1\{D_{i,b} = l\}1\{D_{j,b} = k\}\middle|(X_m)_{m=1}^n\right) X_l X_k\right) - \mathbb{E}(X_i)^2 \\
&= \frac{1}{n}\mathbb{E}(X_i^2) + \frac{n(n-1)}{n^2}\mathbb{E}(X_i)^2 - \mathbb{E}(X_i^2) \\
&= \frac{1}{n} Var(X_i)
\end{aligned}$$

As a consequence by Chebyshev's Inequality (CHI) for any $e > 0$

$$\begin{aligned}
\mathbb{P}(|\frac{1}{n_c}\sum_{i=1}^{n_c} X_{i,b} - \mathbb{E}(X_i)| > e) &\leqslant \frac{Var\left(\frac{1}{n_c}\sum_{i=1}^{n_c} X_{i,b}\right)}{e^2} \\
&= \frac{\frac{1}{n_c^2}\sum_{i=1}^{n_c} Var(X_{i,b}) + \frac{1}{n_c^2}\sum_{i,j\neq i} Cov(X_{i,b}, X_{j,b})}{e^2} \\
&= \frac{\frac{1}{n_c} Var(X_i) + \frac{n_c - 1}{n_c n} Var(X_i)}{e^2} \to 0
\end{aligned}$$

where the convergence to 0 is due to the fact that $n_c \underset{n\to+\infty}{\to} +\infty$. Therefore $plim\ \frac{1}{n_c}\sum_{i=1}^{n_c} X_{i,b} = \mathbb{E}(X_i)$. □

Lemma 3.10

Consider an iid sample of random variables $(X_i)_{i=1}^n$ with finite second moments and a bootstrapped sample $(X_{i,b})_{i=1}^{n_c}$ and let $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $\hat{\sigma}_X = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$, then

$$dlim\ \frac{1}{\sqrt{n_c}} \sum_{i=1}^{n_c} \frac{X_{i,b} - \overline{X}}{\hat{\sigma}_X} = \mathcal{N}(0,1), \qquad dlim\ \frac{1}{\sqrt{n_c}} \sum_{i=1}^{n_c} \frac{X_{i,b} - \mathbb{E}(X_i)}{\sqrt{Var(X_i)}} = \mathcal{N}(0, 1+c^2)$$

where $lim\ \sqrt{\frac{n_c}{n}} = c$. Furthermore if $\overline{X} = \mathbb{E}(X_i) = 0$ or if $lim\ \sqrt{\frac{n_c}{n}} = 0$ then

$$dlim\ \frac{1}{\sqrt{n_c}} \sum_{i=1}^{n_c} \frac{X_{i,b} - \mathbb{E}(X_i)}{\sqrt{Var(X_i)}} = \mathcal{N}(0,1)$$

Proof. Let $Y_{i,b} = \frac{X_{i,b} - \overline{X}}{\hat{\sigma}_X}$ then based on the proof of lemma 3.9 notice that

$$Y_{i,b} = \sum_{j=1}^{n} 1\{D_{i,b} = j\} Y_j$$

where $D_{i,b}$ has support $\{1; \ldots; n\}$, for any j, $\mathbb{P}(D_{i,b} = j) = \frac{1}{n}$ and $D_{i,b}$ is independent of $D_{j,b}$ and $(Y_m)_{m=1}^n$. Thus conditionally on $(X_m)_{m=1}^n$, $(Y_{i,b})_{i=1}^{n_c}$ is iid. Consequently for any $t \in \mathbb{R}$

$$\begin{aligned}
\mathbb{E}(exp(it\hat{S})) &= \mathbb{E}\left(exp\left(i\frac{t}{\sqrt{n_c}} \sum_{i=1}^{n_c} Y_{i,b} \right) \right) \\
&= \mathbb{E}\left(\mathbb{E}\left(exp\left(i\frac{t}{\sqrt{n_c}} Y_{i,b} \right) \middle| (X_m)_{m=1}^n \right)^{n_c} \right) \\
&\equiv \mathbb{E}\left(\mathbb{E}\left(\varphi\left(\frac{t}{\sqrt{n_c}} \right) \middle| (X_m)_{m=1}^n \right)^{n_c} \right)
\end{aligned}$$

Next because $\frac{t}{\sqrt{n_c}} \to 0$ by Taylor's theorem

$$\begin{aligned}
\varphi\left(\frac{t}{\sqrt{n_c}} \right) &= \varphi(0) + \varphi'(0)\frac{t}{\sqrt{n_c}} + \varphi''(0)\frac{t^2}{2n_c} + o\left(\frac{t^2}{n_c} \right) \\
&= 1 + iY_{i,b}\frac{t}{\sqrt{n_c}} - Y_{i,b}^2\frac{t^2}{2n_c} + o\left(\frac{t^2}{2n_c} \right)
\end{aligned}$$

Moreover reusing arguments of the proof of lemma 3.9

$$\mathbb{E}(Y_{i,b} | (X_m)_{m=1}^n) = \overline{Y} = 0, \quad \mathbb{E}(Y_{i,b}^2 | (X_m)_{m=1}^n) = \frac{\hat{\sigma}_X^2}{\hat{\sigma}_X^2} = 1$$

Thus

$$\mathbb{E}(exp(it\hat{S})) = \mathbb{E}\left(\left(1 - \frac{t^2}{2n_c} + o\left(\frac{t^2}{n_c}\right)\right)^{n_c}\right) = \left(1 - \frac{t^2}{2n_c}\right)^{n_c} + o\left(\frac{t^2}{n_c}\right) \to exp\left(\frac{-t^2}{2}\right)$$

which is the characteristic function of a standard normal.

Next let $\tilde{Y}_{i,b} = \frac{X_{i,b} - \mu_X}{\sigma_X}$ where $\mu_X = \mathbb{E}(X_i)$ and $\sigma_X^2 = Var(X_i)$. Then notice that

$$\begin{aligned}
\frac{1}{\sqrt{n_c}}\sum_{i=1}^{n_c}\tilde{Y}_{i,b} - \frac{1}{\sqrt{n_c}}\sum_{i=1}^{n_c}Y_{i,b} &= (\sigma_X\hat{\sigma}_X)^{-1}\frac{1}{n_c}\sum_{i=1}^{n_c}(X_{i,b}(\hat{\sigma}_X - \sigma_X) + \overline{X}\sigma_X - \mu_X\hat{\sigma}_X) \\
&= (\sigma_X\hat{\sigma}_X)^{-1}\frac{1}{\sqrt{n_c}}\sum_{i=1}^{n_c}\left((X_{i,b} - \overline{X})(\hat{\sigma}_X - \sigma_X) + (\overline{X} - \mu_X)\hat{\sigma}_X\right) \\
&= \frac{\hat{\sigma}_X - \sigma_X}{\hat{\sigma}_X\sigma_X}n_c^{-1/2}\sum_{i=1}^{n_c}(X_{ib} - \overline{X}) + \sqrt{\frac{n_c}{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - \mu_X}{\sigma_X} \\
&= o_P(1) + O_P\left(\sqrt{\frac{n_c}{n}}\right)
\end{aligned}$$

Thus if $\sqrt{\frac{n_c}{n}} \to 0$ or if $\overline{X} = \mu_X = 0$ then $\frac{1}{\sqrt{n_c}}\sum_{i=1}^{n_c}Y_{i,b}$ has the same limit as $\frac{1}{\sqrt{n_c}}\sum_{i=1}^{n_c}\tilde{Y}_{i,b}$. On the other hand if $\sqrt{\frac{n_c}{n}} \to c$ then

$$dlim\ \frac{1}{\sqrt{n_c}}\sum_{i=1}^{n_c}Y_{i,b} = \ dlim\ \left(\frac{1}{\sqrt{n_c}}\sum_{i=1}^{n_c}\tilde{Y}_{i,b} + \frac{c}{\sqrt{n}}\sum_{i=}^{n}Y_i\right) = \mathcal{N}(0, 1 + c^2)$$

□

Lemma 3.11

Consider B resampled bootstrap samples $(X_{i,b})_{i=1}^{n_c}$ of random variables $(X_i)_{i=1}^{n}$ and a resampled statistic $\hat{S}_b = K((X_{i,b})_{i=1}^{n_c})$ such that $Var(\hat{S}_b) < +\infty$, $((X_{i,b})_{i=1}^{n_c})_{b=1}^{B}$ is identically distributed across b, and $B \underset{n\to+\infty}{\to} +\infty$. Then

$$plim\ \left|\frac{1}{B}\sum_{b=1}^{B}\hat{S}_b - \mathbb{E}(\hat{S}_b|(X_i)_{i=1}^{n})\right| = 0$$

Furthermore if the B samples are independent or if for any b $\sum_{b'=1}^{B}Cov(\hat{S}_b, \hat{S}_{b'}) \leqslant \sum_{n^*=0}^{n_c}Var(\hat{S}_b)c^{n_c-n^*}$ for some $c \in (0;1)$ then

$$plim\ \left|\frac{1}{B}\sum_{b=1}^{B}\hat{S}_b - \mathbb{E}(\hat{S}_b)\right| = 0$$

Proof. By Chebyshev's inequality and using the fact that conditionally on $(X_i)_{i=1}^n$ $((X_{i,b})_{i=1}^{n_c})_{b=1}^B$ and therefore $(\hat{S}_b)_{b=1}^B$ is iid across b, for any $e > 0$

$$\begin{aligned}\mathbb{P}(|\frac{1}{B}\sum_{b=1}^{B}\hat{S}_b - \mathbb{E}(\hat{S}_b|(X_i)_{i=1}^n)| > e\big|(X_i)_{i=1}^n) &\leqslant \frac{\frac{1}{B^2}Var(\sum_{b=1}^B \hat{S}_b|(X_i)_{i=1}^n)}{e^2} \\ &= \frac{\frac{1}{B}Var(\hat{S}_b|(X_i)_{i=1}^n)}{e^2} \\ \Rightarrow \mathbb{P}(|\frac{1}{B}\sum_{b=1}^{B}\hat{S}_b - \mathbb{E}(\hat{S}_b|(X_i)_{i=1}^n)| > e) &\leqslant \frac{\frac{1}{B}\mathbb{E}(Var(\hat{S}_b|(X_i)_{i=1}^n))}{e^2} \to 0\end{aligned}$$

because $\mathbb{E}(Var(\hat{S}_b|(X_i)_{i=1}^n)) \leqslant Var(\hat{S}_b) < +\infty$. On the other hand note that

$$\sum_{b'=1}^{B} Cov(\hat{S}_b, \hat{S}_{b'}) \leqslant \sum_{n^*=0}^{n_c} Var(\hat{S}_b)c^{n_c-n^*} = Var(\hat{S}_b)c^{n_c}\left(1 + \frac{1-c^{-n_c-1}}{1-c^{-1}}\right) = Var(\hat{S}_b)\left(c^{n_c} + \frac{c^{n_c}-c^{-1}}{1-c^{-1}}\right)$$

$$\Rightarrow \frac{1}{B^2}\sum_{b,b'}^{B} Cov(\hat{S}_b, \hat{S}_{b'}) \leqslant Var(\hat{S}_b)\left(\frac{c^{n_c}}{B} + \frac{\frac{c^{n_c}}{B} - \frac{1}{Bc}}{1-\frac{1}{c}}\right) \to 0$$

Consequently if the B samples are independent then $Cov(\hat{S}_b, \hat{S}_{b'}) = 0$ so that

$$\mathbb{P}\left(\left|\frac{1}{B}\sum_{b=1}^{B}\hat{S}_b - \mathbb{E}(\hat{S}_b)\right| > e\right) \leqslant \frac{\frac{1}{B}Var(\hat{S}_b)}{e^2} \to 0$$

And if $\sum_{b'=1}^B Cov(\hat{S}_b, \hat{S}_{b'}) \leqslant \sum_{n^*=0}^{n_c} Var(\hat{S}_b)c^{n_c-n^*}$ for any (b, b') then

$$\mathbb{P}(|\frac{1}{B}\sum_{b=1}^{B}\hat{S}_b - \mathbb{E}(\hat{S}_b)| > e) \leqslant \frac{\frac{1}{B^2}\sum_{b,b'} Cov(\hat{S}_b, \hat{S}_{b'})}{e^2} \to 0$$

□

Lemma 3.12

Under assumptions A, B, and C, for $k \in \{EXO; PMSE; MSE\}$, $\mathbb{E}(\hat{R}_k(S)^2) < +\infty$ and $\hat{R}_k(S)$ is uniformly integrable.

Under the same conditions, for any S such that $b_S \geqslant 1/2$, $\mathbb{E}(n^2\hat{R}_{EXO}(S)^2) < +\infty$ and $\hat{R}_{EXO}(S)$ is uniformly integrable.

Proof. □

Lemma 3.13

Under assumptions A, B, and C, for $k \in \{EXO; PMSE; MSE\}$ and for any S

$$plim\ |\hat{R}_k(S) - R_k(S)| = 0$$

moreover under the same conditions, for any S

$$plim\ |n_c\hat{R}_{EXO}(S) - \mathbb{E}(\tilde{R}_{EXO}(S))| = 0$$

Proof. First note that for $k \in \{EXO; PMSE; MSE\}$ and for any $e > 0$

$$\begin{aligned}\mathbb{P}(|\hat{R}_k(S) - R_k(S)| > e) \leqslant \mathbb{P}(|\hat{R}_k(S) - \mathbb{E}(\hat{R}_k(S))| > e/2) \\ + \mathbb{P}(|\mathbb{E}(\hat{R}_k(S)) - R_k(S)| > e/2)\end{aligned}$$

because $\mathbb{P}(A + B > e) \leqslant \mathbb{P}(A > e/2) + \mathbb{P}(B > e/2)$ for any random variables (A, B). Next the second term on the right converges to zero because

$$lim\ |\mathbb{E}(\hat{R}_k(S)) - R_k(S)| \leqslant\ lim\ |\mathbb{E}(\hat{R}_{k,app}(S)) - \mathbb{E}(R_k(S))| + lim\ |\mathbb{E}(\hat{R}_{k,app}(S)) - R_k(S)| = 0$$

by lemma 3.4, lemma **??**, and lemma 3.5 and using the fact that (almost) sure convergence implies convergence in probability.

Next because the resampled data is identically distributed therefore $\mathbb{E}(\hat{R}_{k,b}(S))$ is the same for any b, and by lemma 3.11 whose assumptions are satisfied

$$\mathbb{P}(|\hat{R}_k(S) - \mathbb{E}(\hat{R}_k(S))| > e/2) = \mathbb{P}\left(\left|\frac{1}{B}\sum_{b=1}^{B}\hat{R}_{k,b}(S) - \mathbb{E}(\hat{R}_{k,b}(S))\right| > e/2\right) \to 0$$

Therefore for any S for $r \in \{bt; cv; oob\}$, for any $k \in \{EXO; PMSE; MSE\}$

$$plim\ |\hat{R}_k(S) - R_k(S)| = 0$$

With a similar argument, using lemma 3.6 on the limit of $\mathbb{E}(\hat{R}_w(S))$ and $\mathbb{E}(\hat{R}_o(S))$ and lemma 3.11 on the convergence of risk estimators, for any S for any $r \in \{cv; oob\}$

$$plim\ |n\hat{R}_{EXO,bt}(S) - \mathbb{E}(\hat{R}_w(S))| = 0, \qquad plim\ |n\hat{R}_{EXO,r}(S) - \mathbb{E}(\hat{R}_o(S))| = 0$$

□

C.4 Proof of Theorem 4.1

The proof is in 2 steps. First I prove that for $k \in \{EXO; PMSE; MSE\}$ for any $S \in \mathcal{S}$, $plim\ \frac{|\hat{R}_k(S) - R_k(S)|}{R_k(S)} = 0$. Second I prove that this implies $plim\ \frac{\hat{R}_k(\hat{S}_{\hat{R}_k})}{\min\limits_{S\in\mathcal{S}} R_k(S)} = 1$.

- If $k \in \{PMSE; MSE\}$ notice that for any $\tilde{\beta} \in \mathbb{R}$

$$\mathbb{E}((u^* - v^*\tilde{\beta})^2) = 0 \Leftrightarrow \sigma_u^2 + \sigma_v^2\tilde{\beta}^2 - 2\rho\tilde{\beta}$$

But discriminant $4(\rho^2 - \sigma_u^2\sigma_v^2)$ is strictly negative thus $\forall \tilde{\beta} \in \mathbb{R}\ \mathbb{E}((u^* - v^*\tilde{\beta})^2) > 0$. Hence for $k \in \{PMSE; MSE\}$, looking at the decomposition of $R_k(S)$ from section 4.1, there exists $c > 0$ such that $\forall S\ \mathbb{P}(R_k(S) \leqslant c) = 0$. Consequently for any $S \in \mathcal{S}$, for any $e > 0$

$$\mathbb{P}\left(\left|\frac{\hat{R}_k(S) - R_k(S)}{R_k(S)}\right| > e\right) \leqslant \mathbb{P}\left(|\hat{R}_k(S) - R_k(S)| > ec\right)$$

which converges to zero by lemma 3.13. If $k = EXO$ then consider instead $\tilde{R}_{EXO}(S) = \mathbb{E}\left(\mathbb{E}_n\left((y^* - x^*\hat{\beta}_S)z_S^{*'}\Sigma_S^{-1}z_S^*(y^* - x^*\hat{\beta}_S)\right)\right)$. Again there exists $c > 0$ such that $\forall S\ \mathbb{P}(R_k(S) \leqslant c) = 0$, thus

$$\mathbb{P}\left(\left|\frac{n_c\hat{R}_{EXO}(S) - \tilde{R}_{EXO}(S)}{\tilde{R}_{EXO}(S)}\right| > e\right) \leqslant \mathbb{P}\left(|n_c\hat{R}_{EXO}(S) - \tilde{R}_{EXO}(S)| > ec\right)$$

which converges to zero from lemma 3.13.

- If for any S and for $k \in \{EXO; PMSE; MSE\}$ $plim\ \frac{\hat{R}_k(S) - R_k(S)}{R_k(S)} = 0$ then $plim \max_{S \in \mathcal{S}} \frac{\hat{R}_k(S) - R_k(S)}{R_k(S)} = 0$ because $\mathcal{S}$ is finite. Then denote by $S^* = Argmin\ R_k(S)$ and notice that

$$\hat{R}_k(S^*) \geqslant R_k(S^*), \qquad \hat{R}_k(S^*) \geqslant \hat{R}_k(\hat{S}_{\hat{R}_k}), \qquad (R_k(\hat{S}_{\hat{R}_k}))^{-1} \leqslant (\alpha R_k(\hat{S}_{\hat{R}_k}) + (1-\alpha)R_k(S^*))^{-1}$$

for some $\alpha \in (0; 1)$. Therefore

$$\begin{aligned}
\frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*)}{R_k(\hat{S}_{\hat{R}_k})} &\leqslant \frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*)}{\alpha R_k(\hat{S}_{\hat{R}_k}) + (1-\alpha)R_k(S^*)} \leqslant \frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*) + \hat{R}_k(S^*) - \hat{R}_k(\hat{S}_{\hat{R}_k})}{\alpha R_k(\hat{S}_{\hat{R}_k}) + (1-\alpha)R_k(S^*)} \\
&\leqslant \frac{|R_k(\hat{S}_{\hat{R}_k}) - \hat{R}_k(\hat{S}_{\hat{R}_k})| + |\hat{R}_k(S^*) - R_k(S^*)|}{\alpha R_k(\hat{S}_{\hat{R}_k}) + (1-\alpha)R_k(S^*)} \\
&\leqslant \frac{2}{\alpha} \max_{S \in \mathcal{S}} \frac{|\hat{R}_k(S) - R_k(S)|}{R_k(S)} \xrightarrow{\mathbb{P}} 0
\end{aligned}$$

On the other hand because $R_k(\hat{S}_{\hat{R}_k}) \geqslant R_k(S^*)$, $\frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*)}{R_k(\hat{S}_{\hat{R}_k})} \geqslant 0$ thus

$$\frac{R_k(\hat{S}_{\hat{R}_k}) - R_k(S^*)}{R_k(\hat{S}_{\hat{R}_k})} = 1 - \frac{R_k(S^*)}{R_k(\hat{S}_{\hat{R}_k})} \xrightarrow{\mathbb{P}} 0$$

By the CMT this implies $\frac{R_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} \xrightarrow{\mathbb{P}} 1$ thus again by the CMT

$$\frac{\hat{R}_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} = \frac{R_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} + \frac{\hat{R}_k(\hat{S}_{\hat{R}_k}) - R_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} = \frac{R_k(\hat{S}_{\hat{R}_k})}{R_k(S^*)} + o_P(1) \xrightarrow{\mathbb{P}} 1$$

C.5 Proof of Theorem 4.2

For $k \in \{EXO; PMSE; MSE\}$, from lemma **??**, lemma **??** and lemma **??** on the rankings between IV sets based on different expected risks, if there exists some $S \in \mathcal{S}_c$ then for any $S \in \mathcal{S}_c$, for any $S' \notin \mathcal{S}_c$ the inequality $lim\ R_k(S) < lim\ R_k(S')$ holds so that $lim\ R_k(S) = lim \min_{S \in \mathcal{S}} R_k(S)$. At the same time from lemma 3.13 on the convergence of risk estimators $plim \min_S \hat{R}_k(S) = lim \min_S R_k(S)$. As a consequence $plim\ \hat{S}_{\hat{R}_k} = plim\ Argmin_{S \in \mathcal{S}} \hat{R}_k(S) = lim\ Argmin_{S \in \mathcal{S}} R_k(S)$ and at the limit $Argmin_{S \in \mathcal{S}} R_k(S)$ belongs to $\mathcal{S}_{cv}$ (surely), therefore $\hat{S}_{\hat{R}_k}$ belongs to $\mathcal{S}_{cv}$ with probability one at the limit.

Using the same argumentation and the fact that from lemma **??**, lemma **??** and lemma **??** for $k \in \{EXO; PMSE; MSE\}$ if there exists some $S \in \mathcal{S}_{an}$ then for any $S \in \mathcal{S}_{an}$, for any $S' \notin \mathcal{S}_{an}$ the inequality $lim\ R_k(S) < lim\ R_k(S')$ holds so that $\hat{S}_{\hat{R}_k}$ belongs to $\mathcal{S}_{an}$ with probability one at the limit.

Similarly if there exists some $S \in \mathcal{S}_r$ then $\hat{S}_{\hat{R}_k}$ belongs to $\mathcal{S}_r$ with probability one at the limit unless $\hat{R}_k$ has been normalised as presented in section 3.1.

D Additional Theoretical Results

In this section are formal results of statements made in the paper. First are formal characterizations of the target parameter sets from section 2.3. Second are formal proofs that the "true models" from section 3.1 are well defined and well characterized by conditions on a_S and b_S. Third is the step-by-step decomposition of the risks from section 4.1. Fourth I provide different assumptions for the consistency of the risk estimators using k-class estimators instead of 2SLS. Refer to appendix C for notations and conventions.

D.1 IV Estimation Target Sets

In this subsection I characterize the target parameter set in case the objective behind IV estimation is minimizing the weighted sum of exclusion restrictions or minimizing the mean squared error after projection of the endogenous variable on the IVs from section 2.3.

First let $\beta_S = \underset{\beta \in \mathcal{S}}{Argmin}\ \mathbb{E}((y_i - x_i\tilde{\beta})z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta}))$. It turns out that the target space is the whole real line if the IVs being considered are irrelevant, a pseudo true value if the IVs considered are endogenous but not irrelevant, and the true causal effect β in case the IVs are relevant and exogenous. This is summarized in the following proposition.

Proposition 4.1
Let $\beta_S = \underset{\beta \in \mathcal{S}}{Argmin}\ \mathbb{E}((y_i - x_i\tilde{\beta})z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta}))$ *then assuming that* $(y_i, x_i, z_{iS})_{i=1}^n$ *is iid such that (2.4) and (2.5) hold*

$$\begin{aligned}
&\beta_S = \mathbb{R} && \text{if} \quad \pi_S = 0 \\
&\beta_S = \beta + (\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha && \text{if} \quad \pi_S \neq 0 \\
&\beta_S = \beta && \text{if} \quad \pi_S \neq 0 \quad \text{and} \quad \mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = 0
\end{aligned}$$

Proof. The objective function can be decomposed in the following way

$$\begin{aligned}
\Omega(\tilde{\beta}) &\equiv \mathbb{E}((y_i - x_i\tilde{\beta})z'_{iS})\Sigma^{-1}\mathbb{E}(z_{iS}(y_i - x_i\tilde{\beta})) \\
&= \mathbb{E}((u_i + z'_{i\overline{E}}\alpha + x_i(\beta - \tilde{\beta}))z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}(u_i + z'_{i\overline{E}}\alpha + x_i(\beta - \tilde{\beta}))) \\
&= \alpha'\mathbb{E}(z_{i\overline{E}}z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha + 2(\beta - \tilde{\beta})\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha + (\beta - \tilde{\beta})^2\pi'_S\Sigma_S\pi_S
\end{aligned}$$

Then case by case

- If $\pi_S = 0$ then

$$\Omega(\tilde{\beta}) = \alpha'\mathbb{E}(z_{i\overline{E}}z'_{iS})\Sigma_S^{-1}\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha \Rightarrow\ \beta_S = \underset{\tilde{\beta}}{Argmin}\ \Omega(\tilde{\beta}) = \mathbb{R}$$

- If $\pi_S \neq 0$ then taking the FOC yiels

$$\beta_S : -2\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha - 2(\beta - \beta_S)\pi'_S\Sigma_S\pi_S = 0 \Leftrightarrow \beta_S = \beta + (\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha$$

- If $\pi_S \neq 0$ and $\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = 0$ then using the result in case $\pi_S \neq 0$

$$\beta_S = \beta$$

□

On the other hand if $\beta_S = \underset{\beta \in \mathcal{S}}{Argmin}\ \mathbb{E}((y_i - z'_{iS}\pi_S\tilde{\beta})^2)$ then the parameter target space turns out to be exactly the same, see the following proposition.

Proposition 4.2
Let $\beta_S = \underset{\beta \in \mathcal{S}}{Argmin}\ \mathbb{E}((y_i - z'_{iS}\pi_S\tilde{\beta})^2)$ *then assuming that* $(y_i, x_i, z_{iS})_{i=1}^n$ *is iid such that (2.4) and (2.5) hold*

$$\begin{aligned}
&\beta_S = \mathbb{R} && \text{if} \quad \pi_S = 0 \\
&\beta_S = \beta + (\pi'_S\Sigma_S\pi_S)^{-1}\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha && \text{if} \quad \pi_S \neq 0 \\
&\beta_S = \beta && \text{if} \quad \pi_S \neq 0 \quad \text{and} \quad \mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = 0
\end{aligned}$$

Proof. The objective function can be decomposed in the following way

$$\begin{aligned}
\Omega(\tilde{\beta}) &\equiv \mathbb{E}((y_i - z'_{iS}\pi_S\tilde{\beta})^2) \\
&= \mathbb{E}((u_i + z'_{i\overline{E}}\alpha + v_i\beta + z'_{iS}\pi_S(\beta - \tilde{\beta}))^2) \\
&= \mathbb{E}((u_i + v_i\beta)^2) + 2(\beta - \tilde{\beta})\pi'_S\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha + (\beta - \tilde{\beta})^2\pi'_S\Sigma_S\pi_S
\end{aligned}$$

Note that except for the first component which doesn't depend on $\tilde{\beta}$ the criterion has exactly the same decomposition as in the criterion considered in proposition 4.1 and therefore the same solutions. □

D.2 True Models in the Linear IV Context

In this subsection I prove that the conditions given in section 3.1 on the level of strength a_S and the level of endogeneity b_S of the IVs in the sets $\mathcal{S}_{id}$, $\mathcal{S}_{cv}$, $\mathcal{S}_{an}$, and $\mathcal{S}_r$ are right. As long as there exists some set S such that it is exogenous and relevant then β is identified. If there is some S such that the IVs are not weak and their endogeneity level is sufficiently low relative to their strength level then 2SLS will converge. If there is some S such that the IVs are not weak and their endogeneity level is low then 2SLS will be asymptotically normal in the sense that a standard t-test confidence interval will have nominal coverage asymptotically. If there is some S such that endogeneity is sufficiently low then there exists a valid inference procedure for β. This is summarized in the following proposition.

Proposition 4.3
Assuming that $(y_i, x_i, z_i)_{i=1}^n$ *is iid such that (2.1) and (2.2) hold where for any* $S \in \mathcal{S}$, $\pi_S = n^{-a_S}\kappa_S$, $\mathbb{E}(z_{iS}z'_{i\overline{E}})\alpha = n^{-b_S}\delta_S$, $\kappa_S \in \mathbb{R}_*^S$ *is fixed and* $\delta_S \in \mathbb{R}_*^S$ *is fixed then*

- *β is identified if $\mathcal{S}_{id} \neq \emptyset$*
- *There exists some S such that plim $\hat{\beta}_S = \beta$ if $\mathcal{S}_c \neq \emptyset$*
- *There exists some S such that $\frac{\hat{\beta}_S - \beta}{\sqrt{(x' P_{z_S} x)^{-1} \hat{\sigma}_u^2}} \xrightarrow{d} \mathcal{N}(0,1)$ if $\mathcal{S}_{an} \neq \emptyset$*
- *There exists some S such that a valid inference method exists if $\mathcal{S}_r \neq \emptyset$*

Proof. Case by case:

- Identification directly follows from the fact that for some S, $\pi_S \neq 0$ and $\alpha_E = 0$. Indeed for any such S, β can be expressed as $\beta = \mathbb{E}(\omega' z_{iS} x_i)^{-1} \mathbb{E}(\omega' z_{iS} y_i)$ for some non-random vector ω such that $\mathbb{E}(\omega' z_{iS} x_i) \neq 0$. Therefore for β to be identified $\mathcal{S}_{id}$ must be non-empty.
- Recall that $\hat{C}_S \equiv \hat{\beta}_S - \beta$ then for any S such that $a_S \geqslant 1/2$

$$dlim\ \hat{C}_S \neq 0$$

which is proven in lemma 3.1 and lemma 3.2. This also prevent proper inference. Therefore β can be consistently estimated only if there is some S such that $a_S < 1/2$.

Next recall that $u_{iS} = z'_{i\overline{E}} \alpha + u_i$ where $\mathbb{E}(u_i | z_i) = 0$ and $\mathbb{E}(z_{iS} z'_{i\overline{E}}) \alpha = n^{-b_S} \delta_S$. Then if $a_S < 1/2$ with a slight abuse of the O_P notations

$$\begin{aligned}
\hat{C}_S = \frac{x' P_{z_S} u_S}{x' P_{z_S} x} &= n^{2a_S - 1} \frac{n^{-a_S} \kappa'_S z'_S u + n^{-a_S} \kappa_S z'_S z_{\overline{E}} \alpha + v' P_{z_S} u + v' P_{z_S} z_{\overline{E}} \alpha}{n^{-1} \kappa'_S z'_S z_S \kappa_S + 2 n^{a_S - 1/2} \kappa'_S z'_S v + n^{2a_S - 1} v P_{z_S} v} \\
&= n^{2a_S - 1} \frac{O_P(n^{1/2 - a_S}) + O_P(n^{1 - a_S - b_S}) + O_P(1) + O_P(n^{1/2 - b_S})}{O_P(1)} \\
&= \frac{O_P(n^{a_S - 1/2}) + O_P(n^{a_S - b_S}) + O_P(n^{2a_S - 1}) + O_P(n^{2a_S - b_S - 1/2})}{O_P(1)}
\end{aligned}$$

Consequently $\hat{C}_S = o_P(1)$ if and only if $a_S < 1/2$ and $b_S - a_S > 0$ which are the conditions which characterize the sets in $\mathcal{S}_c$.

- For the t-statistic to be asymptotically normal $\hat{\beta}_S$ must be consistent, so $a_S < 1/2$, then with a slight abuse of the O_P notations the statistic can be written as

$$\begin{aligned} t &= \frac{\hat{\beta}_S - \beta}{\sqrt{(x'P_{z_S}x)^{-1}\hat{\sigma}_u^2}} = \hat{\sigma}_u^{-1}\frac{x'P_{z_S}u_S}{\sqrt{x'P_{z_S}x}} \\ &= \hat{\sigma}_u^{-1}n^{a_S-1/2}\frac{n^{-a_S}\kappa_S'z_S'u + n^{-a_S}\kappa_S z_S'z_{\overline{E}}\alpha + v'P_{z_S}u + n^{-b_S}v'P_{z_S}z_{\overline{E}}\alpha}{\sqrt{n^{-1}\kappa_S'z_S'z_S\kappa_S + 2n^{a_S-1/2}\kappa_S'z_S'v + n^{2a_S-1}vP_{z_S}v}} \\ &= \hat{\sigma}_u^{-1}n^{a_S-1/2}\frac{O_P(n^{1/2-a_S}) + O_P(n^{1-a_S-b_S}) + O_P(1) + O_P(n^{1/2-b_S})}{O_P(1)} \\ &= \hat{\sigma}_u^{-1}\frac{O_P(1) + O_P(n^{1/2-b_S}) + O_P(n^{a_S-1/2}) + O_P(n^{a_S-b_S})}{O_P(1)} \end{aligned}$$

  Clearly $dlim\ t = \ dlim\ \hat{\sigma}_u^{-1}\frac{\frac{1}{\sqrt{n}}\kappa_S'z_S'u}{\sqrt{\frac{1}{n}\kappa_S'z_S'z_S\kappa_S}}$ if and only if $b_S > 1/2$ and $a_S < 1/2$. Finally

$$\begin{aligned} \hat{\sigma}_u^2 &= \frac{1}{n}\sum_{i=1}^n(y_i - x_i\hat{\beta}_S)^2 = \frac{1}{n}\sum_{i=1}^n u_i^2 - \frac{2\hat{C}_S}{n}\sum_{i=1}^n u_i x_i + \frac{\hat{C}_S^2}{n}\sum_{i=1}^n x_i^2 \\ &= \frac{1}{n}\sum_{i=1}^n u_i^2 + o_P(1) \end{aligned}$$

  because $a_S < 1/2$ and $b_S - a_S > 0$ imply $\hat{C}_S = o_P(1)$. Thus if $a_S < 1/2 < b_S$ then

$$dlim\ t = \ dlim\ \sigma_u^{-1}\frac{\frac{1}{\sqrt{n}}\kappa_S'z_S'\varepsilon_S}{\sqrt{\frac{1}{n}\kappa_S'z_S'z_S\kappa_S}} = \mathcal{N}(0,1)$$

- All weak-identification robust inference procedures are based on the fact that under $H_0 : \beta = \beta_0$ the statistic

$$S = \frac{(y - x\beta_0)'z_S(z_S'z_S)^{-1/2}}{\sqrt{\frac{1}{n}(y - x\beta_0)'M_{z_S}(y - x\beta_0)}}$$

  converges in distribution towards $\mathcal{N}(0, I_s)$. Consider the numerator, with an abuse of O_P notations it can be written as

$$\begin{aligned} (y - x\beta_0)'z_S(z_S'z_S)^{-1/2} &= \frac{1}{\sqrt{n}}u_S'z_S(\frac{1}{n}z_S'z_S)^{-1/2} = \frac{1}{\sqrt{n}}u'z_S(\frac{1}{n}z_S'z_S)^{-1/2} + n^{-1/2}\alpha'z_{\overline{E}}'z_S(\frac{1}{n}z_S'z_S)^{1/2} \\ &= O_P(1) + O_P(n^{1/2-b_S}) \end{aligned}$$

  Therefore if and only if $b_S > 1/2$ can the nominator converges to a Gaussian asymptotically. With similar arguments it can be proven that the denominator converges to σ_u^2 if and only if $b_S > 0$.

Consequently weak-identification robust inference can only be performed if there exists some S such that $b_S > 1/2$, ie if and only if $\mathcal{S}_r \neq \emptyset$.

□

D.3 Risks Decomposition

In this subsection I show prove the statements on the decomposition of the risks from section 4.1. Assumption A is maintained throughout the subsection.

The three risk can be decomposed into quadratic forms which depend on $z_{\overline{E}}^{*\prime}\alpha$ and $\hat{\beta}_S - \beta$. This is shown using the orthogonality between the errors (u_i, v_i) and the IVs z_i and using the independence between (y^*, x^*, z^*) and $(y_i, x_i, z_i)_{i=1}^n$, see the subset model (2.4) and (2.5).

Starting with R_{EXO}

$$\begin{aligned}
R_{EXO}(S) &= \mathbb{E}\left(\mathbb{E}_n\left((y^* - x^*\hat{\beta}_S)z_S^{*\prime}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(y^* - x^*\hat{\beta}_S)\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left((u_S^* - x^*(\hat{\beta}_S - \beta))z_S^{*\prime}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(u_S^* - x^*(\hat{\beta}_S - \beta))\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left((u^* + z_{\overline{E}}^{*\prime}\alpha - (z_S^{*\prime}\pi_S + v^*)(\hat{\beta}_S - \beta))z_S^{*\prime}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(u^* + z_{\overline{E}}^{*\prime}\alpha - (z_S^{*\prime}\pi_S + v^*)(\hat{\beta}_S - \beta))\right)\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left((z_{\overline{E}}^{*\prime}\alpha - z_S^{*\prime}\pi_S(\hat{\beta}_S - \beta))z_S^{*\prime}\right)\Sigma_S^{-1}\mathbb{E}_n\left(z_S^*(z_{\overline{E}}^{*\prime}\alpha - z_S^{*\prime}\pi_S(\hat{\beta}_S - \beta))\right)\right) \\
&= \alpha'\mathbb{E}(z_{\overline{E}}^* z_S^{*\prime})\Sigma_S^{-1}\mathbb{E}(z_S^* z_{\overline{E}}^{*\prime})\alpha + \mathbb{E}((\hat{\beta}_S - \beta)^2)\pi_S'\Sigma_S\pi_S - 2\pi_S'\Sigma_S^{-1/2}\mathbb{E}(z_S^* z_{\overline{E}}^{*\prime})\alpha \\
&= \mathbb{E}\left(||\Sigma_S^{-1/2}\mathbb{E}(z_S^* z_{\overline{E}}^{*\prime})\alpha - \Sigma_S^{1/2}\pi_S(\hat{\beta}_S - \beta)||^2\right)
\end{aligned}$$

Then with R_{PMSE}

$$\begin{aligned}
R_{PMSE}(S) &= \mathbb{E}\left(\mathbb{E}_n\left(y^* - z_S^{*\prime}\pi_S\hat{\beta}_S\right)^2\right) = \mathbb{E}\left(\mathbb{E}_n\left(x^*\beta + u_S^* - z_S^{*\prime}\pi_S\hat{\beta}_S\right)^2\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left(u^* + v^*\beta + z_{\overline{E}}^{*\prime}\alpha - z_S^{*\prime}\pi_S(\hat{\beta}_S - \beta)\right)^2\right) \\
&= \mathbb{E}\left((u^* + v^*\beta)^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*\prime}\alpha - z_S^{*\prime}\pi_S(\hat{\beta}_S - \beta)||^2\right)
\end{aligned}$$

And finally with R_{MSE}

$$\begin{aligned}
R_{MSE}(S) &= \mathbb{E}\left(\mathbb{E}_n\left(y^* - x^*\hat{\beta}_S\right)^2\right) = \mathbb{E}\left(\mathbb{E}_n\left(x^*\beta + u_S^* - x^*\hat{\beta}_S\right)^2\right) \\
&= \mathbb{E}\left(\mathbb{E}_n\left(u^* - v^*(\hat{\beta}_S - \beta) + z_{\overline{E}}^{*\prime}\alpha - z_S^{*\prime}\pi_S(\hat{\beta}_S - \beta)\right)^2\right) \\
&= \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*\prime}\alpha - z_S^{*\prime}\pi_S(\hat{\beta}_S - \beta)||^2\right)
\end{aligned}$$

Strong and Endogenous IVs In case IVs subset S is strong and endogenous, as in $a_S = b_S = 0$, the difference between the IV subset estimator and β is

$$\hat{\beta}_S - \beta = \frac{\pi_S' z_S' z_{\overline{E}} \alpha}{\pi_S' z_S' z_S \pi_S} + o_P(1) = \frac{\pi_S' \mathbb{E}(z_{iS} z_{i\overline{E}}')\alpha}{\pi_S' \Sigma_S \pi_S} + o_P(1) = \frac{\pi_S' \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha}{\pi_S' \Sigma_S \pi_S} + o_P(1)$$

The proof is omitted, a more general result is derived for the proof of the main asymptotic results in appendix C. Then the risks can be rewritten as a quadratic functions of α and $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha$.

For any $S \in \mathcal{S}$ such that $a_S = b_S = 0$ the risk R_{EXO} can be rewritten as

$$\begin{aligned}
R_{EXO}(S) &= \mathbb{E}\left(||\Sigma_S^{-1/2} \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha - \Sigma_S^{1/2} \pi_S \frac{\pi_S' \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha}{\pi_S' \Sigma_S \pi_S} ||^2 \right) + o_P(1) \\
&= \mathbb{E}\left(|| \left(\Sigma_S^{-1/2} - \frac{\Sigma_S^{1/2} \pi_S \pi_S'}{\pi_S' \mathbb{E}(z_{iS} z_{iS}') \pi_S} \right) \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha ||^2 \right) + o_P(1) \\
&= \mathbb{E}\left(|| \left(I_s - \frac{\Sigma_S^{1/2} \pi_S \pi_S' \Sigma_S^{1/2}}{\pi_S' \Sigma_S \pi_S} \right) \Sigma_S^{-1/2} \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha ||^2 \right) + o_P(1) \\
&\equiv \mathbb{E}\left(||M_{\Sigma_S^{1/2}\pi_S} \Sigma_S^{-1/2} \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha ||^2 \right) + o_P(1) \\
R_{EXO}(S) &\equiv \alpha' \mathbb{E}(z_{\overline{E}}^* z_S^{*'}) M_1 \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha + o_P(1)
\end{aligned}$$

where $M_{\Sigma_S^{1/2}\pi_S} = I_s - \Sigma_S^{1/2}\pi_S(\pi_S'\Sigma_S\pi_S)^{-1}\pi_S'\Sigma_S^{1/2}$ is the projection matrix on the space orthogonal to $\Sigma_S^{1/2}\pi_S$, and $M_1 = \Sigma_S^{-1/2} M_{\Sigma_S^{1/2}\pi_S} \Sigma_S^{-1/2} = \Sigma_S^{-1} - \pi_S'(\pi_S'\Sigma_S\pi_S)^{-1}\pi_S$ is a symmetric positive semi-definite matrix of rank $s-1$ by properties of projection matrices.

Similarly for any $S \in \mathcal{S}$ such that $a_S = b_S = 0$ R_{PMSE} can be rewritten as

$$\begin{aligned}
R_{PMSE}(S) &= \mathbb{E}\left((u^* + v^*\beta)^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2 \right) \\
&= \mathbb{E}\left((u^* + v^*\beta)^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S \frac{\pi_S' \mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha}{\pi_S' \Sigma_S \pi_S} ||^2 \right) + o_P(1) \\
&= \mathbb{E}\left((u^* + v^*\beta)^2\right) + \mathbb{E}\left(||BLOP(z_{\overline{E}}^* | z_S^{*'}\pi_S)'\alpha ||^2 \right) + o_P(1) \\
&= \mathbb{E}\left((u^* + v^*\beta)^2\right) + \alpha' \mathbb{E}\left(BLOP(z_{\overline{E}}^* | z_S^{*'}\pi_S) BLOP(z_{\overline{E}}^* | z_S^{*'}\pi_S)' \right) \alpha + o_P(1) \\
R_{PMSE}(S) &= \mathbb{E}\left((u^* + v^*\beta)^2\right) + \alpha' M_2 \alpha + o_P(1)
\end{aligned}$$

where $BLOP(z_{\overline{E}}^* | z_S^{*'}\pi_S) = z_{\overline{E}}^* - \mathbb{E}(z_{\overline{E}}^* z_S^{*'})\pi_S(\pi_S'\Sigma_S\pi_S)^{-1} z_S^{*'}\pi_S$ is the best linear projection of $z_{\overline{E}}^*$ on the space orthogonal to $z_S^{*'}\pi_S$, and $M_2 = \Sigma_E - \mathbb{E}(z_{\overline{E}}^* z_S^{*'})\pi_S(\pi_S'\Sigma_S\pi_S)^{-1}\pi_S'\mathbb{E}(z_S^* z_{\overline{E}}^{*'})$ is a symmetric positive semi-definite matrix by properties of projection matrices.

And finally for any $S \in \mathcal{S}$ such that $a_S = b_S = 0$ R_{MSE} can be rewritten as

$$R_{MSE}(S) = \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_{\overline{E}}^{*'}\alpha - z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$
$$R_{MSE}(S) = \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \alpha' M_2 \alpha + o_P(1)$$

Exogenous IVs In case the IVs subset S is exogenous, ie if $\mathbb{E}(z_S^* z_{\overline{E}}^{*'})\alpha = \alpha = 0$, the three risks can be rewritten as

$$R_{EXO}(S) = \mathbb{E}\left(||\Sigma_S^{1/2}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$
$$R_{PMSE}(S) = \mathbb{E}\left((u^* + v^*\beta)^2\right) + \mathbb{E}\left(||z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$
$$R_{MSE}(S) = \mathbb{E}\left((u^* - v^*(\hat{\beta}_S - \beta))^2\right) + \mathbb{E}\left(||z_S^{*'}\pi_S(\hat{\beta}_S - \beta)||^2\right)$$

D.4 k-class Estimators

TBC